A Hybrid Method for Learning Edge-Minimal Gaussian I-maps

Oliver Schulte and Gustavo Frigo
School of Computing Science
Simon Fraser University
Vancouver-Burnaby V5A 1S6
Canada

Russell Greiner
Department of Computing Science
University of Alberta
Edmonton, AB T6G 2E8
Canada

Abstract

This paper presents a hybrid algorithm for structure learning in linear Gaussian models whose structure is a directed graph. The algorithm performs a local search for a model that meets the following criteria: (1) The Markov blankets in the model should be consistent with dependency information from statistical tests. (2) Minimize the number of edges subject to the first constraint. (3) Maximize a given score function subject to the first two constraints. Our local search is based on Graph Equivalence Search (GES); we also apply the recently developed SIN statistical testing strategy to help avoid terminating the search in a local minimum. Simulation studies with GES search and the BIC score provide evidence that for nets with about 10 or more variables, the hybrid method selects simpler graphs whose structure is closer to the target graph.

1 Introduction

Bayes nets [18] are a widely used formalism for representing and reasoning with uncertain knowledge. A Bayes net (BN) model is a directed acyclic graph (DAG) $G = \langle \mathbf{V}, \mathbf{E} \rangle$ whose nodes $\mathbf{V}$ represent random variables and whose edges $\mathbf{E}$ represent statistical dependencies, together with conditional probability tables that specify the distribution of a child variable given an instantiation of its parents. In this paper we consider Gaussian Bayes networks with the following properties: (1) all variables are continuous, (2) a child variable is a linear function of its parent variables plus a Gaussian error term, (3) all error terms are independent. In econometrics, such models are called recursive structural equation models (SEMs). SEM models are widely employed in economics, psychology, sociology and genetics [14],[17, Ch 4.1], [19, Ch.5].

There are two well established general approaches to learning BN structure. Constraint-based (CB) methods employ a statistical test to detect conditional (in)dependencies given a sample d, and then compute a BN G that fits the (in)dependencies [23]. Score-based methods search for models that maximize a model selection score [13]. Hybrid methods aim to combine the strengths of both approaches [24, 8, 12]. Evaluations have shown that for DAGs with *discrete* variables, the best hybrid methods outperform both purely score-based and purely constraint-based methods [24]. We introduce a new hybrid model selection criterion and develop a novel search strategy for the criterion that integrates statistical tests and score functions. Our new criterion combines constraints and score functions as follows: (1) A DAG G should satisfy the *Markov boundary condition*, meaning that for any two nodes X and Y, no statistically significant correlation is found between X and Y given the neighbors and spouses of X. (2) The model G should have the minimum number of edges among the graphs that satisfy the boundary condition. (3) Among the minimum-edge graphs satisfying the boundary condition, our criterion selects the ones that maximize a given score.

Motivation

There is theoretical, statistical and computational motivation for this composite selection criterion. A BN model that represents the target or operating distribution generating the data must satisfy the Markov boundary condition. It is widely considered than an acceptable graphical model G of the target distribution should be edge-minimal, meaning that no subgraph of G represents the target distribution [18, Ch.3.3], [17, Ch.2.4]. Minimizing the number of edges implies edge-minimality. [21] provides a learning-theoretic justification for minimizing the number of edges as a small-sample selection criterion. *Statistical motivation* is provided by the observation that standard model selection criteria like the Bayes Information Criterion (BIC; [17, Ch.8.3.2]) tend to favor overly complex models when applied to linear models [20]. We give further empirical evidence to support this finding.

One reason why standard model scores tend to overfit in continuous domains, but not with discrete variables, is that the penalty term for model complexity in the score is generally a function of the number of parameters in the model; in continuous-variable

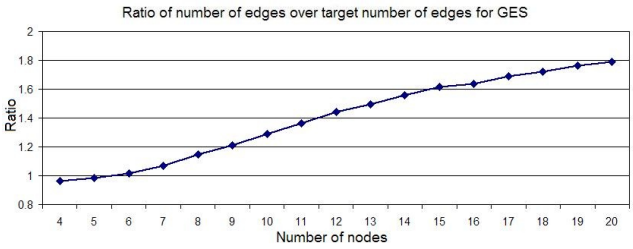

Figure 1: The overfitting factor of the GES algorithm. The figure shows the number of edges returned by GES over the average number of edges in the target DAG. a ratio of 1 is ideal indicating that GES has the same number of edges as the target DAG. Values below 1 suggest underfitting, and values over 1 indicate overfitting. The trent in the upper right shows a clear tendency that as the DAG becomes larges, so does the overfitting factor. The figures also seems to suggest that for the larger graphs, the overfitting factor grows with the sample size.

models with linear dependencies of a child variable on its parents, the number of parameters is linear in the number of nodes, whereas in discrete-variable models it is typically exponential. Our composite criterion addresses overfitting by assigning higher priority to minimizing the number of edges than to maximizing the score. Thus the criterion favors adding an edge only if this is necessary for representing a statistically significant correlation found in the data, even if adding the edge improves the model selection score. A *computational motivation* for adding the model selection score is that the problem of finding minimum-edge graphs consistent with a set of given dependencies is NP-hard [2, Lm. 4.5]; the score serves as a heuristic for exploring the search space.

Overview of Hybrid Search Method.

The goal of our search method is to combine the strengths of both score-based and constrained-based learning so as to ameliorate the weaknesses of each. Our general approach is to treat the information from statistical tests as a constraint on the model selection search that effectively reduces the search space. The main issue with constraint-based methods is their sensitivity to type II error, that is, false acceptances of the independence null hypothesis, which leads them to falsely remove links [11, 12, 26]. We address this problem with the approach of [**?**], which is to *rely on rejection of the null hypothesis as indicating dependencies, but draw no conclusion from failure to reject.* The key idea in our hybrid search system is to constrain the addition of edges: given a current candidate graph G, a local search method may add an edge only if this leads to a graph G' that entails a statistically significant dependency `d` that is not entailed in G. Thus the local search is prevented from fitting statistically insignificant correlations even if this would lead to a higher score. We provide a general schema for adapting *any* hill-climbing search algorithm with a given score function for constrained search. The adapted algorithm can be seen as a forward-backward selection strategy for discovering a minimal Markov boundary. Our method also integrates one of the most recent CB algorithms, the "condition on nothing and everything else" strategy of SIN graphical model selection [9]: For any two variables X and Y, test the unconditional correlation betwen X and Y and the correlation conditional on all other variables (i.e., $\mathbf{V} - \{X, Y\}$).

