A New Hybrid Method for Bayesian Network Learning With Dependency Constraints

Oliver Schulte, Gustavo Frigo, Russell Greiner, Wei Luo, and Hassan Khosravi

***Abstract*— A Bayes net has qualitative and quantitative aspects: The qualitative aspect is its graphical structure that corresponds to correlations among the variables in the Bayes net. The quantitative aspects are the net parameters. This paper develops a hybrid criterion for learning Bayes net structures that is based on both aspects. We combine model selection criteria measuring data fit with correlation information from statistical tests: Given a sample d, search for a structure G that maximizes $score(G, \mathsf{d})$, over the set of structures G that satisfy the dependencies detected in d. We rely on the statistical test only to accept conditional *dependencies*, not conditional independencies. We show how to adapt local search algorithms to accommodate the observed dependencies. Simulation studies with GES search and the BDeu/BIC scores provide evidence that the additional dependency information leads to Bayes nets that better fit the target model in distribution and structure.**

I. Introduction

Bayes nets [1] are a widely used formalism for representing and reasoning with uncertain knowledge, with many applications ranging from medical diagnosis to scientific discovery. A Bayes net (BN) model is a directed acyclic graph $G = \langle \mathbf{V}, \mathbf{E} \rangle$ 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. A sparse BN compactly represents the joint probability distribution over a set of random variables. In this paper we consider Bayes nets with discrete variables only.

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 [2], [3], [4]. By constrast, score-based methods search for models that maximize a model selection score [3], [4]. Recent research into *hybrid methods* aims to combine the strengths of both approaches [5], [6]. Additional motivation for the hybrid approach comes from cognitive science and observations of human intelligence: Psychological studies have shown that people infer causal models on the basis of observed correlations [7], [3]. At the same time, they evaluate the importance of associations based on the observed frequencies of events.

A natural approach to a hybrid system is to treat the information from statistical tests as a constraint on the model selection search that effectively reduces the search space [8]. In this paper we propose a new *hybrid criterion*: find a Bayes net that maximizes the score given the constraint that the net must satisfy the dependencies detected by a suitable statistical test.

We provide a general schema for adapting *any* hill-climbing search algorithm with a given score function for the hybrid criterion. The adapted algorithm can be seen as a two-phase strategy for discovering a minimal Markov boundary: The growth phase performs hill-climbing with the given score function to add edges to the BN structure until for any two nodes X and Y, no statistically significant correlation is found between X and Y given the neighbors and spouses of X. The shrink phase performs hill-climbing to remove edges from the BN structure, maintaining the Markov boundary condition, until a local score optimum is reached.

For experimental evaluation, we adapted the state-of-the-art GES search procedure [9], [10] for constrained optimization; we refer to the resulting procedure as IGES (for "I-map + GES"). We report a number of simulations comparing GES search with and without dependency constraints, based on the well-established BDeu and BIC score functions [11], [10]. Simulation results compare the graphs learned with and without dependency constraints to the target graph. These simulations illustrate how for small to medium sample sizes, adding dependency constraints corrects some of the underfitting tendency of score functions. In our simulations, the constrained search model has lower KL-divergence and a better structure metric than the unconstrained search model. Most of the improvement occurs for graph sizes of 10 or less, and sample sizes of 1,000–2,000. It depends on the choice of the significance level α for controlling the type I error rate (falsely accepted dependencies); our experiments show that the interaction between error rate, sample size and graph size is crucial for learning performance.

a) Paper Organization: The next section reviews basic notions from Bayes net theory. Section III discusses the major design choices in our system, including our adaptation of GES search. Section IV presents simulation studies that compare constrained GES search with the BDeu score to regular GES search with the same score. As this paper proposes a hybrid model selection criterion, we compare our approach to both constraint-based and score-based methods.

Oliver Schulte, Gustavo Frigo, Wei Luo and Hassan Khosravi are with the School of Computing Science, Simon Fraser University, Canada (email: {oschulte,gafrigo,wluoa}@cs.sfu.ca). Russell Greiner is with the Department of Computing Science, University of Alberta, Canada (email: greiner@cs.ualberta.ca)

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.

b) Related Work. CB and Hybrid Methods: There are many constraint-based algorithms that employ statistical tests to discover BN structure [2], [12], [13]. Many of these methods use the "single link deletion" strategy [14]: if a significance test does not reject an independence null hypothesis X is independent of Y given S, then infer a conditional independence and mark variables X and Y as nonadjacent. As we do not infer independence from failure to reject, our approach does not rely on the single edge deletion strategy. To motivate this, observe that (at small to medium sample sizes) a rejection is a quite reliable indicator that the null hypothesis is false, but failure to reject is a less reliable indicator that the null hypothesis is true. Many statisticians recommend against inferring the truth of the null hypothesis when the null hypothesis is not rejected [15]; our use of statistical tests follows this recommendation and is more conservative than the use of tests in previous CB algorithms. For more discussion of independence tests in CB algorithms, see [3, p.593], [2, Sec.5.6], [16]. A recent hybrid method (max-min hill climbing) that incorporates the single link deletion strategy is presented in [5]. While this work indicates that independence constraints from a statistical test can improve a score-based search, the analysis of [8] shows that max-min hill climbing is still sensitive to errors of the independence test (type II errors). While the single edge deletion strategy has to address type II error, the issue for our system is type I error. Other previous hybrid BN learning algorithms (e.g., [6], [17]) 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. Our algorithm can be seen as a hybrid version of the Grow-Shrink procedure [12]. 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 employes the score function to select the next candidate structure.

