Modeling and Analysis of Non-Linear Dependencies using Copulas, with Applications to Machine Learning

Kiran Karra

Dissertation submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Electrical Engineering

Lamine Mili, Chair
T. Charles Clancy
Guoqiang Yu
Sanjay Raman
Narendran Ramakrishnan

July 26, 2018
Arlington, Virginia

Keywords: Copula, Machine Learning, Statistical Association

Modeling and Analysis of Non-Linear Dependencies using Copulas, with Applications to Machine Learning

Kiran Karra

(ABSTRACT)

Many machine learning (ML) techniques rely on probability, random variables, and stochastic modeling. Although statistics pervades this field, there is a large disconnect between the copula modeling and the machine learning communities. Copulas are stochastic models that capture the full dependence structure between random variables and allow flexible modeling of multivariate joint distributions. Elidan was the first to recognize this disconnect, and introduced copula based models to the ML community that demonstrated magnitudes of order better performance than the non copula-based models Elidan [2013]. However, the limitation of these is that they are only applicable for continuous random variables and real world data is often naturally modeled jointly as continuous and discrete. This report details our work in bridging this gap of modeling and analyzing data that is jointly continuous and discrete using copulas.

Our first research contribution details modeling of jointly continuous and discrete random variables using the copula framework with Bayesian networks, termed Hybrid Copula Bayesian Networks (HCBN) [Karra and Mili, 2016], a continuation of Elidan's work on Copula Bayesian Networks Elidan [2010]. In this work, we extend the theorems proved by Nešlehová [2007] from bivariate to multivariate copulas with discrete and continuous marginal distributions. Using the multivariate copula with discrete and continuous marginal distributions as a theoretical basis, we construct an HCBN that can model all possible permutations of discrete and continuous random variables for parent and child nodes, unlike the popular conditional linear Gaussian network model. Finally, we demonstrate on numerous synthetic datasets and a real life dataset that our HCBN compares favorably, from a modeling and flexibility viewpoint, to other hybrid models including the conditional linear Gaussian and the mixture of truncated exponentials models.

Our second research contribution then deals with the analysis side, and discusses how one may use copulas for exploratory data analysis. To this end, we introduce a nonparametric copula-based index for detecting the strength and monotonicity structure of linear and nonlinear statistical dependence between pairs of random variables or stochastic signals. Our index, termed Copula Index for Detecting Dependence and Monotonicity (*CIM*), satisfies several desirable properties of measures of association, including Rényi's properties, the data processing inequality (DPI), and consequently self-equitability. Synthetic data simulations reveal that the statistical power of *CIM* compares favorably to other state-of-the-art measures of association that are proven to satisfy the DPI. Simulation results with real-world data reveal *CIM*'s unique ability to detect the monotonicity structure among stochastic signals to find interesting dependencies in large datasets. Additionally, simulations show that *CIM* shows favorable performance to estimators of mutual information when discovering Markov network structure.

Our third research contribution deals with how to assess an estimator's performance, in the scenario where multiple estimates of the strength of association between random variables need to be rank ordered. More specifically, we introduce a new property of estimators of the strength of statistical association, which helps characterize how well an estimator will perform in scenarios where dependencies between continuous and discrete random variables need to be rank ordered. The new property, termed the estimator response curve, is easily computable and provides a marginal distribution agnostic way to assess an estimator's performance. It overcomes notable drawbacks of current metrics of assessment, including statistical power, bias, and consistency. We utilize the estimator response curve to test various measures of the strength of association that satisfy the data processing inequality (DPI), and show that the *CIM* estimator's performance compares favorably to *kNN*, *vME*, *AP*, and H_{MI} estimators of mutual information. The estimators which were identified to be suboptimal, according to the estimator response curve, perform worse than the more optimal estimators when tested with real-world data from four different areas of science, all with varying dimensionalities and sizes.

Modeling and Analysis of Non-Linear Dependencies using Copulas, with Applications to Machine Learning

Kiran Karra

(GENERAL AUDIENCE ABSTRACT)

Many machine learning (ML) techniques rely on probability, random variables, and stochastic modeling. Although statistics pervades this field, many of the traditional machine learning techniques rely on linear statistical techniques and models. For example, the correlation coefficient, a widely used construct in modern data analysis, is only a measure of linear dependence and cannot fully capture non-linear interactions. In this dissertation, we aim to address some of these gaps, and how they affect machine learning performance, using the mathematical construct of copulas.

Our first contribution deals with accurate probabilistic modeling of real-world data, where the underlying data is both continuous and discrete. We show that even though the copula construct has some limitations with respect to discrete data, it is still amenable to modeling large real-world datasets probabilistically. Our second contribution deals with analysis of non-linear datasets. Here, we develop a new measure of statistical association that can handle discrete, continuous, or combinations of such random variables that are related by any general association pattern. We show that our new metric satisfies several desirable properties and compare it's performance to other measures of statistical association. Our final contribution attempts to provide a framework for understanding how an estimator of statistical association will affect end-to-end machine learning performance. Here, we develop the estimator response curve, and show a new way to characterize the performance of an estimator of statistical association, termed the estimator response curve. We then show that the estimator response curve can help predict how well an estimator performs in algorithms which require statistical associations to be rank ordered.

Dedication

This work is dedicated to my wife, my parents, my brother, my sister, my coworkers, and my mentors who have been a constant source of encouragement, support, and inspiration throughout this process. I could not have done this without you, and am forever indebted to each and every one of you for your unique and individual contributions. This work is as much yours as it is mine, and I hope you are all proud of the way I have expressed your energies and efforts through this work.

My best, Kiran

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Machine learning can be broadly defined as the research and development of algorithms which give computers the ability to learn from data. These algorithms can be categorized into supervised, unsupervised, and reinforcement learning techniques. Supervised machine learning can be viewed as inferring a functional model from labeled training data, while unsupervised learning can be viewed as organization of unlabeled data. Reinforcement learning departs slightly from this paradigm, and its goals are to learn optimal behavior and decisions in an on-line environment. Each category of algorithms is applied based on the unique needs of the problem and the available data.

A fundamental construct present in all these types of learning is probability, random variables, and dependence. This is due to the fact that machine learning deals with real world data and phenomenon, which are never deterministic or noise free. However, although probability theory underpins most learning algorithms, the construct of copulas, which are stochastic models that capture the full dependence structure between random variables, has largely been ignored by the machine learning community. In this report, we explore how copulas can be used for both supervised and unsupervised machine learning techniques. This work builds off the pioneering work of Elidan [2013], who first recognized the overlap of objectives between the copula modeling and

the machine learning communities. Several seminal papers appeared in scientific literature during this period and showed the promise of copula modeling in machine learning, including Copula Bayesian Networks [Elidan, 2010], Copula Network Classifiers [Elidan, 2012], and Copula Mixture Models [Rey and Roth, 2012]. However, although these techniques have been shown to be effective, their limitation is that they can only be applied to data modeled as continuous random variables safely [1]. Careful analysis and modification are required to apply these developed methods for discrete data.

Our work to date has been to bridge this gap, and bring the copula construct to discrete and hybrid data, and apply these modified constructs to machine learning problems. The motivation for this comes from the fact that most real world data to be modeled and/or analyzed is often jointly continuous and discrete. The specific contributions of our research involve three major topics: 1) extending copula based probabilistic graphical models to account for continuous and discrete random variables simultaneously [Karra and Mili, 2016], 2) developing a new index of nonlinear dependence between continuous, discrete, or hybrid random variables [Karra and Mili, 2018a], and 3) developing a new way to assess an estimator of association's performance in the scenario where multiple estimates of association need to be rank ordered [Karra and Mili, 2018b].

This report is organized as follows. Chapter 2 introduces the mathematical and statistical concepts required for the subsequent developments and provides an introduction to copulas, discrete and hybrid copulas, and statistical dependence. Chapter 3 details our first contribution to copulas and machine learning, and discusses large dimensional probabilistic graphical models with discrete and continuous data, termed Hybrid Copula Bayesian Networks (HCBN). We then discuss exploratory data analysis using copulas in Chapter 4, and detail our contributions to detection of nonlinear dependence of continuous, discrete, and hybrid random variables. Finally, we discuss a new way to assess the performance of estimators when multiple estimates of the strength of association need

[1] This topic is treated in detail in Section 2.3

to be rank ordered, which is shown to be empirically useful in Chapter 5. We then summarize and conclude in Chapter 6, and provide future research direction.

Chapter 2

Copulas

In this chapter, we discuss the mathematical preliminaries required for subsequent chapters. We begin by introducing copula modeling for continuous random variables in Section 2.1. Section 2.2 then describes several copula models that are important from a modeling perspective. We then discuss copula modeling with discrete and hybrid random variables in Section 2.3. Finally, we discuss measures of association based on copulas in Section 2.4.

2.1 Copulas

A d-copula C is a functional mapping from $\mathbf{I}^d \rightarrow \mathbf{I}$, where $\mathbf{I}^d$ is the unit hypercube of dimensionality d, which has the following properties:

1. For every $\mathbf{u}$ in $\mathbf{I}^d$, $C(\mathbf{u}) = 0$ if atleast one coordinate of $\mathbf{u}$ is 0.

2. $C(\mathbf{u}) = u_k$ if all coordinates of $\mathbf{u}$ are 1 except u_k.

3. For every $\mathbf{a}$ and $\mathbf{b}$ in $\mathbf{I}^d$ such that $\mathbf{a} \leq \mathbf{b}$, $V_C([\mathbf{a}, \mathbf{b}]) \geq 0$. Here, $V_C([\mathbf{a}, \mathbf{b}]) = \sum sgn(\mathbf{u})C(\mathbf{u})$,

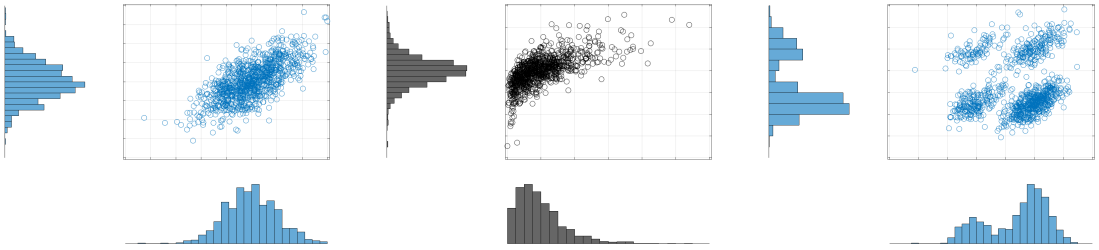
where the sum is taken over all vertices $\mathbf{u}$ of $[\mathbf{a}, \mathbf{b}]$ and

$$sgn(\mathbf{u}) = \begin{cases} +1, \text{ if } u_k = a_k \text{ for an even number of } k\text{'s} \\ -1, \text{ if } u_k = a_k \text{ for an odd number of } k\text{'s} \end{cases}$$

These three properties are also properties of joint distribution functions; resultingly, a d-copula C can be interpreted as a joint distribution function on the support $\mathbf{I}^d$. The link between the functional interpretation and the probabilistic interpretation of a copula C was discovered by Sklar [1959]. Sklar's theorem states that for any collection of random variables $X_1, \ldots, X_d$, the relationship between the marginal distributions $F_{X_1}(x_1), \ldots, F_{X_d}(x_d)$ and the joint distribution $H(\cdot)$ can be expressed by the copula C as

$$C(F_{X_1}(x_1), \ldots, F_{X_d}(x_d)) = H(x_1, \ldots, x_d) \tag{2.1}$$

Sklar's theorem guarantees the unicity of the copula C when $X_1, \ldots, X_d$ are continuous random variables, and it follows that the copula C captures the unique dependency structure between any continuous random variables $X_1, \ldots, X_d$. Conversely, a copula model can be applied to chosen marginal distributions to generate distribution functions that have a desired dependency structure with heterogeneous marginals. These dual views make copulas a powerful construct in stochastic modeling and simulation. Figure 2.1 shows an example of generating three different joint distributions with a Normal copula and heterogeneous marginal distributions.



(a) (b) (c)

Figure 2.1: 500 Samples of multivariate distributions generated from a Normal Copula ($\rho = 0.7$) and various marginal distributions. In (a), $X \sim \mathcal{N}(0,1)$ and $Y \sim \mathcal{N}(0,1)$. In (b), $X \sim Gamma(2,1)$ and $Y \sim T(5)$. In (c), $X \sim 0.3\mathcal{N}(-2,1) + 0.7\mathcal{N}(2,1)$ and $Y \sim 0.7\mathcal{N}(-1,1) + 0.3\mathcal{N}(4,1)$.

2.2 Important Copula Models

Several important copula models which capture typical dependence structures between random variables, have been discovered by copula practitioners. The most important of these are the Fréchet-Hoeffding bounds and the independence copula. When two random variables, X and Y, are comonotonic, meaning that Y is almost surely an increasing function of X, their copula is given by the Fréchet-Hoeffding upper bound

$$M(u,v) = min(u,v)$$

Conversly, when two random variables, X and Y, are countermonotonic, meaning that Y is almost surely a decreasing function of X, their copula is given by the Fréchet-Hoeffding lower bound

$$W(u,v) = max(0, u+v-1)$$

Finally, when two random variables X and Y are independent, their copula is given by the independence copula

$$\Pi(u,v) = uv$$

Together, the Fréchet-Hoeffding bounds and the independence copula encompass all possible extremes of dependency structures between random variables. Although these model extremes of dependence, real data often exhibits stochastic rather than deterministic dependence. By stochastic dependence, we mean that the assocation tends positively or negatively, but is not perfectly associated, as in the comonotonic and countermonotonic cases. For this reason, various families of copulas, which encompass a range of stochastic dependencies have been discovered. A family of copulas is termed to be a comprehensive if the family encompasses the dependence structures described by the M, W and Π copulas.

Several comprehensive families of copulas exist, including the Fréchet and Gaussian families of copulas. The Fréchet family is a two parameter family that is defined as a convex combination of the Fréchet-Hoeffding bounds and the independence copula, and written by

$$C_{\alpha,\beta}(u,v) = \alpha M(u,v) + (1-\alpha-\beta)\Pi(u,v) + \beta W(u,v) \quad \forall \alpha, \beta \in \mathbf{I}, \alpha+\beta \leq 1$$

The Gaussian family of copulas describes the entire range of linear dependence structures through a single parameter, the correlation coefficient ρ, and is given by

$$C(u,v) = \frac{1}{2\pi\sqrt{1-\rho^2}} \int_{-\infty}^{\Phi^{-1}(u)} \int_{-\infty}^{\Phi^{-1}(v)} exp\left[\frac{-(s^2-2\rho st+t^2)^2}{2(1-\rho^2)}\right] dsdt$$

where $\Phi^{-1}(\cdot)$ denotes the inverse standard normal distribution function. Many other families including Archimedean, Plackett, and Cuadras-Augé exist, and have been useful in different modeling scenarios.

Given a set of data, the copula modeling process typically involves choosing a family and estimating the copula parameter such that the likelihood of the model fitting the chosen copula family and parameter is high. Model selection is either performed using maximum likelihood techniques,

or a-priori knowledge of the dependence structure characteristics. The copula parameter can then be estimated by computing a measure of concordance on the data, and applying the functional relationship between the measure of concordance and the copula parameter.

2.3 Discrete and Hybrid Copulas

Sklar's theorem, given by (2.1), guarantees that the mapping between any joint distribution H and a copula C is unique if the marginal distributions are comprised of continuous random variables. When the marginal distributions are discrete random variables, the copula C is no longer unique. The difficulty here stems from the fact that the associated cumulative probability distribution function will exhibit discontinuities at discrete values where it has non-zero probability. This is easily seen when Sklar's theorem is expressed in terms of the joint cumulative distribution function $H(\cdot)$ as

$$C(u_1, \ldots, u_d) = H(F_{X_1}^{(-1)}(u_1), \ldots, F_{X_d}^{(-1)}(u_d)) \tag{2.2}$$

where u_i for all $i = 1, \ldots d$ is a uniform random variable taking on values in the unit interval $[0, 1]$. In this representation of Sklar's theorem, $H(\cdot)$ is a function of u_i for $i = 1 \ldots d$ that takes discrete values over the unit interval $[0, 1]$. Therefore, there exist many copulas that satisfy the relations given by (2.1) and (2.2), i.e. the copula is not unique. In order to address this issue, Denuit and Lambert [2005] propose continuing the discrete random variable X with a random variable U valued on $(0, 1)$ independent of X with a strictly increasing cumulative distribution function $L_U(u)$, on $(0, 1)$ and sharing no parameters with X. Nešlehová [2007] generalizes the foregoing

construction by defining the transformation $\psi : [-\infty, \infty] \times [0, 1] \rightarrow [0, 1]$

$$\psi(x, u) = P[X < x] + uP[X = x], \tag{2.3}$$

where X denotes a random variable that can be continuous or discrete and U denotes a uniform continuous random variable that is independent of X, and u is a realization of U. It is then shown in the bivariate case that transforming the discrete random vector (X, Y) to the vector $\mathbf{\Psi}(\mathbf{X}, \mathbf{U}) = (\psi(X, U), \psi(Y, V))$ yields a possible copula describing the original joint discrete distribution function of (X, Y). Additionally, if U and V are independent, the copula so defined is again the standard extension copula of Schweizer and Sklar [1974]. Consequently, it preserves the concordance properties of the original discrete distribution function of (X, Y) [Denuit and Lambert, 2005, Nešlehová, 2007]. Concordance is a form of dependence, which measures the degree to which two random variables are associated with each other; a pair of random variables are said to be concordant if large values of one random variable tend to be associated with large values of the other random variable, and small values of one random variable are associated with small values of the other variable [Nelsen, 2006]. Due to this and other properties and generality, we use the ψ transformation defined by Nešlehová [2007] in this report.

We consider the ψ function to take any admissible form; this will allow us to encompass not only the continuous extension defined by Denuit and Lambert [2005], but also the case of continuous random variables following any probability distribution. For instance, if X is a continuous random variable, then we have

$$\psi(X, U) = X, \tag{2.4}$$

because $P[X \leq x] = P[X < x] + uP[X = x] = P[X < x]$ since $P[X = x] = 0$. In summary, the transformation $\psi(X, U)$ preserves continuous random variables and transforms discrete random variables into continuous uniform random variables over $[0, 1]$.

