The IMAP Hybrid Method for Learning Gaussian Bayes Nets

Oliver Schulte[1], Gustavo Frigo[1], Russell Greiner[2], and Hassan Khosravi[1]

[1] School of Computing Science, Simon Fraser University,
Burnaby, B.C., Canada V5A 1S6
{oschulte,gafrigo,hkhosrav}@cs.sfu.ca
[2] Department of Computing Science, University of Alberta,
Edmonton, Alberta Canada T6G 2E1
greiner@cs.ualberta.ca

Abstract. This paper presents the I-map hybrid algorithm for selecting, given a data sample, a linear Gaussian model 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 local minima. Simulation studies with GES search and the BIC score provide evidence that for nets with 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 each 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.

There are two well established general approaches to learning a BN structure. Constraint-based (CB) methods employ a statistical test to detect conditional (in)dependencies given a sample $\mathtt{d}$, 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

A. Farzindar and V. Keselj (Eds.): Canadian AI 2010, LNAI 6085, pp. 123–134, 2010.

statistical tests and score functions in the context of continuous variables. 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 (co-parents) 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 scoring function.

There are theoretical, statistical and computational motivations for this composite selection criterion. It is well-known in Bayes net theory that a BN model that represents the target or operating distribution generating the data must satisfy the Markov boundary condition. It is widely accepted that a 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. Schulte et al. provide a learning-theoretic justification for minimizing the number of edges as a small-sample selection criterion [22]. *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 [19]. Our simulations provide further empirical evidence to support this finding. Our composite criterion addresses overfitting by assigning higher priority to minimizing the number of edges rather 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 [4, Lm. 4.5]; the score serves as a heuristic for exploring the search space.

For experimental evaluation, we adapted the state-of-the-art Graph Equivalence Search (GES) procedure [16, 5]. 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 graphs with 10 nodes and greater, we observe that BIC significantly overfits the data in the sense that it produces graphs with too many adjacencies. Our simulations 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 structure is closer to the target graph, as measured by the number of correctly/incorrectly placed edges. Our source code is available for anonymous ftp access at ftp://ftp.fas.sfu.ca/pub/cs/oschulte/imap/.

Paper Organization. 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. It provides a proof of consistency (asymptotic correctness) for our hybrid search procedure. Section 4 presents simulation studies

that compare constrained GES search with the BIC score to regular GES search with the same score.

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 CMU's Tetrad system [6]. (3) Other standard criteria like AIC penalize complex structures less than BIC so the tendency of BIC towards complex models corrected by our algorithm is even stronger with these criteria.

Hybrid Methods. Tsamardinos et al. [24] recently presented a hybrid method (max-min hill climbing) for discrete variables that treats the tests of statistical outcomes as constraints. While this work indicates that independence constraints from a statistical test can improve a score-based search, Hay et al. [12] show that because it accepts independence null hypotheses, max-min hill climbing is sensitive to type II errors. This paper extends our earlier work [21] as it treats only dependencies (rejections of the null hypothesis) as "hard" constraints. However, the previous algorithm addressed the problem of *underfitting* in score-based BN learning with discrete variables, whereas the problem in BN learning in Gaussian models is overfitting. Therefore the previous method adds more adjacencies than regular score-based search, whereas the method of this paper adds 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.

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 and variance σ_Y^2. The variance of ε_Y and the coefficients a_i are parameters of the model. The mean and variance of each root node 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 $BIC(G, \mathsf{d}) = L(\hat{G}, \mathsf{d}) - par(G) \cdot ln(m)/2$

where $\hat{G} = \hat{G}(d)$ 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 $X \rightarrow Y \leftarrow Z$. We assume familiarity with the notion of d-separation [18]. We write $(\mathbf{X} \perp\!\!\!\perp \mathbf{Y}|\mathbf{S})_G$ to denote that two disjoint sets $\mathbf{X}$ and $\mathbf{Y}$ of vertices are d-separated by a third set $\mathbf{S}$ in G. If two sets $\mathbf{X}$ and $\mathbf{Y}$ are not d-separated by $\mathbf{S}$ in graph G, then $\mathbf{X}$ and $\mathbf{Y}$ are **d-connected** by $\mathbf{S}$ in G, written $(\mathbf{X} \not\perp\!\!\!\perp \mathbf{Y}|\mathbf{S})_G$. We write $\mathcal{D}(G)$ for the set of all d-connections $(\mathbf{X} \not\perp\!\!\!\perp \mathbf{Y}|\mathbf{S})_G$ or conditional dependencies 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. 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 $MB_G(X)$. Given its Markov blanket $MB(X)$, each node X is d-separated from all other nodes outside of the Markov blanket.

