NeuroImage 54 (2011) 875–891

Contents lists available at ScienceDirect

NeuroImage

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

Network modelling methods for FMRI

Stephen M. Smith [a,*], Karla L. Miller [a], Gholamreza Salimi-Khorshidi [a], Matthew Webster [a], Christian F. Beckmann [a,b], Thomas E. Nichols [a,c], Joseph D. Ramsey [d], Mark W. Woolrich [a,e]

[a] FMRIB (Oxford University Centre for Functional MRI of the Brain), Dept. Clinical Neurology, University of Oxford, UK
[b] Department of Clinical Neuroscience, Imperial College London, UK
[c] Departments of Statistics and Manufacturing, Warwick University, UK
[d] Department of Philosophy, Carnegie Mellon University, Pittsburgh, PA, USA
[e] OHBA (Oxford University Centre for Human Brain Activity), Dept. Psychiatry, University of Oxford, UK

ARTICLE INFO

Article history:
Received 22 May 2010
Revised 19 August 2010
Accepted 25 August 2010
Available online 15 September 2010

Keywords:
Network modelling
FMRI
Causality

ABSTRACT

There is great interest in estimating brain "networks" from FMRI data. This is often attempted by identifying a set of functional "nodes" (e.g., spatial ROIs or ICA maps) and then conducting a connectivity analysis between the nodes, based on the FMRI timeseries associated with the nodes. Analysis methods range from very simple measures that consider just two nodes at a time (e.g., correlation between two nodes' timeseries) to sophisticated approaches that consider all nodes simultaneously and estimate one global network model (e.g., Bayes net models). Many different methods are being used in the literature, but almost none has been carefully validated or compared *for use on FMRI timeseries data*. In this work we generate rich, realistic simulated FMRI data for a wide range of underlying networks, experimental protocols and problematic confounds in the data, in order to compare different connectivity estimation approaches. Our results show that in general correlation-based approaches can be quite successful, methods based on higher-order statistics are less sensitive, and lag-based approaches perform very poorly. More specifically: there are several methods that can give high *sensitivity to network connection detection* on good quality FMRI data, in particular, partial correlation, regularised inverse covariance estimation and several Bayes net methods; however, accurate estimation of *connection directionality* is more difficult to achieve, though Patel's τ can be reasonably successful. With respect to the various confounds added to the data, the most striking result was that the use of functionally inaccurate ROIs (when defining the network nodes and extracting their associated timeseries) is extremely damaging to network estimation; hence, results derived from inappropriate ROI definition (such as via structural atlases) should be regarded with great caution.

Introduction

Neuroimaging is used to study many aspects of the brain's function and structure; one area of rapidly increasing interest is the mapping of functional networks. Such mapping typically starts by identifying a set of functional "nodes", and then attempts to estimate the set of connections or "edges" between these nodes. In some cases, the directionality of these connections is estimated, in an attempt to show how information flows through the network.

There are many ways to define network nodes. In the case of electrophysiological data, the simplest approach is to either consider each recorded channel as a node, or instead use spatial sources after source reconstruction (spatial localisation) has been carried out. In the case of FMRI, nodes are often defined as spatial regions of interest (ROIs), for example, as obtained from brain atlases or from functional localiser tasks. Alternatively, independent component analysis (ICA) can be run to define independent components (spatial maps and associated timecourses), which can be considered network nodes, although the extent to which this makes sense depends on the number of components extracted (the ICA dimensionality). If a low number of components is estimated (Kiviniemi et al., 2003), then it makes more sense to think of each component itself as a network. This will often include several non-contiguous regions, all having the same timecourse according to the ICA model, and hence *within-component* network analysis is not possible without further processing, such as splitting the components and re-estimating each resulting node's timeseries. Furthermore, *between-component* network analysis is quite possibly not reasonable, as each component will in itself constitute a gross, complex functional system. However, if a higher number of components is estimated (Kiviniemi et al., 2009), these are more likely to be smaller, isolated regions (functional parcels), which can more sensibly be then considered as nodes for use in network analysis.

Once the nodes are defined, each has its own associated time-course. These are then used to estimate the connections between

* Corresponding author. Fax: +44 1865 222 717.
E-mail address: steve@fmrib.ox.ac.uk (S.M. Smith).

1053-8119/$ – see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.neuroimage.2010.08.063

nodes—in general, the more similar the timecourses are between any given pair of nodes, the more likely it is that there is a functional connection between those nodes. Of course, *correlation* (between two timeseries) does not necessarily imply either *causality* (in itself it tells one nothing about the direction of information flow), or whether the functional connection between two nodes is *direct* (there may be a third node "in-between" the two under consideration, or a third node may be feeding into the two, without a direct connection existing between them). This distinction between apparent correlation and true, direct functional connection (sometimes referred to as the distinction between *functional* and *effective* connectivity respectively; Friston, 1994) is very important if one cares about correctly estimating the network. For example, in a 3-node network where $A \rightarrow B \rightarrow C$, and with external inputs (or at least added noise that feeds around the network) for all nodes, then all three nodes' timeseries will be correlated with each other, so the "network estimation method" of simple correlation will incorrectly estimate a triangular network. However, another simple estimation method, partial correlation, can correctly estimate the true network; this works by taking each pair of timeseries in turn, and regressing out the third from each of the two timeseries in question, before estimating the correlation between the two. If B is regressed out of A and C, there will no longer be any correlation between A and C, and hence the spurious third edge of the network (A–C) is correctly eliminated.

The question of *directionality* is also often of interest, but in general is harder to estimate than whether a connection exists or not. For example, many methods, such as the two mentioned above (full correlation and partial correlation) give no directional information at all. The methods that do attempt to estimate directionality fall into three general classes. The first class is "lag-based", the most common example being Granger causality (Granger, 1969). Here it is assumed that if one timeseries looks like a time-shifted version of the other, then the one with temporal precedence *caused* the other, giving an estimation of connection directionality. The second class is based on the idea of *conditional independence*, and generally starts by estimating the (zero-lag) covariance matrix between all nodes' timeseries (hence such methods are based on the same raw measure of connectivity as correlation-based approaches—but attempt to go further in utilising this matrix to draw more complex inferences about the network). Such methods may look at the probability of pairs of variables conditional on sets of other variables; for example, Bayes net methods (Ramsey et al., 2010) in general estimate directionality by first orienting "unshielded colliders" (paths of the form $A \rightarrow B \leftarrow C$) and then drawing inferences based on algorithm-specific assumptions regarding what further orientations are implied by these colliders. The third class of methods utilises higher order statistics than just the covariance; for example, Patel's pairwise conditional probability approach (Patel et al., 2006) looks at the probability of A given B, and B given A, with asymmetry in these probabilities being interpreted as indicating causality.

A large number of network estimation methods have been used in the neuroimaging literature, with varying degrees of validation. The closer a given modality's data is to the underlying neural sources, the simpler it is to interpret the data and analyses resulting from it. In the case of FMRI, the data is a relatively indirect measure of the neural activity, being distanced from the underlying sources by many confounding stages, particularly the nonlinear neuro-vascular coupling that adds (generally unknown amounts of) significant blurring and delay to the neural signal (Buxton et al., 1998). This means that very careful validation is necessary before network estimation methods applied to FMRI data can be safely interpreted, and, unfortunately, it is too often the case that careful, sufficiently rich, validation is not carried out before real data is analysed and interpreted. Several approaches have been applied to electrophysiological data, and have been well validated for that application domain; however, because FMRI data is so much further removed from the underlying sources of interest than is generally the case with the various electrophysiological modalities, FMRI-specific validations are of particular importance. We concentrate solely on FMRI data in this paper. We simulate resting FMRI data, although the results will also in general be relevant for task FMRI (in fact, the input timings generated in the simulations could equally be viewed as simulating an event-related task FMRI experiment).

The purpose of this work is to apply a rich biophysical FMRI model to a range of network scenarios, in order to provide a thorough simulation-based evaluation of many different network estimation methods. We have compared their relative sensitivities to finding the presence of a direct network connection, their ability to correctly estimate the direction of the connection, and their robustness against various problems that can arise in real data. We find that some of the methods in common use are not effective approaches, and even can easily give erroneous results.

Methods: Simulations

Networks of varied complexity were used to simulate rich, realistic BOLD timeseries. The simulations were based upon the dynamic causal modelling (DCM; Friston et al., 2003) FMRI forward model, which uses the nonlinear balloon model (Buxton et al., 1998) for the vascular dynamics, sitting on top of a neural network model. We now describe in detail how our simulations were generated. Specific simulation parameters given are true in general for most of the evaluations, except where particular evaluations change one parameter in order to investigate its effect on network estimability—for example, in one particular evaluation the haemodynamic lag variability was removed.

Each node has an external input that is binary ("up" or "down") and generated based on a Poisson process that controls the likelihood of switching state. Neural noise/variability of standard deviation 1/20 of the difference in height between the two states is added. The mean durations of the states were 2.5 s (up) and 10 s (down), with the asymmetry representing longer average "rest" than "firing" durations; the final results did not depend strongly on these choices (for example, reducing these durations by a factor of 3 made almost no difference to the final results). These external inputs into each node can be viewed equivalently as either a *signal* feeding directly into each node, or as *noise* appearing at the neural level.

The neural signals propagate around the network using the DCM neural network model, as defined by the A network matrix:

$$\dot{z} = \sigma A z + C u \tag{1}$$

where z is the neural timeseries, $\dot{z}$ is its rate of change, u are the external inputs and C the weights controlling how the external inputs feed into the network (often just the identity matrix). The off-diagonal terms in A determine the network connections between nodes, and the diagonal elements are all set to -1, to model within-node temporal decay; thus σ controls both the within-node (neural) temporal inertia/smoothing and the neural lag between nodes.[1] The original DCM forward model includes a prior on σ that results in a mean 1 s lag between neural timeseries from directly connected nodes; this unrealistically long lag was originally coded into DCM for

[1] Although this neural model does not include neural lags as an explicit distinct process, the effect of the within-node dynamics (exponential temporal decay) is to create a lag between the input and output of every node. We verified this by generating a simple network of 4 nodes, with $A \rightarrow B \rightarrow C \rightarrow D$, and a single impulse input applied into A. Each node's output neural timeseries was indeed a blurred and delayed version of the previous node, with the delay controlled directly by σ. We also tested whether replacing the neural and haemodynamic forward model with simple shifts and linear HRF convolutions affected any of our final results, and found no significant differences. Finally, we tested that the neural lags were as expected by confirming that Granger causality, applied to the *neural* timeseries, gave correct network edge directions.

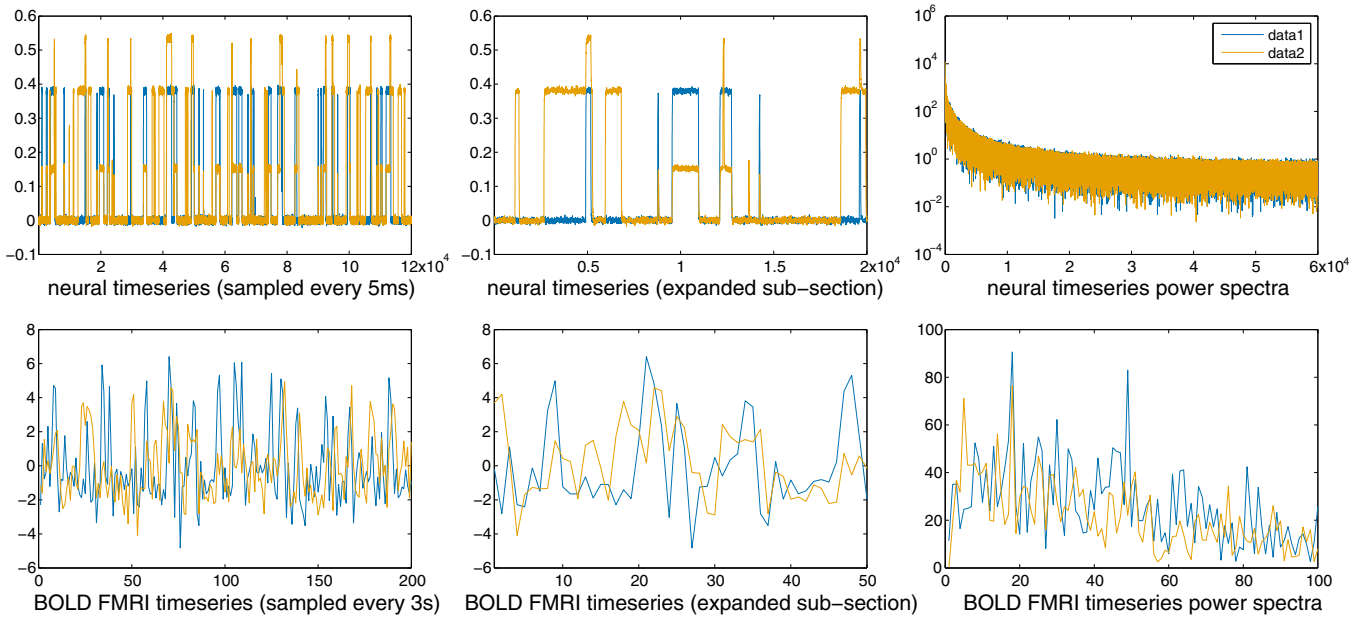

Fig. 1. Example simulated neural and FMRI BOLD timeseries, for a simple 2-node network, where node 1 feeds into node 2 with strength 0.4, and both nodes have random external inputs as described in the main text. The *y*-axis units are arbitrary. Blue shows the neural/BOLD data at node 1, and orange shows node 2.

practical algorithmic purposes in the Bayesian modelling. Although this is not a problem when DCM is applied to real data (as the data overwhelms this weak prior), it produces unrealistic lags in a simulation based on this model. Hence we changed this to a more realistic time constant, resulting in a mean neural lag of approximately 50 ms. This is chosen to be towards the upper end of the majority of neural lags generally seen,[2] in order to evaluate lag-based methods in a best-case scenario, while remaining realistic. (The reason for not also testing the lag-based methods with lower, more realistic neural lags is that, as seen below, even with a relatively long lag of 50 ms, performance of these methods is poor.)

