NeuroImage xxx (2011) xxx–xxx

Contents lists available at ScienceDirect

NeuroImage

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

Multi-subject search correctly identifies causal connections and most causal directions in the DCM models of the Smith et al. simulation study ☆

Joseph D. Ramsey [a,*], Stephen José Hanson [b], Clark Glymour [a,c]

[a] Department of Philosophy, Carnegie Mellon University, USA
[b] Department of Psychology, Rutgers University, USA
[c] Florida Institute for Human and Machine Cognition, USA

ARTICLE INFO

Article history:
Received 27 April 2011
Revised 20 June 2011
Accepted 23 June 2011
Available online xxxx

Keywords:
FMRI
IMaGES
Group analysis
Effective connectivity
Causal modeling
Directionality

ABSTRACT

Smith et al. report a large study of the accuracy of 38 search procedures for recovering effective connections in simulations of DCM models under 28 different conditions. Their results are disappointing: no method reliably finds and directs connections without large false negatives, large false positives, or both. Using multiple subject inputs, we apply a previously published search algorithm, IMaGES, and novel orientation algorithms, LOFS, in tandem to all of the simulations of DCM models described by Smith et al. (2011). We find that the procedures accurately identify effective connections in almost all of the conditions that Smith et al. simulated and, in most conditions, direct causal connections with precision greater than 90% and recall greater than 80%.

Introduction

Estimates of effective connections of brain areas using imaging time series promise a more nuanced understanding of neural processes than do measures of localized activity alone. Fulfilling that promise requires that, when applied to fMRI time series, the procedures for specifying causal structure be reasonably accurate. Smith et al. (2011) have investigated the accuracy of a large number of search procedures on more than 21,000 simulated fMRI signals generated from DCM models (Friston et al., 2003; Buxton et al., 1998), for each of 50 simulated subjects. Smith et al. find that with 50 simulated regions of interest represented by a directed acyclic graph with 50 vertices and 61 directed edges, no method tested is very good at identifying the graphical adjacencies, and no method tested is much better than chance at finding the direction of causal influence for the edges. Better results are obtained with 5 or 10 regions of interest, but no procedure Smith et al. tested is much better than chance at identifying *both* the true adjacencies *and* the true directions of influence in the generating models. These results could reasonably lead fMRI researchers to the conclusion that available search algorithms for causal structure are of little or no aid in finding directions of causal connections, and with large variable sets all available methods are likely to be misleading even for identifying causal connections without regard to direction of influence. Our results with the Smith et al. data give strong evidence that such conclusions would be premature and that the situation is more complex and more promising.

We consider a combination of methods, one of which, IMaGES (Ramsey et al., 2010), is a multi-subject method Smith et al. noted but explicitly did not test. The other, LOFS, is an adaptation of an idea in the LiNGAM algorithm Smith et al. did test. Like IMaGES, LOFS is used with multiple subjects.[1] We show that on the 50 variable data Smith et al. use, the adjacencies in the generating graph are discovered *almost perfectly* by IMaGES—100% precision and 98% recall—used with exactly the parameters used in Ramsey et al.(2010). Of the edges directed by IMaGES, 87% are correct. IMaGES does not orient more than 20% of the edges. When IMaGES is supplemented by a LOFS postprocessor, the precision of orientations is greater than 90% and the precision of recall greater than 80%—i.e., more edges are directed than with IMaGES alone, and with no loss of accuracy. IMaGES and LOFS are superior as well on the Smith et al. data from smaller graphs.

One signal difference between IMaGES and the methods Smith et al. tested is that IMaGES and LOFS estimate causal relations from multiple time series, extracting full information from the multiple

☆ We thank the James S. McDonnell Foundation for support of this research and Steve Smith for helpful comments.
* Corresponding author at: Department of Philosophy, Carnegie Mellon University, Pittsburgh, PA 15213, USA. Fax: +1 412 268 1440.
E-mail address: jdramsey@andrew.cmu.edu (J.D. Ramsey).

[1] "LOFS" abbreviates "LiNG Orientation, Fixed Structure." Linear, nongaussian ("LiNG") models are assumed, though acyclicity is not.

1053-8119/$ – see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.neuroimage.2011.06.068

data sets simultaneously without producing artifacts that might result if the time series were pooled or appended to one another (Ramsey et al., 2011). Smith et al. suggest applying LiNGAM to combined multiple data sets. We carry out that study with the same data samples used for IMaGES/LOFS. We find the LiNGAM accuracies, while better than any of those in the methods tested by Smith et al., are markedly inferior to the IMaGES/LOFS accuracies. The comparison of simulations in the Smith et al. study gives further evidence that the IMaGES advantage is due in part to how the samples are analyzed, and not merely to the increased sample size afforded by using multiple subjects.

We suggest that these results, and others we will report here, have implications for the design of fMRI studies, for the methodology that should be used in making causal inferences from them, and for research directions on machine learning for fMRI.

The Smith et al. simulations

Smith et al., provide 28 simulations (Table 1) based on DCM models using effective connectivity graphs, or variations of them, as shown in Fig. 1.

Understanding why all methods fail on some simulations requires attention to details of the simulations, and we cannot do better than quoting Smith et al. at length:

> "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." (p. 2)

Table 1
Summary of the 28 simulation specifications (Smith et al., 2011).

Sim	# nodes	Session length	TR (s)	Noise %	HRF sigma	Other factors
1	5	10	3	1	0.5	
2	10	10	3	1	0.5	
3	15	10	3	1	0.5	
4	50	10	3	1	0.5	
5	5	60	3	1	0.5	
6	10	60	3	1	0.5	
7	5	250	3	1	0.5	
8	5	10	3	1	0.5	Shared inputs
9	5	250	3	1	0.5	Shared inputs
10	5	10	3	1	0.5	Global mean confound
11	10	10	3	1	0.5	ROI time series intermixed
12	10	10	3	1	0.5	Random time series mixed in
13	5	10	3	1	0.5	2 cycles added
14	5	10	3	1	0.5	5 cycles: 1→5 reversed
15	5	10	3	0.1	0.5	Stronger effective connections
16	5	10	3	1	0.5	Triangulated connections
17	10	10	3	0.1	0.5	
18	5	10	3	1	0	
19	5	10	0.25	0.1	0.5	Neural lag = 100 ms
20	5	10	0.25	0.1	0	Neural lag = 100 ms
21	5	10	3	1	0.5	Two coefficient groups
22	5	10	3	0.1	0.5	Non-stationary connection strengths
23	5	10	3	0.1	0.5	
24	5	10	3	0.1	0.5	One strong noise input
25	5	5	3	1	0.5	
26	5	2.5	3	1	0.5	
27	5	2.5	3	0.1	0.5	
28	5	5	3	0.1	0.5	

> "… [W]e changed [the DCM neural lag] 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, 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. The main network topologies are shown in [Fig. 1].[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. 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." (p. 3)

Smith et al. tabulate a summary of the simulations as shown in Table 2 (with some modification of the descriptions).

A "subject" is an fMRI time series for the relevant variables in a simulation. Thus each simulation consists of 50 "subjects." Simulation 13 introduces randomly selected 2-cycles into the S5 graph of each subject. Thus in this simulation while the 1→5 edge is never reversed, otherwise the only invariant structure across "subjects" is the adjacency structure among the five variables. Simulation 14 reverses the 1→5 edge in S5 to form a cyclic graph. Simulation 16 introduces extra edges into the S5 graph, forming triangles. In simulation 22 connection strengths vary dramatically with time within a subject. In

[2] This is Smith et al.'s original Fig. 2.

