

A million variables and more: the Fast Greedy Equivalence Search algorithm for learning high-dimensional graphical causal models, with an application to functional magnetic resonance images

Joseph Ramsey¹ · Madelyn Glymour¹ · Ruben Sanchez-Romero¹ · Clark Glymour¹

Received: 7 September 2016 / Accepted: 6 November 2016 / Published online: 1 December 2016
© Springer International Publishing Switzerland 2016

Abstract We describe two modifications that parallelize and reorganize caching in the well-known Greedy Equivalence Search algorithm for discovering directed acyclic graphs on random variables from sample values. We apply one of these modifications, the Fast Greedy Equivalence Search (fGES) assuming faithfulness, to an i.i.d. sample of 1000 units to recover with high precision and good recall an average degree 2 directed acyclic graph with one million Gaussian variables. We describe a modification of the algorithm to rapidly find the Markov Blanket of any variable in a high dimensional system. Using 51,000 voxels that parcellate an entire human cortex, we apply the fGES algorithm to blood oxygenation level-dependent time series obtained from resting state fMRI.

Keywords Greedy Equivalence Search · Scalable · Parallelized · Markov Blanket · FMRI

1 Introduction

High-dimensional data sets afford the possibility of extracting causal information about complex systems from samples. Causal information is now commonly represented by directed graphs associated with a family of probability distributions. Fast, accurate recovery is needed for such systems.

Electronic supplementary material The online version of this article (doi:10.1007/s41060-016-0032-z) contains supplementary material, which is available to authorized users.

✉ Joseph Ramsey
jsph.ramsey@gmail.com

¹ Carnegie Mellon University, Pittsburgh, PA, USA

An acyclic directed graphical (DAG) model $\langle G(\mathbf{V}, \mathbf{E}), P \rangle$ consists of a directed acyclic graph, G , whose vertices v are random variables with directed edges, e , and a probability distribution, P , satisfying the Markov factorization for all assignments of values to variables in \mathbf{V} having positive support:

$$\prod_{v \in \mathbf{V}} P(v) | \mathbf{PA}(v) \quad (1)$$

where $\mathbf{PA}(v)$ denotes a value assignment to each member of the set of variables with edges directed into v (the parents of v). The Markov Equivalence Class (MEC) of a DAG G is the set of all DAGs G having the same adjacencies as G and the same “ v -structures”—substructures $x \rightarrow y \leftarrow z$, where x and y are not adjacent in G .

Such graphs can be used simply as devices for computing conditional probabilities. When, however, a causal interpretation is appropriate, search for DAG models offers insight into mechanisms and the effects of potential interventions, finds alternatives to a priori hypotheses, and in appropriate domains suggests experiments. Under a causal interpretation of G , a directed edge $x \rightarrow y$ represents the proposition that there exist values of all variables in $\mathbf{V} \setminus \{x, y\}$ such that if these variables were (hypothetically) fixed at those values (not conditioned on those values), some exogenous variation of x would be associated with variation in y .

A causal interpretation is inappropriate for Pearson correlations of time series, because the correlations are symmetric and do not identify causal pathways leading from one variable to another; correlation searches will return not only an undirected edge (adjacency) between the intermediate variables in a causal chain, but also adjacencies for the transitive closure of a chain of causal connections. Causal interpretations are usually not correct for Markov Random Field (MRF) models, including those obtained for high dimensions by var-

ious penalized inverse covariance methods such as GLASSO and LASSO [1,2], in part because like correlation graphs MRF graphs are undirected, and in the large sample limit specify an undirected edge between two variables that are conditionally independent but dependent when a third variable is conditioned on. This pattern of dependencies is characteristic of circumstances in which two conditionally independent variables are jointly causes of a third. It should be of interest therefore whether DAG search procedures can be speeded up to very high dimensions, with what accuracies. Ideally, one should be able to analyze tens of thousands of variables on a laptop in a work session and be able to analyze problems with a million variables or more on a super-computer overnight. Such a capability would be useful, for example, in cognitive neuroscience and in learning DAGs from high-dimensional biomedical datasets.

2 Strategies

Various algorithmic strategies have been developed for searching data for DAG models. One strategy, incorporated in GES, iteratively adds edges starting with an empty graph according to maximal increases in a score, generally the BIC score [3] for Gaussian variables, and the BDeu score [4] for multinomial variables, although many other scores are possible, including modifications of the Fisher Z score [5] for conditional independence tests. The algorithms return a MEC of models, with the guarantee that if the distribution P is in the MEC for some DAG over the sample variables then asymptotically in probability the models returned have the highest posterior probability. There is no guarantee that on finite samples the model or models returned are those with the highest posterior probability, which is known to be an NP-hard problem [6].

The GES algorithm, as previously implemented, has trouble with large variable sets. Smith et al. gave up running GES on low dimensional problems (e.g., 50 variables and 200 datapoints) on the grounds that too much time was required for the multiple runs their simulations required [7].