In sum, the novelty in our use of statistical tests is the combination of (1) relying only on dependency information rather than independencies, (2) using the Markov blanket concept to select an informative set of independence hypotheses to test in a given graph, and (3) interleaving the testing strategy with a score-based search.

c) Related Work. Score+Search Methods: Several previous studies have observed the tendency of many score-based methods towards graphs that are sparser than the target structure [6], [18], [10]. The following simple experiment illustrates how standard model-selection scores can fail to capture statistically significant associations on small-to-medium sample sizes. It is meant only to elucidate the issue; more comprehensive studies appear in Section IV below. The target graph is the 3-node graph $X \to Z \leftarrow Y$ with ternary variables. Each simulation considered a sample size (ranging from 100 to 2500), and for each sample size we generated 1000 random parameter assignments for the target structure (with a uniform distribution over the [0,1] domain) and a random sample of the given size for the distribution defined by the parameter assignment. We used the Tetrad package [19], one of the most prominent software environments for Bayes nets based at CMU. The simulations investigated the BDeu score, with parameters set to the Tetrad default values following [10]: structure prior = 0.001, ESS = 10. The BDeu score has a Bayesian theoretical foundation, and has been shown to be a competitive score for learning BN models [11], [10]. The hybrid BDeu method combines the BDeu score with the same parameters with the requirement of fitting the statistically significant correlations (details in Section III).

Figure 1 shows that, in the simple 3-node target graph, the BDeu score often fails to fit statistically significant dependencies, and that fitting the dependencies often leads to recovering the target adjacencies. This improvement is statistically significant (using a two-tail paired t-test with p-value at 5%).

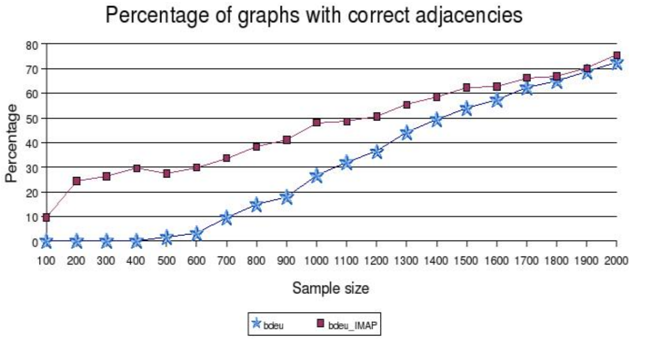


Fig. 1. An experiment to illustrate how a standard model selection score may not fit all statistically significant correlations. The target structure is $\boxed{X \to Z \leftarrow Y}$. The applied score function is the BDeu score, compared to hybrid BDeu that enforces fitting observed statistically significant correlations. The y-axis shows, for each of the two methods, how often the method discovers the correct adjacencies in the target graph.

Score-based functions trade off the global complexity of a structure with how well it fits the data overall; with some, as in the BDeu score, the trade-off is controlled by parameters for the score. It is difficult for a user to know a priori which balance between model complexity and data fit is best. A novel contribution of our hybrid method is using the results of statistical tests to aid the learning algorithm in detecting correlations in a in a *dynamic, data-driven* manner without the need for a user to set parameter values in advance.

II. Basic Definitions

We consider Bayes nets for a set of variables $\mathbf{V} = \{X_1, \ldots, X_n\}$ where each X_i has a *finite* number of values or states. 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 joint probability distribution over $\mathbf{V}$. Two nodes X, Y are **adjacent** in a BN if G contains an edge $X \to Y$ or $Y \to X$; 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** [1, 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 $(X \not\!\perp\!\!\!\perp Y|\mathbf{S})_G$. If $\mathbf{X}, \mathbf{Y}$ and $\mathbf{Z}$ are three disjoint sets of variables, then $\mathbf{Z}$ d-separates $\mathbf{X}$ and $\mathbf{Y}$ if for all variables $X \in \mathbf{X}$ and $Y \in \mathbf{Y}$, the set $\mathbf{Z}$ d-separates X and Y. We write $\mathcal{D}(G)$ for the set of all d-connections that hold in a graph G.

Let P be a joint distribution over 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 $(\mathbf{X} \perp\!\!\!\perp \mathbf{Y}|\mathbf{S})_P$, if $P(\mathbf{X}, \mathbf{Y}|\mathbf{S}) = P(\mathbf{X}|\mathbf{S})\,P(\mathbf{Y}|\mathbf{S})$ whenever $P(\mathbf{S}) > 0$. A BN structure G is an **I-map** of distribution P if $(\mathbf{X} \not\!\perp\!\!\!\perp \mathbf{Y}|\mathbf{S})_P$ implies $(\mathbf{X} \not\!\perp\!\!\!\perp \mathbf{Y}|\mathbf{S})_G$ for all variables X, Y and variable sets $\mathbf{S}$ disjoint from X, Y. For a given BN structure G and joint distribution P, there is a parametrization θ_G such that P is the joint distribution over $\mathbf{V}$ defined by $\langle G, \theta \rangle$ if and only if G is an I-map of P; see [3, Th.1.4,1.5].