2.4 Measures of Association based on Copulas

From (2.1) and (2.2), it is evident copulas capture the entire dependency structure between random variables. Thus, it is reasonable to think that mappings can be defined between copulas and measures of dependence. A popular measure of dependence that is based on the copula is concordance. Concordance is a form of dependence which measures the degree to which two random variables are associated with each other. More precisely, points in $\mathbb{R}^2$, (x_i, y_i) and (x_j, y_j), are concordant if $(x_i - x_j)(y_i - y_j) > 0$ and discordant if $(x_i - x_j)(y_i - y_j) < 0$ [Nelsen, 2006]. This can be probabilistically represented by the concordance function, Q, defined as

$$Q = P[(X_1 - X_2)(Y_1 - Y_2) > 0] - P[(X_1 - X_2)(Y_1 - Y_2) < 0] \tag{2.5}$$

where (X_1, Y_1) and (X_2, Y_2) are independent vectors of continuous random variables with distribution functions H_1 and H_2 with common margins of F (of X_1 and X_2) and G (of Y_1 and Y_2). Nelsen [2006] then shows that the concordance function can be written in terms of the copulas of (X_1, Y_1) and (X_2, Y_2), C_1 and C_2 respectively, rather than the joint distribution function as

$$Q(C_1, C_2) = 4 \int \int_{\mathbf{I}^2} C_2(u, v) dC_1(u, v) - 1 \tag{2.6}$$

Note that (2.6) is the general definition of concordance that measures the concordance between two random vectors (X_1, Y_1) and (X_2, Y_2) with identical marginals but different dependency structures. To compute the concordance between two random variables X and Y, the concordance function in (2.6) would be applied as $Q(C, C)$ where C is the copula of the joint distribution (X, Y).

Many metrics of association are based on the concept of concordance, with the two most popular being Kendall's τ [Kendall, 1938] and Spearman's ρ [Spearman, 1904]. Kendall's τ is defined in terms of the concordance function as $\tau = Q(C, C)$, where C is the copula of the joint distribu-

tion (X, Y), and interpreted as the scaled difference in the probability between concordance and discordance. Spearman's ρ is defined in terms of the concordance function as $\rho = 3Q(C, \Pi)$, where C is the copula of the joint distribution (X, Y) and Π is the independence copula, and can be interpreted as a rank correlation coefficient. Concordance-based measures of association such as Kendall's τ are ideal for detecting linear and nonlinear monotonic dependencies because they are rank statistics. These measures have the desirable properties of being margin independent and invariant to strictly monotonic transforms of the data [Scarsini, 1984, Nelsen, 2006].

Although these statistics work well for measuring monotonic association between continuous random variables, adjustments need to be made to account for discrete and hybrid random variables. This is encapsulated by Sklar's theorem, which does not guarantee the unicity of the copula C when the marginal distribution functions are non-continuous. If some or all of the marginal distribution functions are associated with discrete random variables, then many copulas satisfy (2.1) due to ties in the data. Consequently, the measure of concordance becomes margin-dependent (i.e, cannot be expressed solely in terms the joint distribution's copula as in (2.6)) and in many cases cannot not reach 1 or -1 in scenarios of perfect monotonocity and comonotonicity, respectively [Genest and Nešlehová, 2007].

Chapter 3

Hybrid Copula Bayesian Networks

In this chapter, we discuss a new model, based on hybrid copulas described in Section 2.3, for large multivariate modeling using graphical models with continuous and discrete data, termed Hybrid Copula Bayesian Networks (HCBN). We begin by discussing multivariate modeling and the state-of-the-art in this field from a copula perspective. The limitations of the existing models are then described, and the new model is introduced. We then show experimental results which shows this newly defined model as a viable alternative to the existing conditional linear Gaussian (CLG) and mixture of truncated exponentials (MTE) models.

3.1 Multivariate Modeling

Modeling large multivariate probability distributions is an essential research activity in many fields, ranging from engineering to computational biology to economics. Their advantage is that they allow us to understand the dependencies and interactions between random variables, which is arguably instrumental in making intelligent decisions from data. Although modeling large multivariate distributions has many advantages, it is a difficult problem from an analytical and computational

perspective. Closed form expressions for multivariate continuous random variables are limited in number; common ones include the multivariate Gaussian, t, and chi-squared distributions. Although analytically tractable, these standard models do not fully capture either the true marginal distribution or the statistical dependency for datasets which do not follow these parametric models. Computationally, one runs into the curse of dimensionality with many real-life datasets.

For discrete random variables, the situation is better because multinomial random variables can represent arbitrary discrete distributions. However, a difficulty arises when the data is modeled using multivariate probability distributions of both discrete and continuous random variables, referred to henceforth as hybrid random variables. One approach to modeling them is to factorize the joint distribution into a conditional distribution of one set of outcomes and a marginal distribution of the other set [Koller and Friedman, 2009, de Leon and Wu, 2011]. While this is a valid approach, it suffers from the problem of identifying probability distributions for each discrete outcome. When multiple discrete random variables are combined, the number of conditional distributions to identify explodes combinatorially.

Bayesian networks are graphical models used to estimate large dimensional joint probability distribution functions by estimating products of smaller joint probability distributions under conditional independence assumptions [Koller and Friedman, 2009]. In an all continuous network, these joint probability distribution functions are typically chosen from parametric models such as the multivariate normal probability distribution function. For hybrid networks, where nodes can be modeled as both discrete and continuous random variables, popular models such as the CLG network have additional limitations; in particular, they do not allow a continuous parent to have a discrete child [Koller and Friedman, 2009]. By contrast, the MTE approach removes the parent/child random variable type restrictions by effectively piecewise surface-fitting the underlying distribution [Moral et al., 2001]; although effective from a modeling perspective, it still retains the combinatorial problem of identifying conditional distributions, and could introduce inaccuracies at the

inflection points of the multivariate distribution due to the nature of the exponential distribution. This motivates the need for new hybrid models.

Elidan [2010] overcomes the limitations of analytically tractable models by using copula theory to model Bayesian networks, termed Copula Bayesian Networks (CBNs). By using copulas, Elidan [2010] shows that CBN's can effectively model the underlying data while reducing the number of variables to be estimated jointly.

3.2 Limitations of the CBN Approach

The main limitation of the CBN framework in its current form is that it can only model nodes in the Bayesian network as continuous random variables. This is due to the fact that Sklar's theorem, given by (2.1), guarantees the existence of a unique copula only when the marginal distribution functions represent continuous random variables. For a given joint distribution function with marginal distribution functions that represent discrete random variables, there exist many copulas that satisfy (2.1). Copula network classifiers (CNCs) Elidan [2012] build upon the CBN model and define conditional copulas, allowing for copula networks of mixed random variables. However, CNCs have the same structure limitations as the CLG model.

To overcome this difficulty, several methods have been proposed in literature to construct copulas for discrete probability distributions. For instance, following Schweizer and Sklar [1974], Denuit and Lambert [2005] and Nešlehová [2007] develop continuous extensions and transformations, respectively, of discrete random variables to define unique copulas with discrete marginal distributions that retain many of the familiar properties of copulas with continuous marginal distributions. By contrast, de Leon and Wu [2011] develop a conditional probability distribution-based methodology, while Smith and Khaled [2012] propose a latent variable approach based on MCMC. As for Kolesarova et al. [2006], they introduce a method based on sub-copulas over domains defined

by uniform discrete distributions. In this report, we concentrate on the approach investigated by Nešlehová [2007] and propose a new model for constructing hybrid Bayesian networks with copulas, termed Hybrid Copula Bayesian Networks (HCBN).

3.3 Required Constructions

The building block of our HCBN construction will be the hybrid copula, which is a copula function that joins both continuous and discrete random variables. The discrete random variables under consideration are count or ordinal discrete random variables, with some ordering between values in the domain of the discrete random variable [1]. Continuing the developments in Section 2.3, let us construct the multivariate hybrid copula by defining the random vector $\mathbf{X} = (X_1, \ldots, X_n)$, and without loss of generality let $X_1, \ldots, X_k$, $k \leq n$ be continuous random variables, and $X_{k+1}, \ldots, X_n$ be ordinal or count discrete random variables. Following the construction of Nešlehová [2007], let us define the transformed random vector as

$$\begin{aligned} \Psi(\mathbf{X}, \mathbf{U}) &= (\psi(X_1, U_1), \ldots, \psi(X_k, U_k), \psi(X_{k+1}, U_{k+1}), \ldots, \psi(X_n, U_n)) \\ &= (X_1, \ldots, X_k, \psi(X_{k+1}, U_{k+1}), \ldots, \psi(X_n, U_n)). \end{aligned} \tag{3.1}$$

Proposition 1 *The unique copula $C_{\mathbf{\Psi}(\mathbf{X},\mathbf{U})}$ of $\mathbf{\Psi}(\mathbf{X}, \mathbf{U})$ is a possible copula of $\mathbf{X}$.*

Proof: See Appendix A.

Proposition 2 *If $U_1 \ldots U_n$ are independent, then $C_{\mathbf{\Psi}(\mathbf{X},\mathbf{U})}$ is the standard extension copula of the form specified by Schweizer and Sklar, extended to a k-linear interpolation.*

[1] This point is discussed in further detail in Section 3.3.1

Proof: Follows directly from Proposition 1 and Nešlehová [2007].

Additionally, due to (2.4), we can directly apply the results of Mesfioui and Quessy [2010] to obtain the concordance properties of the multivariate mixed copula. More formally, we have

$$\tilde{Q}(H_1, H_2) = \mathbf{T}Q(H_1, H_2), \tag{3.2}$$

where $\mathbf{T} = \prod_{l=1}^{d} \left(\frac{T_l+1}{2}\right)$, $T_l P[\mathbf{X}_1 < \mathbf{X}_2] = P(X_{11} < X_{21}, \ldots, X_{1l} < X_{2l}, \ldots, X_{1n} < X_{2n}]$, $\mathbf{X}_1 = (X_{11}, \ldots, X_{1d})$, $\mathbf{X}_2 = (X_{21}, \ldots, X_{2d})$, $\tilde{Q}(H_1, H_2) = Q(C_1^*, C_2^*)$, and C_1^* and C_2^* are the unique copulas associated with $\Psi(\mathbf{X_1}, \mathbf{U_1})$ and $\Psi(\mathbf{X_2}, \mathbf{U_2})$ respectively. Equation 3.2 shows us that the multivariate mixed copula has the desirable consequence of retaining the concordance properties of the original joint distribution $\mathbf{X} = (X_1, \ldots, X_k, X_{k+1}, \ldots, X_n)$.

3.3.1 Applicability of the Multivariate Extension Framework

The mixed copulas and their properties introduced above apply to count or ordinal discrete data. This is because copulas contain information about the dependency structure between two or more random variables. If there is no ordering of the discrete data, then the concept of two random variables behaving together in either a concordant or discordant way (in order words, having some dependency structure) does not exist. This however does not prevent one from constructing copulas (details of copula construction are provided in Section 3.4.2) with mixed random variables where the discrete random variables are unordered, such as categorical data. To construct a copula with categorical data, one assigns arbitrary ordinal values to the categories. This works because the continued random variable CDFs given by (3.1) agree with the original discrete CDF in the discrete domain. However, interpretation of any dependence structure from the computed copula only makes sense and is valid if the discrete random variables are count data or ordinal data. This is not a limitation of copulas or the HCBN framework, but rather an inherent property of the discrete

data being modeled.

3.4 HCBN Model

Having established the theoretical properties and framework for the multivariate hybrid copula, we are now ready to formally define the HCBN. An HCBN is a tuple $\mathcal{C}_{HCBN} = (\mathcal{G}, \Theta_C, \Theta_f)$ that encodes the joint density $f_{\mathcal{X}}(x)$. $\mathcal{G}$ is the graph structure, Θ_C is the set of mixed copula functions for each copula family (i.e. a child and its parents), and Θ_f is the set of parameters representing the marginal densities $f_{X_i}(x_i)$. This is similar to the CBN definition in Elidan [2010], with the difference that Θ_C is a matrix in which the i^{th} column represents the parameters of the i^{th} copula, rather than a 1-D vector with the i^{th} element describing the dependence parameter of the i^{th} copula. This is because the copulas joining all continuous variables only require one parameter (assuming one parameter families are used) to describe the copula, but copulas joining continuous and discrete variables will need more.

To express the overall joint density $\mathcal{X}$ represented by the HCBN framework, let us first define the local density of a family via the HCBN framework. In a local density containing n nodes, without loss of generality, define k to be the continuous marginals, and the remaining $n - k$ to be discrete marginals. The local joint density of the i^{th} family can then be written as

$$f_i(\mathbf{x_i}) = \prod_{l=1}^{k} f_{X_l}(x_l) \times \sum_{j_{k+1}=1}^{2} \cdots \sum_{j_n=1}^{2} (-1)^{j_{k+1}+\cdots+j_n} \times C_i^k(F_{X_1}(x_1), \ldots, F_{X_k}(x_k), u_{k+1,j_n+1}, \ldots, u_{n,j_n}) \tag{3.3}$$

where

$$C_i^k = \frac{\partial^k}{\partial u_1 \partial u_2 \ldots \partial u_k} C_i(u_1, \ldots, u_n) \tag{3.4}$$

, $u_{j,1} = F_{X_j}(x_j^-)$, $u_{j,2} = F_{X_j}(x_j)$. The local density in (3.3) is a product of the continuous marginal distributions, the partial derivative of the encompassing copula function with respect to the continuous variables, and the C-volume of the discrete marginal distributions; the C-volume is computed via the summation of the encompassing copula function in the discrete dimensions over $\{1,2\}$, following (2.2.3) in Nelsen [2006]. By defining the local density, and defining that $C_i^k = 1$ in the case of root nodes, we derive the full joint density, $\mathcal{X}$, using the conditional independence assumptions of Bayesian networks as follows

$$f_{\mathcal{X}}(\mathbf{x}) = \prod_{i=1}^{D} f_i(\mathbf{x_i}) \tag{3.5}$$

This is different than the CBN parametrization of $\mathcal{X}$ as it does not rely on the copula densities alone [Elidan, 2010]. This is because the CBN model depends upon the fact that for continuous distributions, the relation given by

$$f(x_1, \ldots, x_N) = c(F_{X_1}(x_1), \ldots, F_{X_N}(x_N)) \prod_{i=1}^{N} f_{X_i}(x_i) \tag{3.6}$$

is true. However, (3.6) does **not** hold true for discrete or mixed distributions [Panagiotelis et al., 2012]. This is the main reason why the CBN framework cannot be directly used for hybrid data. The local density described by (3.3) is a generalization of (3.6), and thus applicable to the mixed random variable scenario.

3.4.1 HCBN Structure Learning

The steps common to any existing structure learning algorithm required to construct the HCBN are:

1. Preprocess each discrete random variable X_i with the transformation $\psi(X_i, U_i)$.

2. Compute empirical marginal distributions for each node in the Bayesian network.

If score-based approaches such as greedy hill climbing are used for learning, candidate structures scores would be computed using the scoring function given by given by

$$\mathcal{L}(\mathcal{D} : \mathcal{C}_{HCBN}) = \sum_{m=1}^{M} \sum_{i} \log f_i(\mathbf{x_i}[\mathbf{m}]),$$

where f_i is the local density defined above in (3.3) and the best scoring structure taking into account graph complexity would be chosen. The HCBN framework is also compatible with constraint based approaches, and statistical tests of independence and conditional independence for mixed random variables developed in Chapter 4 would need to be applied.

It is worth noting the flexibility of the ψ transform in the context of HCBN construction; the transformation ψ is applied without any a-priori knowledge of the Bayesian network structure. One consequence of this is that for each node to which the transformation $\psi(X_i, U_i)$ is applied, an independent U_i is used. The advantage of using an independent U_i for each discrete node is that the constructed copula inherits the dependence properties of the data it is trying to model, as shown by (3.2). The other advantage is that continuing the discrete random variables and the structure learning are independent, allowing for computational efficiency.

3.4.2 Copula Density Estimation

In the CBN construction described by Elidan [2010], standard copula models including the Frank and Gaussian copulas are used to join the continuous marginal distributions. The HCBN construction can also use standard copula models when all the nodes of a copula family are modeled as continuous random variables. In the continuous marginals case, this approach works well because

the chosen copula's dependency parameter θ can be set such that the empirical Kendall's $\hat{\tau}$ of the dataset matches the Kendall's τ of the copula, using known relationships between τ and θ and thereby capturing the underlying concordance properties of the data (although copula model selection itself is another problem). However, using the relationship between a copula's dependency parameter θ and empirical Kendall's $\hat{\tau}$ with discrete marginal distributions leads to inconsistent and biased estimates of the dependence parameter [Genest and Nešlehová, 2007].

Although algorithms to estimate parameters for standard copulas with discrete marginals without bias have been explored [Smith and Khaled, 2012], to the best of the authors' knowledge, model selection for copulas in the hybrid scenario has not yet been explored. For this reason, we do not use standard copulas for modeling mixed random variables. Propositions 1 and 2 show that copulas constructed by transforming the discrete random variables with ψ and independent U_i's yield unique copulas that follow the concordance properties of the underlying data. Although this amounts to creating a unique copula to model each local mixed distribution, we take this approach for it's theoretical soundness. Equivalently, this unique copula model can be viewed as a nonparametric approach to estimating the dependency structure of the corresponding copula family. This nonparametric approach also avoids the copula compatibility problem, because overlapping copula marginals remain the same due to them being estimated directly from the data, avoiding the marginal skewness problem outlined in the CBN construction [Elidan, 2010]. However, this advantage comes at the computational cost of estimating the copula density nonparametrically.

To estimate the multivariate empirical copula density, the beta-kernel approach is taken [Charpentier et al., 2007]. Beta-kernels are an ideal way to estimate copula density functions due to their bounded support. For the purposes of HCBN, we estimate the copula density from the data $\mathbf{x}$

directly with

$$\hat{c}_h(\mathbf{u}) = \frac{1}{M}\sum_{m=1}^{M}\prod_{d=1}^{D}\beta(F_{X_d}(x_d(m)), \frac{u}{h}+1, \frac{1-u}{h}+1), \tag{3.7}$$

where h is a bandwidth parameter, β is the probability distribution function of the Beta distribution, and $F_{X_d}(x_d(m))$ is the m^{th} pseudo-observation for the d^{th} dimensional data point. Equation (3.7) is a multivariate extension of the bivariate copula density estimated by beta-kernals [Charpentier et al., 2007].

