Lightning-speed Structure Learning of Nonlinear Continuous Networks

Gal Elidan
Department of Statistics, The Hebrew University

Abstract

Graphical models are widely used to reason about high-dimensional domains. Yet, learning the structure of the model from data remains a formidable challenge, particularly in complex continuous domains. We present a highly accelerated structure learning approach for continuous densities based on the recently introduced Copula Bayesian Network representation. For two common copula families, we prove that the expected likelihood of a building block edge in the model is *monotonic* in Spearman's rank correlation measure. We also show numerically that the same relationship holds for many other copula families. This allows us to perform structure learning while bypassing costly parameter estimation *as well as* explicit computation of the log-likelihood function. We demonstrate the merit of our approach for structure learning in three varied real-life domains. Importantly, the computational benefits are such that they open the door for practical scaling-up of structure learning in complex nonlinear continuous domains.

1 Introduction

Probabilistic graphical models, and in particular directed Bayesian networks (BNs) [Pearl, 1988], have become increasingly popular as a flexible and intuitive framework for modeling multivariate densities. An important super-exponential challenge is that of learning the graph structure $\mathcal{G}$ of these models from training data. Unfortunately, even when using on a simple greedy procedure, structure learning can be computationally prohibitive. This is particularly true for continuous nonlinear domains. In practice, with as few as tens of variables, learning any continuous model beyond the simple linear Gaussian BN one can be computationally demanding (see Section 6 for a recent exception). Unfortunately, for many domains, the linear parameterization can be too restrictive. Our goal is to overcome this barrier and scale up structure learning of complex continuous high-dimensional distributions.

A copula function [Nelsen, 2007] links *any* univariate marginals (e.g., nonparametric) into a multivariate joint distribution. A joint distribution parameterized by a copula is often easier to estimate and less prone to over-fitting than a fully nonparametric one. At the same time, copulas offer great flexibility in capturing nonlinear and multi-modal distributions. Recently, Elidan [2010] introduced the Copula Bayesian Networks (CBNs) model that fuses the copula and BN formalisms, allowing for the construction of high-dimensional graph-based distributions, while retaining the flexibility of copulas. In the context of multivariate density estimation, the construction has led to appealing performance gains. In this work we show that the CBN model opens the door for accelerated and effective structure learning.

Structure learning is most commonly carried out via a greedy search that is guided by a model selection score that is used to assess the merit of candidate structures (e.g., BIC [Schwarz, 1978]). The computational difficulty is in the evaluation of the log-likelihood function that, for BNs, equals a constant plus the mutual information between variables and their parents in the network. In the case of two jointly Gaussian variables, the expected information is *monotonic* in the absolute value of Pearson's correlation [Cover and Thomas, 1991]), so that the empirical correlation can be used as a surrogate model selection measure (this was recently used by Goldberger and Leshem [2011] to estimate Gaussian tree approximations). In this work we propose a more general proxy to the expected likelihood of the model that can be used to substantially speed-up structure learning of nonlinear CBNs.

Appearing in Proceedings of the 15^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2012, La Palma, Canary Islands. Volume XX of JMLR: W&CP XX.

Spearman's rank correlation coefficient ρ_s is a scale-invariant measure of association that is closely related to copulas. Specifically, it is a simple linear function of the integral of the copula (cumulative) distribution (see Section 2 for details). In this work we prove that the absolute value of Spearman's rho is monotonic in the expected likelihood of the model for two important copula families: the Gaussian and the Farlie-Gumbel-Morgenstern (FGM) copulas. We also show numerically that a similar monotonicity holds for many other common copula families.

The monotonicity result implies that, in place of costly computation, the easy to compute empirical Spearman's rho can serve as an accurate proxy to the benefit of an edge in a network. This allows us to bypass costly parameter estimation *as well as* explicit computation of the log-likelihood, thereby substantially speeding up the building block computation of structure learning. For few copula families (Gaussian,Plackett,FGM [Nelsen, 2007]), an explicit relationshop is known between the copula parameter and ρ_s, so that parameter estimation can already be carried out efficiently (e.g., [Genest and Favre, 2007]). As we show in the experimental evaluation, *even in these cases*, our approach results in substantial computational speed-ups. For other copula families, the running time benefits can be even more considerable.[1]

We use our approach to learn the structure of CBNs in three varied real-life domains, and demonstrate dramatic running time improvements. For the largest domain, we are able to learn a CBN model that offers considerable generalization benefits while taking no longer to learn than a simple linear Gaussian BN. This is in contrast to the typical learning scenario where a more expressive model that generalizes well goes hand in hand with increased computational demands. Importantly, the computational benefits are such that they facilitate practical scaling up of structure learning in complex nonlinear continuous domains.

2 Background

We briefly review copulas and the recently introduced Copula BN model [Elidan, 2010]. We start with the necessary notation. Let $\mathcal{X} = \{X_1, \ldots, X_N\}$ be a finite set of real-valued random variables and let $F_{\mathcal{X}}(\mathbf{x}) \equiv P(X_1 \leq x_1, \ldots, X_n \leq x_N)$ be a (cumulative) distribution over $\mathcal{X}$, with lower case letters denoting assignment to variables. For compactness, we use $F_i(x_i) \equiv F_{X_i}(x_i) = P(X_i \leq x_i, X_{\mathcal{X}/X_i} = \infty)$ and $f_i(x_i) \equiv f_{X_i}(x_i)$. When there is no ambiguity we sometimes abuse notation and use $F(x_i) \equiv F_{X_i}(x_i)$, and similarly for densities and for sets of variables.

