NeuroImage 49 (2010) 1545–1558

Contents lists available at ScienceDirect

NeuroImage

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

Six problems for causal inference from fMRI

J.D. Ramsey [a,*], S.J. Hanson [b], C. Hanson [b], Y.O. Halchenko [b], R.A. Poldrack [c], C. Glymour [d]

[a] Department of Philosophy, Carnegie Mellon University, Pittsburgh, PA 15213
[b] Department of Psychology, Rutgers University, Rumba Lab
[c] Imaging Research Center and Departments of Psychology and Neurobiology, University of Texas at Austin
[d] Department of Philosophy, Carnegie Mellon University, and Florida Institute for Human and Machine Cognition

ARTICLE INFO

Article history:
Received 13 February 2009
Revised 7 August 2009
Accepted 31 August 2009
Available online 9 September 2009

ABSTRACT

Neuroimaging (e.g. fMRI) data are increasingly used to attempt to identify not only brain regions of interest (ROIs) that are especially active during perception, cognition, and action, but also the qualitative causal relations among activity in these regions (known as *effective connectivity*; Friston, 1994). Previous investigations and anatomical and physiological knowledge may somewhat constrain the possible hypotheses, but there often remains a vast space of possible causal structures. To find actual effective connectivity relations, search methods must accommodate indirect measurements of nonlinear time series dependencies, feedback, multiple subjects possibly varying in identified regions of interest, and unknown possible location-dependent variations in BOLD response delays. We describe combinations of procedures that under these conditions find feed-forward sub-structure characteristic of a group of subjects. The method is illustrated with an empirical data set and confirmed with simulations of time series of non-linear, randomly generated, effective connectivities, with feedback, subject to random differences of BOLD delays, with regions of interest missing at random for some subjects, measured with noise approximating the signal to noise ratio of the empirical data.

Six problems for causal inference from fMRI

Functional Magnetic Resonance data are increasingly used to attempt to identify not only brain regions of interest (ROIs) that are especially active during perception, cognition, and action, but also the causal relations among activity in these regions (known as "effective connectivities"; Friston, 1994). Modeling of this kind is typically done either by positing a parameterized causal structure a priori and estimating the parameters from data, or by statistical comparison of a few alternative models. Previous investigations and anatomical and physiological knowledge may somewhat constrain the possible hypotheses, but there often remains a vast space of possible models. Some automated search procedures have been tried, notably multiple regression, modification indices (Bullmore et al., 2000) and exhaustive search (Hanson et al., 2007). Recent investigations (e.g., Zheng and Rajapake, 2006) have used quasi-Bayesian model scoring methods to search parts of this space of alternative models, but none of these automated search procedures addresses a combination of fundamental problems particular to fMRI applications. In this paper we outline some of the well-known, central challenges inherent in modeling causal relations in such data, and consider a set of partial solutions.

It is now customary to represent effective connectivity relations abstractly as directed graphs, where nodes in the graph represent brain regions and directed edges in the graph represent relatively direct causal influences of one region on another. More precisely, in a directed graph G, with vertex/node/variable-set $\mathbf{V}$, a directed edge $V_j \rightarrow V_k$, represents the proposition that there are values of the variables in $\mathbf{V}$, leaving aside V_j or V_k, such that if those variables were (hypothetically) to be held fixed at those values, then some hypothetical intervention that varies values of V_j, would produce an associated set of values for V_k. V_k is said to be a *child* of $V_j$and V_j a *parent* of V_k. If there is a directed path from V_j to V_r, V_r is said to be a *descendant* of V_j. A graphical causal model may contain explicit unmeasured ("latent") variables as well as measured variables. In the search for brain mechanisms from imaging data, the goal is to identify the causal relations among the (unmeasured) neuronal populations whose activity gives rise to observed fMRI signals in spatially localized regions of interest (ROIs).

Graphical causal models–sometimes called "causal Bayes nets"–combine a directed graph with a joint probability distribution on the graph nodes that represent random variables. The graphical structure is intended to capture both the compositional structure of the causal relations and general aspects of all probability distributions that factor according to that structure. For directed graphs without cycles (DAGs)

* Corresponding author.
E-mail address: jdramsey@andrew.cmu.edu (J.D. Ramsey).

1053-8119/$ – see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.neuroimage.2009.08.065

the defining property of a graphical causal model is the Causal Markov Property (Spirtes et al., 1993), characterized as follows (boldface denotes sets of variables):

"*Causal Markov Condition*: Let *G* be causal graph with vertex set **V** and *P* be a probability distribution over the vertices in **V** generated by the causal structure represented by *G*. *G* and *P* satisfy the Causal Markov Condition if and only if for every *W* in **V**, *W* is independent of **V****Descendants**(W) ∪ **Parents**(*W*) given **Parents**(*W*)."

The Markov property implies that for any value assignments to the variables, the joint distribution of the variables is equal to the product, over all variables, of the probability of the value of each variable conditional on the values of its parent variables, a so-called *Markov factorization* of the distribution.

A graphical causal model may be linear or non-linear, may form a time series, and may have feedback relations (Spirtes et al., 2000; Glymour, 2003). The graph itself is a non-parametric simultaneous representation of conditional independence constraints (for whatever probability distribution accords with the graph topology) and qualitative effective connectivity relations. Feedback relations can be represented by directed graphs with cycles or by time series with acyclic graphs. Cyclic graphs for linear simultaneous equations with independent disturbances do not satisfy the Markov condition, but when used to represent linear systems they do satisfy a generalization of the Markov condition (Spirtes, 1995) that permits a factorization and computation of the vanishing partial correlations (and for Gaussian distributions, the conditional independence relations) that are satisfied by all linear parameterizations of the graph.

Problem 1: searching over models

A first challenge in characterizing causal relations between brain regions arises from the astronomical size of the possible space of alternative causal models. Represented graphically, the number of alternative possible causal structures relating *N* ROIs (disallowing further variables as common causes) is the number of directed graphs (including 2-cycles) on *N* vertices, which is $4^{((N-1)\,N)/2}$. Even confined to directed *acyclic* graphs (DAGs), the number of alternative graphs grows super-exponentially with increasing numbers of vertices. The empirical data we will consider for illustration consist, respectively, of subsets of 5, 9 and 11 ROIs from a single study. Even for singly connected DAGs with a designated input variable and 10 other variables–which is a proper subset of the space we will search in our empirical example and in some of our simulations–there are 2,609,107,200 alternative causal structures (the set of permutations of rooted trees on 10 vertices). Reliable and computationally tractable methods of searching over the spaces of alternative models seem urgent if imaging data is to warrant novel claims about effective connectivity.

Search over a space of alternative models is a statistical estimation problem. In searches for the graphical structure of causal relations, the search space is some set of directed graphs, subject to some joint distribution constraints. The estimates of graphical representation of causal structure that arise from these searches can be characterized using the same kinds of methods applied to routine statistical estimators (e.g., consistency or bias criteria). These and other desiderata are the subject of proofs of properties for estimators under assumptions limiting the hypothesis space—i.e., the graph structures, parameters and probability distribution families. Where analytic results are not available the behavior of statistical estimators is increasingly studied by simulation and the same applies to search procedures that estimate graphical causal structure. Just as a statistical estimator may not provide an estimate of all of the information of interest about a probability distribution, a search procedure over graphical causal models may not provide all of the information of interest about the processes generating the data. Thus, for reasons to be explained, the procedures we will describe estimate only feed-forward substructures of effective connectivity relations, even when the underlying process generating the data contains effective back-projections. Further, the graphical causal models sought by the search procedures we will describe are non-parametric. They do not contain explicit parameters such as linear coefficients, although they can be parameterized according to the investigator's judgment or other considerations, and data can then be used to estimate the parameter values. Further, the graphs we obtain are intended to be characteristics shared by a group of subjects (as with Chen and Herskovitz, 2007), and a number of subjects is required for reliable discovery. Finally, the procedures we will describe do not always return a unique DAG. Multiple DAGs, implying different causal structures, may imply the very same set of conditional independence relations when the Markov condition is applied to each DAG. Such a collection of DAGs is a *Markov Equivalence Class*. For Gaussian distributions, the methods we will describe (and in the Gaussian case, all known consistent search methods for graphical causal models that are DAGs) return only a Markov Equivalence Class. That class may, however, sometimes be a singleton, as in the empirical example we will describe later.

The parallel equivalence relation for directed cyclic graphs representing linear systems is well understood, and algorithms for deciding whether two cyclic graphs imply the same conditional independence relations for all of their linear parameterizations are known (Richardson, 1996; Lacerda et al., 2008). The cyclic equivalence classes are, however, more complex and members of the same equivalence class share less structure than in the acyclic case. In particular, all DAGs in a Markov Equivalence Class share the same *adjacencies*–if $A \rightarrow B$ is in one such graph, then $A \rightarrow B$ or $A \leftarrow B$ is in any other graph in the same Markov Equivalence class–but that is not true for equivalence classes of cyclic graphs. For these reasons, we attend here to the more limited problem of finding feed-forward substructures from fMRI data.

Problem 2: indirect measurements

Structural equation models, or SEMs, posit linear causal relations among variables with independent, unobserved noises. Introduced in social science in the late 19th century, they were proposed as models for neural influences measured by fMRI signals by McIntosh and Gonzalez-Lima in 1994, and continue to be recommended for that purpose (Demarco et al., 2009). It has been correctly objected (for example, by Penny et al., 2004) that the measured variables in SEM models of fMRI data are in reality indirect effects of the underlying non-linear relations of interest (effective connectivities). SEM models can, however, be given non-linear forms and so the remark invites a question: *why cannot models using variables defined by aggregates of voxel measurements of hemodynamic responses capture the underlying structure of aggregate neural influences between regions?* One answer is that the Markov factorization of a graphical causal model specifies conditional independence relations; the distribution of measured variables is a marginal of a joint distribution including latent variables, and conditional independence relations are not always preserved by marginalization. A search procedure for causal structure may therefore be unbiased (have an expected graph structure in the Markov equivalence class of the true structure) and consistent (that is, converge in probability in the large sample limit to correct information) if the causal relations are only among the measured variables, but biased and not consistent if the causal relations are among latent variables that produce values of measured variables subject to random disturbances. Thus, for concreteness, suppose that the fMRI signal recorded from three regions (X_1, X_2, X_3) reflects the true net synaptic activity in those regions (reflected in latent variables L_1, L_2, and L_3) along with random disturbances in the fMRI signal (represented by error variables $\varepsilon_{X1} - \varepsilon_{X3}$) and random disturbances in the influence between regions at the neural level (represented by error variables $\varepsilon_{L1} - \varepsilon_{L3}$). Fig. 1 shows a graphical representation of this model.

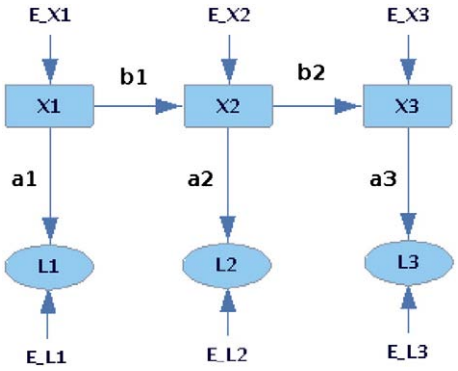

Fig. 1. An indirectly measured causal chain.