Please cite this article as: Ramsey, J.D., et al., Multi-subject search correctly identifies causal connections and most causal directions in the DCM models of the Smith et al. simu..., NeuroImage (2011), doi:10.1016/j.neuroimage.2011.06.068

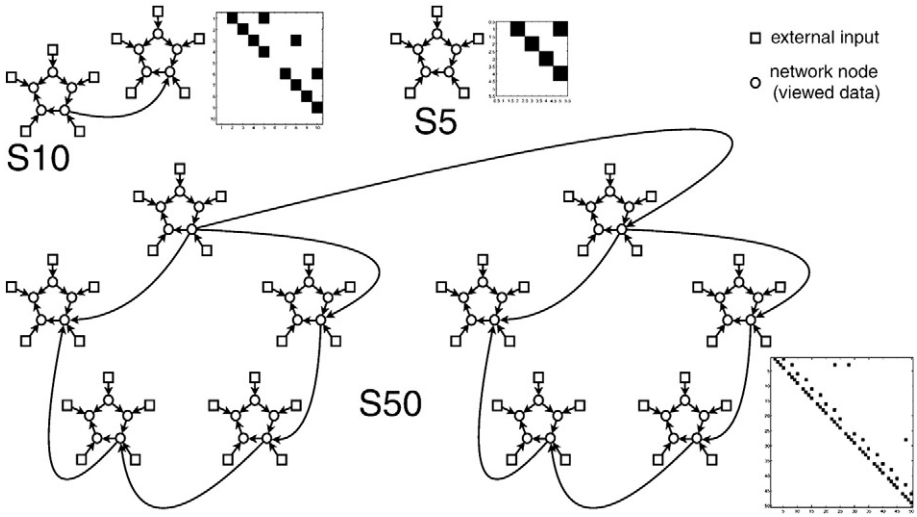

Fig. 1. From Smith et al. (2011), Fig. 2. Quote: "The main network topologies fed into the FMRI data simulations. For each network graph the corresponding connection matrix is shown, where an element in the upper diagonal of the matrix implies a directed connections from a lower-numbered node to a higher-numbered one."

simulation 21 subjects are divided into 2 groups with differing strengths of effective connections. In addition, various simulations test different TRs, time series lengths, and noise values.

Smith et al. test 38 search methods, including five "Bayes net" methods, on simulations with 5, 10 or 15 vertex graphs, and they test 33 methods on the 50-vertex graph. They evaluate edge direction accuracy by the percentage of correct orientations among the correctly estimated adjacencies, choosing the most frequent orientation when conflicting orientations are found by a search method for different subjects in the same simulation. Adjacency correctness is evaluated, for each subject, by computing a Z score for each edge and computing the fraction of true positive adjacencies whose Z scores are greater than the 95 percentile of the false positive adjacencies. These ratios are then averaged over the 50 subjects in a simulation. Their findings can be briefly summarized.

For simulation 1, with 10 minute sessions, 3 second TRs, the best identification of directions is a version of Patel's τ with less than 70% accuracy. The average Z score ratio for separable true positives for this method is 0.2. In other words, the method produces a lot of false positives, but between 65 and 70% of its true positives are correctly oriented. Five "Bayes net" methods have average separable true positive Z score ratios >0.9 but are at chance in directing edges. Other simulations with S5 give similar results. The results are essentially the same for simulation 2, with 10 nodes and simulation 3 with 15 nodes. Accuracy is worse for adjacencies of Bayes net methods with variable effective connection coefficients, slightly worse for all methods with latent confounding, worse for all methods (except LiNGAM orientations) with 2 cycles or 5 cycles. Triangulation makes little difference. Noise values make little difference except that accuracy falls when a single neural variable is given a very large exogenous input and others are minimized. For simulation 4, which is S50 with 50 nodes, the ratio of separable true positives for Patel's τ falls to 0.1 and the orientation accuracy of those true adjacencies the method finds is just above 60%. For reasons we do not understand, but which Smith et al. say are computational,[3] none of the Bayes net methods were tested on simulation 4, the only 50 node simulation.

The Smith et al. report is an impressive and valuable effort. There are nonetheless some important limitations. The actual neural structures supporting cognition are apt to be more complex than the small-world neural network geometry that Smith et al. used. The causal structure is not varied much in the simulations and there is not much feedback. Real experiments usually have an input variable time series that can be convolved with an HRF and meshed by an appropriate lag with the neural time series. Since the direction of any causal connection of the input variable with neural variables is known, an input variable provides valuable information that can be used by some methods to infer causal directions of effective connections. The methods that can effectively use prior information are thus handicapped. Poisson noise sources, whether to neural inputs or to neural outputs may be larger and white noise proportionately smaller that in the simulations. The fMRI variables in the simulation, some of which have a small right skew, may be too Gaussian to represent real cases. The nearly Gaussian variables favor correlation methods and disadvantage methods that exploit higher moments. Further research in this direction might consider more complex and more realistic causal connections and, and further characterize the sensitivity of the search methods to variations in BOLD noise distributions.

Most important for our discussion, the Smith study omits the IMaGES algorithm. They write:

> "… We concentrate here on evaluating network modelling methods for single subject (single session) data sets, and only utilise multiple subjects' data sets in order to characterise variability of results across multiple random instantiations of the same underlying network simulation."

> "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 data sets; we felt that this is a primary question of interest, needing some clear answers before considering possible multi-subject modeling 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."

[3] We conjecture that in order to accommodate the Bayes net algorithms, Smith et al. were working across platforms, which could potentially slow down processing. The GES algorithm, as implemented in Tetrad IV, was very slow for the 50-variable case. This has since been reimplemented. We have no difficulty running IMaGES on up to 500 variables, and we are currently using PC and other "Bayes net" methods to analyze data from 5000+ voxels.

Please cite this article as: Ramsey, J.D., et al., Multi-subject search correctly identifies causal connections and most causal directions in the DCM models of the Smith et al. simu..., NeuroImage (2011), doi:10.1016/j.neuroimage.2011.06.068

Subject to the limitations just noted, we think the Smith study gives a fairly decisive answer to the question they pose about the accuracy of the 38 methods with a single subject and 10 minute time series under the conditions of their simulations: *none of the methods provide correct information about both adjacencies and directions*. Several methods provide correct information about adjacencies. Further, their simulation 7 with 250 minute series gives evidence that neither LiNGAM nor any of the other methods they test work well even with an unrealistically long time series. Their results naturally invite the question of how various methods would compare when run with multiple subjects on their simulated data; the results of our experiments provide an answer to that question.

IMaGES

The IMaGES algorithm is described in detail in Ramsey et al. (2010) and we will only sketch its strategy here. The input to the algorithm is any number of multivariate time series of approximately equal length. The procedure starts with an empty graph. Each model with one possible directed edge is then evaluated by computing the residuals from the model in each time series and, with them, assigning a BIC score to the model for that data set. The BIC scores for the model from the several data sets are then averaged, and the edge with the highest average score is selected. Next, models with two directed edges are considered, etc. At each stage, the search is not over particular edge additions to the model of "Markov equivalence class" resulting from an addition—see Ramsey, et al. (2010) for details. This allows the procedure to reverse directions of some previously posited edges, but does not eliminate adjacencies posited in previous steps. When further additions do not improve the score, the forward procedure is stopped, and a backward procedure begins, removing edges by an analogous method. The entire search stops when the backward procedure cannot improve the average BIC score. The output is often not a directed graph, but a "pattern" with some undirected edges. Patterns represent Markov equivalence classes of graphs (Pearl, 2000). The GES algorithm which IMaGES modifies provably converges to the true Markov equivalence class (essentially under the assumption that positive and negative effects do not perfectly cancel one another) when there are no latent common causes and no cycles; Ramsey et al. (2010) extend this result to IMaGES.

The BIC score contains a penalty term for the number of edges. Ramsey et al. (2010) repeat the search with an increasing penalty until a "non-triangular" model is produced. A model is "non-triangular" if there it contains no 3 clique. Their justification for this constraint is that the very fact that BOLD signals are indirect measures of causally related variables can easily produce false 3 cliques. Using an input variable, they show that in an extensive simulation with nonlinear ROI connections, empirically modeled noise, thorough feedback, and 2 second simulated TRs, the procedure recovers the nontriangular feedforward (from the input) graphical structure with considerable accuracy. They also show that the structure is recovered if a fraction of the ROIs is missing at random in the time series. They apply the procedure to empirical data sets with missing variables in some of the time series and recover a plausible structure.

The advantage of the IMaGES procedure is that it makes full use of the information in multiple data sets without producing artifacts that might result if the time series were appended to one another, and without requiring ad hoc statistical methods for missing values when some ROIs are not recorded for some subjects. The disadvantage is that it should work best when subjects have the same effective connectivity structure, although not necessarily the same parameter values—effective connective parameters and variances can vary across subjects. In principle, the method should work even when subjects differ in their effective connectivity structure, so long as connections are not in contrary directions, though this has not been tested. Detailed conditions are in Ramsey et al.(2011).