2.1 Copulas

A copula function [Sklar, 1959] links marginal distributions to form a multivariate one. Formally,

Definition 2.1: Let $U_1, \ldots, U_N$ be real random variables marginally uniformly distributed on $[0, 1]$. A copula function $C : [0, 1]^N \to [0, 1]$ is a joint distribution

$$C_\theta(u_1, \ldots, u_N) = P(U_1 \leq u_1, \ldots, U_N \leq u_N),$$

where θ are the parameters of the copula function. ∎

Sklar's seminal theorem states that *any* joint distribution $F_{\mathcal{X}}(\mathbf{x})$ can be represented as a copula function C of its univariate marginals

$$F_{\mathcal{X}}(\mathbf{x}) = C_\theta(F_1(x_1), \ldots, F_N(x_N)).$$

When the univariate marginals are continuous, C is uniquely defined. The constructive converse, which is of central interest from a modeling perspective, is also true: *any* copula function taking *any* marginal distributions $\{F_i(x_i)\}$ as its arguments, defines a valid joint distribution with marginals $\{F_i(x_i)\}$. Thus, copulas are "distribution generating" functions that allow us to separate the choice of the univariate marginals and that of the dependence structure, encoded in the copula function C. Importantly, this flexibility often results in a construction that is beneficial in practice.

Assuming C has Nth order partial derivatives (true almost everywhere when continuous), the joint density can be derived from the copula function using the derivative chain rule

$$\begin{aligned} f(\mathbf{x}) &= \frac{\partial^N C_\theta(F_1(x_1), \ldots, F_N(x_N))}{\partial F_1(x_1) \ldots \partial F_N(x_N)} \prod_i f_i(x_i) \\ &\equiv c_\theta(F_1(x_1), \ldots, F_N(x_N)) \prod_i f_i(x_i), \quad (1) \end{aligned}$$

where $c_\theta(\cdot)$ is called the *copula density*.

Example 2.2: Perhaps the most commonly used is the Gaussian copula [Embrechts et al., 2003]:

$$C_\Sigma(\{F_i(x_i)\}) = \Phi_\Sigma\big(\Phi^{-1}(F_1(x_1)), \ldots, \Phi^{-1}(F_N(x_N))\big), \quad (2)$$

where Φ is the standard normal distribution and Φ_Σ is a zero mean normal distribution with correlation matrix Σ. Figure 1 shows samples from this copula using two different marginals. As can be seen, even with a

[1] We note that an explicit relationship between copulas and Kendall's τ is used in practice to perform parameter estimation for a few additional copula families (e.g., [Genest and Rivest, 1993]). However, computation of τ (cubic in the number of instances) is considerably slower than ρ_s and is typically not faster than an efficient conjugate gradient procedure, as was verified in our implementation.

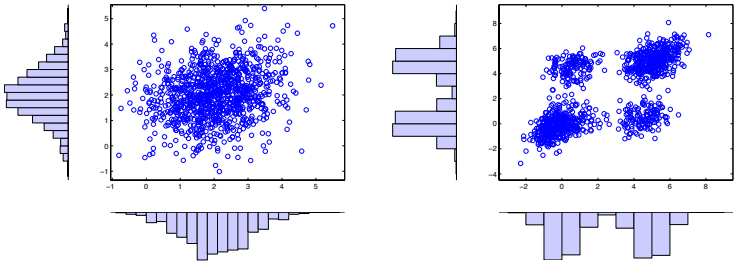

Figure 1: Samples from the bivariate Gaussian copula with correlation $\theta = 0.25$. (left) with unit variance Gaussian marginals; (right) with a mixture of Gaussian marginals.

Normal marginals Mixture of Gaussians marginals

simple elliptical copula, a variety of markedly different and multi-modal distributions can be constructed. More generally, and without any added computational difficulty, we can use different marginals for each variable, and can also mix and match marginals of different forms with *any* copula function.

2.2 Copulas and Spearman's Rho

Spearman's rank correlation coefficient or Spearman's rho is a nonparametric measure of statistical dependence that measures how well the relationship between two variables can be described using a monotonic function. Formally, let $U \equiv F_X(x)$ be the rank (quantile) of X and $V \equiv F_Y(y)$ be the rank of Y. Spearman's rho, denoted by ρ_s, is defined as the standard Pearson's correlation between U and V

$$\rho_s(X,Y) \equiv \frac{cov(U,V)}{\sigma(U)\sigma(V)},$$

where σ denotes the standard deviation. Copula functions, which are aimed at capturing the dependence structure between variables, are closely related to ρ_s (as well as other rank correlation measures). Concretely, if $F_{X,Y}(x,y) = C_\theta(U,V)$ then it can be easily shown (e.g., [Nelsen, 2007]) that

$$\rho_s(X,Y) = \rho_s(C_\theta) \equiv 12\iint C_\theta(U,V)dudv - 3, \qquad (3)$$

where $\rho_s(C_\theta)$ is a notation used to emphasize that ρ_s can be computed directly from the copula function. As can be expected, $\rho_s = 1$ if and only if X and Y exhibit perfect monotonic dependence; $\rho_s = -1$ when X and Y are perfectly negatively correlated.