The disturbance terms are positive variance and jointly independent of each other. L_1 is independent of L_3 conditional on L_2, but there is no corresponding independence among the X variables. The point is general and not dependent either on linearity or Normality, but it is easily demonstrated with a linear Gaussian system: standardize the X and L variables to unit variance and 0 mean. Let $X_i = a_i L_i + \varepsilon_i$, where the ε_i are independent of each other and of the L_i, and let $L_2 = b_2 L_1 + \varepsilon_{L2}$ and $L_3 = b_3 L_2 + \varepsilon_{L3}$, where the ε_L variables are jointly independent and independent of the ε_i variables. Because the distribution is Gaussian, conditional independence is equivalent to vanishing partial correlation. The numerator of the partial correlation, $\rho_{L1,\ L3.\ L2}$, equals $\rho_{L1,\ L3} - \rho_{L1,L2}\rho_{L2,\ L3} = b_2\ b_3 - (b_2)\ (b_3) = 0$, but the partial correlation of X_1 and X_3 controlling for X_2 is proportional to $a_1 b_2 b_3 a_3 - (a_1 b_2 a_2)(a_2 b_3 a_3) = a_1 b_2 b_3 a_3 (1 - a_2^2)$ which is not equal to 0. Causal structure estimators that implicitly or explicitly separate variables by vanishing partial correlations or by conditional independence will tend to find a direct dependence of X_3 on X_1, which does not reflect the relation between L_3 and L_1. Thus, any causal inference methods to be applied to fMRI must be able to address the statistical consequences of indirect measurement and latent sources. There is one bright spot. In the example of Fig. 1, the spurious association of X_1 and X_3 conditional on X_2 will be weaker than the association of X_1 with X_2 and weaker than the association of X_2 with X_3. Spurious associations involving conditioning on multiple intermediate variables in a causal chain will be weaker still. We will exploit differences of these kinds.

In principle, this problem could be solved by estimating the values of the latent variables. "Deconvolution" methods attempt to do that; if, however, the deconvolution estimates are noisy, the same difficulty–spurious estimated connections–may arise.

Problem 3: modeling causal structure across individuals

Imaging studies are frequently run on multiple subjects, and separate laboratories may carry out experiments with essentially identical stimuli. Potentially, such repetitions considerably increase the sample size and should decrease dimensions of the variance of causal structure estimators, but there are serious theoretical difficulties. While the several participants in a study may share a common abstract processing structure–i.e., which regions of the brain influence which other regions in response to a stimulus–they are apt to vary in the strengths of regional responses to stimuli and in the strengths of influences from one region to another. Measurements with the same stimulus but with different magnets or other instrumentation may also result in different measurement sensitivities to brain activity. For all of these reasons, as well as because of measurement error, the probability distribution of neural responses, and the directed graphs representing those distributions, can be expected to differ across participants. Even when there is a common causal structure shared by multiple subjects, combining two or more joint distinct probability distributions for the same variables changes the distributions and destroys information that may be shared by all of the distributions (Yule, 1919). Except in special cases, independence and conditional independence relations that hold in each of several probability distributions will not hold in the combined distribution. Consequently, data from multiple subjects cannot be pooled to form a unified data base for model search. Some authors working in fMRI analysis have emphasized the need for random effects models in estimating localized activities with multiple subjects (e.g., Lazar et al., 2002; Mumford and Poldrack, 2007). Stephan et al. (2009) have recently proposed assessing and comparing particular models with multiple subject data using a hierarchical Bayesian model and updating a Dirichlet distribution.

Problem 4: distinct but overlapping variable sets

An experiment can yield different ROIs for different subjects. In the most unhappy case, that could be because different individuals have different processing mechanisms for an experimental task, but more optimistically it may be because of individual differences in strengths of BOLD responses and thresholding effects or due to anatomical variation (cf. Poline, 2003) or chance disturbances.

Using Gaussian distributions, Chen and Herskovitz (2007) developed a Bayes net procedure for identifying small sets of active voxels characteristic of ROIs for a group of subjects, differentiating such cluster sets between groups of different neurotypes, and estimating cluster associations within groups. Using only the ROIs shared by a group may yield an accurate model of causal connections among the shared ROIs, but, if omitted ROIs act as common causes of shared ROIs, the omission may seriously misrepresent the relationships among the shared variables. If an omitted ROI Y is merely an intermediary in a causal pathway from shared ROI X to shared ROI Z, neglect of Y will lose information of interest, but it will not invalidate the inference that X causes Z. If, however, Y causes X and Z but X and Z have no influence on one another, omission of Y will leave a spurious association between X and Z which may be difficult to separate from a real causal relation. Alternatively, neglect of data sets not containing some variables will, all else being equal, reduce the accuracy or informativeness of inferences.

Problem 5: varying delays in BOLD response

Brain tissues may vary in the time delay of hemodynamic responses to neural activity. A consequence of the indirect measurement of neural processes in fMRI studies is that these time delays may not be known. The synchronization of fMRI measurements of variables from distinct ROIs may therefore not be correct for the sequences of the underlying neural activities. Various authors (e.g., Breakspear et al., 2003) have noted that this fact poses a problem for causal inferences based on time series analysis, but it likewise poses a problem for every search method whether or not the resulting models have time-indexed variables. Causes cannot precede effects, and any method, such as multiple regression, that presumes a time order among the variables will therefore risk bias except in cases where such an order is actually known independently.

Problem 6: equilibrium or time series?

The BOLD signal from a neural region is a time series that is usually sampled intermittently, and in the course of an experiment may be influenced by a sequence of stimuli. The reconstruction of activation influences among regions–effective connectivity–may be attempted by modeling the joint, time-stamped measurements of variables associated with several ROIs as a time series, or by modeling them as "equilibrium" values resulting from an exogenous stimulation. (The latter might be viewed as a time series analysis with lag 0.) Each approach carries risks. The equilibrium approach is unrealistic if

measured BOLD responses to one exogenous experimental stimulation endure and influence measurements after a subsequent stimulus presentation. Since the interval between consecutive samples of voxels in the same ROI is typically slower than the neural influences to be modeled, the time series approach to model specification implicitly requires that lagged influences can be factored out, as by regression. We are aware of no comparative studies of results of these alternative approaches with real or simulated data; adequate studies would have to address the other problems noted here as well.

These problems certainly do not exhaust the challenges. They do not include, for example, the art of ROI selection. Again, ideally, one would like to use measures of the BOLD response to estimate variables that quantify neural activity in regions of interest, and use those estimates to search for causal models. Accurate estimates require prior knowledge of the mathematical relation between quantified variables representing aggregate neural activity over time and the BOLD signals they produce. Even if the neural activity/BOLD response is known and the function inverted, measurement errors in the fMRI signal may produce a component of variance in the estimate of neural activity that will, in principle, eliminate conditional independencies such as that illustrated above for L_1 and L_3 conditional on L_2. Whether in practice modeling is best done by assuming a definite mathematical relationship and estimating, or instead by using the fMRI measurements of the BOLD signal as noisy surrogates for neural activity, seems an open question. We will illustrate various methods without deconvolution estimates, but the solutions we propose would apply to deconvoluted estimates as well. A further issue, noted by a reviewer, is that fMRI acquisition and preprocessing schemes can result in the mixing of signals acquired at different points in time within a single region of interest, and it is likely that this occurred in the real fMRI data presented here (due to the use of interleaved slice acquisition). In principle, this could cause particular problems for methods, such as Granger causality, that rely upon relative timing of responses across regions. We will, however, provide some evidence that the interleaving of slices does not affect the accuracy of the solutions we offer below.

Addressing the problems using graphical causal models

We propose that modifications of machine learning techniques for graphical causal modeling available for more than a decade (Meek, 1997) provide a basis for addressing the problems we have sketched above. Appendix A presents a more detailed description of graphical causal models and more detail on the algorithms we use to learn these models from data. The software we used is available as freeware in the TETRAD IV suite of algorithms with a graphical user interface at www.phil.cmu.edu/projects/tetrad. Under some general assumptions (Chickering, 2002), Meek's method using the Bayes Information Criterion (BIC) provably provides consistent estimates in the large sample limit for a single data set. We generalize the procedure to multiple data sets, some possibly with missing variables, and provide a proof (see Appendix A) that the BIC score generalizes to this case. We will illustrate our adaptations with various data sets selected from an experiment with 13 subjects, and we will show, sequentially, how the first five problems can be addressed. The last problem will be addressed by constructing both 0 lag and 1 lag time series analyses as we proceed. Finally, we offer some simulations to confirm the methods.

Meek's Greedy Equivalence Search (GES) (Meek, 1997) begins with an empty graph whose vertices are the recorded variables and proceeds to search forward, one new connection at a time, over Markov Equivalence classes of DAGs. Each class of models with an additional edge is scored using BIC (Schwarz, 1978): $-2\ln(ML) + k \ln(n)$, where ML is the maximum likelihood estimate, k is the dimension of the model (in our cases, the number of directed edges plus the number of variables), and n is the sample size. The algorithm searches forwards from the empty graph until no improvement in BIC score is possible, and then backwards, and outputs a description of a Markov Equivalence class. In practice, the algorithm requires a computation of a series of maximum likelihood estimates, and is limited to cases where approximations to such estimates can be rapidly obtained. The implementation we use estimates maximum likelihood on a Gaussian distribution hypothesis, but this is evidently a second moment approximation. The graphical models output by GES may be parameterized and tested in conventional ways, for example as linear models with asymptotic chi square tests, but for reasons explained later, we do not recommend doing so for fMRI applications. For sparse graphs with the number of variables considered here (11 or fewer), GES runs on a conventional personal computer in less than a second.

The examples that follow employing real fMRI data are from a study (Xue and Poldrack, 2007) in which subjects performed rhyming judgments on words or pseudowords across separate blocks. Details regarding the design, acquisition, and data processing are provided in Appendix B.

Problem 1

We illustrate the ability of the GES algorithm (see Appendix A) to extract causal structure from fMRI data and compare it to a well-known search method that uses a Lagrangian modification index search (Bullmore et al., 2000). The data in this example are from an arbitrarily chosen one of 13 subjects performing rhyming judgments on either words or pseudowords across separate blocks (see Appendix B for more details). We consider for the moment only 5 left hemisphere ROIs (LOCC: left occipital cortex, LMTG: left middle temporal gyrus, LACC: left anterior cingulate, LIFG: left inferior frontal gyrus, LIPL: left inferior parietal) along with an input variable that represents the task, convolved with a canonical hemodynamic response. Both searches are conducted with the prior constraint that the input variable, *I,* is not an effect of the other variables, and, in the Lagrangian case, with the assumption that there is no feedback. The models obtained by each method are presented in Fig. 2, panels A and B. The simpler GES model has a p value of 0.31; the more complex model obtained with modification indices has a p value of 0.17, each by an asymptotic chi square test with the model as the null hypothesis. The example does not fully indicate the disadvantages of Lagrangian modification indices as a search method. Search by modification index fails to find the correct structure (no matter the sample size) even in simple linear systems with joint Gaussian distributions, for example in a system of four variables, $X_1,\ldots, X_4$ in which X_1 and X_2 are independent and each directly influences X_2 and X_3, neither of which influences the other.