Each node's neural timeseries was then fed through the nonlinear balloon model for vascular dynamics responding to changing neural demand. The amplitude of the neural timeseries were set so that the amount of nonlinearity (nonlinearity here being potentially with respect both to changing neural amplitude and duration) matched what is seen in typical 3 T FMRI data, and BOLD % signal change amplitudes of approximately 4% resulted (relative to mean intensity of simulated timecourses). The balloon model parameters were in general set according to the prior means in DCM. However, it is known that the haemodynamic processes vary across brain areas and subjects, resulting in different lags between the neural processes and the BOLD data, with variations of up to at least 1 s (Handwerker et al., 2004; Chang et al., 2008). We therefore added randomness into the balloon model parameters at each node, resulting in variations in HRF (haemodynamic response function) delay of standard deviation 0.5 s. This is towards the lower end of the variability reported in the literature, in order to evaluate lag-based methods in a best-case scenario while remaining reasonably realistic. Finally, thermal white noise of standard deviation 0.1–1% (of mean signal level) was added.

The BOLD data was sampled with a TR of 3 s (reduced to 0.25 s in a few simulations), and the simulations comprised 50 separate realisations (or "subjects"), all using the same simulation parameters, except for having randomly different external input timeseries, randomly different HRF parameters at each node (as described above) and (slightly) randomly different connection strengths as described below. Each "subject's" data was a 10-min FMRI session (200 timepoints) in most of the simulations. Example simulated neural and FMRI BOLD timeseries can be seen in Fig. 1, for a simple 2-node network, where node 1 feeds into node 2 with strength 0.4, and both nodes have external inputs as described above.

The main network topologies are shown in Fig. 2. The first network, *S5*, was 5 nodes in a ring (though not with cyclic causality—see arrows within the figure), with one independent external input per node, and connection strengths set randomly to have mean 0.4, standard deviation 0.1 (with maximum range limited to 0.2:0.6). *S10* took two networks like *S5*, connected via one link only (a simple "small-world" network). *S50* used 10 sub-networks, again with "small-world" topology. Each *N*-node network can also be represented as an $N \times N$ connection matrix (see examples in Fig. 2), where each element (i,j) determines the presence of a connection from node i to node j, and directed connections are represented by asymmetry in the elements—if (i,j) is nonzero and (j,i) is zero, then there is directionality from node i to node j.

Our evaluations looked at the distribution of estimated network results over the 50 simulated subjects, to estimate false-positive and false-negative rates for the various methods tested. Some aspects of our simulation framework are similar to that used to evaluate the "greedy equivalence search" (GES) network modelling method in Ramsey et al. (2010). One difference is that, whereas Ramsey et al. (2010) developed and tested methodology for explicit *cross-subject* network modelling, we concentrate here on evaluating network modelling methods for single subject (single session) datasets, and only utilise multiple subjects' datasets in order to characterise variability of results across multiple random instantiations of the same underlying network simulation. There is also a similar approach to simulated data generation in Witt and Meyerand (2009), where the DCM forward model is used to generate a simple 3 node network, with evaluation of DCM, structural equation modelling and autoregressive modelling (including Granger causality). In Marrelec et al.

[2] de Pasquale et al. (2010) note that intrahemispheric delays are typically 5–10 ms, while Ringo et al. (1994) note that unmyelinated interhemispheric fibres can result in delays of up to 300 ms, but in many long fibres are between 5 and 35 ms. Event-related potentials are often reported as being a few hundred milliseconds, but this is generally the delay between external stimulation and the response generated by a higher cognitive area; hence, this period will typically have involved communication between several distinct functional units, and is unlikely to reflect a single network "connection" that is of interest for investigation with imaging-based network modelling.

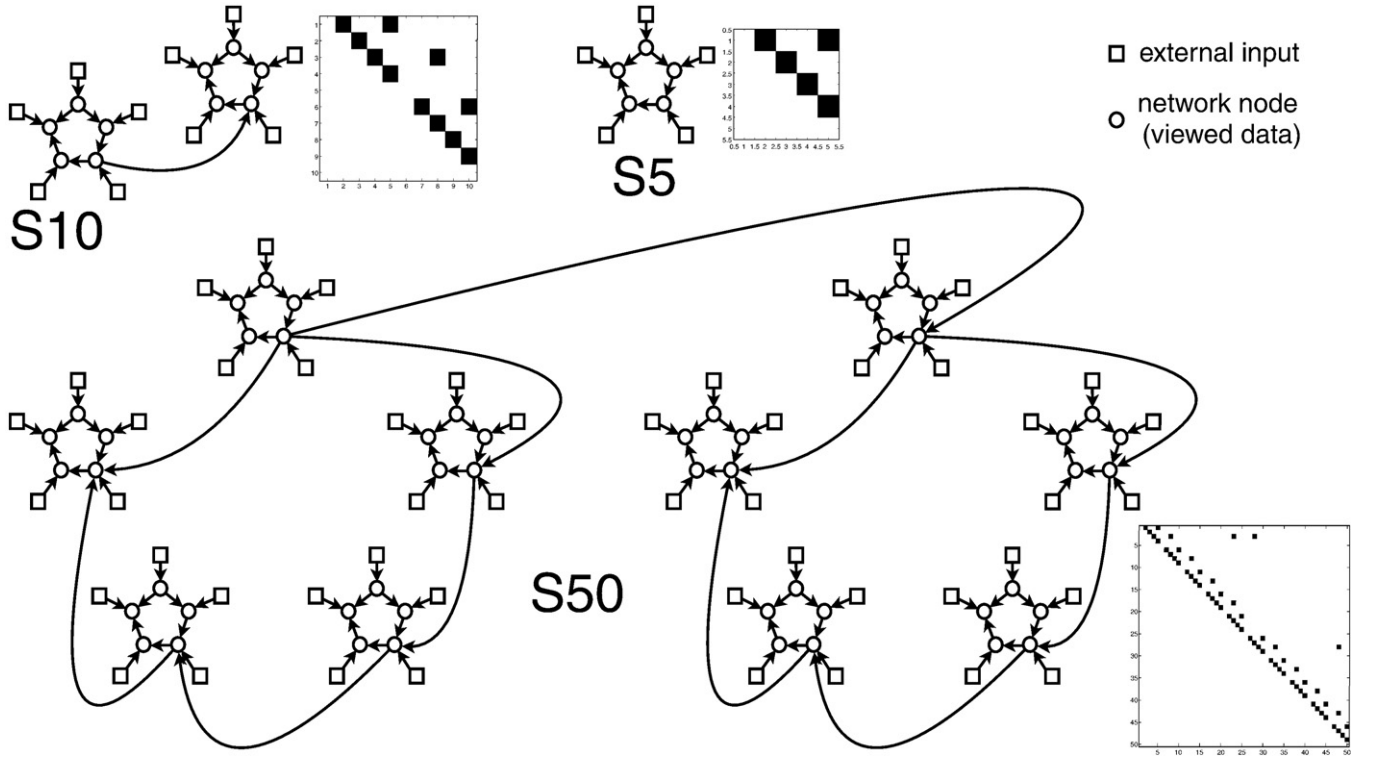

Fig. 2. The main network topologies fed into the FMRI data simulations. For each network graph the corresponding connection matrix is shown, where an element in the upper diagonal of the matrix implies a directed connection from a lower-numbered node to a higher-numbered one.

(2009) there is another somewhat similar simulation of BOLD timeseries, using a large-scale neural model of seven network nodes under different task conditions, followed by linear haemodynamic convolution; this is used to investigate the performance of partial correlation, and compare against structural equation modelling.

Methods: Network modelling methods tested

We now give a brief description of each of the methods tested. Where minor variants of each main method (including alternative choices in controlling parameters) performed universally worse than other variants, we exclude the unsuccessful variants from further consideration in the paper, in order to maximise the clarity of presentation. We describe all variants tested (including descriptions of those that were rejected) within this section.

Not tested: DCM and SEM

There are two major network modelling approaches which we have not included in this paper—dynamic causal modelling (DCM; Friston et al., 2003) and structural equation modelling (SEM; Wright, 1920; McIntosh and Gonzales-Lima, 1994). The primary reason for the exclusion of both methods is mathematical and computational feasibility—neither method is able to effectively *search* across the full range of possible network topologies, as well as both being mathematically poorly conditioned when attempting to fit the most general (unconstrained) network model to data. In general, both approaches need (at most) a few potential networks to be hypothesised and compared with their respective modelling approaches, though see Freenor and Glymour (2010) for early work on searching over DCM models.

The second reason why DCM is not appropriate here is that we are interested in modelling *resting* as well as task FMRI data. While there is some early work on stochastic DCMs that *may* be able to address this (Daunizeau et al., 2009), established DCM methods require that the "input" timings be specified in advance—something that is clearly not known for resting data.

Correlation and Partial correlation

The simplest measure of pairwise similarity between two timeseries is covariance. If the timeseries are normalised to unit variance this measure becomes (normalised) correlation, which we will refer to as *Full correlation*, to distinguish this from partial correlation.

We also evaluated full correlation, applied after bandpass filtering all timeseries data, to investigate if certain BOLD frequency bands contain more useful information for connectivity modelling. We did this because both the signal and the noise are potentially frequency-dependent (for example, the haemodynamics reduce the power in the signal at the highest frequencies). We first filtered the data keeping the lower and (separately) upper halves of the full frequency range, and also filtered the data into eight bands, each covering 1/8th of the full frequency range. Results from most frequency bands performed less well than using the unfiltered data, with network connection sensitivity becoming increasingly poorer at higher frequencies. The two frequency bands that did show some interesting results were the bottom-half of the range (*bandpass1/2*) and the second-lowest of the eight frequency bands (*bandpass2/8*).

Partial correlation refers to the normalised correlation between two timeseries, after each has been adjusted by regressing out all other timeseries in the data (all other network nodes). One attractive feature of doing this is that it attempts to distinguish direct from indirect connections, as discussed above. Partial correlation has been advocated (e.g., Marrelec et al., 2006, where light regularisation of the partial correlation matrix is applied using an uninformative prior), as being a good surrogate for SEM, which makes sense if one thinks of SEM as being a multiple regression where the data is related to itself via a network matrix, the link being that parameter estimation in a multiple regression framework is always driven by the unique (orthogonal) components of any given regressor within the model.

Regularised inverse covariance

An efficient way to estimate the full set of partial correlations is via the inverse of the covariance matrix (Marrelec et al., 2006). Under the constraint that this matrix is expected to be sparse, regularisation can be applied, for example, using the Lasso method (Banerjee et al., 2006; Friedman et al., 2008). This shrinks entries that are close to zero more than those that are not, and can be useful in the context of Bayes nets/graphical models. For example, this method (which we refer to as ICOV: Inverse COVariance) is expected to be useful when there are a limited number of observations at each node, such as with shorter FMRI scanning sessions.

If we consider a continuum of methods, with one extreme being pure pairwise methods (e.g., full correlation) and the other extreme global network modelling approaches (e.g., Bayes nets), then partial correlation (which uses all nodes' data in its calculations) sits somewhere in the middle, with ICOV also being intermediate, but a little closer to the more global modelling extreme. SEM could be thought of as being one option at the global modelling extreme, where more sophisticated/explicit modelling allows (under certain model assumptions) the estimation of model aspects such as the involvement of "latent variables" (external inputs which are not directly measured, but inferred from the viewed data).

We use an implementation of *ICOV* referred to as L1precision (www.cs.ubc.ca/~schmidtm/Software/L1precision.html), which requires the setting of the regularisation-controlling parameter λ. We tested a range of λ values: 5, 10, 20, 50, 100, and 200 (higher λ gives greater regularisation).[3] Final results showed that values of 10, 20, 50, and 200 never gave the best results, and so were discarded, keeping the values of 5 and 100. We also tested the value of 0 (no regularisation), to confirm that this gave the same results as partial correlation, which indeed was the case, and hence this was not reported on below.

Mutual information

Mutual information (MI) (Shannon, 1948) quantifies the shared information between two variables, and can reflect both linear and nonlinear dependencies. It is calculated by comparing the individual and joint histograms, and is high when one variable predicts characteristics of the other. As *MI* is sensitive to higher order statistics than is correlation (which only considers second order), this measure may be able to detect some forms of network connectivity that correlation cannot.

We used the implementation in the Functional Connectivity Toolbox (Zhou et al., 2009) (groups.google.com/group/fc-toolbox). As well as estimating mutual information, we also estimated *Partial MI*, for each pair of timecourses after all other timecourses in a given dataset were regressed out of the pair of interest.

Granger causality and related lag-based measures

Granger causality (Granger, 1969) defines a statistical interpretation of causality in which *A* is said to cause *B* if knowing the past of *A* can help predict *B* better than knowing the past of *B* alone. This is implemented using multivariate vector autoregressive modelling (MVAR). Early use of Granger causality for neuroimaging data can be found in Goebel et al. (2003) and Roebroeck et al. (2005). There has been some criticism of this approach (e.g., Friston, 2009), in part due to the lack of a biologically based generative model, but also with specific problems raised such as the likelihood of spurious estimated "causality" being in fact caused by systematic differences across brain regions in haemodynamic lag (this being a problem for FMRI, but not in general for more direct electrophysiological modalities).

We tested four implementations of Granger causality. For the first, we used the "Causal Connectivity Analysis" toolbox (Seth, 2010) (www.anilseth.com). This implements "conditional" Granger causality (Geweke, 1984), where "one variable causes a second variable if the prediction error variance of the first is reduced after including the second variable in the model, with all other variables included in both cases" (Guo et al., 2008). As this requires the specification of the "model order" (number of recent timepoints to include in the autoregressive model) we tested (separately) the use of 1, 2, 3, 10 and 20 previous observations. These results are referred to below as *Granger An* where *n* is the model order. The toolbox also allows for the estimation of "partial" Granger causality (Guo et al., 2008), which attempts to further reduce the deleterious effects of latent (unrecorded) confounding processes by making greater utilisation of the off-diagonal covariance matrix terms than the "conditional" approach does.