In the context of HCBN, estimating the copula density directly is preferred to finding the copula function and then differentiating. This is due to two primary reasons: 1.) although the empirical copula function estimate introduced by Deheuvels [1979] provides a robust and universal way to compute an empirical copula estimate, the discontinuous features of the estimator introduce difficulties when differentiating the empirical copula, even with respect to the continuous random variables, and 2.) the inverse function of H in (2.2) at some point $\frac{i}{T}$ may be chosen arbitrarily between the defined points of the dataset from which the empirical copula is constructed [Charpentier et al., 2007]. Once the copula density is estimated, the discrete dimensions are numerically integrated to find C_i^k for the i^{th} family.

3.4.3 Accuracy of Density Estimation in the Hybrid Context

In the hybrid random variable context, (3.7) becomes a better estimator of the true copula density as the number of discrete outcomes increases. This is because in this scenario, less volume in the unit hypercube is filled in with the uniform random variable in $\psi(X, U)$. Figure 3.1 shows pseudo-observations of two different bivariate mixed random variables, with the discrete random variable continued by (3.1). Both have a dependency structure described by the same copula whose pseudo-observations are shown in the middle plot; the mixed random variable with two discrete outcomes

is shown in the left plot and the mixed random variable with ten discrete outcomes is shown in the right plot. It is seen that as as the number of outcomes for the discrete random variable increases, the empirical pseudo-observations are closer to the true pseudo-observations, and thus the copula and dependency structure is better estimated.

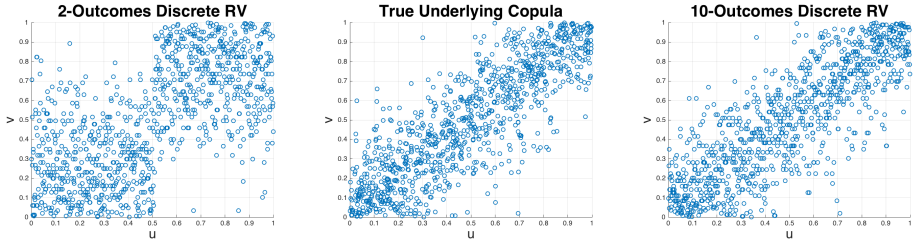
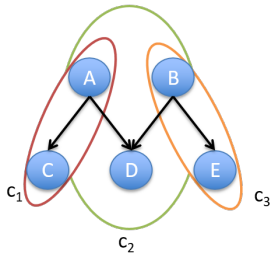


Figure 3.1: Pseudo-Observations plot for different discrete random variables with the same copula compared to true pseudo-observations of copula.

This intuition is also corroborated by theory in Genest and Nešlehová [2007], Nešlehová [2007], Denuit and Lambert [2005]. Conversely, as the number of discrete outcomes increases, the MTE and CLG models need to define an exponentially growing number of conditional distributions with a decreasing number of samples per conditional distribution. Therefore, from a computational and modeling perspective, the HCBN can be recommended over the CLG or MTE models as the number of discrete outcomes per local density (copula family in the HCBN context) increases.

3.5 Results

We test the HCBN construction with various experiments on various synthetic datasets and a real-life dataset. In order to compare the HCBN's performance against the CLG, MTE, and multinomial BN models, we first create a synthetic data-set generated from a structure with the same I-map as the Bayesian Network structure showed in Figure 3.2, where nodes A and B are multinomial discrete random variables, and C, D, and E are continuous random variables. This structure restriction allows us to compare the CLG model fairly to the other models.



Nodes A/B Multinomial Probabilities	Nodes C/D/E PDF Types	Dependency Structure Models (C1/C2/C3)	Dependency Strengths for C1/C2/C3
[0.5 0.5]/[0.5 0.5]	N(2,0.5)/N(2,0.5)/N(2,0.5)	**Linear Dependency** Gaussian/ Gaussian/ Gaussian	**Strong**
	U(-2,2)/U(-2,2)/U(-2,2)		
[0.25 0.25 0.25 0.25]/ [0.25 0.25 0.25 0.25]	N(-2,0.3)+N(2,0.8)/ N(-2,0.3)+N(2,0.8)/ N(-2,0.3)+N(2,0.8)		
	T(3)/T(3)/T(3)		
	N(-2,0.3)+N(2,0.8)/ U(-2,2)/ N(-2,0.3)+N(2,0.8)	**Non-Linear Dependency** Frank/ Gaussian/ Frank	**Weak**
	N(2,0.5)/ N(-2,0.3)+N(2,0.8)/ U(-2,2)		
	U(-2,2)/U(-2,2)/T(3)		
	N(-2,0.3)+N(2,0.8)/ N(2,0.5)/U(-2,2)		

Figure 3.2: Bayesian Network used for synthetic data testing.

To test each model, every permutation of the different configurations for the discrete nodes, continuous nodes, and the dependencies between them outlined in Figure 3.2 were tested for the synthetic data evaluation for the CLG, MTE, multinomial BN (data was discretized using a histogram approach), and HCBN models. All models were given the true structure in Figure 3.2, in order to isolate and quantify the expressiveness of each model. Figure 3.3 shows the experimental bias and variance of each model for strong and weak, linear and nonlinear dependency structures for the graphical model with copulas C_1, C_2, C_3 in Figure 3.2.

The results show that in general, for weak dependencies, the copula approach is on par with or exceeds the performance of the MTE models. For strong dependencies, it is seen that the MTE model has better performance characteristics than the HCBN model. The strong dependency results are explained by the reasoning provided in Section 3.4.3. The weak dependency results can be explained in a similar light; in a weak dependency scenario, the pseudo-observations are more spread out and resultantly, the domain of the discrete random variable has less impact in the copula estimation of the true dependency structure. As expected, the CLG model displays high bias due to cases in the experiments where the data does not follow the Gaussian distribution. It is seen that the multinomial approach exhibits relatively low bias but high variance for each category tested, which can be explained by the discretization process. It remains to be explored how sensitive the

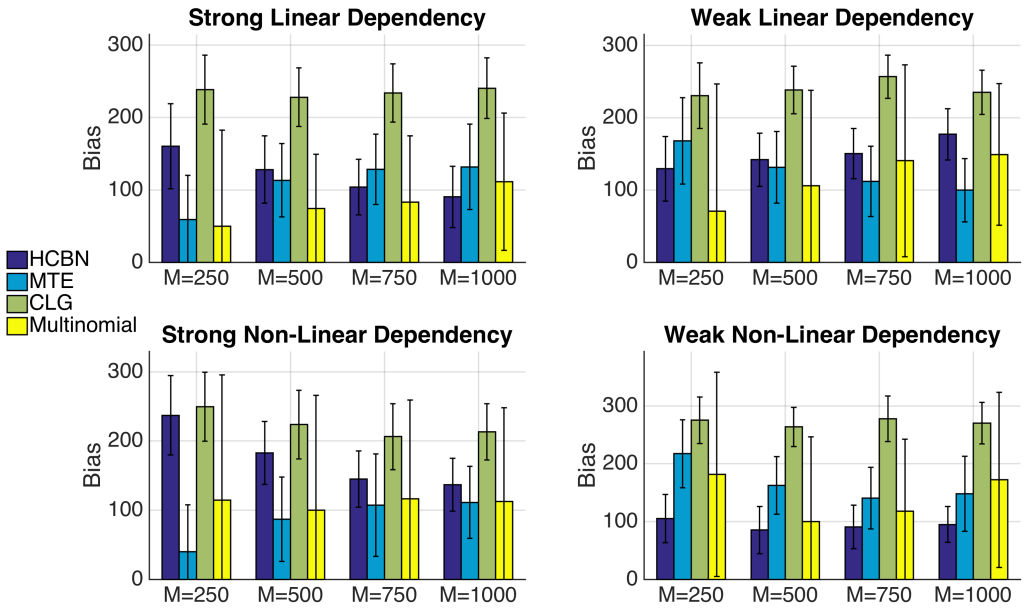


Figure 3.3: Monte Carlo simulations of Hybrid Models against synthetic data.

multinomial approach is to other discretization approaches.

In addition to the synthetic data experiments described above, we model the 1994 Census dataset [Lichman, 2013] using the HCBN. Figure 3.4 shows samples generated from the HCBN Bayesian network, versus the collected census data. The left handed columns in Figure 3.4 are scatter plots of two random variables from the actual data, and the right handed columns are samples from the HCBN generative model of that data. It is seen from Figure 3.4 heuristically, that both continuous and discrete random variables are modeled expressively and accurately with the HCBN model.

3.6 Alternative Approaches

The local density construction via empirical copula density estimation given in (3.3), although accurate, may be computationally expensive with large dimensional local densities and prone to overfitting. As mentioned above, Smith and Khaled [2012] explore the idea of estimating the

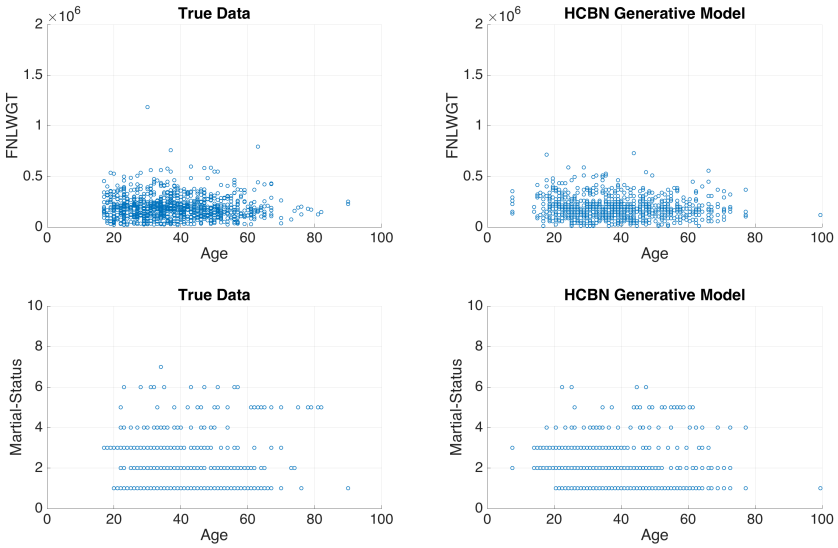


Figure 3.4: The top left figure shows a scatter plot of Age (Continuous) vs. FNLWGT (Continuous) in the 1994 Census data. The top right figure shows samples produced from the HCBN generative model of Age vs. FNLWGT. The bottom left figure shows a scatter plot of Age (Continuous) vs Martial-Status (Discrete-Categorical) in the 1994 Census data. The bottom right figure shows samples produced from the HCBN generative model of Age vs. Martial-Status.

copula dependence parameter for discrete marginals.

As an alternative, the developments in Chapter 4, and more specifically τ_{KL}, can be used to fit standard copula models to hybrid data. This is because, as will be shown in Section 4.2, τ_{KL} exhibits low bias in estimating copula parameters for discrete, hybrid, or continuous marginals, and thus the unique functional relationship between τ_{KL} and the copula's dependency parameter(s) can be inverted to derive the copula parameters. The question of model selection still remains, although one approach could be to use a maximum likelihood based technique. In this proposed method, one would first compute the τ_{KL} of the dataset, and find the copula parameters associated with this value of τ_{KL} for all candidate copula models. The model which maximizes the likelihood would be chosen. This suggested alternative approach, which trades off computational complexity for data fitting accuracy needs to be assessed more rigorously. The answer will likely ultimately depend on the dataset being modeled.

A distinct advantage of using copula families is that efficient inference becomes possible. By using the conditional relationship

$$f(\mathbf{x}|\mathbf{y}) = c(\mathbf{u}|\mathbf{v})\Pi_i f(x_i)$$

, where $\mathbf{x}$ is the set of unobserved nodes, and $\mathbf{y}$ is the set of observed nodes and c is the copula density, the densities for unobserved nodes can be readily calculated.

3.7 Conclusion

In this chapter, we have shown a new method for constructing multivariate copulas with discrete and continuous random variables, and applied this construction to Bayesian networks to create the HCBN. The main contribution is the extension of proofs by Nešlehová [2007] to the multivariate case to ensure the validity of the constructed hybrid copulas and incorporating them into the CBN framework proposed by Elidan [2010]. The defined framework has the ability to:

1. Effectively model multivariate hybrid distributions, while removing the restrictions of graph structure imposed by the CLG model [Koller and Friedman, 2009].

2. Avoid defining a combinatorial number of conditional distributions.

3. Efficiently sample the Bayesian network through copula sampling algorithms.

The empirical evaluation shows that the HCBN construction compares favorably to the CLG model and performs similarly to the MTE model. The estimated copula contains information about the underlying dependency structure, which may eventually be useful for causality related studies. Although both of these topics are out of the scope of this paper, they motivate research activities in estimating the copula density of the family via the beta-kernel based technique.

Chapter 4

Exploratory Data Analysis using Copulas

This chapter introduces a nonparametric copula-based index for detecting the strength and monotonicity structure of linear and nonlinear statistical dependence between pairs of random variables or stochastic signals. Our index, termed Copula Index for Detecting Dependence and Monotonicity (*CIM*), satisfies several desirable properties of measures of association, including Rényi's properties, the data processing inequality (DPI), and consequently self-equitability. Synthetic data simulations reveal that the statistical power of *CIM* compares favorably to other state-of-the-art measures of association that are proven to satisfy the DPI.

4.1 Dependence as an Exploratory Data Analysis Tool

A fundamental problem in exploratory data analysis involves understanding the organization and structure of large datasets. An unsupervised approach to this problem entails modeling the features within these datasets as random variables and discovering the dependencies between them using measures of association. Many measures of association have been introduced in the literature, including the correlation coefficient [Pearson, 1895], *MIC* [Reshef et al., 2011], the *RDC* [Lopez-Paz

et al., 2013], the *dCor* [Székely et al., 2007], the *Ccor* [Chang et al., 2016], and *CoS* [Ben Hassine et al., 2016]. In addition, many estimators of mutual information such as the *kNN* [Kraskov et al., 2004], the *vME* [Kandasamy et al., 2015], and the *AP* [Darbellay and Vajda, 1999] are used as measures of association.

However, properties of the dataset such as whether the data are discrete or continuous, linear or nonlinear, monotonic or nonmonotonic, noisy or not, and independent and identically distributed (*i.i.d.*) or not, to name a few, are important factors to consider when deciding which measure(s) of association one may use when performing exploratory data analysis. Because these properties are typically not known a priori, the task of selecting a single measure of association is difficult. Additionally, a measure of association should also satisfy certain desirable properties, namely Rényi's properties [Rényi, 1959], the data processing inequality (DPI) [Kinney and Atwal, 2014], and equitability [Reshef et al., 2011]. However, no measure satisfies all these properties while simultaneously being able to handle the different types and properties of data described. For example, the most commonly used measure of statistical dependence, the correlation coefficient, only measures linear dependence. Others such as the *RDC* exhibit high bias and relatively weak statistical power for the basic (and arguably the most important [Kinney and Atwal, 2014, Simon and Tibshirani, 2014]) linear dependency structure, due to overfitting. Finally, estimators of mutual information do not have a theoretical upper bound, meaning that the values can only be used in a relative sense. Even though each of the aforementioned measures of association perform well in the conditions for which they were designed, they cannot be used as an omnibus solution to an exploratory data analysis problem.

To help address these shortcomings, we introduce a new index of nonlinear dependence, *CIM*. This index is based on copulas and the rank statistic Kendall's τ [Kendall, 1938], that naturally handles linear and nonlinear associations between continuous, discrete, and hybrid random variables (pairs of random variables where one is continuous and the other is discrete) or stochastic signals. Ad-

ditionally, *CIM* provides good statistical power over a wide range of dependence structures and satisfies several desirable properties of measures of association including Rényi's properties and the data processing inequality. Furthermore, it uniquely identifies regions of monotonicity in the dependence structure which provide insight into how the data should be modeled stochastically. Due to these properties, *CIM* is a powerful tool for exploratory data analysis.

This paper is organized as follows. Section 4.2 discusses modifying Kendall's τ to account for discrete and hybrid random variables. Section 4.3 introduces *CIM* index, which builds upon the extension of Kendall's τ in Section 4.2 to handle both monotonic and non-monotonic dependencies and proposes an algorithm to estimate it. Here, important properties which theoretically ground *CIM* as a desirable measure of association are proved. Additionally, an estimation algorithm and it's properties are discussed and it is shown that the algorithm is robust to hyperparameter selection. Next, Section 4.4 provides simulations to exercise the developed metric against other state-of-the-art dependence metrics, including *MIC*, the *RDC*, the *dCor*, the *Ccor*, and *CoS* and measures of information including *kNN*, *vME*, and *AP* using synthetic data. These simulations reveal that *CIM* compares favorably to other measures of association that satisfy the Data Processing Inequality (DPI). Simulations with real-world data show how *CIM* can be used for many exploratory data analysis and machine learning applications, including probabilistic modeling, discovering interesting dependencies within large datasets, and Markov network discovery. These simulations show the importance of considering the monotonicity structure of data when performing probabilistic modeling, a property that only *CIM* can measure. The favorability of using *CIM* when performing Markov network discovery is shown through the **netbenchmark** simulation framework. Concluding remarks are then provided in Section 3.7.

4.2 Extension of Kendall's τ for Hybrid Random Variables

Although rank statistics work well for measuring monotonic association between continuous random variables, adjustments need to be made to account for discrete and hybrid random variables. In that case, Sklar's theorem does not guarantee the unicity of the copula C and many copulas satisfy (2.1) due to ties in the data. Consequently, the measure of concordance becomes margin-dependent (i.e, cannot be expressed solely in terms the joint distribution's copula as in (2.6)) and in many cases cannot reach $+1$ or -1 in scenarios of perfect comonotonicity and countermonotonicity, respectively [Genest and Nešlehová, 2007].

Several proposals for adjusting Kendall's τ for ties have been made, including τ_b [Kendall, 1945], τ_{VL} [Vandenhende and Lambert, 2003], and τ_N [Nešlehová, 2007]. The common theme among these proposals is that they use different scaling factors to account for ties in the data. However, even with scaling, perfect monotone dependence does not always imply $|\tau_b| = 1$, and τ_{VL} is not interpretable as a scaled difference between the probabilities of concordance and discordance [Genest and Nešlehová, 2007]. Nešlehová [2007] overcomes both of these limitations and defines the non-continuous version of Kendall's τ, denoted by τ_N [see Nešlehová, 2007, Definition 9].