Empirical Evaluation For experimental evaluation, we adapted the state-of-the-art Graph Equivalence Search (GES) procedure [16, 4]. We report a number of measurements comparing GES and our constrained GES, based on the well-established BIC score function. Simulation results for both randomly generated and real-world target BN structures compare the graphs learned with and without (in)dependency constraints to the target graph. For node sizes of 10 and greater, we observe that BIC significantly overfits the data in the sense that it produces graphs with too many adjacencies. Our experiments illustrate how adding (in)dependency constraints corrects some of this overfitting tendency of the BIC score function. The constrained search produces simpler models (i.e., with fewer adjacencies) whose graphs whose structure is closer to the target graph, as measured by the standard Hamming distance.

The paper is organized as follows. The next section reviews basic notions from Bayes net theory. Section 3 discusses the major design choices in our system, including our adaptation of GES search. Section 4 presents simulation studies that compare constrained GES search with the BIC score to regular GES search with the same score.

Contributions

Key novel features of our algorithm include the following. (1) To our knowledge, the first development and evaluation of a hybrid structure learning algorithm for continuous-variable Bayes nets. (2) Making use of dependency constraints primarily rather than independency constraints. (3) A new approach for selecting informative statistical hypotheses to test based on the Markov boundary.

Related Work *Score-based Methods.* A number of score functions are widely used in structural equation modelling, such as AIC and model chi-square [14]. We focused our study on the BIC information criterion, for several reasons. (1) BIC is one of the best established in the SEM literature. (2) BIC is widely used for evaluating Bayes nets in computer science studies [8, 25]; it is the default score for Gaussian models in the Tetrad system. (3) Other standard criteria like AIC penalize complex structures less than BIC so the overfitting tendency of BIC corrected by our algorithm is even stronger with these criteria.

Hybrid Methods. A recent hybrid method (max-min hill climbing) that treats the tests of statistical outcomes as constraints is presented in [24]. While this work indicates that independence constraints from a statistical test can improve a score-based search, the analysis of [12] shows that because it accepts independence null hypotheses, max-min hill climbing is sensitive to type II errors. The method of [**?**] is similar to ours in that it treats only dependencies (rejections of the null hypothesis) as "hard" constraints. However, [**?**] addressed the problem of *underfitting* in score-based BN learning with discrete variables, whereas the problem in BN learning in Gaussian models is overfitting. Specifically, [**?**] requires a local search method to add more adjacencies until all statistically significant correlations are entailed by the graph, whereas the method of this paper constrains the search method to add fewer adjacencies. Other previous hybrid BN learning algorithms (e.g., [8, 11]) consider statistical measures (*e.g.*, mutual information), but do not incorporate the outcome of a statistical test as a constraint that the learned model must satisfy. To our knowledge, the hybrid methods whose description and evaluation have been published to date deal with discrete variables rather than continuous ones. Our algorithm can be seen as a hybrid version of the Grow-Shrink procedure [15]. The main difference is that Grow-Shrink relies on a fixed ordering of variables to select the next candidate structure and the next statistical hypothesis to test. Our method employs the score function to select the next candidate structure.

2 Basic Definitions

The definition and theorems cited in this section are standard; for further details see [17, 18, 23]. We consider Bayes nets for a set of random variables $\mathbf{V} = \{X_1, \ldots, X_n\}$ where each X_i is real-valued. A **Bayes net structure** $G = \langle \mathbf{V}, \mathbf{E} \rangle$ for a set of variables $\mathbf{V}$ is a directed acyclic graph (DAG) over node set $\mathbf{V}$. A Bayes net (BN) is a pair $\langle G, \theta_G \rangle$ where θ_G is a set of parameter values that specify the probability distributions of each variable conditioned on instantiations of its parents. A BN $\langle G, \theta_G \rangle$ defines a p.d.f over $\mathbf{V}$. In a linear Gaussian BN, each child Y is a linear function of its parents $X_1, ..., X_k$ so $Y = \sum_{i=1}^{k} a_i X_i + \varepsilon_Y$, where the error term ε_Y has a normal distribution with mean 0. The variance of ε_Y and the coefficients a_i are parameters of the model. For a source node X, its mean and variance are further parameters of the model. We make the standard assumption that the error terms for different variables are uncorrelated. The BIC score is defined as $BIC(G, \mathsf{d}) = L(\hat{G}, \mathsf{d}) - par(G)/2 \times ln(m)$ where $\hat{G}$ is the BN G with its parameters instantiated to be the maximum likelihood estimates given the sample d, the quantity $L(\hat{G}, \mathsf{d})$ is the log-likelihood of $\hat{G}$ on the sample d, the sample size is denoted by m, and $par(G)$ is the number of free parameters in the structure G.

Two nodes X, Y are **adjacent** in a BN if G contains an edge $X \rightarrow Y$ or $Y \rightarrow X$; an adjacency is a pair of adjacent nodes. An **unshielded collider** in G is a triple of nodes connected as $X \rightarrow Y \leftarrow Z$, where X and Z are not adjacent. The **pattern** $\pi(G)$ of DAG G is the partially directed graph over $\mathbf{V}$ that has the same adjacencies as G, and contains an arrowhead $X \rightarrow Y$ if and only if G contains an unshielded collider $X \rightarrow Y \leftarrow Z$.Every BN structure defines a separability relation between nodes X, Y relative to a set of nodes $\mathbf{S}$, called **d-separation** [18, Ch.3.3]. We assume familiarity with d-separation.

We write $(X \perp\!\!\!\perp Y | \mathbf{S})_G$ if X and Y are d-separated by $\mathbf{S}$ in graph G. If two nodes X and Y are not d-separated by $\mathbf{S}$ in graph G, then X and Y are **d-connected** by $\mathbf{S}$ in G, written $(X \not\perp\!\!\!\perp Y | \mathbf{S})_G$. We write $\mathcal{D}(G)$ for the set of all d-connections $(X \not\perp\!\!\!\perp Y | \mathbf{S})_G$ that hold in a graph G. Two DAGs G and G' satisfy exactly the same dependencies iff they have the same patterns (*i.e.*, $\mathcal{D}(G) = \mathcal{D}(G')$ iff $\pi(G) = \pi(G')$ [17, Th.2.4]). We take the set of dependencies associated with a pattern π to be the set of dependencies in any DAG G whose pattern is π.