The second implementation that we tested was pairwise Granger causality estimation, using the Bayesian Information Criterion to estimate the model order, up to a specified maximum (www.mathworks.co.kr/matlabcentral/fileexchange/25467-granger-causality-test). We again set the maximum lags considered to 1, 2, 3, 10 and 20 (though in this case, the model order chosen by the use of BIC may well be less than the maximum allowed). This set of tests is referred to as *Granger Bn*.

The third and fourth implementations that we tested were from the BioSig toolbox (biosig.sourceforge.net): *DC* ("directed Granger causality") and *GGC* ("Geweke's Granger causality"). We tested the same range of maximum MVAR model orders as listed above, and also did this with the other lag-based methods from BioSig mentioned below. GGC can be restricted to be applied to frequencies of interest in the data; we found that the highest frequencies gave the best results and so report just a single set of results (for each MVAR model order), taken from the top end of the data frequencies.

The default Granger measure of causality for *A* causing *B* is an *F*-statistic F_{AB}, and the measure for the reverse causality is F_{BA}. In Roebroeck et al. (2005) it is suggested that a more robust variant of the Granger causality measure for *A* causing *B* is to subtract the two measures, i.e., use $F_{AB} - F_{BA}$. We therefore tested this "causality-difference" measure for all of the above Granger evaluations, in addition to the raw direction measures. (In our directionality evaluations we did in any case subtract the two directions' estimates for *all* measures, rendering this unnecessary for those tests, but that does not fully remove the value of adding in the above difference measures, to allow connection *strength* to be evaluated separately for non-differenced and differenced Granger measures.)

We also tested Partial directed coherence (PDC), which measures the "relationships (direction of information flow) between multivariate time series based on the decomposition of multivariate partial coherences computed from multivariate autoregressive models…[this reflects] a frequency-domain representation of the concept of Granger causality" (Baccalá and Sameshima, 2001). We used the implementation of PDC in the BioSig toolbox. PDC is a function of the frequencies (in the data) that are investigated; we found that all frequencies except the very highest gave identical results, and so only report one set of results (for each MVAR model order) for PDC.

Directed transfer function (DTF) is another frequency-domain method that "describes propagation between channels furnishing at the same time information about their directions and spectral characteristics. It makes possible the identification of situations where different frequency components are propagating differently" (Kamiński et al., 1997). We used the implementation in BioSig. As with PDC, DTF is a function of the frequencies investigated in the data; again we found the results almost identical across frequencies, with a

[3] The L1precision code does not enforce scale invariance for the covariance vs. λ; hence we utilised the following call to this code, in order to normalise the overall scaling of the covariance, and adjust λ accordingly. Use of this call allows the reader to interpret the λ values quoted here correctly: icov=L1precisionBCD(cov(X)/mean(diag(cov(X))), λ/1000)

slight preference for the higher frequencies, and hence only report one set of results (for each MVAR model order) for DTF. We also tested a "modified version" of DTF implemented in BioSig ("ffDTF"), but this gave the same results as DTF, so we do not report further on this method.

All of the BioSig-based measures gave equivalent or better results when run pairwise, compared with feeding in the entire sets of all nodes' timeseries; hence we only report the pairwise results below.

An extra pre-processing option for lag-based methodologies is to detrend, demean and normalise to unit temporal standard deviation all timecourses before passing them into causality estimation. We repeated all of the above tests with such pre-processing included, but found that this made no appreciable difference to any results, and so discarded those evaluations from further consideration.

From all of the Granger evaluations, using the "causality-difference" measures were never better than the raw directional measures, although in many cases they were very similar. We therefore do not report these further. From the *Granger A* tests, we found that the "partial" Granger causality evaluations were always similar to or worse than the "conditional" measures, so we do not present those in our detailed results below. The 2 and 10 model order evaluations for *Granger A* were discarded as they did not perform as well as the other choices. In the *Granger B* tests, all results were identical across different maximum model orders, implying that the BIC was always dictating a model order of 1; we thus discarded all higher-order tests. For *GGC* we kept model orders 1 and 10, as the other model orders never performed as well. All data frequencies from 0.01 to 0.1 Hz were similar, with 0.01 Hz performing slightly better, and so we kept only the *GGC* results from 0.01 Hz filtering. For *DC*, we discarded model orders of 1 and 20, as these did not perform as well as orders 2, 3, and 10. For *PDC* we discarded model orders of 2 and 20, and for *DTF*, we discarded 1, 2 and 20, in both cases because they never performed as well as other model orders.

Coherence

Two signals are said to be coherent if they have constant relative phase, or, equivalently, if their power spectra correlate, for a given time and/or frequency window. This measure is therefore insensitive to a fixed lag between two timeseries. Coherence is typically either estimated for a single (often narrow) frequency range, or estimated within multiple frequency ranges, with the multiple results then combined with each other.

We tested two implementations of coherence. For the first we used wavelet transform coherence from the Crosswavelet and Wavelet Coherence Matlab toolbox (Grinsted et al., 2004) (www.pol.ac.uk/home/research/waveletcoherence). This allows the estimation of coherence between two signals as a function of both time and frequency, and was used recently in Chang and Glover (2010) to investigate nonstationary effects in resting FMRI data. Continuous wavelet transforms are used to estimate phase-locked behaviour in two timeseries, and we average our different coherence measures over all estimated time windows. We estimate mean (over time) coherence for 0.15 Hz (close to the highest possible frequency), 0.017 Hz (close to the lowest), average over all frequencies, average over 25–50% of the frequency range, and average over the lower half of the frequency range; these are referred to as *Coherence A1* to *A5* respectively. For measures 3–5 we also estimated the 95th percentile (across the entire set of values from all times and frequencies within the ranges described) instead of the mean. We did this in the hope that we might show increased sensitivity to nonstationarities in the BOLD data (in the simulation where we generated nonstationary correlations); however *Coherence A3* (mean over all frequencies) always performed better than the other options, so the rest were discarded.

The second implementation of coherence that we used was that provided in the Functional Connectivity Toolbox mentioned above. This estimates the normalised cross-spectral density at a range of frequencies (up to Nyquist). We tested the same set of frequencies and frequency ranges as above, though we do not have separate measurements at separate timepoints (as we set the time window for spectral estimation to be equal to the timeseries length); these are referred to as *Coherence B1–5*. *Coherence B3* (mean over all frequencies) always performed better than the other options, so the rest were discarded.

Generalised synchronisation

Generalised (or nonlinear) synchronisation "evaluates synchrony by analysing the interdependence between the signals in a state space reconstructed domain" (Dauwels et al., 2010). We used the implementation available at www.vis.caltech.edu/~rodri/programs/synchro.m, which provides three related measures of nonlinear interdependence utilising generalised synchronisation; for detailed descriptions of these measures and the differences between them, see Quian Quiroga et al. (2002) and Pereda et al. (2005). These are referred to in our results as *Gen Synch S/H/N*. We found that methods *H* and *N* always gave very similar results, so we just report *H* below.

The three primary measures generated by the generalised synchronisation code are directional. In Quian Quiroga et al. (2002) there is discussion of the interpretability of asymmetries in these directional measures. It is stated that the "asymmetry can give information about driver-response relationships, but can also reflect the different dynamical properties of the data", and indeed, we did often find that the direction of the asymmetry was not consistent across the three tested synchronisation measures.

We used the default parameters of embedding dimension = 10, number of nearest neighbours = 10, Theiler correction = 50. We tested the effects of doubling and halving each of these parameters; this caused either unchanged or worse network estimation performance, so we left these default values unchanged. With respect to the time lag parameter, we tested both the default of 2, and also tested time lag = 1 (hence the numbers 2 and 1 in our results below).

Because the three measures are directional, we also averaged both directions' measures as a further test of connection strength, but this did not improve any results, and so is not reported on further. Finally, we also estimated *Gen Synch* measures for each pair of timecourses after all other timecourses in a given dataset were regressed out of the pair of interest, but this did not improve results, so we do not report further on those tests.

Patel's conditional dependence measures

The conditional dependence proposed in Patel et al. (2006) simply looks at an imbalance between $P(x|y)$ and $P(y|x)$, to arrive at a measure of connectivity/causality. This makes most sense (as a data model) when applied to fundamentally binarised data, however it can also be applied to continuous data. Here we mapped each timeseries into the range 0:1, by limiting data under the 10th percentile to 0, and data over the 90th percentile to 1, and linearly mapping data in between to the range 0:1. We then calculate the conditional dependencies directly from these "normalised" timeseries. There are two measures that can be derived from the conditional dependences: κ, a measure of connection strength, and τ, a measure of connection directionality.

As well as the results based on the continuous data, we also binarised the timeseries at a range of thresholds (0.25, 0.5, 0.75, 0.9) after mapping the data into the range 0:1 as discussed above. Each of these resulted in a separate evaluation. We found that binarisation at 0.25, 0.5 and 0.9 always performed worse or equal to binarisation at 0.75 or no binarisation, so these results were discarded.

Bayes net methods

A range of Bayes Net modelling algorithms are implemented in the Tetrad IV toolbox (www.phil.cmu.edu/projects/tetrad/tetrad4.html). We tested CCD, CPC, FCI, PC, and GES. PC ("Peter and Clark"; Meek, 1995) searches for causal graphs under the assumption that the true causal model forms a directed acyclic graph (DAG), which entails that there are no cycles and that all common causes of variables in the graph are in the graph (causal sufficiency). PC uses an efficient search for the graph's adjacencies, first computing unconditional independencies, then independencies conditional on one variable, and so on, with a specific set of rules for determining orientation. CPC (Conservative PC; Ramsey et al., 2006) uses a similar adjacency search to PC, though more conservative, limiting the number of false orientations. GES (Greedy Equivalence Search; Chickering, 2003; Ramsey et al., 2010) is a score-based search under the same assumptions as PC. It works by adding edges that most improve the score until no more edges can be added, then removing edges whose removal most improves the score, until no more edges can be removed, using the BIC (Bayesian Information Criterion) cost function. FCI (Fast Causal Inference; Zhang, 2008), unlike PC, CPC, and GES, allows for the existence of latent (or unmeasured) variables, producing a more complicated output. CCD (Cyclic Causal Discovery; Richardson and Spirtes, 2001), unlike PC, CPC, or GES, allows for the existence of cycles, also producing a more complicated output.

There is a very slight edge-direction bias in the CCD implementation that tends to direct connections from lower-numbered nodes to higher-numbered, rather than vice versa (the node "numbering" simply referring to the order in which timeseries are input to the algorithm). This slight bias only becomes apparent when the data does not support directionality strongly, and when combining results across a large number of tests. We eliminated the bias by randomising the node ordering when feeding test data into the Bayes net methods, and then undoing this reordering upon reading the results back in.

These global network modelling approaches output *binarised* network matrices, which may contain connection direction information, but generally not strength. The sensitivity of each method is determined by a single controlling input parameter for each method. However, in order to be able to test these methods within the same evaluation framework as all other methods, we ideally wanted to assign different connection strengths to different network edges in the output network matrices. To achieve this, we ran each method with approximately 100 different specificity settings (logarithmically spaced), and assigned the estimated strength of each connection to be the (−log of the) most conservative specificity (input parameter) that resulted in that element being reported as a connection. In other words, if a network edge is only reported when the modelling is run with a high-sensitivity, low-specificity (i.e., liberal) controlling parameter, that network edge is assigned a relatively low connection strength, and vice versa.

LiNGAM

The LiNGAM (Linear, Non-Gaussian, Acyclic causal Models) algorithm is a global network model that is different from Bayes net approaches in that it is not based directly on conditional independences between nodes' timeseries (typically derived via the covariance matrix), but utilises higher-order distributional statistics, via ICA, to estimate the network connections (Shimizu et al., 2006). The assumption is made that each viewed node has its own external input, and that all external inputs have distinct, non-Gaussian, distributions. Under this assumption, temporal ICA (applied to the full set of nodes' timeseries) can be used to estimate the external inputs, and the ICA "mixing matrix" can then be manipulated to estimate the network connections.

We used the original implementation of LiNGAM available from (www.cs.helsinki.fi/group/neuroinf/lingam) which includes FastICA. We used default parameters, except for using symmetric decorrelation and the "skew" nonlinearity. This latter option gave better results than other ICA nonlinearities, probably because it is simpler and hence able to function with more limited amounts of data. Because ICA requires a large number of datapoints in the relevant dimension, and LiNGAM uses *temporal* ICA, the limited number of timepoints in typical BOLD data is a potential problem for LiNGAM.

Results

We begin by explaining how we summarised the outputs from testing the different network modelling approaches. (For a summary of the specifications for all 28 simulations see Table 1.) Results from one of the most "typical" network scenarios (*Sim2*) are shown in Fig. 3, which has 10 nodes, 10 min FMRI sessions for each subject, TR = 3 s, measurement noise (thermal noise added onto the BOLD signal) of 1%, and HRF lag variability of ±0.5 s.

For some of these plots we use the raw connection strength values as estimated by each network modelling method, and for others, we have converted the connection strengths into *Z* scores, through the use of an empirical null distribution. The latter is to make the plots more qualitatively interpretable, as the connection strengths are then more comparable across the different methods. The conversion from raw connection strengths to *Z* scores (such as seen in row 1) is achieved by utilising the null distribution of connection strengths; we feed in truly null timeseries data into each of the modelling methods. The null data was created by testing for connections between timeseries from *different* subjects' datasets, which have no causal connections between them (i.e., we randomly shuffled the subject

Table 1
Summary of the 28 simulations' specifications.