Nešlehová [2007] then defines an estimator of the non-continuous version of τ_N as

$$\hat{\tau}_N = \frac{\text{\# concordant pairs - \# discordant pairs}}{\sqrt{\binom{n}{2} - u}\sqrt{\binom{n}{2} - v}} \tag{4.1}$$

where $u = \sum_{k=1}^{r} \binom{u_k}{2}$, $v = \sum_{l=1}^{s} \binom{v_l}{2}$, r is the number of distinct values observed in x and s is the number of distinct values observed in y, u_k is the number of times the k^{th} distinct element occurred in the u dimension, v_l is the number of times the l^{th} distinct element occurred in the v dimension. $\hat{\tau}_N$ achieves $+1$ or -1 in the comonotonic and countermonotic cases, respectively, for discrete random variables by subtracting the number of ties for each variable u and v independently

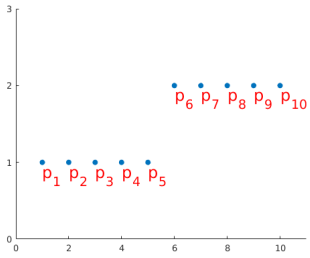


Figure 4.1: The hybrid random variable pair (X, Y) is comonotonic, with X the continuous random variable and Y the discrete random variable. In the computation of $\hat{\tau}$, the pairs of points $[p_i, p_j]$ for $i = 1:5, j = 1:5$ and $i = 6:10, j = 6:10$ are not counted as concordant. Only the pairs of points $[p_i, p_j]$ for $i = 1:5, j = 6:10$ are, leading to $\hat{\tau}$ not reaching $+1$ in the perfectly comonotonic case for hybrid random variables.

from the denominator. The accounting of ties is required due to the strict inequalities used for concordance and discordance in (2.5). In the continuous case, there are no ties and τ_N reduces to the original Kendall's τ.

Although the measure defined by τ_N is valid for continuous, discrete, or hybrid random variables, the estimator $\hat{\tau}_N$ in (4.1) does not achieve a value of $+1$ or -1 in the perfectly comonotonic and countermonotonic cases, respectively, for hybrid random variables. In order to make $\hat{\tau}_N$ equal to $+1$ and -1 in these cases respectively, we propose to use the maximum number of ties as a correction factor. This is because in the hybrid case, the numerator of $\hat{\tau}_N$ does not count the increasing continuous variables as concordant (or decreasing as discordant). Fig. 4.1 illustrates this counting in an example, and shows why $\hat{\tau}_N$ fails to achieve $+1$ or -1 in the hybrid random variable case for perfectly comonotonic/countermonotonic random variables respectively. In it, the pairs of samples along the continuous dimension x within a discrete value ($[p_i, p_j]$ for $i = 1:5, j = 1:5$ and $i = 6:10, j = 6:10$) are not counted as comonotonic. To overcome this drawback, our proposed extension to $\hat{\tau}_N$ is defined as

$$\hat{\tau}_{KL} = \begin{cases} \dfrac{\text{\# concordant pairs - \# discordant pairs}}{\binom{n}{2}} & \text{for continuous random variables} \\ \dfrac{\text{\# concordant pairs - \# discordant pairs}}{\sqrt{\binom{n}{2}-u}\sqrt{\binom{n}{2}-v}} & \text{for discrete random variables} \\ \dfrac{\text{\# concordant pairs - \# discordant pairs}}{\sqrt{\binom{n}{2}-t}\sqrt{\binom{n}{2}-t}} & \text{for hybrid random variables} \end{cases} \tag{4.2}$$

where $t = \mathbf{max}(u, v) - K$, and where u and v are the same as in $\hat{\tau}_N$, and $K = \binom{u'}{2} \times v'$, u' denotes the number of overlapping points in the continuous dimension and between different discrete values in the discrete dimension, and v' denotes the number of unique elements in the discrete dimension. K is zero for perfectly monotonic hybrid random variables, but takes nonzero values for copula-based dependencies; it helps to reduce the bias of $\hat{\tau}_{KL}$ when hybrid random variable samples are drawn from a copula dependency.

The performance of $\hat{\tau}_{KL}$, compared to $\hat{\tau}_b$ and $\hat{\tau}_N$ for perfectly comonotonic random variables is shown in Table 4.1. It is seen that the proposed modifications to the τ_N estimate in (4.2) do indeed reduce the bias for the hybrid random variables case. It is also observed that the bias of the $\hat{\tau}_b$ and $\hat{\tau}_N$ is reduced as the number of discrete levels is increased. However, in all these cases, $\hat{\tau}_{KL}$ still maintains a null bias.

Discrete Levels	$\hat{\boldsymbol{\tau}}_b$	$\hat{\boldsymbol{\tau}}_N$	$\hat{\boldsymbol{\tau}}_{KL}$
2	0.58	0.71	1.00
4	0.84	0.87	1.00
8	0.93	0.93	1.00

Table 4.1: Step function dependency with various levels of discretization; it is seen that τ approaches 1 as the number of discretization levels increases, but without the bias correction described in (4.2), dependence between continuous and discrete random variables is not measured accurately by τ_b and τ_N.

Figs. 4.2 (a),(b),(c), and (d) show the bias and variance between the estimated value of $\hat{\tau}_{KL}$ and the value of τ that generates the corresponding copula. Here, samples of $X = F_X^{-1}(U)$ and $Y = F_Y^{-1}(V)$ are drawn from a Gaussian distribution and from a uniform discrete distribution, respectively, and joined together with four different dependency structures captured by the Gaussian, Frank, Gumbel, and Clayton copulas. This follows the methodology described by Madsen and Birkes [2013] for simulating dependent discrete data. Figs. 4.2 shows that $\hat{\tau}_{KL}$ achieves low bias and variance among all proposed modifications to estimators of τ for hybrid random variables with copula-based dependencies. The null distribution of $\hat{\tau}_{KL}$ under independence, denoted by

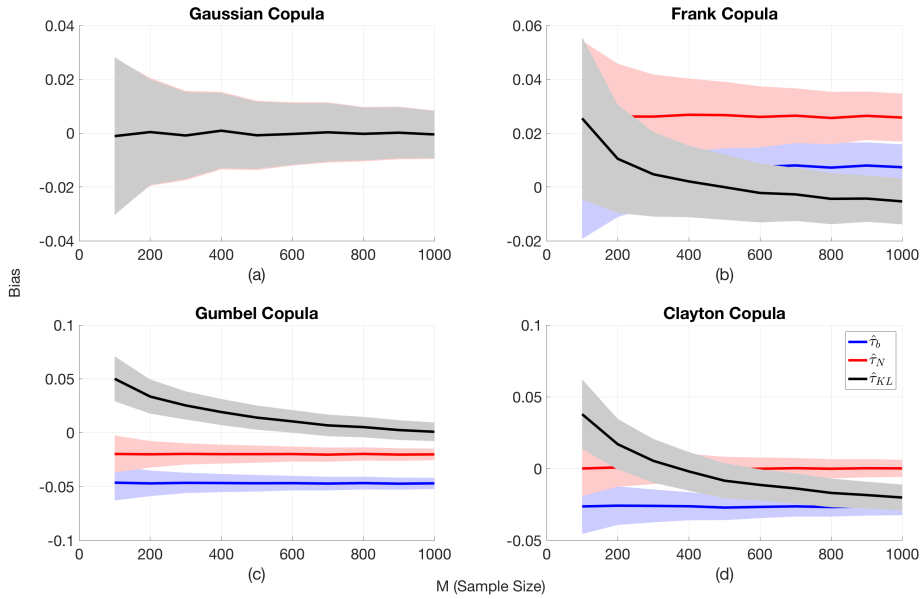


Figure 4.2: Average bias and standard deviation of $\hat{\tau}_b$, $\hat{\tau}_N$, and $\hat{\tau}_{KL}$ for varying number of samples of copula based with hybrid random variables. The bias and variance for each sample size was averaged over the entire range of copula dependencies (for the Gaussian copula, the copula parameter θ was varied from $[-1, 1]$ and for Archimedean copulas, the copula parameter α was varied from $[1, 10]$) for 300 Monte-Carlo simulations.

$X \perp\!\!\!\perp Y$, is depicted in Fig. 4.3. It shows that $\hat{\tau}_{KL}$ is Gaussian with a sample mean of approximately zero and a decreasing sample variance as M increases for continuous, discrete, and hybrid random variables.

4.3 Copula Index for Detecting Dependence and Monotonicity between Stochastic Signals

In the previous section, we described an extension to the estimator of τ_N to account for hybrid random variables. However, τ_N is still a rank statistic and thus cannot measure nonmonotonic dependencies. Here, we describe *CIM*, which is an extension of τ_N to detect nonlinear, nonmonotonic statistical dependencies that satisfies both Rényi's properties and the data processing inequality

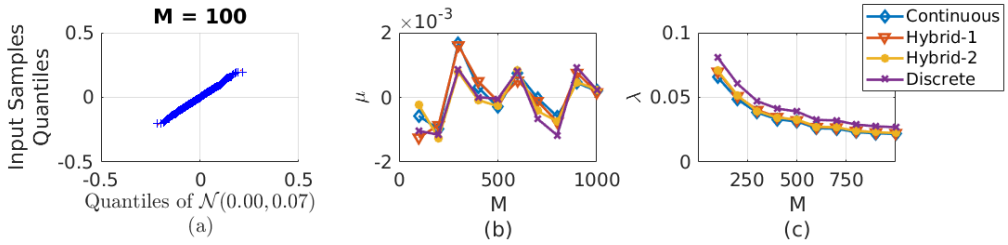


Figure 4.3: (a) QQ-Plot of $\hat{\tau}_{KL}$ for continuous random variables with $X \perp\!\!\!\perp Y$ and $M = 100$, (b) QQ-Plot of $\hat{\tau}_{KL}$ for continuous random variables with $X \perp\!\!\!\perp Y$ and $M = 1000$, (c) The sample mean of the distribution of $\hat{\tau}_{KL}$ for $X \perp\!\!\!\perp Y$ as a function of M (sample size), (d) The sample variance of the distribution of $\hat{\tau}_{KL}$ for $X \perp\!\!\!\perp Y$ as a function of M (sample size). Note: Hybrid-1 refers to a discrete X and continuous Y, Hybrid-2 refers to a continuous X and discrete Y.

(DPI). The motivation for this development comes from the need to assess the strength of association for any general dependence structures that may not be monotonic, when exploring real-world datasets for both analysis and stochastic modeling perspectives, and constructing Markov networks from data, to name a few. The theoretical foundations of this methodology are first developed. We then describe the properties of *CIM* and propose an algorithm to estimate it.

4.3.1 Theoretical Foundations of *CIM*

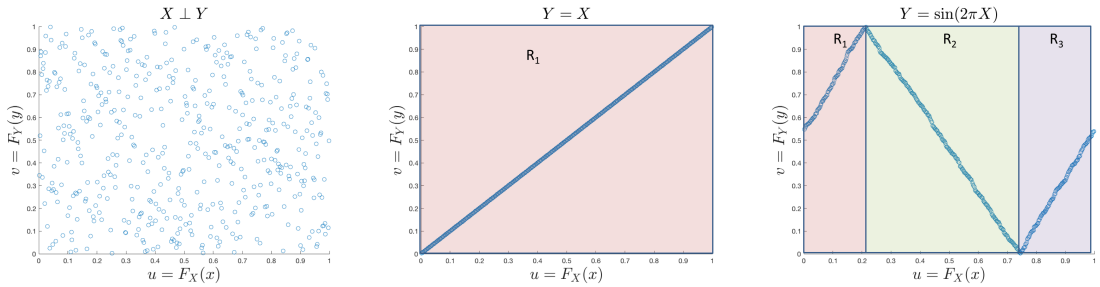
CIM detects statistical dependencies by leveraging concepts from concordance, defined above in (2.5). However, measures of concordance do not perform well for measuring nonmonotonic dependencies. This is because two random variables can be perfectly associated, while having the probability of concordance, $P[(X_1 - X_2)(Y_1 - Y_2) > 0]$, equal to the probability of discordance, $P[(X_1 - X_2)(Y_1 - Y_2) < 0]$, yielding a null concordance function Q. An example of such an association is $Y = X^2$, with $X \sim U[-1, 1]$. Thus, in order to use concordance as a measure of nonmonotonic dependence, one must consider regions of concordance and discordance separately; this provides the basis of *CIM*, which computes a weighted average of $|\tau_N|$ for each of these regions.

To develop *CIM*, we begin by proving that a set of observations drawn from any mapping can be grouped into concordant and discordant subsets of pseudo-observations that are piecewise linear functions of each other. Let $F_{X_d}(x_d(m))$ be the m^{th} pseudo-observation for the d^{th} dimensional data point and denote the range-space of (X, Y), where X and Y are random variables, to be the subset of $\mathcal{R}^2$ which encompasses every pair of values that the bivariate random variable (X, Y) can take on. We can then state the following theorem:

Theorem 3 *Suppose X and Y are random variables that are associated through the mapping defined by $g(\cdot)$, where $g(\cdot)$ is monotone over each of the regions Ω_i $\forall i = 1, \ldots, n$ that form a partition of the range-space of (X, Y). Define the random variables $U = F_X(x)$ and $V = F_Y(y)$. Then, V is a piecewise linear function of U.*

The proof is provided in B. Theorem 3 shows that if two random variables are associated in a deterministic sense, their Cumulative Distribution Functions (CDFs) are piecewise linear functions of each other. This implies that the pseudo-observations of realizations of these dependent random variables can be grouped into regions of concordance and discordance. Furthermore, in each region, the dependent variable's pseudo-observations are linear functions of the independent ones, contained in the unit square $\mathbf{I}^2$. Using this as a basis, *CIM* detects dependencies by identifying regions of concordance and discordance after transforming the original data, x and y, into the pseudo-observations, $F_X(x)$ and $F_Y(y)$, respectively.

As displayed in Fig. 4.4a, by definition of concordance, in the independence scenario, no regions of concordance or discordance exist. Similarly, as depicted in Fig. 4.4b, for monotonic dependencies only one region, $\mathbf{I}^2$, exists. Finally, for nonmonotonic dependencies, many regions may exist. As an example, Fig. 4.4c displays the pseudo-observations of sinusoidal functional dependence. Here, it is easy to see that R_1 and R_3 are regions of concordance, and R_2 is a region of discordance.



(a) (b) (c)

Figure 4.4: Regions of concordance and discordance for three different scenarios: (a) shows two independent random variables, in which case by definition there are no regions of concordance or discordance; (b) shows comonotonic random variables, in which there is one region of concordance, R_1; (c) shows a sinusoidal dependence between two random variables, in which there are two regions of concordance, R_1 and R_3, and one region of discordance, R_2.

The foregoing examples motivate the following definition of *CIM*:

$$CIM = \sum_i \left(w_i |\tau_N^i| \right), \tag{4.3}$$

where $|\tau_N^i|$ is the absolute value of τ_N for the i^{th} region and w_i is the ratio of the area of region R_i to $\mathbf{I}^2$. From (4.3) and the properties of τ_N, *CIM* reduces to τ for monotonic continuous random variables, and zero for independent random variables. It should be noted that (4.3) defines *CIM* metric, but an algorithm is required in order to identify each region for which τ_N is computed. In Section 4.3.3, we propose an algorithm to identify these regions.

4.3.2 Properties of *CIM*

In this section, we describe the properties of *CIM* defined in (4.3). We begin by discussing Rényi's seven properties of dependence measures, and show that *CIM* satisfies all of them. We then prove that *CIM* satisfies the Data Processing Inequality (DPI), which implies that it satisfies self-equitability. Finally, we briefly discuss Reshef's definition of equitability and it's application

to *CIM*.

Dependence Metric Properties

Rényi [1959] defined seven desirable properties of a measure of dependence, $\rho^*(X, Y)$, between two random variables X and Y:

1. $\rho^*(X, Y)$ is defined for any pair of non-constant random variables X and Y.
2. $\rho^*(X, Y) = \rho^*(Y, X)$.
3. $0 \leq \rho^*(X, Y) \leq 1$.
4. $\rho^*(X, Y) = 0$ iff $X \perp\!\!\!\perp Y$.
5. For bijective Borel-measurable functions, f, g: $\mathbb{R} \to \mathbb{R}$, $\rho^*(X, Y) = \rho^*(f(X), g(Y))$.
6. $\rho^*(X, Y) = 1$ if for Borel-measurable functions f or g, $Y = f(X)$ or $X = g(Y)$.
7. If $(X, Y) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then, $\rho^*(X, Y) = |\rho(X, Y)|$, where ρ is the correlation coefficient.

CIM satisfies these seven properties. The proof is provided in E.

Self Equitability and the Data Processing Inequality

As noted by Kinney and Atwal [2014], the DPI and self equitability are important, desirable properties of a dependence metric. In this section, we prove that $|\tau|$ and *CIM* both satisfy the DPI, and are thus both self-equitable for continuous random variables. We show that the scaling factors proposed in (4.1) and (4.2) to account for discrete and continuous random variables, unfortunately, does not satisfy the DPI. We then propose a solution to allow *CIM* to satisfy the DPI, even in the discrete and hybrid scenarios.

The DPI is a concept that stems from information theory. It states that if random variables X, Y, and Z form a Markov chain, denoted by $X \to Y \to Z$, then $I(X;Y) > I(X;Z)$, where $I(X;Y)$ is the mutual information between X and Y defined as

$$I(X;Y) = \int_Y \int_X f_{XY}(x,y) \mathbf{log} \frac{f_{XY}(x,y)}{f_X(x) f_Y(y)} dxdy, \tag{4.4}$$

where $f_{XY}(x,y)$ is the joint distribution of X and Y, and $f_X(x)$ and $f_Y(y)$ are the marginal distributions of X and Y, respectively [Cover and Thomas, 2006]. Intuitively, it asserts that information is never gained when being transmitted through a noisy channel [Kinney and Atwal, 2014]. As an analog to the information theoretic definition of the DPI, Kinney and Atwal [2014] define a dependence metric D to satisfy the DPI if and only if $D(X;Y) \geq D(X;Z)$, whenever the random variables X, Y, and Z form the Markov chain, $X \to Y \to Z$. Here, we prove that ***CIM***, as defined by (4.3), satisfies the DPI for continuous random variables, and in the sequel we show that a modified version of ***CIM***, termed CIM^S satisfies the DPI for discrete and hybrid random variables. More precisely, we have

Theorem 4 *If the continuous random variables X, Y, and Z form a Markov chain $X \to Y \to Z$, then $CIM(X,Y) \geq CIM(X,Z)$.*

The proof is given in C. An immediate implication of ***CIM*** satisfying DPI is that it is a self-equitable statistic. A dependence measure $D(X;Y)$ is said to be self-equitable if and only if it is symmetric, that is, ($D(X;Y) = D(Y;X)$), and satisfies $D(X;Y) = D(f(X);Y)$, whenever f is a deterministic function, X and Y are variables of any type, and $X \to f(X) \to Y$, implying that they form a Markov chain [Kinney and Atwal, 2014]. Self equitability implies that $CIM(X,Y)$ is invariant under arbitrary invertible transformations of X or Y [Kinney and Atwal, 2014], which is in-fact a stronger condition than Rényi's 5^{th} property given in Section 3.2.1.

