Fast Greedy Search (FGES) Algorithm for Continuous Variables

This document provides a brief overview of the FGES algorithm, focusing on a version of FGES that works with continuous variables called FGES-continuous (FGESc).

Purpose

FGESc is an algorithm that takes as input a dataset of continuous variables, greedily searches over selected causal Bayesian network (CBN) structures[1] (models), and outputs the most probable model it finds. The model FGESc returns serves as a data-supported hypothesis about causal relationships that exist among the variables in the dataset. The model is intended to help scientists form hypotheses and guide the design of controlled experiments to investigate these hypotheses.

Methodological Approach

FGES is an optimized and parallelized version of an algorithm developed by Meek [Meek, 1997] called the Greedy Equivalence Search (GES). The algorithm was further developed and studied by Chickering [Chickering, 2002]. GES is a Bayesian algorithm that heuristically searches the space of CBNs and returns the model with highest Bayesian score it finds. In particular, GES starts its search with the empty graph. It then performs a forward stepping search in which edges are added between nodes in order to increase the Bayesian score. This process continues until no single edge addition increases the score. Finally, it performs a backward stepping search that removes edges until no single edge removal can increase the score.

FGESc uses the Bayesian Information Criterion (BIC) [Raftery, 1995] to score models, which approximates the marginal likelihood of the data given a graph structure *M*: *P*(*data* | *M*). More precisely, it approximates the natural logarithm of the marginal likelihood. In particular, BIC is defined as follows:[2]

$$BIC = 2 \cdot \ln P(data|\hat{\theta}, M) - c \cdot k \cdot \ln(n),$$

where *M* denotes a CBN structure, $\hat{\theta}$ denotes the value of the CBN parameters (i.e., the probabilities) that maximize the data, *k* are the number of parameters in the CBN, *n* is the number of samples (cases), and *c* is a constant that is 1 in the traditional definition of BIC, but which we allow to be greater than one, as discussed below.

Input Data and Parameters

FGESc has the following requirements for data input:

---

[1] A CBN structure is a directed acyclic graph in which nodes represent variables and arcs represent direct causation among the nodes, where the meaning of *direct* is relative to the nodes in the CBN. For further information about CBNs, see [Spirtes, 2010; Lagani, 2016; Pearl 2016].

[2] This equation is actually the arithmetic negation of the traditional BIC. In this equation, the higher the score, the more likely is the model *M*.

- the (training) data are in a table in which columns represent variables, rows represent samples, and the value of each variable in a sample is continuous (i.e., a real number).
- the first row of the table lists the variable names, in order and unique; the data and variable names are separated by a delimiter (default: tab).
- there are no missing values in the table.
- there are no linear dependencies in the data. The user has the option to turn off the linear dependency check.
- there are no variables that have zero variance. The user has the option to turn off the zero variance checks.

FGESs takes the following parameters, which modify the behavior of the algorithm:

- penalty-discount - the specification of a complexity penalty parameter *c* that is shown in the BIC equation above (FGESc uses *c* = 4 by default).
- search depth - specifies the maximum number of edges to orient into a node during a single orientation step (FGESc uses a default search depth of -1, which indicates unlimited depth). Small search depth values will reduce search time.
- heuristic-speedup – apply a heuristic speedup that does not assume faithfulness condition; this condition is discussed briefly below (FGESc by default applies this heuristic speedup).
- prior knowledge - the user may specify knowledge by providing a file that describes precedence and required and/or forbidden edges in the structure. By default, the algorithm assumes no prior knowledge about the causal graph structure
- thread – by default the algorithm will run in a parallel fashion using as many threads as are needed and available on the system. The user has the option to specify a smaller number of threads.
- graph_ml - the user has the option to output the graph in standard graphml format. By default the program will output a text file describing the causal graph and the search path.