Sim	# nodes	Session duration (min)	TR (s)	Noise (%)	HRF std. dev. (s)	Other factors
1	5	10	3.00	1.0	0.5	
2	10	10	3.00	1.0	0.5	
3	15	10	3.00	1.0	0.5	
4	50	10	3.00	1.0	0.5	
5	5	60	3.00	1.0	0.5	
6	10	60	3.00	1.0	0.5	
7	5	250	3.00	1.0	0.5	
8	5	10	3.00	1.0	0.5	shared inputs
9	5	250	3.00	1.0	0.5	shared inputs
10	5	10	3.00	1.0	0.5	global mean confound
11	10	10	3.00	1.0	0.5	bad ROIs (timeseries mixed with each other)
12	10	10	3.00	1.0	0.5	bad ROIs (new random timeseries mixed in)
13	5	10	3.00	1.0	0.5	backwards connections
14	5	10	3.00	1.0	0.5	cyclic connections
15	5	10	3.00	0.1	0.5	stronger connections
16	5	10	3.00	1.0	0.5	more connections
17	10	10	3.00	0.1	0.5	
18	5	10	3.00	1.0	0.0	
19	5	10	0.25	0.1	0.5	neural lag = 100 ms
20	5	10	0.25	0.1	0.0	neural lag = 100 ms
21	5	10	3.00	1.0	0.5	2-group test
22	5	10	3.00	0.1	0.5	nonstationary connection strengths
23	5	10	3.00	0.1	0.5	stationary connection strengths
24	5	10	3.00	0.1	0.5	only one strong external input
25	5	5	3.00	1.0	0.5	
26	5	2.5	3.00	1.0	0.5	
27	5	2.5	3.00	0.1	0.5	
28	5	5	3.00	0.1	0.5	

labels for each node in the network). We always generated null datasets with the same number of nodes as being tested in a given simulation, as this could affect the generation of the correct null distribution for methods that consider all nodes simultaneously. The cumulative density function of an analytical approximation to the estimated null was then used to convert raw connection strengths into *Z* scores.

The top row shows, for each modelling approach, the distribution of estimated network *Z* score values, taken from the points in the network matrix that are *true connections*. In the case of methods that estimate directionality, we use the higher of the two directions' measures to be the estimated connection strength. The distributions are over all 50 simulated "subjects" and over all correct network edges; higher is better, although this plot does not take into account the false positives—the values estimated in the network matrix that should be empty. The plots, known as "violin plots", are simply (vertically oriented) smoothed histograms, reflected in the vertical axis for better visualisation. We did not use standard boxplots because many of the distributions were not unimodal.

The second row shows the distribution of estimated network (*Z* score) strengths for the network matrix elements that should be empty, according to the ground truth network matrix. This

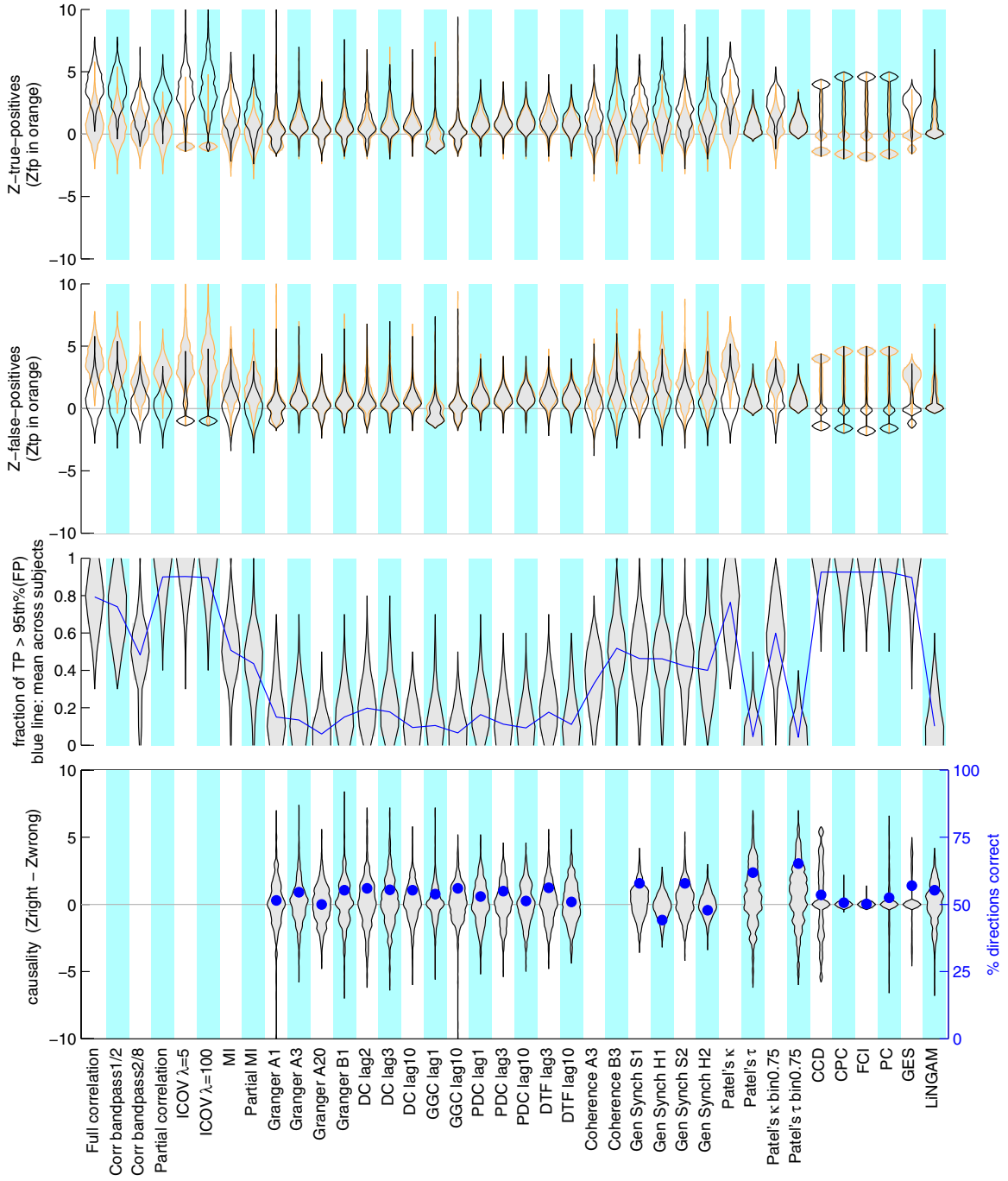

Fig. 3. Results from one of the most representative simulations, showing sensitivity to correct connection detection in the top row, sensitivity to false positives in the second, sensitivity to connection detection after correcting for false positives in the third row, and accuracy of directionality estimation in the bottom row.

distribution of "false positive" (FP) values should ideally be non-overlapping with the TP distribution, for successful methods. In the top row the TP distribution is shown in black, with the FP distribution shown in orange, for ease of comparison, and vice versa in row 2.

The fact that the FP values shown in row 2 are not all zero-mean unit-standard-deviation (as one might expect from the fact that the values have been all converted to *Z* scores using the null distribution derived from truly null data) reflects the fact that the presence of true connections will often induce increased false positive estimation (indirect connections). For example, note that the FP distribution for *Full correlation* is higher than that for *Partial correlation*, because the latter takes steps to reduce the estimation of indirect (FP) connections.

Row 3 combines the information from the top two rows—it shows the fraction of true positives that are estimated with higher connection strengths than the 95th percentile of the false positive distribution. This is therefore a measure of the success in separating the TP from the FP estimated connection strengths. The FP-based threshold is estimated separately for each subject, and the number of TP values above this counted, and then divided by the total number of true connections. The violin-plot distributions are thus showing variation across subjects. We do not need to use the *Z* scores here, but use the raw values, for greater accuracy. The blue line plots the mean of the distributions of TP fraction. As this value will be discussed many times in the following text, we shall refer to this mean fractional rate of detecting true **c**onnections as "c-sensitivity"; this is the most important, quantitative evaluation of how sensitive the different methods are to estimating the *presence* of a network connection (as opposed to its directionality).

The general story regarding c-sensitivity told by this particular simulation is reasonably reflective of our simulations in general. *Partial correlation*, *ICOV* and the Bayes net methods perform excellently, with a c-sensitivity of more than 90%. *Full correlation* and *Patel's* κ perform a little less well. *MI*, *Coherence* and *Gen Synch* are significantly less sensitive, with a c-sensitivity of about 50%. The lag-based methods (*Granger*, etc.) perform extremely poorly, with c-sensitivity of under 20%. LiNGAM also performs poorly, as it requires a larger number of timepoints to function well.

Row 4 presents results of the connection *directionality* estimation. For each true connection (positive element in the true network matrix), we took the estimated connection strength for that element (*i*,*j*) and subtracted the corresponding estimated strength in the reverse direction (*j*,*i*). The violin-plot distributions of subtracted-*Z* are over all true connections and over all subjects. A distribution showing a majority of positive values designates success in estimating directionality (causality). The blue dots indicate the overall percentage of causality directions that are correct, with 50% being the level of chance. We shall refer to this mean fractional rate of detecting the correct directionality of true connections as "d-accuracy". Some methods (e.g., *Full correlation*) only output symmetric network matrices, hence no results are shown for these methods.

Again, the general story told by this particular simulation is reasonably reflective of many of our simulations. None of the methods is very accurate, with *Patel's* τ performing best at estimating directionality, reaching nearly 65% d-accuracy, all other methods being close to chance.

We now discuss various sets of simulations, each testing a different aspect of real-world data and experimental designs. The full set of plots is included in the Supplementary Information; in the text below we summarise the relevant findings from each simulation.

Basic simulation results

We start with a set of basic experimental/data scenarios. As discussed above, *Sim2* has 10 nodes, 10 min FMRI sessions for each subject, TR = 3 s, final added noise of 1%, and HRF variability of ± 0.5 s. *Sim1* is the same, but with just 5 nodes. These two simulations are referred to frequently in the following sections, as "baselines" against which to compare the various changes we make in the network scenarios. *Sim3* is also the same, but with 15 nodes, organised in 3 clusters of 5 nodes. *Sim4* is the same, but with 50 nodes, as shown in Fig. 2.

The *Sim2* results were summarised above; in particular, *Partial correlation*, *ICOV* and the Bayes net methods perform excellently, with a c-sensitivity of above 90%, while at the other end of the spectrum, lag-based methods (*Granger*, etc.) perform extremely poorly, detecting less than 20% of true network connections. None of the methods is very accurate at estimating directionality, with *Patel's* τ performing best, reaching nearly 65% d-accuracy. The results with 5 and 15 nodes are extremely similar to those with 10.

With 50 nodes, the Bayes Net methods and *Granger A* were too computationally intensive to be practical in our testing, and so are not reported.[4] *Full correlation*, *ICOV* $\lambda = 100$ and *Patel's* κ all perform excellently, at over 90% c-sensitivity. *Full correlation* performs slightly better than *Partial correlation* (just over 80%), which is not very surprising as now the overall fraction of indirect connections is lower than with the smaller networks, aiding full correlation's results, while the number of timeseries to regress out of each pair being tested by partial correlation is much larger than with the smaller networks, which presumably removes more signal from each pair. However, with increased regularisation from the higher λ of 100, *ICOV* was able to ameliorate this, and perform better than *Full correlation*. No methods gave impressive results in estimating directionality.

Sim17 shows results from the 10 node scenario, but this time with reduced noise added to the BOLD data—just 0.1%. This reduction in amplitude of noise by a factor of 10 reflects what one might achieve by averaging timeseries over multiple voxels—in the case of thermal noise that is spatially uncorrelated, averaging over just 100 voxels would be expected to give this reduction. This might, for example, be a result of defining a spatial ROI, or by utilising one component's timecourse from ICA, which in effect is averaging timecourses over voxels present in a given component's spatial map. Somewhat surprisingly, the results are very similar to *Sim2*, suggesting that (at least in the range 0.1–1%), the exact amount of noise in the data does not significantly affect network estimation). Slight exceptions are that *MI* and *Coherence B3* have an increased c-sensitivity of 78%; clearly these measures benefit more than others from reduced noise levels, but they are still significantly less sensitive than the best methods, which have now reached as high as 97% c-sensitivity.

Effect of FMRI session length

We now consider the effect of varying the FMRI session length. *Sim5* contains 5 nodes and 60-min sessions (in reality this would not be very pleasant for the subjects, but in general is possible). Comparing this with *Sim1*, we see that most methods have improved sensitivity, with all methods except for *LiNGAM* and the majority of the lag-based methods achieving better than 80% c-sensitivity. The single lag-based result that reaches 80% is *Granger A3*; however, from the fact that the d-accuracy in this case is almost no better than chance, we must conclude that it is not the lag-based causality information that is driving this result, but simply the fact that correlation between timeseries is bleeding into the Granger causality measure (see further investigation of this effect below).

With respect to the estimation of directionality, the lag-based methods are still performing poorly with the hour-long sessions.

[4] Our tests require running each Bayes net method many thousands of times, which is why we had to exclude these from our 50-node testing; however, if only needing to be run once (i.e., on real data), the Bayes net methods are in general usable, completing on 50 nodes in less than an hour, and, judged qualitatively, performing well.

LiNGAM has now improved to 77%, with the increased number of timepoints starting to help the temporal ICA to function well. Patel's τ is largely unchanged, with all other methods still under 70% d-accuracy. We also simulated hour-long sessions for 10 nodes (*Sim6*). The results are very similar to those for 5 nodes, though *LiNGAM* is reduced slightly, to 67%.

When we increased the sessions further, to just over 4 h (*Sim7*), *LiNGAM* was able to achieve the highest d-accuracy across all methods in all of our tests (90%). *Patel's τ bin0.75* was also impressive (79%), with *CCD* at 71%.

The *Gen Synch H* directionality is well below chance (making a significant number of directionality estimates in the wrong direction). As discussed above, asymmetry in the *Gen Synch* measures is expected to be potentially driven by other factors than just the underlying causality, and this appears to be the case here.