When X, Y, and Z are discrete or hybrid random variables, the scaling factors proposed in

(4.1) and (4.2) to account for discrete and hybrid random variables cannot be guaranteed to satisfy the DPI. Indeed, for a Markov chain $X \rightarrow Y \rightarrow Z$, the scaling factor for $\tau_N(X,Y)$ is $\sqrt{(1-E[\Delta F_X(X)])(1-E[\Delta F_Y(Y)])}$. However, the relationship between $E[\Delta F_Y(Y)]$ and $E[\Delta F_Z(Z)]$ is not obvious from $X \rightarrow Y \rightarrow Z$ because the scaling factors are dependent on the marginal distributions, which are related to, but not derivable from, knowledge of the joint distribution's copula alone. To enable *CIM* to satisfy these properties, we propose to remove the scaling factors defined in (4.2) and to compute $|\hat{\tau}|$ of the standard extension copula, for both discrete and hybrid random variables (i.e. the numerator in (4.1) and (4.2)). More specifically, define

$$\tau_S(X,Y) = 4 \int C^S_{XY} dC^S_{XY} - 1,$$

where C^S_{XY} is the standard extension copula [see Nešlehová, 2007, Equation 2]. From [Denuit and Lambert, 2005] and [Nešlehová, 2007], we infer that C^S_{XY} follows the concordance ordering. Combining this property with Theorem 4, it is straightforward to show that $\tau_S(X,Y)$ does indeed satisfy the DPI, and thus CIM^S defined as

$$CIM^S = \Sigma_i \left(w_i |\tau^i_S| \right), \tag{4.5}$$

also satisfies the DPI, where w_i is defined as before in (4.3). The consequences of not using the scaling factor are that τ_S does not reach $+1$ or -1 for either the perfectly comonotonic, or countermonotonic discrete, or hybrid cases. However, from the perspective of ranking dependencies, as long as concordance order is satisfied, the absolute value of τ_S is irrelevant; only the relative values of $\tau_S(X,Y)$ and $\tau_S(X,Z)$ are pertinent. It should be noted that the effect of the scaling factors is decreased in two scenarios: 1) the support of the discrete random variables is large, and 2) the probability mass is more evenly distributed among the support of the discrete random variable [Genest and Nešlehová, 2007]. In these scenarios, it is safe to use *CIM* as defined in (4.3)

when applying it to an algorithm that requires the DPI principle, as the effect of the scaling factor is negligible in the presence of noisy data. However, in scenarios where the support of the discrete random variables are small, or the discrete distribution is skewed, it is advisable to use (4.5) when comparing the relative strengths of dependencies.

Equitability and Noise Properties

Equitability is a measure of performance of a statistic under noise. Notionally, an equitable statistic assigns similar scores to equally noisy relationships of different types [Reshef et al., 2011]. Kinney and Atwal [2014] formalize this concept as R^2-equitability. Recall that a dependence measure $D[X;Y]$ is R^2 equitable if and only if, when evaluated on a joint probability distribution $p(X,Y)$, that corresponds to a noisy functional relationship between two real random variables X and Y, the relation given by

$$D[X;Y] = g(R^2([f(X);Y])) \tag{4.6}$$

holds true, where g is a function that does not depend on $p(X,Y)$, R^2 denotes the squared Pearson correlation measure, and f is the function defining the noisy functional relationship, namely $Y = f(X) + \eta$, for some random variable η.

Through simulations, we observe that τ is not an equitable metric. Following Reshef et al. [2015], we compute the equitability curves, which show the relationship between τ and R^2 for different relationships, for the two association patterns $Y = X$ and $Y = e^X$. These are displayed in Fig. 4.5. The worst interpretable interval, which can be informally defined as the range of R^2 values corresponding to any one value of the statistic is represented by the red hashed line. Fig. 4.5 depicts a large interval, which is indicative of the lack of R^2-equitability of this estimator. From a theoretical perspective, this can be understood from (4.2), which shows that the distances between

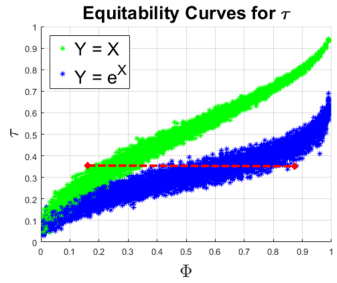


Figure 4.5: Equitability curves for Kendall's τ for two functional dependencies, where $X \sim U[2, 10]$ and $Y = X$ in green and $Y = e^X$ in blue. Here, we see that the worst interpretable interval, shown by the red hashed line, is large, indicating lack of equitability of $\hat{\tau}$.

points are not considered, only their relative rankings. Because τ is not equitable and *CIM* is based on τ, the latter is also not equitable according to (4.6). Additionally, the distance argument leads to the conclusion that any concordance-based measures are not R^2-equitable.

4.3.3 Proposed Algorithms

In this section, we propose an algorithm to estimate *CIM* metric. From (4.3), *CIM* estimator can be defined as

$$\widehat{CIM} = \sum_i \left(w_i |\hat{\tau}^i_{KL}| \right), \tag{4.7}$$

where $|\hat{\tau}^i_{KL}|$ is the absolute value of (4.2) for the i^{th} region and w_i is the ratio of the number of samples in R_i to the total number of samples being considered. The steps in Listing 1 present a high level outline of estimating *CIM* index.

Listing 1: *CIM* Estimation Overview

1. Transform the original data to pseudo-observations.
2. Compute $\hat{\tau}_{KL}$ on the pseudo-observations of increasing subsets of the data in the unit-square.
3. If a large change in $\hat{\tau}_{KL}$ between the previous subset and the current subset is detected,

declare a new region boundary.

4. Apply (4.7) after all region boundaries have been dentified.

More specifically, the first step in approximating *CIM* statistic is to transform the data by applying the probability integral transform, via the empirical cumulative distribution function, to both dimensions of the data independently, generating the pseudo-observations. Next, the unit square is scanned to identify regions of concordance and discordance. The output of this step for independent, linear, and sinusoidal association patterns is shown in Figs. 4.4 (a), (b), and (c), respectively. The scanning is performed by creating an empty region in the unit square, increasing the size of that region by a defined amount, si, in the u dimension, and estimating the $\hat{\tau}_{KL}$ metric over all points contained within that region, denoted by $\hat{\tau}_{KL}$. Every $\hat{\tau}_{KL}$ is compared to the previous value, denoted as $\hat{\tau}'_{KL}$. If

$$|\hat{\tau}_{KL}| < |\hat{\tau}'_{KL}| - \frac{\sigma_{\hat{\tau}_{KL}}}{\sqrt{M}} u_{1-\frac{\alpha}{2}}, \tag{4.8}$$

where $\sigma_{\hat{\tau}_{KL}}$ is the standard deviation of the $\hat{\tau}_{KL}$ and $u_{1-\frac{\alpha}{2}}$ is the quantile of the standard normal distribution at a significance level of α, then a new region boundary is declared. Stated differently, if the current value $|\hat{\tau}_{KL}|$ has decreased by more than a defined amount of confidence, α, from its previously estimated value $|\hat{\tau}'_{KL}|$, then the algorithm declares this a new boundary between monotonic regions. Fig. 4.6 pictorially depicts these steps. In Fig. 4.6 (a), R_1 that has been identified by the algorithm as the one that contains points of concordance, noted by $\hat{\tau}'_{KL}$. Additionally, the green region in Fig. 4.6 (a) shows the region under consideration by the algorithm, which is an increment of the one identified by si. $\hat{\tau}'_{KL}$ and $\hat{\tau}_{KL}$ are compared according to the criterion given above. In Fig. 4.6 (a), the criterion in (4.8) yields the decision that the points in the green region belong to the same region, denoted by R_1. In Fig. 4.6 (b), the same criterion in (4.8) yields the decision that the points in green belong to a new region, R_2, as depicted in Fig. 4.4c.

In order to maximize the power of *CIM* estimator against the null hypothesis that $X \perp\!\!\!\perp Y$, the

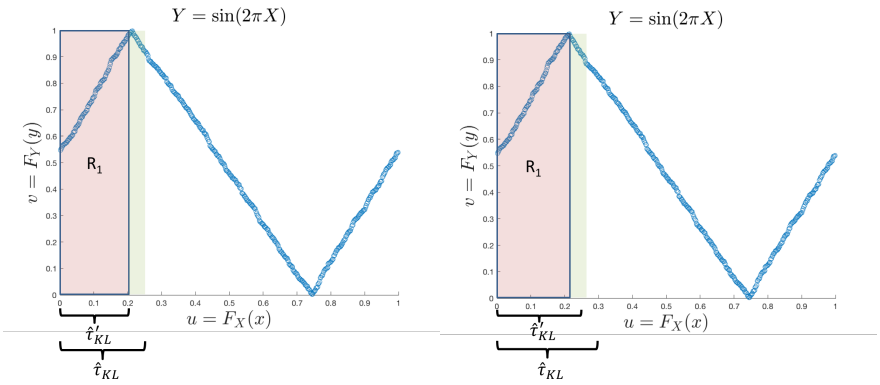


Figure 4.6: Operation of *CIM* algorithm. In (a), *CIM* algorithm decides that the green region belongs to the same region as R_1. In (b), *CIM* algorithm decides that green region belongs to a new region, different from R_1.

scanning process is conducted for multiple values of si, both orientations of the unit-square (u-v, and v-u), and sub-intervals of u and v separately. The scanning and orientation of the unit square, which maximizes the dependence metric, is the approximate value of *CIM*. The minimum scanning increment (width of the green region in Fig. 4.6 (a)), msi, and the confidence level, α, are the two hyperparameters for the proposed algorithm. The value of msi used in all the simulations, except the sensitivity study, is $\frac{1}{64}$. The denominator of this value bounds the size and frequency of changes to the monotonicity that the algorithm can detect. By choosing $\frac{1}{64}$, it is found that all reasonable dependence structures can be captured and identified. The value of α used in all the simulations is 0.2, which was found to be a good tradeoff between overfitting and detecting new regions experimentally from a statistical power perspective. The experiments conducted in Section 4.1 and 4.4.2 corroborate these choices. The complete pseudocode for estimating *CIM* index is shown in Algorithm 1 in F; additionally, a reference implementation is also provided [1].

[1] `https://github.com/stochasticresearch/depmeas/blob/master/algorithms/cim_cc.m`

Algorithm Validity

In this section, we discuss the theoretical validity of the convergence of Algorithm 1. From [Schmock, 2010], we have

$$P\left[\hat{\tau}_{KL} - \frac{\sigma_{\hat{\tau}_{KL}}}{\sqrt{M}} u_{1-\frac{\alpha}{2}} \leq \tau_{KL} \leq \hat{\tau}_{KL} + \frac{\sigma_{\hat{\tau}_{KL}}}{\sqrt{M}} u_{1-\frac{\alpha}{2}}\right] \xrightarrow{M\to\infty} 1-\alpha, \tag{4.9}$$

where M is the number of samples available to estimate τ_{KL} and the other variables were defined above in (4.8). From (4.9), we can state the following:

Theorem 5 *The region detection criterion, $|\hat{\tau}_{KL}| < |\hat{\tau}'_{KL}| - \frac{\sigma_{\hat{\tau}_{KL}}}{\sqrt{M}} u_{1-\frac{\alpha}{2}}$, guarantees that as $M \to \infty$, a change in monotonicity in the dependence structure will be detected with a probability of $1-\alpha$, where α is a configurable confidence level, and M is the number of samples available to estimate τ_{KL}.*

The proof is given in D. Theorem 5 guarantees that if the unit square is scanned across v for the full-range of u, any injective or surjective association pattern's changes in monotonicity will be detected with probability of $1-\alpha$ as $n \to \infty$. For association patterns which map multiple values of y to one value of x (such as the circular pattern), the range of u is divided and scanned separately. Because the dependence structure is not known a-priori, various scans of the unit-square are performed at different ranges of u and v. As stated above, the configuration that maximizes the dependence metric is then chosen amongst all the tested configurations.

Algorithm Performance

In this section we investigate the performance of Algorithm 1 using various synthetic datasets. We show that the proposed algorithm is robust to both input hyperparameters, msi and α. We also

investigate the convergence properties and speed of convergence of $\widehat{CIM}$ as estimated by Algorithm 1. Because the algorithm performance depends heavily on how well it detects the regions of concordance and discordance, we begin by characterizing the region detection performance.

To test the region detection performance, we simulate noisy nonmonotonic relationships of the form

$$Y = 4(X - r)^2 + \mathcal{N}(0, \sigma^2), \tag{4.10}$$

where $X \sim U(0, 1)$. By varying r and the number of samples, M, that are drawn from X, nonmonotonic relationships of this form comprehensively test the algorithm's ability to detect regions for all types of association. This is because r directly modulates the angle between the two piecewise linear functions at a region boundary, and the number of samples and the noise level test the performance of the decision criterion specified previously in (4.8) as a function of the number of samples. After generating data according to (4.10) for various values of r, M, and σ, Algorithm 1 is run on the data and the boundary of the detected region is recorded, for 500 Monte-Carlo simulations. A nonparametric distribution of the detected regions by Algorithm 1 for different values of r and M is displayed in Fig. 4.7. It is seen that on average, the algorithm correctly identifies the correct region boundary. In the scenario with no noise, the variance of the algorithm's detected region boundary is small, regardless of the sample size. For larger levels of noise, the variance decreases with the sample size, as expected.

Next, we investigate the sensitivity of Algorithm 1 to the hyperparameter msi. For various dependency types, we compute the maximum deviation of the *CIM* value over 500 Monte-Carlo simulations for sample sizes, M, ranging from 100 to 1000 for msi taking on one the values of the set $\{\frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{32}, \frac{1}{64}\}$, for $\alpha = 0.2$. Fig. 4.8 shows the maximum deviation of the estimated CIM value for each value of noise over sample sizes ranging from 100 to 1000 for eight different association patterns for these values of msi. The results show that when the dependency is not masked by the msi parameter, the algorithm's maximum deviation over the noise range, sample

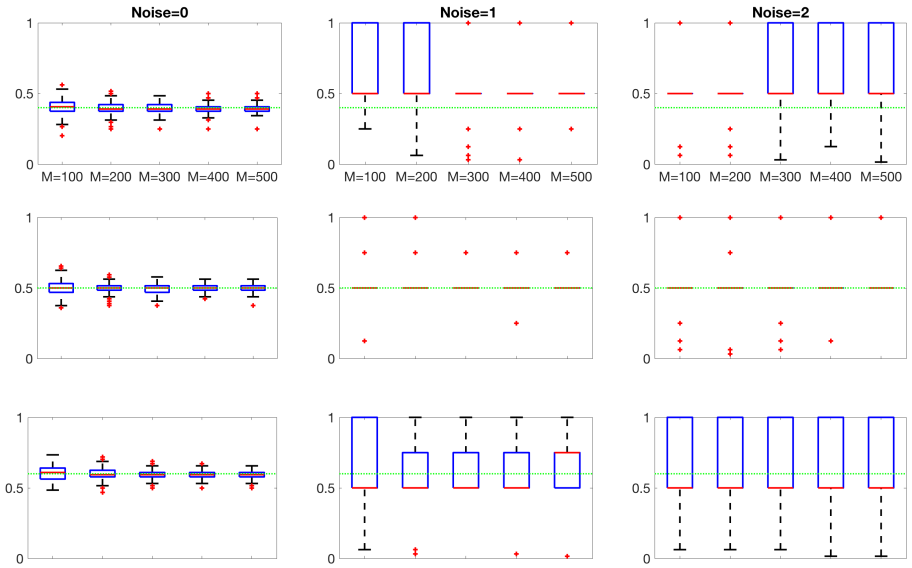


Figure 4.7: Region boundaries detected by Algorithm 1 for various noise levels and sample sizes. The hashed green line represents the actual region boundary, r, and the box and whisker plots represent the non-parametric distribution of the detected region boundary by Algorithm 1, for an $msi = \frac{1}{64}$ and $\alpha = 0.2$.

sizes, and dependencies tested is no greater than 4×10^{-3}. This is shown by the blue lines for the linear, quadratic, fourth-root, circular, and step function dependnecies, and by the red lines in the cubic and sinusoidal dependencies. When the dependency is masked by the msi, as expected, the algorithm is sensitive to the chosen value of msi. As seen in Fig 4.8, the maximum deviation of the algorithm for low-noise levels can reach a value close to 0.5 for the low-frequency sinusoidal dependency. From this, we can infer that small values of msi should be chosen for more robust results for estimating $\widehat{CIM}$, as they empirically have minimal effect on measuring association patterns that do not have many regions of monotonicity, but have a positive effect on detecting and measuring dependencies with many regions of monotonicity. The only drawback of choosing very small values of msi is that they require more computational resources.

Next, we test the sensitivity of the algorithm to various values of α. More specifically, for the various dependence structures that are considered, we compute the maximum deviation of the *CIM* estimation over 500 Monte-Carlo simulations for sample sizes, M, ranging from 100 to 1000 for α taking on one of the values of the set $\{0.05, 0.10, 0.15, 0.20, 0.25, 0.30\}$, for $msi = \frac{1}{64}$. Fig. 4.9 displays the maximum deviation of the estimated CIM value for each value of noise over sample sizes ranging from 100 to 1000 for eight different association patterns. The results show that the algorithm is minimally sensitive to the value of α for dependencies that have more than a small number of monotonic regions, such as the sinusoidal dependencies. This is easily explained by (4.9) and (4.8) which show that the upper bound of the variance of the τ estimate is high with small sample sizes and that the small number of samples combined with a large α prevent reliable detection of region boundaries.