Although the consistency proof for GES assumes no latent confounding, and models obtained with the algorithm will often admit alternatives with latent variables, in this example most edges in the model output cannot be replaced or supplemented by latent common causes without reducing the model fit. The consistency proof for GES also assumes that the true graph is acyclic, but feed-forward effects in non-linear cyclic systems can be identified by methods that call the GES algorithm, as we will illustrate later.

BOLD signals are sometimes modeled as an AR(1) or higher order autoregressive time series, with causal relations estimated by modified versions of Granger's (1969) criterion: X is regarded as a cause of Y in multivariate series X_t, Y_t, Z_t provided Y_{t+1} is better predicted (with least squares loss) by (X_{t-}, Y_{t-}, Z_{t-}) than by $\{Y_{t-}, Z_{t-}\}$, where subscript $t-$ denotes t and all preceding time steps. More generally, X is a Granger–cause of Y if Y_{t+1} is dependent on X_t– conditional on $\{Y_{t-}, Z_{t-}\}$. In practice the conditional dependence is estimated by multiple regression of each time series variable on a specified number of lags of itself and other variables. The Granger procedure, or variations of it, has been used in previous studies of fMRI data (e.g., Roebroeck et al., 2005). In the machine learning literature, statistical analogues of Granger's criterion have been given

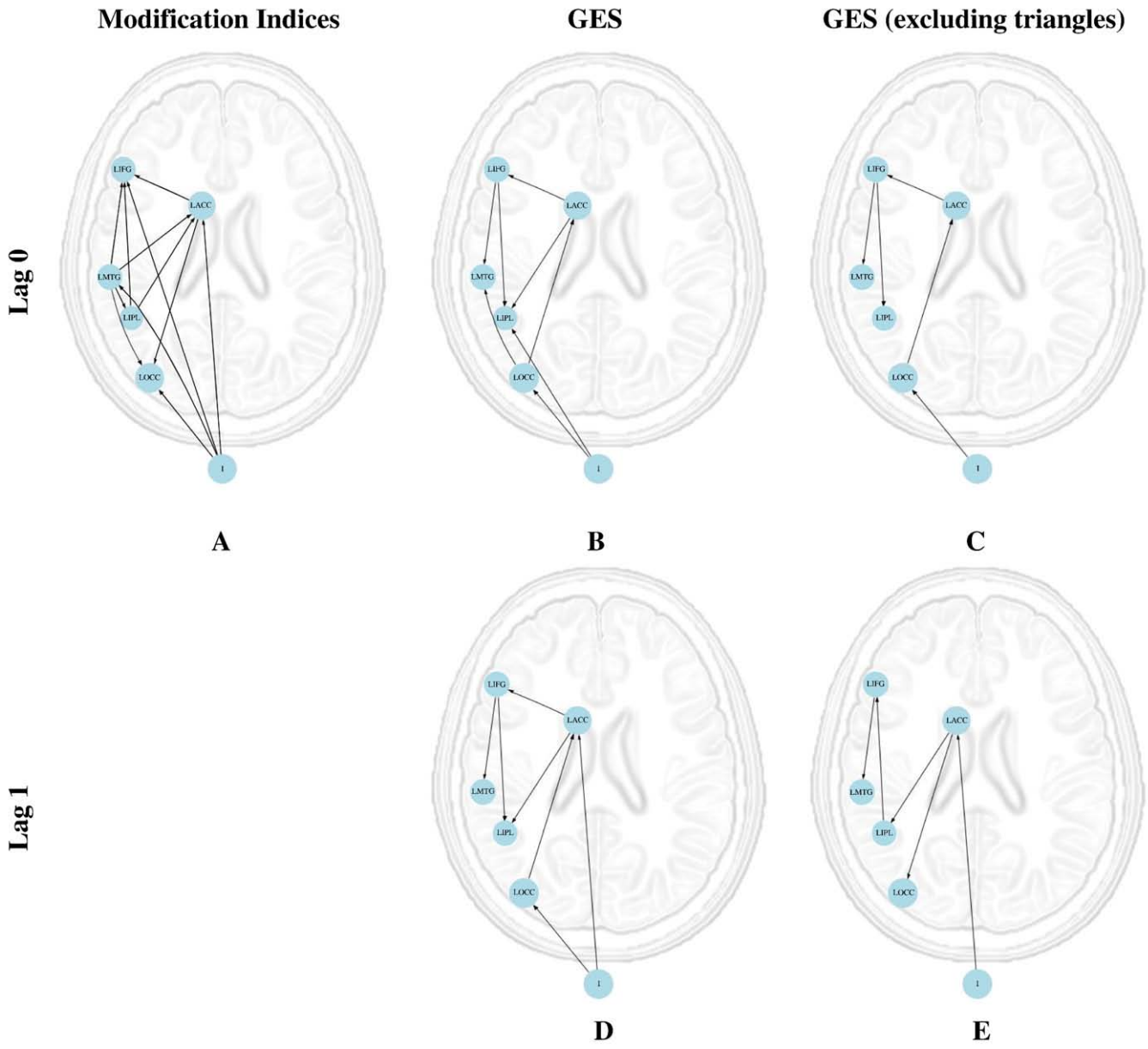

Fig. 2. Best-fitting lag 0 models for modifications indices (Panel A, $p = 0.17$) and GES (Panel B, $p = 0.31$), $c = 1$, and, as well as the first model for GES excluding triangles (Panel C, $p = 0.001$). For lag 1, the GES model for $c = 1$ is shown (Panel D, $p = 0.00003$), along with the first GES model excluding triangles (Panel E, $p = 0$).

graphical representations under the description "Dynamic Bayes Networks." Eichler (2005), provides a graphical generalization to allow for latent variables and gives an fMRI application. Granger himself later noted that his procedure can fail to capture causal relations when processes occur faster than the sampling rate, which is certainly the case when we wish to extract causal relations in neural activity from fMRI signals. For example, if X_t causes Y_{t+1} which in turn causes Z_{t+2}, and measures are taken only at intervals of 2 time steps or greater, Granger's criterion will yield the information that X causes Y and Y causes Z and X causes Z but will not yield the information that Y is the intermediate variable in the causal chain: an additional spurious direct connection between X and Z will be found, because Y_{t+1} is unobserved and therefore not conditioned on in the Granger regressions. To remedy such problems, Swanson and Granger (1996) proposed regressing the time series variables on lags and applying an earlier search procedure for causal models (Glymour et al., 1987) limited to finding causal chains, to the residuals of the time series regression. For example, X_{t+2}, Y_{t+2}, Z_{t+2} may be regressed on X_t, Y_t, Z_t . The regression coefficients are used to obtain predictions X^*_t, Y^*_t, Z^*_t for each time point t. The procedure should leave only associations produced by processes occurring more rapidly than the sampling interval. The residual values $(X_t - X^*_t)$ etc. at each time point are then treated as cases for analysis by a causal search procedure. Demiralp and Hoover (2003) instead applied a more general search procedure not restricted to causal chains (the PC algorithm, Spirtes and Glymour, 1991) to the Swanson–Granger residuals. We illustrate the results for the data from the previous data set discussed above, applying GES to the residuals obtained by regressing all time series variables on all variables one step previously. The lag 1 time series residual graph is shown in Panel D of Fig. 2; a linear model of the lag 1 residual graph fails a chi square test ($p = 0.001$).

Problem 2

The second problem is essentially that the class of cases considered in the discussion of problem 1 is the wrong one: we want the causal connections among the unobserved neural activities that are causes of the measured hemodynamic responses.

The noisy observation of indirect effects creates a difficult problem for the estimation of latent causal relations. As illustrated for Fig. 1 above, GES produces false "triangulations" of measured variables that are manifestations of a chain of three unmeasured variables. The problem is not particular to GES, but is a general problem about inference to latent structure.

With GES, the triangulation problem can be mitigated by increasing the penalty function in the BIC score. The ease with which the GES procedure finds directed edges is controlled by the penalty term $k\ln(n)$ of the BIC score. The penalty can be multiplied by any constant c greater than 1 and the causal estimator will be less likely to posit spurious transitive closures for chains of three variables. At the risk of missing some real causal connections, spurious causal connections can be reduced or eliminated by tuning the BIC penalty in GES so as to eliminate triangulated connections while retaining single connections and allowing direct connections between variables also connected by a path with more than one intermediate variable. The p values of the usual chi square test for linear Gaussian models will no longer be even a rough guide to model preference, since the search is designed to produce models that do not account for all of the conditional associations in the data.

For the left hemisphere lag 0 data from subject 1, the GES algorithm, excluding triangles, produces the model shown in panel C of Fig. 2. This model is a subgraph of the one in panel B. The graph obtained by treating the same data as a lag 1 time series, taking residuals and varying the penalty function until there are no triangles, is shown in Panel E. It seems plausible that the input directly influences the occipital region rather than influencing it via the anterior cingulate, and on those grounds the model of Panel C would be preferred to the model of Panel E. However, the experiment involved 13 subjects, and the analyses so far have used data for only 1 subject. Results should change, and do, when all subjects sharing a common set of ROIs are considered. The distribution of the disturbance terms is of interest for the adequacy of the Gaussian approximation, and this subject is typical. Estimating the distribution of the disturbances by the distribution of residuals after linear regression of each variable on its parents for the model in Panel C, the input and the occipital disturbances are not Gaussian, but the disturbance terms for other variables have Gaussian distributions.

Problem 3

Because directly combining datasets can result in statistical dependencies in the combined data that do not exist in any of the individual datasets, we developed a modification of the GES procedure (which we call IMaGES to abbreviate *Independent Multiple-sample Greedy Equivalence Search*) that allows for modeling of multiple datasets. Suppose there are m data sets D_i of sizes n (the issue is more complicated if the sample sizes differ considerably). For any graph G, let $\ln(D_i, G)$ be the natural log of the likelihood of the data determined by the maximum likelihood estimate of the free parameters in G, which for Gaussian linear systems are the variances of exogenous and error variables and all coefficients without fixed values. IMaGES follows the GES procedure but each stage scores each graph by using the average of the BIC scores for that graph in each of the individual data sets:

$$\text{IMscore} = -(2/m)\sum_i \ln(D_i, G) + ck\ln(n)$$

The IMscore is a BIC score with the number of independent dimensions multiplied by the number of data sets (see Appendix A). Markov equivalent graphs should receive the same IMscore when the distribution is Gaussian, but may receive slightly different scores for non-Gaussian distributions (which for BOLD maybe be more typical, Hanson and Bly, 2000). For independent samples from the same distribution, the IMscore is consistent whenever GES is consistent but with increasing size of each sample IMaGES converges more rapidly than does GES applied to only one data set of the same size.

The IMaGES models so obtained are non-parametric, although a parametric assumption is used to obtain maximum likelihood estimates in the BIC and IMscore. The graph expresses only hypotheses as to the causal connections and conditional independence relations, whatever the true underlying distribution may be. Consistent with the graph representation, two or more ROIs may influence another ROI interactively, as in our simulations described below. The proof of convergence of the BIC score to a representation of the posterior distribution assumes the parameters used in the likelihood computations are independent, but the IMaGES graphical models imply no posterior distributions over the parameters. The combination of parametric maximum likelihood estimates with a non-parametric generalization makes the IMaGES procedure heuristic, although the linear, Gaussian parametrization should identify the correct Markov equivalence classes when the true distributions imply that independence and conditional independence are accompanied by zero correlations and partial correlations.