The basic LiNGAM procedure (Shimizu et al., 2005) assumes a linear model with independent, non-Gaussian disturbances. The substantive variables are resolved into independent components. Dependencies between the latent independent components and measured variables are then pruned by a threshold. Matrix manipulation then results in a partial order on the measured variables determining a directed acyclic graph and an estimate of the linear coefficients. PC-LiNGAM (Hoyer et al., 2008), which is the inspiration for our LOFS procedures described below, dispenses with the independent components analysis. It takes as input a data set and a pattern, enumerates each DAG G in the equivalence class, and calculates a nonlinear score from the residuals of each variable in the model G. The DAG with the highest score is chosen, where edges connecting variables with Gaussian residuals are not additionally oriented by the procedure. The search over all DAGs in an input equivalence class can make the procedure quite slow, and the measure of non-Gaussianity used is not particularly sensitive to variations from a Normal distribution. Both limitations are addressed in the LOFS procedures.

LOFS

The LiNGAM family of algorithms exploits the fact that the residuals of the correct linear model with independent non-Gaussian errors will be less Gaussian that the residuals of any incorrect model. That can be seen from two facts: (1) a sum of i.i.d. non-Gaussian variables is (usually) closer to Normal than any of the terms in the sum; and (2) the regression residual of a variable X on a false orientation of its adjacent variables is a weighted sum of the error term for X and the error terms for the variables of mis-oriented edges—whereas on the correct orientation the residual for X is just the error term for X.

The first point can be seen for standardized cumulants such as skew (the third standardized cumulant) and excess kurtosis (the fourth standardized cumulant) of weighted sums of residuals. Let $\kappa_n(e)$ be the nth cumulant and let $\kappa_n(e)/(\kappa_2(e))^{n/2}$ be the nth standardized cumulant. Let $r = \Sigma_i a_i e_i$, where e_i, e_j are i.i.d., with $a_i \neq 0$, $i = 1, \ldots, m$, $m \geq 2$. Then $\kappa_n(r)/(\kappa_2(r))^{n/2} = (\Sigma_i a_i^n/(\Sigma_i a_i^2)^{n/2})(\kappa_n(e_1)/(\kappa_2(e_1))^{n/2})$, from which it follows that the nth standardized cumulant of r is closer than the nth standardized cumulant of e_1 to the nth standardized cumulant of a Normal distribution (i.e., zero) just in case $|\Sigma_i a_i^n/(\Sigma_i a_i^2)^{n/2}| < 1$, for $n \geq 3$. This condition is always true for skew and excess kurtosis.

The second point can be seen with a simple example. Suppose in truth that $X = aY + e_X$, with e_X independent of $Y = e_Y$. Regressing X on Y leaves the residual $R_{X|Y} = e_X$. Regressing Y on X leaves the residual R_X for X as X itself, that is $R_X = aY + e_X = ae_Y + e_X$. More generally, assume the true system is linear with i.i.d. non-Gaussian errors. That is,

$$X = BX + e$$

where $X = <x_1, \ldots, x_m>$ are the variables of the linear system, $e = <e_1 \ldots, e_m>$ the errors, and B the coefficient matrix. For any incorrect model of the variables that directly influence or are directly influenced by X, we have a system of the form

$$X = CX + r$$

where C is the alternative model, $r = <r_1, \ldots, r_m>$ the residuals of that model. Assuming the relevant matrices are invertible, it follows that

$$r = (I–C)(I–B)^{-1}e$$

from which we conclude that the residuals of the incorrect model are linear combinations of the residuals (errors) of the correct model. (It is not necessary for elements of r to be independent.) In addition, $B = C$ just in case $r = e$, as expected.

These properties of linear models imply two rules, either of which can be combined with a measure, "NG(X)", of the "non-Gaussianity"

Please cite this article as: Ramsey, J.D., et al., Multi-subject search correctly identifies causal connections and most causal directions in the DCM models of the Smith et al. simu..., NeuroImage (2011), doi:10.1016/j.neuroimage.2011.06.068

of the distribution of X in order to orient undirected edges. The NG measure we use is the A^{2*} statistic of the Anderson–Darling test of the hypothesis that a distribution is Gaussian (Anderson and Darling, 1952; D'Agostino, 1986). In the case where variance and mean are unknown, $A^{2*} = A^2(1 + 4/n - 25/n^2)$ where

$$A^2 = -n\text{–}1 / n \; \Sigma_i(2i\text{–}1)(ln\mathrm{F}(\mathrm{Y}i) + ln(1\text{–}\mathrm{F}(\mathrm{Y}n + 1 - i)))$$

where $F(Y_i)$ is the Normal(0, 1) CDF evaluated at Y_i and Y_i is the ith standardized data point in ascending order, with n the sample size.

We denote the evaluation of NG for the conditional distribution of X on a set S of variables by NG(X|S), and when the evaluations are for a list of data sets $D = <D_1, \ldots, D_k>$, we write NG(X|S; D). The rules are as follows:

Rule R1 *Given a set P of variables that are connected to a variable X by undirected edges, choose the orientations of those edges that maximize NG(X|P).*

Rule R2 *Let X and Y be adjacent, and let the candidate parents of X other than Y be O_X, and the candidate parents for Y other than X be O_Y. Orient as $X \leftarrow Y$ if $NG(X|Y,O_X) > NG(Y|X,O_Y)$, $NG(Y|O_Y) > NG(X|O_X)$, $NG(X|Y,O_X) > NG(X|O_Y)$ and $NG(Y|O_Y) > NG(Y|X,O_Y)$, and vice-versa for $X \rightarrow Y$.*[4]

Rule 1 is a straightforward implication of the facts about sums of residuals just reviewed. The rule allows that an edge X − Y may be directed into X and into Y—which can be interpreted either as a 2 cycle between X and Y or as a latent common cause of X and Y. The algorithm cannot distinguish these cases.

For each node X in an undirected graph, for each set S of nodes that are adjacent to X, and each list of data sets $<D_1, \ldots, D_k>$, let NG(X|S; $<D_1, \ldots, D_k>$) be calculated as follows. Form a set of regression residuals R as follows. For each i = 1, …, k, if X and all nodes in S are measured in D_i, calculate the multiple regression residual r_i of X conditional on S in D_i and add r_i—mean(r_i) to R. Concatenate the vectors in R, producing a vector r. NG(X|S; $<D_1, \ldots, D_k>$) equals the A^{2*} statistic of r.

Procedure LOFS-R1(G, D1 … Dk)

1. Let G′ be empty.
2. For each node X in G,
   2.1. Find the subset S of the nodes adjacent to X in G that maximizes NG(X|S; <D1, …, Dk>)
   2.2. For each Y in S, add Y → X to G′
3. For each pair of nodes Z and W adjacent in G but not in G′
   3.1. Add Z—W to G′
4. Return G′

Step 3 is required because step 2 may not generate some of the adjacencies that are in graph G. The algorithm is purely local and very fast.

LOFS-R2 takes the undirected graph from IMaGES, or the mixed graph from R1, and applies Rule 2 to each edge in the graph, evaluating NG on multiple data sets as in the LOFS-R1 algorithm.

The result is four possible procedures:

1. IMaGES alone
2. IMaGES + LOFS-R1
3. IMaGES + LOFS-R1 + LOFS-R2 applied to the unoriented edges of IMaGES + LOFS-R1
4. IMaGES + LOFS-R1 + LOFS-R2 applied to the unoriented edges and the 2 cycles of IMaGES + LOFS-R1.

Procedure 2 may leave some edges unoriented and may produce 2-cycles which may be due to actual feedback between the two ROIs or to unrecorded common causes of the variables in the 2-cycles, or both. Procedure 4 will produce the fewest undirected edges, but it assumes that there are no unobserved common causes of ROIs and no 2-cycles or latent common causes influencing two ROIs.