Finally, following Theorem 5, we demonstrate through simulations that Algorithm 1 converges to the true *CIM* value. The results of the algorithm's convergence performance are displayed in Fig. 4.10. The subtitles for each subplot indicate the number of samples required such that the error between the estimated value, $\widehat{CIM}$, and the true value, *CIM*, over all computed noise levels is less

than 0.01 over 500 Monte-Carlo simulations. It can be seen that for the dependencies with small numbers of regions of monotonicity, Algorithm 1 converges very quickly to the true value over all noise levels. On the other hand, the dependencies with a large number of regions of monotonicity, such as the high frequency sinusoidal relationship depicted in the fifth subplot, a larger number of samples is required in order to ensure convergence. This can be explained from the fact that the variance of the $\widehat{CIM}$ increases as the number of samples decreases. Thus, with a smaller number of samples in a dependency structure, the increased variance leads Algorithm 1 to make incorrect decisions regarding the region boundaries. As the number of samples increases, the detection performance increases.

Sampling Distribution of *CIM*

Simulations were also conducted in order to characterize the null distribution of *CIM* approximation provided by Algorithm 1. It is found experimentally that the proposed algorithm produces a statistic whose probability distribution can be approximated by the Beta distribution, as displayed in Fig. 4.11. Figs. 4.11 (b) and (c) both show that as the sample size increases, the α shape parameter remains relatively constant while the β shape parameter increases linearly as a function of M. This roughly corresponds to a distribution converging to a delta function centered at zero. This is a desirable property because it implies that *CIM* approximation algorithm yields a value close to 0 for data drawn from independent random variables with a decreasing variance as the number of samples used to compute *CIM* increases. It is interesting to note that the Beta distribution is intimately connected to rank statistics. More precisely, one can show that if $\{U_1, \ldots, U_n\}$ are n independent random variables, each following $\sim U[0, 1]$, then the distribution of the k^{th} order statistic, $U_{(k)}$ follows Beta$(k, n + 1 - k), \forall\ k = [1, n]$ [Ahsanullah et al., 2013].

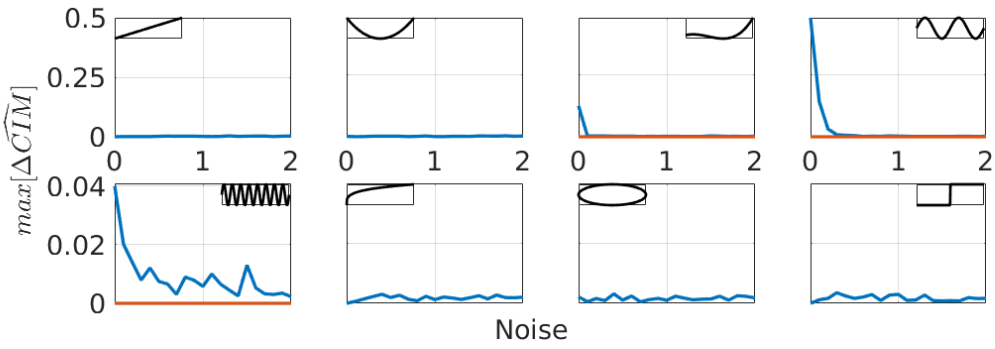


Figure 4.8: The maximum sensitivity of Algorithm 1 for various association patterns (shown in the upper left inset) swept over different values of noise for sample sizes (M) ranging from 100 to 1000 and msi taking on one of the values in the set $\{\frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{32}, \frac{1}{64}\}$, with $\alpha = 0.2$. The red lines show the maximum sensitivity when the msi value does not mask the dependence structure for the cubic and sinusoidal dependencies.

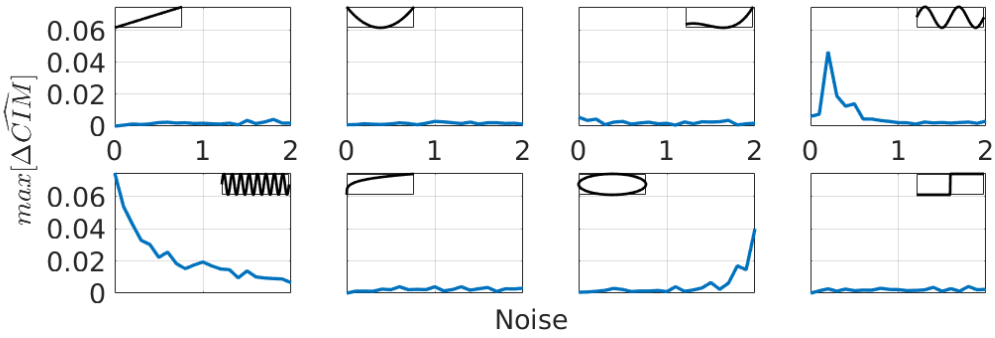


Figure 4.9: The maximum sensitivity of Algorithm 1 for various association patterns (shown in the upper left inset) swept over different values of noise for sample sizes,M, ranging from 100 to 1000 and α taking on one of the values in the set $\{0.05, 0.10, 0.15, 0.20, 0.25, 0.30\}$, with $msi = \frac{1}{64}$.

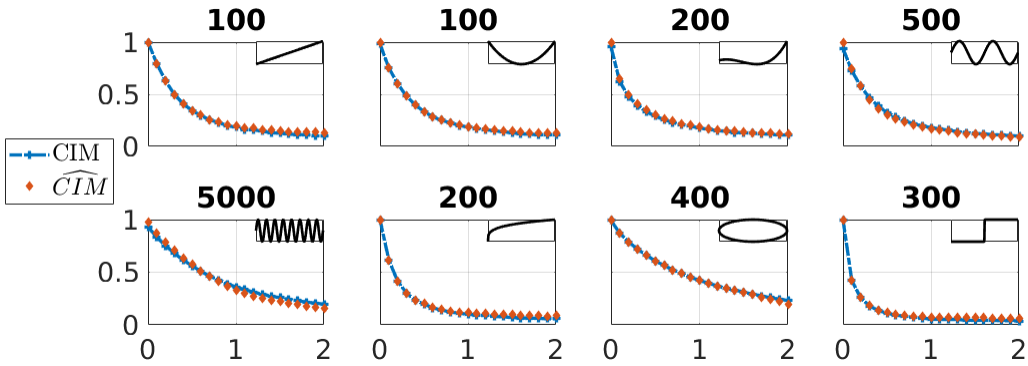


Figure 4.10: Theoretical and estimated values of *CIM* for various association patterns shown in the upper right inset swept over different noise levels. The subtitle shows the minimum number of samples for $\widehat{CIM}$ to be within 0.01 of *CIM* over all noise levels tested for 500 Monte-Carlo simulations. The simulations were conducted with $\alpha = 0.2$ and $msi = \frac{1}{64}$.

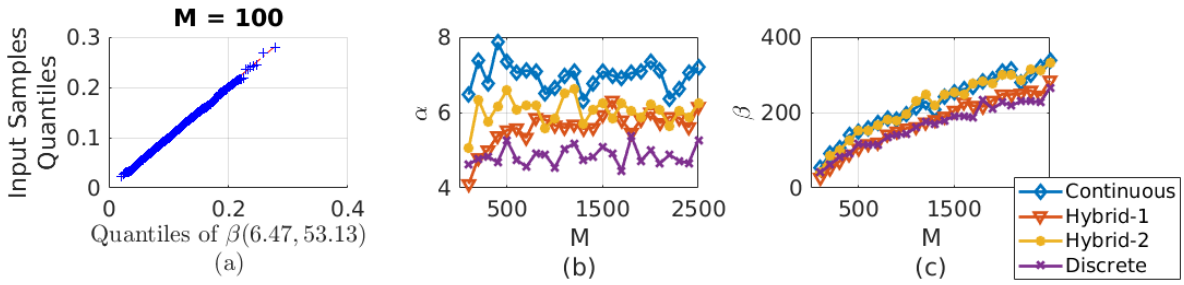


Figure 4.11: (a) QQ-Plot of CIM for continuous random variables X and Y such that $X \perp\!\!\!\perp Y$ and $M = 100$, (b) α of the distribution of *CIM* as a function of M, (c) β of the distribution of *CIM* as a function of M

Computational Complexity

In this section we describe the computational complexity of computing *CIM* algorithm above. We propose a new algorithm to compute $\hat{\tau}_{KL}$ to achieve a computational complexity of $\mathcal{O}(n^2)$ when estimating *CIM* for continuous and discrete random variables, and $\mathcal{O}(n^3)$ when estimating *CIM* for hybrid random variables.

The core of Algorithm 1, described earlier, consists of repeated computations of $\hat{\tau}_{KL}$. If one were to näively compute this, by recomputing the number of concordant and discordant pairs every time a new region was tested, the operations required to compute the number of concordant and discordant samples would exponentially increase. Instead, we propose another algorithm to compute $\hat{\tau}_{KL}$ efficiently while accumulating new samples into the batch of data for which the value of $\hat{\tau}_{KL}$ is desired (i.e. when expanding the region by si). The essence of this algorithm is that it pre-sorts the data in the direction being scanned, so that the number of concordant and discordant samples do not need to be recomputed in every iteration of the scanning process. Instead, the sorted data allows us to store in memory the number of concordant and discordant samples, and update this value every time a new sample is added to the batch of samples being processed. Additionally, during the sorting process, the algorithm converts floating point data to integer data by storing the statistical ranks of the data rather than the data itself, allowing for potentially efficient FPGA or GPU based

implementations. The efficient algorithm to compute $\hat{\tau}_{KL}$ for continuous and discrete data, given a new sample, is described in the CONSUME function of Algorithm 2, which is detailed in G. From Algorithm 2, it is seen that if n samples are to be processed, then the CONSUME function is called n times. For clarity of exposition, the remaining helper functions are not presented; however, their operation is only to initialize the variables.

The CONSUME function has a computational complexity of $\mathcal{O}(n)$, due to lines 9 and 10 in Algorithm 2, which require computation over a vector of data. The consume function is called n times by Algorithm 1 in order to process all the samples, yielding a total complexity of $\mathcal{O}(n^2)$. It should be noted that lines 9 and 10 in Algorithm 2 are vectorizable operations, and the initial presorting is an $\mathcal{O}(nlog(n))$ operation. For hybrid data, additional calculations are required in the CONSUME function in order to count the number of overlapping samples between discrete outcomes in the continuous domain, as described in (4.2). This requires an additional $\mathcal{O}(n)$ operations, bringing the computational complexity of Algorithm 1 to process hybrid random variables to $\mathcal{O}(n^3)$. For clarity, the pseudocode to compute the overlapping points is not shown in Algorithm 2, but a reference implementation to compute $\hat{\tau}_{KL}$ is provided[2].

4.4 Simulations

In this section, we compare *CIM* to other metrics of dependence and analyze their performance. We begin by conducting synthetic data experiments to understand the bounds of performance for all state-of-the-art dependence metrics. We then apply *CIM* to real world datasets from various disciplines of science, including computational biology, climate science, and finance.

[2] https://github.com/stochasticresearch/depmeas/blob/master/algorithms/taukl_s.m

4.4.1 Synthetic Data Simulations

Following Simon and Tibshirani [2014], we begin by comparing the statistical power of *CIM* against various estimators of mutual information, including k-nearest neighbors estimation [Kraskov et al., 2004], adaptive partitioning MI estimation [Darbellay and Vajda, 1999], and MI estimation based on von Mises expansion [Kandasamy et al., 2015]. The motivation for this simulation stems from Section 4.3.2, where it was proved that *CIM* satisfied the DPI and thus, could be substituted for measures of mutual information. Fig. 4.12 compares these metrics and shows that *CIM* outperforms the compared estimators of mutual information for all dependency types considered [3]. The results displayed in Fig. 4.12 are from simulations with a sample size of $M = 500$. Although we do not include additional plots here, even for small sample sizes such as $M = 100$ (which are typical for biological datasets where estimators of the *MI* are commonly used), *CIM* outperforms the compared estimators of *MI* for all the association patterns tested. These simulations suggest that *CIM* can indeed replace estimators of the *MI* when used with algorithms which rely on the DPI, such as ARACNe [Margolin et al., 2006] or MRNET [Meyer et al., 2007].

We also investigate the power characteristics of *CIM* and estimators of mutual information as a function of the sample size. The green asterisk in Fig. 4.12 displays the minimum number of samples required to achieve a statistical power of 0.8 for the different dependency metrics considered for a noise level of 1.0. A green plus symbol is shown if the number of samples required is beyond the scale of the plot. It is seen that *CIM* outperforms the compared estimators for all dependency types considered. In general, *CIM* displays good small sample performance because it is based on Kendall's τ, which is shown to have superior small sample performance as compared to other metrics of monotonic dependence [Bonett and Wright, 2000, Helsel and Hirsch, 2002].

[3]The source code for Shannon Adaptive Partitioning and von Mises based MI estimators is from the ITE Toolbox [Szabó, 2014]. K-NN based MI estimation source code is from `https://www.mathworks.com/matlabcentral/fileexchange/50818-kraskov-mutual-information-estimator`

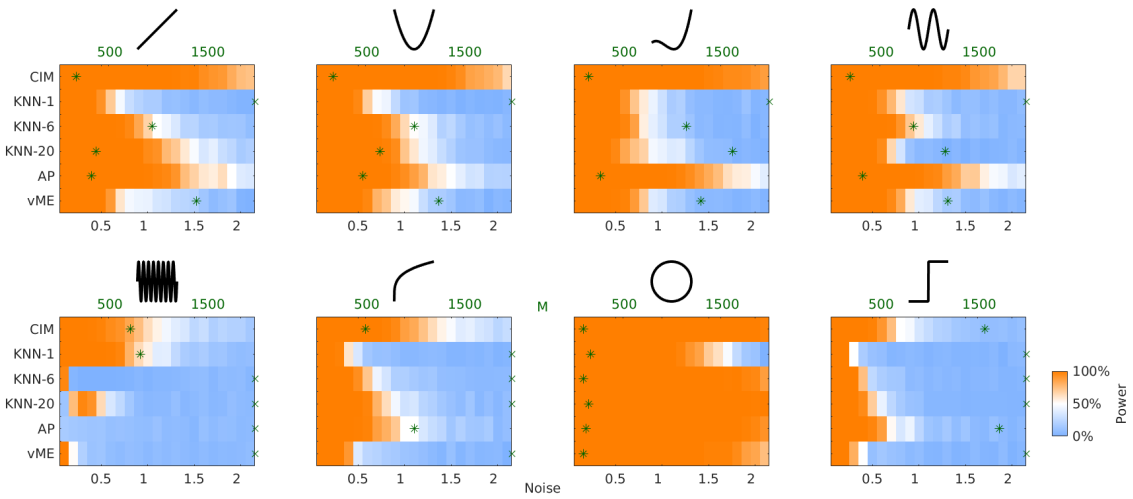


Figure 4.12: Statistical power of *CIM* and various estimators of mutual information including the KNN-1, the KNN-6, the KNN-20, Adaptive Partitioning, and von Mises Expansion for sample size $M = 500$ and computed over 500 Monte-Carlo simulations. Noise-free form of each association pattern is shown above each corresponding power plot. The green asterisk displays the minimum number of samples required to achieve a statistical power of 0.8 for the different dependency metrics considered for a noise level of 1.0. A green plus symbol is shown if the number of samples required is beyond the scale of the plot.

Next, we compare *CIM* to other state-of-the-art dependence metrics, which are not proven to satisfy the DPI. We begin by comparing the estimated indices for various functional and stochastic dependencies for continuous and discrete marginals. The results, displayed in Fig. 4.14, show that *CIM* performs equivalently to other leading measures of dependence, including MIC_e, the *RDC*, the *dCor*, the *Ccor*, and *CoS* for continuous and discrete random variables in the absence of noise. *CIM* achieves $+1$ for all functional dependencies with continuous marginals (Fig. 4.14 (a), (c)) and for monotonic functional dependencies with discrete marginals (Fig. 4.14 (a), (c)), and values close to $+1$ for nonmonotonic functional dependencies with discrete marginals (Fig. 4.14 (d), (e), (f)). Only the *RDC* shows similar performance. However, as shown in Fig. 4.14 (b) and (e), the *RDC* has the highest bias in the independence case. Discrete random variables are not tested for the *Ccor* and *CoS* metrics because they were not designed to handle discrete inputs.

Fig. 4.13 compares the statistical power of *CIM* to other state-of-the-art dependence metrics which

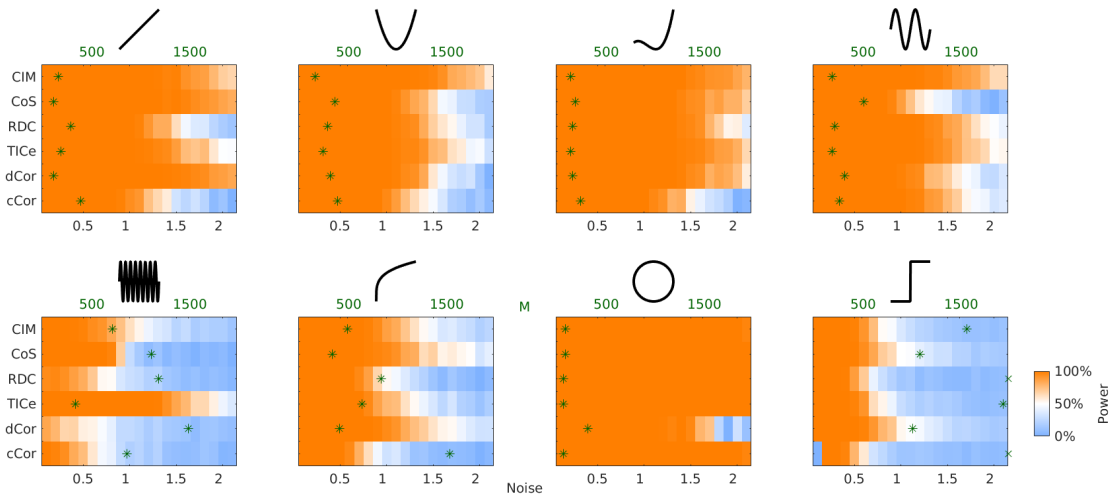


Figure 4.13: Statistical power of *CIM* and various measures of dependence including CoS, the RDC, TICe, the dCor, and the cCor for sample size $M = 500$ and computed over 500 Monte-Carlo simulations. Noise-free form of each association pattern is shown above each corresponding power plot. The green asterisk displays the minimum number of samples required to achieve a statistical power of 0.8 for the different dependency metrics considered for a noise level of 1.0. A green plus symbol is shown if the number of samples required is beyond the scale of the plot.

are not proven to satisfy the DPI. The results in Fig. 4.12 show that *CIM* displays the best performance for quadratic, cubic, and sinusoidal dependence. For linear, fourth-root, and step function dependence, it performs better than the *RDC*, *TIC*, and the *Ccor*, but is beaten by *CoS* and the *dCor*. In the high frequency sinusoidal case, it is more powerful than the *RDC* but less powerful than the *TIC*. This can be explained by the fact that the region configuration which maximizes the dependence (*lines 25-26* in Algorithm 1) becomes more ambiguous as the noise level increases when multiple partitions of the range space of $X - Y$ are needed. Our general observations are that *CoS* and the *dCor* are the best for monotonic dependencies, *CIM* is the best for small numbers of monotonic regions, and *TIC* performs extremely well for high frequency sinusoidal dependencies. The sample size requirements, again shown with the green asterisk and plus symbols, reflect these same observations.