2.3 Copula Bayesian Networks

We now briefly describe the multivariate density model proposed by Elidan [2010] that fuses the copula and Bayesian networks [Pearl, 1988] formalisms. Let $\mathcal{G}$ be a directed acyclic graph whose nodes correspond to the random variables $\mathcal{X} = \{X_1, \ldots, X_N\}$, and let $\mathbf{Pa}_i = \{\mathbf{Pa}_{i1}, \ldots, \mathbf{Pa}_{ik_i}\}$ be the parents of X_i in $\mathcal{G}$. As for standard BNs, we use $\mathcal{G}$ to encode the independence statements $I(\mathcal{G}) = \{(X_i \perp ND_i \mid \mathbf{Pa}_i)\}$, where $\perp$ denotes the independence relationship, and ND_i are nodes that are not descendants of X_i in $\mathcal{G}$.

Definition 2.3: A Copula Bayesian Network (CBN) is a triplet $\mathcal{C} = (\mathcal{G}, \Theta_C, \Theta_f)$ that defines $f_\mathcal{X}(\mathbf{x})$. $\mathcal{G}$ encodes the independencies $(X_i \perp ND_i \mid \mathbf{Pa}_i)$, assumed to hold in $f_\mathcal{X}(\mathbf{x})$. Θ_C is a set of local copula functions $C_i(F(x_i), F(\mathbf{pa}_{i1}), \ldots, F(\mathbf{pa}_{ik_i}))$ that are associated with the nodes of $\mathcal{G}$ that have at least one parent. In addition, Θ_f is the set of parameters representing the marginal densities $f_i(x_i)$ (and distributions $F_i(x_i)$). The joint density $f_\mathcal{X}(\mathbf{x})$ then takes the form

$$f_\mathcal{X}(\mathbf{x}) = \prod_{i=1}^{N} R_{c_i}\left(F(x_i), F(\mathbf{pa}_{i1}), \ldots, F(\mathbf{pa}_{ik_i})\right) f_i(x_i),$$

where, if X_i has at least one parent in the graph $\mathcal{G}$, the term $R_{c_i}\left(F(x_i), F(\mathbf{pa}_{i1}), \ldots, F(\mathbf{pa}_{ik_i})\right)$ is defined as

$$R_{c_i}(\cdot) \equiv \frac{c_i(F(x_i), F(\mathbf{pa}_{i1}), \ldots, F(\mathbf{pa}_{ik_i}))}{\frac{\partial^K C_i(1, F(\mathbf{pa}_{i1}), \ldots, F(\mathbf{pa}_{ik_i}))}{\partial F(\mathbf{pa}_{i1}) \ldots \partial F(\mathbf{pa}_{ik_i})}}$$

When X_i has no parents in the graph $\mathcal{G}$, $R_{c_i}\left(F(x_i), F(\mathbf{pa}_{i1}), \ldots, F(\mathbf{pa}_{ik_i})\right) \equiv 1$. ∎

The term $R_{c_i}\left(F(x_i), F(\mathbf{pa}_{i1}), \ldots, F(\mathbf{pa}_{ik_i})\right) f_i(x_i)$ is always a valid conditional density $f(x_i \mid \mathbf{pa}_i)$ and can be easily computed. In particular, when the copula density $c(\cdot)$ has an explicit form, so does this term.

Elidan [2010] showed that a CBN defines a valid joint density, and further that the product of local ratio terms R_{c_i} defines a joint copula over $\mathcal{X}$. Thus, like other graphical models, a CBN takes advantage of the independence assumptions to represent $f_\mathcal{X}(\mathbf{x})$ compactly via a product of local terms. Differently from a regular BN, a CBN has an explicit marginal representation. This can result in substantial practical advantages (see Elidan [2010] for more details).

Given a complete dataset $\mathcal{D}$ of M instances where all of the variables $\mathcal{X}$ are observed in each instance, the log-likelihood of the data given a CBN model $\mathcal{C}$ is

$$\ell(\mathcal{D} : \mathcal{C}) \quad = \quad \sum_{m=1}^{M} \sum_{i=1}^{N} \log f(x_i[m]) + \log R_{c_i}[m],$$

where $R_{c_i}[m]$ is a shorthand for the value that the copula ratio $R_{c_i}(\cdot)$ takes in the m'th instance. While this objective appears to fully decompose according to the structure of $\mathcal{G}$, each marginal distribution $F_i(x_i)$ can appear in several local copula terms (of X_i and its children in $\mathcal{G}$). Thus, to facilitate efficient estimation, we adopt the common approach where the marginals are estimated first [Joe and Xu, 1996]. Given $F_i(x_i)$, we can then estimate the parameters of each local copula function *independently* of the others using standard procedures (e.g., conjugate gradient when a closed form solution is not available).

3 Spearman's rho as a Proxy to Expected Likelihood

As discussed, structure learning is computationally demanding since, to evaluate the merit of each of the numerous candidate structures, we need to find the maximum-likelihood parameters and then compute the log-likelihood function given these parameters. For many parametric forms, the parameter estimation stage dominates computations, while in other cases the computation of the log-likelihood function *given* the maximum likelihood parameters can take longer than parameter estimation (e.g., for the linear Gaussian model). Our goal is to bypass the bulk of these computations by using a proxy model selection measure that can be computed efficiently. Ideally, we would like a proxy that is monotonic in the expected maximum log-likelihood function, so that it can be used to accurately rank competing models. In this section we identify such a measure.