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, written $MB_G(X)$. Given its Markov blanket $MB_G(X)$, each node X in G is d-separated from all other nodes outside of the Markov blanket. We refer to the set of independencies $\{X \perp\!\!\!\perp Y|MB_G(X) : Y \notin MB_G(X)\}$ as the **set of Markov blanket independencies** for graph G. If a graph G is an I-map of a distribution P, then all the Markov blanket independencies in G hold in P [1, Ch.3.3]. 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.

III. 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 dependencies, then integrating statistical testing with a score-based local search.

A. Use of Statistical Tests for Detecting Conditional Dependencies

I-map learning requires a statistical significance test for conditional independence hypotheses of the form $X \perp\!\!\!\perp Y|\mathbf{S}$. As with CB methods, the test can be chosen to suit the type of available data and application domain. We used the traditional χ^2 test for categorical data [20, Ch.9]. Since I-map learning treats the results of the statistical test as hard constraints, it is important that the decisions of the test be reliable, even on small to medium sample sizes. To this end, our system follows two principles for applying the significance test. (1) Accept rejection of the independence null hypothesis as indicating dependence, but draw no conclusion from failure to reject. (2) Require a minimum sample coverage for the χ^2 test [20, Ch.9.1]: the expected number of samples in each cell C_i must be at least 5; that is, $m \times p_i \geq 5$, where p_i is the probability of cell C_i according to the null hypothesis, and m is the sample size. The coverage condition implies that the χ^2 distribution is a reliable approximation to the distribution of the test statistic. If the sample coverage condition is not met, we draw no conclusion from the outcome of the test.

If a suitable test rejects the hypothesis that $X \perp\!\!\!\perp Y|MB_G(X)$ for two nodes X, Y, this is evidence that the graph G is not correct. I-map learning relies on a procedure `find-new-dependencies`(G) that takes as input a new graph G adopted during the local search, tests the new Markov blanket 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 2.

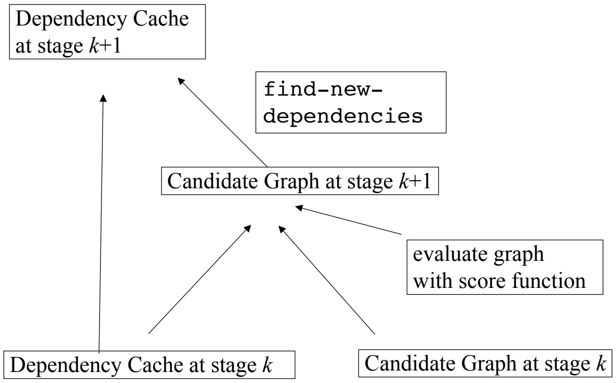

Dependency Cache at stage k+1
find-new-dependencies
Candidate Graph at stage k+1
evaluate graph with score function
Dependency Cache at stage k
Candidate Graph at stage k

Fig. 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 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 [21], [22]. Many constraint-based and hybrid systems simply carry out multiple hypotheses at the same fixed significance level [2], [6], [12]. Our simulations follow this approach, to facilitate comparisons with the competitor systems, and to investigate our hybrid model selection criterion in a relatively simple implementation first before examining more complicated systems. The next section describes how our statistical testing strategy is interleaved with a local search.

B. Heuristic Search Algorithm for I-map learning

We describe a general schema for adapting any local hill-climbing search procedure L with score function $score(G, \mathsf{d})$ [3, Ch.9.1] to perform optimization of the score in combination with a statistical testing strategy implemented by the procedure `find-new-dependencies`. The procedure can be applied to search in any DAG-based space, such as the space of patterns (Sec. II) or the space of DAGs; in what follows we simply refer to graphs. We call the constrained version of the L search procedure IL search (for I-map + L). If the current state of the search is a graph G, a local search procedure L moves to the highest scoring graph G' in a neighborhood $\mathsf{nbdh}(G)$ provided that $score(G', \mathsf{d}) > score(G, \mathsf{d})$. The *neighborhood constrained by dependencies* $\mathcal{D}$ is defined as follows. A graph G' is a member of $\mathsf{nbdh}^{\mathcal{D}}(G)$ if

1) $G' \in \mathsf{nbdh}(G)$ and $(\mathcal{D}(G') \cap \mathcal{D}) \supseteq (\mathcal{D}(G) \cap \mathcal{D})$, and
2) $score(G', \mathsf{d}) > score(G, \mathsf{d})$ or $(\mathcal{D}(G') \cap \mathcal{D}) \supset (\mathcal{D}(G) \cap \mathcal{D})$.