Another area of interest for scaling up MEC search has been the search for Markov Blankets of variables, the minimal set of variables, not including a target variable t , conditional on which all other recorded variables are independent of t [8–10]. This has an obvious advantage for scalability; if the nodes surrounding a target variable in a large data set can be identified, the number of relations among nodes that need to be assessed is reduced. Since the literature has not suggested a way to estimate Markov Blankets using a GES-style algorithm, we provide one here. It is scalable for sparse graphs and yields estimates of graphical structure about the target that are simple restrictions of the MEC of the graphical causal structure generating the data.

3 Fast Greedy Equivalence Search

The elements of a GES search and relevant proofs, but no precise algorithm, were given by Chickering [11]; the basic idea is given by Meek [12]. Free public implementations of GES are available in the *pcalg* package in *R* and the Tetrad software suite. GES searches iteratively through the MECs of one-edge additions to the current MEC, beginning with the totally disconnected graph on the recorded variables. At each stage of the forward phase, all alternative one-edge additions are scored; the edge with the best score is added and the resulting MEC formed, until no more improvements in the score can be made by a single-edge addition; in the backward stage, the procedure is repeated but this time removing edges, starting with the MEC of the forward stage, until no more improvements in the score can be made by single-edge deletions. For multinomial distributions, the algorithm has two intrinsic tuning parameters: a “structure prior” that penalizes for the number of parents of a variable in a graph, and a “sample prior” required for the Dirichlet prior distribution. For Gaussian distributions, the algorithm has one tuning parameter, the complexity penalty in the BIC score. The original BIC score has a penalty proportional to the number of variables, but without changing asymptotic convergence to the MEC of DAGs with the maximum posterior probability, this penalty can be increased or decreased (to a positive fraction) to control false positive or false negative adjacencies. Increasing the penalty forces sparser graphs on finite samples.

It is important to note that the proof of correctness of GES assumes causal sufficiency—that is, it assumes that a common cause of any pair of variables in the set of variables over which the algorithm reasons is in the set of variables. That is, if the algorithm is run over a set of variables \mathbf{X} , and x and y are in \mathbf{X} , and $x \leftarrow L \rightarrow y$ for some variable L , then L is in \mathbf{X} as well. Thus, GES assumes that there are no unmeasured common causes. If there are unmeasured common causes, then GES will systematically introduce extra edges. GES also assumes that there are no feedback loops that would have to be represented by a finite cyclic graph.

The fGES procedure uses a similar strategy with some important differences. First, in accord with (1), the scores are decomposable by graphical structure, so that the score for an entire graph may be the sum of scores for fragments of it. The updating of the score for a potential edge addition is done by caching scores of potential edge additions from previous steps, and where a new edge addition will not (for graphical reasons) alter the score of a fragment of the graph, uses the cached score, saving time in graph scoring. This requires an increase in memory, since the local scores for the relevant fragments must be stored.

Second, each step of the fGES procedure may be parallelized. For most of the running time of the algorithm, a parallelization can be given for which all processors are used,

even on a large machine, radically improving the running time of the algorithm.

Third, if the BIC score is used, the penalty for this score can be increased, forcing estimation of a sparser graph. The sparser the graph, the faster the search returns but at the risk of false negative adjacencies.

Fourth, one can make the assumption that an edge $x \rightarrow y$, where x and y are uncorrelated, is not added to the graph at any step in the forward search. This is a limited version of the so-called faithfulness assumption [5] and allows a search procedure to be speeded up considerably at the expense of allowing graphs to violate the Markov factorization, (1), for the conditional dependence and independence relations estimated from the sample. This weak faithfulness assumption can be included in fGES, or not. For low dimensional problems where speed is not an issue, there is no compelling reason to assume this kind of one-edge faithfulness, but in high dimensions the search can be much faster if the assumption is made. There are two situations in which assuming one-edge faithfulness leads to incorrect graphs. The first is perfectly canceling paths. Consider a model $A \rightarrow B \rightarrow C \rightarrow D$, with $A \rightarrow D$, where the two paths exactly cancel. fGES without the weak faithfulness assumption will (asymptotically) identify the correct model; with that assumption, the $A \rightarrow D$ edge will go missing. In this regard, fGES with the assumption behaves like any algorithm whose convergence proof assumes faithfulness, such as the PC algorithm [5], although only with regard to single-edge violations of faithfulness. Faithfulness conditions for perfectly canceling paths have been shown to hold measure 1 for continuous measures over the parameter spaces of either discrete or continuous joint probability distributions on variable values, but in estimates of joint distributions from finite samples faithfulness can be violated for much larger sets of parameter values [5]. We will use this one-edge faithfulness assumption in what follows.

fGES is implemented, with publicly accessible code, in the development branch of <https://github.com/cmu-phil/tetrad>. Pseudo-code is given in the supplement to this paper.

