entropy

MDPI

Article

Learning Continuous Decomposable Models Using Mutual Information and Statistical Copulas

Luiz Desuó Neto [1,*], Henrique de Oliveira Caetano [2], Matheus de Souza Sant'Anna Fogliatto [2] and Carlos Dias Maciel [1]

[1] Department of Electrical Engineering, São Paulo State University (UNESP), Guaratinguetá 12516-410, SP, Brazil; carlos.maciel@unesp.br
[2] Department of Electrical and Computer Engineering, University of São Paulo (USP), São Carlos 13566-590, SP, Brazil; henriquecaetano1@usp.br (H.d.O.C.); matheusfogliatto@usp.br (M.d.S.S.F.)
* Correspondence: luizdesuo@gmail.com

Abstract

Learning dependence graphs from multivariate continuous data is challenging when marginal distributions are heterogeneous, since likelihood-based nonparametric scores can be sensitive to smoothing choices and can confound marginal irregularities, including non-identifiability, with dependence. This work studies structure learning in the class of decomposable (chordal) Markov random fields, where junction tree factorizations enable tractable inference and local score updates. Our first contribution is a theoretical result showing that, under decomposability, mutual information can be expressed as a difference of clique/separator copula entropies, yielding a dependence-only decomposition aligned with the clique/separator structure. Building on this identity, we define an information-theoretic objective for decomposable graphs with a complexity penalty that preserves clique/separator additivity, and we derive closed-form local score differences for chordality-preserving single-edge insertions and deletions. To make the score computable from data, we instantiate clique/separator copula entropies using pseudo-observations and a probit-transformed kernel density estimator with predictive log score evaluation to mitigate boundary effects on the unit hypercube. The resulting nonparametric greedy procedure improves edge recovery accuracy on synthetic chordal benchmarks compared with a likelihood-driven nonparametric baseline, and it produces interpretable dependence summaries on an airway epithelial gene expression dataset. Concretely, this paper contributes (1) a decomposable mutual information identity via clique/separator copula entropies, (2) a copula information score with an additive complexity penalty for decomposable graphs, (3) a closed-form local score, enabling greedy chordal add or delete search, (4) a practical nonparametric copula entropy estimation pipeline, and (5) empirical gains on synthetic and real data.

Keywords: decomposable models; structure learning; mutual information; statistical copulas

Check for updates

Academic Editor: Nikolai Leonenko

Received: 10 January 2026
Revised: 28 February 2026
Accepted: 1 March 2026
Published: 4 March 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Dependence structure learning is increasingly used to convert high-dimensional continuous measurements into interpretable interaction networks that support downstream inference, hypothesis generation, and decision making [1]. In this context, both undirected MRFs and BNs are routinely used as structured representations of multivariate dependence, trading off interpretability, computational tractability, and (in the directed case) a

natural language for causal hypotheses [2]. However, for continuous, heterogeneous data, one recurring difficulty is to combine a dependence score that is robust and marginally invariant with a graph class that supports scalable exact inference and local search updates without committing to a full directed learning pipeline [2]. In many scientific applications, this task must balance statistical fidelity with computational tractability because unconstrained structure learning is combinatorial, while the datasets motivating it are often large and heterogeneous [1].

Generally, calculating the partition function (normalizing factor that reestablishes the unit measure assumption from Kolmogorov's probability axioms) of MRFs is an intractable problem [3]. Given the substantial computational burden, often exceeding feasibility, the cost-effectiveness of this strategy remains unclear. Without independent statistical interpretation for Gibbs potentials, attributing meaning to their parameters presents a significant challenge [4]. In contrast, the statistical meaning arises only from the joint probability density function, which is composed of their products. In light of the computational challenges and difficulty in interpreting the local parameters of Gibbs potentials, decomposable networks emerge as a relevant alternative due to their inherent advantages. For this reason, addressing non-chordal structure learning is intentionally left to work outside the scope of this work.

Although decomposable models admit both undirected and directed representations [4], this paper focuses on learning the undirected decomposable structure. This emphasis is deliberate; chordality yields a junction tree representation and clique/separator factorizations that enable scalable exact inference and genuinely local score updates during structure searching. BN learning (i.e., orienting edges and enforcing acyclicity, followed by a dedicated DAG search) can be treated as a complementary downstream step once a tractable, decomposable dependence scaffold has been learned. We do not develop that step here and instead refer to existing work on relating BNs and decomposable Markov networks [2,5].

A particularly attractive compromise is to restrict attention to decomposable (chordal) MRFs, since chordality yields a junction tree representation and enables clique/separator factorizations with efficient local computations [6]. This decomposable restriction has long served as a pragmatic foundation for scalable modeling, and it continues to motivate algorithmic research on both optimization and sampling over chordal graph spaces [6]. The work proposed by Olsson et al. [6] reinforced this perspective by developing explicit junction tree-based samplers, such as sequential sampling schemes that operate directly on junction tree structures to explore distributions over decomposable graphs. In parallel, recoverability analyses for tree-structured Markov random fields have clarified how noise and finite-sample effects can fundamentally limit structure identification even in low-treewidth settings [7].

However, these advances do not yet resolve the practical problem of structure recovery for continuous decomposable models under heterogeneous marginals. Junction tree samplers enable global exploration and uncertainty quantification over decomposable structures, but they do not by themselves provide a default, computationally dominant score for high-dimensional continuous data, where repeated clique/separator evaluations must be both stable and marginally robust [6]. At the same time, the sharpest modern identifiability and sample complexity results remain concentrated in tree-structured settings, leaving a gap in understanding how estimation error in local statistical primitives propagates through chordality-constrained search when the separator size exceeds one [7,8]. Finally, while the benchmarking infrastructure for structure learning is improving, end-to-end, decomposable graph recovery benchmarks for continuous regimes remain comparatively

under-standardized, complicating reproducibility and fair comparison across score and search choices [1].

Despite recent progress in chordality-preserving search, learning continuous decomposable models remains fragile in practice because many widely used scores conflate dependence with one-dimensional marginal features and are sensitive to smoothing and boundary effects [9]. In particular, likelihood-driven nonparametric decomposable models based on clique-wise kernel density estimation avoid discretization but can become unstable under a bandwidth choice and can explain marginal irregularities by adding edges, which degrades structural interpretability [9].

Conversely, exact or near-exact combinatorial approaches for decomposable structure learning have been developed mainly for discrete formulations, where discretization or coarsening and state space growth can dominate the error budget when the underlying variables are genuinely continuous [10]. Many modern information-theoretic scoring ideas for continuous data have been developed primarily in the BN setting, where acyclicity and Markov equivalence shape both the score and the search space [2]. As a consequence, even when the target family is decomposable, it is nontrivial to obtain an undirected score that is simultaneously marginally invariant and junction tree-local, i.e., it admits clique/separator decomposition and inexpensive updates under chordality-preserving edits [9].

Moreover, on the BN side, widely used single-link look-ahead searches can fail in pseudo-independent domains, while multi-link look-ahead remedies are often prohibitively expensive because the DAG search space is far larger than the chordal and decomposable graph space, motivating intermediate-step strategies that first learn a decomposable MRF to guide BN learning at reduced complexity [2,5]. Together, these issues motivate a central open problem: to construct a dependence-oriented scoring criterion for continuous decomposable MRFs that is invariant to strictly monotone marginal reparameterizations, decomposes into clique/separator contributions to enable local score updates under chordality-preserving moves, and can be estimated robustly enough that repeated local comparisons do not translate smoothing artifacts into persistent false positives [2].

Copula theory offers a direct route to marginal invariance by separating marginal behavior from dependence through Sklar's representation [11]. In particular, mutual information can be written as (negative) copula entropy, implying that a dependence score can be defined entirely with the copula density rather than the original marginal densities [12]. This identity has already been exploited in continuous structure learning for copula BNs via information-theoretic criteria and BIC-like penalties, suggesting that information-based scoring can be both practical and statistically targeted [13]. However, transferring this idea to decomposable undirected models requires a score that respects junction tree locality and admits efficient updates under chordality-preserving edge edits [9].

This paper addresses that gap by developing a copula information structure learning framework for continuous decomposable models that combines information theory, copula-based invariance, and chordal graph locality [12]. The proposed approach (1) rewrites the objective in terms of copula entropy contributions defined on clique and separator marginals, (2) introduces a CBIC-type complexity penalty that preserves clique/separator additivity, and (3) optimizes the resulting criterion via greedy chordality-preserving searching on decomposable graphs [13]. To support repeated evaluation of high-dimensional copula entropy terms, the method adopts a nonparametric copula density estimation strategy based on pseudo-observations and a probit transformation, which is designed to mitigate boundary bias on multidimensional unit hypercubes [14]. Algorithmically, chordality is maintained throughout the search using junction tree representations and dynamic chordal graph criteria so that each accepted move remains within the decomposable model class [15].

The significance of this contribution is twofold. It yields a dependence-focused learning criterion that is invariant to monotone marginal transformations, and it remains computationally aligned with decomposable graph search through local clique/separator updates [12]. Empirically, the resulting method is evaluated on controlled synthetic experiments (with a known chordal ground truth) and on a gene expression dataset, illustrating both improved structure recovery and practical applicability in noisy continuous measurements [1]. The approach taken in this work is to first establish a decomposable model identity that expresses mutual information through clique/separator copula entropies (Theorem 1) and then to use that identity to design a CBIC-type objective and a chordality-preserving greedy optimizer whose score differences depend only on the small set of affected clique/separator margins. We focus on learning the undirected decomposable structure; constructing an explicit BN from the learned chordal graph is complementary and is not pursued here (see [2,5]). The main contributions are as follows:

- Theoretical results: We prove a decomposable model identity showing that mutual information can be written as a difference of clique and separator copula entropies (Theorem 1), yielding an information decomposition aligned with a junction tree structure.
- New score design for decomposable graphs: Using Theorem 1, we define an information-theoretic objective for continuous decomposable models with a CBIC-type complexity penalty that preserves clique/separator additivity (Section 3.1), adapting the mutual information CBIC motivation from copula BNs [13] to the decomposable undirected settings.
- Local score delta for greedy chordal search: We derive a closed-form local update inspired by [9,15] for single-edge insertions and deletions under the proposed score (Section 3.1), adapting junction tree locality arguments from log-likelihood chordal and decomposable local search to the new copula information objective.
- Nonparametric estimation pipeline: For repeated evaluation of clique and separator terms, we instantiate the score using pseudo-observations and a probit-transformed, KDE-based copula entropy estimator with predictive log score evaluation (Section 3.3), leveraging the established copula entropy and boundary correction methodology [14,16].
- Empirical findings: On chordal synthetic benchmarks and a gene expression application, the resulting NG-CBIC procedure improves structure recovery (higher F1 and lower SHD) relative to likelihood-based nonparametric baselines under the same chordality-preserving search framework (Section 5).

The remainder of this paper is organized as follows. Section 2 develops the theoretical foundations, culminating in Theorem 1, which shows how mutual information can be expressed through clique/separator copula entropies in decomposable models. Section 3 turns this decomposition into a practical learning pipeline by defining the CBIC-type copula information score, deriving local score updates for chordality-preserving edge edits and specifying the nonparametric copula entropy estimator and greedy search procedure. Section 4 describes the experimental protocol and evaluation metrics, Section 5 reports the results on synthetic chordal benchmarks and a gene expression case study, and Section 6 concludes with limitations and directions for future work.

2. Theoretical Foundations

This section fixes the main probabilistic and graphical preliminaries used throughout the paper and clarifies how dependence is represented both analytically (via copulas) and structurally (via decomposable graphs). The guiding theme is that copulas separate marginal behavior from dependence, while decomposable (chordal) graphs make global dependence constraints tractable through clique/separator decompositions. These two

perspectives will later be combined to produce information-theoretic measures that can be evaluated and updated locally during structure search.

2.1. Mutual Information as Statistical Copula Entropy

This subsection recalls the copula decomposition of a continuous multivariate distribution and uses it to rewrite mutual information in a form that depends only on the dependence structure. Under Sklar's theorem, the joint law can be expressed through its marginals and a copula, isolating the part of the distribution that encodes rank dependence from the part that encodes one-dimensional behavior. This separation makes it natural to reinterpret information-theoretic quantities—originally defined via densities on $\mathbb{R}^d$ or other continuous supports—in terms of the copula density on $[0,1]^d$.

Let $(\Omega, \mathscr{A}, \mathbb{P})$ be a probability space, and let $\mathbf{X} = (X_1, \ldots, X_d)^{\mathrm{T}}$ be a continuous random vector defined on this space with values in $\mathbb{R}^d$. We denote its joint CDF by

$$F_{\mathbf{X}}(x_1, \ldots, x_d) \coloneqq \mathbb{P}(X_1 \leq x_1, \ldots, X_d \leq x_d),$$

which we abbreviate as $F(\mathbf{x})$ for $\mathbf{x} = (x_1, \ldots, x_d)^{\mathrm{T}} \in \mathbb{R}^d$. Assume that F is absolutely continuous with respect to the Lebesgue measure λ^d and that its joint density $p_{\mathbf{X}} \colon \mathbb{R}^d \to \mathbb{R}_+$ satisfies

$$p_{\mathbf{X}}(\mathbf{x}) = \frac{\partial^d F(\mathbf{x})}{\partial x_1 \cdots \partial x_d}.$$

For each component $i \in \{1, \ldots, d\}$, we denote the marginal distribution function by $F_i(x_i) \coloneqq \mathbb{P}(X_i \leq x_i)$, with the corresponding marginal density $p_i(x_i) = F_i'(x_i)$.