Let ρ be a joint probability density function (p.d.f) for variables $\mathbf{V}$. If $\mathbf{X}, \mathbf{Y}$ and $\mathbf{Z}$ are three disjoint sets of variables, then $\mathbf{X}$ and $\mathbf{Y}$ are **stochastically independent given** $\mathbf{S}$, denoted by $(\mathbf{X} \perp\!\!\!\perp \mathbf{Y} | \mathbf{S})_\rho$, if $\rho(\mathbf{X}, \mathbf{Y} | \mathbf{S}) = \rho(\mathbf{X} | \mathbf{S})\, \rho(\mathbf{Y} | \mathbf{S})$ whenever $\rho(\mathbf{S}) > 0$. A BN structure G is an **I-map** of p.d.f. ρ if for any three disjoint sets of variables $\mathbf{X}, \mathbf{Y}$ and $\mathbf{Z}$ we have $(\mathbf{X} \perp\!\!\!\perp \mathbf{Y} | \mathbf{S})_G$ implies $(\mathbf{X} \perp\!\!\!\perp \mathbf{Y} | \mathbf{S})_\rho$. For a given BN structure G and joint density function ρ, there is a parametrization θ_G such that ρ is the joint density for $\mathbf{V}$ defined by $\langle G, \theta \rangle$ only if G is an I-map of P.

For a node X, we refer to the set of its parents, children and co-parents (*i.e.*, other parents of its children) as **the Markov blanket** of X in G, written $MB_G(X)$. If the graph G is fixed by context or irrelevant, we also simply write $MB(X)$. Given its Markov blanket $MB(X)$, each node X is d-separated from all other nodes outside of the Markov blanket. We refer to the set of independencies $\{X \perp\!\!\!\perp Y | MB(X) : Y \notin MB(X)\}$ as the **set of Markov blanket independencies** for a graph. If a graph G is an I-map of a joint density ρ, then all the Markov blanket independencies in G hold in ρ. As the characteristic feature of our approach is searching for a graph that satisfies this condition, we refer to it as "I-map learning". The next section describes an implementation of I-map learning.

3 Algorithm Design for I-map Learning

This section describes the major design choices in our system. We first discuss employing statistical tests for detecting conditional (in)dependencies, then integrating statistical testing with a score-based local search.

3.1 Use of Statistical Tests

I-map learning requires a statistical significance test for conditional independence hypotheses of the form $X \perp\!\!\!\perp Y|\mathbf{S}$. Our system architecture is modular, so the test can be chosen to suit the type of available data and application domain. We followed other CB methods and used Fisher's z-statistic for testing whether a given partial correlation is 0 [23, Ch.5.5].

For a given graph G, say that node Y is a *proper spouse* of node X if X and Y have a common child but are not adjacent. The set of *nonchildren* of X and Y are the nodes that are adjacent to X or Y but not children of either ; denote this set by $NC_G(X, Y)$. Our basic test selection strategy applies the chosen significance test to the following independence hypotheses, for each ordered pair of nodes (X, Y).

1. The **M**arkov blanket independencies $\{X \perp\!\!\!\perp Y|MB_G(X) : Y \notin MB(X)\}$.
2. The **s**pousal independencies $\{X \perp\!\!\!\perp Y|NC_G(X) : Y \in MB(X)\}$.

These independence tests are well-suited for pattern-based search since the Markov blanket and common children are determined by the pattern alone. The spousal independencies help to distinguish nodes on the Markov blanket that are both neighbors and spouses from nodes that are spouses only.

If a suitable test rejects a Markov blanket or a spousal independency hypothesis, this is evidence that the graph G is not correct. I-map learning implements the Markov blanket testing strategy through a procedure `find-new-dependencies(G)` that takes as input a new graph G adopted during the local search, tests the new Markov blanket and spousal hypotheses for the graph G, and returns the set of rejected independence hypotheses. Every time the local search moves to a new graph structure G, the procedure `find-new-dependencies` is applied to G to augment the cache of observed dependency constraints; see Figure 1. The procedure `find-new-dependencies` tests a set of independence hypotheses, so issues of multiple hypothesis testing arise. Our system architecture is modular, so any multiple hypothesis testing method can be employed to implement the functionality of `find-new-dependencies`, such as the methods described in [1, 9]. Many constraint-based and hybrid systems simply carry out multiple hypotheses at the same fixed significance level [23, 8, 15]. Our experiments follow this approach to facilitate comparisons with the competitor systems.

3.2 Heuristic Search Algorithm for I-map learning

For our experiments we adapt the GES (Greedy Equivalence Search) local search algorithm. GES is a state-of-the-art BN search strategy that satisfies optimality guarantees in the large sample limit and has been extensively evaluated [4]. Since our goal is to investigate whether adding dependency constraints improves the quality of learned models, we want to employ a high-quality score-based method such as GES. We describe GES only in sufficient detail to indicate how we adapt it. During its growth phase, GES moves from a current candidate pattern π to the highest-scoring pattern π' in the upper neighborhood $\mathsf{nbdh}^+(\pi)$. A pattern π' in $\mathsf{nbdh}^+(\pi)$ contains exactly one more adjacency than π, and may have arrows reversed, subject to several conditions that ensure that $\mathcal{D}(\pi) \subset \mathcal{D}(\pi')$, i.e., π' entails a strict superset of the dependencies entailed by π. The growth phase terminates with a pattern π when no graph in $\mathsf{nbdh}^+(\pi)$ has higher score than π. During the subsequent shrink phase, GES moves from a current candidate pattern π to the highest-scoring pattern π' in the lower neighborhood $\mathsf{nbdh}^-(\pi)$. A pattern π' in $\mathsf{nbdh}^-(\pi)$ contains exactly one less adjacency than π, and may have arrows reversed, subject to several conditions that ensure that $\mathcal{D}(\pi') \subset \mathcal{D}(\pi)$, i.e., π' entails is a strict subset of the dependencies entailed by π. GES terminates with a pattern π when no graph in $\mathsf{nbdh}^-(\pi)$ has higher score than π.