[3]The code for these dependency metrics and simulations is provided here: `https://github.com/stochasticresearch/depmeas`

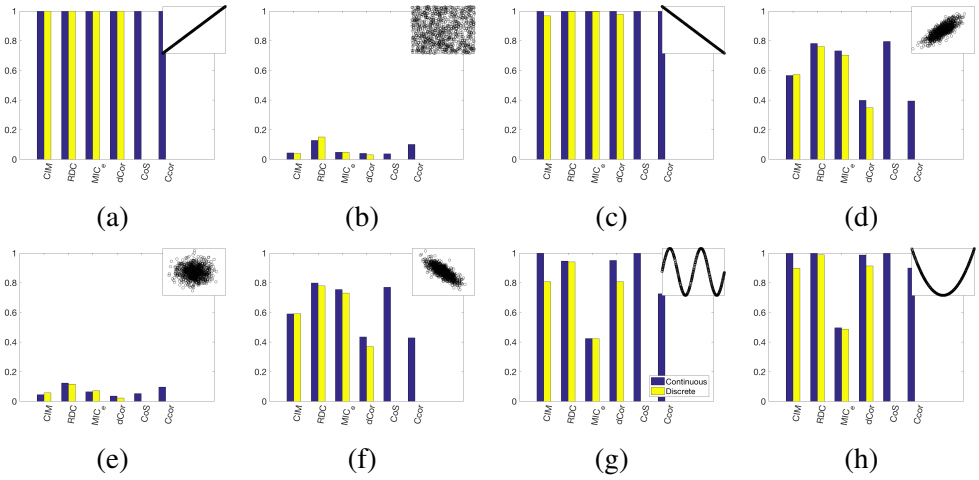


Figure 4.14: Values attained by various dependence metrics for various noiseless functional associations (a),(c),(g),(h), and (i) and Gaussian copula associations (d), (e), and (f). (b) is the independence case, and (e) is the Gaussian copula with $\rho = 0$.

4.4.2 Real Data Simulations

In this section, we apply *CIM* metric to various data exploration and machine learning problems using real-world data, including discovering interesting dependencies in large datasets, Markov network modeling, stochastic modeling of random variables.

Data Exploration

We begin by applying *CIM* metric to real data with the primary goal of characterizing the monotonicity structure of data from different areas of science. This is motivated by both the fields of joint probabilistic data modeling and data exploration. More explicitly, for joint probabilistic modeling of high dimensional datasets, many copula-based techniques are beginning to be adopted in practice, including copula Bayesian networks [Elidan, 2010] and vine copula models [Bedford and Cooke, 2002] due to their flexibility in modeling complex nonlinear relationships between random variables. The authors of these methods advocate the use of parametric copula families for modeling local joint probability distributions. The main reason for this is that it is computation-

ally efficient to estimate a parametric copula for a joint dataset using the relationship between the copula parameter, θ, and a measure of concordance such as Kendall's τ. However, popular copula families such as the Archimedean and Gaussian families only capture monotonic dependencies. Thus, if datasets being modeled are nonmonotonic, these copulas will fail to model all the dynamics of the underlying data. Conversely, if the dependencies within these datasets are monotonic, these efficient procedures can be used and to fit the data to known copula families, and computationally expensive techniques such as estimating empirical copulas can be ignored. Thus, to know whether a parametric copula family can be used, the monotonicity structure must be understood. Therefore, from a copula modeling and analysis perspective, knowledge of the monotonicity structure provides more actionable information than Reshef's proposed nonlinearity coefficient, defined as

$$\theta_{Reshef} = MIC - \rho, \tag{4.11}$$

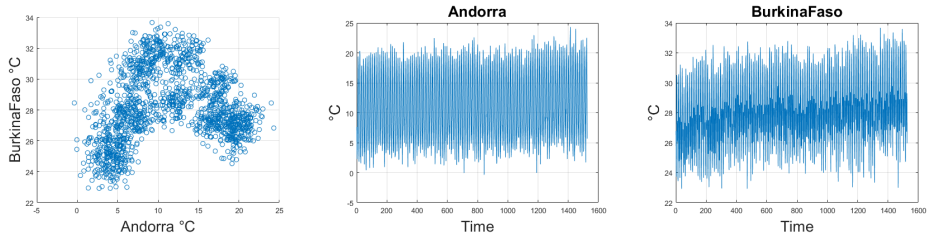
where ρ is the Pearson's correlation coefficient [Pearson, 1895]. Interestingly, copulas can capture monotonic nonlinear relationships while the nonlinearity coefficient defined in (4.11).

In order to answer these questions, we process pairwise dependencies for multiple datasets related to gene expression data, financial returns data, and climate features data[4]. For each pairwise dependency within a dataset, we count the number of monotonic regions by examining the number of regions detected by Algorithm 1. Additionally, to prevent overfitting, we decide that a pairwise dependency only has one monotonic region if the value of $\hat{\tau}_{KL}$ is within 5 % of the estimated value of *CIM*. When time-series data is compared, we only include results of dependencies where the data is considered stationary by the Dickey-Fuller test, at a significance level of $\alpha = 0.05$, and ensure time coherency between the series being compared. Due to the CIM's reliance on copulas, the only requirement is that the data be identically distributed; independence between samples is not required because a copula can capture both inter-dependence and serial dependence within

[4]Details of the datasets used and how they were processed are provided in H

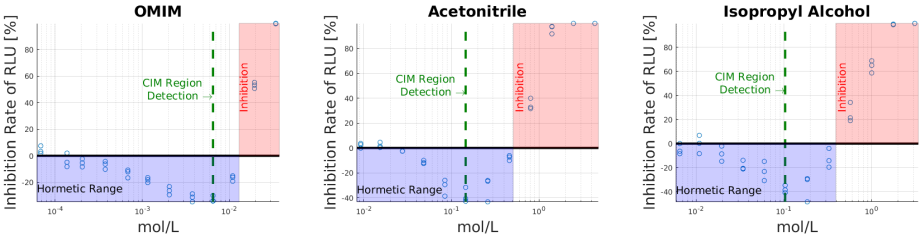
realizations of a random variable. Additionally, we only count dependencies if the dependence metric is statistically significant at a level of $\alpha = 0.05$ and the dependence strength exceeds a value of 0.4 as measured by *CIM* estimation algorithm. Dependencies are only calculated for all unique combinations of features *within* each dataset. With these procedures, after processing 7765 pairwise dependencies which meet the criterion above for various cancer datasets, we find that 96% of gene expression indicators within a cancer dataset are in fact monotonic. Similarly, we process 73 pairwise dependencies between closing price returns data for 30 major indices over a period of 30 years. We find that 99% of the dependencies are monotonic. Finally, we process over 42185 pairwise dependencies of El-Nino indicators in the Pacific ocean, land temperatures of major world cities over the past 200 years, and air quality indicators in major US cities in the past 15 years. In these datasets, termed climate related datasets, we find that 97% of the dependencies within each dataset that meet the criterion above are monotonic. The prevalence of monotonicity in these datasets suggests that techniques that use copula modeling with popular copula families such as the Gaussian or Archimedean families will tend to capture the underlying dynamics of the data properly.

Conversely, *CIM*'s unique ability to identify regions of monotonicity can be used to identify "interesting" dependence structures that may warrant closer analysis by subject matter experts. As an example, Fig. 4.15 shows a nonmonotonic association pattern that was automatically discovered by *CIM* between temperature patterns in Andorra and Burkina Faso, while mining over $27,000$ pairwise dependencies. As shown in Fig. 4.15, the time-series patterns do not clearly reveal this nonmonotonic dependency structure. This example serves to highlight the ability of *CIM* to discover these kinds of dependence structures automatically.



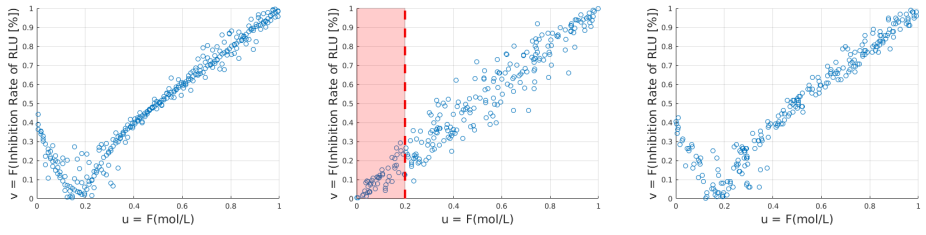
(a) (b) (c)

Figure 4.15: (a) Scatter plot of time-aligned temperature data from Andorra and Burkina Faso, which reveals a nonmonotonic association pattern (b) Time-series of the temperature data from Andorra (c) Time-series of the temperature data from Burkina Faso.



(a) (b) (c)

Figure 4.16: (a) The hormetic effect of 1-octyl-3-methylimidazolium chloride ([OMIM]Cl, CAS RN. 64697-40-1) on firefly luciferase after 15 min exposure (b) the hormetic effect of acetonitrile (CAS RN. 75-05-8) on photobacteria Vibro-qinghaiensis sp. Q67 after 15 min exposure, and (c) the hormetic effect of NaBF4 (CAS RN.13755-29-8) on Vibro-qinghaiensis sp. Q67 after 12 h exposure. The blue and red regions indicate the hormetic and inhibition regions of the dependence structure, respectively, as indicated by toxicological experts. The green hashed line indicates the region boundary, detected by *CIM* algorithm.



(a) (b) (c)

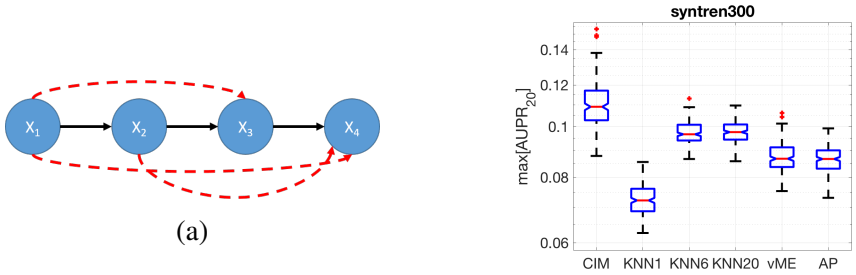
Figure 4.17: (a) **OMIM** data from Fig. 4.16a, interpolated with noise to provide more data-points for modeling purposes. (b) Pseudo-Observations of a Gaussian copula model of data in (a). The red highlighted region represents pseudo-observations which are incorrectly modeled by the Gaussian copula model. (c) Pseudo-Observations of an empirical copula model of data in (a).

Stochastic Modeling

To highlight the importance of nonmonotonic dependence structures from a stochastic modeling perspective, we examine nonmonotonic dose response data from Zhu et al. [2013]. The data are displayed in Fig. 4.16, and regions of importance of the relationship between the data as labeled by scientific experts in the field of toxicology is highlighted in the blue and pink regions. Additionally, the unique ability of *CIM* to automatically identify these regions is shown by the hashed green line. The regions detected by *CIM* correspond to where the monotonicity changes in the dependence structure.

To understand why regions of monotonicity are important from a data modeling perspective, we take the **OMIM** data from Fig. 4.16a and show the difference between modeling it with a Gaussian copula and an empirical copula. Fig. 4.17b shows pseudo-observations drawn from a Gaussian copula model of the data displayed in Fig. 4.16a, which are used to estimate the empirical copula model shown in Fig. 4.17c. The red highlighted region represents pseudo-observations that are incorrectly modeled by the Gaussian copula model. This problem will in fact occur with any popular copula model, including copulas from the Archimedean family, due to the fact that the latter only capture monotonic dependence structures.

We conclude by recognizing that although generalizations about all datasets cannot be drawn from these findings, it is prudent to understand the details of the dataset being analyzed. More specifically, in the context of assessing dependency and modeling joint behavior probabilistically, the simulations conducted show the importance of understanding whether dependencies are monotonic or not.



(a) (b)

Figure 4.18: (a) The true Markov Chain $X_1 \rightarrow X_2 \rightarrow X_3 \rightarrow X_4$; because *CIM* satisfies DPI, indirect interactions represented by the red arrows will be removed as edges in the network discovery algorithm for both ARACNe and MRNET. (b) Results MRNET applied to SYNTREN300 dataset with a global noise level of 10 and a local noise level of 5, using *CIM* and various estimators of mutual information (*MI*), for 200 Monte-Carlo simulations. The median performance of *CIM* exceeds the next best estimator of *MI*, the *KNN20* by 11.77%, which corresponds to accurate detection of 55 more edges from the true network.

Markov Network Modeling

An immediate implication of *CIM* satisfying the DPI, from Section 4.3.2, is that it can be used for network modeling and information flow of data through Markov chains. This is done by performing repeated Max-Relevance Min-Redundancy (*mRMR*) feature selection [Peng et al., 2005] for each variable in the dataset and construct a Maximum Relevancy Network (MRNET) [Meyer et al., 2007]. In principle, for a random variable $X_j \in X$, *mRMR* works by ranking a set of predictor variables $X_{S_j} \subseteq \{X \setminus X_j\}$ according to the difference between the mutual information (*MI*) of $X_i \in X_{S_j}$ with X_j (the relevance) and the average *MI* with the selected variables in X_{S_j} (the redundancy). By choosing the variable that maximizes this difference, a network can be constructed in which direct interactions between variables imply edges. By virtue of Theorem 4, *CIM* and average *CIM* can be substituted for the *MI* and the average *MI* to apply *CIM* to the MRNET reconstruction. As for Fig. 4.18a, using the MRNET algorithm and Theorem 4, we can readily say that

$$CIM(X_1, X_2) - 0.5[CIM(X_1, X_3) + CIM(X_1, X_4)] \geq$$
$$CIM(X_1, X_3) - 0.5[CIM(X_1, X_2) + CIM(X_1, X_4)],$$

and

$$CIM(X_1, X_2) - 0.5[CIM(X_1, X_3) + CIM(X_1, X_4)] \geq$$
$$CIM(X_1, X_4) - 0.5[CIM(X_1, X_2) + CIM(X_1, X_3)],$$

yielding the connection between X_1 and X_2 in Fig. 4.18a. Similar reasoning can be applied to the other network connections. Simulation results discussed in Section 4.4.1 motivate the use of *CIM* as a substitute for the *MI*. In that section, we compare the statistical power of *CIM* to various estimators of the *MI* including: 1) k-nearest neighbors (*k-NN*) estimation [Kraskov et al., 2004], 2) adaptive partitioning (*AP*) MI estimation [Darbellay and Vajda, 1999], and 3) MI estimation via von Mises expansion (*vME*) [Kandasamy et al., 2015], and show that *CIM* is more powerful. This suggests that *CIM* is indeed a viable alternative for use in the estimation of Markov networks from datasets.

We explore the utility of *CIM* for Markov network modeling in the domain of computational biology by using the *MRNET* algorithm with *CIM* and the *MI* estimators previously described using the gene regulatory network benchmarking tool **netbenchmark**. That tool uses over 50 datasets of known gene regulatory networks and compares the performance of a provided algorithm when different amounts of noise are added to the datasets in order to assess in a standardized way, the performance of *MI* based network reconstruction algorithms [Bellot et al., 2015]. The datasets used by **netbenchmark** are different than the gene expression datasets we previously analyzed for monotonicity. The area under the precision-recall curve of the 20 most confident predictions

(AUPR20) is shown for *MRNET* in Fig. 4.18b using CIM and the various estimators of the *MI*, for a global noise level of 10 and a local noise level of 5. The results reveal that for the 200 different variations of the **syntren300** dataset that were compared, the median performance of the *MRNET* is greater when using *CIM* by 11.77%, which corresponds to accurate detection of 55 more edges from the true network. Although not shown here, for a sweep of both global and local noise levels between 10 and 50, *CIM* consistently showed greater performance. On average, *CIM* was able to discover 5.18% more edges over these noise ranges, which corresponds to 24 more edges in the **syntren300** network. These results are not surprising, and are corroborated by the analysis and the curves displayed in Fig. 4.12.

4.5 Conclusion

In this chapter, we have introduced a new statistic of dependence between discrete, hybrid, and continuous random variables and stochastic signals termed *CIM*. We showed that this index follows Rényi's properties for a metric of dependence, satisfies the DPI, and is self-equitable. The implications of satisfying the DPI are discussed in the context of the Markov network construction using the DPI measures. *CIM* is then compared to other measures of mutual information and state-of-the-art nonparametric measures of dependence. It is shown to compare favorably and similarly to these compared metrics, respectively, in various synthetic data experiments. Table 4.2 summarizes the various dependence measures' properties that were discussed in this chapter.

A unique output of *CIM* estimation algorithm, the identification of the regions of monotonicity in the dependence structure, is used to analyze numerous real world datasets. The results reveal that among all the datasets compared, at least 96% of the statistically significant dependencies are indeed monotonic. The simulations highlight the need to fully understand the dependence structure before applying statistical techniques.