Experimental procedure

All four of the procedures described in the previous section were applied to the Smith, et al., simulations. In addition, for comparison purposes the Greedy Equivalence Search (GES) was run both as Smith et al. ran it and also with BIC penalty adapted to prevent 3-cycles—which is equivalent to IMaGES as Ramsey et al.(2010) ran the program but with a single subject. Finally, LiNGAM was run on data from 10 subjects at a time appended as Smith et al. suggest.

IMaGES was applied to the data from the Smith et al., simulations in the following way. The program was given random samples respectively of 1 or of 10 distinct subjects drawn with replacement from the 50 subjects in each of the Smith simulations. IMaGES was run with the 0-lag, nontriangular search exactly as in Ramsey et al. (2010). The precision and recall of adjacencies and directed edges, and the number of edges left undirected, were counted. For each simulation, the procedure was repeated fifty times, sampling with replacement, and the averages of the various accuracies and numbers of undirected edges were calculated for each simulation.

In addition, in each simulation, for each sample of subjects given to IMaGES, the LOFS-R1 and two LOFS-R1 + R2 procedures were applied to the undirected graphical skeleton of the IMaGES output. For example, if 10 subjects were selected as input to IMaGES, LOFS-R1 was applied to the undirected edges in the IMaGES output from all 10 subjects.

The IMaGES and LOFS algorithms were run on all of the Smith, et al. data with the exception that for experiment 13, only the 10 subjects sharing the same graphical structure were used.

Results

The results of our experiments are shown in five graphs (Figs. 2 through 6). Tabulated results are given in the appendix.

Discussion

Although not shown, the error rates decrease monotonically with the number of data sets (subjects) used as input to IMaGES and to LOFS. Even five subjects dramatically decrease the error rates, and with ten subjects most of the adjacency error rates are essentially zero. In most simulations with 10 subjects the correct directions with LOFS procedures are 80% or above, and few of the edges remain undirected after applying the LOFS procedures.

We note especially simulation 4, with 50 variables, for which all of the search methods Smith et al. tested performed very badly. With 10 subject inputs to IMaGES the adjacency results are almost perfect, and with the LOFS post processors on average about 1 edge is left unoriented, and 86% of the oriented edges are correctly directed.

The IMaGES/LOFS procedures remain highly accurate when the time series have a global mean confound, when there is an unrecorded common cause or causes, when the time series are mixed with a random time series, when there is a 5 cycle in the generating structure, and when subjects in a simulation consist of two groups with different strengths of effective connections.

Results that are in one or another respect problematic are obtained for simulations 11, 13, 15, 16, 22, 23 and 24.

Simulation 11 produces poor results on every evaluation dimension, by every method. The reason is that the simulated ROI activities have been combined by Smith et al. to produce simulated measured

[4] Optionally, one may exclude the conditions that $NG(X|Y,O_X) > NG(X|O_Y)$ and $NG(Y|O_Y) > NG(Y|X,O_Y)$; the rule with these conditions is a stronger test, the rule without a weaker test of directionality.

Please cite this article as: Ramsey, J.D., et al., Multi-subject search correctly identifies causal connections and most causal directions in the DCM models of the Smith et al. simu..., NeuroImage (2011), doi:10.1016/j.neuroimage.2011.06.068

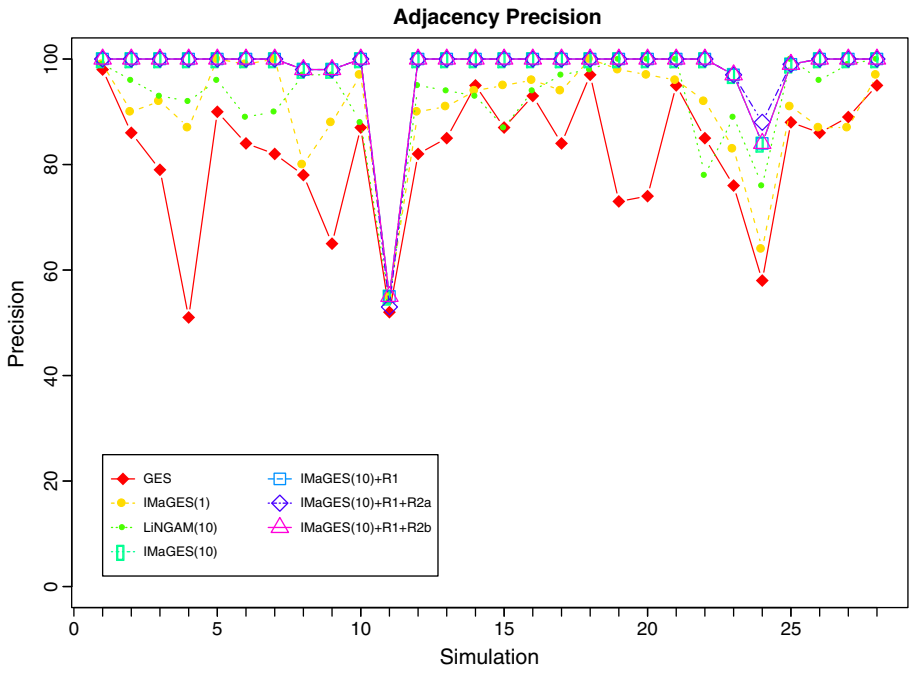

Fig. 2. Adjacency precision = # true adjacencies in output / # adjacencies in output, for each of simulations 1 through 28, for each algorithm being compared. See text for details. The algorithms are: GES, I(1) = IMaGES, first nontriangular, with one subject's data as input, L(10) = LiNGAM using concatenated data from 10 subjects, I(10) = IMaGES, first nontriangular, with 10 subjects' data as input, I(10) + R1 = I(10) followed by LOFS rule R1, I(10) + R1 + R2a = I(10) + R1 followed by LOFS rule R2 applied to undirected edges only, I(10) + R1 + R2b = I(10) + R1 followed by LOFS rule R2 applied to undirected edges and 2-cycles.

outputs, presenting a worst case of ROI selection and providing another example of the old theorem: garbage in, garbage out.

Orientation results for simulation 13 are not tabulated because they were obtained in an unusual way. In this simulation, 2 cycles were introduced at random for subjects. Thus most of the 50 subjects have different graphical structures and there would be no common orientations shared by randomly selected groups of 10 subjects. One structure was shared by 10 subjects, and IMaGES + LOFS-R1 was run on that single set of 10 subjects. All orientations were correct and the 2-cycle was identified.

Simulation 15 produces poor orientation results by all methods because the suppressed exogenous inputs and increased effective connections make the system nearly deterministic. Simulation 16 has extra edges added to S5 forming triangles, one edge of each of which IMaGES is forced to exclude, resulting in high adjacency precision but poor adjacency recall—i.e., a high proportion of false negative

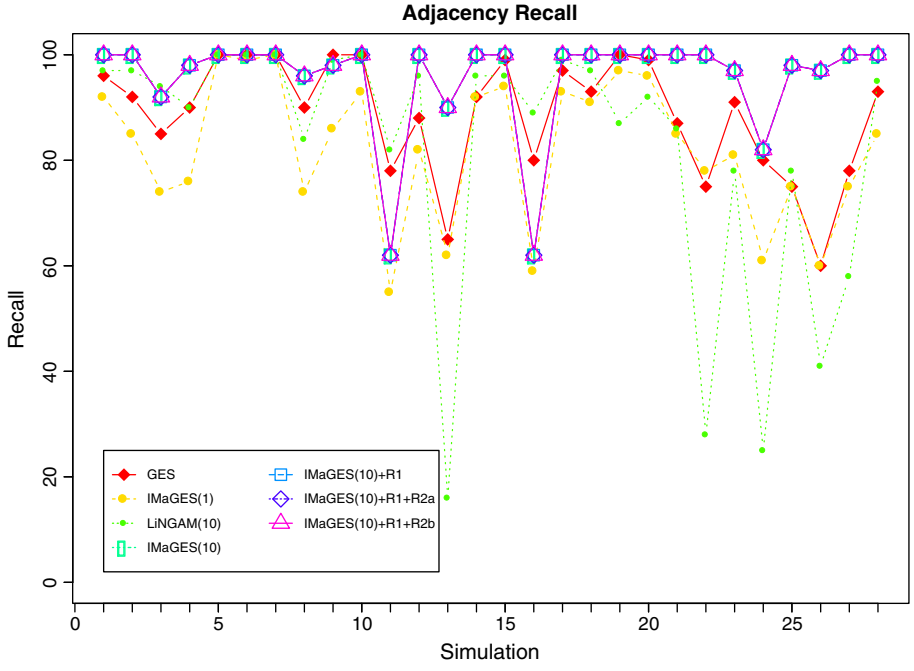

Fig. 3. Adjacency recall = # true adjacencies in output / # adjacencies in the true model, for each of simulations 1 through 28, for each algorithm being compared. See text for details. The algorithms are: GES, I(1) = IMaGES, first nontriangular, with one subject's data as input, L(10) = LiNGAM using concatenated data from 10 subjects, I(10) = IMaGES, first nontriangular, with 10 subjects' data as input, I(10) + R1 = I(10) followed by LOFS rule R1, I(10) + R1 + R2a = I(10) + R1 followed by LOFS rule R2 applied to undirected edges only, I(10) + R1 + R2b = I(10) + R1 followed by LOFS rule R2 applied to undirected edges and 2-cycles.