Sim25 contains 5-min sessions, and *Sim28* has 5-min sessions with noise reduced to 0.1%. The reduced session time does not greatly affect the results, but does reduce estimation quality a little, with the best approaches still being *Partial correlation*, *ICOV* and the Bayes net methods (c-sensitivity 70–78%). With the reduced noise level, results improved towards the values seen in *Sim1* (84–89%). As with *Sim1*, directionality estimation is not very impressive from any of the methods. This pattern develops further with 2.5-min sessions; *Sim26* contains 2.5-min sessions, with the same best three methods having c-sensitivity 57-59%, and *Sim27* has 2.5-min sessions with noise 0.1%, with the same best three methods having 71–76%.

To summarise the dependence of the best methods' c-sensitivity on session duration, for the 5 nodes, 1% BOLD noise case: 60 min:100%, 10 min:95%, 5 min:77%, 2.5 min:59%.

Effect of global additive confound

The first "problem" that we introduced into the data was a global additive confound—adding the same random timeseries to all nodes' BOLD timeseries. There has been much discussion in the literature regarding the nature of the global mean timeseries (i.e., whether it is primarily valid neural-related signal, or uninteresting non-neural physiological confound) and whether it should be subtracted before any further analyses (Fox et al., 2009). The general consensus would currently appear to be that as many specific confounds as possible should be removed from the data (e.g., by estimating signal in white matter and cerebrospinal fluid (CSF), and regressing this out of all timeseries), but there is not clear agreement as to whether global timeseries removal should be carried out.

Here we investigated the effect of an additive global timeseries confound. From a real resting-FMRI dataset (36 subjects' 4D data concatenated temporally in a common space), we estimated the timeseries from approximately 100 functional parcels. The data had originally been "cleaned" through the use of confound regressors derived from CSF and white matter masks, as well as head motion parameters. Further, we discarded ICA components which were clearly artefactual. We took the mean of all 100 timecourses, and estimated the standard deviation of this, std(mean(T_i)). We then estimated the mean of the standard deviations of each individual timecourse, mean(std(T_i)). The ratio std(mean(T_i))/mean(std(T_i)) was approximately 0.3, whereas in pure random noise data it would be expected to be around $1 / \sqrt{100} = 0.1$, roughly suggesting an *upper limit* to the additive global effect of approximately 20% of the raw data standard deviation.

Initial results with a global mean confound having 20% of the amplitude of the "uncorrupted" data showed, somewhat surprisingly, almost no change in the final c-sensitivity and d-accuracy results, so we increased the confound fraction to 50%, generating *Sim10*. The results are still almost unchanged, compared with *Sim1*, suggesting that the potential presence of a global confound is not a problem. However, looking at the separate distributions of TP and FP values, we see that for some methods the distributions are significantly shifted upwards, as one would expect from adding in a *global* confound. This is particularly noticeable for *Full correlation*, not surprisingly. Because *both* TP and FP distributions are shifted upwards, the c-sensitivity results do not show a significant worsening, and this raises a potential problem in our interpretation of these results (and later use of certain network modelling methods); if we do not already know what the global confound signal is, we cannot adjust for the altered FP distribution, and hence cannot achieve the c-sensitivity results seen. We are only able, here, to estimate c-sensitivity because we already know what the ground truth is, but in real experiments we would not. However, all is not lost—we can see that with other modelling methods (in particular the methods that performed the best in the basic simulations—*Partial correlation*, *ICOV* and the Bayes net methods), there is *not* a significant shift in the raw TP and FP distributions. This is not surprising—for example, partial correlation will be expected to remove (some fraction of) the global confound from any pair of timeseries before they are correlated, because it is present in all the other timeseries (Marrelec et al., 2006). These results and considerations provide added strength to the argument for using these approaches, as opposed to (e.g.) *Full correlation*.

In those simulations that add confounding processes into the data, we did not regenerate the null distributions for the connectivity measures, but used those derived from the matched simulations without the confounds. We did this in order to make the TP and FP Z score distributions more easily comparable across simulations, but this has no effect on the quantitative measures of success (c-sensitivity and d-accuracy).

Effect of shared inputs

The next "problem" that we introduced into the data was mixing of the external inputs feeding into the network. So far we have assumed that each viewed node has its own independent external input. These inputs can be thought of as neuronal "noise", or as distinct sensory inputs, or as inputs from other parts of the brain not included in the set of viewed nodes. In the case of the latter two scenarios, there is the real possibility that an external input could feed into more than one viewed node, which could be expected to have a deleterious effect on the network modelling, if this "sharing" of inputs is not modelled (Larkin, 1971). For example, this could arise if one has imaged (or considered) only certain parts of the brain.

Sim8 simulates this problem. In the previous simulations, the external inputs feed into the viewed nodes with strength 1; in this simulation, in addition, each external input may feed into any other viewed node with strength 0.3, with the random probability of this occurring being 20% for any one possible connection between external input and viewed node. The results, compared with *Sim1*, show that the presence of shared inputs is quite deleterious to all estimation methods. *Partial correlation* and *ICOV* $\lambda = 5$ fall respectively to 69% and 67% c-sensitivity, with all other methods being below 60%. The directionality estimates, however, are largely unchanged from the *Sim1* results.

Sim9 applies the same sharing of inputs to 4-h sessions. Compared with *Sim7*, We see all methods' c-sensitivity fall, but particularly the Bayes net methods, which fall to less than 40%. However, the methods that showed the best directionality estimation results in *Sim7* (*LiNGAM* and *Patel's τ bin0.75*) still performed quite well (78% and 73% respectively).

Effect of inaccurate ROIs

The next problem we considered is the effect of mixing the BOLD timeseries with each other. Whereas the previous section discussed the mixing of *inputs*, this is in effect a mixing of *outputs* from the data simulation. This scenario would arise, for example, if the spatial ROIs

used to extract average timeseries for a brain region did not match well the actual functional boundaries. This is very likely to happen to some extent when using predefined ROIs that are not derived from the data, for example when using atlas-based ROIs.

Sim11 simulates 10 nodes, and for each node's timeseries, mixes in a relatively small amount of one other node's timeseries (randomly chosen, but the same for all subjects), in proportion 0.8:0.2. The results are extremely bad—every method gives lower than 20% c-sensitivity.

A less serious related scenario is when we have incorrect ROIs, but these just result in mixing in of signals from brain areas not already present in the set of ROIs used. For example, if the ROIs are spread sparsely across the brain and not touching each other, then incorrect ROI specification will in general not mix them together, but will mix in new, unrelated signals. *Sim12* shows the effects of mixing in unrelated timeseries into each timeseries of interest (achieved, for each subject, by using data from another subject), again in the ratio of 0.8:0.2. In this case the additional "confounds" have almost no effect at all on the network modelling, with the results looking almost identical to *Sim2*. This is not very surprising, given that we have already established that the results are not very sensitive to added noise, which is effectively all we are adding here (albeit with more temporal structure than previously).

Effect of backwards connections

It is rarely the case that two brain regions are connected in one direction only; there will generally be connections in both directions, including those with "negative" connection strengths (implying inhibition). However, it is not clear, in terms of gross, averaged behaviour of information flow around a network, whether the presence of backwards connections is in practice relevant. It is very hard also to know how to insert such connections into our simulations —for example, how strong should the backwards connections be, and should they be positive or negative? In the absence of clear answers to these questions, we chose, somewhat arbitrarily, to randomly select half of the forwards connections, and add into those a negative backwards connection of equal average strength (0.4 ± 0.1).

These results are reported in *Sim13*. Compared with *Sim1*, all the approaches that were performing well have heavily reduced sensitivity, with the best methods being the Bayes Net approaches, at 64% c-sensitivity. Interestingly, *Coherence*, *Gen Synch* and *MI* are no longer significantly worse than the correlation measures, *ICOV* and *Patel's* κ, all being around 50–55% c-sensitivity; this presumably is because the former set of methods are relatively robust against the potentially more complex relationships induced by the backward connections. No methods performed well at estimating directionality in this simulation (maximum d-accuracy being 62%), where estimated directionality is compared against the positive, "forward" connection direction —though of course, the interpretation in the cases where a backwards connection is present is difficult.

Effect of cyclic connections

Sim14 shows the results from reversing the direction of the connection between nodes 1 and 5 in the 5 node simulation. This generates "cyclic causality", which is theoretically a problem for many of the global network modelling approaches such as most of the Bayes net methods, as this breaks their modelling assumptions.

Somewhat surprisingly, compared with *Sim1*, this change makes virtually no difference to any of the c-sensitivity measures. It does reduce the d-accuracy values, although these were already low in *Sim1*.

Effect of more connections

Sim16 shows the effect of using a denser set of connections in the 5 node simulation. As well as the original 5 out of 10 possible network edges, we add in a further two more, leaving just 3 "missing" edges. This therefore simulates the scenario of a very highly connected small network. The results are very similar to *Sim1*, with a slight reduction in c-sensitivity in the best methods. There is also very little change in the d-accuracy results.

Effect of stronger connections, and investigation of lag-based directionality estimation

Sim15 shows the effect of increasing the strength of the network connections to a mean of 0.9 instead of 0.4. We set the noise to be 0.1%, so these results should be compared against *Sim17*. In this case the previously most sensitive methods (*Partial correlation*, *ICOV* and the Bayes net methods) remain relatively unchanged in c-sensitivity, falling to 90–95%. *Full correlation* and *Patel's* κ fall further, to around 60%—presumably the increased connection strengths have increased the sensitivity to detection of *indirect* connections. *MI*, *Coherence* and *Gen Synch* are relatively unchanged, although notably, *Partial MI* is increased to 85%, presumably because it is less sensitive to indirect connections than *MI*. Lag-based methods are still performing very poorly, the best reaching 34% c-sensitivity. With respect to estimating directionality, *Patel's* τ, has increased to 78%, with *Gen Synch* rising to 68%.

The lag-based methods have improved d-accuracy, with a maximum of 72% for *GGC lag-10* and *DC lag-10*. However, the fact that the *existence* of the connections was only estimated with a maximum c-sensitivity of 34% suggests that the directionality results may not be trustworthy (i.e. truly reflecting an estimation of causality based on lag). In order to investigate more interpretably whether concern here is justified, we simulated a two-node network (node 1 feeding into node 2), with otherwise the same parameters as *Sim15*. The results were consistent with the above finding; some Granger methods gave only chance-level results (e.g., *Granger A1*), but most gave a reasonable level of correct directionality estimation. However, when we reduced the neural lag to 0 and re-ran the simulation, the results were qualitatively unchanged—the Granger methods that had previously reported the "correct" causal direction continued to do so, despite the absence of any lag between the two neural timeseries. Furthermore, when we returned the lag to 50 ms, but now increased the added noise to 1.5% for the first node only, the estimated causality direction then became *negative* (i.e., the wrong direction was estimated) for all Granger methods. All of these results were replicated if we replaced the DCM forward model (including nonlinear dynamics) with a simple shift for the neural lag, followed by linear haemodynamic convolution. Unfortunately this demonstrates how the interaction of haemodynamic smoothing and measurement noise renders the lag-based methods generally unreliable for FMRI data. *Gen Synch* at model order 1 also showed the same problems as seen with Granger, but did not show this with model order 2.

Effect of having only one strong external input

Sim24 is the same as *Sim15* (strong connections and low noise), except that all nodes apart from node 1 had their external input strengths reduced from 1 to 0.1. With this reduction in external inputs and the low noise level, this simulation is passing the one primary external source around the network with very little disturbance from other sources or noise. The results become poor—none of the methods had a c-sensitivity greater than 50%, and none had a d-accuracy greater than 61%. One hopes that this particular simulation is not too representative of reality! In general every node is so highly correlated with every other node that it is hard for the methods to distinguish direct from indirect connections. Of the methods that have so far been performing the best, *Partial correlation* and *ICOV* $\lambda = 5$ do still generate amongst the best results (c-sensitivity 46–48%), but not better than

MI, *Coherence* and *Gen Synch*; the Bayes net models perform badly (27%).

Sensitivity to connection strength changes

Sim21 tests how sensitive the different methods are at detecting *changes* in connection strength across different subjects. In this case we started with the *Sim1* simulation, but halved the strength of the network connections in 25 of the 50 subjects. We then performed a two-group *t*-test on the raw TP connection strength measures. This test is of relevance to researchers wishing to discriminate between subjects on the basis of network connectivity, for example comparing patients vs. controls. While we have only evaluated "univariate" group comparisons here (i.e., testing each edge independently from each other), there is a good chance that a more sensitive general approach to network change detection will be "multivariate" (i.e., using all edges simultaneously to find *patterns* of change).

The most sensitive method was *Patel's* κ, with $t=7.4$. Other sensitive methods, with $t>5$, were *Full correlation*, *Partial correlation*, *ICOV*, *Gen Synch* and most of the Bayes net methods.

Effect of nonstationary connection strength

Sim22 and *Sim23* investigate the effect of nonstationarity of connection strength between nodes. At the neural level, there is evidence of connection strength varying over time (Popa et al., 2009; de Pasquale et al., 2010); this was investigated for FMRI data in Chang and Glover (2010).

In *Sim23*, we used 5 nodes, noise of 0.1%, strong connections (mean 0.9) and reduced strength of 0.3 for all external inputs apart from node 1. There was no nonstationarity present in this simulation. *Partial correlation* and *ICOV* perform the best, at around 80% c-sensitivity, but the Bayes net methods do not perform so well, falling to 60%. For directionality estimation, *Patel's* τ and *Gen Synch* perform reasonably, having d-accuracy of around 70%.