A sensible heuristic strategy is to run IMaGES with a systematically increasing penalty constant, c, until triangulations of variables disappear or the input becomes disconnected–whichever comes first–placing more trust in edges that are more robust as the penalty increases.

After excluding two ROIs missing from the data of four individuals, IMaGES was applied to the data from the 9 of 13 subjects who had activation clusters in all of the ROIs, including both the left and right hemispheres. The results are shown in Fig. 3, for both lag 0 and for lag 1 residual data. When applied to lag 0 data, the results become strikingly reasonable as the cost function is increased to eliminate triangles. These graphs were obtained only with prior specification that the input variable is not an effect of the anatomical variables but with no priors on the anatomical connections, and no specification, prohibition, or priors for influences of the input variable on the anatomical variables. Without any such information, IMaGES was able to reconstruct the canonical left-hemisphere language network and its influence over the right-hemisphere homologs.

IMaGES applied to the lag 1 residuals produces no triangles when $c = 6$, and the resulting graph (Fig. 3D) is nearly identical to that obtained from the lag 0 data (Fig. 3C), with one edge reversed.

Problem 4

An even more difficult problem arises when data are not available for all regions of interest across all subjects; this can occur, for example, if activation clusters are used to define the regions but some subjects do not have clusters in some regions of interest. In statistics, multiple data sets with distinct but overlapping sets of variables have been treated by "multiple imputation" (Rubin, 1987). However, this procedure assumes a model in which the unshared variables are linear functions of the shared variables, and can produce erroneous results when this assumption is false.

The full data set for the experiment we are considering has 13 subjects and 11 variables; although some subjects were missing data from some regions (due to the requirement of significant activation to create the ROI), variables for all ROIs were present in data for 5 of the subjects. We applied IMaGES just as before, but for each data set we computed IMscores only for edges connecting the variables in that data set. If it assumed only that the multiple samples share a common factorization up to Markov equivalence, we conjecture, but have not proved, that the IMaGES procedure converges on a collection of data sets whenever GES converges on each data set singly, provided at least one data set contains all variables in any data set.

The results with the lag 0 time series data and also with residuals from a lag 1 time series are shown in Fig. 4 (LTMG and RTMG are left and right temporal medial gyrus, respectively). The results are similar

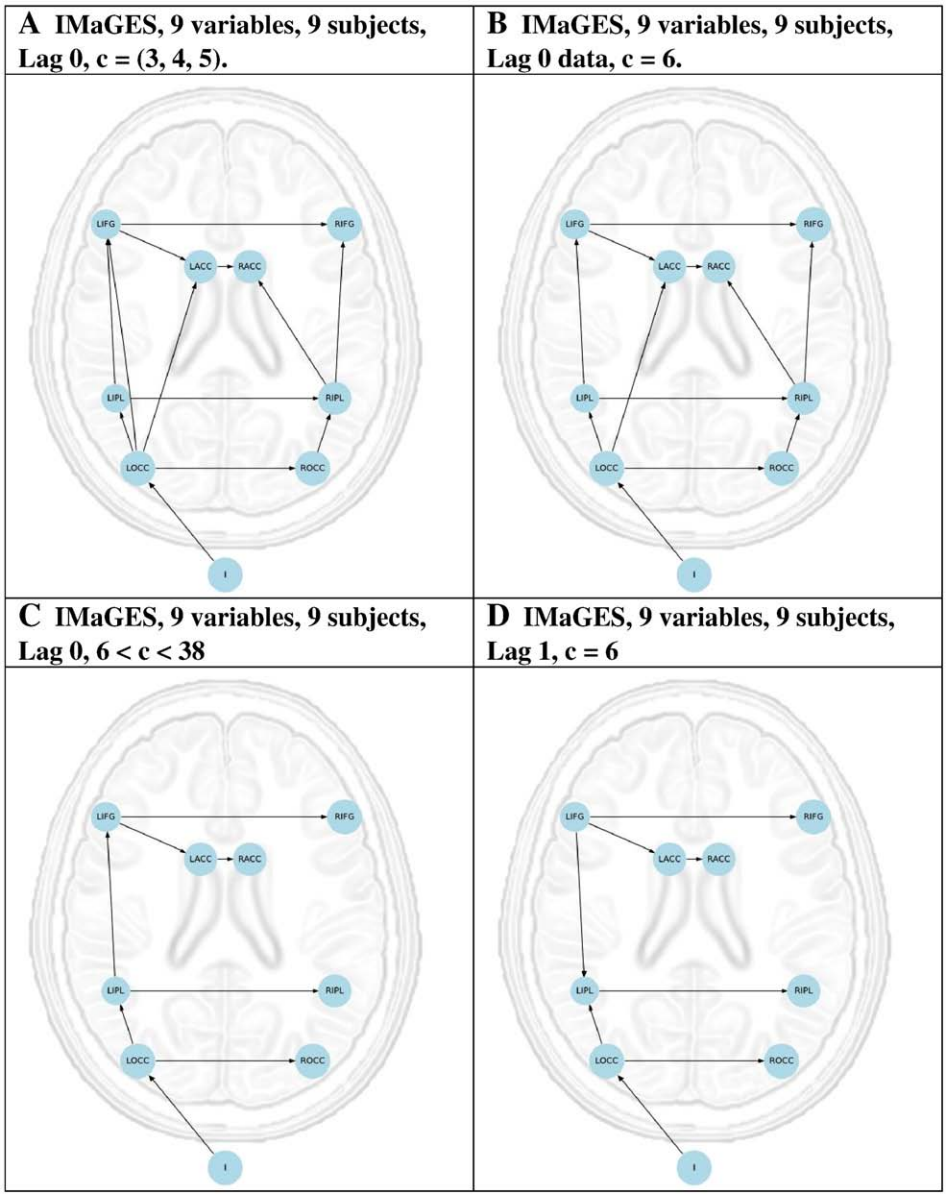

Fig. 3. Best fitting IMaGES models of 9-variable, 9-subject lag 0 data as c increases (Panels A, B, C) and lag 1 data for $c = 6$ (Panel D). (See text for details.)

to those obtained in Fig. 3 using common ROIs, but also detect additional connections, including some connections (e.g., LOCC→LACC) between regions that were suggested in the earlier analysis of left hemisphere variables.

Problems 5 and 6

A fundamental challenge to causal modeling is that we wish to infer the direction of influence between variables from sample data alone, apparently in contradiction to the truism that correlation does not imply causation. All of the DAGs within the same Markov Equivalence Class share the same variables and the same adjacencies, but the DAGs within a class may differ in the directions given to one or more edges. In some cases, as with the results of our analyses of the experimental data, but not always with the simulated data we will later describe, all directions are uniquely determined in a Markov Equivalence Class. When a unique direction is given to an edge in the output of the search, the direction is not arbitrarily chosen: any DAG that implies the same independence and conditional independence relations and has no unrepresented common causes must so direct that edge.

The indirect measurement of neural activity through BOLD responses produces a further problem of timing. It is well known that the parameters of hemodynamic responses (including delays) may differ substantially across regions within an individual (e.g. Miezin et al., 2000). This means that the BOLD signal at time t in one region may reflect neuronal activity at time $t - \Delta_1$, whereas the BOLD signal at the same time in another region may reflect neuronal activity at time $t - \Delta_2$, where Δ_1–Δ_2 may be on the order of seconds. Thus, unknown to the search algorithm or the investigator, the fMRI data may reflect neuronal signals at:

$$x_{t+1}, y_t, z_t$$
$$x_{t+2}, y_{t+1}, z_{t+1}$$

etc., whereas the actual events are in the sequence

$$x_t, y_t, z_t$$
$$x_{t+1}, y_{t+1}, z_{t+1}$$

etc. If the second of these sequences, the actual one, is i.i.d., then in the first, measured sequence, X will be independent of Y and of Z, even if

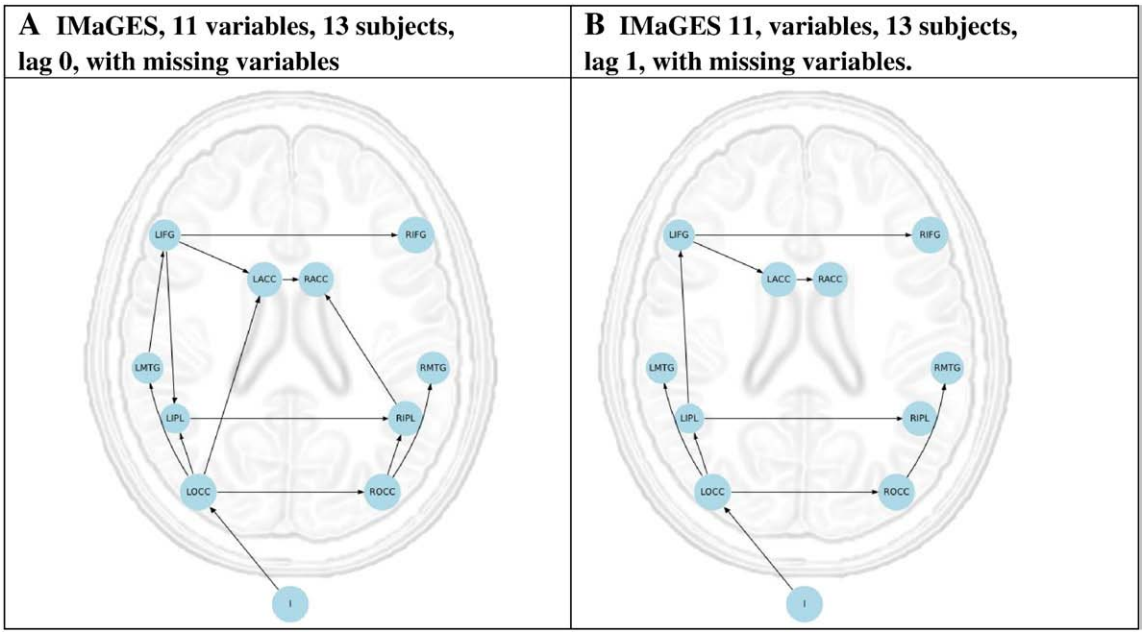

Fig. 4. Best fitting lag 0 and lag 1 IMaGES models of 11-variables, 13-subject data, allowing for missing ROIs across individuals, reporting the first model excluding triangles as c increases.

there are strong causal relations connecting X with Y and with Z. In that case, no matter the sample size, GES will be biased against any causal connections involving X, and so will any other search method that treats the simultaneously measured variables as due to simultaneous causes. In fMRI designs, however, each variable is typically autocorrelated: the value of X at t is not entirely independent of the value of X at $t+1$ or possibly of other variables. In that case, mis-synchronization will remove some relations of independence conditional on X, and may result in the large sample limit in estimation of one false edge in a triangle of edges.