The first clause requires that a candidate graph G' for constrained optimization must be a candidate graph in the original search space, and that it must cover at least as many of the given dependencies $\mathcal{D}$ as the current graph G. The second clause stipulates that a candidate graph G' must make progress, in that G' has a higher score or covers more of the given dependencies. From a current graph G, the constrained IL search moves to the neighboring candidate graph $G' \in \mathsf{nbdh}^{\mathcal{D}}(G)$ with maximum score. Note that IL search may move to a graph with lower score G' if G' covers more dependencies and all the neighbors of G have a lower score than G. The search terminates with graph G when there are no more candidate graphs, that is, when $\mathsf{nbdh}^{\mathcal{D}}(G) = \emptyset$. Given the modified definition of neighborhood, this schema can be extended in an obvious way to local search strategies more complex than hill climbing. 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. Algorithm 1 gives pseudocode for IL search.

Algorithm 1 The IL procedure adapts a local BN search procedure based on a neighborhood structure `nbdh`.

```
Input: data sample d for random variables $\mathbf{V}$.
Calls: score evaluation function score($G$,d), statistical
testing procedure find-new-dependencies($G$,d).
Output: BN structure that maximizes score function given
dependencies.
initialize with the disconnected graph $G$ over $\mathbf{V}$.
for all Variables $X, Y$ do
    test the hypothesis $X \perp\!\!\!\perp Y$
    if $X \perp\!\!\!\perp Y$ is rejected by statistical test, add to detected
    dependencies stored in $\mathcal{D}$
end for
while $\mathsf{nbdh}^{\mathcal{D}}(G, \mathcal{D})$ is not empty do
    choose $G'$ in $\mathsf{nbdh}^{\mathcal{D}}(G, \mathcal{D})$ with maximum score
    $\mathcal{D} := \mathcal{D} \cup$ find-new-dependencies($G$,d)
end while
```

Analysis of Search Algorithm. The next observation asserts that constrained local search finds a local optimum for our hybrid criterion, provided that the basic operations of the local search procedure make it possible to reach an I-map for any set of dependencies $\mathcal{D}$. This is the case if single edge addition is one of the local operations; all local search algorithms that we know of consider single edge additions.

Observation 1: Let L be a local optimization procedure for a score function, with single edge addition as one of its basic operations. Then on any sample d, the constrained local search IL terminates with a local score optimum G that satisfies its Markov blanket independencies for a given statistical test. That is, if two nodes X and Y are d-separated in G given $MB_G(X)$, then the statistical test applied to sample d does not reject the hypothesis that $X \perp\!\!\!\perp Y | MB_G(X)$, and every neighbor of G has lower score or fails to satisfy a statistically significant dependency found during the search.

The proof is omitted due to space constraints.

The *computational overhead* compared to regular local score optimization is the number of statistical calls. For a graph G with n nodes, the Markov blanket independence hypotheses form an informative set of size $O(\binom{n}{2})$—two tests for each pair of nodes X, Y that are 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 (as in the analysis of the PC algorithm [2, 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 IV).

Adapting GES Search for Constrained Optimization. For our experiments we adapt the GES (Graph 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 [10]. 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; for a full description see [10]. During its "growth phase", GES adds an edge to a current pattern π, subject to several conditions, until reaching a local score maximum. Adding an edge to a pattern π leads to a pattern π' that covers strictly more conditional dependencies. During the subsequent "shrink phase", GES deletes an edge from a current pattern π, subject to several conditions, until reaching a local score maximum. GES is particularly natural for I-map learning because the set of entailed dependencies grows monotonically, so during its growth phase the second clause of the IL search condition is always satisfied. It is possible to show that if the conditional dependencies $\mathcal{D}$ constraining IGES hold in the target distribution, then in the sample size limit GES and IGES find the same graph (proof omitted due to space constraints). This implies that the asymptotic convergence guarantees for GES established by Chickering also hold for IGES [10].

Our motivation for using statistically significant dependencies is our expectation that fitting these dependencies will speed up convergence to a correct graph structure. The next section presents evidence from simulation studies based on GES that validate this expectation.

IV. Experimental comparison of Hybrid Criterion With Standard Search+Score Method

We begin with our experimental design and performance metrics. Most of our evaluation focuses on augmenting the BDeu score with statistical tests. We examine how the false acceptance rate, which is controlled by the choice of significance level, interacts with sample and graph size. We also summarize results for the BIC score, and for two real-word datasets (Insurance and Alarm).

A. Experimental Design and Performance Metrics

Our code is written in Java and uses many of the tools in the Tetrad package [19]. The target models considered were randomly generated networks with 4–10 binary nodes. We used Tetrad's random DAG generating functions to build the networks. A parent and a child are chosen at random, and if the corresponding edge does not violate graph constraints, it is added to the random graph. The number of edges is also determined randomly; the graphs are constrained so that the number of edges is at most twice the number of nodes. This skew towards sparser graphs favors standard model selection criteria rather than the hybrid criterion. The parameters for a given graph structure are uniformly and independently drawn from the [0,1] interval. For each graph, we drew samples of various sizes (ranging from 100 to 8000). 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. The following two learning methods were compared.