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 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. 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

We first discuss employing statistical tests for detecting conditional (in) dependencies, then integrating statistical testing with a score-based local search.

Use of Statistical Tests. I-map learning requires a statistical significance test for testing 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 pattern 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 $NC_G(X, Y)$. (In a completely directed graph, these are just the parents of X and Y; our definition applies to partially directed patterns as well.) 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 **Markov blanket independencies** $\{X \perp\!\!\!\perp Y | MB_G(X) : Y \notin MB_G(X)\}$.
2. The **spousal independencies** $\{X \perp\!\!\!\perp Y | NC_G(X) : Y$ is a proper spouse of $X\}$.

These independence tests are well-suited for pattern-based search since the Markov blanket, adjacencies, and common children are determined by the pattern alone. The spousal independencies distinguish nodes on the Markov blanket that are both neighbors and spouses from nodes that are spouses only. If a graph entails a Markov blanket hypothesis (resp, spousal independency hypothesis), but a suitable test rejects the independency hypothesis, this is evidence that the graph 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 (cf. [21]). The procedure `find-new-dependencies` tests a set of independence hypotheses, so issues of multiple hypothesis testing arise. Any multiple hypothesis testing method can be employed to implement the functionality of `find-new-dependencies` [2, 9]. Like many other constraint-based and hybrid systems, we simply carry out multiple hypotheses at the same fixed significance level [23, 8, 15]. At an intermediate stage, 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 (1) the unconditional correlation betwen X and Y and (2) the correlation conditional on all other variables.

Heuristic Search Algorithm for I-map Learning. For our simulations we adapt the state-of-the art GES (Greedy Equivalence Search) local search algorithm. 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. The constrained version IGES (for I-map + GES) constrains the GES neighborhoods so they satisfy a given set of observed dependencies. Formally, the *growth*

Algorithm 1. The IGES procedure adapts GES based on the neighborhood structures nbdh^+ and nbdh^- to perform constrained score optimization with a statistical testing method

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 $\mathbf{V}$, and the empty dependency set $\mathcal{D}$.
2: **for all** Variables X, Y **do**
3: test the hypothesis $X \perp\!\!\!\perp Y$ on sample $\mathtt{d}$
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 $\mathsf{nbdh}^+_{\mathcal{D}}(\pi, \mathcal{D})$ **do**
8: choose π' in $\mathsf{nbdh}^+_{\mathcal{D}}(\pi, \mathcal{D})$ 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 $\mathsf{nbdh}^-_{\mathcal{D}}(\pi, \mathcal{D})$ with greater score than current pattern π **do**
13: choose π' in $\mathsf{nbdh}^-_{\mathcal{D}}(\pi, \mathcal{D})$ with maximum score
14: **end while**
15: {prune pattern π further with "nothing and everything else" SIN 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 once (lines 6-14).
18: Return the current pattern π.

neighborhood constrained by dependencies $\mathcal{D}$ is defined as follows:

$$\pi' \in \mathsf{nbdh}^+_{\mathcal{D}}(\pi) \quad \text{if and only if} \quad \pi' \in \mathsf{nbdh}^+(\pi) \text{ and } (\mathcal{D}(\pi') \cap \mathcal{D}) \supset (\mathcal{D}(\pi) \cap \mathcal{D}).$$

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. 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:

$$\pi' \in \mathsf{nbdh}^-_{\mathcal{D}}(\pi) \text{ if and only if } \pi' \in \mathsf{nbdh}^-(\pi) \text{ and } (\mathcal{D}(\pi') \cap \mathcal{D}) \supseteq (\mathcal{D}(\pi) \cap \mathcal{D}).$$

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.