Sim22 is the same as *Sim23*, except that the connection strength is modulated over time by additional random processes. The strength of connection between any two connected nodes is either unaffected, or reduced to zero, according to the state of a random external bistable process that is unique for that connection. The transition probabilities of this modulating input are set such that the mean duration of interrupted connections is around 30 s and the mean time of full connections is about 20 s. The results are quite similar to *Sim22*, the main difference being that the Bayes net methods improve quite a bit, achieving the highest c-sensitivity (78%), slightly ahead of *Partial correlation* and *ICOV* (both at 73%). We had expected that perhaps one of the *Coherence* measures would show particular value (compared with other methods) when faced with nonstationarity, but this was not the case. The directionality estimates are largely unchanged, though many of the lag-based methods *reverse* the mean estimated directionality, another example of problematic inference using temporal lag with FMRI.

Effect of HRF variability and low TR

A common criticism of lag-based methods for FMRI is that any neural lag information is likely to be a) swamped by the haemodynamic smoothing and b) rendered inestimable because of variabilities in haemodynamic-induced delays. Our results so far support these views, in that there is no evidence of the lag-based methods working on realistic simulated FMRI data. In the following simulations, we attempted to push the parameters of the simulations in different ways to give the lag-based methods even better chances of succeeding (though at the risk of becoming rather unrealistic in representing real biology, experimental design and FMRI acquisitions).

Sim18 is the same as *Sim1*, except that we have removed all haemodynamic lag variability. The results are unchanged—all lag-based methods still perform very poorly both with respect to c-sensitivity and d-accuracy.

Sim19 reduces the TR to 0.25 s (currently impractical to achieve in whole-brain FMRI, but achievable for a few slices of data), sets the noise to 0.1% and increases the neural lag to 100 ms. HRF variability is set to 0.5 s. *Sim20* is identical, except that the HRF variability is removed. The results in the two cases are quite similar. The highest c-sensitivity is achieved by *Partial correlation*, *ICOV* and *GES* (95–99%), but some of the Granger approaches are close behind, achieving up to 89%, and giving impressive d-accuracy results. Because of our results described above (and because it is suspicious that including HRF delay variability that is large compared with the neural lag does not greatly affect results), we ran simpler, two-node simulations, starting with the same simulation parameters as in *Sim19* and *Sim20*. When including HRF variability, we still found that lag = 0 results gave the same "correct" causality, and that adding extra noise (as little as 0.15%) onto node 1 reversed the estimated causality direction. In these very-low-TR, low-noise, 2-node tests, it is not until we remove all HRF variability, or increase the neural lag to 0.3 s, that we start to see stronger (correct) causality estimation than the lag = 0 results, and, even in these cases, simply adding extra noise onto node 1 still reverses the estimated causality direction in the majority of tests. We also found similar general problems with *Gen Synch*.

Summary across all simulations

Table 1 summarises the specifications for all 28 simulations, and Figs. 4 and 5 show summaries of the dependence of all methods' c-sensitivity and d-accuracy on the different simulation parameters. Simulations 4 and 21 were excluded from these calculations (the former because not all methods were run in that case, and the latter because it was used to test group-difference performance rather than group-mean performance). For each modelling method, and for each "subject", a multiple regression was fit to the 28 simulations' c-sensitivity and (separately) d-accuracy values, with the different simulation parameters as model covariates. We used binary indicator variables for the cases where complex changes were applied, such as in the "global mean confound" case. Simulations 19 and 20 reduced TR and increased neural lag, so the effects of these are not separable, and the third regressor models both effects together. The model fitting was carried out separately for each of the 50 "subjects", and the parameter estimates from each regression were then summarised across subjects in terms of their effect size (mean/standard deviation). The distribution of the data is probably not Gaussian, and the relationship to certain covariates not likely to be linear, but the final summary statistic (combining across subjects) is still well-conditioned, and at least gives a semi-quantitative view of the relative strengths and signs of the dependencies. The results are mostly as one would expect; for example, session duration correlates strongly and positively with c-sensitivity for virtually all methods. For the most successful methods, session duration is more important than TR, which in turn is more important than noise level (the fact that neural lag was also varied along with the TR is not relevant in these cases). Some patterns are more complex, such as the effect of the number of nodes and the addition of a global mean confound. The effect of bad ROIs is notable as being particularly deleterious to the results from almost all methods. With respect to directionality, the most striking result is the strong dependency of the most successful methods on session duration.

We conclude with two figures that summarise the overall performances of the different network modelling methods across all simulations (except *Sim4*, which was not run for all methods). For each class of method we take the best result, for each simulation, over all variants of that method (for example, putting all the Bayes net methods together into a single class). This creates a slight bias in favour of those methods with many variants tested (although in general not in a significant way, as

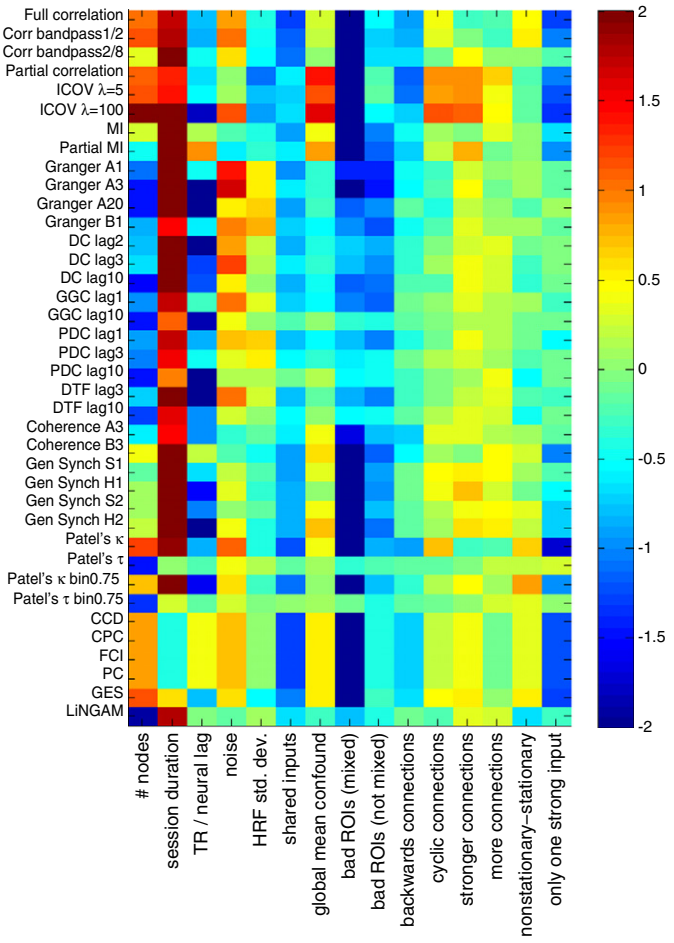

Fig. 4. Dependence of all methods' c-sensitivity on simulation parameters. For each method a multiple regression was fit, with the different simulation parameters as covariates. This was done separately for each of the 50 "subjects", and then the parameter estimates were summarised across subjects, with the colour scale showing statistical effect size (mean/standard deviation).

most variants of a given class of method performed similarly), but allows us to create simple summary figures. Fig. 6 shows the summary results for all methods' sensitivity to correctly detecting the presence of a network connection. The thick black line shows the median result across the simulations. Fig. 7 shows the summary results for all methods' accuracy in detecting network connection direction. Again the thick black line shows the median result across simulations. The grey line shows the approximate level above which accuracy is significantly better than chance, i.e., the 95th percentile of the appropriate binomial distribution (the level shown is correct for 50 subjects × 5 connections, and falls very slightly for larger numbers of nodes). *Patel's τ* performs the best overall, and *LiNGAM* is the only method to achieve over 80% accuracy (in the case of the longest timeseries). As discussed above, we have to doubt whether the *Gen Synch* and *Granger* results here are valid results driven by the true underlying causality (as opposed to confounding asymmetries in the timeseries characteristics).

Discussion

Although some of the different data scenarios generated quite variable sets of results, a general picture does emerge, that should be applicable across a large fraction of real FMRI experiments.

With respect to estimating the presence of a network connection, the overall results suggest that the "Top-3" (*Partial correlation*, *ICOV* and the Bayes net methods) often perform excellently, with a sensitivity of more than 90% on "typical" data. The Bayes net methods are as good as, or slightly better than, *Partial correlation* and *ICOV* in many scenarios, but do not seem to be in general quite as robust when faced with certain problematic scenarios. *Partial correlation* and *ICOV* often give very similar results, but with an increasing number of nodes, the regularisation in *ICOV* starts to show some benefit, as long as λ is increased accordingly. *Full correlation* and *Patel's κ* perform a little less well than the Top-3 on typical data, and often suffer more when faced with problematic scenarios. *MI*, *Coherence* and *Gen Synch* are significantly less sensitive, often detecting less than 50% of true connections. Lag-based methods (*Granger*, *PDC* and *DTF*) perform extremely poorly, with c-sensitivity of under 20%. LiNGAM performs poorly for "typical-length" FMRI timeseries, as it requires a larger number of timepoints to function well.

The sensitivity of detecting the presence of a network connection does depend on the length of the FMRI sessions, but already achieving (for the best methods) 95% with 10 min sessions. To summarise the dependence of the best methods' c-sensitivity on session duration, for the 5 nodes, 1% BOLD noise case: 60 min:100%, 10 min:95%, 5 min:77%, 2.5 min:59%.

With respect to the estimation of network connection *directionality*, high levels of accuracy are harder to achieve than just estimating the *presence* of the connection. For "typical" datasets, none of the methods is *very* accurate, though *Patel's τ* does achieve around 70% accuracy (one should bear in mind that 50% is chance level). Overall, this method performed better than all others. We can hope to improve on this accuracy in direction estimation, particularly through minimising noise levels and maximising session duration. Lag-based methods and general synchronisation often give spurious directionality estimation, driven by other factors in the data than the neural lag

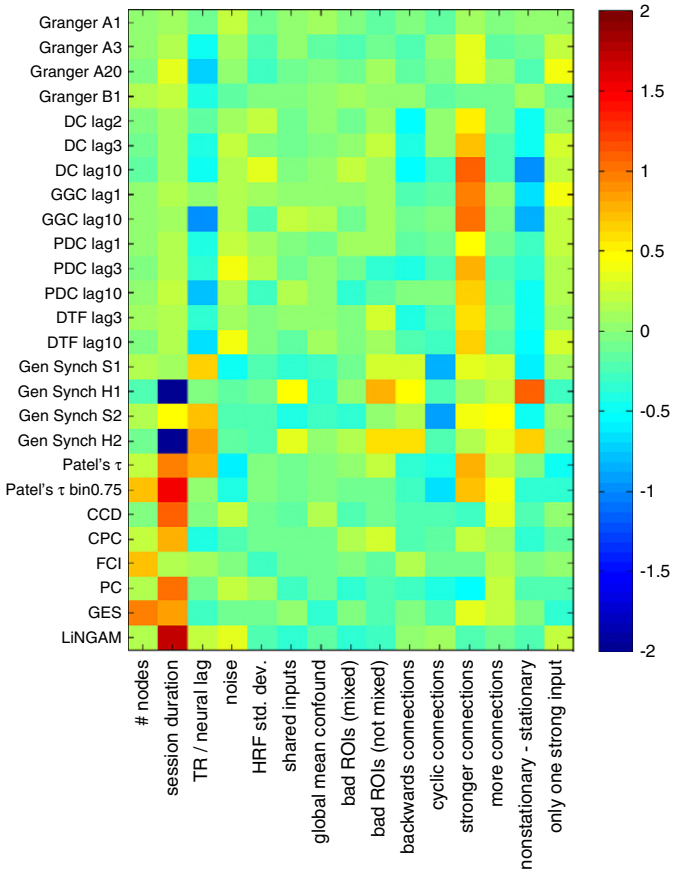

Fig. 5. Dependence of all methods' d-accuracy on simulation parameters. For each method a multiple regression was fit, with the different simulation parameters as covariates. This was done separately for each of the 50 "subjects", and then the parameter estimates were summarised across subjects, with the colour scale showing statistical effect size (mean/standard deviation).

or true direction of information flow; the sensitivity to effects such as small noise differentials in different nodes is a serious problem.

The fact that the methods based on covariance/correlation (between different nodes' timeseries) performed significantly better than those based on higher-order statistics, phase, or temporal lag, is indicative that the relevant signal in BOLD data is relatively close to being Gaussian, i.e., most of the useful signal lies in the variance. This is primarily because the haemodynamic blurring removes much of the structured signal of interest from the original neural processes. In terms of estimating directionality, no aspects of the data contain very powerful cues, but it would seem that the best we can do (given typical current datasets) is to estimate directionality not from lag or complex modelling of the full network covariance structure, but from the higher-order statistics (kurtosis, skew, etc.). This is in effect what is driving both LiNGAM and Patel's τ. Future work might look to optimise the use of the higher-order statistics specifically for the scenario of estimating directionality from BOLD data.

With respect to specific confounds that may appear in the data; some confounds do surprisingly little damage to the best modelling approaches, while others render all methods mostly unusable. A global additive confound, even with relatively high amplitude, does little harm to the accuracies of the Top-3 methods. "Shared" external inputs are slightly more of a problem for Bayes net methods than *Partial Correlation* and *ICOV*, although directionality estimation by the Bayes net methods was not badly affected. Inaccurate ROI specification, even by as little as 20%, gives extremely poor results—every method gives lower than 20% sensitivity to detecting the presence of connections. This emphasises the great importance of determining ROIs that are appropriate for the data, and speaks against using atlas-based ROIs. However this issue is less of a problem when the ROIs of interest are not spatially neighbouring, because in this scenario ROI inaccuracy just adds noise, rather than mixing the timeseries together.

Cyclic connections are not too much of a problem for connection-presence sensitivity, and neither was an increased "density" of connections. Increased network connection strengths (~0.9 instead of ~0.4) did not change the overall results greatly, although the accuracy of estimating directionality does suffer, with *Patel's* τ performing the best. With only one strong input to the network (implying also no neural "noise" feeding into any nodes other than node 1), all methods suffer badly, with *Partial correlation* and *ICOV* performing the best. In the presence of network-connection-strength temporal nonstationarity, the Top-3 methods are still performing the best, and achieve good results.