Please cite this article as: Ramsey, J.D., et al., Multi-subject search correctly identifies causal connections and most causal directions in the DCM models of the Smith et al. simu..., NeuroImage (2011), doi:10.1016/j.neuroimage.2011.06.068

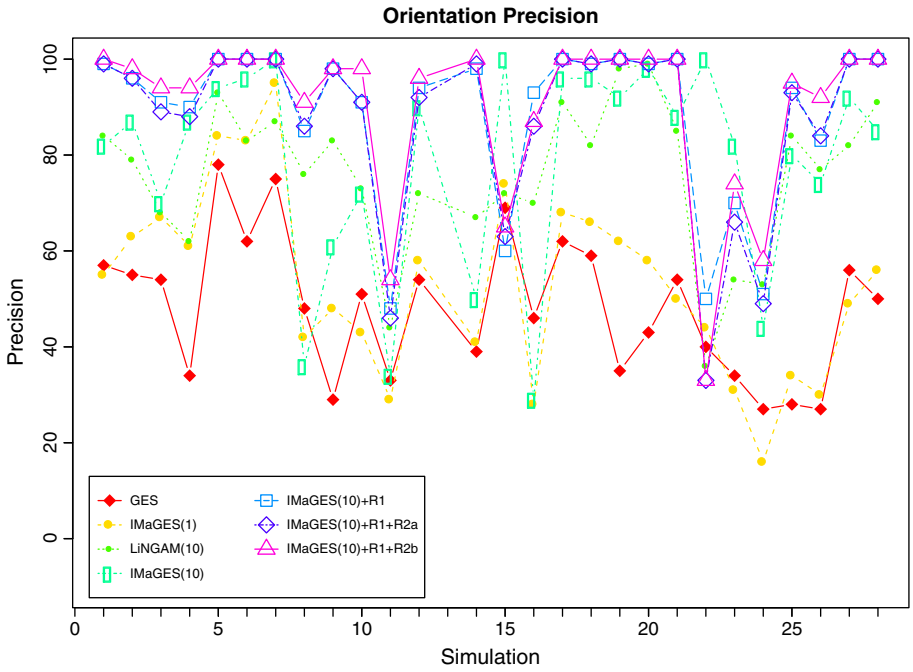

Fig. 4. Orientation precision = # true orientations in output / # orientations in output, for each of simulations 1 through 28, for each algorithm being compared. See text for details. The algorithms are: GES, I(1) = IMaGES, first nontriangular, with one subject's data as input, L(10) = LiNGAM using concatenated data from 10 subjects, I(10) = IMaGES, first nontriangular, with 10 subjects' data as input, I(10) + R1 = I(10) followed by LOFS rule R1, I(10) + R1 + R2a = I(10) + R1 followed by LOFS rule R2 applied to undirected edges only, I(10) + R1 + R2b = I(10) + R1 followed by LOFS rule R2 applied to undirected edges and 2-cycles.

adjacencies. IMaGES + LOFS-R1 gives orientation precision greater than 90%, but the orientation recall is poor—as it must be given that a high proportion of the adjacencies is missing.

Simulation 21 has a doubling in connection strength for half of the subjects. Although Smith et al. used this design for a different purpose, we found it relevant for testing the reliability of the search procedure when the data given to the search algorithms is from subjects with the same causal structure but different connection strengths. The variation had essentially no effect on the accuracy of the search, which remained nearly perfect for IMAGES + LOFS procedures.

Simulation 22 introduced modulation variables which independently switched off each effective connection 3/5 of the time on average. The effect of this extreme modulation was that in each subject the number of time points in which each effective connection exists was reduced by 60%, from 10 min to 4 min. Worse, because the "off" periods were independent for each edge, the percentage of time all edges in the graph simultaneously

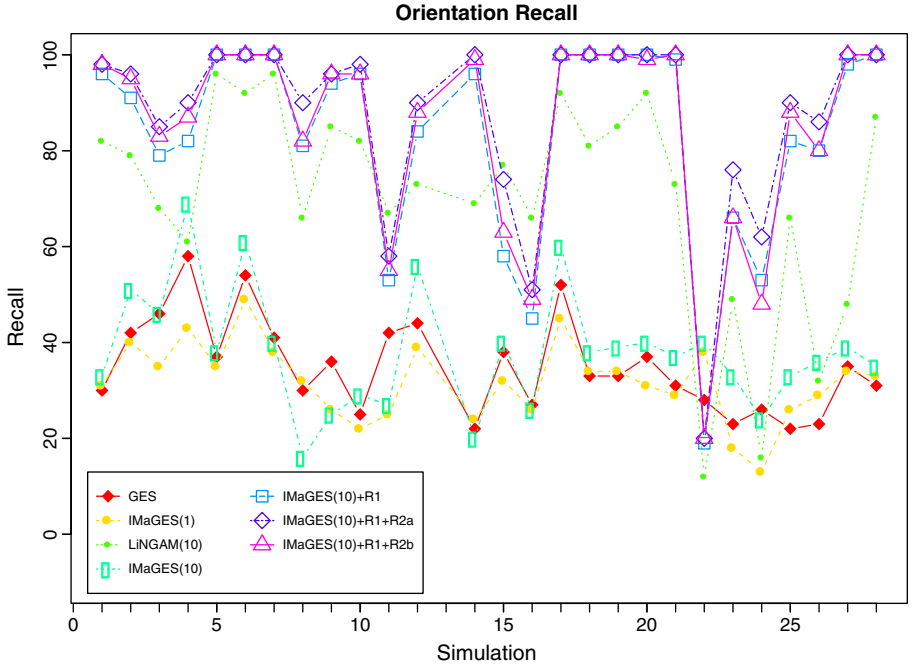

Fig. 5. Orientation recall = # true orientations in output / # orientations in true model, for each of simulations 1 through 28, for each algorithm being compared. See text for details. The algorithms are: GES, I(1) = IMaGES, first nontriangular, with one subject's data as input, L(10) = LiNGAM using concatenated data from 10 subjects, I(10) = IMaGES, first nontriangular, with 10 subjects' data as input, I(10) + R1 = I(10) followed by LOFS rule R1, I(10) + R1 + R2a = I(10) + R1 followed by LOFS rule R2 applied to undirected edges only, I(10) + R1 + R2b = I(10) + R1 followed by LOFS rule R2 applied to undirected edges and 2-cycles.