The continuous BIC score for the difference between a graph with a set of parents \mathbf{Z} of y , $\mathbf{Z} \rightarrow y$, and a graph with an additional parent x of Y , $\mathbf{Z} \cup \{x\} \rightarrow y$, is $\text{BIC}(\mathbf{Z} \cup \{x\}, y) - \text{BIC}(\mathbf{Z}, y)$, where for an arbitrary set \mathbf{X} of variables, $\text{BIC}(\mathbf{X}, y) = 2L - ck \ln n$, with L the log likelihood of the linear model $\mathbf{X} \rightarrow y$ with Gaussian residual (equal to $-n/2 \ln s^2 + C$), $k = |\mathbf{X}| + 1$, n is the sample size, s^2 is the variance of the residual of y regressed on \mathbf{X} , and c is a constant (“penalty discount”) used to increase or decrease the penalty of the score. Chickering [11] shows this score is positive if it is not the case that $(x \perp\!\!\!\perp y|\mathbf{Z})$ and negative if $x \perp\!\!\!\perp y|\mathbf{Z}$.

The discrete BDeu score is as given by Chickering [11]. We use differences of these scores as for the continuous case

above. We use a modified structure prior, as described in Sect. 4.

The correctness of the implementation from d -separation [13] can be tested using Chickering’s “locally consistent scoring criterion” theorem. Using d -separation, any DAG can be consistently scored by +1 if it is not the case that $d\text{sep}(x, y|\mathbf{Z})$ and -1 if $d\text{sep}(x, y|\mathbf{Z})$. This score may be used to run fGES directly from a DAG using d -separation facts alone and can be used to check the correctness of the implementation of the algorithm without using data and the difference scores. Running fGES using the graph score should recover the MEC of G . We have tested the implementation in this way on more than 100,000 random DAGs with 10 nodes and 10 or 20 edges without an error.

4 Multiple samples with fGES

Appending data sets can lead to associations among variables that are independent in each data set. For that reason, the IMaGES algorithm [14] runs each step of GES separately on each data set and updates the inferred MEC by adding (or in the backwards phase, deleting) the edge whose total score summed over the scores from the several data sets is best. The same strategy can be used with fGES. The scoring over the separate data sets can of course be parallelized to reduce runtime almost in proportion to the number of data sets, but we have not yet done so in our published software.

5 Simulations with BIC and BDeu

Our simulations generated data from directed acyclic graphs parameterized with linear relations and independent, identically distributed Gaussian noise, or with categorical, three valued variables, parameterized as multinomial. For continuous distributions, all coefficients were sampled uniformly from $(-1.5, -.5) \cup (.5, 1.5)$. For categorical variables, for each cell of the conditional probability table, random values were drawn uniformly from $(0, 1)$ and rows were normalized. All simulations were conducted using Pittsburgh Supercomputing Center (PSC) Blacklight computer. The number of processors (hardware threads) used varied from 2 (for 1000 variables) to 120 (for one million variables). All simulations used samples of size 1000. Where repeated simulations were done for the same dimension, a new graph, parameterization, and sample were used for each trial. The BIC penalty was multiplied by 4 in all searches with continuous data. The BDeu score used is as by Chickering [11], except that we used the following structure prior (for each vertex in a possible DAG):

$$(e/(v-1))^p + (1-e/(v-1))^{v-p-1} \quad (2)$$

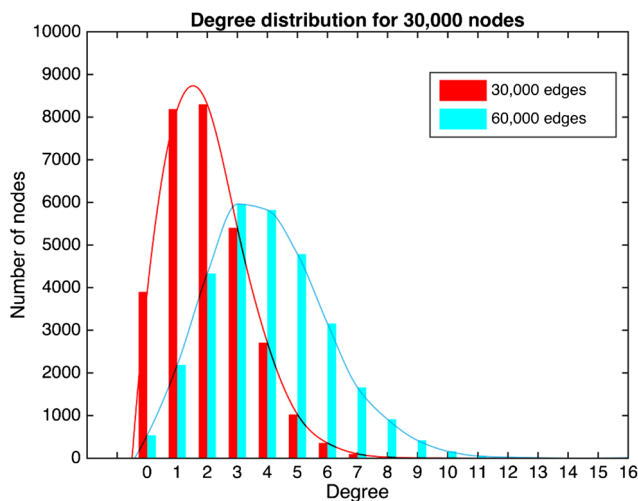


Fig. 1 Node degree distribution for simulated graphs with 30,000 nodes and 30,000 or 60,000 edges

where v is the number of variables in a DAG, p is the number of parents of that vertex and e is a prior weight, by default equal to 1. In (2), whether a node X_i has a node X_j as a parent is modeled as a Bernoulli trial with probability p_j of X_j being a parent of X_i and $(1 - p_j)$ of X_j not being a parent of X_i . In simulations, we have found this structure prior yields more accurate estimates of the DAG than the structure prior suggested by Chickering [11]. We used a sample prior of 1.