It is hard to know specifically how to simulate "backwards" neural connections for FMRI data, but our limited results suggest that these can reduce the accuracy of the network modelling. However, the Top-3 approaches were still the strongest, with Bayes nets slightly ahead. When interpreting network connections that are estimated as negative, one should bear in mind that these do not necessarily represent "inhibition", but can also be induced to be negative (in fact, more generally, reverse sign) by certain analysis methodologies; for example, under certain circumstances, network matrix elements that would be estimated as positive by full correlation may be estimated as negative by partial correlation (Marrelec et al., 2009). Until further experimental and simulation validations are carried out that specifically look into these issues, including the interpretability of negative

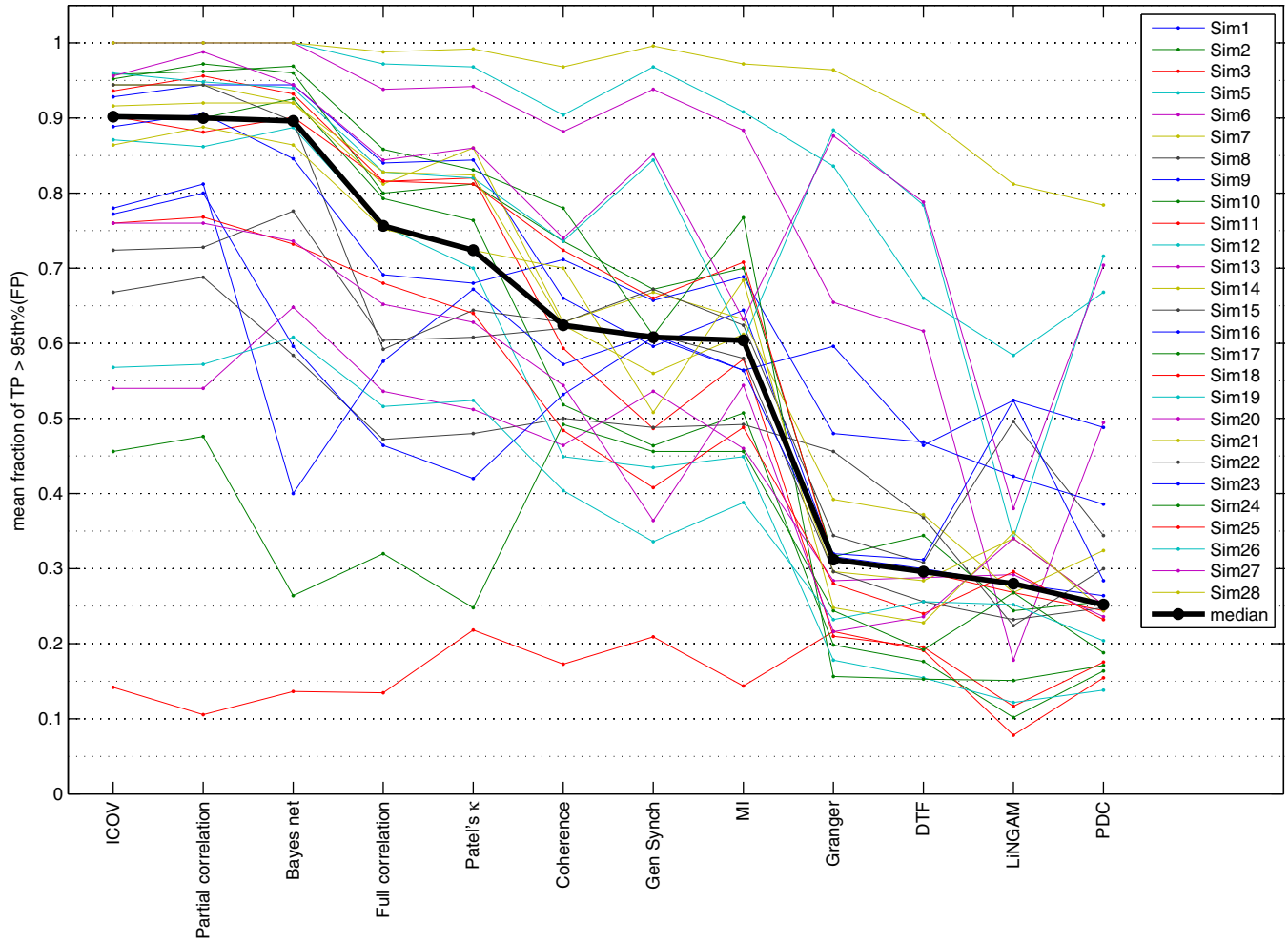

Fig. 6. Sensitivity to correctly detecting the presence of a network connection; summary over simulations and methods. The thick black line shows the median result over simulations.

("inhibitory") connections, it may not be wise to assume that apparently significant negative connections are real.

The lag-based methods performed very poorly, and this deserves a little discussion, as these methods are currently used regularly in FMRI analysis. It has been noted (Friston, 2009) that variations in haemodynamic lag across brain regions are likely to swamp any causal lag in the underlying neural timeseries, as they are generally at least an order of magnitude larger (except possibly in rare cases where certain cognitive experiments may induce the largest lags between functional units). This concern was noted in one of the earliest papers applying Granger to FMRI data (Roebroeck et al., 2005), where it was stated: "…one should rule out the possibility that influence found from one area to another based on temporal difference in signal variation is due to a systematic difference in the hemodynamic lag at the two areas. A possible approach to exclude this confound is to show that the measured influence varies with experimental condition or cognitive context." Such an approach relies on a linear (or well-modelled) transfer function between neural activity and the BOLD signal, including full knowledge of how varying experimental condition changes this transfer function. In the case of resting FMRI timeseries, and in many task FMRI experiments, we do not have the luxury of varying the experimental context in a suitable way that guarantees no changes in haemodynamics. Even without the HRF variability, the blurring effect of the HRF is expected to reduce any lag information present in the neural timeseries to insignificance, for the temporal neural lags seen in *most* experiments. These issues are discussed further in recent papers (Roebroeck et al., in press-a, in press-b; Friston, in press; David, in press). However the major concerns described above have still not been satisfactorily answered in the literature, except to point to the results of certain simulations, which would appear to show successful causality estimation, but which are likely to be spurious, as discussed below.

The spurious causality estimation that is still seen in the absence of HRF variability most likely relates to various problems described in the Granger literature (Tiao and Wei, 1976; Wei, 1978; Weiss, 1984; Nalatore et al., 2007; Nolte et al., 2008); it is known that measurement noise can reverse the estimation of causality direction, and that temporal smoothing can mean that correlated timeseries are estimated to "cause" each other. These known theoretical results probably explain the various problematic results we saw in our simulations even in the absence of HRF variability, such as the estimation of causality even when neural lag was reduced to zero, and the adding of (small/realistic amounts of) measurement noise reversing the estimated direction of causality. It is certainly possible that non-neural noise components (both physiological and scanner-related) can cause different image regions to have different noise characteristics (including amplitude), and hence this concern is a real one in practice, not just in theory and simulations. There is hope that *these* issues could be ameliorated through more sophisticated modelling (Nalatore et al., 2007; Havlicek et al., 2010; Deshpande et al., 2010a), for example, using Kalman-based noise modelling, and through Granger measures that model out the effects of lag-zero covariance. However, it seems unlikely that (uncertainty in) haemodynamic lag can be robustly distinguished from lags caused by neural delays through such modelling, and hence one of the major problems would remain.

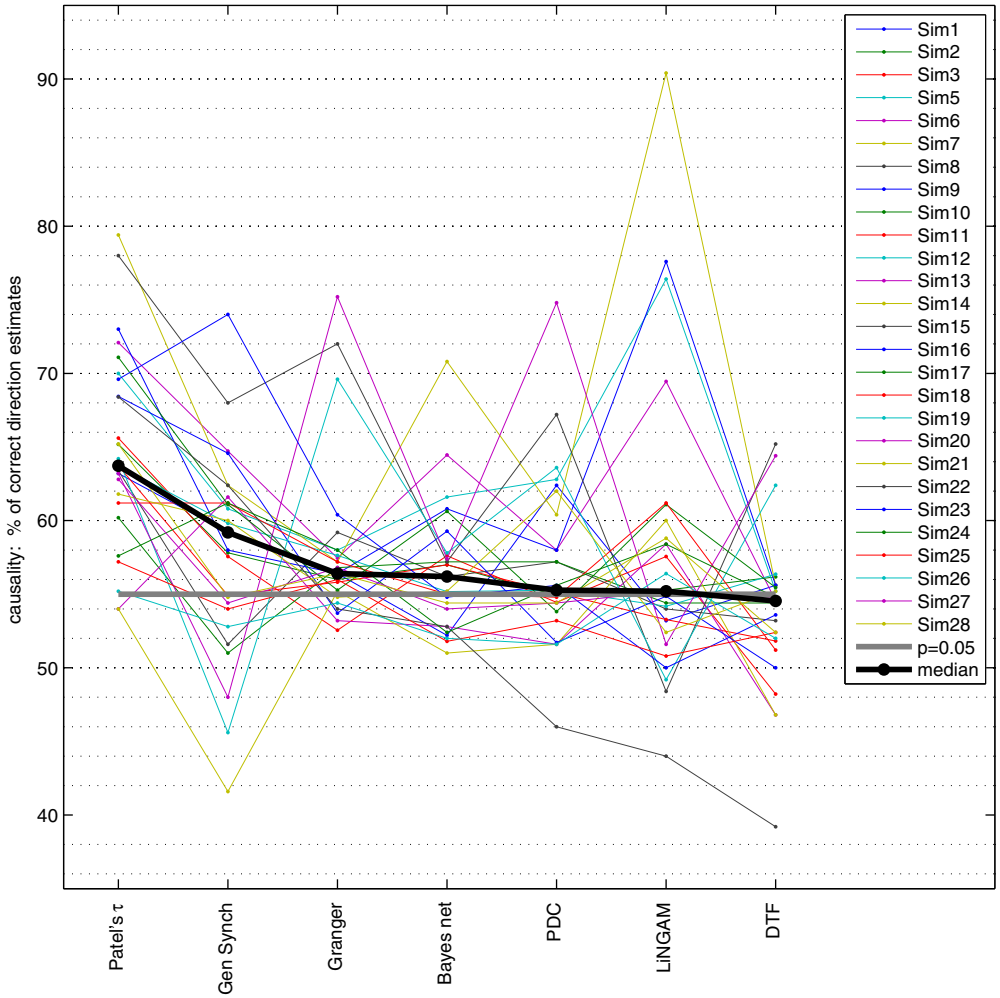

Fig. 7. Accuracy in detecting network connection direction; summary over simulations and methods. The thick black line shows the median result over simulations. The grey line shows the level above which accuracy is significantly better than chance.

Previously published simulations suggesting successful estimation of lag-based causality in FMRI data most likely suffer from the various problems discussed above, including: not adding HRF variability into the simulation, adding neural lags that are unrealistically long for the majority of neural connections, and not testing whether reducing the neural lag to zero in the simulations shows spurious causality. For example, in Deshpande et al. (2010b), the causality graphs, worryingly, often do not fall to zero for zero neural lag, and the results in general seem to only support robust (safe) causality estimation for neural lags of hundreds of milliseconds. Furthermore, in their four-node network simulations, two of the four connections were apparently simulated with lags of *tens* of seconds.[5] Another example is Havlicek et al. (2010), where an impressive methodology is developed that aims to deal with nonstationary effects in the data. The authors begin by testing their methodology against simulated data; however, the simulations are problematic, as an unrealistically long lag (2 s) is inserted between the simulated timeseries, the method is not tested against zero-lag simulated data (to confirm that null data gives a null result), and it is not shown whether the method can distinguish between neural lags and varying haemodynamic lags across regions.

Although our results suggest that *in general* lag-based methods are not useful when applied to FMRI data, this is not to say that they can *never* work with carefully generated and interpreted FMRI data. It is possible that for those (relatively rare) functional connections whose effective neural lags are greater than 100 ms, then with very low TR (probably limiting the field-of-view) and high SNR, and if the haemodynamic lag can be accurately pre-characterised for each region of interest and for every experimental context of relevance, then in theory lag-based neural causality from FMRI data may be estimable. For example, David et al. (2008) show that lag-based causality can give reasonable results after deconvolving the HRF, with the timing of the HRF estimated through the use of electrophysiological data acquired simultaneously. In another example, Rogers et al. (2010) show that neural lags as short as 100 ms should be estimable between two areas from the BOLD timecourses in the case when the two areas in question have (presumably) identical HRFs (left and right V1), and using high field strength (7 T) and low TR (250 ms).

By generating realistic simulated data we are able to test different methods' sensitivities to detecting network connections, comparing estimated results against the ground truth used to drive the simulations. We have thus been able to control not only the "full null" false positive rate of connection detection *in the absence of any true positives*, but also correct for the (generally increased) rate of false positives *induced by the presence of true connections* (i.e., evaluating the specificity of each method in distinguishing direct from indirect connections). When a network modelling method is applied to new, real data, it is important to be able to estimate the *significance* of estimation connections (whether considering connection presence, strength or directionality). For methods where the two kinds of null distributions particularly diverge (e.g., full correlation), it can potentially be hard to make accurate inference, as the null distribution *in the presence of true connections* may be hard to estimate. Certain network scenarios will exacerbate this problem, as seen in our results, such as increasing connection strengths and density of connections. The use of surrogate data ("null data" generated to have similar characteristics to the real data) may help, possibly in conjunction with a network simulation such as those utilised in this work; however, valid surrogate data can be difficult to generate, as seen for example in the (surrogate-corrected) problematic results in Deshpande et al. (2010b), described above. The methods that gave the best results overall in our tests beat the "second-best" methods (such as full correlation) primarily because they were the most successful at distinguishing direct from indirect connections, and for the same reason their "full null" distributions need the least correction when true connections are present (making it easier to achieve valid inference on their outputs when applied to real data). Closely related to these questions is whether any given method's own estimates of statistical significance (for example, *p*-values derived from parametric assumptions regarding a method's estimated connection strength or directionality) are accurate. We have not needed to utilise any method's own associated (or "built-in") approach for estimating specificity, because we have been able to use the ground truths to estimate false-positive rates empirically; it would be interesting to investigate whether different methods' own claimed *p*-values are accurate, but this is outside the scope of this paper (and of course can be confounded, for some methods, by the additional false positives that can be induced by true positives, as discussed above).