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. The standard analysis of CB methods assumes the correctness of the statistical tests, which holds in the sample size limit [7, 23]. Under these assumptions, our local search method is consistent. The proof is available at [20].

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.*

Number of Statistical Calls. 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 target graph has constant degree (cf. [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 simulations, and summarize the main findings. More details are available in an extended version [20]. Our code is written in Java and uses many of the tools in the Tetrad package [6]. 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 $\alpha = 5\%$. We also refer to this as the I-map pruned method.

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 [6] as follows. (1) A parent and a child are chosen at random and the corresponding edge is added to the random graph unless it causes a cycle in the resulting graph. The number of edges is also determined randomly, with the constraint that there are at most twice as many edges as nodes. (2) Linear coefficients are drawn uniformly from the union of the intervals $(-0.5, -1.5)$ and $(0.5, 1.5)$. Variance parameters are drawn uniformly from the interval $(1.0, 3.0)$. Means are drawn from a standard normal distribution with mean 0 and variance 1. For each graph, we drew samples of various sizes (ranging from 100 to 20,000). We repeated the simulation 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.

Model Complexity and Structure. Our key findings are graphed in Figures 1 and 2. Figure 1 shows that the hybrid criterion together with the SIN tests 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. Without the SIN tests, the improvement is not as great. We measured the quality of the graph structure by combining adjacencies in the target structure (true positive) vs. adding adjacencies not present in the target structure (false positive) using the F-measure from information retrieval [26, p.146], which is defined as

$$\frac{2(\text{True Positive})}{2(\text{True Positive}) + (\text{False Positive}) + (\text{False Negative})}.$$

Higher F-measures are better. In general, the GES search produces more false positives than IGES search and fewer false negatives, as our edge-ratio measurements confirm. Figure 2 shows that the adjacency F-measure for the hybrid criterion is slightly worse for graphs with less than 10 nodes. This is because the overfitting tendency of the BIC score is small for small graphs, as our edge-ratio measurements confirm, so the overall balance of false positives and false negative is slightly favorable for unconstrained GES search. As the graph size increases, so does the number of false positives relative to graph size in GES, which means that the F-measure balance becomes favorable for the hybrid criterion.

Performance of Statistical Testing Strategy. A number of measurements concern the behavior of the testing strategy. A standard measure for the performance of a multiple hypothesis testing method is the *false discovery rate* (FDR) [2], 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. 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%

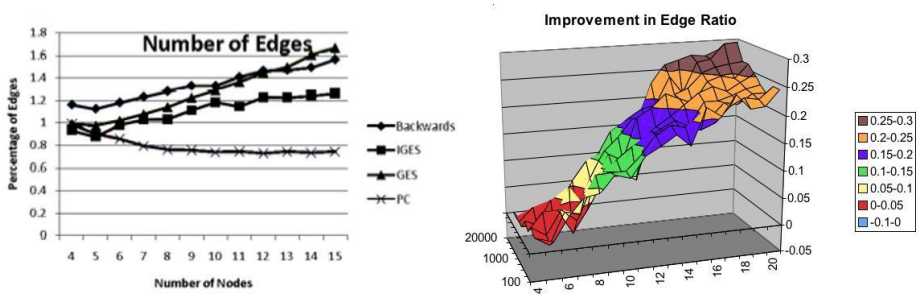

Fig. 1. Left: The figure shows the distribution of the edge ratio for the comparison methods, defined as #edges in target graph/#edges in learned graph. A ratio of 1 is ideal. The x-axis indicates the number of nodes, the y-axis the average edge ratio over all sample sizes for the given graph size (30 graphs per sample size and number of nodes). 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. Right: The improvement of the edge ratio attained by IGES+SIN; the y-axis shows edge-ratio(IGES+SIN)-edge-ratio(GES). The improvement increases with sample size and network size.