We address this problem by time shifting the series of values for one or more variables backwards in various ways to obtain a collection of data sets with the same backshift, applying IMaGES to each resulting collection of data sets, and choosing the resulting model with the best BIC score. Backshifting reduces the sample size by the maximum over the variables of the number of shifts. Ideally, every plausible combination of time shifts of variables should be considered, but combinatoric considerations prevent that. A greedy heuristic is to time shift variables one at a time and if the BIC score is not improved by a shift to delete that shift, but we have found that procedure unreliable. Instead, we bound the size of the set of shifted variables and consider all shifts over those variables up to a maximum shift. The run time requirements for a search of this kind, for all 13 subjects, bounding the cardinality of each set of shifted variables at 3, is about 10 min. The procedure can obviously be implemented in parallel with reduced run time. This method was applied to the rhyme judgment dataset presented above. Both for 0 lag and for lag 1 residuals analysis, for all of the 13 subject data sets and for each combination of 1, 2 or 3 of the 11 variables, except I, we backshifted the variables in all combinations of 0.5, 1, 1.5 or 2 recorded time steps, corresponding to shifts between 1 and 4 s. In all cases the IMaGES models without shifts had the best BIC scores. This suggests that, for the blocked design data examined here, there are no detectable differences in timing between regions of interest.

Simulation studies

In view of the theoretical difficulties with inference to latent, non-linear relations between sets of neurons, some investigation by simulation of the accuracy of the IMaGES procedures is needed. To address this, we simulated causal processes in which a large number of unrecorded latent variables, L (e.g., synaptic activity in each of a set of cortical columns), influence one another non-linearly, with feedback, and combine with a hemodynamic response function to produce a multivariate time series of values of the recorded X variables (fMRI data in regions of interest). This simulation framework is meant to reproduce, so far as possible, the circumstances of the Xue and Poldrack experiment, but with randomly selected effective connections among the ROIs. It is not meant to simulate faithfully the details of neuronal activity, but rather as an abstract representation of the relation between neuronal activity in voxels, aggregated voxel activities within a ROI, and fMRI signals. Typical signal to noise ratios were estimated for the Xue and Poldrack data and parameters of the simulation were adjusted approximately to match. The empirical disturbance variances were estimated as the residuals after regressing each variable in the GES model of Fig. 1 on its parents. Signal to noise was represented as the difference of the variance of a variable and its disturbance variance, divided by its disturbance variance. Further details regarding the simulation model are provided in Appendix C.

Two kinds of simulated data sets were generated, one for 9 variables and 9 subjects and one for 11 variables and 13 subjects, using randomly generated graphs of ROIs respectively with 9 and with 11 edges, in addition to a backprojection for each feedforward connection. For each experiment, the entire simulation (with a new randomly generated DAG representing feed-forward effective connectivity) and search were repeated 10 times and the error rates averaged over the ten runs. Table 1 shows the results of the experiment with 9 variables, 9 subjects, no missing ROIs and no time shifts. Table 2 shows the results of the experiment with 11 variables, 13 subjects, and up to 3 variables missing at random for the data for any subject on any run. Table 3 is for an experiment like that of Table 1, except in each run the records of three randomly chosen variables are randomly shifted forward in time 0.5, or 1, or 1.5 or 2 sampling intervals for all subjects. Table 4 is for an experiment like that of Table 2, except again in each run the records of three randomly chosen variables randomly shifted forward in time 0.5, or 1, or 1.5 or 2 sampling intervals for all subjects, representing the simulation of BOLD delay differences between 1 and 4 s. The error rates can be obtained by dividing the number of errors by the number of feed-

Table 1
Errors for graphs recovered by IMaGES on simulated 9-variable, 9-subject data, with no missing ROIs across individuals, without voxelwise or ROI shifts.

Run	Lag 0, first nontriangular			Lag 1, $c=1$			Lag 1, first nontriangular		
	AFP	AFN	DE	AFP	AFN	DE	AFP	AFN	DE
1	1	2	2	0	1	0	0	1	0
2	1	2	0	0	1	2	0	1	2
3	0	0	1	2	0	2	0	5	1
4	0	0	4	3	1	2	0	1	2
5	1	2	1	1	2	1	1	2	1
6	0	0	1	0	1	1	0	1	1
7	1	2	2	2	0	2	0	2	2
8	1	0	1	1	0	1	0	2	1
9	0	0	2	0	1	1	0	1	1
10	1	0	1	2	2	2	2	2	2
AVG	0.6	0.8	1.5	1.1	0.9	1.4	0.3	1.8	1.3

AFP = adjacency false positives, AFN = adjacency false negatives, and DE = orientation errors for directed edges only (number of directed edges in the output that are oriented differently from the Markov equivalence class). See text for details.

Table 3
Errors for ROI shifting experiments for 9-node, 9-subject simulated experiments, with no missing ROI's across individuals, without voxelwise shifts.

Run	Shifted, lag 0, no triangle			Backshifted, lag 0, no triangle			Backshifted, lag 1, no triangle		
	AFP	AFN	DE	AFP	AFN	DE	AFP	AFN	DE
1	2	2	2	4	1	1	0	1	0
2	1	2	0	0	0	0	1	5	1
3	1	0	0	0	0	1	1	2	2
4	3	1	4	0	0	4	2	1	1
5	1	2	1	1	1	1	1	5	1
6	0	0	2	0	0	2	0	1	1
7	2	2	2	1	1	2	3	1	2
8	2	2	1	1	0	1	0	0	1
9	1	1	2	0	0	2	1	2	2
10	4	2	1	1	0	1	0	2	0
AVG	1.7	1.4	1.5	0.8	0.3	1.5	0.9	2	1.1

AFP = adjacency false positives, AFN = adjacency false negatives, and DE = orientation errors for directed edges only (number of directed edges in the output that are oriented differently from the Markov equivalence class). See text for details.

forward edges in the simulation, 9 for Tables 1 and 3, and 11 for Tables 2 and 4. BOLD delays were selected at random for simulated (in terms of the sampling interval) 1–s, 2-s, 3-s and 4-s delays, but the search methods shifted measurements only by 2 s or 4 s in order to represent circumstances where the true BOLD delay is a fraction of the shifts used in search. Fractional delays are most difficult for search when the delays occur at mid-point between the variable shifts.

Errors are computed with respect to the feed-forward (from the Input variable) substructure of the randomly generated effective connectivity graphs. None of the 3 methods we consider dominates the others. Error rates for stopping at the first non-triangulated variable using a 0 lag vary from 6.6% for false positive adjacencies with 9 variables and no BOLD delay shifts, no missing variables, to 16.6% for 11 variables with 3 variables missing at random and up to 3 variables shifted in their BOLD delays by as much as 4 s. For the same method, errors of direction of edges are in all cases no higher than 16.6 %. Error rates for false negatives are in all cases and by all methods inflated because all three methods deliberately search for a sparse, connected substructure without triangulated variables, but the simulation procedure is allowed to generate graphs with such triangles. Again for the 0 lag procedure, the false negative adjacency error rate varies from 9% for 9 variables without missing ROIs and without shifted BOLD delays, to 25.5% for 11 variables with missing ROIs and varying BOLD delays. (A more informative count of false negatives, which we have not done, might include only edges in a non-triangulated substructure of the feed-forward graph).

One reviewer has suggested that significant BOLD delays may occur between voxels within the same ROI. We simulated that possibility by allowing all variables within a ROI to be shifted at random according to a uniform distribution on a simulated [0, 0.5] second interval, representing completely asynchronous activity of voxels within a half-second interval. The results, when this asynchronous voxel shifting is combined with the random delays in BOLD response for up to 3 ROIs (resulting in a maximum BOLD delay difference for voxels in different ROIs of up to 4.5 s), and up to 3 variables are missing at random for each of 11 subjects, is shown in Table 5. The error rates increase, for example for the 0 lag method to just under 25% for false positive adjacencies, just under 14% for orientation errors (not counting the orientations of false adjacencies in the numerator or denominator in the ratio of errors to cases), and just under 20% for false negative adjacencies.

To illustrate the value of applying IMaGES to a collection of data sets from multiple subjects, rather than applying a search such as GES to individual subjects, we have simulated a 9 variable, 9 subject version of the Xue and Poldrack experiment, using the graph of Fig. 3B to represent effective connectivity relations. IMaGES was run on the collection of data sets and GES was run separately on each individual data set. The comparisons of edges produced by these various procedures with those in the true graph are given in Table 6.

The table shows remarkable effects of analyzing the data collectively via IMaGES rather than, for example, running GES individually and using a voting procedure for edges. Only one of the

Table 2
Errors for graphs recovered by IMaGES on simulated 11-variable, 13-subject data, with missing ROIs across individuals, without voxelwise or ROI shifts.

Run	Lag 0, first nontriangular			Lag 1, $c=1$			Lag 1, first nontriangular		
	AFP	AFN	DE	AFP	AFN	DE	AFP	AFN	DE
1	0	0	2	1	0	3	0	3	1
2	0	1	0	1	3	0	1	6	0
3	2	2	2	2	2	3	0	5	0
4	2	1	2	0	1	0	0	1	0
5	1	2	1	2	3	2	0	5	1
6	1	1	1	0	2	1	0	2	1
7	2	3	2	2	0	3	0	2	2
8	2	3	2	1	1	2	0	5	1
9	2	3	1	1	4	0	1	4	0
10	0	1	0	1	2	3	1	2	3
AVG	1.2	1.7	1.3	1.1	1.8	1.7	0.3	3.5	0.9

AFP = adjacency false positives, AFN = adjacency false negatives, and DE = orientation errors for directed edges only (number of directed edges in the output that are oriented differently from the Markov equivalence class). See text for details.

Table 4
Errors for ROI shifting experiments for 11-node, 13-subject simulated experiments, with missing ROI's across individuals, without voxelwise shifts.

Run	Shifted, Lag 0, no triangle			Backshifted, Lag 0, no triangle			Backshifted, Lag 1, no triangle		
	AFP	AFN	DE	AFP	AFN	DE	AFP	AFN	DE
1	1	0	2	0	0	2	1	2	3
2	0	1	0	0	1	0	0	3	1
3	2	3	2	1	2	2	0	3	0
4	3	2	2	2	1	2	1	3	0
5	2	3	1	2	3	1	2	4	1
6	2	2	1	2	2	1	1	4	2
7	1	4	0	1	4	0	2	0	2
8	2	3	1	1	2	3	1	0	0
9	2	3	2	2	3	2	0	4	1
10	0	2	0	0	1	0	2	5	1
AVG	1.5	2.3	1.1	1.1	1.9	1.3	1.0	2.8	1.1

AFP = adjacency false positives, AFN = adjacency false negatives, and DE = orientation errors for directed edges only (number of directed edges in the output that are oriented differently from the Markov equivalence class). See text for details.

Table 5
Errors for ROI shifting experiments for 11-node, 13-subject simulated experiments, with missing ROI's across individuals, with additional random voxelwise shifts.

Run	Shifted, Lag 0, no triangle			Backshifted, Lag 0, no triangle			Backshifted, Lag 1, no triangle		
	AFP	AFN	DE	AFP	AFN	DE	AFP	AFN	DE
1	0	0	1	1	2	1	1	1	2
2	7	4	0	2	3	1	4	2	2
3	5	1	2	5	1	2	1	1	1
4	3	3	2	3	3	3	0	1	4
5	4	3	1	3	3	1	1	4	2
6	1	2	1	1	1	0	4	1	0
7	2	2	3	2	2	1	1	5	0
8	1	1	0	1	1	0	4	2	3
9	2	2	1	3	3	2	2	3	0
10	2	3	0	1	2	0	1	1	3
AVG	2.7	2.1	1.1	2.2	2.1	1.1	1.9	2.1	1.7

AFP = adjacency false positives, AFN = adjacency false negatives, and DE = orientation errors for directed edges only (number of directed edges in the output that are oriented differently from the Markov equivalence class). See text for details.