According to Sklar's theorem [11], there exists a copula function $C \colon [0,1]^d \to [0,1]$ such that for all $\mathbf{x} \in \mathbb{R}^d$, we have

$$F_{\mathbf{X}}(x_1, \ldots, x_d) = C(F_1(x_1), \ldots, F_d(x_d)).$$

If the marginals F_i are continuous, then the copula C is unique. Furthermore, if C is absolutely continuous with respect to the Lebesgue measure λ^d on $[0,1]^d$, then it admits a copula density $c \colon [0,1]^d \to \mathbb{R}_+$ defined by

$$c(u_1, \ldots, u_d) = \frac{\partial^d C(u_1, \ldots, u_d)}{\partial u_1 \cdots \partial u_d}.$$

Under these assumptions, the joint density $p_{\mathbf{X}}$ factorizes as follows:

$$p_{\mathbf{X}}(x_1, \ldots, x_d) = c(F_1(x_1), \ldots, F_d(x_d)) \prod_{i=1}^{d} p_i(x_i). \tag{1}$$

The differential entropy of X_i and of the joint vector $\mathbf{X}$ are

$$h(X_i) = -\,\mathbb{E}[\log(p_i(X_i))], \qquad h(\mathbf{X}) = -\,\mathbb{E}[\log(p_{\mathbf{X}}(\mathbf{X}))].$$

Now, we define the distributional transform $U_i = F_i(X_i)$ for each $i \in \{1, \ldots, d\}$. Each U_i is uniformly distributed on $[0,1]$, and the transformed random vector $\mathbf{U} = (U_1, \ldots, U_d)^{\mathrm{T}}$ has a joint density $c(\mathbf{u}) \equiv c(u_1, \ldots, u_d)$. The copula entropy is then defined by

$$h_c(\mathbf{U}) \coloneqq -\,\mathbb{E}[\log(c(\mathbf{U}))].$$

Finally, following Ma and Sun [12], if $h(X_i)$ and $h(\mathbf{X})$ are finite, then the mutual information of $\mathbf{X}$ can be expressed equivalently as follows:

$$I(\mathbf{X}) = \sum_{i=1}^{d} h(X_i) - h(\mathbf{X}) = -h_c(\mathbf{U}). \tag{2}$$

This identity demonstrates that the mutual information between continuous random variables equals the negative entropy of their associated copula density, thereby directly linking the dependence structure (via the copula) to information-theoretic principles. As a result, estimating mutual information reduces to estimating the (negative) entropy of the copula density. This perspective is particularly valuable for structure learning because it isolates conditional dependence patterns while remaining invariant to marginal artifacts.

2.2. Decomposable Models

This subsection reviews decomposable (chordal) undirected models as a structural restriction that enables exact clique/separator factorizations and efficient inference. The central point is that chordality is equivalent to the existence of a junction tree (clique tree) representation, in which global consistency is enforced through separator intersections (running intersection). This representation will later serve as the algorithmic scaffold for both scoring and admissible local graph edits.

PGMs describe how a joint distribution factorizes according to an underlying structure encoding conditional (in)dependencies. BNs are associated with DAGs, while MRFs use UGs as their representation. DMs occupy the intersection of these two classes; they are probability distributions that admit a perfect map in a chordal (triangulated) graph and hence can be represented both as DAGs and as undirected graphical models [4].

Let $G = (V, E)$ be an undirected graph with a vertex set $V = \{v_1, \ldots, v_d\}$ and an edge set $E \subseteq V \times V$, where each vertex $v_i \in V$ is associated with the component i of the random vector $\mathbf{X} = (X_1, \ldots, X_d)^{\mathrm{T}}$. We will abbreviate sub-vectors using coordinate projections. For the d-dimensional random vector $\mathbf{X}$ and index set $A \subset \{1, \ldots, d\}$, we define the projection $\pi_A : \mathbb{R}^d \to \mathbb{R}^{|A|}$ by $\pi_A(x) := (x_i)_{i \in A}^{\mathrm{T}}$ and write $\mathbf{X}_A := \pi_A(\mathbf{X})$ and $\mathbf{X}_{-A} := \mathbf{X}_{A^{\mathrm{c}}}$. The absence of an edge $\{v_i, v_j\} \notin E$ corresponds to the pairwise Markov statement $X_i \perp\!\!\!\perp X_j \mid \mathbf{X}_{-\{i,j\}}$.

A subset $K \subset V$ is called a clique if every pair of vertices in K is adjacent, and it is maximal if it is not strictly contained in any larger clique. Let $\mathcal{K} \subset 2^V$ denote the collection of all maximal cliques of G. An undirected graph G is called chordal (or triangulated) if every cycle with a length of at least four has a chord, i.e., an edge joining two nonconsecutive vertices on the cycle. Moreover, G is decomposable if and only if it is chordal, and this is equivalent to the maximal cliques in the family $\mathcal{K}$ satisfying the running intersection property. Specifically, G is decomposable if and only if there exists an ordering $(K_1, \ldots, K_m)$ of the cliques in $\mathcal{K}$ such that for every $l \in \{2, \ldots, m\}$, there exists an index $\eta(l) < l$ (the predecessor of K_l) with

$$S_l \equiv K_l \cap \bigcup_{k=1}^{l-1} K_k = K_l \cap K_{\eta(l)}. \tag{3}$$

In other words, the intersection of each clique K_l with all previously listed cliques is contained in a single predecessor clique $K_{\eta(l)}$. The sets S_l are called the separator sets, with the convention $S_1 := \emptyset$. For a decomposable graph G, such an ordering of maximal cliques exists and induces a junction tree representation of G.

Assume now that G is decomposable and encodes a Markov random field for $\mathbf{X}$, with a joint density $p_{\mathbf{X}}(\mathbf{x}) > 0$. With the ordering of maximal cliques $(K_1, \ldots, K_m)$ and separator

sets $(S_1, \ldots, S_m)$ defined in Equation (3), the joint density factorizes with respect to the decomposable graph G if and only if

$$p_{\mathbf{X}}(\mathbf{x}) = \prod_{l=1}^{m} p(\mathbf{x}_{K_l \setminus S_l} \mid \mathbf{x}_{S_l}) = \prod_{l=1}^{m} \frac{p(\mathbf{x}_{K_l})}{p(\mathbf{x}_{S_l})}, \tag{4}$$

where $S_1 = \emptyset$ so that the first factor reduces to $p(\mathbf{x}_{K_1})$.

The factorization in Equation (4) shows how a strictly positive MRF on a decomposable graph can be written as a product of clique terms normalized by separator terms, avoiding double counting on overlaps. Importantly, this decomposition implies that many global quantities become sums (or differences of sums) over local subsets, provided they respect the same clique/separator structural constraints. The next subsection uses this idea to extend information measures defined on the copula into the decomposable setting.

2.3. Copula-Based Information Measures in Decomposable Models

This subsection combines the two previous threads; copulas encode dependence, and decomposable graphs organize multivariate objects through cliques and separators. The key technical step is to formalize how clique copulas induce separator copulas via marginalization (expressed here through γ marginals) so that the same junction tree overlaps are reflected at the copula level. Given this construction, information measures derived from copulas can inherit a clique/separator additivity principle.

Recalling the distributional transform $U_i := F_i(X_i)$, we write $\mathbf{U} = (U_1, \ldots, U_d)^{\mathrm{T}}$. Let C denote the copula of $\mathbf{X} = (X_1, \ldots, X_d)^{\mathrm{T}}$, and let c denote its copula density. For any index set $A \subset \{1, \ldots, d\}$, we write $\mathbf{U}_A := \pi_A(\mathbf{U})$ and denote by C_A the distribution of $\mathbf{U}_A$ with an (absolutely continuous) density c_A. To work with maximal clique and separator copulas in decomposable models, we now introduce the notion of a marginal copula.

Definition 1 (γ-marginal of a copula function [17])**.** *Let $C : [0,1]^d \to [0,1]$ be a d-dimensional copula function, and let $\gamma = \{j_1, \ldots, j_r\} \subset \{1, \ldots, d\}$ with $1 \le r \le d-1$ and $j_1 < \cdots < j_r$. The γ-marginal of C is the function $C^\gamma : [0,1]^{|A|} \to [0,1]$ defined by*

$$C^\gamma(u_1, \ldots, u_r) := C(v_1, \ldots, v_d),$$

where it holds for each $j \in \{1, \ldots, d\}$ that

$$v_j = \begin{cases} u_k, & \text{if } j = j_k \text{ for some } k \in \{1, \ldots, r\}, \\ 1, & \text{otherwise.} \end{cases}$$

If C^γ is absolutely continuous, then we define its copula density by

$$c^\gamma(u_1, \ldots, u_r) := \frac{\partial^{|A|}}{\partial u_1 \cdots \partial u_r} C^\gamma(u_1, \ldots, u_r).$$

Now, let $G = (V, E)$ be decomposable with maximal cliques $(K_1, \ldots, K_m)$ and separators $(S_1, \ldots, S_m)$ as shown in Equation (3), where $S_1 = \emptyset$. For each clique $K_l \in \mathcal{K}$, we fix a bijection $\alpha_l : \{1, \ldots, |K_l|\} \to K_l$, which specifies the argument order used for the clique copula. For $l \ge 2$, we define the positional index set, selecting the separator coordinates inside clique K_l via $\gamma_l := \alpha_l^{-1}(S_l) \subset \{1, \ldots, |K_l|\}$. With this convention, the copula associated with the separator variables $\mathbf{U}_{S_l}$ is the γ_l marginal of the clique copula:

$$C_{S_l} \equiv C_{K_l}^{\gamma_l}, \qquad \text{and hence} \qquad c_{S_l} \equiv c_{K_l}^{\gamma_l}.$$

Proposition 1 (Markov property under marginal transforms). *Assume that each F_i is continuous and strictly increasing, and define $U_i = F_i(X_i)$. Then, for any disjoint $A, B, S \subseteq \{1, \ldots, d\}$, we have*

$$\mathbf{X}_A \perp\!\!\!\perp \mathbf{X}_B \mid \mathbf{X}_S \Leftrightarrow \mathbf{U}_A \perp\!\!\!\perp \mathbf{U}_B \mid \mathbf{U}_S.$$

Consequently, if $\mathbf{X}$ is Markov with respect to G, then so is $\mathbf{U}$.

Proof. Since F_i is continuous and strictly increasing, it is a bijective Borel-measurable map $\mathbb{R} \to \mathbb{R}$ with a Borel-measurable inverse F_i^{-1}. Thus, $U_i = F_i(X_i)$ is $\sigma(X_i)$-measurable and $X_i = F_i^{-1}(U_i)$ is $\sigma(U_i)$-measurable, and hence

$$\sigma(U_i) \subseteq \sigma(X_i) \quad \text{and} \quad \sigma(X_i) \subseteq \sigma(U_i),$$

such that $\sigma(X_i) = \sigma(U_i)$.

Now, we fix disjoint $A, B, S \subseteq \{1, \ldots, d\}$. Because $\mathbf{X}_A$ and $\mathbf{U}_A$ generate the same σ field, similar to B and S, one has

$$\sigma(\mathbf{X}_A) = \sigma(\mathbf{U}_A), \quad \sigma(\mathbf{X}_B) = \sigma(\mathbf{U}_B), \quad \sigma(\mathbf{X}_S) = \sigma(\mathbf{U}_S).$$

Conditional independence of random vectors is defined in terms of their generated σ fields: $\mathbf{X}_A \perp\!\!\!\perp \mathbf{X}_B \mid \mathbf{X}_S \Leftrightarrow \sigma(\mathbf{X}_A) \perp\!\!\!\perp \sigma(\mathbf{X}_B) \mid \sigma(\mathbf{X}_S)$ [18].

Using the equalities of the σ fields above, this is equivalent to

$$\sigma(\mathbf{U}_A) \perp\!\!\!\perp \sigma(\mathbf{U}_B) \mid \sigma(\mathbf{U}_S),$$

which is to say that $\mathbf{U}_A \perp\!\!\!\perp \mathbf{U}_B \mid \mathbf{U}_S$.

The same argument in reverse shows the converse implication, and hence

$$\mathbf{X}_A \perp\!\!\!\perp \mathbf{X}_B \mid \mathbf{X}_S \Leftrightarrow \mathbf{U}_A \perp\!\!\!\perp \mathbf{U}_B \mid \mathbf{U}_S$$

for all disjoint $A, B, S \subseteq \{1, \ldots, d\}$, completing the proof. □

Building upon this notion of γ marginals, the main structural message for copula-decomposable models is that the copula inherits the same clique/separator organization as the classical decomposable factorization in Equation (4). In a junction tree, adjacent cliques overlap only through their separator sets, so if one naively multiplies clique copula densities, the contribution of the shared variables is counted multiple times. The remedy is the familiar separator normalization; dividing by the corresponding separator copula density removes exactly the overcounted dependence associated with the overlap. Formally, this cancellation is justified by the running intersection property together with an induction along the chosen clique ordering. Variables that appear earlier can only reappear through separators and therefore vanish between the numerator and denominator terms, while the remainder sets $K_l \setminus S_l$ contribute only once. As a result, all marginal effects cancel out, leaving a factorization that isolates the dependence structure entirely through ratios of clique and separator copula densities. A complete, more standard proof (cast in the usual junction tree formalism) is provided in [19].