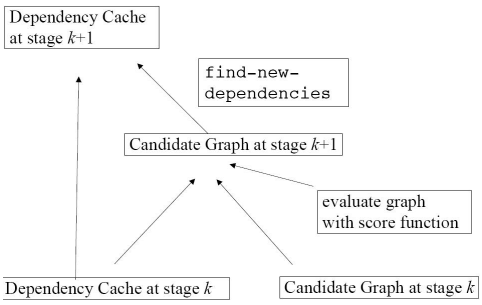

Figure 2: Integrating a local search for a score-maximizing graph structure with testing for statistically significant dependencies. Once a candidate structure G_k is chosen that maximizes the score function given the dependencies observed at stage k, the procedure `find-new-dependencies` applies the Markov blanket concept to test new independence hypotheses entailed by G_k, and adds rejected independence hypotheses to the global cache for stage $k+1$.

The constrained version IGES (for I-map + GES) constrains the GES neighborhoods so they satisfy a given set of observed dependencies. Formally, the *growth neighborhood constrained by dependencies* $\mathcal{D}$ is defined as follows:
$\pi' \in \mathsf{nbdh}^+_{\mathcal{D}}(\pi)$ iff $\pi' \in \mathsf{nbdh}^+(\pi)$ and $(\mathcal{D}(\pi') \cap \mathcal{D}) \supset (\mathcal{D}(\pi) \cap \mathcal{D})$.

So the growth phase keeps expanding a candidate structure to entail more of the observed dependencies $\mathcal{D}$, and terminates when all observed dependencies are covered. Note that the search may terminate even when a local operation (e.g., adding an edge) increases the score, if the local operation does not contribute to covering more statistically significant dependencies. This termination condition is appropriate when the score tends to overfit the data with overly complex

structures. To check if a graph expansion covers strictly more dependencies, we keep a cache of dependencies that have not yet been covered during the growth phase, and go through these dependencies in order to see if any of them are covered by a candidate graph. The *shrink neighborhood constrained by dependencies* $\mathcal{D}$ is defined as follows:
$\pi' \in \mathtt{nbdh}^-_{\mathcal{D}}(\pi)$ if and only if $\pi' \in \mathtt{nbdh}^-(\pi)$ and $(\mathcal{D}(\pi') \cap \mathcal{D}) \supseteq (\mathcal{D}(\pi) \cap \mathcal{D})$.

So the shrink phase moves to higher-scoring patterns in the GES lower neighborhood, subject to the constraint of fitting the observed dependencies, until a local score maximum is reached. Algorithm 1 gives pseudocode for IGES search. Our approach to constraining a local search with a given set of dependencies $\mathcal{D}$ applies to any hill-climbing search S that moves to the highest scoring graph in am S-neighborhood $\mathtt{nbdh}(G)$: First, modify the definition of $\mathtt{nbdh}$ to define a constrained growth neighborhood $\mathtt{nbdh}^+(G)$ and a constrained shrink-neighborhood $\mathtt{nbdh}^-_{\mathcal{D}}$. Then apply Algorithm 1 with the constrained S-neighborhoods; the result is a two-stage grow-shrink Markov blanket search based on the hill-climbing strategy S. This schema can be extended to beam search and other local search strategies more complex than hill climbing.

3.3 Analysis of Search Procedure

A score function is consistent if as the sample size increases indefinitely, with probability 1 all graphs that maximize the score are I-maps of the target distribution. The score function is decomposable if the score of a graph can be computed from scores for each node given its parents (for definitions of consistency and decomposability, see [17].) The standard analysis of CB methods assumes the correctness of the statistical tests, which holds in the sample size limit [6, 23]. Under these assumptions, our local search method is consistent.

Proposition 1 *Suppose that the statistical test returns only valid dependencies in target graph G during an execution of Algorithm 1 (with or without SIN testing), and that the score function is consistent and decomposable. Then as the sample size increases indefinitely, with probability 1, the algorithm terminates with an I-map π of the target distribution defined by G.*

Proof Outline. Let π be the final pattern in the growth phase of IGES. The correctness proof for the grow-shrink algorithm [15] can be adapted to show that the Markov blanket of each node X is identified correctly, in the sense that if Y is in $MB_G(X)$, then Y is in $MB_\pi(X)$. Then the only way π can fail to be an I-map is if it contains an undirected triple $X - \mathbf{Z} - Y$ that is oriented as $X \rightarrow \mathbf{Z} \leftarrow Y$ in the target graph G. Since the score is decomposable, adding an adjacency $X - Y$ increases it. (Otherwise GES search would fail to correctly identify the graph $X \rightarrow \mathbf{Z} \leftarrow Y$ in the sample size limit.) Since $\mathbf{Z}$ is a common child in the true graph G, and Y is a proper spouse of X in π, the spousal test shows a dependence between X and Y that is not covered in π. So IGES search adds an adjacency $X - Y$, contrary to the hypothesis that it terminates with π. [4] shows that if the growth phase of GES search with a consistent score terminates with an I-map of the target distribution, then so does the shrink phase. So GES search never selects a smaller graph that fails to cover a true dependency. Thus any moves from pattern π to pattern π' selected by GES during its shrink phase covers the dependencies $\mathcal{D}(\pi)$ found by the statistical test. Therefore in the sample size limit the shrink phase of GES returns an I-map of the target distribution.

The *computational overhead* compared to regular local score optimization is the number of statistical calls. For a graph G with n nodes, the number of Markov blanket independence hypotheses is on the order of $O(\binom{n}{2})$—two tests for each pair of nodes X, Y that are not in each other's Markov blanket. By taking advantage of the structure of the local search procedure, we can often reduce the set of hypotheses to be tested to an equivalent but smaller set. For example, if the local search adds a single edge $X \rightarrow Y$ to a graph G, the only nodes whose Markov blanket has been affected are X, Y and the parents of Y. Assuming that the tar-

Algorithm 1 The *IGES* procedure adapts GES based on the neighborhood structures $\mathtt{nbdh}^+$ and $\mathtt{nbdh}^-$.