Output

FGESc outputs the most probable CBN structure it finds, according to the BIC measure. More precisely, it outputs a "pattern"[3] [Chickering, 2002] containing arcs (→), which represent direct causation, and undirected edges (—), where such an edge indicates there is a causal arc, but its direction cannot be determined.

Algorithmic Assumptions

This section describes a sufficient set of assumptions for the application of FGESc to achieve the guarantees described in the next section. While the pattern output by FGESc may still include correct edges (and perhaps many correct edges) even if one or more of these assumptions are violated, there are no theoretical guarantees it will do so.

A sufficient set of conditions for recovering the causal structure of the data-generating process in the large sample limit (i.e., as the sample size grows without bound) is as follows. Assume that the causal process generating the data *D* given to FGESc is accurately modeled by a CBN

[3] Patterns are also known as PDAGs, essential graphs, and maximally oriented graphs.

containing only measured variables, which we call ***G***. Assume that each variable (node) in ***G*** is a linear function of its parents, plus a finite additive Gaussian noise term. Finally, assume that each case in *D* was obtained by randomly sampling all the variables in ***G*** from the joint distribution defined by ***G***.

While the above procedure is simple, it includes several assumptions that may not be immediately obvious. Key among them are:

- cases (samples) in the data *D* are independent and identically distributed.
- the data are being generated by processes that are accurately modeled as linear relationships among the measured variables with finite Gaussian noise.
- the causal Markov condition holds [Spirtes, 2010]. This condition states that a variable is independent of its non-effects, given its direct causes (parents). It expresses a form of local causality.
- the causal faithfulness condition holds [Spirtes, 2010]. This condition states that all the independence relationships among the measured variables are implied by the causal Markov condition.
- there are no missing data. The user must fill in missing data before running FGESc. Many statistical packages provide methods for handling missing values, including imputing them.
- there are no hidden confounders of the measured variables. That is, none of the measured variables have a common hidden variable as one of their direct causes (relative to the variable set). This is perhaps the strongest assumption of FGESc, because hidden confounders are typically replete in scientific data. Nonetheless, the output of FGESc may still provide helpful clues about the causal relationships among the measured variables. The Tetrad system [Tetrad, 2016] currently contains several algorithms that model latent confounders, including FCI and RFCI. The Center for Causal Discovery plans to release a relatively efficient algorithm called GFCI that models hidden confounders and uses FGES as an initial step.
- there is no selection bias. This means that the chance a case (sample) was selected from the population for inclusion in dataset *D* did not depend on the values of any of the measured variables in the data.
- there are no feedback cycles among the measured variables. Extensions to CBNs, such as causal Dynamic Bayesian Networks (DBNs) [Neapolitan, 2003], do allow feedback cycles, but they are not currently implemented in FGESc.

Structure Learning Performance Guarantees

If the assumptions in the previous section hold, then in the large sample limit, the CBN structure output by FGESc will contain (1) an arc X → Y if and only if X causes Y; (2) an edge (—) if and only if either X causes Y or Y causes X; and (3) no edge between X and Y if and only if X and Y have no direct causal relationship between them.

Practice Dataset

Here we apply FGESc to a very small dataset to illustrate its use. The user may wish to apply FGESc to this dataset and verify that the CBN structure obtained is the same as the one shown

in Figure 1. This example dataset will provide the results shown using the default parameters of FGESc.

Table 1. Simple practice dataset

X	Y	Z
1.3997	4.6118	-1.5362
-0.7004	-4.0433	1.3182
-0.5749	2.0839	-1.6841
1.9082	6.0967	-2.4738
-2.1138	-1.9683	-1.3704
-1.4251	-0.5521	-0.5197
-1.5074	-2.4464	0.4334
0.2097	1.1319	-1.3497
-0.3749	-0.3251	-0.9903
-1.5654	-0.8384	-0.5516
0.565	1.9158	-0.3094
-1.3864	-4.1625	1.2169
-0.087	-0.2636	0.9364
-0.7332	-2.8329	1.2629
-0.2929	-0.8709	0.4784
2.089	2.9733	-1.1307
2.1261	5.3042	-1.7263
0.7871	-1.2154	1.4612
0.6513	3.5145	-0.9141
3.1871	6.0871	-0.7839

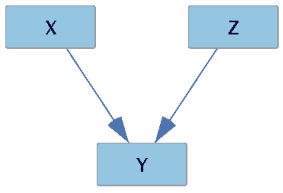

Figure 1. The CBN structure output by FGESc when given the data shown in Table 1.