Equivalently, the same conclusion follows directly from the decomposable (junction tree) factorization applied to the copula-transformed variables. According to Proposition 1, the componentwise transform $U_i = F_i(X_i)$ preserves conditional independences, and thus $\mathbf{U}$ is Markov with respect to the same decomposable graph G. Any Markov distribution on a decomposable graph admits a junction tree marginal factorization into clique marginals divided by separator marginals (see Equation (4)). Since the joint density of $\mathbf{U}$ coincides with the copula density c, and for any A the marginal density of $\mathbf{U}_A$ is c_A, applying this fac-

torization to $\mathbf{U}$ yields the decomposable copula factorization in Equation (5). In particular, the global copula density admits the decomposable factorization

$$c(\mathbf{u}) = \prod_{l=1}^{m} \frac{c_{K_l}(\mathbf{u}_{K_l})}{c_{K_l}^{\gamma_l}(\mathbf{u}_{S_l})}, \qquad \mathbf{u} \in [0,1]^d, \tag{5}$$

with the convention that the denominator equals 1 for $l = 1$ (since $S_1 = \emptyset$).

As an immediate consequence of Equation (5), mutual information admits a copula entropy representation that is compatible with the clique/separator structure of G [19].

Theorem 1 (Mutual information via copula entropy in decomposable models)**.** *Let $\mathbf{X} = (X_1, \ldots, X_d)^{\mathrm{T}}$ be a random vector with continuous marginals, let $\mathbf{U} = (U_1, \ldots, U_d)^{\mathrm{T}}$ with $U_i := F_i(X_i)$, and let c denote the copula density of $\mathbf{X}$. Assume that the dependence graph G is decomposable with maximal cliques $(K_1, \ldots, K_m)$ and separator sets $(S_1, \ldots, S_m)$ (with $S_1 = \emptyset$) as shown in Equation (3) and that the copula density factorizes as shown in Equation (5). For every non-empty $A \subseteq \{1, \ldots, d\}$, let c_A be the density of $\mathbf{U}_A := \pi_A(\mathbf{U})$ and define its copula entropy as*

$$h_{c_A}(\mathbf{U}_A) := -\mathbb{E}[\log(c_A(\mathbf{U}_A))],$$

whenever the expectation is finite.

If $\mathbb{E}[|\log(c(\mathbf{U}))|] < \infty$, $\mathbb{E}[|\log(c_{K_l}(\mathbf{U}_{K_l}))|] < \infty$, and $\mathbb{E}[|\log(c_{S_l}(\mathbf{U}_{S_l}))|] < \infty$ for all $l \in \{1, \ldots, m\}$, then

$$I(\mathbf{X}) = \sum_{l=1}^{m} h_{c_{S_l}}(\mathbf{U}_{S_l}) - \sum_{l=1}^{m} h_{c_{K_l}}(\mathbf{U}_{K_l}), \tag{6}$$

with $c_{S_l} \equiv c_{K_l}^{\gamma_l}$.

Proof. Under the identity $I(\mathbf{X}) = -h_c(\mathbf{U})$ from Equation (2), one has

$$I(\mathbf{X}) = \mathbb{E}[\log(c(\mathbf{U}))].$$

Using the decomposable copula factorization from Equation (5), we obtain

$$\log(c(\mathbf{U})) = \sum_{l=1}^{m} \log(c_{K_l}(\mathbf{U}_{K_l})) - \sum_{l=1}^{m} \log(c_{S_l}(\mathbf{U}_{S_l})),$$

where $c_{S_l} \equiv c_{K_l}^{\gamma_l}$ by construction and $S_1 = \emptyset$ contributes no term.

Taking the expectations and using linearity (justified by the assumed integrability) yields

$$I(\mathbf{X}) = \sum_{l=1}^{m} \mathbb{E}\big[\log(c_{K_l}(\mathbf{U}_{K_l}))\big] - \sum_{l=1}^{m} \mathbb{E}\big[\log(c_{S_l}(\mathbf{U}_{S_l}))\big].$$

Finally, since $\mathbb{E}[\log(c_A(\mathbf{U}_A))] = -h_{c_A}(\mathbf{U}_A)$ by definition of copula entropy for each relevant margin, rearranging this gives Equation (6). □

The decomposable copula factorization in Equation (5) is the structural statement that makes the proposed learning algorithm possible; it rewrites a global dependence density as a product of clique copula terms divided by separator copula terms, matching the usual junction tree normalization logic. On this basis, mutual information decomposes into a difference of sums of copula entropies over separators and cliques, as stated in Theorem 1.

Taken together, the results in this section justify an information-centric representation of dependence that is simultaneously invariant to monotone marginal transforms and compatible with junction tree factorizations. This is precisely the combination needed for scalable learning; copula entropy links directly to mutual information, and decomposability

converts global objects into sums of local clique and separator contributions. The remainder of the paper treats this decomposition not just as a theoretical identity but as a blueprint for defining scores and computing local updates under chordality-preserving moves. The next section leverages this compatibility to define a practical nonparametric scoring criterion and a greedy search strategy that maintains chordality at every step.

3. Nonparametric Structure Learning for Decomposable Models Using Information-Based Measures

This section turns the theoretical copula–graph compatibility into a concrete structure-learning procedure for decomposable graphs. The goal is to optimize a global objective that is motivated by mutual information but is computable from data using nonparametric copula entropy estimators. Because the search is restricted to chordal graphs, the algorithm can maintain a junction tree throughout and exploit clique/separator locality to evaluate score changes efficiently.

Let $\mathbf{X} = (X_1, \ldots, X_d)^{\mathrm{T}}$ be a continuous random vector as defined previously. The objective is to learn a decomposable graph $G = (V, E)$ that encodes the conditional independence structure of $\mathbf{X}$. This is formulated as optimizing a global criterion derived from the copula-based information measures introduced in Theorem 1. The search is constrained to the space of chordal (decomposable) graphs and proceeds via local edge modifications that preserve decomposability.

3.1. Information-Theoretic Scoring Function

In this section, we define the population-level score that will drive structure learning, using copula entropy as the primitive building block. The definition is engineered to mirror the clique/separator decomposition of mutual information in decomposable models so that the score is both globally meaningful and locally decomposable. A CBIC-type penalty is then added to balance fit and complexity while preserving the same additivity pattern across cliques and separators.

Let G be a decomposable graph with maximal cliques $(K_1, \ldots, K_m)$ and separators $(S_1, \ldots, S_m)$ satisfying the running intersection property (Equation (3)), where $S_1 = \emptyset$. For any nonempty index set $A \subseteq \{1, \ldots, d\}$, recall the copula entropy $h_{c_A}(\mathbf{U}_A) = -\mathbb{E}[\log(c_A(\mathbf{U}_A))]$. Define the functional

$$\psi(A) := -h_{c_A}(\mathbf{U}_A) = \mathbb{E}[\log(c_A(\mathbf{U}_A))]. \tag{7}$$

In direct analogy to Equation (6), we define the global information of G as

$$\mathcal{I}(G) := \sum_{l=1}^{m} \psi(K_l) - \sum_{l=1}^{m} \psi(S_l). \tag{8}$$

According to Theorem 1, if the copula density c factorizes according to G as in Equation (5), then $\mathcal{I}(G) = I(\mathbf{X})$, representing the mutual information of the full vector.

Lasserre et al. [13] proposed an improved CBIC criterion for structure learning in copula BNs, in which the log-likelihood is rewritten in terms of mutual information (equivalently, copula entropy) contributions. Motivated by this construction, we introduce an analogous global information-based score for decomposable models. Assume that the graph complexity functional $\Lambda(G)$ admits the clique/separator decomposition

$$\Lambda(G) := \sum_{l=1}^{m} \lambda(K_l) - \sum_{l=1}^{m} \lambda(S_l), \tag{9}$$

for some nonnegative set functional $\lambda\colon 2^{\{1,\dots,d\}} \to \mathbb{R}_+$ (for instance, $\lambda(A)$ may be the dimension of a parametric family or built from clique/separator dimensions $\lambda(A) = |A|(|A|+1)/2$).

Given the global information functional $\mathcal{I}(G)$ in Equation (8), we define the population CBIC-type score as

$$\mathcal{S}_{\mathrm{CBIC}}(G) := n\,\mathcal{I}(G) - \frac{\log(n)}{2}\,\Lambda(G), \tag{10}$$

where n is the sample size. The scaling by n and $\log(n)$ mirrors the usual large-sample form of CBIC, while the leading term remains purely information-theoretic through $\mathcal{I}(G)$.

The resulting criterion $\mathcal{S}_{\mathrm{CBIC}}(G)$ is designed such that (1) its leading term targets the dependence structure through information and (2) its penalty term can be chosen to reflect model capacity in a graph-local way [13]. This ensures that changing the graph only perturbs a small number of set function evaluations, which is crucial for greedy search. The next subsection makes this locality explicit by deriving closed-form score differences for single-edge updates.

3.2. Local Search and Score Decomposition

We incorporate the global score into a practical optimization procedure by focusing solely on single-edge modifications that preserve decomposability. The central observation is that under a junction tree representation, an admissible insertion or deletion only modifies a small neighborhood of cliques and separators [9], causing most score terms to cancel out in differences. As a result, it becomes possible to evaluate many candidate moves cheaply, without recomputing the entire score from scratch.

The search algorithm relies on evaluating the score change induced by a single edge addition or removal that preserves decomposability. Let G and G' be two decomposable graphs differing by exactly one such edge operation. Their respective score difference $\mathcal{S}_{\mathrm{CBIC}}(G') - \mathcal{S}_{\mathrm{CBIC}}(G)$ (Equation (10)) depends only on the cliques and separators whose composition changes such that most clique/separator contributions cancel out.

Let $G = (V, E)$ be decomposable, and let $(K_1, \dots, K_m)$ be a clique ordering satisfying the running intersection property with separators $S_l = K_l \cap K_{\eta(l)}$ as shown in Equation (3). Recall that the sum $\sum_{l=1}^{m} \psi(S_l)$ treats separators with their multiplicity induced by the chosen clique tree (equivalently, by the edges of a junction tree). This is relevant because different clique edges may yield identical separator sets as subsets of V.

A decomposable graph G may admit multiple junction tree (clique tree) representations and multiple running-intersection orderings. At the population level, when the copula density factorizes with respect to G, the resulting clique/separator decomposition (and hence $\mathcal{I}(G)$) is a property of G and does not depend on the particular junction tree. In the finite-sample setting, however, $\widehat{\psi}$ is an approximation, and numerical values can differ slightly across equivalent junction tree representations because separators enter the score with their edge-induced multiplicity. In our implementation, the score is therefore defined with respect to the current maintained junction tree $T = (\mathcal{K}, \mathcal{F})$; the separator contributions correspond to the multiset $\{K \cap K' : (K, K') \in \mathcal{F}\}$, and this same representation is used consistently throughout the search and for computing local score updates (see Section 3.4).

Denote by $\mathfrak{G}_{\mathrm{dec}}$ the class of decomposable graphs on V. For $G \in \mathfrak{G}_{\mathrm{dec}}$, we define its valid local neighborhood $\mathcal{N}(G) \subset \mathfrak{G}_{\mathrm{dec}}$ as the set of decomposable graphs obtained from G by either inserting one edge $\{v_i, v_j\} \notin E$ while preserving decomposability or deleting one edge $\{v_i, v_j\} \in E$ while preserving decomposability. In practice, the admissibility of insert and delete moves can be checked using dynamic chordal graph criteria expressed on a maintained junction tree [15]. In particular, a deletion $\{v_i, v_j\} \in E$ preserves decomposability if and only if $\{v_i, v_j\}$ is contained in exactly one maximal clique of G, and an

analogous junction tree criterion exists for insertion moves [9]. Assume that $G \in \mathfrak{G}_{\mathrm{dec}}$, and let $G' \in \mathcal{N}(G)$ be obtained by a single valid edge insertion. Assume that the move can be represented on some junction tree of G as follows. There exist two adjacent maximal cliques K_k and K_l with the separator

$$S = K_k \cap K_l,$$

such that $v_i \in K_k \setminus S$ and $v_j \in K_l \setminus S$, and after inserting $\{v_i, v_j\}$, a new maximal clique

$$K^* = S \cup \{v_i, v_j\}$$

is created, replacing the separator S with the two separators $S \cup \{v_i\}$ and $S \cup \{v_j\}$ in the updated junction tree [9].

Proposition 2 (Local score update under edge insertion [9,15])**.** *Let $G \in \mathfrak{G}_{\mathrm{dec}}$, and let G' be obtained from G by a single valid insertion of the edge $\{v_i, v_j\} \notin E$, with the associated separator S and new clique $K^* := S \cup \{v_i, v_j\}$ as described above. Then, the score difference*

$$\Delta_{ij}(G) := \mathcal{S}_{\mathrm{CBIC}}(G') - \mathcal{S}_{\mathrm{CBIC}}(G)$$

can be written as

$$\Delta_{ij}(G) = \kappa\big(S \cup \{v_i, v_j\}\big) - \kappa\big(S \cup \{v_i\}\big) - \kappa\big(S \cup \{v_j\}\big) + \kappa(S), \tag{11}$$

where, for any nonempty $A \subseteq \{1, \ldots, d\}$, we have

$$\kappa(A) := n\,\psi(A) - \frac{\log n}{2}\,\lambda(A), \tag{12}$$

with $\psi(A)$ as in Equation (7) and $\lambda(A)$ being the local complexity term appearing in Equation (9). In particular, $\Delta_{ij}(G)$ depends only on the four subsets S, $S \cup \{v_i\}$, $S \cup \{v_j\}$, and $S \cup \{v_i, v_j\}$.