For continuous data, we have run fGES with the penalized BIC score defined above for (1) DAGs with 1000 nodes and 1000 or 2000 edges, for 100 repetitions, (2) DAGs with 30,000 nodes and 30,000 or 60,000 edges, for 10 repetitions, and (3) a DAG with 1,000,000 nodes and 1,000,000 edges, in all cases assuming the weak faithfulness condition noted above. For discrete data, we have run fGES with the BDeu only for circumstances (1) and (2) above. Random graphs were generated by constructing a list of variables and then adding forward edges by random iteration, avoiding cycles. The node degree distribution returned by the search from samples from random graphs with 30,000 nodes and 30,000 or 60,000 edges generated by our method is shown in Fig. 1. Continuous results are tabulated in Table 1; discrete results are tabulated in Table 2. Where M1 is the true MEC of the DAG and M2 is the estimated MEC returned by fGES, precision for adjacencies was calculate as $TP/(TP + FP)$; recall as

$TP/(TP + FN)$, where TP is the number of adjacencies shared by the true (M1) and estimated (M2) MECs of the DAG, FP is the number of adjacencies in M2 but not in M1, and FN is the number of adjacencies in M1 but not in M2.

Arrowhead precision and recall were calculated in a similar way. An arrowhead was taken to be in M1 and M2 for each variable A and B such that $A \rightarrow B$ in both M1 and M2, and an arrowhead was taken to be in one MEC but not the other for each variable A and B such that $A \rightarrow B$ in one but $A \leftarrow B$ in the other, or $A - B$ in the other, or A and B are not adjacent in the other.

In all cases, we generate a DAG G randomly with the given number of nodes and edges, parameterize it as a linear, Gaussian structural equation model, run fGES as described using the penalized BIC score with BIC penalty multiplied by 4, and calculate average precision and recall for adjacencies and arrowheads with respect to the MEC of G . We record average running time. Times are shown in hours (h), minutes (min), or seconds (s).

In all cases, we generated a DAG G randomly with the given number of nodes and edges, parameterize it as a multinomial model with 3 categories for each variable, ran fGES as described using the penalized BDeu score with sample prior 1 and structure prior 1, and calculated average precision and recall for adjacencies and arrowheads with respect to the MEC of G . We record average running time.

For time comparisons (accuracies were the same), searches were also run on a 3.1 GHz MacBook Pro laptop computer with 2 cores and 4 hardware threads, for 1000 and 30,000 variable continuous problems. Runtime for 1000 nodes with 1000 edges was 3.6 s. 30,000 variable problems with 30,000 edges required 5.6 min. On the same machine, 1000 variable categorical search with 2000 edges required 3.9 s, and 30,000 variables with 60,000 edges required (because of a 16 GB memory constraint) 40.3 min.

The results in Tables 1 and 2 use the one-edge faithfulness assumption, above. Using the implementation described in the Appendix of Supplementary Material, this assumption does not need to be made if time is not of the essence. The running time is, however, considerably longer, as shown in Table 3 for the case of 30,000 node graphs. Accuracy is not appreciably better.

Table 1 Accuracy and time results for continuous data

# Nodes	# Edges	# Rep	Adj prec (%)	Adj rec (%)	Arr prec (%)	Arr rec (%)	# Processors	Elapsed
1000	1000	100	98.92	94.77	98.92	90.05	2	1.2 s
1000	2000	100	98.43	88.04	96.27	85.74	4	8.5 s
30,000	30,000	10	99.77	94.60	99.04	89.97	120	53.5 s
30,000	60,000	10	99.81	86.72	99.23	84.47	120	3.4 min
1,000,000	1,000,000	1	93.90	94.83	83.11	90.57	120	11.0 h

Table 2 Accuracy and time results for discrete data

# Nodes	# Edges	# Rep	Adj prec (%)	Adj rec (%)	Arr prec (%)	Arr rec (%)	# Processors	Elapsed
1000	1000	100	99.65	74.98	89.52	43.79	2	2.1 s
1000	2000	100	48.48	82.70	82.70	28.41	4	2.2 s
30,000	30,000	10	99.96	72.18	92.46	37.97	120	2.6 min
30,000	60,000	10	99.97	45.85	86.39	23.89	120	3.2 min

Table 3 Accuracy and time results for the 30,000 node cases, for continuous data, not assuming one-edge faithfulness

# Nodes	# Edges	# Rep	Adj prec (%)	Adj rec (%)	Arr prec (%)	Arr rec (%)	# Processors	Elapsed
30,000	30,000	1	99.75	94.55	98.88	89.91	120	54.8 s
30,000	60,000	1	99.83	87.74	99.32	85.63	120	8.9 min

6 Markov Blanket search

A number of algorithms have been proposed to calculate the Markov Blanket of an arbitrary target variable t without calculating the graphical structure over the nodes. Other algorithms attempt to identify the graphical structure over these nodes (that is, the parents of t , the children of t , and the parents of children of t). (An excellent overview of the field as of 2010 is given by Aliferis et al. [9]; see also [16]). A simple restriction of fGES, fGES-MB, finds the Markov Blanket of any specified variable and estimates the graph over those variables including all orientations that fGES would estimate for edges in that graph.