Performance on Simulated Data

We evaluated the performance of FGESc on simulated data. We first created a random CBN with a given number of nodes and edges, which we call CBN_{gen}. We then randomly sampled the

distribution defined by CBN_{gen} to generate a set of training data *D consisting of 1000 samples*. We provided that data to FGESc to obtain the CBN structure that it output, which we call CBN_{out}. Both CBN_{gen} and CBN_{out} are patterns. We compared CBN_{gen} with CBN_{out} to derive arc precision and arc recall. Arc recall is the fraction of arcs in CBN_{gen} that appear in CBN_{out}. Arc precision is the fraction of arcs in CBN_{out} that appear in CBN_{gen}. We also recorded the CPU time in minutes that FGCc required to derive CBN_{out} when using a given number of processors on a machine at the Pittsburgh Supercomputing Center. The precision, recall, and timing results are averages over multiple repetitions of the process just described, as shown in Table 2 below.

Table 2. Performance results for FGESc on simulated data.

# Nodes	# Edges	# Repetitions	Average Arc Precision	Average Arc Recall	# Processors	Average Learning Time (minutes)
1,000	1,000	10	99.9	94.9	3	0.02
1,000	2,000	10	98.1	87.7	6	0.15
10,000	10,000	10	99.9	94.7	120	0.21
10,000	20,000	10	99.7	86.8	120	1.39
100,000	100,000	10	99.3	94.6	120	7.34
100,000	200,000	10	99.7	86.7	120	17.40

The results in Table 2 provide benchmarks that may be helpful in estimating the performance of FGESc when it is applied to real datasets. We emphasize, however, that the recall and precision results obtained with such simulated data may be higher than those obtained with real datasets.

References

Chickering DM. Optimal structure identification with greedy search. *Journal of Machine Learning Research* 3 (2002) 507-554.
http://www.jmlr.org/papers/volume3/chickering02b/chickering02b.pdf

Lagani V, Triantafillou S, Ball G, Tegner J, Tsamardinos I. Probabilistic computational causal discovery for systems biology. *Uncertainty in Biology* 17 (2016) 33-73.
http://www.mensxmachina.org/files/publications/Probabilistic%20Causal%20Discovery%20for%20Systems%20Biology_prePrint.pdf

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

Neapolitan RE. *Learning Bayesian Networks* (Pearson, 2003).

Pearl J, Glymour M, Jewell NP. *Causal Inference in Statistics – A Primer* (John Wiley & Sons, 2016).
https://books.google.com/books?hl=en&lr=&id=IqCECwAAQBAJ&oi=fnd&pg=PT1&dq=Causal+Inference+in+Statistics&ots=NPpnh1N4lC&sig=-CyGyDAsTQP1vFstnAZh3dt-Ih8#v=onepage&q=Causal%20Inference%20in%20Statistics&f=false

Ramsey J. Scaling up Greedy Equivalence Search for continuous variables (2015). http://arxiv.org/ftp/arxiv/papers/1507/1507.07749.pdf

Raftery AE. Bayesian model selection in social research. *Sociological Methodology* 25 (1995) 111-163. https://www.stat.washington.edu/raftery/Research/PDF/socmeth1995.pdf

Spirtes P. Introduction to causal inference. *Journal of Machine Learning Research* 11 (2010) 1643-1662. http://jmlr.org/papers/volume11/spirtes10a/spirtes10a.pdf

Tetrad system (2016). http://www.phil.cmu.edu/tetrad/current.html