Proof. Write $\mathcal{S}_{\mathrm{CBIC}}(G)$ in the clique/separator form in Equation (10), with separator multiplicities given by the chosen clique ordering (or, equivalently, by the junction tree). Under the described insertion move, all clique terms and all separator terms that are not incident to the modified junction tree edge cancel out in the difference $\mathcal{S}_{\mathrm{CBIC}}(G') - \mathcal{S}_{\mathrm{CBIC}}(G)$. The only changes are that the new clique K^* is added; the old separator S is removed; and the two new separators $S \cup \{v_i\}$ and $S \cup \{v_j\}$ are added. Substituting these changes into Equation (10) yields Equation (11). □

Conversely, suppose that $\{v_i, v_j\} \in E$ is contained in exactly one maximal clique K of G such that deleting $\{v_i, v_j\}$ preserves decomposability, thus resulting in G'. Let $S := K \setminus \{v_i, v_j\}$. Then, the deletion splits the clique K into two cliques, $S \cup \{v_i\}$ and $S \cup \{v_j\}$, connected by a separator S. In this case, the score change is the negative of Equation (11).

The update formula in Equation (11) provides a relevant operational result; it expresses $\Delta_{ij}(G)$ in terms of only four local subsets defined by the affected separator and the two incident vertices. Such local updates are what make greedy search over chordal graphs scalable when combined with efficient admissibility checks. The remaining task is to specify how the local information terms $\psi(A)$ are estimated from data in a manner that is stable under the repeated evaluations required by the search procedure.

3.3. Nonparametric Copula Entropy Estimator

This section specifies the estimator used for the copula information functional, which is the statistical core of the scoring function. Because $\psi(A)$ depends only on the copula of $\mathbf{X}_A$, the procedure constructs pseudo-observations from empirical marginals, estimates a copula density on $(0,1)^{|A|}$ via transformation-based kernel smoothing, and evaluates a predictive log-score [12,14].

We adopt the nonparametric approach of Ma and Sun [12] and the related copula density-based method [14,16] to estimate the functional $\psi(A) = -h_{c_A}(\mathbf{U}_A)$. Let $\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(n)}$ be an IID sample of $\mathbf{X} = (X_1, \ldots, X_d)^{\mathrm{T}}$, and fix a nonempty index set $A \subseteq \{1, \ldots, d\}$. Since $\psi(A)$ depends only on the copula of $\mathbf{X}_A$, the estimation reduces to constructing pseudo-observations for $\mathbf{U}_A$ and estimating the corresponding copula density and copula entropy [12,20].

We partition the training index set J_{tr} into k disjoint folds $J_1, \ldots, J_k$ and define, for each fold $r \in \{1, \ldots, k\}$, the corresponding CV training index set:

$$J_{-r} := J_{\mathrm{tr}} \setminus J_r.$$

We then define the indicator function of a subset $B \subseteq D$ as $\chi_B \colon D \to \{0,1\}$ with

$$\chi_B(b) := \begin{cases} 1, & \text{if } b \in B, \\ 0, & \text{otherwise.} \end{cases}$$

For each $i \in \{1, \ldots, d\}$ and each fold $r \in \{1, \ldots, k\}$, we define the empirical marginal CDF

$$\widehat{F}_{i,r}(x) := \frac{1}{|J_{-r}|} \sum_{s \in J_{-r}} \chi_{(-\infty, x]}\big(X_i^{(s)}\big), \qquad x \in \mathbb{R}.$$

To obtain pseudo-observations in the unit interval that remain strictly away from the endpoints and are therefore suitable for probit transformation, we follow the standard pseudo-observation scaling used in copula estimation and add an explicit truncation. For each observation index $t \in \{1, \ldots, n\}$, we define the fold r pseudo-observations by

$$\widehat{U}_{i,r}^{(t)} := \min\left\{1 - \varepsilon,\ \max\left\{\varepsilon,\ \frac{|J_{-r}|}{|J_{-r}| + 1}\, \widehat{F}_{i,r}\big(X_i^{(t)}\big)\right\}\right\}.$$

The factor $|J_{-r}|/(|J_{-r}| + 1)$ is equivalent to the pseudo-observation construction based on marginal ranks, which is widely used in semiparametric copula inference [14]. The additional truncation at $\varepsilon > 0$ is a deterministic numerical convention that ensures $\widehat{U}_{i,r}^{(t)} \in [\varepsilon, 1 - \varepsilon]$. This guarantees that the maximum observation in a fold never maps exactly to one and, similarly, that the minimum never maps exactly to zero.

The truncation is essential because the probit transformation Φ^{-1} is only defined on the open interval $(0,1)$, with limiting behavior

$$\lim_{u \to 0^+} \Phi^{-1}(u) = -\infty, \qquad \lim_{u \to 1^-} \Phi^{-1}(u) = +\infty.$$

Applying the probit transform to pseudo-observations exactly at zero or one would therefore yield infinite values, rendering kernel density estimation impossible. By enforcing $\widehat{U}_{i,r}^{(t)} \in [\varepsilon, 1 - \varepsilon]$, the transformed observations $\widehat{Z}_{i,r}^{(t)} = \Phi^{-1}(\widehat{U}_{i,r}^{(t)})$ are guaranteed to be finite real numbers for all sample points, ensuring the probit transformation remains numerically

well-defined throughout the procedure [14]. We define the corresponding vectors and subvectors by

$$\widehat{\mathbf{U}}_r^{(t)} := (\widehat{U}_{1,r}^{(t)}, \ldots, \widehat{U}_{d,r}^{(t)})^{\mathrm{T}}, \qquad \widehat{\mathbf{U}}_{A,r}^{(t)} := \pi_A(\widehat{\mathbf{U}}_r^{(t)}) \in (0,1)^{|A|}.$$

Kernel smoothing on $(0,1)^{|A|}$ is prone to boundary bias. Following Geenens et al. [14], we therefore apply the probit transform component-wise. Let Φ and ϕ denote the standard normal CDF and PDF, respectively, and define

$$\widehat{Z}_{i,r}^{(t)} := \Phi^{-1}(\widehat{U}_{i,r}^{(t)}), \qquad \widehat{\mathbf{Z}}_r^{(t)} := (\widehat{Z}_{1,r}^{(t)}, \ldots, \widehat{Z}_{d,r}^{(t)})^{\mathrm{T}} \in \mathbb{R}^d, \qquad \widehat{\mathbf{Z}}_{A,r}^{(t)} := \pi_A(\widehat{\mathbf{Z}}_r^{(t)}) \in \mathbb{R}^{|A|}.$$

Let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d) \in \mathbb{R}_+^d$ be a global diagonal bandwidth vector, and write $\boldsymbol{\theta}_A := \pi_A(\boldsymbol{\theta})$. We define the Gaussian kernel on $\mathbb{R}^{|A|}$ by

$$g(\mathbf{z} \mid \boldsymbol{\theta}_A) = (2\pi)^{-|A|/2} |\mathrm{diag}(\boldsymbol{\theta}_A)|^{-1/2} \exp\left(-\frac{1}{2}\mathbf{z}^{\mathrm{T}} \mathrm{diag}(\boldsymbol{\theta}_A)^{-1} \mathbf{z}\right).$$

For each fold r and subset $A \subseteq \{1, \ldots, d\}$, define the probit transformation KDE estimator of the copula density c_A by

$$\widehat{c}_{A,r}(\mathbf{u}_A \mid \boldsymbol{\theta}_A) := \frac{\frac{1}{|J_{-r}|} \sum_{s \in J_{-r}} g(\mathbf{z}_A - \widehat{\mathbf{Z}}_{A,r}^{(s)} \mid \boldsymbol{\theta}_A)}{\prod_{i \in A} \phi(z_i)}, \qquad \mathbf{u}_A \in (0,1)^{|A|}, \tag{13}$$

where $z_i = \Phi^{-1}(u_i)$ for $i \in \{1, \ldots, d\}$, $\mathbf{z} = (z_1, \ldots, z_d)^{\mathrm{T}} \in \mathbb{R}^d$ and the corresponding sub-vector on margin index set A is $\mathbf{z}_A = \pi_A(\mathbf{z})$. The denominator in Equation (13) is the Jacobian factor arising from the change in variables between the probit-transformed space and the copula domain. We then define the estimator of $\psi(A)$ as the average predictive log-score per observation:

$$\widehat{\psi}_{\mathrm{KDE}}(A) := \frac{1}{n} \sum_{r=1}^{k} \sum_{t \in J_r} \log \widehat{c}_{A,r}(\widehat{\mathbf{U}}_{A,r}^{(t)} \mid \boldsymbol{\theta}_A). \tag{14}$$

For each fold r, the empirical marginals $\widehat{F}_{i,r}$ and the kernel centers $\widehat{\mathbf{Z}}_{A,r}^{(s)}$ are estimated using the training indices J_{-r} only, while the predictive log-score averages $\log \widehat{c}_{A,r}(\widehat{\mathbf{U}}_{A,r}^{(t)} \mid \boldsymbol{\theta}_A)$ over the held-out indices $t \in J_r$. This construction ensures that the predictive log-score is strictly out of fold and avoids information leakage from test folds into the fitted copula model.

Following the predictive-assessment KDE approach of Schwaighofer et al. [9], we select a single global bandwidth vector $\boldsymbol{\theta}$ before the structure search and subsequently keep it fixed. Here, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d) \in \mathbb{R}_+^d$ is a vector with one component per dimension, and for any subset A, we use the restriction $\boldsymbol{\theta}_A = \pi_A(\boldsymbol{\theta})$ in Equation (13). The same global vector is shared across all cliques and separators.

We choose $\boldsymbol{\theta}$ by maximizing a leave-one-out CV predictive log-score on the probit-transformed full training sample $\{\widehat{\mathbf{Z}}^{(t)}\}_{t \in J_{\mathrm{tr}}} \subset \mathbb{R}^d$ without conditioning on a specific subset A. For numerical convenience, we optimize over the log-bandwidth parameters $\boldsymbol{\xi} \in \mathbb{R}^d$, defined component-wise by $\xi_j := \log(\theta_j)$, so that $\boldsymbol{\theta}$ is automatically in $\mathbb{R}_+^d$. We solve a box-constrained problem using an L-BFGS-B quasi-Newton method [21–23]:

$$\widehat{\boldsymbol{\xi}} \in \underset{\boldsymbol{\xi} \in [\boldsymbol{\ell}, \mathbf{h}]}{\arg\max} \sum_{t \in J_{\mathrm{tr}}} \log\left(\frac{1}{|J_{\mathrm{tr}}| - 1} \sum_{s \in J_{\mathrm{tr}} \setminus \{t\}} g(\widehat{\mathbf{Z}}^{(t)} - \widehat{\mathbf{Z}}^{(s)} \mid \exp(\boldsymbol{\xi}))\right), \qquad \widehat{\boldsymbol{\theta}} = \exp(\widehat{\boldsymbol{\xi}}).$$

The box constraints are chosen component-wise as

$$\ell_j = \log\big(\widehat{\mathrm{Var}}(Z_j)\,\beta_{\min}\big), \qquad h_j = \log\big(\widehat{\mathrm{Var}}(Z_j)\,\beta_{\max}\big),$$

for fixed constants $0 < \beta_{\min} < \beta_{\max}$, where $\widehat{\mathrm{Var}}(Z_j)$ is the empirical variance of the jth coordinate in $\{\widehat{\mathbf{Z}}^{(t)}\}_{t\in J_{\mathrm{tr}}}$. The optimizer is terminated when either a prescribed maximum number of iterations is reached or the relative change in the objective falls below a tolerance parameter.

This leave-one-out CV optimization is performed once before structure learning. In all experiments reported in Section 4, its wall clock time was negligible compared with the subsequent greedy search, which requires repeated evaluation of $\widehat{\psi}_{\mathrm{KDE}}(A)$ for many overlapping subsets A.

The truncation $\widehat{U}_{i,r}^{(t)} \in [\varepsilon, 1-\varepsilon]$ guarantees that the probit transform $\widehat{Z}_{i,r}^{(t)} = \Phi^{-1}(\widehat{U}_{i,r}^{(t)})$ is always finite and avoids singularities in the Jacobian factor $\prod_{i\in A}\phi(z_i)^{-1}$. In the computation of the KDE sums entering the leave-one-out CV objective and $\widehat{c}_{A,r}$, we evaluate log-densities using log-sum-exp identities to avoid numerical underflow in the tails and for large values of $|A|$.

To further guard against numerical artifacts when $\widehat{c}_{A,r}$ is extremely small, we apply a fixed-density floor. Throughout, we replace

$$\log \widehat{c}_{A,r}(\mathbf{u}_A \mid \boldsymbol{\theta}_A) \;\leftarrow\; \max\Big\{\log \widehat{c}_{A,r}(\mathbf{u}_A \mid \boldsymbol{\theta}_A),\, \log \delta_{\mathrm{floor}}\Big\}$$

for a small constant $\delta_{\mathrm{floor}} > 0$. If a candidate evaluation yields non-finite values due to numerical issues, then we treat that candidate as inadmissible and deterministically prevent that configuration from being selected during greedy search. Because Δ_{ij} in Equation (11) depends only on differences in a small number of local terms, and the same flooring rule is applied uniformly across all candidate edges and graphs, this safeguard only affects terms that are already well below machine precision and did not lead to unstable edge decisions in our empirical study.