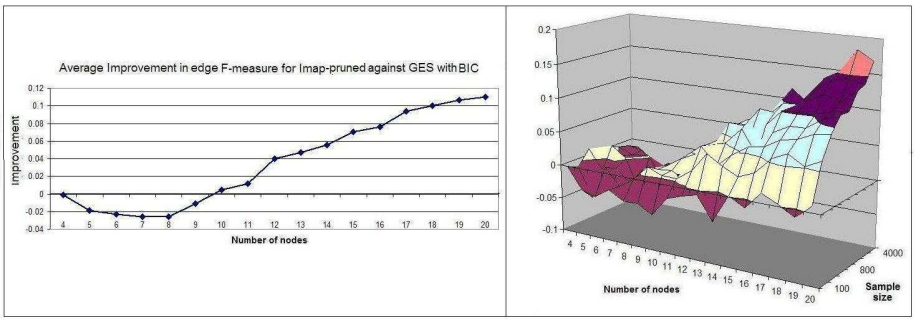

Fig. 2. Left: Average improvement in adjacency F-measure of IGES+SIN over the GES algorithm (both using BIC score) plotted against number of nodes. The x-axis indicates the number of nodes, the y-axis the difference IGES-GES for the average edge ratio over all sample sizes for the given graph size (30 graphs per sample size and number of nodes). Starting around 10 nodes, the average F-measure for IGES + SIN is better than for GES, which has a tendency towards overly complex models. Right: The improvement in adjacency F-measure increases with sample size and network size.

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. Both FAR and FDR decrease with sample size. The FDR also depends on the size of the graph, as it increases somewhat with larger graphs.

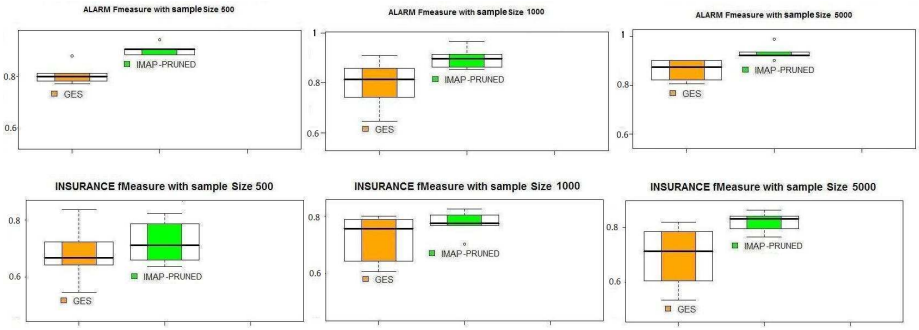

Fig. 3. Boxplots comparing the F-measure in the ALARM and INSURANCE networks for 3 different sample sizes, for GES search vs. IGES+SIN search (= Imap-pruned). Higher F-Measure values indicate a closer fit to the target structure. This plot shows the average F-measures over 5 random samples drawn for the given sample size. To better display the differences for each setting, the top figures uses a different scale from the bottom figure.

We also examined the *computational overhead* incurred by carrying out statistical testing in addition to score-based search. Our results show that the number of independence tests is roughly linear in the length of the search. 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 for each graph generated.

Simulations with Real World Networks. Our simulations with real-word BNs with more nodes—Alarm [1] (37 nodes) and Insurance [3] (25 nodes)—confirm that with larger graphs, the difference in model quality increases.[1] We observed an improvement in adjacency F-measure for the constrained method, both on average and in the variance of the results, as illustrated in Figure 3.

Conclusion and Future Work

This paper presented a hybrid method for learning linear Gaussian BN structures. 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. Our testing strategy was based on the Markov blanket. We treated only rejections

[1] These networks models were originally constructed with discrete variables. We followed the approach of Schmidt et al. [19] of using the same graph structure with continuous domains for the nodes.

of independence hypotheses as hard constraints on the score-based search. This makes our hybrid method less sensitive to the failures of independence tests, which are known to be the main problem for constraint-based methods.

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 appropriately complex in that it fits the target graph structure better than the model selected by unconstrained learning. 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 address the weakness of each.

Acknowledgements. This research was supported by NSERC discovery grants to RG and OS. RG gratefully acknowledges financial support from the Alberta Ingenuity Centre for Machine Learning. We are grateful to Joseph Ramsey for providing us with the source code of the Tetrad project.