The building-block task of structure learning is the evaluation of the merit of an edge $X \rightarrow Y$, independently of other edges (in Section 4 we explain how this is utilized when learning a full structure). The part of the model selection score for a CBN that depends on this edge is

$$\sum_{m=1}^{M} \log c_{\hat{\theta}}(F_X(x[m]), F_Y(y[m])),$$

where $\hat{\theta}$ are the estimated parameters, and the sum is over training instances. When data is generated from the copula, as $M \rightarrow \infty$, the above expression approaches the negative (differential) entropy

$$-H(C_\theta(U,V)) = \int c_\theta(u,v) \log c_\theta(u,v) du dv, \quad (4)$$

where, as before, U, V are the ranks of X, Y, respectively. Thus, if we find an efficient surrogate for the computation of this entropy, we will have a surrogate for the expected log-likelihood of the model.

We now prove that the magnitude of Spearman's rank correlation coefficient $|\rho_s(X,Y)|$ is monotonic in the expected likelihood for two important copula families. We then show numerically that the same relationship holds for many other common copulas.

3.1 The Gaussian Copula

We start by proving the result for the undoubtedly most popular copula, namely the Gaussian copula defined in Eq. (2).

Theorem 3.1 : $-H(C_\theta(U,V))$ is monotonic in $|\rho_s(X,Y)|$ for the bivariate Gaussian copula

Proof: For the Gaussian copula, Eq. (3) has a known explicit form: $\rho_s = \frac{6}{\pi} \sin^{-1} \frac{\theta}{2}$ (see, for example [Genest and Favre, 2007])). Similarly to the case of the standard bivariate Gaussian [Cover and Thomas, 1991], it is easy to show that entropy of the copula is $\frac{1}{2}\log(1-\theta^2) + A$, where A does not depend on θ. The result follows from the fact that the entropy is monotonic in θ^2, and the monotonicity of the absolute value of the sine function for $\theta \in [-1,1]$. ∎

3.2 The Farlie-Gumbel-Morgenstern Copula

The Farlie-Gumbel-Morgenstern (FGM) copula function is defined as

$$C_\theta(u,v) = uv + \theta uv(1-u)(1-v), \quad (5)$$

for $\theta \in [-1,1]$. This family has been widely used despite its limited dependency range due to its analytical simplicity (see [Nelsen, 2007, Joe, 1997] for properties and [Hutchinson and Lai, 1990] for applications).

Theorem 3.2 : $-H(C_\theta(U,V))$ is monotonic in $|\rho_s(X,Y)|$ for the FGM family of copulas

Proof: The FGM family is continuous in θ, and it is easy to show that every copula in the FGM family can be represented as a convex combination of its extreme members $C_{-1}(u,v)$ and $C_1(u,v)$ [Nelsen, 2007]. Further, $C_0(u,v) = uv$ is the independence copula. The entropy is concave in its argument and thus, since it is maximal for $C_0(u,v)$, it must increase monotonically in the range $\theta \in [-1,0]$ and similarly decrease monotonically in the range $\theta \in [0,1]$. It follows that the entropy is monotonic in the absolute value of the copula parameter θ. Now, like most common copulas, the FGM family is also positively ordered so that $C_{\theta_2}(u,v) > C_{\theta_1}(u,v)$ for all u, v whenever $\theta_2 > \theta_1$ [Nelsen, 2007, Joe, 1997]. Our result follows from the fact that ρ_s grows monotonically with $C(\cdot)$ (see Eq. (3)) so that the absolute value of ρ_s increases with θ for $\theta > 0$ and decreases with θ for $\theta < 0$. ∎

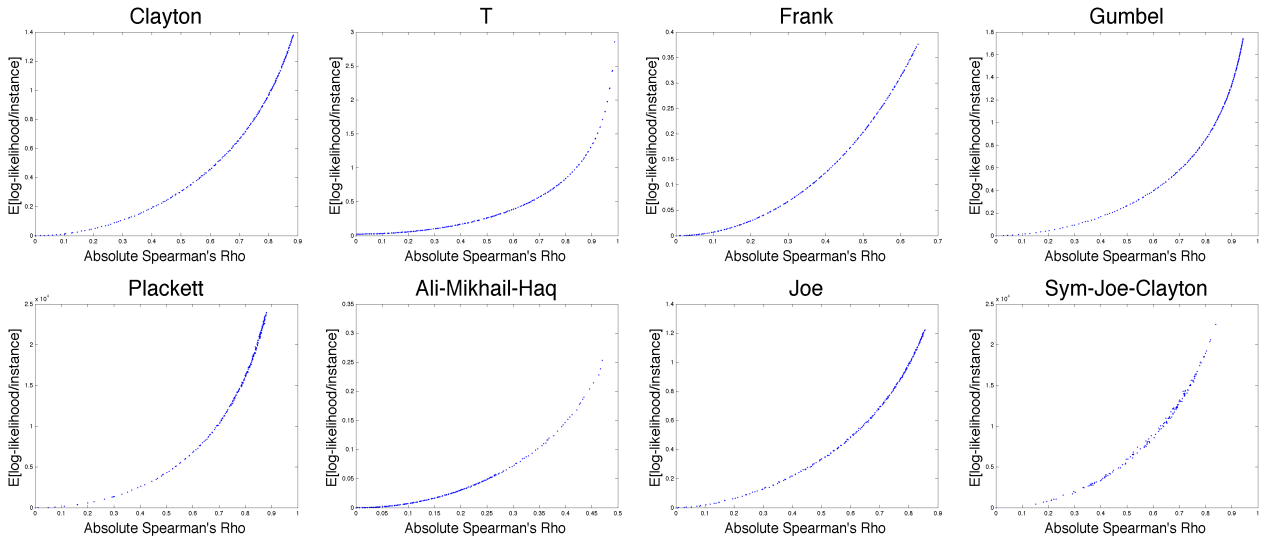

Figure 2: Numerical experiments demonstrating the monotonic relationship between the absolute value of Spearman's rho (x-axis) and the expected likelihood (y-axis) for eight varied copula families. For each copula, using 25000 samples, shown is the empirical log-likelihood function vs. Spearman's rho for 500 parameter values. The T copula shown has 5 degrees of freedom (results were similar for values from 2 to 20).