Input: data sample $\mathtt{d}$ for random variables $\mathbf{V}$.
Calls: score evaluation function $\mathtt{score}(\pi, \mathtt{d})$, statistical testing procedure $\mathtt{find\text{-}new\text{-}dependencies}(\pi, \mathtt{d})$.
Output: BN pattern constrained by (in)dependencies detected in the data.

1: initialize with the disconnected pattern π over $\mathbf{V}$.
2: **for all** Variables X, Y **do**
3: test the hypothesis $X \perp\!\!\!\perp Y$
4: if $X \perp\!\!\!\perp Y$ is rejected by statistical test, add to detected dependencies stored in $\mathcal{D}$
5: **end for**
6: {begin growth phase}
7: **while** there is a pattern π' in $\mathtt{nbdh}^+_{\mathcal{D}}(\pi, \mathcal{D})$ **do**
8: choose π' in $\mathtt{nbdh}^+_{\mathcal{D}}(\pi, \mathcal{D})$ with maximum score
9: $\mathcal{D} := \mathcal{D} \cup \mathtt{find\text{-}new\text{-}dependencies}(\pi', \mathtt{d})$
10: **end while**
11: {begin shrink phase}
12: **while** there is a pattern π' in $\mathtt{nbdh}^-_{\mathcal{D}}(\pi, \mathcal{D})$ with greater score than current pattern π **do**
13: choose π' in $\mathtt{nbdh}^-_{\mathcal{D}}(\pi, \mathcal{D})$ with maximum score
14: **end while**
15: {prune pattern π further with "nothing and everything else" tests}
16: for any two variables X and Y that are adjacent in π, if $X \perp\!\!\!\perp Y$ or $X \perp\!\!\!\perp Y | \mathbf{V} - \{X, Y\}$ are not rejected by the statistical test, remove the link between X and Y.
17: repeat growth phase and shrink phase (lines 6-10).
18: Return the current pattern π.

get graph has constant degree (as in the analysis of the PC algorithm [23, Ch.5.4.2.1]), only a linear number of new independence tests is required at each stage of the search. Thus we expect that in practice, the order of independence tests required will be $O(n \times ca)$ where ca is the total number of candidate structures examined during the local search. Our simulations provide evidence for this hypothesis (Section 4).

4 Empirical Evaluation of Hybrid Criterion With Standard Search+Score Method

We performed a large number of experiments, but restrict ourselves to a few key findings due to space constraints. Our code is written in Java and uses many of the tools in the Tetrad package [5].

4.1 Experiments with Synthetic Data

The target models considered were randomly generated networks with 5-20 variables. We used Tetrad's random DAG generating functions to build the networks: A parent and a child are chosen at random, and the corresponding edge is added to the random graph unless it violates graph constraints. The number of edges is also determined randomly, with the constraint there are at most twice as many edges as nodes. For each graph, we drew samples of various sizes (ranging from 100 to 20000). We repeated the experiment 30 times, resulting in 30 random graphs for each combination of sample size and node count. Our graphs and tables display the average of the 30 networks for all measurements. The following learning methods were applied with the BIC score function.

1. Score-based search: GES starting with the empty graph.
2. Constraint-based search: PC algorithm [23] with z test and significance level $\alpha = 5\%$.
3. Backward Selection [10]: start with the complete DAG with all edges, apply the shrink phase of GES search.
4. Hybrid search method. IGES + SIN search with z test and significance level $\alpha = 5\%$.

Model Complexity and BIC score Our key findings are graphed in Figure 3. Our simulations show that the hybrid criterion effectively reduces the overfitting tendency of the regular score-based criterion, as measured by the number of edges in the learned model versus the number in the true graph. We found that IGES without the SIN tests leads to a small improvement in average number of edges; because our random networks are relatively small, the overfitting tendency of the score is not as strong as with the real-world structures. In the smaller networks the SIN tests aid substantially in reducing the model complexity. Our simulations show that hybrid search achieves a BIC score about as high as regular GES search on average. The high BIC score indicates that IGES + SIN fits the data as well as regular GES with fewer edges.

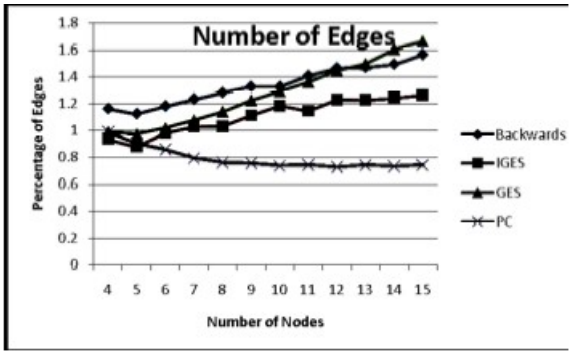

Figure 3: The figure shows the distribution of the edge ratio for the comparison methods, defined as #edges in target graph/#edges in learned graph. A ratio of 1 is ideal. The x-axis indicates the sample size, the y-axis the average edge ratio over 30 networks drawn at the given sample size. OUR method in this experiment is IGES + SIN. The average edge ratio for IGES + SIN is closer to 1 than for GES, which has a clear tendency towards more complex models. The improvement increases with sample size and network size.

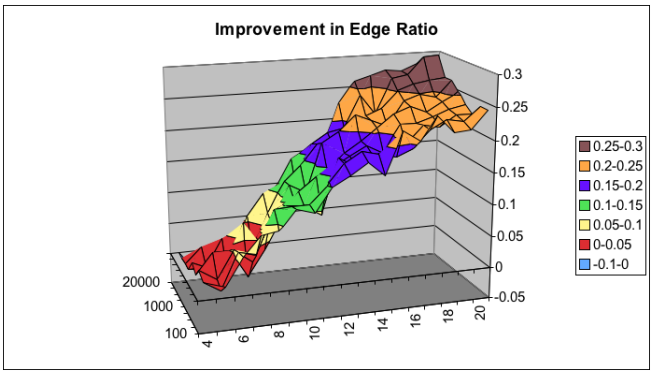

Figure 4: The improvement of the edge ratio attained by IGES.