For the Markov blanket search, one calculates a set \mathbf{A} of adjacencies x - y about t as follows. First, given a score $I(a, b, C)$, one finds the set of variables x such that $I(x, t, \emptyset) > 0$ and adds x - t to \mathbf{A} for each x found. Then for each such variable x found, one finds the set of variables y such that $I(y, x, \emptyset) > 0$ and adds y - x to \mathbf{A} for each such y found. The general form of this calculation is familiar for Markov blanket searches, though in this case it is carried out using a score difference function. One then simply runs the rest of fGES, restricting adjacencies used in the search to the adjacencies in \mathbf{A} in the first pass and then marrying parents as necessary and re-running the backward search to get the remaining unshielded colliders, inferring additional orientations as necessary. The resulting graph may then simply be trimmed to the variables in \mathbf{A} and returned.

That this procedure returns the same result as running fGES over the entire set of variables and then trimming it to the variables in \mathbf{A} is implied by the role of common causes in the search, assuming one-edge faithfulness. \mathbf{A} consists of all variables d -connected to the target or d -connected to a variable d -connected to the target. But any common cause of any two variables in this Markov blanket of t is immediately in this set and so will be taken account of by fGES-MB. Trimming the search over \mathbf{A} to the nodes that are adjacents of t or spouses of t will then produce the same result (and

the correct Markov blanket) as running fGES over all of the (causally sufficient) variables and then trimming that graph in the same way to the Markov blanket of t .

For example, consider how fGES-MB is able to orient a single parent of t . Let $\{x, y\} \rightarrow w \rightarrow r \rightarrow t$ be the generating DAG. The step above will form a clique over x, y, w, r , and t (except for the edge x - y). It will then orient $x \rightarrow w \leftarrow y$, either directly or by shielding in the forward step and then removing the shield and orienting the collider in the backward step. This collider orientation will then be propagated to t by the Meek rules, yielding graph $\{x, y\} \rightarrow w \rightarrow r \rightarrow t$. This graph will then be trimmed to $r \rightarrow t$, the correct restriction of the fGES output to the Markov blanket of t as judged by fGES.

As with constraint-based Markov blanket procedures, fGES-MB can be modified to search arbitrarily far from the target and can trim the graph in different ways, either to the Markov blanket, or the adjacents and adjacents of adjacents, or more simply, just to the parents and children of the target. One only needs to construct an appropriate set of adjacencies over which fGES should search and then run the rest of the steps of the fGES algorithm restricted to those adjacencies. Even for several targets taken together, the running time of fGES-MB over the appropriate adjacency set will generally be much less than the running time of the full fGES procedure over all variables and can usually be carried out on a laptop even for very large datasets.

To illustrate that fGES-MB can run on large data sets on a laptop, Tables 4 and 5 show running times and average counts for fGES-MB for random targets selected from random models with 1000 nodes and 1000 or 2000 edges, 30,000 nodes with 30,000 or 60,000 edges, for continuous and discrete simulations, in the same style as for Tables 1 and 2. For the continuous case, a simulation with 1,000,000 nodes and 1,000,000 edges is included as well. Because the number of variables in the Markov Blanket varies considerably, accuracies are given in average absolute numbers of edges rather than in percentages.

Table 4 Average accuracy and time results for fGES-MB run on continuous data in simulation

# Nodes	# Edges	# Rep	# Correct adj	# Correct arr	# FP adj	# FN adj	Elapsed
1000	1000	100	3.1	3.1	0.0	0.2	99 ms
1000	2000	100	7.5	7.1	0.5	2.4	3.7 s
30,000	30,000	10	5.3	4.7	0.0	0.3	418 ms
30,000	60,000	10	5.4	5.4	0.0	1.8	13.2 s
1,000,000	1,000,000	1	2.0	2.0	0.0	1.0	42.9 s

Table 5 Accuracy and time results for fGES-MB run on discrete data in simulation

# Nodes	# Edges	# Rep	# Correct adj	# Correct arr	# FP adj	# FN adj	Elapsed
1000	1000	100	3.1	3.1	0.0	0.2	99 ms
1000	2000	100	3.0	2.3	0.1	6.3	65 ms
30,000	30,000	10	2.2	2.0	0.0	0.6	1.5 s
30,000	60,000	10	5.3	4.3	0.0	6.8	3.2 min

For each row of the table, a single DAG was randomly generated and parameterized as a linear, Gaussian structural equation model as described in the text. Then random nodes in the model were selected and their Markov blanket MECs assessed using the fGES-MB algorithm, using the penalized BIC score with BIC penalty multiplied by 4; these were then compared to the restriction of the true MEC of the generating DAG, restricted to the Markov blanket variables. Average number of correct adjacencies (# Correct Adj), correctly oriented edges (# Correct Arr), number of false positive adjacencies (# FP Adj) and number of false negative adjacencies (# FN Adj) are reported. Times are shown in milliseconds (ms) and seconds (s). The average number of misdirected edges for each row of Table 4 is given by the difference of the Correct Adjacency value and the Correct Arrowhead values.