1) Search + score method. Score function: BDeu, with parameters set to the Tetrad default values, which were also used in [10]: structure prior = 0.001, ESS = 10. Search method: GES search (Section III-B).
2) Hybrid search, score and test method. Score function: BDeu, as above. Search method: IGES search (Section III-B) with χ^2 test and significance level $\alpha = 5\%$.

Performance Measures. The motivation for I-map learning is to achieve a better fit to the target distribution before convergence. As in other Bayes net learning studies (*e.g.*, [6], [18]), the distributional criterion considered is the Kullback-Leibler (KL) divergence of the fitted model to the true distribution [23]. Given a target distribution f that generates the training sample, and a DAG G inferred from the sample, let $\hat{f}_G$ be the fitted distribution (with MLE estimation of parameters [4]). Then the KL divergence of $\hat{f}_G$ to f is defined as $\text{KLD}(f, \hat{f}_G) = \mathbb{E}_f(\log f/\hat{f}_G)$ where $\mathbb{E}_f$ denotes the expectation with respect to distribution f. Our simulations use an exact method to compute KL divergence.

We also consider the structural difference between the target graph and the learned graph. The main effect of fitting dependencies is to add adjacencies. The KL-divergence simulation shows that these adjacencies are useful for fitting the target distribution. Our measurements examine the trade-off between adding adjacencies in the target structure (true positive) vs. adding adjacencies not present in the target structure (false positive). To aid interpretation of the experimental results, we combine positives and negatives using the F-measure [24, p.146], defined as

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

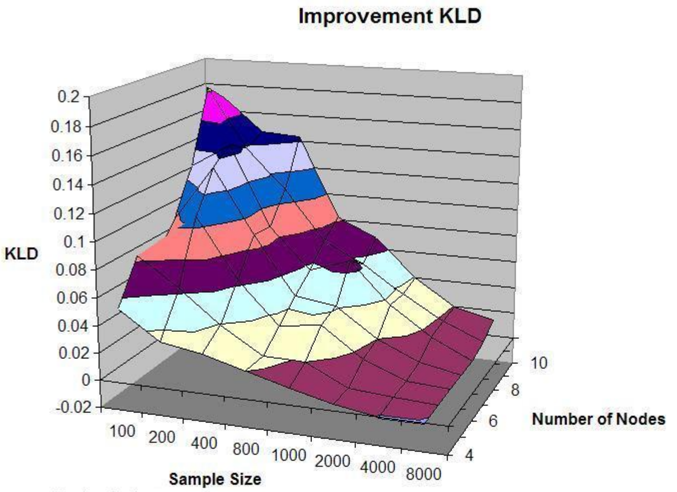
Higher F-measures are better.

B. Performance Measurements for BDeu Model Selection vs. Constrained BDeu Model Selection

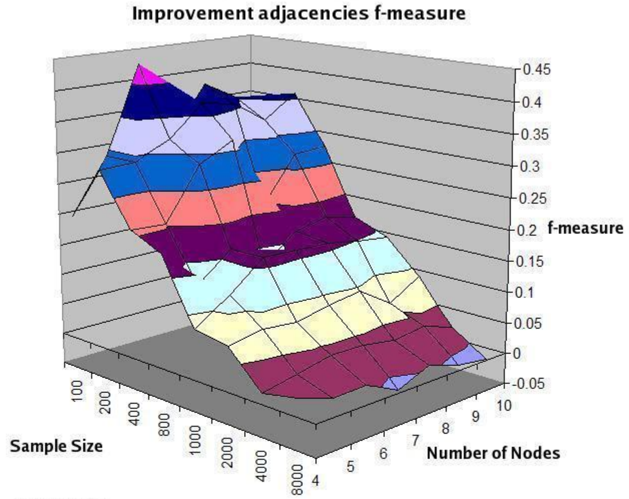
1) Simulations with Random Data: Figure 3 shows a uniform improvement for IGES search over regular GES search *on both the distributional and structural measures.* As expected, the improvements are especially pronounced for smaller sample sizes. The improvements are statistically significant for 47 of the 56 experimental constellations involving sample sizes from 100 to 8,000 (using a two-tail paired t-test with p-value at 5%).

Dependencies and Statistical Testing Strategy. The next measurement considers the strength of the dependencies present in the target distribution, detected by the statistical test, but missed by the model selection score. For example, if the target structure is $\boxed{X \rightarrow Y \quad Z}$, the score selects the structure $\boxed{X \quad Y \quad Z}$, and the test rejects the independence hypothesis $X \perp\!\!\!\perp Y$, then the dependency $X \not\perp\!\!\!\perp Y$ is in this category. For each dependency $X \not\perp\!\!\!\perp Y | \mathbf{Z}$ missed by the model selection score, we computed the conditional mutual information $I(X, Y | \mathbf{Z})$. Figure 4 confirms our expectation that explicit statistical testing helps the score-based learner to more quickly converge to dependencies of medium strength: At small sample sizes, the score is able to identify stronger associations only (mutual conditional information $>$ 0.06), and misses weaker yet statistically significant dependencies many of which are included in the graphs learned by the hybrid method. As the sample size increases, the score becomes sensitive to more and more of the weaker statistically significant dependencies, but only after the hybrid method has detected them.