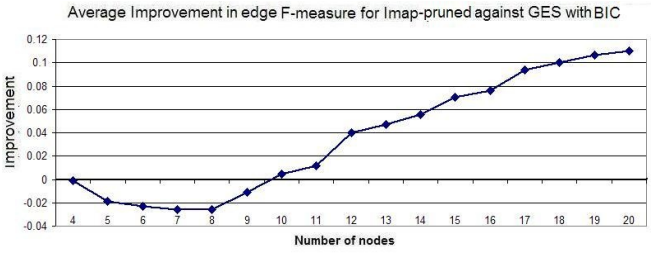

Figure 5: Average improvement in edge F-measure of Imap-pruned over the GES algorithm (both using BIC score) plotted against number of nodes.

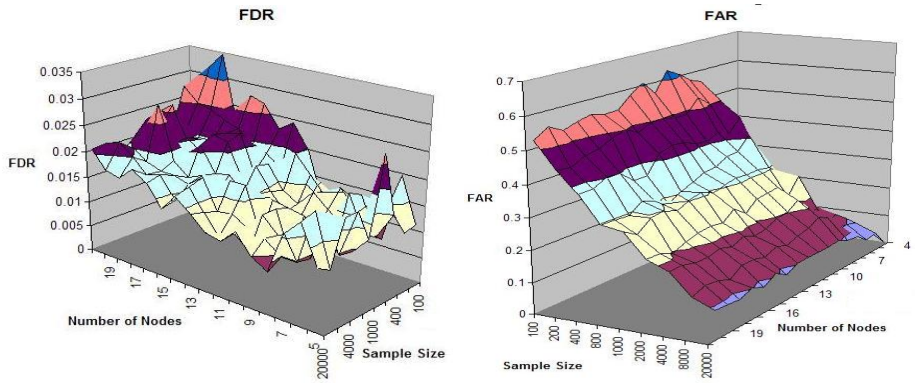

Figure 6: Left: False Discovery Rate for IGES, defined by # rejected true independence hypotheses/#tested independence hypotheses. The FDR is smaller than the significance level $\alpha = 5\%$. Right: False Acceptance Rate for IGES, defined by # accepted false independence hypotheses/# independence hypotheses.

Performance of Statistical Testing Strategy A number of measurements concerns the behavior of the testing strategy. A standard measure for the performance of a multiple hypothesis testing method is the *false discovery rate* (FDR) [1], which is defined as #rejected true independence hypotheses/#tested independence hypotheses. For the SIN independence hypotheses we also measured the *false acceptance rate* (FAR), defined as #false accepted independence hypotheses/#tested independence hypotheses. Figure 4.1 shows that in our simulations, with the significance level fixed at $\alpha = 5\%$, the FDR in random graphs was on average no greater than α, which is a good result in light of the Bonferroni inequality. In fact, for most experimental constellations the FDR was below 1.5%; it peaks at 3.5% with sample size = 100, number of nodes = 4. For sample size 1,000 the average FAR is about 20%, and decreases linearly to about 5% for sample size 10,000. The results support our strategy of treating rejections of the null hypothesis as much more reliable than acceptances.

We also examined the computational overhead incurred by carrying out statistical testing in addition to score-based search. The theoretical analysis of Section 3.1 suggests that the number of independence tests should be linear in the length of the search. Our results confirm this expectation. The exact slope of the line depends on the sample and graph sizes; averaging over these and plotting the number of independence tests as a function of number of candidate graphs examined during the search, we find that the number of tests performed is about 6 times the number of graphs generated. For off-line analysis of a dataset, the testing overhead seems acceptable given the improvement in the quality of the learned model. As a side benefit, the observed correlations are often of interest in themselves to the user, and they help to explain the construction of the learned structure.

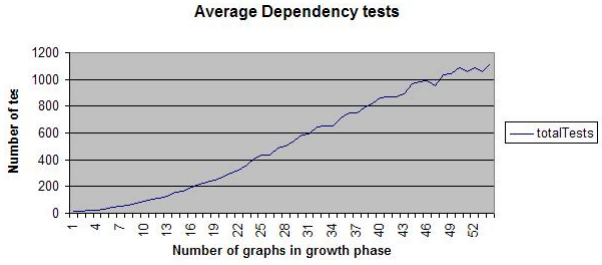

Figure 7: The number of independence tests as a function of the number steps in the growth phase. There is a linear relationship between the number of tests and the number of steps performed in the growth phase.

4.2 Simulations with Real World Networks

We examined a famous real-world network used used in the SEM Literature to model social alienation [22]. The network consists of 9 variables, and 9 edges connecting them. For one sample of 1000 data points our method recovers 8 of these edges without adding any additional adjacencies; in contrast, GES overfits the model by outputting 11 edges (9 true edges and 2 false ones). The two graphs are given in the supplementary material; they illustrate the typical difference between GES and IGES. Simulations with more samples (about 5) and different sample sizes (200; 4000) yield similar results. We note that our experiments indicate that with larger graphs, the difference in model quality would be still greater.

Our experiments with real-word BNs with more nodes—Alarm [**?**] (37 nodes) and Insurance [**?**] (25 nodes)—indicate that for larger graphs, the significance level should be adjusted downward to maintain a suitable false discovery rate for the testing strategy. A static approach is to use a fixed conservative α such as 1% or 0.1% (cf. [8]). With both $\alpha = 1\%$ or 0.1%, we observed a uniform improvement in KL-divergence