References

[1] Beinlich, I., Suermondt, H., Chavez, R., Cooper, G.: The ALARM monitoring system. In: AIME 1989, pp. 247–256 (1989)
[2] Benjamini, Y., Hochberg, Y.: Controllling the false discovery rate—a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society 57(1), 289–300 (1995)
[3] Binder, J., Koller, D., Russell, S., Kanazawa, K.: Adaptive probabilistic networks with hidden variables. Machine Learning 29 (1997)
[4] Bouckaert, R.R.: Bayesian belief networks: from construction to inference. PhD thesis, Universiteit Utrecht (1995)
[5] Chickering, D.: Optimal structure identification with greedy search. JMLR 3, 507–554 (2003)
[6] The Tetrad project: Causal models and statistical data (2008), `http://www.phil.cmu.edu/projects/tetrad/`
[7] Cooper, G.: An overview of the representation and discovery of causal relationships using Bayesian networks. In: Glymour, C., Cooper, G. (eds.) Computation, Causation, and Discovery, pp. 4–62. MIT, Cambridge (1999)
[8] de Campos, L.: A scoring function for learning Bayesian networks based on mutual information and conditional independence tests. JMLR, 2149–2187 (2006)
[9] Drton, Perlman: A SINful approach to Bayesian graphical model selection. Journal of Statistical Planning and Inference 138, 1179–1200 (2008)
[10] Edwards, D.: Introduction to Graphical Modelling. Springer, New York (2000)
[11] Friedman, N., Pe'er, D., Nachman, I.: Learning Bayesian network structure from massive datasets. In: UAI, pp. 206–215 (1999)
[12] Hay, M., Fast, A., Jensen, D.: Understanding the effects of search constraints on structure learning. Technical Report 07-21, U Mass. Amherst CS (April)
[13] Heckerman, D.: A tutorial on learning with Bayesian networks. In: NATO ASI on Learning in graphical models, pp. 301–354 (1998)

[14] Klein, R.: Principles and practice of structural equation modeling. Guilford, New York (1998)
[15] Margaritis, D., Thrun, S.: Bayes. net. induction via local neighbor. In: NIPS, pp. 505–511 (2000)
[16] Meek, C.: Graphical Models: Selecting causal and statistical models. PhD thesis, CMU (1997)
[17] Neapolitan, R.E.: Learning Bayesian Networks. Pearson Education, London (2004)
[18] Pearl, J.: Probabilistic Reasoning in Intelligent Systems. Morgan Kaufmann, San Francisco (1988)
[19] Schmidt, M., Niculescu-Mizil, A., Murphy, K.: Learning graphical model structure using L1-regularization path. In: AAAI (2007)
[20] Schulte, O., Frigo, G., Greiner, R., Khosravi, H.: The IMAP hybrid method for learning Gaussian Bayes nets: Full version, `ftp://ftp.fas.sfu.ca/pub/cs/oschulte/imap/imap-linear.pdf`
[21] Schulte, O., Frigo, G., Greiner, R., Khosravi, H.: A new hybrid method for Bayesian network learning with dependency constraints. In: Proceedings IEEE CIDM Symposium, pp. 53–60 (2009)
[22] Schulte, O., Luo, W., Greiner, R.: Mind change optimal learning of bayes net structure. In: Bshouty, N.H., Gentile, C. (eds.) COLT. LNCS (LNAI), vol. 4539, pp. 187–202. Springer, Heidelberg (2007)
[23] Spirtes, P., Glymour, C., Scheines, R.: Causation, Prediction, and Search. MIT Press, Cambridge (2000)
[24] Tsamardinos, I., Brown, L.E., Aliferis, C.F.: The max-min hill-climbing bayesian network structure learning algorithm. Machine Learning 65(1), 31–78 (2006)
[25] van Allen, T., Greiner, R.: Model selection criteria for learning belief nets: An empirical comparison. In: ICML, pp. 1047–1054 (2000)
[26] Witten, I.H., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques, 2nd edn. Morgan Kaufmann, San Francisco (2005)