Please cite this article as: Ramsey, J.D., et al., Multi-subject search correctly identifies causal connections and most causal directions in the DCM models of the Smith et al. simu..., NeuroImage (2011), doi:10.1016/j.neuroimage.2011.06.068

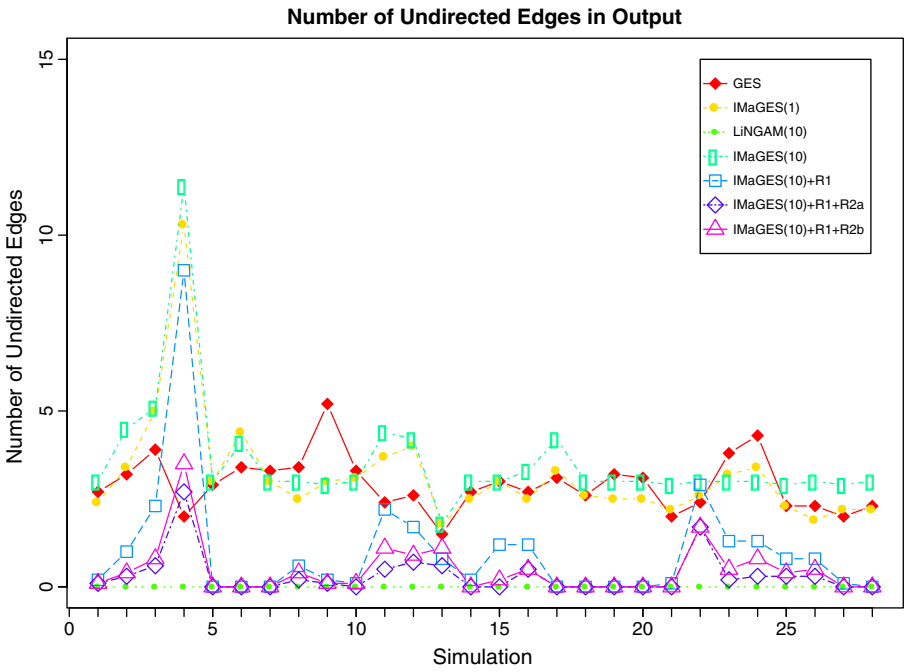

Fig. 6. Number of unoriented edges in output. See text for details. The algorithms are: GES, I(1) = IMaGES, first nontriangular, with one subject's data as input, L(10) = LiNGAM using concatenated data from 10 subjects, I(10) = IMaGES, first nontriangular, with 10 subjects' data as input, I(10) + R1 = I(10) followed by LOFS rule R1, I(10) + R1 + R2a = I(10) + R1 followed by LOFS rule R2 applied to undirected edges only, I(10) + R1 + R2b = I(10) + R1 followed by LOFS rule R2 applied to undirected edges and 2-cycles.

corresponded to effective connections is much less than 60%. The exact frequency is not calculable from the information given. Nonetheless, IMaGES found the adjacencies perfectly in simulation 22 with non-stationary changes in the effective connection coefficients. All of the edges IMaGES oriented were oriented correctly, but on average fewer than half of the edges were oriented. iMaGES orientation accuracy was actually better with the non-stationary coefficients of simulation 22 than with the fixed coefficients of an otherwise comparable design in simulation 23. The LOFS procedures were a disaster in simulation 22 but somewhat better than IMaGES alone on simulation 23. Simulation 24 has a much larger exogenous input for one variable than for the other variables. IMaGES adjacency accuracies are decreased but remain fairly good; orientation precision and recall are poor for all procedures possibly because if the causal structure is A→B p="0.12"/>←e then as the influence of e on B is increased, other things equal, the A, B association will diminish.

The accuracy of IMaGES alone or with IMaGES + LOFS with 10 subjects strongly dominates that of any of the search procedures on the tests reported by Smith et al. One might suppose that IMaGES + LOFS procedures with 10 subject input does better than the methods tested in Smith's study simply because IMaGES + LOFS receives a sample multiplied by 10 in comparison with the single subject applications. That hypothesis can be tested in two ways. First, the same multiples of 10 subjects were given to LiNGAM, with the time series for the 10 subjects appended to one another as Smith et al., suggest. The results are systematically inferior to IMaGES + LOFS procedures.

The second test is by comparison of IMaGES + LOFS on simulation 1 and the Smith et al., results in simulation 7. In simulation 7 the time series is for 250-minute runs under conditions that are otherwise the same as in simulation 1, which has 10 minute runs. Since the parameters are the same, each simulation 7 subject is the statistical equivalent of 25 simulation 1 subjects. Again, since the probability distributions and graphical structures are the same in all simulation 1 and simulation 7 subjects, simulation 7 provides a best possible case for how the various methods in the Smith et al. study would do using appended data for 25 subjects, each of whom was exposed to 10-minute sessions. Thus the Smith et al. results for simulation 7 for the various methods can be compared with our results for IMaGES + LOFS procedures for simulation 1 using 10 subject inputs to IMaGES. The comparison is in fact biased in favor of the methods Smith et al. study, since their sample size for each subject in simulation 7 is 2.5 times as large as the 10-subject sample used by IMaGES + LOFS. In simulation 7 in the Smith et al. experiments, our run of GES finds all adjacencies but on average 1 false positive edge, with poor orientation results. IMaGES + LOFs procedures are essentially perfect in simulation 1.

To illustrate the application of the LOFS procedures to real data we compare the graph found for Xue and Poldrack data (Xue and Poldrack, 2007) with IMaGES in Ramsey, et al., 2010 (Fig. 7), with the graph found for the same data by IMaGES + LOFS-R1 (Fig. 8). The difference in one adjacency is due to improvements in IMaGES implementation since 2010. The 2-cycles found by LOFS-R1 may be due to genuine feedback relations, or to unrecorded common causes, or both. LOFS-R1 cannot distinguish among those alternatives. Differences in orientations in the two graphs are due to the fact that IMaGES decides

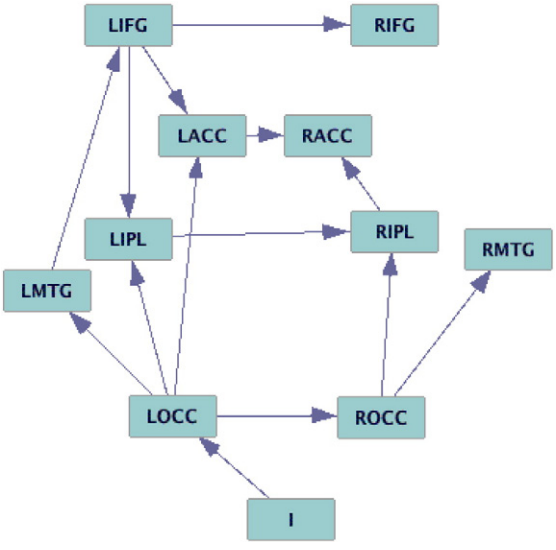

Fig. 7. IMaGES from Ramsey et al.(2010).

Please cite this article as: Ramsey, J.D., et al., Multi-subject search correctly identifies causal connections and most causal directions in the DCM models of the Smith et al. simu..., NeuroImage (2011), doi:10.1016/j.neuroimage.2011.06.068

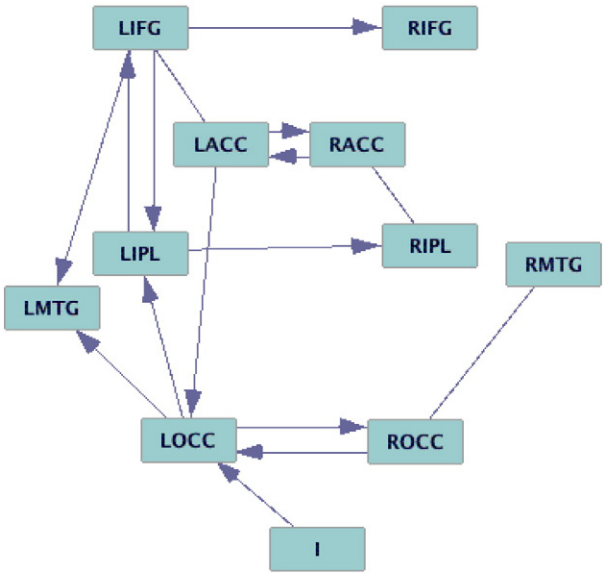

Fig. 8. IMaGES + LOFS-R1.

orientations chiefly by conditional independence and dependence relations, while IMaGES + LOFS bases orientation decisions on measures of deviation of residuals from Normal.

The concatenated 10-subject residuals from simulations in the Smith, et al. experiments are closer to Gaussian in skew and kurtosis than and the concatenated residuals from the subjects in the Xue and Poldrack study. For example, for a selection of 10 subjects from Simulation 2 in Smith et al., among parent orientations supplied by rule R1, the maximum skew of appended residuals is 0.28, and the maximum excess kurtosis is 0.51. By contrast, for the analysis given in Fig. 8, the maximum skew of appended residuals is 1.0 and the maximum excess kurtosis 39.3.