Our next two measurements concerns the behavior of the testing strategy. A standard measure for the performance of



Number of nodes / Sample size	4	5	6	7	8	9	10
100	0.054	0.083	0.075	0.096	0.128	0.183	0.170
200	0.033	0.063	0.107	0.136	0.147	0.138	0.146
400	0.029	0.039	0.064	0.089	0.087	0.109	0.144
800	0.020	0.021	0.033	0.046	0.056	0.072	0.082
1000	0.012	0.020	0.037	0.041	0.067	0.056	0.069
2000	0.005	0.005	0.015	0.020	0.036	0.036	0.040
4000	0.000	0.003	0.003	0.006	0.015	0.018	0.020
8000	0.000	0.001	0.001	0.001	0.001	0.005	0.013



Number of nodes / Sample size	4	5	6	7	8	9	10
100	0.385	0.287	0.366	0.312	0.325	0.347	0.349
200	0.238	0.259	0.324	0.331	0.317	0.367	0.334
400	0.228	0.221	0.299	0.237	0.260	0.255	0.239
800	0.141	0.144	0.134	0.155	0.160	0.179	0.157
1000	0.099	0.101	0.142	0.158	0.166	0.143	0.135
2000	0.091	0.065	0.089	0.076	0.063	0.074	0.055
4000	0.022	0.022	-0.005	0.035	0.019	0.034	0.025
8000	-0.006	0.006	0.004	0.006	0.004	0.002	-0.019

Fig. 3. The figure summarizes the performance of the tested methods with respect to KL divergence and adjacency f-measure as a function of sample size and the number of nodes in the graph (average over 30 trials for each sample size/node number pair). The z-coordinate is the difference between the measurements for GES/BDeu and IGES/BDeu, where positive values indicate better performance by the hybrid method due to the use of statistical tests.

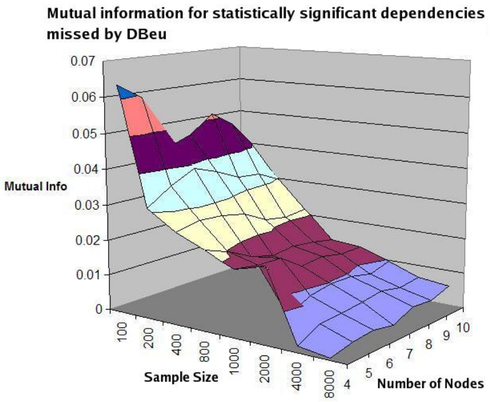


Fig. 4. The conditional mutual information for the true dependencies detected by the statistical test but missed by the score-based search without testing.

a multiple hypothesis testing method is the *false discovery rate* (FDR) [21], [25], which is defined as #rejected true independence hypotheses/#tested independence hypotheses. Figure 5 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 = 8,000, number of nodes = 4.

Our next measurement examines the computational overhead incurred by carrying out statistical testing in addition to score-based search. The theoretical analysis of Section III-A 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

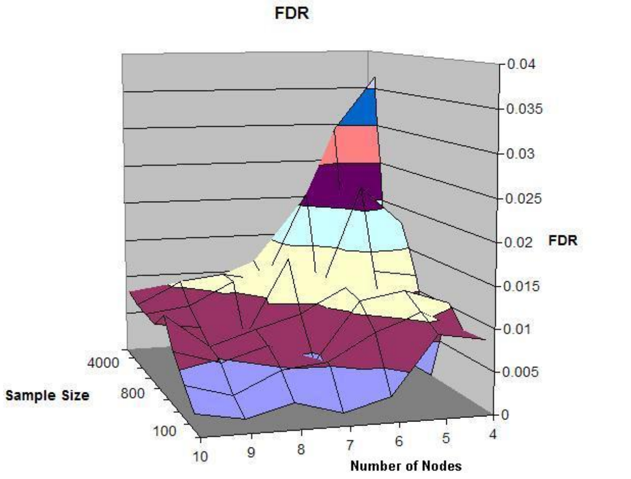


Fig. 5. False Discovery Rate for BDEU/IGES, defined by # rejected true independence hypotheses/#tested independence hypotheses. The FDR is smaller than the significance level $\alpha = 5\%$.