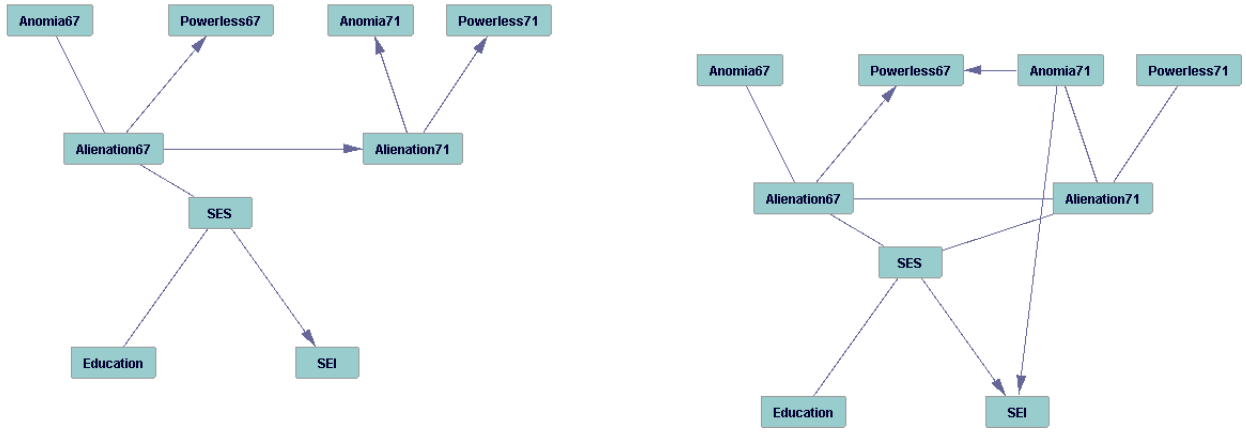

Figure 8: An experiment on reconstructing the social alienation model from the SEM literature. Left: The IMAP algorithm returns a model that differs from the true model by only one edge (between SES and Alienation71). Right: In contrast, GES overfits the model by adding 2 additional (false) edges to the true model.

for BDeu/IGES over BDeu/GES that is statistically significant, but whose magnitude is less than with the smaller random graphs. The adjacency f-measures are virtually the same (Graphs are available at [**?**].) It appears that the statistical testing leads to the introduction of correct adjacencies that lower the KL-divergence, but also false adjacencies that balance out the overall adjacency f-measure. We expect further improvement from a dynamic strategy for controlling the FDR of multiple hypothesis testing, such as the BH procedure [1].

5 Conclusion and Future Work

This paper presented a hybrid method for learning linear Gaussian BN structures or structural equation models. Our hybrid method combines strengths from both score-based and constraint-based BN learning approaches. Compared to traditional score-based approaches, the statistical testing performed by a hybrid method detects regularities in the data that constrain the search and can guide it towards a better model. Compared to traditional constraint-based methods, the model selection score serves as a heuristic to search for a structure that satisfies the observed (in)dependency constraints. Also, a hybrid method can adopt a strategy for selecting statistical hypotheses that focuses on a relatively small set of tests that can be performed reliably. In this paper our testing strategy was based on the Markov blanket, and we treated only rejections of independence hypotheses as hard constraints on the score-based search. Thus our hybrid method is less sensitive to the failures of independence tests that are the main problem for constraint-based methods. For small graphs, we attained further improvement by applying the recent SIN testing strategy, treating SIN independencies as soft constraints for the score-based search.

We showed how to adapt a generic local search+score procedure for the constrained optimization required by the hybrid criterion. Evidence from simulation studies with the well-established BIC criterion indicates that, when the number of variables exceeds about 10, the additional constraints from statistical tests help select a model that is less complex yet fits the data as well as the model selected by unconstrained learning. A recent direction in learning directed models has been to apply L1-regularization to the BIC score [20] for Markov blanket selection. An avenue for future research is to apply an L1-penalized score with our hybrid search method instead of the original BIC score. In sum, our hybrid method appears to be a principled and effective way to address overfitting in learning Gaussian Bayes networks that combines ideas from both score-based and constraint-based learning to addresses the weakness of each.

References

[1] Y. Benjamini and Y. Hochberg. Controllling the false discovery rate—a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, 57(1):289–300, 1995.

[2] R. R. Bouckaert. *Bayesian belief networks : from construction to inference*. PhD thesis, Universiteit Utrecht, 1995.

[3] J. Cheng and R. Greiner. Learning bayesian networks from data: An information-theory based approach. *Artificial Intelligence*, 137:43–90, 2002.

[4] D. Chickering. Optimal structure identification with greedy search. *JMLR*, 3:507–554, 2003.

[5] The tetrad project: Causal models and statistical data, 2008. http://www.phil.cmu.edu/projects/tetrad/.

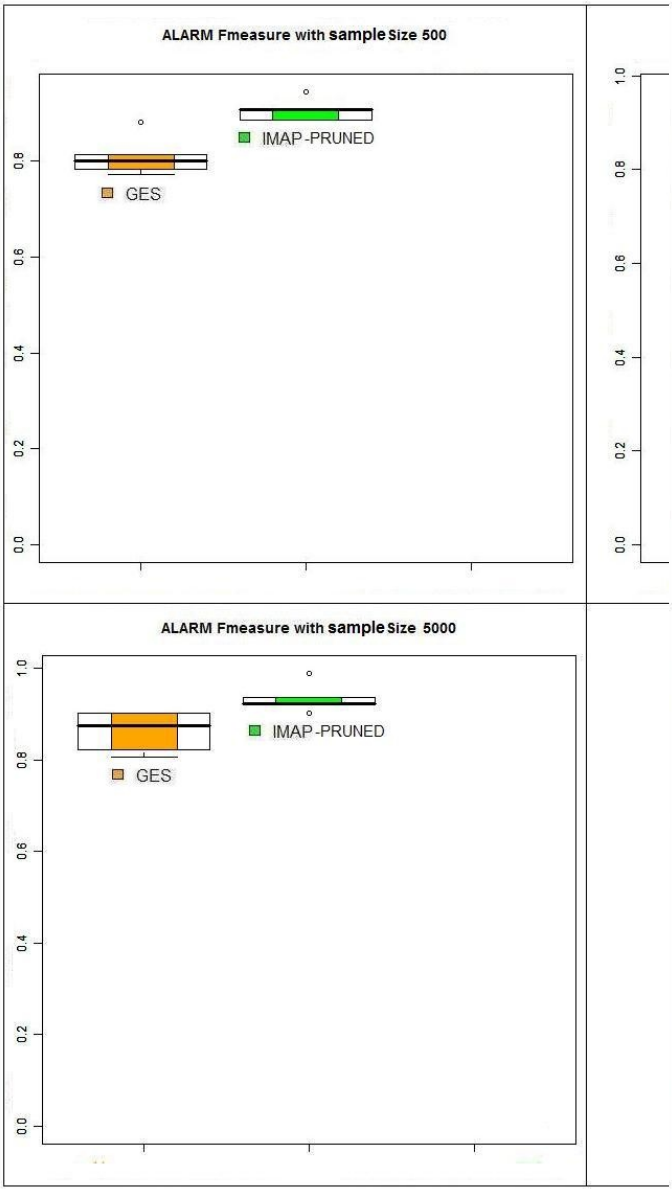

Figure 9: Boxplots comparing the F-measure measure in the alarm network for 3 different sample sizes. Higher F-Measure values indicate a closer fit to the target structure. Note: the scale for each boxplot pair in not the same. This was done in order to better display the differences for each setting.