The estimator $\widehat{\psi}_{\mathrm{KDE}}(A)$ inherits the standard large-sample properties of probit-transformed kernel copula density estimators. Under suitable smoothness conditions on c_A, and for bandwidths $\boldsymbol{\theta}$ shrinking at appropriate rates, $\widehat{c}_{A,r}$ is pointwise consistent (and, on compact subsets of the interior, uniformly consistent), with asymptotic bias and variance governed by the usual kernel-smoothing orders in the transformed domain [14,24]. Moreover, consistency of the plug-in predictive log-score for the copula information functional $\psi(A) = \mathbb{E}[\log c_A(\mathbf{U}_A)]$ holds under additional regularity, ensuring that the logarithm is well behaved. For example, c_A is strictly positive on the interior and $\mathbb{E}|\log c_A(\mathbf{U}_A)| < \infty$, together with sufficiently strong (e.g., uniform-on-compacts) convergence of $\widehat{c}_{A,r}$ so that $\log \widehat{c}_{A,r} \to \log c_A$ on those regions [25,26].

The use of k-fold out-of-fold predictive log-scores rather than exact leave-one-out optimization modifies the fitted estimators through the smaller training size in each fold and hence can affect finite-sample variability (and, slightly, the finite sample bias), but for fixed k values, the difference is asymptotically negligible as $n \to \infty$ in standard cross-validatory logarithmic assessment settings [27,28]. The empirical marginal pseudo-observation construction implies that $\widehat{\psi}_{\mathrm{KDE}}(A)$ is invariant under strictly monotone transformations of each marginal X_i, and thus the resulting score depends only on the copula of $\mathbf{X}_A$ [14]. Finally, the endpoint truncation $\widehat{U}_{i,r}^{(t)} \in [\varepsilon, 1-\varepsilon]$ ensures that the probit transform and its Jacobian factor remain finite; if $\varepsilon \to 0$, and $\log c_A(\mathbf{U}_A)$ is integrable, then the truncation perturbs the target functional by a vanishing amount as $|J_{-r}| \to \infty$, akin to the trimming and truncation devices commonly used to control log-density singularities in kernel-based entropy estimation [25,26].

3.4. Search over Decomposable Structures

This subsection details the optimization strategy: a greedy forward and backward procedure over chordal graphs that enforces decomposability at every step. The junction tree is treated as the central state representation, since it simultaneously encodes the clique/separator structure for scoring and supports graph-theoretic queries for deciding whether an insertion or deletion is admissible. The algorithm is therefore best viewed as an interplay between local information gains (via $\Delta_{ij}(G)$) and local structural feasibility (via dynamic chordal criteria).

We optimize the global score $\mathcal{I}(G)$ in Equation (10) over $\mathfrak{G}_{\text{dec}}$ via the greedy stepwise algorithm of Schwaighofer et al. [9], which alternates forward selection and backward elimination within a single run while preserving decomposability (chordality) at every step. In the original work [9], the score is a predictive assessment criterion based on a cross-validated log-likelihood for a decomposable kernel density estimator, where the clique/separator factorization enables local accumulation; however, the search mechanism is agnostic to the particular scoring functional. Starting from the empty graph, each iteration evaluates admissible single-edge insertions and deletions and applies the move that yields the largest score improvement, with the central design requirements being that valid local moves can be enumerated efficiently within $\mathfrak{G}_{\text{dec}}$ and that each move admits a local score update as shown in Equation (11). Because unrestricted edge edits may violate chordality, candidate moves are explicitly restricted to chordality-preserving operations using a scheme inspired by fully dynamic chordal graph algorithms, allowing enumeration of valid edges in amortized $O(d^2 \log(d))$ time while maintaining the clique/separator representation needed to compute local score changes throughout the search.

To represent the current structure and test the admissibility of local moves (i.e., preserving decomposability), we maintain a junction tree $T = (\mathcal{K}, \mathcal{F})$ associated with the current chordal graph G, where $\mathcal{K}$ is the set of maximal cliques of G and $\mathcal{F}$ is a set of edges connecting these cliques so that T satisfies the clique intersection property (i.e., T is a clique tree representation of G). We keep a single junction tree representation of the current graph and update it dynamically after each accepted move; all clique and separator terms (including their multiplicities) are computed from this maintained junction tree (equivalently, separator terms correspond to the multiset $\{K \cap K' : (K, K') \in \mathcal{F}\}$). For prospective insertion of a non-existing edge $\{v_i, v_j\}$, the method considers the closest cliques K_k and K_l in the current join tree such that $v_i \in K_k$ and $v_j \in K_l$ and uses a weight function $w\colon \mathcal{F} \to \mathbb{Z}_+$ on the join tree edges defined by

$$w(K_k, K_l) = |K_k \cap K_l|.$$

Insertion is valid precisely when the minimum-weight edge along the path between K_k and K_l has a weight equal to $w(K_k, K_l)$, ensuring that a clique tree for the updated graph exists [9,15]. For deletion of an existing edge $\{v_i, v_j\}$, validity is characterized more simply; $\{v_i, v_j\}$ can be removed while preserving decomposability if and only if the current graph has exactly one maximal clique containing both v_i and v_j [9,15]. To make these checks efficient, the join tree is maintained using a dynamic tree representation based on splay trees, supporting link and cut and shortest-path queries in amortized $O(\log(d))$ time. Since a chordal graph on d variables has at most d cliques, the required join tree operations inherit this bound [9,15].

Maintaining chordality throughout the search ensures that every intermediate graph admits a valid clique/separator decomposition, and thus score evaluation never leaves the decomposable regime the theory assumes. Moreover, the dynamic junction tree machinery makes it possible to enumerate and test candidate moves without resorting to expensive

global triangulation at each step. With the learning procedure fully specified, attention can now shift to empirical behavior, or how well the method recovers edges from finite samples and how its runtime scales under repeated local updates.

The learning problem is reduced to three modular components: a decomposable information score, a local-move mechanism that preserves chordality, and a nonparametric estimator for the clique/separator copula contributions. This modularity is essential; it separates statistical modeling choices (how to estimate $\psi(A)$) from graph-theoretic constraints (among which moves are admissible) and from optimization strategy (greedy updates with local deltas). The next section validates this pipeline empirically on controlled synthetic settings and on a real dataset.

Let k be the number of folds used in the predictive log-score estimator of Section 3.3, let $\omega + 1$ denote the maximum clique size over graphs visited by the search, and write N_{iter} and $N_{\text{cand}}^{(q)}$ for the number of greedy iterations and the number of admissible candidate moves at iteration q, respectively. For a fixed subset A with $|A| \leq \omega + 1$, the k-fold predictive log-score $\widehat{\psi}_{\text{KDE}}(A)$ in Equation (14) is obtained by repeatedly evaluating the probit-KDE in Equation (13) on held-out indices $t \in J_r$ using all training indices $s \in J_{-r}$. A direct implementation therefore requires $O(n^2|A|) \leq O(n^2\omega)$ kernel evaluations per $\widehat{\psi}_{\text{KDE}}(A)$. Proposition 2 shows that each candidate edge $\{v_i, v_j\}$ in $\mathcal{N}(G^{(q)})$ requires only the four subsets, and thus one local score difference $\Delta_{ij}(G^{(q)})$ can be computed in $O(n^2\omega)$ time up to a constant factor. Since the dynamic tree machinery of Section 3.4 yields all admissible candidates $\mathcal{N}(G^{(q)})$ and supports the corresponding junction tree updates in amortized $O(d^2 \log d)$ and $O(\log d)$ time, the dominant cost of iteration q is $O(d^2 \log d) + O(N_{\text{cand}}^{(q)}\, n^2\omega)$, and a worst-case bound for the whole greedy run is $O(N_{\text{iter}}\, d^2 \log d) + O(n^2\omega \sum_{t=1}^{N_{\text{iter}}} N_{\text{cand}}^{(q)})$, highlighting that the overall complexity is driven primarily by the number of evaluated candidates and the $O(n^2\omega)$ cost of each KDE-based local score.

4. Experimental Evaluation

This section empirically assesses both the ability of the proposed copula information score to recover decomposable dependence graphs from finite samples and the computational behavior of the resulting greedy chordal search [9,15]. The experiments are designed to isolate the effect of the copula information score under controlled ground-truth decomposable graphs while also evaluating end-to-end performance within a realistic greedy chordal search loop. In addition to synthetic studies with a known truth, a real dataset experiment demonstrates how the learned structures can be applied in a practical setting. All experiments are based on undirected decomposable (chordal) structures and compare learned graphs using standard edge recovery metrics. The experiments were conducted on an 8-core Intel Core i7 machine with 16 GB of RAM and the code in Python 3.10.

4.1. Synthetic Data Generation

Now, we define a controlled data-generating process in which the ground-truth dependence structure is known and explicitly decomposable. The construction proceeds in two stages; the first involves sampling a random graph and then applying a chordal completion to guarantee decomposability, and the second involves sampling observations from a Gaussian copula Markov random field whose latent conditional independences match that chordal structure, while the observed univariate marginals are imposed by applying the corresponding quantile functions. This set-up provides a clean benchmark for recovery, since any deviation in the learned edges can be attributed to finite-sample effects, estimation error in the score, or search sub-optimality.

The number of variables $d \in \mathbb{N}$ and an edge probability $\rho \in (0,1)$ are fixed. We first draw an Erdős–Rényi graph $G_0 \sim \mathfrak{G}(d,\rho)$ and then obtain a chordal (decomposable)

ground-truth graph $G = \tau(G_0)$ via a triangulation (chordal completion) procedure [29]. We treat the resulting chordal graph $G = (V, E)$ with $V = \{v_1, \ldots, v_d\}$ as the ground-truth conditional-independence structure.

Conditional on G, we draw IID latent weights

$$Y_{ij} \sim \text{Gamma}(\alpha = 2, \beta = 1), \qquad 1 \le i < j \le d,$$

where α denotes the shape parameter such that $\mathbb{E}[Y_{ij}] = \alpha/\beta$. Define the symmetric weighted adjacency matrix $W \in \mathbb{R}^{d\times d}$ by

$$W_{ij} = \chi_E(\{v_i, v_j\})\, Y_{ij}, \quad i < j, \qquad W_{ji} = W_{ij}, \qquad W_{ii} = 0.$$

Let $D = \text{diag}(d_1, \ldots, d_d)$ be the diagonal degree matrix with $d_i = \sum_{k=1}^{d} W_{ik}$, and let I_d denote the $d \times d$ identity matrix. We construct the precision matrix $Q \in \mathbb{R}^{d\times d}$ as a ridge-regularized graph Laplacian:

$$Q = \nu I_d + D - W, \qquad \nu > 0.$$

This matrix is strictly diagonally dominant and hence positive definite, and its off-diagonal sparsity pattern matches the edge set E, so it can be used as the precision matrix of a copula Gaussian graphical model [30,31]. Let Q^{-1} denote the latent covariance matrix. Since a Gaussian copula is parameterized by a correlation matrix, we normalize Q^{-1} to a correlation matrix using

$$S = \text{diag}(Q^{-1})^{1/2},$$

so that $S^{-1}Q^{-1}S^{-1}$ has a diagonal unit. Note that diagonal rescaling preserves the conditional independence graph; the standardized latent precision is SQS, which has the same zero pattern as Q while keeping the identifiability of the marginal scales. We generate IID latent vectors via

$$\mathbf{Z}^{(1)}, \ldots, \mathbf{Z}^{(n)} \sim \text{Multivariate Normal}\big(\mathbf{0},\, (SQS)^{-1}\big),$$

and form pseudo-observations $U_j^{(t)} = \Phi(Z_j^{(t)}) \in (0, 1)$, where Φ is the standard normal CDF.

Finally, we obtain the observed data by applying the PPF of a chosen univariate distribution component-wise:

$$X_j^{(t)} = F_j^{-1}(U_j^{(t)}), \qquad j \in \{1, \ldots, d\},\ t \in \{1, \ldots, n\}.$$

In maintaining the same pseudo-observations $\{U_j^{(t)}\}$ across marginal regimes (and applying the same marginal distribution to all coordinates), we consider three choices with distinct supports: (1) Normal$(0, 1)$ on $\mathbb{R}$, for which $F^{-1} = \Phi^{-1}$ and therefore $X_j^{(t)} = \Phi^{-1}(U_j^{(t)}) = Z_j^{(t)}$; (2) Exponential$(1)$ on $[0, \infty)$, using the exponential PPF; and (3) Beta$(1/2, 1/2)$ on $[0, 1]$, using the beta PPF.

By generating data from a Gaussian copula graphical model whose latent precision respects the chordal sparsity pattern, the experiment provides a well-understood baseline in which missing edges correspond to conditional independences in the latent Gaussian layer, while the three marginal regimes probe learning performance under unbounded real-valued, strictly positive, and bounded supports. This design isolates marginal heterogeneity; across regimes, we keep the same latent dependence graph (Gaussian copula and conditional independences) and modify only the one-dimensional marginals via component-wise monotone transforms.

4.2. Learning Objectives and Protocol

The experimental protocol is designed so that structure learning and evaluation are clearly separated. A train–test split decouples optimization of the graph structure from the assessment of generalization, while repeated random seeds quantify variability due to both sampling noise and random ground-truth draws. Two scoring strategies are compared within the same greedy chordal search mechanism, ensuring that observed differences can be attributed primarily to the scoring function rather than to the optimizer.