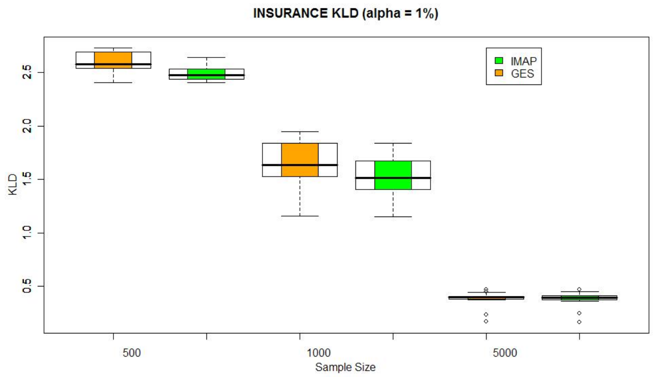


Fig. 6. Comparing GES/BDeu (left) and IGES/BDeu (right) on the Insurance network structure. For each sample size of 500, 1000, and 5000, we generated 14 random samples and compared the outputs of GES search with the standard Tetrad parameters settings for BDeu, with and without statistical testing. The significance level for the tests was fixed at $\alpha = 1\%$. As the adjacency f-measure does not show a statistically significant difference between the two methods, we plot the results only for KL-divergence. The box plot shows an improvement for the hybrid method IGES/BDeu up to sample size 5,000.

number of graphs generated. We omit the graph due to space constraints. 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.

2) Simulations with Insurance and Alarm Networks: We followed the same simulation protocol for generating samples and testing learning methods on two well-known real world BNs: Alarm [26] (37 nodes) and Insurance [27] (25 nodes) networks. We found 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. [6]). With both $\alpha = 1\%$ or 0.1%, we observed a uniform improvement in KL-divergence for BDeu/IGES over BDeu/GES that is statistically significant, but whose magnitude is less than with the smaller random graphs. Moreover, the additional statistical testing reduces the variance of the KL-divergence, suggesting that the extra constraints make density estimation more stable. The adjacency f-measures are virtually the same. We plot the results for the Insurance network in Figure 6; they are similar for Alarm.

3) Summary: For smaller graphs (10 nodes or less) and small to medium sample sizes (1,000-2,000), our measurements show a clear and uniform improvement with a standard choice of significance level, $\alpha = 5\%$. With larger graphs (20 or more nodes), the number of falsely accepted dependencies appears to lead to overly complex structures, unless the significance level is reduced, which in turn diminishes the gains from I-map learning. For larger graphs, we expect further improvement from a dynamic strategy for controlling the FDR of multiple hypothesis testing, such as the BH procedure [21] or the recent SIN approach for graphical models [28].

C. Comparison with BIC Score

Figure 7 (overleaf) compares the BDeu score with the widely-used BIC score [3, Ch.8.3.2]. To compare absolute performance metrics, we fix the number of nodes at 10; the results are similar for smaller graph sizes. The results show that the incremental improvement of adding testing to the BIC criterion (i.e., BIC/IGES vs. BIC/GES), while statistically significant, is not as great as the improvement with the BDeu criterion. The improvement is greater for ternary than for binary variables. The absolute performance is best with the BDeu/IGES combination, in the sense that this combination gets a better structural measure and a similar KL divergence, meaning that it approximates the target distribution using more correct adjacencies.

Discussion. This finding illustrates three general points. (1) Most model selection scores trade off model complexity vs. data fit. For scores that place relatively more weight on data fit, such as BIC, the incremental improvement from including observed dependencies is relatively less. (2) While scores that place more weight on data fit benefit less from additional statistical testing, their overall performance is often worse, at least in the sense that they produce overly complex models. Taking the BIC score as an example, a replication of the v-structure experiment from Section I-.0.c shows that it often introduces adjacencies whose corresponding dependencies are not statistically significant. In other simulations we observed a strong overfitting tendency of the BIC score in linear models with continuous variables, because in these models the number of parameters is relatively small compared to the complexity of the graph structure (the number of parameters for each node is linear in the number of its parents). (3) Whether a model selection criterion strikes the right balance between model complexity and data fit depends on the application domain. For a user it is therefore difficult to know a priori which model selection criterion is best. An advantage of I-map learning is that it expands a sparse model in a *dynamic, data-driven* manner by accommodating the observed statistically significant correlations. Thus the user can start with a score that has a relatively strong bias towards simple models (e.g., BDeu), and use the hybrid method to expand the model as necessary to accommodate the data. The spirit of our approach is similar to regularization methods, which introduce a scalar λ to weight the likelihood (data fit) term, and use data-driven methods (e.g., cross-validation) to estimate this weight.

V. Conclusion and Future Work

This paper introduced a new criterion for learning Bayes nets: find the graph G that maximizes a given score, subject to the constraint that G cover the dependencies detected by a statistical test. This is a hybrid criterion that combines the basic idea behind CB approaches—to treat the output of a statistical test as constraints on the learned structure—with the search+score framework. The practical motivation for the hybrid criterion is that many standard scoring criteria based