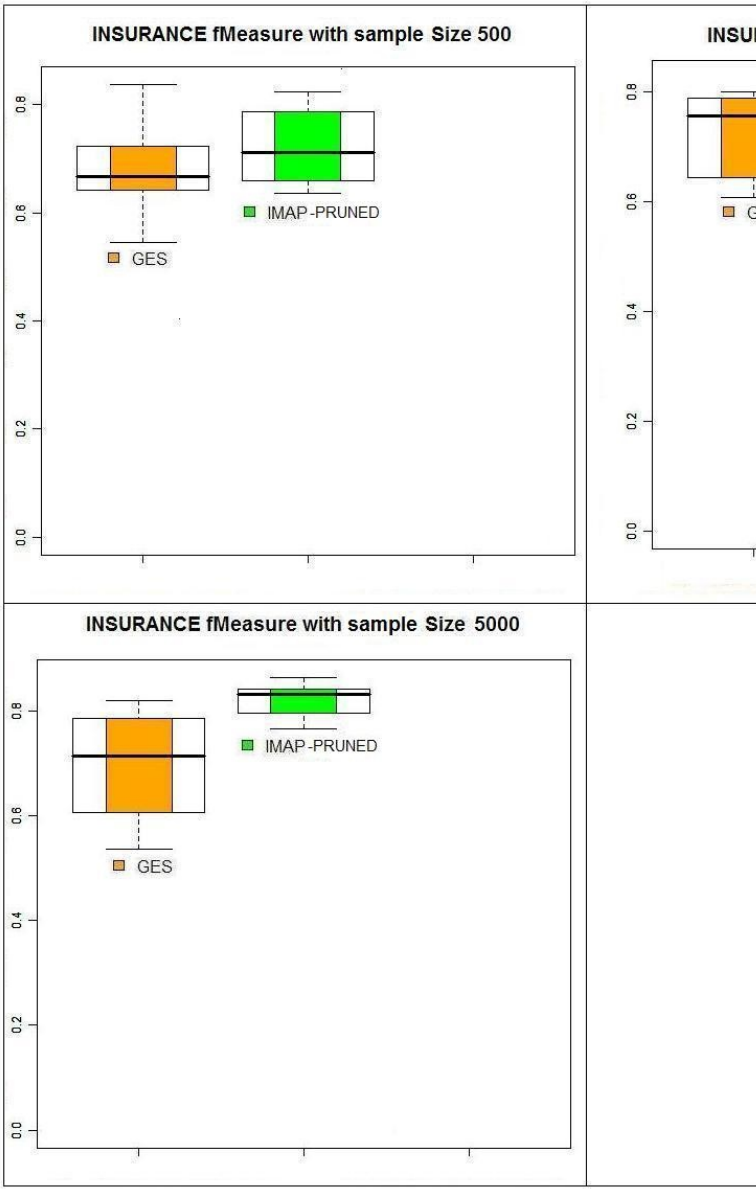

Figure 10: Boxplots comparing the F-measure measure in the insurance network for 3 different sample sizes. Higher F-Measure values indicate a closer fit to the target structure.

[6] G. Cooper. An overview of the representation and discovery of causal relationships using bayesian networks. In C. Glymour and G. Cooper, editors, *Computation, Causation, and Discovery*, pages 4–62. MIT, 1999.

[7] Denver Dash and Marek J. Druzdzel. Robust independence testing for constraint-based learning of causal structure. In Christopher Meek and Uffe Kjærulff, editors, *UAI*, pages 167–174. Morgan Kaufmann, 2003.

[8] L. de Campos. A scoring function for learning bayesian networks based on mutual info. and cond. indep. tests. *JMLR*, pages 2149–2187, 2006.

[9] Drton and Perlman. A sinful approach to bayesian graphical model selection. *Journal of Statistical Planning and Inference*, 138:1179–1200, 2008.

[10] D.M. Edwards. *Introduction to Graphical Modelling*. Springer, New York., 2000.

[11] N. Friedman, D. Pe'er, and I. Nachman. Learning bayesian network structure from massive datasets. In *UAI*, pages 206–215, 1999.

[12] Michael Hay, Andrew Fast, and David Jensen. Understanding the effects of search constraints on structure learning. Technical Report 07-21, U Mass. Amherst CS, April 2007.

[13] D. Heckerman. A tutorial on learning with bayesian networks. In *NATO ASI on Learning in graphical models*, pages 301–354, 1998.

[14] R.B. Klein. *Principles and practice of structural*

equation modeling. Guilford, 1998.

[15] D. Margaritis and S. Thrun. Bayes. net. induction via local neighbor. In *NIPS*, pages 505–511, 2000.

[16] C. Meek. *Graphical Models: Selecting causal and statistical models*. PhD thesis, CMU, 1997.

[17] R. E. Neapolitan. *Learning Bayesian Networks*. Pearson Education, 2004.

[18] J. Pearl. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kauffmann, 1988.

[19] J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge university press, 2000.

[20] M Schmidt, A Niculescu-Mizil, and K Murphy. Learning graphical model structure using l1-regularization path. In *AAAI*, 2007.

[21] O. Schulte, W. Luo, and R. Greiner. Mind change optimal learning of bayes net structure. In *20th Annual Conference on Learning Theory (COLT)*, 2007. Unpublished manuscript.

[22] Structural equation modeling with the sem package in r.

[23] P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. MIT Press, 2000.

[24] I. Tsamardinos, L. E. Brown, and C. F. Aliferis. The max-min hill-climbing bayesian network structure learning algorithm. *Machine Learning*, 65(1):3178, 2006.

[25] T. van Allen and R. Greiner. Model selection criteria for learning belief nets: An empirical comparison. In *ICML*, pages 1047–1054, 2000.

[26] Y. Xiang, S. K. Wong, and N. Cercone. Critical remarks on single link search in learning belief networks. In *UAI*, pages 564–57, 1996.