For each row of the table, a single DAG was generated and parameterized as a Bayes Net with multinomial distributed categorical variables each taking 3 possible values. Conditional probabilities were randomly selected from a uniform [0, 1] distribution for each cell and the rows in the data table normalized. Then random nodes in the model were selected and their Markov blanket MECs assessed using the fGES-MB algorithm, using the penalized BIC score with BIC penalty multiplied by 4; these were then compared to the restriction of the true MEC of the generating DAG, restricted to the Markov blanket variables. Average number of correct adjacencies (# Correct Adj), correctly oriented edges (# Correct Arr), number of false positive adjacencies (# FP Adj) and number of false negative adjacencies (# FN Adj) are reported. Altogether for each row in Table 5, 100 nodes were used to calculate the averages shown. Times are shown in milliseconds (ms) and seconds (s).

7 A voxel level causal model of the resting state human cortex

To provide an empirical illustration of the algorithm, we applied fGES to all of the cortical voxels in a resting state

Table 6 Adjacencies retained at decreasing BIC penalty from resting state fGES searches over all cortical voxels

Penalty	# Adjacencies	% in 40	% in 20	% in 10	% in 4
40	6013	1.00	0.99	0.96	0.93
20	16,722		1.00	0.95	0.91
10	40,982			1.00	0.90
4	127,533				1.00

Table 7 Directed edges retained at decreasing BIC penalty from resting state fGES searches over all cortical voxels

Penalty	# Directed edges	% in 40	% in 20	% in 10	% in 4
40	5850	1.00	0.99	0.96	0.93
20	16,453		1.00	0.95	0.91
10	39,999			1.00	0.89
4	127,281				1.00

fMRI scan (approximately 51,000 voxels). Details of measurements and data preparation are given in the supplement. Because of the density of cortical connections, we applied the algorithm at penalties 40, 20, 10, and 4. Lower penalty analyses were not computationally feasible. All runs were done on the Pittsburgh Supercomputing Center (PSC) Bridges computer, using 24 nodes. Runtime for the penalty 4 fGES search was 14 h.

Tables 6 and 7 show the percentage of adjacencies and directed edges retained between each penalty run and the absolute numbers of adjacencies and directed edges returned at each penalty. Note that the great majority of adjacencies are directed.

For the penalty 4 run, the distribution of path lengths (in voxel units) between voxels (not counting zero lengths) is almost Gaussian (Fig. 2). The distribution of total degrees is inverse exponential as expected (Fig. 3). Figure 4 illustrates the parents and children of a single voxel.