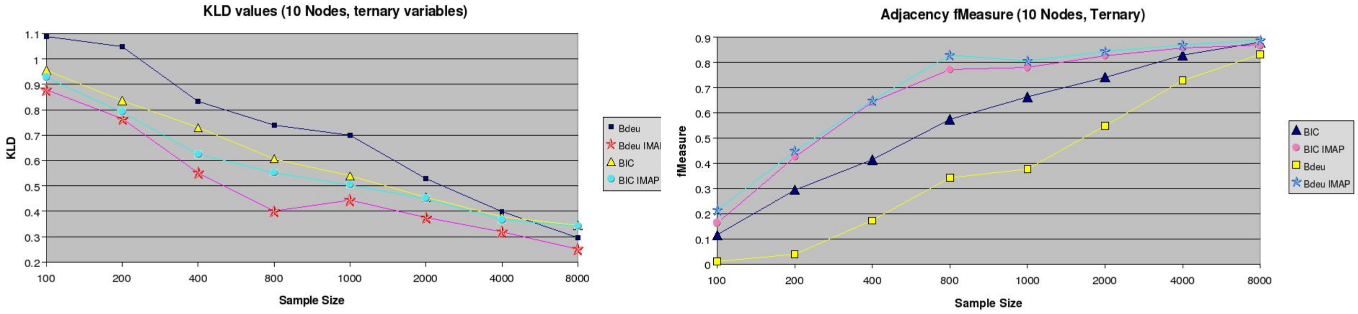


Fig. 7. Comparison of two different scores on synthetic data (10 nodes, ternary variables, averaged over 30 data sets): BDeu and BIC, each with and without statistical hypothesis testing. For each of the four methods, we give the KLD divergence and the f-measure with respect to adjacencies. Of the two model selection criteria without dependency constraints added, BIC performs better than BDeu, but among the hybrid methods, performance is best with the BDeu/IGEs combination.

on parameter counts tend to produce overly sparse graphs; our criterion selects expanded graphs that fit the observed statistically significant correlations.

We showed how to adapt a generic local search+score procedure for the constrained optimization required by the hybrid criterion. For BN structures with discrete variables, the number of parameters is large compared to the number of variables, and thus the penalization term tends to lead to underfitting. Evidence from simulation studies with the well-established BDeu and BIC criteria indicates that fitting statistically significant dependencies leads to better learning, as evaluated both by distributional and topological criteria. Our statistical testing strategy achieves this improvement without the need for the user to apply different model selection scores or select parameters for a given model selection score. We observed the improvement to be greatest when the number of variables is around 10 or less, and the sample size around 1,000-2,000.

In summary, our hybrid criterion is a bridge between the score-based and the constrained-based frameworks that combines frequentist and Bayesian prior-based methods. It appears to be a principled and effective way to address underfitting tendencies in Bayes net learning.

References

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

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

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

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

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

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

[7] A. Gopnik and L. Schulz, Eds., *Causal Learning*. Oxford University Press, 2007.

[8] M. Hay, A. Fast, and D. Jensen, "Understanding the effects of search constraints on structure learning," U Mass. Amherst CS, Tech. Rep. 07-21, April 2007.

[9] C. Meek, "Graphical models: Selecting causal and statistical models," Ph.D. dissertation, CMU, 1997.

[10] D. Chickering, "Optimal structure identification with greedy search," *JMLR*, vol. 3, pp. 507–554, 2003.

[11] D. Heckerman, D. Geiger, , and D. Chickering, "Learning bayesian networks: The combination of knowledge and statistical data," *Machine Learning*, vol. 20, pp. 197–243, 1995.

[12] D. Margaritis and S. Thrun, "Bayes. net. induction via local neighbor." in *NIPS*, 2000, pp. 505–511.

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

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

[15] F. L. Hogben, "Statistical prudence and statistical inference," in *The Significance Test Controversy*. Aldine, 1970.

[16] D. Dash and M. J. Druzdzel, "Robust independence testing for constraint-based learning of causal structure," in *UAI*, C. Meek and U. Kjærulff, Eds. Morgan Kaufmann, 2003, pp. 167–174.

[17] N. Friedman, D. Pe'er, and I. Nachman, "Learning bayesian network structure from massive datasets," in *UAI*, 1999, pp. 206–215.

[18] T. van Allen and R. Greiner, "Model selection criteria for learning belief nets: An empirical comparison." in *ICML*, 2000, pp. 1047–1054.

[19] R. Scheines, P. Spirtes, C. Glymour, C. Meek, and T. Richardson, *TETRAD 3:Tools for Causal Modeling*, CMU Philosophy, 1996.

[20] M. H. Degroot, *Probability and Statistics*, 2nd ed. Addison Wesley, 1986.

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

[22] G. Blanchard and F. Fleuret, "Occam's hammer," in *Learning theory*, ser. LNAI, N. H. Bshouty and C. Gentile, Eds., vol. 4539, COLT. Springer-Verlag Berlin Heidelberg, 2007, pp. 112–126.

[23] S. Kullback and R. Leibler, "On information and sufficiency," *Ann. Math. Stat.*, vol. 22, pp. 79–86, 1951.

[24] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd ed. Morgan Kaufmann, 2005.

[25] J. Listgarten and D. Heckerman, "Determining the number of non-spurious arcs in a learned dag model: Investigation of a bayesian and a frequentist approach." in *Proceedings of Twenty Third Conference on Uncertainty in Artificial Intelligence*. UAI Press, July 2007.

[26] I. Beinlich, H. Suermondt, R. Chavez, and G. Cooper, "The alarm monitoring system," in *AIME'89*, 1989, pp. 247–256.

[27] J. Binder, D. Koller, S. Russell, and K. Kanazawa, "Adaptive probabilistic networks with hidden variables," *Machine Learning*, vol. 29, 1997.

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