3.3 Additional Copula Families

Deriving an explicit expression for the copula entropy is difficult and, aside from the Gaussian copula, we are not aware of any known analytic forms. We can, however, easily evaluate the entropy numerically for different parameter values, and assess whether monotonicity in $|\rho_s|$ holds for additional copula families.

In addition to the Gaussian and FGM copulas, we consider seven popular single parameter copulas, as well as a two parameter family (symmetric Joe-Clayton). Properties of all of these copula families can be found in Joe [1997], Nelsen [2007], Patton [2006]. Note that an explicit relationship between ρ_s and the copula parameter θ is known only for the Plackett family.

For each family, for 500 different parameter values, we generated 25000 samples from the copula distribution. For each set of samples (representing the true distribution), we then computed the log-likelihood function and the absolute value of Spearman's rho. The plots in Figure 2 compare these two measures for the different copula families. A monotonic relationship between the absolute value of Spearman's rho and the expected log-likelihood is evident in all cases.

Given the above theoretical and numerical results, it seems likely that a common property of the copulas considered underlies the monotonicity of the entropy in $|\rho_s|$. We leave the elusive identification of sufficient and/or necessary conditions for future work, and only briefly discuss possible commonalities. All single parameter families are symmetric and define a concordance ordering, i.e., the copula function is monotonic in the dependence parameter. For the symmetric Joe-Clayton family, concordance ordering is conjectured for $\theta > 1$ and is known to hold otherwise. It would indeed be remarkable if concordance ordering implies the monotonicity relationship since most known copulas are symmetric and define a concordance ordering (including all B1-B12 single parameter families in Joe [1997]) . Another property shared by the single parameter families we considered is that the copula distribution function is Schur-concave [Durante and Sempi, 2003]. Although the implications on the copula density are currently unknown, majorization theory may provide the tools needed to prove the monotonicity relationship (see Marshall and Olkin [1979], and Joe [1987] for a generalization for densities) . It is unknown whether the symmetric Joe-Clayton copula is also Schur-concave, and we are not aware of any other commonality between the copulas considered.

4 Lightning-speed Structure Learning

We now briefly describe how the results of Section 3 can be used to learn the structure of a CBN model.

Learning A Tree-structured Network

When the structure $\mathcal{G}$ is constrained so that each random variable has at most one parent, the local copula ratio $R_{c_i}\left(F(x_i), F(\mathbf{pa}_{i1}), \ldots, F(\mathbf{pa}_{ik_i})\right)$ in Eq. (4) reduces to the bivariate copula density $c_i(F(x_i), F(x_j))$, where X_j is the parent of X_i. In this case the log-likelihood function decomposes into pairwise terms.

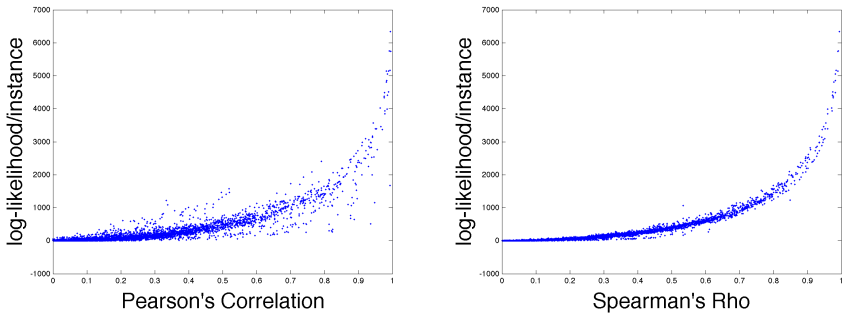

Figure 3: The empirical maximum-likelihood function for pairs of variables in the **Crime** dataset vs. Pearson's correlation (left) and Spearman's rho (right).

Thus, similarly to learning a tree structured BN [Chow and Liu, 1968], given the benefit of each edge independently, the optimal structure can be learned efficiently using a maximum spanning tree algorithm.

The above still requires that we perform costly computations for all $N(N-1)/2$ edges. However, since only bivariate copulas are involved, the developments of Section 3 are applicable directly. That is, we simply replace the pairwise likelihood component with an efficient empirical computation of $\rho_s(X_i, X_j)$. As our experimental results show, despite the fact that training data in practice is finite and is not necessarily generated from the specific copula used, ρ_s serves as an extremely accurate proxy to the merit each edge.

Learning More Complex Structures

More generally, structure learning is computationally difficult [Chickering, 1996], and we typically resort to a greedy search procedure that involves local modifications to the structure (e.g., add/delete/reverse an edge), see Koller and Friedman [2009] for details. In this case, a variable X_i can have more than a single parent and our results do not apply directly.