We have not considered any specific multi-subject modelling approaches here (for example, as seen in Ramsey et al., 2010), as we have concentrated on evaluating the different modelling methods when used to estimate functional brain networks from single subject datasets; we felt that this is a primary question of interest, needing some clear answers before considering possible multi-subject modelling approaches. A question will arise as to whether the methods (such as LiNGAM) that require a relatively large number of timepoints to function well would give good results simply by concatenating timeseries across subjects; this may prove to be the case, although such an approach would then restrict the ability to use simple methods (such as cross-subject mixed-effects modelling of the estimated network parameters) to determine the reliability of the group-estimated network.

To conclude, we have carried out a complex, rich, quantitative set of simulations with a realistic model for FMRI timeseries; we have investigated a wide range of analysis methodologies in order to determine which are the most sensitive at detecting direct network functional connections and the direction of information flow carried by them. We have included various problematic elements in the simulations, in order to evaluate the robustness of the various network modelling methods. Our results show that lag-based approaches perform very poorly. However there are several methods that can give high sensitivity to network connection estimation on good quality FMRI data, in particular partial correlation, regularised inverse covariance estimation and several Bayes net methods. With respect to estimating connection directionality, Patel's τ can be reasonably successful.

All simulated BOLD timeseries and ground-truth network matrices are freely available from www.fmrib.ox.ac.uk/analysis/netsim—the authors will be interested to receive feedback on the simulations, methods tested and results, and will post updates on this website.

Supplementary materials related to this article can be found online at doi:10.1016/j.neuroimage.2010.08.063.

Acknowledgments

We are very grateful to: Gopikrishna Deshpande for helpful discussions, Aslak Grinsted for providing the Crosswavelet and Wavelet Coherence toolbox, Chandler Lutz for providing the pairwise Granger causality code, Alard Roebroeck for helpful discussions, Alois Schlögl for providing the BioSig toolbox, Mark Schmidt for providing the regularised inverse covariance code (and for helpful discussions), Anil Seth for providing the Causal Connectivity Analysis toolbox (and for helpful discussions), Shohei Shimizu, Patrik Hoyer and Aapo Hyvärinen for providing LiNGAM/FastICA (and for helpful discussions), Rodrigo Quian Quiroga for providing the generalised synchronisation code (and for helpful discussions) and Dongli Zhou for providing the Functional Connectivity toolbox. We would like to make clear that any negative results or discussions resulting from relatively poor performance of any of

[5] It is clear from consideration of Figs. 1 and 2 that for two of the four network connections, the neural lags are of the same order as the *lengths* of the timeseries, so either the neural lags are tens of seconds, or the timeseries are too short to be meaningful. In either case, the network simulation results are not easily interpretable.

the methods tested here do not imply any criticism of the original methods themselves or their implementations in the various toolboxes; this paper is purely aiming to investigate the performance of the various methods *when applied to FMRI data*. Many of the methods were originally designed to work with other types of data with different signal and noise characteristics; however, we have felt it important to test all of these methods specifically for FMRI data, because they are being used by the FMRI community.

References

Baccalá, L., Sameshima, K., 2001. Partial directed coherence: a new concept in neural structure determination. Biol. Cybern. 84 (6), 463–474.

Banerjee, O., El Ghaoui, L., d'Aspremont, A., Natsoulis, G., 2006. Convex optimization techniques for fitting sparse Gaussian graphical models. Proceedings of the 23rd International Conference on Machine Learning. ACM, p. 96.

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

Chang, C., Thomason, M., Glover, G., 2008. Mapping and correction of vascular hemodynamic latency in the BOLD signal. Neuroimage 43, 90–102.

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

Daunizeau, J., Friston, K., Kiebel, S., 2009. Variational Bayesian identification and prediction of stochastic nonlinear dynamic causal models. Phys. D Nonlinear Phenom. 238 (21), 2089–2118.

Dauwels, J., Vialatte, F., Musha, T., Cichocki, A., 2010. A comparative study of synchrony measures for the early diagnosis of Alzheimer's disease based on EEG. Neuroimage 49 (1), 668–693.

David, O., in press. fMRI connectivity, meaning and empiricism: Comments on: Roebroeck et al. The identification of interacting networks in the brain using fMRI: model selection, causality and deconvolution. *NeuroImage*, Corrected Proof:–.

David, O., Guillemain, I., Saillet, S., Reyt, S., Deransart, C., Segebarth, C., Depaulis, A., 2008. Identifying neural drivers with functional MRI: an electrophysiological validation. PLoS Biol. 6 (12), e315.

de Pasquale, F., Della Penna, S., Snyder, A., Lewis, C., Mantini, D., Marzetti, L., Belardinelli, P., Ciancetta, L., Pizzella, V., Romani, G., Corbetta, M., 2010. Temporal dynamics of spontaneous MEG activity in brain networks. Proc. Natl Acad. Sci. 107 (13), 6040–6045.

Deshpande, G., Sathian, K., Hu, X., 2010. Assessing and compensating for zero-lag correlation effects in time-lagged Granger causality analysis of fMRI. IEEE Trans. Biomed. Eng. 57 (6), 1446–1456.

Deshpande, G., Sathian, K., Hu, X., 2010. Effect of hemodynamic variability on Granger causality analysis of fMRI. Neuroimage 52 (3), 884–896.

Fox, M., Zhang, D., Snyder, A., Raichle, M., 2009. The global signal and observed anticorrelated resting state brain networks. J. Neurophysiol. 101 (6), 3270–3283.

Freenor, M. and Glymour, C., 2010. Searching the DCM model space, and some generalizations. NeuroImage. In submission.

Friedman, J., Hastie, T., Tibshirani, R., 2008. Sparse inverse covariance estimation with the Graphical Lasso. Biostat 9 (3), 432–441.

Friston, K., 1994. Functional and effective connectivity in neuroimaging: a synthesis. Hum. Brain Mapp. 2, 56–78.

Friston, K., 2009. Causal modelling and brain connectivity in functional magnetic resonance imaging. PLoS Biol. 7 (2), e1000033.

Friston, K., in press. Dynamic causal modeling and Granger causality. Comments on: The identification of interacting networks in the brain using fMRI: model selection, causality and deconvolution. *NeuroImage*, Corrected Proof:–.

Friston, K.J., Harrison, L., Penny, W., 2003. Dynamic causal modelling. Neuroimage 19 (3), 1273–1302.

Geweke, J.F., 1984. Measures of conditional linear dependence and feedback between time series. J. Am. Stat. Assoc. 79 (388), 907–915.

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 (10), 1251–1261.

Granger, C.W.J., 1969. Investigating causal relations by econometric models and cross-spectral methods. Econometrica 37 (3), 424–438.

Grinsted, A., Moore, J., Jevrejeva, S., 2004. Application of the cross wavelet transform and wavelet coherence to geophysical time series. Nonlinear Processes Geophys. 11 (5/6), 561–566.

Guo, S., Seth, A.K., Kendrick, K.M., Zhou, C., Feng, J., 2008. Partial Granger causality—eliminating exogenous inputs and latent variables. J. Neurosci. Meth. 172 (1), 79–93.

Handwerker, D., Ollinger, J., D'Esposito, M., 2004. Variation of BOLD hemodynamic responses across subjects and brain regions and their effects on statistical analyses. Neuroimage 21, 1639–1651.

Havlicek, M., Jan, J., Brazdil, M., Calhoun, V.D., 2010. Dynamic Granger causality based on Kalman filter for evaluation of functional network connectivity in fMRI data. NeuroImage 53 (1), 65–77.

Kamiński, M., Blinowska, K., Szelenberger, W., 1997. Topographic analysis of coherence and propagation of EEG activity during sleep and wakefulness. Electroencephalogr. Clin. Neurophysiol. 102 (3), 216–227.

Kiviniemi, V., Kantola, J.-H., Jauhiainen, J., Hyvärinen, A., Tervonen, O., 2003. Independent component analysis of nondeterministic fMRI signal sources. Neuroimage 19, 253–260.

Kiviniemi, V., Starck, T., Remes, J., Long, X., Nikkinen, J., Haapea, M., Veijola, J., Moilanen, I., Isohanni, M., Zang, Y.-F., Tervonen, O., 2009. Functional segmentation of the brain cortex using high model order group PICA. Hum. Brain Mapp. 30 (12), 3865–3886.

Larkin, P., 1971. This Be The Verse. New Humanist. August.

Marrelec, G., Kim, J., Doyon, J., Horwitz, B., 2009. Large-scale neural model validation of partial correlation analysis for effective connectivity investigation in functional MRI. Hum. Brain Mapp. 30 (3), 941–950.

Marrelec, G., Krainik, A., Duffau, H., Pélégrini-Issac, M., Lehéricy, S., Doyon, J., Benali, H., 2006. Partial correlation for functional brain interactivity investigation in functional MRI. Neuroimage 32, 228–237.

McIntosh, A., Gonzales-Lima, F., 1994. Structural equation modeling and its application to network analysis in functional brain imaging. Hum. Brain Mapp. 2, 2–22.

Meek, C., 1995. Causal inference and causal explanation with background knowledge. Proceedings of the 11th Annual Conference on Uncertainty in Artificial Intelligence, pp. 403–410.

Nalatore, H., Ding, M., Rangarajan, G., 2007. Mitigating the effects of measurement noise on Granger causality. Phys. Rev. E 75 (3), 31123.1–31123.10.

Nolte, G., Ziehe, A., Nikulin, V., Schlögl, A., Krämer, N., Brismar, T., Müller, K., 2008. Robustly estimating the flow direction of information in complex physical systems. Phys. Rev. Lett. 100 (23), 234101.1–234101.4.

Patel, R., Bowman, F., Rilling, J., 2006. A Bayesian approach to determining connectivity of the human brain. Hum. Brain Mapp. 27, 267–276.

Pereda, E., Quiroga, R., Bhattacharya, J., 2005. Nonlinear multivariate analysis of neurophysiological signals. Prog. Neurobiol. 77 (1–2), 1–37.

Popa, D., Popescu, A., Paré, D., 2009. Contrasting activity profile of two distributed cortical networks as a function of attentional demands. J. Neurosci. 29 (4), 1191–1201.

Quian Quiroga, R., Kraskov, A., Kreuz, T., Grassberger, P., 2002. Performance of different synchronization measures in real data: a case study on electroencephalographic signals. Phys. Rev. E 65 (4), 41903.

Ramsey, J., Hanson, S., Hanson, C., Halchenko, Y., Poldrack, R., Glymour, C., 2010. Six problems for causal inference from fMRI. Neuroimage 49 (2), 1545–1558.

Ramsey, J., Zhang, J., Spirtes, P., 2006. Adjacency-faithfulness and conservative causal inference. Proceedings of the 22nd Annual Conference on Uncertainty in Artificial Intelligence.

Richardson, T., Spirtes, P., 2001. Automated discovery of linear feedback models. In: Glymour, C., Cooper, G. (Eds.), Computation, Causation, and Causality. MIT Press.

Ringo, J., Doty, R., Demeter, S., Simard, P., 1994. Time is of the essence: a conjecture that hemispheric specialization arises from interhemispheric conduction delay. Cereb. Cortex 4, 331–343.

Roebroeck, A., Formisano, E., Goebel, R., 2005. Mapping directed influence over the brain using Granger causality and fMRI. Neuroimage 25 (1), 230–242.

Roebroeck, A., Formisano, E., and Goebel, R., in press. Reply to Friston and David: After comments on: The identification of interacting networks in the brain using fMRI: model selection, causality and deconvolution. *NeuroImage*, Corrected Proof:–.

Roebroeck, A., Formisano, E., and Goebel, R., in press. The identification of interacting networks in the brain using fMRI: model selection, causality and deconvolution. *NeuroImage*, Corrected Proof:–.

Rogers, B.P., Katwal, S.B., Morgan, V.L., Asplund, C.L., Gore, J.C., 2010. Functional MRI and multivariate autoregressive models. Magnetic Resonance Imaging 28 (8), 1058–1065.

Seth, A.K., 2010. A MATLAB toolbox for Granger causal connectivity analysis. J. Neurosci. Meth. 186 (2), 262–273.

Shannon, C., 1948. A mathematical theory of communication. Bell Syst. Tech. J. 27, 379–423.

Shimizu, S., Hoyer, P.O., Hyvärinen, A., Kerminen, A., 2006. A linear non-Gaussian acyclic model for causal discovery. J. Mach. Learn. Res. 7, 2003–2030.

Tiao, G., Wei, W., 1976. Effect of temporal aggregation on the dynamic relationship of two time series variables. Biometrika 63 (3), 513–523.

Wei, W., 1978. The effect of temporal aggregation on parameter estimation in distributed lag model. J. Econometrics 8 (2), 237–246.

Weiss, A., 1984. Systematic sampling and temporal aggregation in time series models. J. Econometrics 26 (3), 271–281.

Witt, S., Meyerand, M., 2009. The effects of computational method, data modeling, and TR on effective connectivity results. Brain Imaging Behav. 3, 220–231.

Wright, S., 1920. Correlation and causation. J. Agric. Res. 20, 557–585.

Zhang, J., 2008. On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. Artif. Intell. 172, 1873–1896.

Zhou, D., Thompson, W.K., Siegle, G., 2009. MATLAB toolbox for functional connectivity. Neuroimage 47 (4), 1590–1607.