Table 7
Errors for graphs recovered by IMaGES on simulated 9-variable, 9-subject data, without voxelwise or ROI shifts, but with slice measurements in the order 1, 3, 1/4, 19, 2, 4, …, 20.

Run	Lag 0, first nontriangular			Lag 1, $c=1$			Lag 1, first nontriangular		
	AFP	AFN	DE	AFP	AFN	DE	AFP	AFN	DE
1	2	1	1	1	1	2	0	6	0
2	0	0	0	1	0	2	0	2	2
3	1	2	0	0	5	0	0	5	0
4	0	1	0	0	4	1	0	4	1
5	0	1	0	0	0	2	0	0	2
6	2	3	1	2	1	1	2	1	1
7	2	3	0	3	0	2	0	1	1
8	3	2	1	1	0	1	1	2	2
9	1	2	0	0	6	0	0	6	0
10	1	2	0	3	3	0	2	6	0
AVG	1.2	1.7	0.3	1.1	2	1.1	0.5	3.3	0.9

AFP = adjacency false positives, AFN = adjacency false negatives, and DE = orientation errors for directed edges only (number of directed edges in the output that are oriented differently from the Markov equivalence class). See text for details.

9 true edges occurs in the majority of the GES analyses, but IMaGES omits only two adjacencies and reverses one direction.

Finally, we consider (Table 7) the effect of interleaving slices in signal acquisition. The 9 variable, 9 subject simulation above (without time shifting) was modified in a separate experiment to simulate the order in which individual slices are measured in a typical fMRI experiment. We simulated a measurement sweep of 20 layers but with slices measured in the order 1, 3,…, 19, 2, 4,…, 20, where 1 is the bottom-most slice and 20 is the top-most slice. Again, each ROI R consists of 50 variables, but this time arranged as a 10×5 grid of variables $V(i, j, R)$, $i=1,\ldots, 10$, $j=1,\ldots, 5$. Within R, each variable $V(i, j, i)$ is influenced by variables $V(k, l, R)$ such that $\sqrt{(i-k)^2-(j-l)^2}<1.5$, each such edge being added with 50% probability. R is then positioned randomly in the vertical direction—that is, a random m is chosen, $1\le m\le 15$, and for each $i=1,\ldots, 10$, $j=1,\ldots, 5$, $V(i, j, R)$ is positioned at slice $j+m$. The simulation proceeds otherwise as before, with the exception that variables are recorded in order of slice, in the order 1, 3,…, 19, 2, 4,…, 20, with the sequence repeated until the end of the simulation.

We have not investigated how robust these results are over variations in the vertical distance between ROIs connected by a directed edge in the simulation.

It is essential to the success of GES and IMaGES search procedures that the data generating process is non-linear. Application of these methods to data generated from linear, cyclic SEMs would produce much higher error rates. For that reason, as with DCM models, we do not assign parameter values such as linear coefficients to edges of DAGs obtained from IMaGES or GES searches. Such assignments can be forced, for example by regressing each ROI on its parents using the simulated ROI values, and, using as empirical estimates of these pseudo "latent" parameters the corresponding regressions using the simulated measured variables and the DAG obtained from search. The estimates so obtained tend to be about 50% too high, but they are in any case quite inappropriate estimates of effective connectivity.

Discussion

The IMaGES algorithm is based on a procedure, GES, that has been proven consistent only for feed-forward, acyclic causal structures, whereas the simulated structures have feedback. Despite this, because of the non-linearity of the systems, IMaGES is able to extract the feed-forward structure with reasonable accuracy on large (for fMRI ROI studies) numbers of variables with multiple subject data, missing variables, and varying time delays. The general behavior of the IMaGES algorithm on the simulated data closely resembles the behavior of the same algorithm on the empirical data. With increasing penalty for free parameters, a sparse, connected, partially oriented, graphical structure emerges from multiple data sets. The problem of variations in BOLD response among brain regions can be addressed by explorations of time shifted data. Even with assumptions we think unlikely about varying time delays in BOLD responses between voxels in the same ROI, compounded with unknown BOLD delays between ROIs and missing ROIs in some subjects, the procedures return uncertain but useful information. In our simulations with such realistically difficult data, 75% of the adjacencies in the graphical output of the IMaGES search occur in the "true" graph from which the simulation data were generated.

Table 6
Accuracy of IMaGES versus GES for a single 9-node, 9-subject experiment.

	I → LOCC	ROCC → RIPL	LIPL → RIPL	RIPL → RIFG	LIFG → RIFG	LOCC → ROCC	LOCC → LIPL	LOCC → LACC	LIPL → LIFG	RIPL → RACC
IMaGES	1	1	1	1	1	0	−1	1	0	1
GES1	1	0	0	1	1	0	0	0	1	0
GES2	1	0	1	1	0	0	0	0	0	1
GES3	1	0	0	0	1	0	0	0	0	0
GES4	0	1	0	0	0	1	0	1	0	0
GES5	1	0	0	0	−1	−1	0	1	−1	0
GES6	1	0	0	0	0	0	−1	0	0	0
GES7	1	0	1	1	1	0	0	1	0	1
GES8	1	1	1	1	0	0	−1	1	0	1
GES9	1	0	0	1	1	0	0	0	0	0

The graph in Fig. 3B was used to create a simulation as described in Appendix C. Each column records occurrences of edges in Fig. 3B. The first row records occurrences of these edges for the first nontriangular model found by IMaGES, using all nine data sets; GES1 through GES9 record the first nontriangular model found by GES for each of the nine data sets taken individually. Where X and Y range over the ROIs in Fig. 4B, "1" is recorded in a given row for edge X → Y just in case the model in that row either contains X → Y or X–Y; a "0" is recorded just in case X and Y are no adjacent in the model for that row, and a "−1" is recorded just in case the model for that row contains Y → X.

With the empirical data, analysis with lagged residuals is more conservative than 0 lag analysis. There need not, however, be a decision between analyzing data with 0 lags or with residuals from one (or more) lags where the results of the different analyses are in good agreement or share parts of their structure of interest. In our view, actual conflict of results obtained by the different methods represents uncertainties that can only be resolved by further, independent information.

Our simulation studies leave open a number of issues: to deconvolute or not; whether and how IMaGES accuracy depends on the vertical distance of direct connections of ROIs; and the accuracy of the procedures under varying possible topologies of voxel to voxel connections between ROIs. The "topographic" model (Kaas, 2004) of connections between ROIs we use in our simulations is reasonably established anatomically, but its role in information processing in the brain is not.

The IMaGES procedures can be modified in several ways. As noted previously, they can be used with, or without, deconvolution estimates. Maximum likelihood estimation other than the Gaussian procedure we have used may be used in computing the BIC score, and, in principle, approximations to a posterior distribution other than the BIC score might be used. We have not investigated such alternatives.

Although estimates of linear coefficient parameters and standard errors using Gaussian distribution theory could be given for any of the graphical models we have produced, we deliberately have not done so. In our simulations, inter-regional influences are collective, non-linear and interactive, with no obvious parameterization (as by linear coefficients) that accurately represents strengths of influences and could be estimated from the measured variables.

The reproducibility of the results of search procedures has several dimensions. Besides differences between results in similar experiments from different laboratories and the expected variation in GES results across subjects, GES results may vary within subjects over multiple repetitions of an experiment. When multiple experimental repetitions for each subject are available, and there are multiple subjects, IMaGES reproducibility can be assessed within subjects and across experimental repetitions, across subjects within experimental repetitions, and as a whole. If effective connectivity is eventually to be of use in differentiating clinical groups, systematic studies of reproducibility, and within group and between group variation, in neurotypicals and various patient groups will be essential.

Model scoring search procedures have previously been applied to fMRI data (Zheng and Rajapake, 2006; Chen and Herskovitz, 2007) but the combination of challenges we have reviewed above have not been addressed. Methodological proposals for uncovering effective connectivity from fMRI data should at least take account of the problems of indirect measurement, multiple subjects with varying sets of ROIs, and varying BOLD response delays. Ideally, they should also allow for the possibility of associations among ROIs that are not represented by any recorded ROIs, and they should capture, or at least be robust against, feedback relations. Our search is not robust against confounding of represented ROIs by sources of covariation not represented in the ROIs of any of the subjects, and no search method is currently available that addresses this problem in conjunction with the others just noted. (When, however, the directed graph is uniquely oriented, as in our empirical study, latent confounding of downstream variable pairs would prevent the conditional independence relations the IMaGES search finds.) Independent components methods (Hoyer et al., 2006; Lacerda et al., 2008) produce more precise information than Markov Equivalence Classes, and can learn cyclic graphs, but do not generalize to multiple subjects. When some or all variables are not Gaussian, search methods using conditional independence tests derived from Hilbert kernels yield more informative results than Markov Equivalence classes; they, can be (but have not yet been) extended to make correct inferences for DAGs when there are unknown latent confounders, generalized for multiple independent data sets, or extended to learn cyclic graphs. But no methods are known that correctly search for information about cyclic graphical structure when there may be unknown latent confounders or when the dependencies among measured variables are non-linear. Expanding the information about the neural processes supporting cognition that can be obtained from fMRI investigations will require further development of statistical machine learning methods.

Acknowledgments

This research was supported by a grant from the James S. McDonnell Foundation. We thank Linda Palmer, Peter Spirtes and Richard Scheines for valuable discussions.

Appendix A: Graphical causal models and search algorithms

About DAGs

A set of random variables $\{X_1,\ldots, X_n\}$ is said to factor according to a DAG G, if for all values of $x_1,\ldots, x_n$ of $X_1,\ldots,X_n$, the density $d(x_1,\ldots, x_n) = \Pi_I\, d(x_i \mid \mathbf{PAR}_{G(x1,\ldots, xn)}(x_i))$, where $\mathbf{PAR}_{G(x1,\ldots, xn)}(x_i))$ is the values in x_i of the parents of X_i (i.e., the variables in G with edges directed into X_i). Specific assumptions about linearity, etc. can be added, but the representation is general over arbitrary smooth functional forms with independent disturbances. The factorization determines a unique set of independence and conditional independence relations among the variables. The Markov assumption is that these independence relations hold in any probability distribution for a causal model whose causal structure is given by G. A further assumption, Faithfulness, is that the probability distribution for a model satisfies *only* those independence and conditional independence relations implied by the factorization of the causal structure. For linear systems, Faithfulness amounts to the assumption that the variables are not deterministically related and that the correlations and partial correlations produced by multiple pathways between two variables do not perfectly cancel one another. Let Θ_G be the set of probability distributions that factor according to DAG G. G and G' are said to be Markov equivalent if and only if $\Theta_G = \Theta_{G'}$ or, in other words, if their factorizations are mathematically equivalent. Markov equivalence classes are sometimes singletons but often not.