Yet, intuitively, variables that are highly correlated with X_i (as measured via Spearman's rho) are likely candidates as additional parents. Thus, similarly to the two-stage approach used by Della Pietra et al. [1997] and Friedman et al. [1999], we use ρ_s to crudely yet efficiently pre-rank candidate structure modifications. We then perform exact costly computation of the score only for the K most promising candidates. In our experiments we set $K = 2$ for all domains.

5 Experimental Evaluation

To asses the benefit of our approach for learning the structure of CBNs, we compare the performance and running time to learning CBNs and BNs using a standard approach. For the linear Gaussian BN, we use a closed form estimator; for the nonlinear sigmoid BN, we use a conjugate gradient procedure. For the Gaussian CBN, we perform baseline parameter estimation using Spearman's rho evaluation (e.g., Genest and Favre [2007]). In this case, our method only differs from the baseline in that we bypass the explicit computation of the log-likelihood function *given* the maximum likelihood parameters. We also learn a CBN with the Clayton copula, a representative of the Archimedean family of copulas for which an explicit relationship between Spearman's rho and the copula parameters is not known [Nelsen, 2007]. For this family, we perform baseline parameter estimation using a conjugate gradient procedure (this is faster in practice than estimation using Kendall's τ). For the univariate marginals in both CBN models, we use the standard kernel-based approach [Parzen, 1962] with the common Gaussian kernel (see, for example, [Bowman and Azzalini, 1997] for details). For all models, the network structure was learned using the same search procedure. In all cases, the Bayesian Information Criterion (BIC) of Schwarz [1978] was used to penalize the log-likelihood for the complexity of the model.

We consider three datasets of a markedly different nature and dimensionality:

- **Wine Quality** (UCI repository). 1599 measurements of 11 physiochemical properties and a quality variable of red "Vinho Verde" [Cortez et al., 2009].
- **Dow Jones**. 1508 daily adjusted changes (2001-2005) of the 30 index stocks. To avoid arbitrary imputation, two stocks not traded in all of these days were excluded (KFT,TRV).
- **Crime** (UCI repository). 100 observed variables relating to crime ranging from household size to fraction of children born outside of a marriage, for 1994 communities across the U.S.

Spearman's Rho vs. The Log-likelihood

Our results of Section 3 only apply asymptotically and when the data is generated from a copula. Thus, we begin by examining the real-life relationship between the absolute value of Spearman's rho and the likelihood of the model for the Gaussian copula (results were essentially the same for the Clayton copula). In Figure 3 (right) we empirically compare the two measures for all pairs of variables in the 100 variable **Crime** domain. We also show results for the standard Pearson's correlation coefficient (left).

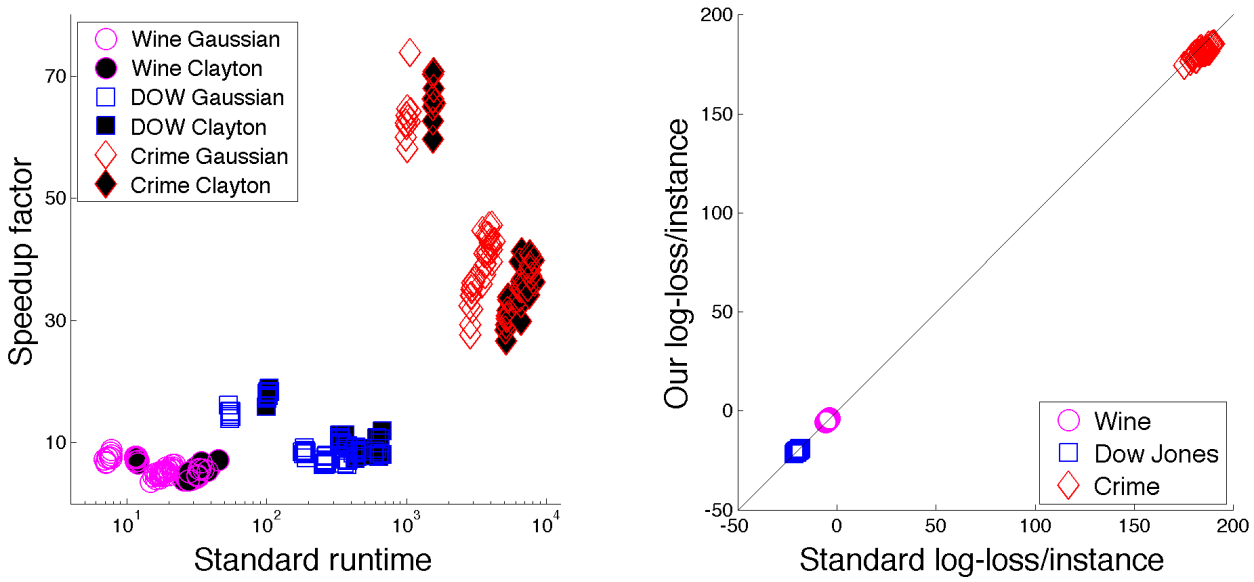

Figure 4: (left) The speedup of our approach (y-axis) as a function of the standard learning time in seconds of a CBN (x-axis) for two copula families and structures with up to 1,2,3, and 4 parents (the tree structures with at most one parent that are learned faster form the upper-left group for each domain). For each setting, shown are results for 10 random test/train partitions. (right) comparison of test performance of the model learned using our procedure (y-axis) vs. standard learning (x-axis), when allowing up to 4 parents for each variable.