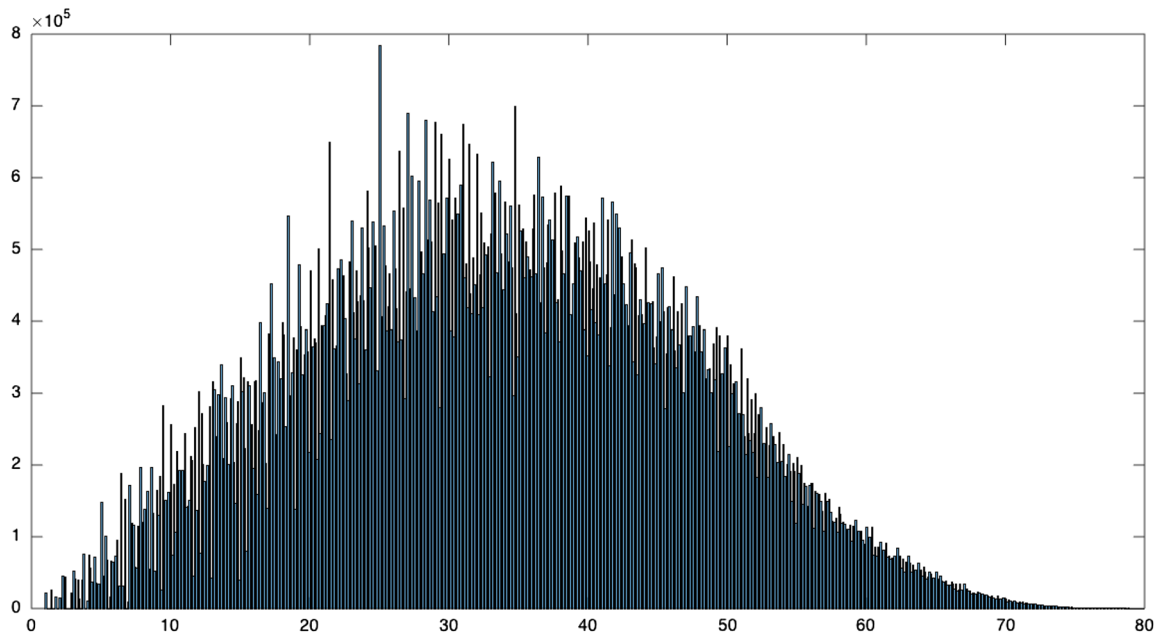


Fig. 2 Distribution of path distances (in voxel units) in penalty 4 run

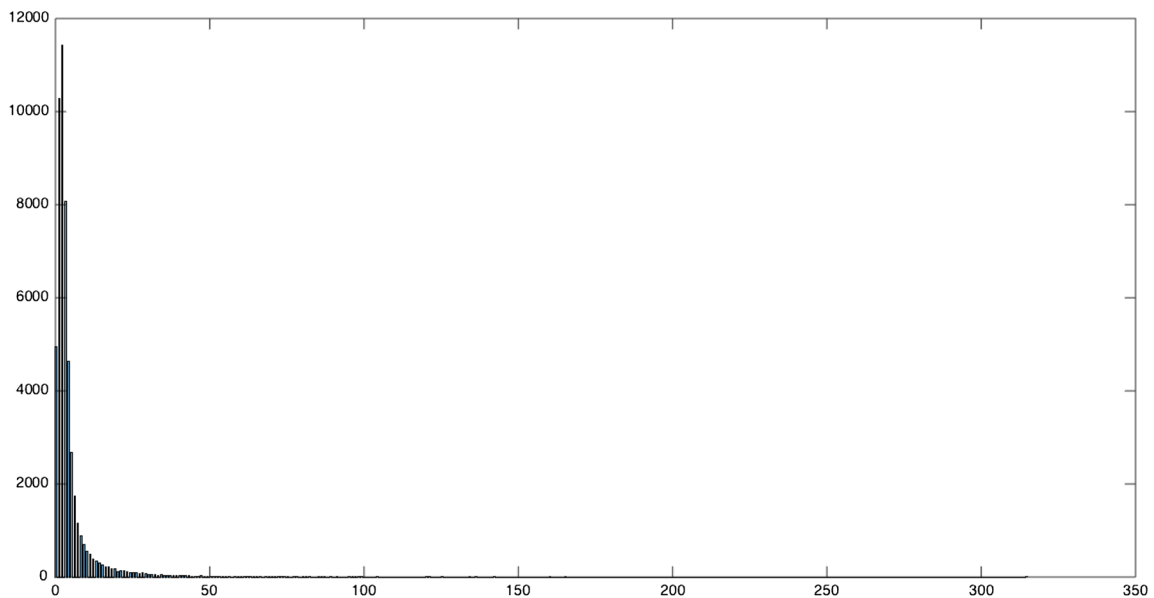


Fig. 3 Histogram of total degrees in penalty 4 run

8 Discussion

For the sparse models depicted in Tables 1 and 2, for fGES, we see high precisions throughout for both adjacencies and arrowheads, although for the million node case with continuous models precision suffers, and for the discrete models direction of edge precision suffers for the denser models. For discrete models, sample sizes are on the low side for the conditional probability tables we have estimated, and it is not surprising that recall is comparably low. If we

had not assumed the weak faithfulness condition, run times would have been considerably longer; for small models, recall would have been higher, though for large models our experience suggests they would have been approximately the same.

Tables 3 and 4, for fGES-MB, show excellent and fast recovery of Markov blankets for the sparse continuous case, less recall for the denser continuous case, and less recall than that for the discrete cases. The runtime for fGES-MB, for single targets, is a fraction of the runtime for fGES for

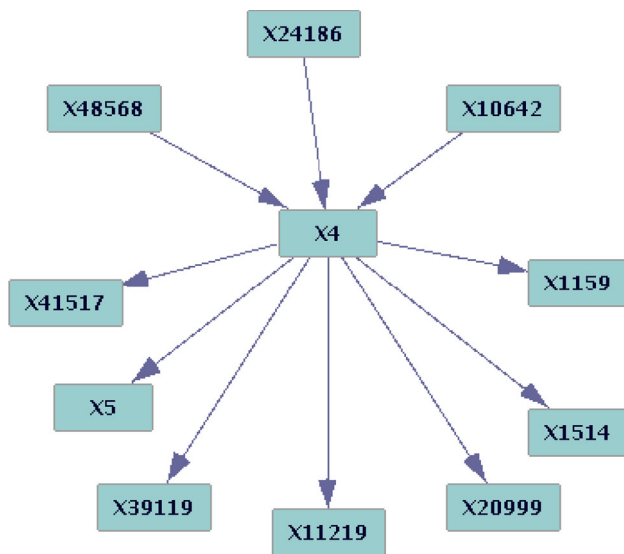


Fig. 4 Parents and children of a voxel

all variables on a large data set—for the continuous, sparse, million node case, it is reduced from 11 h to 42.9 s for single runs of those algorithms. If only the structure around a target node is needed, or structure along a path, it makes more sense to use fGES-MB or some simple variant of it than to estimate the entire graph and extract the portion of it that is needed. If a great deal of the structure is needed, it may make sense to estimate the entire graph.