Each synthetic run used $(d, n, \rho) = (30, 2000, 0.1)$, and we repeated the experiment over 20 random seeds. For each run, to decouple structure learning from held-out evaluation, the full sample index set $\{1, \ldots, n\}$ was split into disjoint training (70%) and test (30%) sets, denoted by J_{tr} and J_{te} with sizes $n_{\text{tr}} := |J_{\text{tr}}|$ and $n_{\text{te}} := |J_{\text{te}}|$ (and $n = n_{\text{tr}} + n_{\text{te}}$), respectively. All cross-validation steps described in Section 3.3 were performed on the training set J_{tr} only. In the outer train and test evaluation, pseudo-observations for the test points were computed by applying the training empirical marginals to the test values (fit on train and transform test). Structure learning was performed via greedy forward and backward search over decomposable graphs, starting from the empty graph and applying only decomposability-preserving single-edge insertions and deletions until no improving move existed.

We compared three structure-learning procedures. NG-LL [9] and NG-CBIC both use the same greedy chordality-preserving search described above, differing only in the score used to accept or reject candidate edge insertions and deletions. In particular, NG-LL uses a predictive (cross-validated) log-likelihood score built from nonparametric KDE estimates of the clique and separator marginals under the decomposable factorization (i.e., it aggregates clique log-likelihood contributions and subtracts separator contributions). This provides a direct contrast with NG-CBIC, which scores candidate graphs using our copula information objective (copula entropy and mutual information terms) together with a CBIC-type complexity penalty. The third baseline, IC-ILP [10], operates on a discretized version of the data obtained by equal-frequency (quantile) binning of each variable and learns a decomposable structure by solving the resulting discrete optimization problem that maximizes the BIC score over decomposable graphs via an ILP-based procedure. This was evaluated on the same train and test splits and seeds.

Within the NG framework, we considered two alternative scoring functionals for greedy structure learning over decomposable graphs: a predictive-assessment criterion based on clique/separator KDEs evaluated using k-fold cross-validated log-likelihood under clique/separator KDEs (NG-LL) and a copula information criterion with a CBIC-type complexity penalty of the form in Equation (10) using $\lambda(A) = |A|(|A| + 1)/2$.

This protocol enables the study of both accuracy and stability, namely accuracy through edge recovery metrics against a known truth and stability through variability across seeds under fixed (d, n, ρ) values. Because the search started from the empty graph and stopped at a local optimum, the results also indirectly reflect how informative and well-behaved the local deltas were under each score. The following subsection formalizes the recovery metrics used to quantify these outcomes.

4.3. Graph Recovery Metrics

The metrics used to compare the learned graph $\widehat{G}$ with the ground-truth graph G are defined as follows. Because the output is an undirected structure, edge-wise discrepancies were summarized using both an edit distance SHD and classification-style measures (precision, recall, and F1 score) computed from TP, FP, and FN. Taken together, these metrics distinguish between conservative learners (high precision, low recall) and aggressive learn-

ers (high recall, low precision), a contrast that is relevant when scoring functions include explicit complexity penalties.

Let $W \in \{0,1\}^{d\times d}$ denote the symmetric adjacency matrix of an undirected graph $G = (V, E)$, with $W_{ii} = 0$ and $W_{ij} = W_{ji} = 1 \Leftrightarrow \{v_i, v_j\} \in E$. Let G be the ground truth and $\widehat{G}$ be the learned graph, with adjacency matrices W and $\widehat{W}$.

We used the standard SHD, i.e., the number of edge additions and deletions needed to transform $\widehat{G}$ into G [32]. For undirected graphs, this equals

$$\mathrm{SHD}(G, \widehat{G}) = \frac{1}{2}\sum_{i=1}^{d}\sum_{j=1}^{d}\left|W_{ij} - \widehat{W}_{ij}\right|. \tag{15}$$

Define the (upper-triangular) edge sets E and $\widehat{E}$ and the counts:

$$\mathrm{TP} = |\widehat{E} \cap E|, \qquad \mathrm{FP} = |\widehat{E} \setminus E|, \qquad \mathrm{FN} = |E \setminus \widehat{E}|.$$

Then, we have

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}, \qquad \mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}, \qquad \mathrm{F1} = \frac{2\,\mathrm{Precision}\cdot\mathrm{Recall}}{\mathrm{Precision}+\mathrm{Recall}}.$$

The combination of SHD and the F1 score provides complementary views; SHD measures absolute disagreement in edge sets, while the F1 score balances false positives and false negatives in a normalized way. Reporting multiple metrics is especially useful here because different scores may trade off sparsity and fit differently, even when optimized by the same greedy procedure. With these definitions in place, the evaluation extends beyond controlled simulations to a real dataset, where interpretability considerations become central.

4.4. Real Data Experiment

This subsection applies the proposed methodology to a gene expression dataset in which airway epithelial samples are stratified by smoking status (never, former, or current), using the study design and cohort described by Spira et al. [33]. The objective was to assess whether the learning pipeline produced a stable and interpretable chordal dependence summary in the presence of biological heterogeneity and measurement noise, rather than claim causal mechanisms or definitive regulatory structure. Accordingly, the workflow included standard diagnostic checks to characterize dominant variation and identify potential artifacts before fitting a decomposable model. To limit dimensionality (100 variables) and focus on the smoking-related signal, the analysis selected a fixed number of discriminative probes using an ANOVA-type screening criterion, after which structure learning was performed on the reduced probe set.

The learned chordal graph was interpreted as an exploratory conditional (in)dependence representation intended to support downstream biological hypothesis generation under a decomposability constraint. Because microarray platforms can map multiple probes to the same gene, probe-level edges are optionally aggregated to gene-level edges using platform annotation, which improves interpretability and reporting without altering the underlying score, estimator, or search procedure. In this cohort, smoking was reported to induce broad transcriptional changes in airway epithelial cells, including the upregulation of xenobiotic metabolism and redox-related genes, which motivates examining whether highly connected nodes in the learned dependence graph correspond to these functional categories. Overall, the real-data analysis complements the synthetic experiments by demonstrating the feasibility of chordal structure learning and by illustrating how annotation-aware post-processing can improve communication of graph structure in applied transcriptomic settings.

5. Results and Discussion

This section evaluates the proposed copula information learning criterion in both controlled and applied settings, emphasizing empirical structure recovery and qualitative interpretability of the learned decomposable graphs. We first report a synthetic benchmark (Table 1), where the ground-truth chordal structure is known and standard graph-recovery metrics (e.g., F1 score and SHD) can be computed directly. In this benchmark, we additionally vary the univariate marginals (Gaussian, exponential, and beta) via component-wise monotone transforms while holding the copula (and thus the ground-truth conditional independence structure) fixed, directly reflecting the paper's motivation regarding marginal heterogeneity. We then present a real gene expression application based on airway epithelial samples stratified by smoking status [33], using the learned network structure as an exploratory conditional (in)dependence summary and illustrating key local patterns via a hub-centered sub-network visualization (Figure 1).

Table 1. Synthetic benchmark across marginal families from three distinct supports (Section 4.1): edge recovery and runtime for decomposable structure learning. Across the three marginal families, we reused the same pseudo-observations and applied only component-wise quantile transforms, so the copula (and ground-truth conditional independence structure) was held fixed. Mean ± standard deviation over 20 random seeds for a chordal synthetic experiment with $d = 30$ variables and $n = 2000$ samples (with the specified edge probability setting and train and test protocol described in Section 4.2, comparing NG-LL [9] (nonparametric log-likelihood), NG-CBIC (proposed copula information score with CBIC-type penalty), and IC-ILP [10] (discrete or coarsening ILP baseline with quantile binning). Metrics include F1 score, precision, recall, SHD, and wall clock time in seconds. NG-CBIC rows are identical across marginal families, visually demonstrating margin invariance. NG-LL varies across marginals, indicating marginal sensitivity.

Method	Marginals	F1 Score	Precision	Recall	SHD	Time [s]
NG–LL	Gaussian	0.76 ± 0.05	0.68 ± 0.08	0.88 ± 0.06	45.35 ± 9.42	33.67 ± 7.15
NG–CBIC	Gaussian	0.86 ± 0.04	0.94 ± 0.03	0.80 ± 0.07	22.65 ± 10.86	22.25 ± 4.63
IC–ILP	Gaussian	0.47 ± 0.08	0.89 ± 0.06	0.32 ± 0.08	62.3 ± 20.31	8.73 ± 1.51
NG–LL	Exponential	0.66 ± 0.06	0.77 ± 0.07	0.59 ± 0.09	51.25 ± 9.02	16.16 ± 1.55
NG–CBIC	Exponential	0.86 ± 0.04	0.94 ± 0.03	0.80 ± 0.07	22.65 ± 10.86	20.19 ± 3.85
IC-ILP	Exponential	0.47 ± 0.08	0.89 ± 0.06	0.32 ± 0.08	62.3 ± 20.31	8.69 ± 1.40
NG–LL	Beta	0.73 ± 0.03	0.68 ± 0.05	0.81 ± 0.08	47.70 ± 13.15	26.60 ± 4.92
NG–CBIC	Beta	0.86 ± 0.04	0.94 ± 0.03	0.80 ± 0.07	22.65 ± 10.86	21.05 ± 3.82
IC-ILP	Beta	0.47 ± 0.08	0.89 ± 0.06	0.32 ± 0.08	62.3 ± 20.31	7.77 ± 1.37

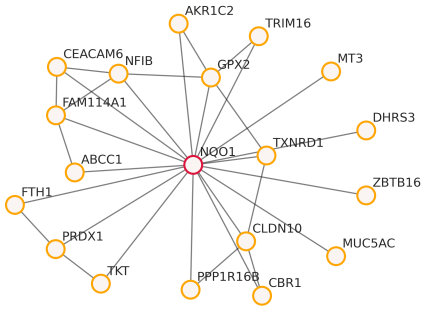


Figure 1. NQO1-centered ego network from the learned airway epithelial dependence graph. Radius-1 ego-graph centered at NQO1 extracted from the decomposable chordal graph learned on the airway epithelial microarray dataset after the real-data preprocessing and feature-screening protocol [33]. Nodes correspond to genes, and edges denote undirected conditional (in)dependence relationships in the fitted model (interpretable as co-expression associations after conditioning on the rest of the selected variables), highlighting a redox and detoxification neighborhood (e.g., GPX2, TXNRD1, PRDX1, and ABCC1) that is statistically linked to epithelial markers such as CEACAM6 and CLDN10.

5.1. Synthetic Data

Table 1 reports the structure recovery performance on a chordal synthetic benchmark with $d = 30$ variables and $n = 2000$ observations, averaged over 20 random seeds. The three marginal blocks (Gaussian, exponential, and beta) were generated by reusing the same pseudo-observations and applying only component-wise monotone quantile transforms, so the copula (and hence the ground-truth conditional independence structure) was held fixed across marginal families. This design separates two effects: (1) Monte Carlo variability of a learning criterion when the dependence structure is unchanged. In particular, methods that depend only on ranks or pseudo-observations recover the same graph within each seed across marginal families, whereas methods that score densities on the original scale can change with the marginal regime.

This pattern is visually clear in Table 1. NG-CBIC was essentially unchanged across Gaussian, exponential, and beta marginals (e.g., identical F1 scores and SHD across families), directly illustrating margin invariance, while NG-LL varied across the same marginal blocks, reflecting sensitivity to marginal distributions. Across all metrics, the proposed NG-CBIC achieved the strongest overall recovery, with the highest F1 score (0.86 ± 0.04) and the lowest SHD (22.65 ± 10.86), and its metrics were identical across the three marginal families under this protocol. In contrast, NG-LL varied substantially with the marginal regime; it attained an $F1 = 0.76 \pm 0.05$ with SHD $= 45.35 \pm 9.42$ under Gaussian marginals but dropped to $F1 = 0.66 \pm 0.06$ with SHD $= 51.25 \pm 9.02$ under exponential marginals. Finally, IC-ILP attained the lowest F1 score (0.47 ± 0.08) and the highest SHD (62.30 ± 20.31), despite its high precision (0.89 ± 0.06), indicating a conservative regime dominated by false negatives.

The precision and recall profiles clarify these differences. NG-LL exhibited high recall under normal marginals (0.88 ± 0.06) but relatively low precision (0.68 ± 0.08), consistent with an edge-including behavior that favors sensitivity at the cost of false discoveries. The proposed NG-CBIC reversed this pattern by achieving extremely high precision (0.94 ± 0.03) while maintaining strong recall (0.80 ± 0.07), which directly explains its superior F1 score and SHD. IC-ILP was the most conservative method (recall 0.32 ± 0.08), and this under-selection dominated its overall error rate, even though its precision remained high.

In Table 1, both IC-ILP and NG-CBIC exhibited identical performance across the normal, exponential, and beta marginal regimes. For IC-ILP, this was expected because equal-frequency (quantile) binning is determined by the ordering of the sample (empirical ranks) [34]. Since ranks are invariant under strictly monotone marginal transformations [35], the discretized dataset, and therefore the ILP objective and recovered graph, remained unchanged across marginal families generated by component-wise strictly increasing transforms of the same latent pseudo-observations. NG-CBIC was computed from pseudo-observations (empirical copula transforms) and thus depended only on ranks. Under the same protocol, this yielded the same learned graph across marginal families for each seed for both methods. Accordingly, the reported standard deviations summarize the Monte Carlo variability across random seeds (independent datasets), rather than variation across marginal families.

The runtime results show that IC-ILP was fastest, while NG-CBIC achieved higher accuracy than NG-LL at a similar runtime. This is consistent with the score and penalty combination leading to fewer unproductive edge modifications during the same chordality-preserving greedy search, thereby improving recovery without increasing the overall computational burden.