It is clear that while the empirical relationship is not perfectly monotonic, Spearman's rho serves as a surprisingly accurate proxy to the log-likelihood function, and is thus extremely effective at correctly ranking two competing edges. As we shall see in the evaluation below, this translates nicely into accurate performance when learning a full CBN model. One may suspect that an effective proxy could also be found in the form of Pearson's correlation. However, as can be seen in Figure 3(left), this is not the case: while generally reasonable, Pearson's correlation can be very high even when the log-likelihood is quite small.

Standard vs. Lightning-speed Learning of Copula Bayesian Networks

We start by considering the speedup factor of our method when learning CBNs of different complexities for two different choices of the local copulas in the model: the standard Gaussian copula and the Clayton copula. Figure 4 (left) shows the the speedup factor of our method as a function of the standard learning time for 10 random train/test partitions for the three domains described above. For each domain and random repetitions, we learned networks with at most 1,2,3 and 4 parents so as to cover a range of network complexities. The upper left "cluster" for each domain corresponds to tree structured networks where learning can be carried out most efficiently. It is easy to see that the speedup of our method is substantial and, as expected, is greatest when learning trees. Importantly, the advantage grows with the number of variables in the domain. For the more complex **Crime** domain, our model offers a speedup of close to two orders of magnitude when learning trees, and a speedup factor of over 30 when learning more complex structures.

To ensure that the speedup of our method does not come with a degradation in performance, Figure 4 (right) compares the test performance our lightning-speed procedure to that of the standard baseline for the Gaussian CBN with up to 4 parents (results were similar for the Clayton copula, and even more accurate for simpler structures). As is clearly evident, we suffer no degradation in performance for the two smaller domains, and only a negligible one for the crime domain. Put together, the runtime and performance results show that our approach offers an appealing lightning-speed alternative to standard learning of CBNs.

Comparison to Learning of Regular BNs

We now compare the performance of our approach to learning a simple linear Gaussian BN, where each variable is normally distributed around a linear combination of its parents $X_i \sim N(\beta_0 + \beta^{\mathbf{T}}\mathbf{pa_i}, \sigma_i)$. We also considered a nonlinear sigmoid BN where $X_i \sim N(\alpha_0 + \alpha_1 \frac{1}{1+e^{\beta_0+\beta^T \mathbf{pa}_i}}, \sigma_i)$. Learning this latter model was substantially slower than learning CBNs (2-3 orders of magnitude) and it performed worse. Thus, for clarity of exposition, we do not report its results.

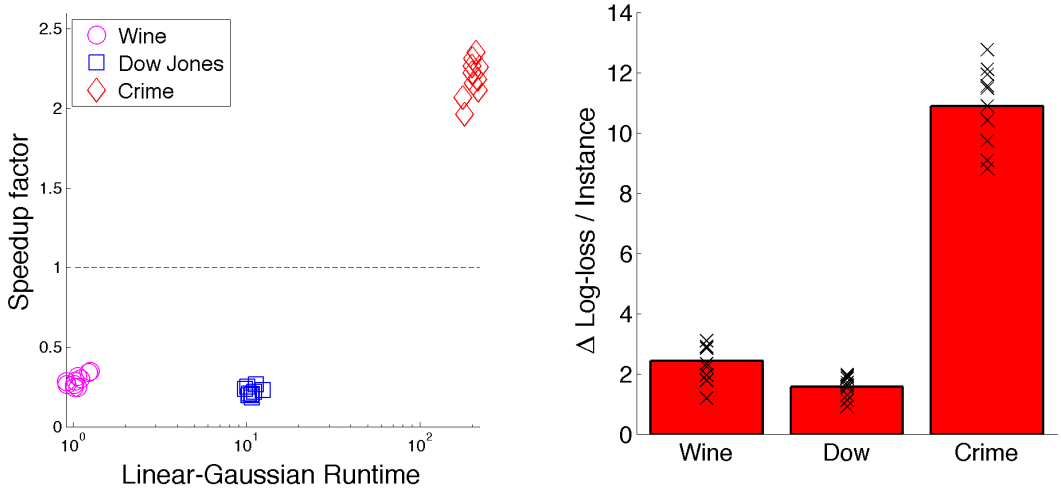

Figure 5: Comparison of our approach to learning a standard linear Gaussian BN with up to 4 parents for each variable. (left) our speedup as a function of the baseline running time. (right) the average (bar) and 10 repetitions (black 'x') test set improvement over the baseline in bits/instances. Results for the sigmoid BN were substantially worse than ours and took 2-3 orders of magnitude longer to learn, and are omitted for clarity.

Figure 5 (left) shows the relative running time of our lightning-speed procedure when learning the structure of a CBN (y-axis) as a function of the learning time of a linear Gaussian BN. For clarity, the result shown is only for the more complex structures where up to four parents for each variable are allowed during structure search (for simpler structures our speedup advantage is greater). While our procedure is still somewhat slower than the baseline BN in the smaller domains, we are able to learn the structure of a complex CBN model for the 100 variable crime domain in about half the time that it takes to learn the simple linear Gaussian BN model. Figure 5 (right) shows the average log-loss improvement of the CBN model on test data relative to the BN baseline. Note that the scale of improvement is in bits/instances so that the performance gains are quite dramatic, particularly for the more complex **Crime** domain. Thus, for sufficiently complex domains, we are able to substantially improve on the performance of the simple BN model, while remarkably requiring less computational resources.