With i.i.d. sampling, or stationary time series, the Markov and Faithfulness Assumptions are sufficient to allow several consistent, unbiased search algorithms (Spirtes et al., 1993, 2000; Demiralp and Hoover, 2003; Meek, 1997; Richardson, 1996; Chickering, 2002; Silva et al., 2006; Zhang and Spirtes, 2008). Another set of algorithms (Ramsey et al., 2006) in principle allows confidence intervals assuming a modification of Faithfulness, although practical computational procedures for computing the intervals are not available. Still another set of algorithms (Hoyer et al., 2006), does not require Faithfulness, but does require i.i.d. sampling, non-determinism, linearity, and that the disturbances or random errors be non-Gaussian. (In the data we consider, most of the variables have Gaussian residuals.) These various algorithms apply to linear and additive non-linear time series, linear or approximately linear equilibrium systems, various non-linear equilibrium systems, linear and approximately linear systems with Gaussian and non-Gaussian variables, feedback systems and systems with latent variables.

GES

The Greedy Equivalence Search (GES) (Meek, 1997), begins with an empty graph whose vertices are the recorded variables and proceeds to search forwards, and then backwards, over Markov

equivalence classes. To approximate a posterior probability distribution, a DAG representative of each considered equivalence class with an additional edge is scored using the Bayesian Information Criterion (BIC) (Schwarz, 1978): $-2\ln(\mathrm{ML}) + k\ln(n)$, where ML is the maximum likelihood estimate, k is the dimension of the model and n is the sample size. For uniform priors on models and smooth priors on the parameters, the posterior probability conditional on the data is a monotonic function of BIC. At each stage, the equivalence class (or more exactly, a representative DAG in the class) with the best BIC score is chosen. This may require reversals of edges directed in a representative of a previous equivalence class. For example if the representative at some stage n is $X \rightarrow Y \rightarrow Z$ and the best edge to add is $W \rightarrow Z$, the DAG representing the new equivalence class will be $X \leftarrow Y \leftarrow Z \leftarrow W$. This forward stage continues until no further additions improve the BIC score. Then a reverse procedure is followed that removes edges according to the same criterion, until no improvement is found. The computational and convergence advantages of the algorithm depend on the fact that it searches over Markov equivalence classes of DAGs rather than individual DAGs. In the large sample limit, GES identifies the Markov equivalence class of the true graph if the assumptions above are met.

Pseudo-code for the GES algorithm is given below, following GES as presented in Chickering (2002). A *pattern* is a graph with directed (→) and undirected (–) edges which represents an equivalence class of directed acyclic graphs (DAGs), as follows. Two nodes X and Y are adjacent in a DAG or pattern G, notated adj(X, Y, G), just in case there is an edge connecting X and Y in G. A *path* (of length $n-1$) is a sequence $\langle X_1,\ldots,X_n\rangle$ of nodes in G such that for each $i = 1, \ldots, n-1$, X_i is adjacent to X_{i+1} in G. A collider in G is a path of length 2 of the form $X \rightarrow Y \leftarrow Z$, for X, Y, and Z in G. In a pattern P, each directed edge in P is so directed in each of the DAGs in the equivalence class, and each collider in each DAG of the equivalence class is a collider in P. A semidirected path from X to Y in G is a path $\langle X_1, \ldots, X_n = Y\rangle$ such that for each $i = 1$ to $n-1$, $X_i \rightarrow X_{i+1}$ or X_i–X_{i+1}.

GES(D)

(1) G ← ∅
(2) S ← the score of G
(3) <G, S> ← ForwardEquivalenceSearch(G, S)
(4) G ← BackwardEquivalenceSearch(G, S)
(5) Return G

ForwardEquivalenceSearch(G, S)

1. Repeat while $E_0 \neq \emptyset$
   1. $E_0 \leftarrow \emptyset$
   2. $T_0 \leftarrow \emptyset$
   3. $S_0 \leftarrow \emptyset$
   4. For each edge E' = X → Y such that ~adj(X, Y, G)
      1. T-Neighbors ← nodes Z such that Z – Y and ~adj(Z, X, G)
      2. For each subset T of T-Neighbors
         1. G′ ← a DAG in G
         2. S′ ← S + ScoreEdgeAddition(G, E′, T)
         3. If S′<S & S'<S_0 & ValidInsert(G, E′, T)
            1. E_0 ← E'
            2. T_0 ← T
            3. S_0 ← S'
   5. If $E_0 \neq \emptyset$
      1. Add E_0 to G
      2. For each T in T_0, if T –Y in G, orient T – Y as T → Y.
      3. S ← S_0
      4. G ← Rebuild(G)
2. Return <G, S>

BackwardEquivalenceSearch(G, S)

1. Repeat while $E_0 \neq \emptyset$
   1. $E_0 \leftarrow \emptyset$
   2. $H_0 \leftarrow \emptyset$
   3. $S_0 \leftarrow \emptyset$
   4. For each edge E' in G, connecting X and Y
      1. H-Neighbors ← nodes Z such that Z – Y and adj(Z, X, G)
      2. For each subset H of H-Neighbors
         1. G' ← a DAG in G
         2. S' ← score that would result from removing the edge from X to Y from G'
         3. If S'<S and S'<S_0 & ValidDelete(G, E', H)
            1. E_0 ← E′
            2. H_0 ← H
            3. S_0 ← S′
      3. If $E_0 \neq \emptyset$
         1. Remove E_0 from G
         2. For each H in H_0, if X – H in G, orient X – H as X→H.
         3. S ← S_0
         4. G ← Rebuild(G)
   5. Return G

ScoreEdgeAddition(X, Y, T, G)

1. NAYX ← Z such that Z–Y and adj(Z, X, G)
2. S2 ← (NAYX ∪ T) ∪ parents(Y)
3. S1 ← S2 ∪ {X}
4. Return score(Y, S1) - score(Y, S2)

ScoreEdgeDeletion(X, Y, H, G)

1. NAYX ← Z such that Z–Y and adj(Z, X, G)
2. S2 ← (NAYX \ H) ∪ parents(Y)
3. S1 ← S2 \ {X}
4. Return score(Y, S1) - score(Y, S2)

ValidInsert(G, E, T)

1. NAYX ← nodes Z such that Z – Y and Z is adjacent to X
2. If NAYX ∪ T is not a clique
   1. Return False
3. If some semidirected path from Y to X does not contain any node in NAYX ∪T
   1. Return False
4. Return True

ValidDelete(G, E, H)

1. NAYX ← nodes Z such that Z – Y and adj(Z, X)
2. If NAYX \ H is not a clique
   1. Return False
3. Else
   1. Return True

Rebuild(G)

1. For each edge X→Y in G
   1. If there does not exist a Z not equal to X such that Z→Y
      1. Orient X→Y as X – Y.
      2. Orient Z→Y as Z – Y.
2. Repeat until no more orientations can be made, for X, Y, Z, W in G
   1. If X→Y, Y – Z and ~adj(X, Z), orient Y–Z as Y→Z
   2. If X→Y, Y→Z, and X–Z, orient X–Z as X→Z.
   3. If X→Y, X→Z, Y→W, Z→W, and X–W, orient X–W as X→W.
3. Return G.

Generalization of the BIC score to the IMscore

Let D be a list of data sets <D_1,…,D_m>, each with sample size n, over a common set of variables V=<D_1,…,V_p>, and let G be a DAG model over V, indexed by j. Let k be the number of nodes plus the number of edges in G. Note that for the linear case,

$$\left(\frac{-m}{2c}\right)\text{imScore} = -(n/c)\sum_{i=1}^{m}\sum_{v=1}^{p} log\left(\sigma_{ip}\right) - (mk/2)log(n)$$

where σ_{ik} is the residual standard deviation of V_i regressed linearly onto the parents of V_i in G for D_i. The first term of this formula is the maximum likelihood of the product of m independent multivariate Normal distributions with modified residual variances $\sigma_{11}^{2/c}, \ldots, \sigma_{mp}^{2/c}$, respectively. Following Schwarz (1978), the function being maximized is:

$$S(Y,n,j) = log\int \alpha_j exp\left(\sum_{i=1}^{m}(Y_i^{\circ}\theta_i - b_i(\theta_i))n\right)d\mu_j(\theta_1)\ldots d\mu_j(\theta_m)$$

where for each i=1,…, m, $Y_i^{\circ}\theta$-$b_i(\theta)$, is distribution of , represented as an element of the exponential family, where α_j, Y, θ, and b are notated as in Schwarz. But since the sum of strictly convex functions over disjoint sets of parameters is strictly convex, by Schwarz's argument, as $n\to\infty$, this is equal to:

$$\begin{aligned} S(Y,n,j) &= log\int \alpha_j exp((A-\lambda|\ \vartheta-\vartheta_0|\ ))d\mu_j(\vartheta_1)\ldots d\mu_j(\theta_m) \\ &= nsup\left(\sum_{i=}^{m}(Y_i\theta_i - b_i(\theta_i))\right) - \left(\sum_{i=1}^{m} k/2\right)log\, n + R \\ &= -(n/c)\sum_{i=1}^{m}\sum_{v=1}^{p} log\left(\sigma_{ip}\right) - (mk/2)log(n) + R \end{aligned}$$

where R is a constant given fixed n, for A, λ, θ, θ_0 and μ_j as notated in Schwarz.

Cyclic graphs

The conditional independence relations of cyclic graphs, parameterized to represent linear systems with independent disturbances, are given by a purely graphical condition, the d-separation algorithm. Two variables, *X, Y* in a directed graph *G* are d-connected with respect to a set **Z** of other variables in the graph if and only if there is a series of pairwise adjacent (in *G*) variables, V_1,…V_n, with V_1=X and V_n=Y, such that for every variable V_i with directed *edges* $V_{i-1}\to V_i\leftarrow V_{i+1}$ there is a directed path from V_i terminating in a variable in **Z**.

The variables are d-separated with respect to a set **Z** if and only if there is no d-connection between them with respect to **Z**. The d-separation property also characterizes the conditional independence relations implied by DAGs.

Cyclic systems that are d-separation equivalent may not share the same adjacencies. For example, the following systems are d-separation equivalent:

$$X = aZ + bY + \varepsilon_x$$
$$Y = cW + dX + \varepsilon_y$$

and

$$X = aW + bY + \varepsilon_x$$
$$Y = cZ + dX + \varepsilon_y$$

Conditional independence relations or vanishing partial correlations cannot distinguish the directions of edges in cycles. That is, for example, if a graph contains the cycle $X\to Y\to Z\to X$, its d-separation class will contain a graph with the cycle $X\leftarrow Y\leftarrow Z\leftarrow W$. A computable, equivalence relation analogous to Markov equivalence for DAGs is available for linear, cyclic systems that imply the same conditional independence relations for all non-zero values of their linear coefficients (Richardson, 1996).

Appendix B: data acquisition and analysis

On each trial, the subject judged whether a pair of visually presented stimuli rhymed or not by pressing one of two response keys. In each 20-s block, 8 pairs of words were presented for 2.5 s each. Blocks were separated by 20-s periods of visual fixation. Four blocks of words were followed by four blocks of pseudowords. The data were acquired on a Siemens Allegra 3T MRI scanner at the UCLA Ahmanson-Lovelace Brain Mapping Center, using the following scanning parameters: TR = 2000 ms, TE = 30 ms, field of view = 200 mm, matrix size = 64 X 64, 33 slices, slice thickness = 4 mm (0 skip), 160 timepoints. For additional details on the subject population and data acquisition, see Xue and Poldrack (2007).