Dependence-focused scoring improved structural fidelity. The synthetic results demonstrate that optimizing decomposable structures using copula information with a CBIC-type

penalty yielded substantially fewer structural errors than a likelihood-driven nonparametric baseline. In particular, NG-CBIC approximately halved the SHD relative to NG-LL while simultaneously increasing the F1 score. This improvement is consistent with the design goal of isolating dependence from one-dimensional marginal effects, which is central to copula-based information measures. The key implication is that for continuous-variable structure learning, a dependence-only objective can yield more reliable edge sets than directly optimizing the predictive likelihood of multivariate KDE factors in cliques and separators.

The copula formulation provides invariance to marginal artifacts. A central theoretical advantage of the proposed approach is that mutual information can be expressed as (negative) copula entropy so that the leading fit term depends only on the copula density rather than on marginal densities. This property makes the score invariant to strictly monotone transformations of individual variables (i.e., reparameterizations that preserve ranks), which directly targets conditional dependence rather than the marginal scale and shape. In practical terms, this invariance mitigates a common failure mode in continuous modeling where marginal irregularities can be spuriously absorbed by additional edges when the likelihood is optimized under flexible, nonparametric density models. The empirical signature of this effect appears in Table 1; NG-LL increased recall but paid a large precision cost, while NG-CBIC maintained high recall without inducing analogous false positive inflation.

Regularization explains the precision gains and stabilizes greedy search. The large precision jump from NG-LL to the proposed method indicates that explicit complexity control is essential, even when the search space is restricted to decomposable (chordal) graphs. Because decomposable models admit clique/separator factorizations, penalties that decompose over cliques and separators preserve locality, enabling score differences to be computed from small set changes induced by chordality-preserving edge edits. This locality is not merely computationally convenient; it also shapes the optimization landscape by discouraging diffuse sequences of small likelihood gains that accumulate into overly dense graphs. The resulting behavior is empirically favorable here; the proposed method sacrifices some recall relative to NG-LL but achieves substantially lower SHD results, meaning that the removed edges are predominantly false positives rather than true dependencies.

Nonparametric copula estimation requires explicit treatment of boundary bias. The statistical reliability of copula entropy estimation hinges on estimating densities supported on multidimensional unit hypercubes, where naive kernel smoothing suffers from boundary distortion. The probit transformation strategy addresses this issue by mapping pseudo-observations to unlimited supports, applying KDE in an unconstrained domain, and correcting via the Jacobian upon mapping back, which is a standard approach in nonparametric copula density estimation. This choice is consistent with the need for repeated evaluation of many overlapping clique/separator subsets during greedy search, where even mild systematic bias can accumulate into persistent edge selection errors. In this sense, the proposed estimator is not an auxiliary detail but a prerequisite for the observed precision improvements, because it reduces the tendency of the score to overreact to artifacts near the boundaries of the copula domain.

Decomposability constrains expressiveness but enables scalable and well-defined learning. Restricting to decomposable graphs is a structural bias that trades representational flexibility for algorithmic tractability and score locality. On the positive side, chordality guarantees a junction tree representation, making global quantities decomposable and enabling efficient updates under local edits, which is exactly the regime required by stepwise search procedures. On the negative side, non-chordal dependence patterns must be approximated by a chordal completion, and thus learned edges can reflect the best

chordal surrogate rather than the exact original sparse structure when the data-generating mechanism is not decomposable. The present synthetic protocol explicitly constructs chordal ground truths via triangulation, and thus the reported gains primarily reflect differences in scoring and estimation rather than a mismatch between the truth and the hypothesis class.

Discrete or coarsening-based approaches can underperform in continuous problems. The conservative behavior of IC-ILP (precision of 0.89 with recall of 0.32) suggests that its optimization biases favor sparse backbones and may fail to recover weaker conditional dependencies in continuous settings when the scoring machinery is not tailored to copula- or density-based continuous criteria. More broadly, discrete decomposable model learning pipelines are susceptible to quantization effects when continuous variables are binned, because discretization can discard dependence information and introduce bin edge artifacts that propagate into structure learning. This issue is especially relevant in scientific measurements, where binning choices can dominate the effective sample size within contingency tables (e.g., biomedical signals). Consequently, the results support a principled division of labor: coarsening and ILP methods remain valuable for large-scale discrete decomposable learning, whereas copula information scoring provides a more appropriate inductive bias for continuous dependence learning.

5.2. Real Data

Inspecting the immediate neighborhood of NQO1 in Figure 1 is useful because, in a learned conditional (in)dependence graph, a hub's one-hop ego-network provides a compact view of the strongest local association structure around a biologically salient marker without requiring global, hard-to-parse topology. In the implied radius-one ego-network figure, NQO1 is the highest-degree node, connecting to an antioxidant- and redox-associated set (GPX2, TXNRD1, and PRDX1), a small metabolic motif (TKT and FTH1 linked via PRDX1 and also adjacent to NQO1), and an epithelial marker- and phenotype-associated branch (CEACAM6, CLDN10, and MUC5AC). Two aspects are especially worth highlighting within this local view; TRIM16 connects into the antioxidant pocket via TRIM16–GPX2, and MT3 (a metallothionein) appears directly adjacent to NQO1, which is notable given that metallothionein behavior can differ between acute and chronic smoke exposure contexts.

In the manuscript's real-data experiment, the airway epithelial gene expression data serve as an applied demonstration for the proposed decomposable structure-learning framework, so the edges should be read strictly as conditional (in)dependence–co-expression relationships under a tractable chordal model rather than as regulatory directionality. At a coarse functional level, the NQO1 neighborhood is consistent with smoking-associated airway epithelial transcriptional shifts reported in the classic bronchial brushing study by Spira et al. [33]. The same paper also reported increased epithelial adhesion markers in smokers, which aligns descriptively with CEACAM6 and CLDN10 sitting in the same immediate neighborhood as the redox/xenobiotic module.

Beyond this corroborative signal, the TRIM16–GPX2 linkage is plausibly interesting because TRIM16 has been described as participating in oxidative and proteotoxic stress programs through the p62–KEAP1–NRF2 axis, making its proximity to a canonical detox or antioxidant pocket a coherent, non-mechanistic highlight [36]. Likewise, MT3's appearance near NQO1 can be highlighted cautiously because airway metallothionein genes have been reported to show discordant responses across exposure regimes, and thus a local association between MT3 and an oxidative stress hub could be context-sensitive rather than universally expected [37]. From a statistical standpoint, these compact neighborhood-level findings illustrate what the proposed method can improve in network analysis (such as in gene expression analysis): dependence-focused edge selection that is more robust to

marginal artifacts (via copula-based information measures) while remaining interpretable and computationally manageable through decomposable structure constraints.

5.3. Limitations

This work deliberately focused on DM because chordality is the structural condition that enables a junction tree representation and avoids computation of the intractable partition function [3,4]. That representation is what makes clique/separator factorization available and supports tractable exact inference together with genuinely local score updates during structure search, which are central to the approach developed here. As a consequence, the method is best understood as learning an interpretable tractable of dependence rather than an unconstrained representation of a conditional-independence structure that may exist in the data [4].

The chordal restriction implies an unavoidable modeling bias when the data-generating dependence is not itself chordal. In real applications, non-chordal dependence patterns may be common, and a decomposable model can only approximate such patterns within the chordal family. In practice, this approximation often manifests through additional edges needed to maintain chordality, which can reduce sparsity and complicate interpretation because some edges may serve primarily to support a chordal scaffold rather than indicate a direct scientific relationship.

The restriction to decomposable graphs also interacts with finite-sample estimation through the clique/separator organization of the score. Section 3 instantiates clique/separator copula entropy terms using pseudo-observations, a probit-based transformation, and predictive evaluation to mitigate boundary effects on the unit hypercube. Even with these design choices, the reliability of repeated nonparametric copula entropy evaluations can deteriorate as clique and separator sizes grow, because the estimation problems become higher-dimensional, more bandwidth-sensitive, and more variable in finite samples. When chordal approximations induce larger cliques than one would obtain under an unconstrained representation, this dimensionality effect can translate into increased runtimes and additional variability in local score comparisons during a search.

The optimization procedure reflects a deliberate trade-off between scalability and global optimality. The proposed method uses greedy, chordality-preserving single-edge moves to keep the search computationally feasible while maintaining decomposability at every step. As with greedy procedures in combinatorial graph spaces, this choice does not guarantee convergence to a global optimum of the objective over the decomposable class, and finite-sample numerical behavior can depend mildly on the maintained junction tree representation because separator contributions enter according to that representation's induced multiplicities, even though the population-level decomposition is a property of the underlying decomposable graph.

Finally, copula-based modeling provides a powerful tool for constructing joint distributions. However, researchers employing this technique should be aware of several limitations. One challenge arises when using copulas with discrete marginal distributions. Sklar's theorem [11] demonstrates that in such cases, a unique copula representation for the joint distribution is not guaranteed. This lack of uniqueness can translate into identifiability issues [38] for the model parameters.

6. Conclusions

This work studied structure learning for continuous decomposable (chordal) Markov random fields using an information-theoretic objective defined on copulas to learn dependence graphs that are robust to marginal artifacts. The central methodological conclusion is that combining copula-based mutual information principles with clique/separator de-

composability yields a scoring rule that remains compatible with chordality-preserving local search while targeting dependence rather than a marginal shape. This conclusion directly addresses the motivating gap that likelihood-driven nonparametric decomposable models can be accurate yet sensitive to marginal irregularities and smoothing choices in high-variance continuous data.

The empirical findings support this dependence-focused perspective on a controlled chordal synthetic benchmark. On this benchmark, the proposed NG-CBIC method achieved the highest F1 score and the lowest SHD, indicating substantially improved recovery of the ground-truth edge set. Compared with the predictive assessment likelihood baseline NG-LL, NG-CBIC approximately halved the SHD while improving the F1 score, which is consistent with reduced false positive inflation under marginally robust scoring. Compared with IC-ILP, NG-CBIC improved the recall while maintaining high precision, showing that the information-based criterion is less prone to the under-selection behavior observed in sparse-backbone learners.

One practical implication is that for continuous-variable structure learning under heterogeneous marginals, copula information scoring with an explicit complexity penalty can yield graphs that are both more accurate and more interpretable than graphs learned by optimizing the nonparametric likelihood alone. A second practical implication is that stable copula entropy estimation is not an implementation detail but a necessary component for reliable edge decisions in repeated local updates, motivating the use of probit-based boundary correction for KDE on multidimensional unit hypercubes. The real-data experiment further indicates that the approach can be applied end to end to noisy gene expression measurements to produce an exploratory dependence summary that supports domain interpretation while remaining explicit about its non-causal status.

Several directions appear particularly important for future works, such as strengthening the complexity control to more directly discourage large cliques while preserving clique/separator additivity; adding stability assessments (for instance, via resampling) to identify edges that are not robust under finite-sample variability; complementing greedy search with junction tree-based sampling to quantify uncertainty over decomposable structures; and developing faster or more stable estimators tailored to repeated clique/separator evaluation. Additionally, a theory is needed to characterize when chordal restrictions and chordal completions affect edge recovery under copula-based scores, especially beyond treewidth-one regimes, where sharp recoverability results are currently better understood. Uncertainty quantification over decomposable structures could be integrated by combining the present dependence score with modern junction tree samplers for decomposable graphs, enabling posterior summaries or stability measures rather than a single selected graph. Scaling the copula entropy evaluation to larger cliques and higher dimensions remains an open computational problem. In addition to parallel programming, the clique/separator decomposition suggests a natural fit with distributed and federated settings, where copula entropy contributions can be computed locally and combined without sharing raw observations, motivating new approximations and caching strategies that preserve the locality exploited by chordal search. Finally, extending the framework to non-IID regimes (e.g., temporal dependence) would broaden applicability but would also require scores and estimators that remain valid under dependence while preserving decomposable updates.

Author Contributions: Conceptualization, L.D.N. and C.D.M.; Methodology, L.D.N.; Software, L.D.N.; Validation, L.D.N. and C.D.M.; Formal analysis, L.D.N. and C.D.M.; Investigation, L.D.N.; Resources, L.D.N. and C.D.M.; Data curation, L.D.N.; Writing—original draft, L.D.N.; Writing—review & editing, L.D.N., H.d.O.C., M.d.S.S.F. and C.D.M.; Visualization, L.D.N.; Supervision, C.D.M.; Project administration, L.D.N. and C.D.M.; Funding acquisition, C.D.M. All authors have read and agreed to the published version of the manuscript.