6 Conclusions and Future Work

In this work we tackled the computationally intensive task of learning the structure of high-dimensional continuous densities using the Copula Bayesian Network (CBN) model. We proved that the expected likelihood of an edge in the model is monotonic in the magnitude of Spearman's rho for two important copula families, and showed numerically that this relationship also holds for many other popular copulas. Motivated by this result, we proposed using the empirical Spearman's rho as a model selection measure. Using our approach for learning the structure of three varied continuous real-life domains, we demonstrated dramatic running time improvements.

Our contribution is twofold. Theoretically, we shed light on the relationship between Spearman's rho correlation measure and the predictive quality of a distribution defined via a copula function. Practically, the highlight of our result is that we are able to learn a complex density that generalizes well without taking any longer to learn than the simplest continuous BN model. This opens the door for effective scaling-up of structure learning in complex continuous domains.

Liu et al. [2010] proposed an approach for learning a high-dimensional nonlinear model that is undirected and is restricted to the Gaussian copula. Technically, they do not define a concrete density so that direct performance comparison to our work requires some adaptation. At a higher level, their work does not aim to use an efficient proxy for estimation but rather focuses on provably consistent estimation and is thus, by construction, slower. Even more broadly, as in the case of Bayesian vs. Markov networks, the undirected and directed representations complement each other theoretically, and are likely to do so in practice.

In future work we plan to tackle two challenges. First, building on the discussion at the end of Section 3.3, we aim to theoretically identify the general necessary and/or sufficient conditions needed to ensure monotonicity of the entropy in the magnitude of Spearman's rho. Second, it would be both interesting and useful to replace the heuristic used when allowing for multiple parents in $\mathcal{G}$ with a theoretically founded approach.

Acknowledgements

The work was support by a Google Research Award. G. Elidan was supported by an Alon fellowship. I thank Amir Globerson and Yair Weiss for their valuable comments on a draft of this work.

References

A. Bowman and A. Azzalini. *Applied Smoothing Techniques for Data Analysis*. Oxford University Press, 1997.

D. M. Chickering. Learning Bayesian networks is NP-complete. In D. Fisher and H. J. Lenz, editors, *Learning from Data: Artificial Intelligence and Statistics V*, pages 121–130. Springer-Verlag, New York, 1996.

C. K. Chow and C. N. Liu. Approximating discrete probability distributions with dependence trees. *IEEE Trans. on Info. Theory*, 14:462–467, 1968.

P. Cortez, A. Cerdeira, F. Almeida, T. Matos, and J. Reis. Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems*, 47(4):547–553, 2009.

T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley & Sons, New York, 1991.

S. Della Pietra, V. Della Pietra, and J. Lafferty. Inducing features of random fields. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 19(4): 380–393, 1997.

F. Durante and C. Sempi. Copula and schur-concavity. *International Mathematics Journal*, 2003.

G. Elidan. Copula bayesian networks. In *Neural Information Processing Systems (NIPS)*, 2010.

P. Embrechts, F. Lindskog, and A. McNeil. Modeling dependence with copulas and applications to risk management. *Handbook of Heavy Tailed Distributions in Finance*, 2003.

N. Friedman, I. Nachman, and D. Pe'er. Learning bayesian network structure from massive datasets: The 'sparse candidate" algorithm. In *Proc. Fifthteenth Conference on Uncertainty in Artificial Intelligence (UAI '99)*, pages 206–215, 1999.

C. Genest and A. Favre. Everything you always wanted to know about copula modeling but were afraid to ask. *Journal of Hydrologic Engineering*, 12:347, 2007.

C. Genest and L. Rivest. Statistical inference procedures for bivariate archimedean copulas. *Journal of the American Statistical Association*, pages 1034–1043, 1993.

J. Goldberger and A. Leshem. Mimo detection for high-order qam based on a gaussian tree approximation. *IEEE Trans. Information Theory*, 57:4973–82, 2011.

T. Hutchinson and C. Lai. *Continuous Bivariate Distributions, Emphasizing Applications*. Rumsby Scientific Publishing, 1990.

H. Joe. Majorization, randomness and dependence for mutivariate distributions. *The Annals of Probability*, 15(3):1217–1225, 1987.

H. Joe. Multivariate models and dependence concepts. *Monographs on Statistics and Applied Probability*, 73, 1997.

H. Joe and J. Xu. The estimation method of inference functions for margins for multivariate models. Technical Report 166, Department of Statistics, University of British Columbia, 1996.

D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques*. The MIT Press, 2009.

H. Liu, J. Lafferty, and L. Wasserman. The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *Journal of Machine Learning Research*, 2010.

A. Marshall and I. Olkin. *Inequalities: Theory of Majorization and Its Applications*. Academic Press, 1979.

R. Nelsen. *An Introduction to Copulas*. Springer, 2007.

E. Parzen. On estimation of a probability density function and mode. *Annals of Math. Statistics*, 33:1065–1076, 1962.

A. Patton. Modelling asymmetric exchange rate dependence. *International Economic Review*, 47(2): 527–556, 2006.

J. Pearl. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, San Francisco, California, 1988.

G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6:461–464, 1978.

A. Sklar. Fonctions de repartition a n dimensions et leurs marges. *Publications de l'Institut de Statistique de L'Universite de Paris*, 8:229–231, 1959.