FMRI data processing was carried out using FEAT (FMRI Expert Analysis Tool) Version 5.98, part of FSL (FMRIB's Software Library, http://www.fmrib.ox.ac.uk/fsl). Following pre-processing steps were applied; motion correction; brain extraction using BET; spatial smoothing using a Gaussian kernel of FWHM 5 mm; grand-mean intensity normalization of the entire 4D dataset by a single multiplicative factor; highpass temporal filtering (Gaussian-weighted least-squares straight line fitting, with = 40.0 s). First level GLM analysis was carried out using FILM with local autocorrelation correction (Woolrich et al., 2001). Second level mixed effect GLM analysis was carried out to detect the activations per condition consistent across subjects. Z(Gaussianised T/F) statistic images were thresholded using clusters determined by Z>2.3 and a (corrected) cluster significance threshold of P = 0.05 (Worsley, 2003).

Thresholded 2nd level GLM statistics maps revealed consistent activations in the following areas: bilateral ACC, IFG, IPL, and OCC; and left MTG. Right MTG activation was detected across a majority of subjects in the 1st level analysis, but was not statistically significant in the 2nd level analysis, hence the RMTG region is only used in analyses that allow variable numbers of regions.

Regions of interest (ROIs) were defined as voxels that were significant in the first-level analysis which also fell within the anatomical regions of interest described above. The first level images were thresholded at t = 2.0 for the analysis presented in Fig. 2, and at z = 4.7 for the analyses presented in all other figures. The mean within these voxels was extracted and used for subsequent analyses. Since some subjects had no supra-thresholded voxels for a particular area, the number of non-degenerate time-series (hence regions) varied across the subjects.

Appendix C: details of simulation

The framework of the simulations creates a time series of ROI values which are sampled at a fixed interval, with the sample measures subject to independent, Gaussian error. Each ROI consists of a collection of 50 variables, which may be thought of as voxels or neurons. Each variable within a ROI influences its neighbors (variables at most two steps forward or backward in a pre-ordered list of variables for the ROI) non-linearly with a simulated 100 ms (1/20th of the sampling interval) time delay. The variables in one ROI may influence variables in another ROI in the same way, but with a longer time delay. For every influence between ROIs in one direction, there is an influence in the opposite direction. If variables in one ROI directly influence variables in another ROI, we represent that dependence by a directed edge in a graph. The "feed-forward" graph is composed of the directed edges on directed paths out of the Input variable. The ROI value at a time is the sum of the values of its constituent variables at that and recent past times, convolved with a hemodynamic response function of time.

We let $K_i(j)$ index values of the ith variable constituent of ROI K at time j. We let P_i range over parents of the ith variable within a ROI and we let $Q_i$range over parents of the ith variable located in ROIs other

than that of i. The values of $K_i(j)$ for each constituent variable i of ROI K, and time j are determined by the equation

$$K_i(j) = 30\tanh\left(\left(\left(\sum_k a_k P_k(j-1) + \sum_m a_m Q_m(j-4)\right) + e_1\right)/30\right) + e_2. \quad (1)$$

with $e_1 \sim N(0, \sigma)$ where for each variable K_i, σ is drawn from U(0,30), and $e_2 \sim N(0, 0.3)$. The different arguments for P and Q represent the different times required for signals to pass to adjacent variables within a ROI (P) as compared with between ROIs (Q). The "30" factors merely increase the scale so that the variances and error variances when measurement error is taken into account will approximate those of the Xue and Poldrack experiment.

The values of ROIs at a time are determined by the following equation:

$$\mathrm{ROI}_K(t_0) = \sum_{j=t_0-100}^{t_0} \sum_{i=1}^{50} K_i(j)\mathrm{HRF}(j, t_0) \quad (2)$$

where HRF is the canonical dual-gamma HRF as used in SPM. Finally, the measured variables are given by:

$$M_K(t) = \mathrm{ROI}_K(t) + e_3 \quad (3)$$

where $e_3 \sim N(0, \sigma)$ where σ is drawn from $U(0, 30)$. Measured values are obtained only every 20 simulated time steps after the first 600 simulated time steps (corresponding to a 60-s "burn in").

For varying delays in BOLD response within ROIs, Eq. (2) is modified for some ROIs to

$$\mathrm{ROI}_K(t_0) = \sum_{j=t_0-100}^{t_0} \sum_{i=1}^{50} K_i(j)\mathrm{HRF}(j, t_0 + \delta_i)$$

where δ_i is drawn from U(0,5).

For varying delays in BOLD response between ROIs, the values given by equation 2 for 3 (randomly) selected ROIs are shifted by 10, 20, 30 or 40 time steps into the future, corresponding to a simulation of 0.5, 1, 1.5 and 2.0 sampling intervals. The forward graphs are selected uniformly at random on the space of all possible 9 variable, 9 edge (respectively 11 variable, 11 edge) DAGs.

The simulated input was treated as a ROI associated with a set of constituent variables, like the ROIs of other variables. In 25-step intervals within 200-step blocks with 200-s rest periods between blocks, 25 values for the constituent variables of the input ROI were selected. Each selected constituent was given a value drawn from N (5.0, 0.5) for the duration of its presentation.

Appendix D. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.neuroimage.2009.08.065.

References

Breakspear, M., Terry, J., Friston, K.J., 2003. Modulation of excitatory synaptic coupling facilitates synchronization and complex dynamics in a biophysical model of neuronal dynamics. Netw. Comput. Neural Syst. 14, 703–732.

Bullmore, E., Horwitz, B., Honey, G., Brammer, M., Williams, S., Sharma, T., 2000. How good is good enough in path analysis of fMRI data. NeuroImage 11, 289–301.

Chen, R., Herskovitz, E., 2007. Graphical model based functional analysis of fMRI images. NeuroImage 35, 635–647.

Chickering, M., 2002. Optimal structure identification with greedy search. Mach. Learn. Res. 3, 507–554.

Demarco, G., Vrignaud, P., Destrieux, C., Demarco, D., Testelin, S., Devauchelle, B., Berquin, P., 2009. Principle of structural equation modeling for exploring functional interactivity within a putative network of interconnected brain areas. Magn. Reson. Imaging 27, 1–12.

Demiralp, S., Hoover, K., 2003. Search for the causal structure of a vector auto-regression. Oxf. Bull. Econ. Stat. 65, 745–767.

Eichler, M., 2005. A graphical approach for evaluating effective connectivity in neural systems. Philos. Trans. R. Soc., B 360, 953–967.

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

Glymour, C., 2003. The Mind's Arrows: Bayes Nets and Graphical Causal Models in Psychology. MIT Press, New York.

Glymour, C., Scheines, R., Spirtes, P., Kelly, K., 1987. Discovering Causal Structure. Academic Press, New York.

Granger, C.W.J., 1969. Investigating causal relation by econometric and cross-sectional method. Econometrica 37, 424–438.

Hanson, S., Bly, B., 2000. The distribution of BOLD susceptibility effects in the brain is non-Gaussian. NeuroReport 12, 1971–1977.

Hanson, S., Hanson, C., Halchenko, Y., Matsuka, T., Zaimi, A., 2007. Bottom-up and top-down brain functional connectivity underlying comprehension of every day visual action. Brain Struct. Funct. 212, 231–244.

Hoyer, P.O., Shimizu, S., Kerminen, A., 2006. Estimation of linear, non-gaussian causal models in the presence of confounding latent variables. Proc.Third European Workshop on Probabilistic Graphical Models (PGM'06). Prague, Czech Republic, pp. 155–162.

Kaas, J., 2004. Topographic maps in the brain. International Encyclopedia of Social and Behavioral Sciences. Elsevier, pp. 15771–15775.

Lacerda, G., Spirtes, P., Ramsey, J., Hoyer, P.O., 2008. Discovering cyclic causal models by independent components analysis. Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence. AUAI Press, Oregon.

Lazar, N., Luna, B., Sweeney, J., Eddy, W., 2002. Combining brains: a survey of methods for statistical pooling of information. NeuroImage 16, 538–550.

Mcintosh, A., Gonzalez-Lima, F., 1994. Structural equation modeling and its application to network analysis in functional brain imaging. [My Copy] Hum. Brain Mapp. 2, 2–22.

Meek, C. (1997). Graphical Models: Selecting causal and statistical models. PhD thesis, Carnegie Mellon University.

Miezin, F., Maccotta, L., Ollinger, J., Petersen, S., Buckner, R., 2000. Characterizing the hemodynamic response: effects of presentation rate, sampling procedure and the possibility of ordering brain activity based on relative timing. NeuroImage 11, 735–759.

Mumford, J., Poldrack, R., 2007. Modeling group fMRI data. Soc. Cogn. Affect. Neurosci. 2, 251–257.

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

Poline, J., 2003. Contrasts and classical inference, In: Frackowiak, R., Friston, K.J., Frith, C., Dolan, R., Price, C., Zeki, S., Ashburner, J., Penny, W. (Eds.), Human Brain Function, 2nd edition. Academic Press.

Ramsey, J., Spirtes, P., Zhang, J., 2006. Adjacency—faithfulness and conservative causal inference. Proceedings of the 22nd Conference on Uncertainty in Artificial Intelligence. AUAI Press, Oregon, pp. 401–408.

Richardson, T. (1996) A discovery algorithm for directed cyclic graphs. Proceedings of the 1996 Conference on Uncertainty in Artificial Intelligence.

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

Rubin, D.B., 1987. Multiple Imputation for Nonresponse in Surveys. John Wiley & Sons, New York.

Schwarz, G., 1978. Estimating the dimension of a model. Ann. Stat. 6, 451–464.

Silva, R., Glymour, C., Scheines, R., Spirtes, P., 2006. Learning the structure of latent linear structure models. J. Mach. Learn. Res. 7, 191–246.

Spirtes, P., 1995. Directed cyclic graphical representation of feedback models. In: Besnard, Philippe, Hanks, Steve (Eds.), Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence. Morgan Kaufmann Publishers, Inc., San Mateo, p. 1995.

Spirtes, P., Glymour, C., 1991. An algorithm for fast recovery of sparse causal graphs. Soc. Sci. Comput. Rev. 9, 62–72.

Spirtes, P., Glymour, C., Scheines, R., 1993. Causation, prediction and search. Springer Lecture Notes in Statistics. .

Spirtes, P., Glymour, C., Scheines, R., 2000. Causation, Prediction and Search, 2nd Edition. MIT Press, Cambridge, MA.

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

Swanson, N., Granger, C., 1996. Impulse response functions based on a causal approach to residual orthogonalization in vector autoregressions. J. Am. Stat. Assoc. 92, 357–367.

Woolrich, M.W., Ripley, B.D., Brady, J.M., Smith, S.M., 2001. Temporal autocorrelation in univariate linear modeling of FMRI data. NeuroImage 14 (6), 1370–1386.

Worsley, K.J., 2003. Statistical analysis of activation images. In: Jezzard, P., Matthews, P.M., Smith, S.M. (Eds.), Functional MRI: An Introduction to Methods. Oxford University Press, New York, NY.

Xue, G., Poldrack, R., 2007. The neural substrates of visual perceptual learning of words: implications for the visual word form area hypothesis. J. Cogn. Neurosci. 19, 1643–1655.

Yule, G.U., 1919. An Introduction to the Theory of Statistics. C. Griffin and Co, London.

Zhang, J., Spirtes, P., 2008. Detection of unfaithfulness and robust causal inference. Minds Mach 7, 239–271.

Zheng, X, Rajapake, J., 2006. Learning functional structure from fMRI images. NeuroImage 31, 1601–1613.