Conclusion

IMaGES run with 10 subjects with settings exactly as described in Ramsey et al.(2010) performs much better than any of the search algorithms as tested by Smith et al. For simulations that are most realistic, the accuracy and informativeness of adjacencies approaches 100%. Orientations of edges by IMaGES are also superior to algorithms in the Smith study, but orientations obtained with IMaGES adjacencies and LOFS orientations are much better. For real data, these adjacencies and orientations are in our experience easier to establish with block designs than with event designs (Chee et al., 2003).

The adjacency information in IMaGES is heuristic and essentially incomplete; it will only give a fragment of the adjacencies in more complex structures. We judge it better to have accurate but partial information rather than more complete but inaccurate output. A correct algorithm for adjacencies and orientations of edges from linear systems with feedback is available (Lacerda, et al., 2008), but its sample size requirements are too large for fMRI work. No search algorithm is known that is provably correct for linear systems with both latent variables and feedback cycles.

We have successfully run IMaGES on simulated data with up to 500 variables. By adjusting search parameters to force sparsity of output, problems with 5000 or more variables can be addressed with other "Bayes nets" algorithms tested by Smith et al. such as the PC algorithm. These algorithms, which use conditional independence tests, can be used with multiple data sets by methods that form a combined p value. In our experience their accuracy for adjacencies is comparable to that of IMaGES.

Taken together, we believe our results belie the conclusion that nothing works and that nothing can work for estimating causal relations from fMRI data. Smith et al. remark that our multiple subject approach "would then restrict the ability to use simple methods (such as cross-subject mixed-effects modeling of the estimated network parameters) to determine the reliability of the group-estimated network." But only slightly more complex methods are available, for example leave one subject out invariance and other resampling tests. Our results do point to directions where further work is needed; in particular we need improved ability to find more dense generating structures and to find confounding by unmeasured factors. And, finally, we think the Smith et al. study and our own together give evidence that computer aids for model specification in fMRI cannot be taken off the shelf from the machine learning and other literatures. The statistical problems fMRI poses are special.

The procedures we have described should not be thought of as buttons that are simply to be pushed. Close examination of the trace of a search will sometimes find LOFS orientation decisions that are very close calls. Resampling may show instability in the estimated causal structures. Examination of residuals may indicate that they are too Gaussian for LOFS procedures to be trusted. And so on. Rather than thinking of the algorithms as replacements for theorizing, they should be thought of as robotic colleagues who can advise on model specification, can provide reasons for their advice (the trace of a LOFS procedure especially can do so), can give advice based on *your* suppositions or prior knowledge, and who can be retired without cost when better robotic colleagues become available.

Software for the algorithms described here Glymour et al., (2011) is available at www.phil.cmu.edu/projects/tetrad.

Appendix A. Tabulated accuracies

Statistics for adjacency and orientation precision and recall, and number of undirected edges, for various algorithms described in the text, applied to Smith et al. simulation data, are given in Tables 2 through 6.

Table 2
Adjacency precision = # true adjacencies in output/# adjacencies in output, for each of simulations 1 through 28, for each algorithm being compared. See text for details. The algorithms are: GES, I(1) = IMaGES, first nontriangular, with one subject's data as input, L(10) = LiNGAM using concatenated data from 10 subjects, I(10) = IMaGES, first nontriangular, with 10 subjects' data as input, I(10).R1 = I(10) followed by LOFS rule R1, I(10).R1.R2a = I(10).R1 followed by LOFS rule R2 applied to undirected edges only, I(10).R1.R2b = I(10).R1 followed by LOFS rule R2 applied to undirected edges and 2-cycles.

Sim	GES	I(1)	L(10)	I(10)	I(10).R1	I(10).R1.R2a	I(10).R1.R2b
1	98	99	99	100	100	100	100
2	86	90	96	100	100	100	100
3	79	92	93	100	100	100	100
4	51	87	92	100	100	100	100
5	90	100	96	100	100	100	100
6	84	99	89	100	100	100	100
7	82	100	90	100	100	100	100
8	78	80	97	98	98	98	98
9	65	88	97	98	98	98	98
10	87	97	88	100	100	100	100
11	52	55	54	55	55	53	55
12	82	90	95	100	100	100	100
13	85	91	94	100	100	100	100
14	95	94	93	100	100	100	100
15	87	95	87	100	100	100	100
16	93	96	94	100	100	100	100
17	84	94	97	100	100	100	100
18	97	100	98	100	100	100	100
19	73	98	100	100	100	100	100
20	74	97	100	100	100	100	100
21	95	96	100	100	100	100	100
22	85	92	78	100	100	100	100
23	76	83	89	97	97	97	97
24	58	64	76	84	84	88	84
25	88	91	100	99	99	99	99
26	86	87	96	100	100	100	100
27	89	87	99	100	100	100	100
28	95	97	100	100	100	100	100

Please cite this article as: Ramsey, J.D., et al., Multi-subject search correctly identifies causal connections and most causal directions in the DCM models of the Smith et al. simu..., NeuroImage (2011), doi:10.1016/j.neuroimage.2011.06.068

Table 3
Adjacency recall = # true adjacencies in output / # adjacencies in true model, for each of simulations 1 through 28, for each algorithm being compared. See text for details. The algorithms are: GES, I(1) = IMaGES, first nontriangular, with one subject's data as input, L(10) = LiNGAM using concatenated data from 10 subjects, I(10) = IMaGES, first nontriangular, with 10 subjects' data as input, I(10).R1 = I(10) followed by LOFS rule R1, I(10).R1.R2a = I(10).R1 followed by LOFS rule R2 applied to undirected edges only, I(10).R1.R2b = I(10).R1 followed by LOFS rule R2 applied to undirected edges and 2-cycles.

Sim	GES	I(1)	L(10)	I(10)	I(10).R1	I(10).R1.R2a	I(10).R1.R2b
1	96	92	97	100	100	100	100
2	92	85	97	100	100	100	100
3	85	74	94	92	92	92	92
4	90	76	90	98	98	98	98
5	100	100	100	100	100	100	100
6	100	99	100	100	100	100	100
7	100	100	100	100	100	100	100
8	90	74	84	96	96	96	96
9	100	86	99	98	98	98	98
10	100	93	100	100	100	100	100
11	78	55	82	62	62	62	62
12	88	82	96	100	100	100	100
13	65	62	16	90	90	90	90
14	92	92	96	100	100	100	100
15	99	94	96	100	100	100	100
16	80	59	89	62	62	62	62
17	97	93	99	100	100	100	100
18	93	91	97	100	100	100	100
19	100	97	87	100	100	100	100
20	99	96	92	100	100	100	100
21	87	85	86	100	100	100	100
22	75	78	28	100	100	100	100
23	91	81	78	97	97	97	97
24	80	61	25	82	82	82	82
25	75	75	78	98	98	98	98
26	60	60	41	97	97	97	97
27	78	75	58	100	100	100	100
28	93	85	95	100	100	100	100

Table 4
Orientation precision = # true orientations in output / # orientations in output, for each of simulations 1 through 28, for each algorithm being compared. See text for details. Asterisks are marked for Simulation 13, since orientations (but not adjacencies) in true models vary by subject. The algorithms are: GES, I(1) = IMaGES, first nontriangular, with one subject's data as input, L(10) = LiNGAM using concatenated data from 10 subjects, I(10) = IMaGES, first nontriangular, with 10 subjects' data as input, I(10).R1 = I(10) followed by LOFS rule R1, I(10).R1.R2a = I(10).R1 followed by LOFS rule R2 applied to undirected edges only, I(10).R1.R2b = I(10).R1 followed by LOFS rule R2 applied to undirected edges and 2-cycles.