Biological systems tend to be scale-free, implying that there are nodes of very high degree. The complexity of the search increases with the number of parents, and recall accuracy decreases. fGES deals better with systems, as in scale-free structures, in which high-degree nodes are parents (or “hubs”) rather than children of their adjacent nodes. If prior knowledge limits the number of parents of a variable, fGES can deal with high-degree nodes by limiting the number of edges oriented as causes of a given node.

These simulations do not address cases in which both categorical and continuous variables occur in datasets. Sedgewick et al. have proposed in such cases to first estimate an undirected graph using a pseudo-likelihood and then pruning and directing edges [15]. It seems worth investigating whether such a mixed strategy proves more accurate than discretizing all variables and using fGES; it would be better still to have a suitable likelihood function for DAG models with mixed variable types.

The fMRI application only illustrates possibilities. In practice, fMRI measurements, even in the same subject, may shift voxel locations by millimeters. Voxel identification across scans is therefore unreliable, which at the voxel level of resolution means that quite different DAGs will be recovered from different scans. Usually in fMRI analyses, clusters of several hundred voxels (regions of interest) are

formed based on anatomy or correlations using any one of the great many clustering algorithms, and connections are estimated from correlations of the average BOLD signals from each cluster. fGES at the voxel level offers the prospect of using causal connections among voxels to build supervoxel regions whose estimated effective connections are stable across scans, a strategy we are currently investigating.

Neural signaling connections in the brain have both feedback and unrecorded common causes that are not captured by any present methods at high dimensions, although low dimensional algorithms have been developed for these problems [5]. It seems important to investigate the possibility of scaling up and/or modifying these algorithms for high-dimensional problems, in part through parallelization, and investigating their accuracies.

Acknowledgements We thank Gregory Cooper for helpful advice and Russell Poldrack for data. Research reported in this publication was supported by grant U54HG008540 awarded by the National Human Genome Research Institute through funds provided by the trans-NIH Big Data to Knowledge (BD2K) initiative. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

- Hastie, T., Tibshirani, R., Friedman, J., Franklin, J.: The elements of statistical learning: data mining, inference and prediction. *Math. Intell.* **27**(2), 83–85 (2005)
- Friedman, J., Hastie, T., Tibshirani, R.: Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9**(3), 432–441 (2008)
- Schwarz, G.: Estimating the dimension of a model. *Ann. Stat.* **6**(2), 461–464 (1978)
- Chickering, D.M.: Optimal structure identification with greedy search. *J. Mach. Learn. Res.* **3**, 507–554 (2003)
- Spirites, P., Glymour, C.N., Scheines, R.: *Causation, Prediction, and Search*, 2nd edn. MIT Press, Cambridge (2000)
- Chickering, D.M., Heckerman, D., Meek, C.: Large-sample learning of Bayesian networks is NP-hard. *J. Mach. Learn. Res.* **5**, 1287–1330 (2004)
- Smith, S.M., Miller, K.L., Salimi-Khorshidi, G., Webster, M., Beckmann, C.F., Nichols, T.E., et al.: Network modelling methods for FMRI. *Neuroimage* **54**(2), 875–891 (2011)
- Aliferis, C.F., Tsamardinos, I., Statnikov, A.: HITON: a novel Markov Blanket algorithm for optimal variable selection. *AMIA Annu. Symp. Proc.* **2003**, 21 (2003)
- Aliferis, C.F., Statnikov, A., Tsamardinos, I., Mani, S., Koutsoukos, X.D.: Local causal and markov blanket induction for causal discovery and feature selection for classification part I: algorithms and empirical evaluation. *J. Mach. Learn. Res.* **11**, 171–234 (2010)
- Aliferis, C.F., Statnikov, A., Tsamardinos, I., Mani, S., Koutsoukos, X.D.: Local causal and markov blanket induction for causal discovery and feature selection for classification part II: analysis and extensions. *J. Mach. Learn. Res.* **11**, 235–284 (2010)
- Chickering, D.M.: Learning equivalence classes of Bayesian network structures. *J. Mach. Learn. Res.* **2**, 445–498 (2002)

12. Meek, C.: Causal inference and causal explanation with background knowledge. Proc. Eleventh Conf. Uncertain. Artif. Intell. **1995**, 403–410 (1995)
13. Pearl, J.: Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference, 1st edn. Morgan Kaufmann Publishers, San Francisco (1988)
14. Ramsey, J.D., Hanson, S.J., Hanson, C., Halchenko, Y.O., Poldrack, R.A., Glymour, C.: Six problems for causal inference from fMRI. *Neuroimage* **49**(2), 1545–1558 (2010)
15. Sedgewick, A.J., Shi, I., Donovan, R.M., Benos, P.V.: Learning mixed graphical models with separate sparsity parameters and stability-based model selection. *BMC Bioinform.* **17**(Suppl 5), 175 (2016)
16. Tsamardinos, I., Aliferis, C.F., Statnikov, A.R., Statnikov, E.: Algorithms for large scale Markov Blanket discovery. In: Proceedings of International Florida Artificial Intelligence Research Society Conference, vol. 2 (2003)