Funding: The authors of this work would like to thank the Center for Artificial Intelligence in Health Management (ciaGsaude); FAPESP 2014/50851-0, CNPq 465755/2014-3, BPE Fapesp 2018/19150-6, FAPESP 2021/12220-1, and CAPES-Brazil.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are openly available in Gene Expression Omnibus Database at https://doi.org/10.1007/978-1-4939-3578-9_5.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

ANOVA	Analysis of variance
BIC	Bayesian information criterion
BN	Bayesian network
CBIC	Continuous Bayesian information criterion
CDF	Cumulative distribution function
CV	Cross-validation
DAG	Directed acyclic graph
DM	Decomposable model
IC	Iterative coarsening
IID	Independent and identically distributed
ILP	Integer linear programming
KDE	Kernel density estimator
LL	Log-likelihood
MRF	Markov random field
NG	Non-parametric greedy
PDF	Probability density function
PGM	Probabilistic graphical model
PPF	Percent-point function
SHD	Structural Hamming distance
UG	Undirected graph
Symbols	
$\mathbb{R}$	Real numbers
$\mathbb{R}_+$	Nonnegative reals
$\mathbb{N}$	Natural numbers
$\mathbb{Z}_+$	Nonnegative integers
$\mathbb{E}$	Expectation
d	Dimension (number of variables or nodes)
n	Sample size
J_{tr}	Training index set
J_{te}	Test index set
n_{tr}	Training sample size
n_{te}	Test sample size
k	Fold count in k-fold cross-validation
i, j	Variable indices $i, j \in \{1, \ldots, d\}$
t	Sample index $t \in \{1, \ldots, n\}$
$(\Omega, \mathscr{A}, \mathbb{P})$	Probability space
$\mathbf{X} = (X_1, \ldots, X_d)^{\mathrm{T}}$	Continuous random vector in $\mathbb{R}^d$
$F_{\mathbf{X}}$, $F(\mathbf{x})$	Joint CDF of $\mathbf{X}$

$p_{\mathbf{X}}$	Joint PDF of $\mathbf{X}$
F_i	Marginal CDF of X_i
p_i	Marginal PDF of X_i
C	Copula function
c	Copula density
U_i	Distributional transform $U_i = F_i(X_i)$
$\mathbf{U}$	Transformed random vector $\mathbf{U} = (U_1, \ldots, U_d)^\mathsf{T}$
$\mathbf{u}$	Generic point in $[0,1]^d$
h	Differential entropy
h_c	Copula entropy
I	Mutual information
A	Index set $A \subseteq \{1, \ldots, d\}$
$\lvert\cdot\rvert$	Cardinality of a set, absolute value
A^{c}	Complement of A in $\{1, \ldots, d\}$
π_A	Coordinate projection $\pi_A : \mathbb{R}^d \to \mathbb{R}^{\lvert A\rvert}$
$\mathbf{X}_A$	Subvector $\mathbf{X}_A = \pi_A(\mathbf{X})$
$\mathbf{X}_{-A}$	Complement subvector $\mathbf{X}_{-A} = \mathbf{X}_{A^{\mathrm{c}}}$
$\mathbf{U}_A$	Subvector $\mathbf{U}_A = \pi_A(\mathbf{U})$
G	Undirected graph
V	Vertex set $V = \{v_1, \ldots, v_d\}$
E	Edge set $E \subseteq V \times V$
v_i	Vertex
$\mathcal{K}$	Maximal clique set $\mathcal{K}$
K_l	l-th maximal clique in in running-intersection ordering
S_l	l-th separator set in running-intersection ordering ($S_1 = \emptyset$)
m	Number of maximal cliques
η	Predecessor index in running-intersection ordering
γ	Copula marginal index set γ (for γ marginal)
α_l	Bijection $\alpha_l : \{1, \ldots, \lvert K_l\rvert\} \to K_l$ (argument order)
γ_l	Separator coordinate positions $\gamma_l = \alpha_l^{-1}(S_l)$
c_A	Copula density on margin set indexed by A
c_{K_l}	Clique copula density
c_{S_l}	Separator copula density $c_{S_l} \equiv c_{K_l}^{\gamma_l}$
ψ	Local information functional
$\mathcal{I}$	Global information
Λ	Complexity functional
λ	Local complexity term
$\mathcal{S}_{\mathrm{CBIC}}$	CBIC-type score
Δ_{ij}	Local move score change
κ	Local score functional
χ	Indicator function
$\widehat{F}$	Empirical marginal CDF
$\widehat{U}$	Pseudo-observation
Φ	Standard normal CDF
ϕ	Standard normal PDF
$\widehat{Z}$	Probit-transformed pseudo-observation
$\boldsymbol{\theta}$	Global bandwidth vector $\boldsymbol{\theta} \in \mathbb{R}_+^d$
$\boldsymbol{\theta}_A$	Bandwidth subvector
g	Gaussian kernel
$\widehat{c}$	Copula density estimator
J_{-r}	k-fold training index set
$\widehat{\psi}$	Local information estimator
$\mathfrak{G}_{\mathrm{dec}}$	Decomposable graph class
$\mathcal{N}(G)$	Valid neighborhood $\mathcal{N}(G) \subset \mathfrak{G}_{\mathrm{dec}}$

https://doi.org/10.3390/e28030293

T	Junction tree $T = (\mathcal{K}, \mathcal{F})$
$\mathcal{F}$	Clique tree edge set
w	Join tree edge weight
ρ	Erdős–Rényi edge probability
τ	Triangulation/chordal completion map
D	Degree matrix
I_d	Identity matrix of size d
Q	Precision matrix
ν	Ridge parameter
W	Adjacency matrix
$\widehat{W}$	Estimated adjacency matrix
$\mathrm{SHD}(G, \widehat{G})$	Structural Hamming distance between G and $\widehat{G}$

References

1. Rios, F.L.; Moffa, G.; Kuipers, J. Benchpress: A Versatile Platform for Structure Learning in Causal and Probabilistic Graphical Models. *J. Stat. Softw.* **2025**, *114*, 1–43. [CrossRef]
2. Kitson, N.K.; Constantinou, A.C.; Guo, Z.; Liu, Y.; Chobtham, K. A survey of Bayesian Network structure learning. *Artif. Intell. Rev.* **2023**, *56*, 8721–8814. [CrossRef]
3. Brémaud, P. *Markov Chains. Gibbs Fields, Monte Carlo Simulation and Queues*, 2nd ed.; Volume 31: Texts in Applied Mathematics; Springer: Cham, Switzerland, 2020. [CrossRef]
4. Pearl, J. Chapter 3—MARKOV AND BAYESIAN NETWORKS: Two Graphical Representations of Probabilistic Knowledge. In *Probabilistic Reasoning in Intelligent Systems*; Pearl, J., Ed.; Morgan Kaufmann: San Francisco, CA, USA, 1988; pp. 77–141. [CrossRef]
5. Huang, Y.; Xiang, Y. Learning Bayesian networks by learning decomposable Markov networks first. In *Engineering Solutions for the Next Millennium. 1999 IEEE Canadian Conference on Electrical and Computer Engineering (Cat. No.99TH8411)*; IEEE: Piscataway, NJ, USA , 1999; Volume 3, pp. 1704–1709. [CrossRef]
6. Olsson, J.; Pavlenko, T.; Rios, F.L. Sequential sampling of junction trees for decomposable graphs. *Stat. Comput.* **2022**, *32*, 80. [CrossRef]
7. Katiyar, A.; Basu, S.; Shah, V.; Caramanis, C. Recoverability Landscape of Tree Structured Markov Random Fields under Symmetric Noise. In Proceedings of the 25th International Conference on Artificial Intelligence and Statistics, Virtual, 28–30 March 2022; Camps-Valls, G., Ruiz, F.J.R., Valera, I., Eds.; Proceedings of Machine Learning Research (PMLR): New York, NY, USA, 2022; Volume 151, pp. 9356–9399.
8. Pfeifer, D.; Kovács, E.A. Vine copula structure representations using graphs and matrices. *Inf. Sci.* **2024**, *662*, 120151. [CrossRef]
9. Schwaighofer, A.; Dejori, M.; Tresp, V.; Stetter, M. Structure Learning with Nonparametric Decomposable Models. In *Proceedings of the Artificial Neural Networks—ICANN 2007*; de Sá, J.M., Alexandre, L.A., Duch, W., Mandic, D., Eds.; Springer: Berlin/Heidelberg, Germany, 2007; pp. 119–128.
10. Orfanides, G.; Pérez, A. Learning decomposable models by coarsening. In Proceedings of the 10th International Conference on Probabilistic Graphical Models, Skørping, Denmark, 23–25 Sepember 2020; Jaeger, M., Nielsen, T.D., Eds.; Proceedings of Machine Learning Research (PMLR): New York, NY, USA, 2020; Volume 138, pp. 317–328.
11. Sklar, M. Fonctions de Répartition à n Dimensions et Leurs Marges. *Ann. l'ISUP* **1959**, *8*, 229–231.
12. Ma, J.; Sun, Z. Mutual Information Is Copula Entropy. *Tsinghua Sci. Technol.* **2011**, *16*, 51–54. [CrossRef]
13. Lasserre, M.; Lebrun, R.; Wuillemin, P.H. Learning Continuous High-Dimensional Models using Mutual Information and Copula Bayesian Networks. *Proc. AAAI Conf. Artif. Intell.* **2021**, *35*, 12139–12146. [CrossRef]
14. Geenens, G.; Charpentier, A.; Paindaveine, D. Probit transformation for nonparametric kernel estimation of the copula density. *Bernoulli* **2017**, *23*, 1848–1873. [CrossRef]
15. Ibarra, L. Fully dynamic algorithms for chordal graphs and split graphs. *ACM Trans. Algorithms* **2008**, *4*, 1–20. [CrossRef]
16. Provost, S.B.; Zang, Y. Nonparametric Copula Density Estimation Methodologies. *Mathematics* **2024**, *12*, 398. [CrossRef]
17. Durante, F.; Sempi, C. *Principles of Copula Theory*, 1st ed.; Chapman and Hall/CRC: Boca Raton, FL, USA, 2015. [CrossRef]
18. Rønn-Nielsen, A.; Hansen, E. *Conditioning and Markov Properties*; Department of Mathematical Sciences, Københavns Universitet: København, Denmark, 2014.
19. Desuó Neto, L. Untying the Gordian Knot of Bayesian Inference in Markov Networks. Ph.D. Thesis, Dynamic Systems, São Carlos School of Engineering, University of São Paulo, São Carlos, Brazil, 2025. [CrossRef]
20. Kraskov, A.; Stögbauer, H.; Grassberger, P. Estimating mutual information. *Phys. Rev. E* **2004**, *69*, 066138. [CrossRef]
21. Zhu, C.; Byrd, R.H.; Lu, P.; Nocedal, J. Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization. *ACM Trans. Math. Softw.* **1997**, *23*, 550–560. [CrossRef]

22. Byrd, R.H.; Lu, P.; Nocedal, J.; Zhu, C. A Limited Memory Algorithm for Bound Constrained Optimization. *SIAM J. Sci. Comput.* **1995**, *16*, 1190–1208. [CrossRef]
23. Morales, J.L.; Nocedal, J. Remark on "algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound constrained optimization". *ACM Trans. Math. Softw.* **2011**, *38*, 1–4. [CrossRef]
24. Segers, J. Asymptotics of empirical copula processes under non-restrictive smoothness assumptions. *Bernoulli* **2012**, *18*, 764–782. [CrossRef]
25. Joe, H. Estimation of entropy and other functionals of a multivariate density. *Ann. Inst. Stat. Math.* **1989**, *41*, 683–697. [CrossRef]
26. Hall, P.; Presnell, B. Density Estimation under Constraints. *J. Comput. Graph. Stat.* **1999**, *8*, 259–277. [CrossRef]
27. Stone, M. An Asymptotic Equivalence of Choice of Model by Cross-Validation and Akaike's Criterion. *J. R. Stat. Soc. Ser. B (Methodol.)* **1977**, *39*, 44–47. [CrossRef]
28. Arlot, S.; Lerasle, M. Choice of V for V-Fold Cross-Validation in Least-Squares Density Estimation. *J. Mach. Learn. Res.* **2016**, *17*, 1–50.
29. Berry, A.; Blair, J.R.S.; Heggernes, P.; Peyton, B.W. Maximum Cardinality Search for Computing Minimal Triangulations of Graphs. *Algorithmica* **2004**, *39*, 287–298. [CrossRef]
30. Dempster, A.P. Covariance Selection. *Biometrics* **1972**, *28*, 157–175. [CrossRef]
31. Jones, B.; West, M. Covariance decomposition in undirected Gaussian graphical models. *Biometrika* **2005**, *92*, 779–786. [CrossRef]
32. Tsamardinos, I.; Brown, L.E.; Aliferis, C.F. The max-min hill-climbing Bayesian network structure learning algorithm. *Mach. Learn.* **2006**, *65*, 31–78. [CrossRef]
33. Spira, A.; Beane, J.; Shah, V.; Liu, G.; Schembri, F.; Yang, X.; Palma, J.; Brody, J.S. Effects of cigarette smoke on the human airway epithelial cell transcriptome. *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 10143–10148. [CrossRef]
34. ouriarhli, B.; Benyacoub, B.; Benazza, H. An Evaluation of Discretization Techniques for HMM-Based Classifiers. *Stat. Optim. Inf. Comput.* **2026** , *15*, 1285–1298. [CrossRef]
35. Brunner, E.; Puri, M.L. 19 Nonparametric methods in design and analysis of experiments. In *Design and Analysis of Experiments*; Volume 13: Handbook of Statistics; Elsevier: Amsterdam, The Netherlands, 1996; pp. 631–703. [CrossRef]
36. Jena, K.K.; Kolapalli, S.P.; Mehto, S.; Nath, P.; Das, B.; Sahoo, P.K.; Ahad, A.; Syed, G.H.; Raghav, S.K.; Senapati, S.; et al. TRIM16 controls assembly and degradation of protein aggregates by modulating the p62-NRF2 axis and autophagy. *EMBO J.* **2018**, *37*, EMBJ201798358. [CrossRef]
37. Billatos, E.; Faiz, A.; Gesthalter, Y.; LeClerc, A.; Alekseyev, Y.O.; Xiao, X.; Liu, G.; ten Hacken, N.H.T.; Heijink, I.H.; Timens, W.; et al. Impact of acute exposure to cigarette smoke on airway gene expression. *Physiol. Genom.* **2018**, *50*, 705–713. [CrossRef] [PubMed]
38. Genest, C.; Nešlehová, J. A Primer on Copulas for Count Data. *ASTIN Bull.* **2007**, *37*, 475–515. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.