Sim	GES	I(1)	L(10)	I(10)	I(10).R1	I(10).R1.R2a	I(10).R1.R2b
1	57	55	84	82	99	99	100
2	55	63	79	87	96	96	98
3	54	67	68	70	91	89	94
4	34	61	62	87	90	88	94
5	78	84	93	94	100	100	100
6	62	83	83	96	100	100	100
7	75	95	87	100	100	100	100
8	48	42	76	36	85	86	91
9	29	48	83	61	98	98	98
10	51	43	73	72	91	91	98
11	33	29	44	34	48	46	54
12	54	58	72	90	94	92	96
13	*	*	*	*	*	*	*
14	39	41	67	50	98	99	100
15	69	74	72	100	60	63	65
16	46	28	70	29	93	86	87
17	62	68	91	96	100	100	100
18	59	66	82	96	99	99	100
19	35	62	98	92	100	100	100
20	43	58	99	98	99	99	100
21	54	50	85	88	100	100	100
22	40	44	36	100	50	33	33
23	34	31	54	82	70	66	74
24	27	16	53	44	51	49	58
25	28	34	84	80	94	93	95
26	27	30	77	74	83	84	92
27	56	49	82	92	100	100	100
28	50	56	91	85	100	100	100

Table 5
Orientation recall = # true orientations in output / # orientations in true model, for each of simulations 1 through 28, for each algorithm being compared. See text for details. Asterisks are marked for Simulation 13, since orientations (but not adjacencies) in true models vary by subject. The algorithms are: GES, I(1) = IMaGES, first nontriangular, with one subject's data as input, L(10) = LiNGAM using concatenated data from 10 subjects, I(10) = IMaGES, first nontriangular, with 10 subjects' data as input, I(10).R1 = I(10) followed by LOFS rule R1, I(10).R1.R2a = I(10).R1 followed by LOFS rule R2 applied to undirected edges only, I(10).R1.R2b = I(10).R1 followed by LOFS rule R2 applied to undirected edges and 2-cycles.

Sim	GES	I(1)	L(10)	I(10)	I(10).R1	I(10).R1.R2a	I(10).R1.R2b
1	30	31	82	33	96	98	98
2	42	40	79	51	91	96	95
3	46	35	68	46	79	85	83
4	58	43	61	69	82	90	87
5	37	35	96	38	100	100	100
6	54	49	92	61	100	100	100
7	41	38	96	40	100	100	100
8	30	32	66	16	81	90	82
9	36	26	85	25	94	96	96
10	25	22	82	29	96	98	96
11	42	25	67	27	53	58	55
12	44	39	73	56	84	90	88
13	*	*	*	*	*	*	*
14	22	24	69	20	96	100	99
15	38	32	77	40	58	74	63
16	27	26	66	26	45	51	49
17	52	45	92	60	100	100	100
18	33	34	81	38	100	100	100
19	33	34	85	39	100	100	100
20	37	31	92	40	100	100	99
21	31	29	73	37	99	100	100
22	28	38	12	40	19	20	20
23	23	18	49	33	66	76	66
24	26	13	16	24	53	62	48
25	22	26	66	33	82	90	88
26	23	29	32	36	80	86	80
27	35	34	48	39	98	100	100
28	31	33	87	35	100	100	100

Table 6
Number of unoriented edges in output model, for each of simulations 1 through 28, for each algorithm being compared. See text for details. The algorithms are: GES, I(1) = IMaGES, first nontriangular, with one subject's data as input, L(10) = LiNGAM using concatenated data from 10 subjects, I(10) = IMaGES, first nontriangular, with 10 subjects' data as input, I(10).R1 = I(10) followed by LOFS rule R1, I(10).R1.R2a = I(10).R1 followed by LOFS rule R2 applied to undirected edges only, I(10).R1.R2b = I(10).R1 followed by LOFS rule R2 applied to undirected edges and 2-cycles.

Sim	GES	I(1)	L(10)	I(10)	I(10).R1	I(10).R1.R2a	I(10).R1.R2b
1	2.7	2.4	0.0	3.0	0.2	0.1	0.1
2	3.2	3.4	0.0	4.5	1.0	0.3	0.4
3	3.9	5.0	0.0	5.1	2.3	0.6	0.8
4	2.0	10.3	0.0	11.4	9.0	2.7	3.5
5	2.9	3.0	0.0	3.0	0.0	0.0	0.0
6	3.4	4.4	0.0	4.1	0.0	0.0	0.0
7	3.3	3.0	0.0	3.0	0.0	0.0	0.0
8	3.4	2.5	0.0	3.0	0.6	0.2	0.4
9	5.2	3.0	0.0	2.9	0.2	0.1	0.1
10	3.3	3.1	0.0	3.0	0.1	0.0	0.1
11	2.4	3.7	0.0	4.4	2.2	0.5	1.1
12	2.6	4.0	0.0	4.2	1.7	0.7	0.9
13	1.5	1.8	0.0	1.8	0.8	0.6	1.1
14	2.7	2.5	0.0	3.0	0.2	0.0	0.0
15	3.0	3.0	0.0	3.0	1.2	0.0	0.2
16	2.7	2.5	0.0	3.3	1.2	0.5	0.5
17	3.1	3.3	0.0	4.2	0.0	0.0	0.0
18	2.6	2.6	0.0	3.0	0.0	0.0	0.0
19	3.2	2.5	0.0	3.0	0.0	0.0	0.0
20	3.1	2.5	0.0	3.0	0.0	0.0	0.0
21	2.0	2.2	0.0	2.9	0.1	0.0	0.0
22	2.4	2.6	0.0	3.0	2.9	1.7	1.7
23	3.8	3.2	0.0	3.0	1.3	0.2	0.5
24	4.3	3.4	0.0	3.0	1.3	0.3	0.8
25	2.3	2.3	0.0	2.9	0.8	0.3	0.4
26	2.3	1.9	0.0	3.0	0.8	0.3	0.5
27	2.0	2.2	0.0	2.9	0.1	0.0	0.0
28	2.3	2.2	0.0	3.0	0.0	0.0	0.0

Please cite this article as: Ramsey, J.D., et al., Multi-subject search correctly identifies causal connections and most causal directions in the DCM models of the Smith et al. simu..., NeuroImage (2011), doi:10.1016/j.neuroimage.2011.06.068

References

Anderson, T.W., Darling, D.A., 1952. Asymptotic theory of certain "goodness-of-fit" criteria based on stochastic processes. Ann. Math. Stat. 23, 193–212.

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

Chee, M.W., Venkatraman, V., Westphal, C., Siong, S.C., 2003. Comparison of block and event-related fMRI designs in evaluating the word-frequency effect. Hum. Brain Mapp. 18 (3), 186–193.

D'Agostino, R.B., 1986. Tests for the normal distribution. In: D'Agostino, R.B., Stephens, M.A. (Eds.), Goodness-of-Fit Techniques. Marcel Dekker, New York.

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

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.

Hoyer, P.O., Hyvärinen, Aapo, Scheines, R., Spirtes, P., Ramsey, J., Lacerda, G., Shimizu, S., 2008. Causal discovery of linear acyclic models with arbitrary distributions. Proceedings of the 24th Annual Conference on Uncertainty in Artificial Intelligence. AUAI Press, Oregon.

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.

Pearl, J., 2000. Causality: Models, Reasoning, and Methods. Cambridge University Press, Cambridge.

Ramsey, J.D., Hanson, S.J., Hanson, C., Halchenko, Y.O., Poldrack, R.A., Glymour, C., 2010. Six problems for causal inference. NeuroImage 49 (2), 1545–1588.

Ramsey, J.D., Spirtes, P., Glymour, C., 2011. On meta-analysis of imaging data and the mixture of records. NeuroImage 57 (2), 323–330.

Shimizu, S., Hyvärinen, A., Kano, Y., Hoyer, P.O., 2005. Discovery of non-Gaussian linear causal models using ICA. Proceedings of the 21st Conference of Uncertainty in Artificial Intelligence. UAUI Press, Oregon.

Smith, S.M., Miller, K.L., Salimi-Khorshidi, G., Webster, M., Beckmann, C., Nichols, T., Ramsey, J., Woolrich, M., 2011. Network modeling methods for FMRI. NeuroImage 54 (2), 875–891.

Glymour, C., Spirtes, P., Scheines, R., Ramsey, J., 2011. Tetrad IV. http://www.phil.cmu.edu/tetrad.

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.

Please cite this article as: Ramsey, J.D., et al., Multi-subject search correctly identifies causal connections and most causal directions in the DCM models of the Smith et al. simu..., NeuroImage (2011), doi:10.1016/j.neuroimage.2011.06.068