Metric	DPI	Rényi	Equitability	Non-Linear	RV Type	$\theta \in [0,1]$	Power
ρ_p	✗	✗	✗	✗	C	✓	●
ρ_s,τ	✗	✗	✗	✓	C	✓	●
Ccor	✓	✗	✓	✓	C	✓	●
CoS	?	?	✗	✓	C	✓	●
dCor	?	✗	✗	✓	CDH	✓	●
RDC	?	✓	✗	✓	CDH	✓	●
MIC	✗	✗	✓	✓	CDH	✓	●
CIM	✓	✓	✗	✓	CDH	✓	●
AP	✓	✗	✗	✓	CDH	✗	●
vME	✓	✗	✗	✓	CDH	✗	●
kNN	✓	✗	✗	✓	CDH	✗	●

Table 4.2: Summary of various dependence measures' properties

While *CIM* is a powerful tool for bivariate data analysis, there are many directions to further this research. A logical first step is to extend *CIM* to a measure of multivariate dependence. To accomplish this, it is important to understand how monotonicity is defined in the multivariate Euclidean space, and how to detect these regions of monotonicity, whether that be directly through a scanning algorithm as was proposed in Algorithm 1 or other means such as projection methods. After algorithms are developed to efficiently and accurately detect these regions, multivariate measures of Kendall's τ such as the proposal made by Joe [1990] can be substituted to create the multivariate version of *CIM*. Additional research can be conducted to improve the performance of *CIM* algorithm for monotonic dependencies, as this is an important class of dependencies. Another area of research is to extend *CIM* to a measure of conditional dependence. By the invariance property of copulas to strictly increasing transforms Embrechts et al. [2001], we can readily state that if $\{\mathbf{Y} \perp\!\!\!\perp \mathbf{Z}\}|\mathbf{X}$, then $\{\mathbf{V} \perp\!\!\!\perp \mathbf{W}\}|\mathbf{U}$, where $\mathbf{U} = (U_1, \ldots, U_d) = (F_{X_1}(x_1), \ldots, F_{X_d}(x_d))$, $\mathbf{V} = (V_1, \ldots, V_k) = (F_{Y_1}(y_1), \ldots, F_{Y_k}(y_k))$, and $\mathbf{W} = (W_1, \ldots, W_n) = (F_{Z_1}(z_1), \ldots, F_{Z_k}(z_n))$, and $\mathbf{X}, \mathbf{Y}$, and $\mathbf{Z}$ are random vectors of arbitrary dimensionality. Due to the invariance property, conditional independence (and dependence) can be measured with the pseudo-observations

by borrowing techniques from partial correlation. Initial results have shown promising results for this application of *CIM*.

Additionally, a more computationally efficient algorithm to count the number of overlapping points for the hybrid case in Algorithm 2 could be researched. From a copula theory and Markov chain perspective, the instability of the W copula, the stability of the Π copula, and the "apex" nature of the M copula as discussed in Appendix C should be further investigated. One area of research here is to understand the speed of convergence to maximum entropy, represented by the Π copula, when the Markov chain is time-homogeneous and non time-homogeneous Additionally, it is well known that the W copula is not a valid copula for $D \geq 3$, where D is the dimensionality of the copula. This seems to correspond loosely to the n-fold Markov product of W corresponding to either W or M depending on whether n is even or odd, and this link should be further investigated.

Another area of research is to extend *CIM* to a measure of conditional dependence. By the invariance property of copulas to strictly increasing transforms Embrechts et al. [2001], we can readily state that if $\{\mathbf{Y} \perp\!\!\!\perp \mathbf{Z}\}|\mathbf{X}$, then $\{\mathbf{V} \perp\!\!\!\perp \mathbf{W}\}|\mathbf{U}$, where $\mathbf{U} = (U_1, \ldots, U_d) = (F_{X_1}(x_1), \ldots, F_{X_d}(x_d))$, $\mathbf{V} = (V_1, \ldots, V_k) = (F_{Y_1}(y_1), \ldots, F_{Y_k}(y_k))$, and $\mathbf{W} = (W_1, \ldots, W_n) = (F_{Z_1}(z_1), \ldots, F_{Z_k}(z_n))$, and $\mathbf{X}, \mathbf{Y}$, and $\mathbf{Z}$ are random vectors of arbitrary dimensionality. Due to the invariance property, conditional independence (and dependence) can be measured with the pseudo-observations.

Chapter 5

The Estimator Response Curve

In this chapter, we discuss how to assess the performance of an estimator of the strength of statistical association, when the goal of measuring the strenght of association between random variables is to rank order many different pairs of random variables. By rank ordering, we simply mean that we would like to order pairwise dependencies according to the strength of association with the output. The goal of rank ordering the strength of association between pairs of random variables is important in machine learning, where features may need to be selected based on which features are most associative with the output to be predicted.

To assess the estimator's performance, we introduce a new property of estimators of the strength of statistical association, which helps characterize how well an estimator will perform in scenarios where dependencies between continuous and discrete random variables need to be rank ordered. The new property, termed the estimator response curve, is easily computable and provides a marginal distribution agnostic way to assess an estimator's performance. It overcomes notable drawbacks of current metrics of assessment, including statistical power, bias, and consistency. We utilize the estimator response curve to test various measures of the strength of association that satisfy the data processing inequality (DPI), and show that the *CIM* estimator's performance com-

pares favorably to *kNN*, *vME*, *AP*, and H_{MI} estimators of mutual information. The estimators which were identified to be suboptimal, according to the estimator response curve, perform worse than the more optimal estimators when tested with real-world data from four different areas of science, all with varying dimensionalities and sizes.

This chapter is organized as follows. We begin by discussing the scenarios where estimators of the strength of dependence fall short. We then discuss the concept of the estimator response curve. We apply the estimator response curve to various measures of the strength of association. We then show the effect of suboptimal estimation of mutual information on feature selection and classification performance. We focus on the scenario where the strength of association needs to be measured between noisy *i.i.d* continuous and discrete random variables (henceforth referred to as hybrid random variables) that are skewed, where the number of unique outcomes of the discrete random variable are small and the dependence structures are nonlinear. This case represents an important subset of problems in machine learning, where real world datasets that often have nonlinear associations between them with skewed marginal distributions, need to be classified according to provided output labels. Additionally, we restrict ourselves to only compare estimators of the strength of statistical association that are proven to satisfy the data processing inequality (DPI); that is, all estimators of mutual information (*kNN*,*vME*,*AP*,H_{MI}) and the index *CIM*. Measures of association that are proven to satisfy the DPI are preferred in machine learning due to the relationship between the DPI and Markov chains Kinney and Atwal [2014]. Furthermore, the DPI assumption is implicit in many machine learning algorithms which utilize measures of the strength of association, such as the maximum-relevance minimum-redundancy (MRMR) algorithm for Markov network discovery and feature selection Margolin et al. [2006], Peng et al. [2005]. We then show real-world examples where we empirically show the utility of the estimator response curve.

5.1 Where do Measures of Dependence Fall Short?

When measuring the strength of association between continuous and discrete random variables, most of the estimators previously mentioned fall short. In general, it becomes more difficult to measure association between continuous and discrete random variables as the number of unique discrete outcomes decreases Genest and Nešlehová [2007]. The case of measuring the strength of association between hybrid random variables, however, is extremely important in machine learning. From classification, clustering, and feature selection perspectives, features are typically amenable to be modeled as continuous random variables, while outcomes or clusters are better modeled as discrete random variables.

Each estimator class above presents a different opportunity for why the hybrid random variable scenario is difficult for estimation. The correlation coefficient, ρ, is actually the standardized linear regression coefficient between random variables X and Y. If Y takes on a small number of unique outcomes, the MMSE objective for solving the regression coefficient does not properly capture the dynamics of the data, and in fact violates an implicit assumption of linear regression, that the dependent variable, Y be continuous. This is illustrated in Fig. 5.1. In it, the independent random variable, X, and the dependent random variable, Y are perfectly associated; the rule

$$Y = \begin{cases} 0, & \text{for } x \leq 500 \\ 1, & \text{for } x > 500 \end{cases}$$

describes the functional relationship in Fig. 5.1. However, the correlation coefficient is 0.86.

As for the Maximal Information Coefficient, *MIC* and other mutual information based methods such as *kNN*, *AP*, and H_{MI}, discretization of the continuous random variable is required in order

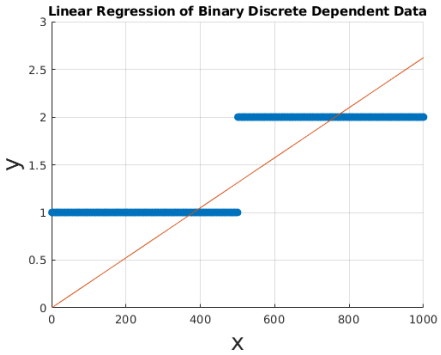


Figure 5.1: Linear Regression of continuous independent variable, X, and discrete dependent variable Y with only two unique outcomes. Here, X and Y are perfectly associated, but the correlation coefficient is computed to be 0.86.

to apply formulas such as (4.4) and

$$I(X,Y) = H(Y) - H(Y|X), \tag{5.1}$$

where

$$H(X) = -\int_{-\infty}^{\infty} p(x)\mathbf{log}[p(x)]dx. \tag{5.2}$$

However, discretization of a random variable cannot be performed optimally without taking into account the end goals of the discretized data, and in addition, there is information loss in the discretization process García et al. [2013].

The copula based estimators mentioned above are also not immune to the hybrid scenario. When modeling discrete random variables with copulas, Sklar's theorem does not guarantee the unicity of the copula C and many copulas satisfy (2.1) due to ties in the data Genest and Nešlehová [2007]. This ambiguity is the reason why copula based estimators also have difficulty measuring association between continuous and discrete random variables. The exception to this is *CIM*, which is based on a bias-corrected measure of concordance, τ_{KL} that can account for continuous and discrete data simultaneously. This is explained in further detail in Section 2.2. of Karra and

Mili's manuscript: *Copula Index for Detecting Dependence and Monotonicity between Stochastic Signals* Karra and Mili [2018a].

5.2 Estimator Response Curve

In this section, we describe an easily measurable property of an estimator, the estimator response curve, that is important in determining its performance, especially in the hybrid random variables scenario. We begin by describing some previously developed metrics of estimator performance, and show why these previously developed metrics do not fully capture the performance of an estimator. We then discuss the estimator response curve, and show why it is important.

There are several properties of estimators of the strength of association which are important in characterizing its performance. Theoretically provable properties include Rényi's seven properties of a dependence measure and the data processing inequality (DPI) for dependence measures Rényi [1959], Kinney and Atwal [2014]. Briefly, Rényi's seven properties of a measure of dependence are important because they establish the space over which the estimator can be used, the exchangeability of the random variables under consideration, and the ranges of possible outputs of the estimators. Similarly, DPI is important for an estimator because it ensures that when measuring the strength of association between random variables in a causal chain, that indirect causes measure weaker than more direct causes Kinney and Atwal [2014]. It is interesting to note that most estimators of mutual information, such as *kNN*, *vME*, *AP*, and H_{MI} do not satisfy Rényi's seven properties. Conversely, most indices of dependence, including *dCor*, *Ccor*, *CoS*, and *RDC* are not proven to satisfy the DPI. The notable exception here is the *CIM*, which is proven to satisfy both Rényi's properties and the DPI Karra and Mili [2018a].

Important empirical properties of an estimator include the statistical power, bias, and consistency. Statistical power is the likelihood that the estimator measures an association between the random

variables, when the random variables are statistically dependent. Usually, the power of an estimator is characterized across a range of dependencies and noise levels to fully assess the estimator's ability to detect different types of dependencies Karra and Mili [2018a]. The importance of power, especially under linear dependence, was originally outlined in Simon and Tibshirani's work Simon and Tibshirani [2014]. Statistical bias is the difference between the true value and the estimated value of the quantity to be measured, and can be defined mathematically as $\text{Bias}_\theta[\hat{\theta}] = E_{x|\theta}[\hat{\theta}] - \theta$, where θ is the true value of the quantity to be estimated, $\hat{\theta}$ is the estimated value, and $E_{x|\theta}$ is the expected value over the conditional distribution $P(x|\theta)$. However, bias is typically only computed under the scenario of independence, where it is known that the value of the θ should be 0. Finally, statistical consistency measures the asymptotic properties of the estimator; an estimator is said to be consistent if the estimated value approaches the true value of the estimator as the sample size grows. Stated mathematically, an estimator of T_n of θ is said to be consistent if $\underset{n\to\infty}{\text{plim}} T_n = \theta$.

While these three empirical properties are important to assess an estimator's performance, none of them capture the notion of the rate of change of an estimated value, as the strength of dependence between the random variables changes. If the rate of increase (or decrease) of an estimated value is not proportional to the rate of increase (or decrease) in the dependence strength between the random variables, then in noisy small sample scenarios, there is a nonzero likelihood that the estimator will incorrectly rank the strength of associations between features and output classes in supervised learning. This becomes especially important in the hybrid random variable scenario, where it is already more difficult to measure the strength of association between two random variables Genest and Nešlehová [2007]. These rates of increase determine the estimators ability to distinguish stronger from weaker relationships, when both relationships are statistically significant. We term the relationship between the actual strength of association between the random variables, X and Y, and the estimated strength of association between X and Y over the entire range of possible strengths of statistical association to be the response curve of an estimator. The response curve

can help explain how an estimator will perform when multiple strengths' of associations need to be measured and ranked, as in mutual information based Markov network discovery and feature selection Margolin et al. [2006], Peng et al. [2005].

An ideal estimator would increase (or decrease) its estimate of the strength of association between random variables X and Y by $\hat{\Delta}$, due to a corresponding increase (or decrease) of Δ of the strength of association between X and Y, across the full range of possible dependence between random variables. If it is desirable to more accurately distinguish stronger dependencies than weaker ones, the ideal response of an estimator across the full range of possible dependencies is a monotonically increasing convex function, with the degree of convexity directly proportional to an increased ability of the estimator to distinguish stronger dependencies apart. This scenario corresponds to $\hat{\Delta} > \Delta$ when the strength of association is high. Conversely, if it is desirable to more accurately distinguish weaker dependencies than stronger ones, the ideal response of an estimator across the full range of possible dependencies is a monotonically increasing concave function, with the degree of concavity directly proportional to an increased ability of the estimator to distinguish weaker dependencies apart. This scenario corresponds to $\hat{\Delta} < \Delta$ when the strength of association is high. The special case of $\hat{\Delta} = \Delta$ is ideal, where the estimator is able to distinguish all dependence types equally well. However, even if an estimator has this kind of response curve, its variance must be low to have a high likelihood that dependencies will be correctly ranked.

Various response curves are shown in Fig. 5.2. The linear response is shown in purple; in it, the estimator attempts to distinguish between all strengths of dependence equally, while in the convex curves shown with o markings in green and blue, stronger dependencies are have a higher likelihood of being ranked correctly. Conversely, in the concave response curves denoted with marks in teal and yellow, the estimator has a higher likelihood of ranking weaker dependencies correctly. The curve is scale-invariant, because it examines the rates of change of an estimator, rather than absolute values. It also shows that nonlinear rescaling of an estimators output may affect

its ability to correctly rank strengths of association. An example of this is Linfoot's informational coefficient of correlation Linfoot [1957]. Here, the mutual information between random variables X and Y is rescaled according to the the relationship

$$r(X,Y) = \sqrt{1 - e^{-2I(X,Y)}},$$

where $I(X,Y)$ is the mutual information. Depending on the variance of the estimator (explained in further detail below), this nonlinear scaling could have an adverse affect on ranking the strengths of association.

The curves in Fig. 5.2 also show the variance of the estimated quantity as a function of the strength of dependence. The variance, along with the concavity/convexity of the estimator determines the probability of correctly ranking dependencies between different pairs of random variables. More specifically, the probability of correctly ranking two different pairs of random variables according to their strength of association is inversely proportional to the area encompassed by the rectangle covering the space between the maximum possible value the estimator can take on for the weaker dependency, denoted by $\hat{\theta}^{\text{max}}_{\text{weaker}}$, and the minimum possible value the estimator can take on for the stronger dependency, denoted by $\hat{\theta}^{\text{min}}_{\text{stronger}}$, if $\hat{\theta}^{\text{max}}_{\text{weaker}} > \hat{\theta}^{\text{min}}_{\text{stronger}}$. For example, in Fig. 5.2, suppose that the true value of the strength of association between X_1 and Y is 0.6, and the true value of the strength of association between X_2 and Y is 0.8, and our goal is to rank them according to their strengths, using an estimator. If the estimator had a response curve similar to the blue curve, then the probability of misranking these dependencies is zero here, because $\hat{\theta}^{\text{max}}_{\text{weaker}}$, denoted by the green $\times$ symbol is less than $\hat{\theta}^{\text{min}}_{\text{stronger}}$, denoted by the green four pointed star. Conversely, if the estimator had a response curve similar to the yellow curve, then $\hat{\theta}^{\text{max}}_{\text{weaker}}$ denoted by the red $\times$ symbol is greater than $\hat{\theta}^{\text{min}}_{\text{stronger}}$, denoted by the red four pointed star. The probability of misranking these dependencies is nonzero and is proportional to the area given by the area in the red shaded rectangle in Fig. 5.2. These probabilities do not need to be exactly computed, but identifying them helps to understand

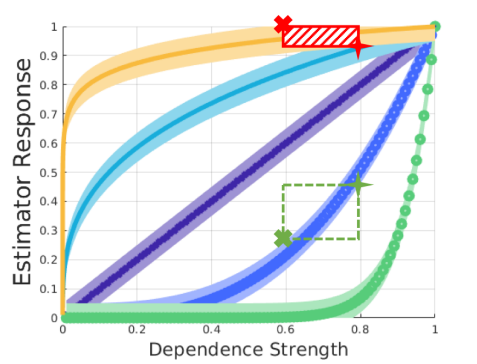


Figure 5.2: Illustration of the various estimator response curves. The linear response is shown in purple; in it, the estimator attempts to distinguish between all strengths of dependence equally, while in the convex curves shown with o markings in green and blue, stronger dependencies are have a higher likelihood of being ranked correctly. Conversely, in the concave response curves denoted with marks in teal and yellow, the estimator has a higher likelihood of ranking weaker dependencies correctly. The red shaded rectangle shows a region of non-zero probability that an estimator with the yellow response curve would have in misranking dependence strengths between two pairs of random variables having strengths of association to be 0.6 and 0.8, respectively. The green hollow rectangle shows the region of zero probability that an estimator with the blue response curve would have in misranking dependence strengths between two pairs of random variables having strengths of association to be 0.6 and 0.8, respectively.

how an estimator may perform for these applications. In summary, the estimator response curve is a quick, marginal distribution invariant method to assess how the estimator will perform under various dependencies.

5.3 Synthetic Simulations

In this section, we detail simulations conducted in order to estimator response curve of the aforementioned estimators of the strength of association which satisfy the DPI constraint. To accomplish this, we use the copula framework which allows us to generate various linear and non-linear dependency structures between random variables, while being agnostic of the marginal distributions.

In our synthetic data simulations, we generate data according to the procedure outlined in Nelsen

for generating dependent random variables with arbitrary dependence structures Nelsen [2006]. We begin by generating data from the Gaussian, Frank, Gumbel, and Clayton copulas. These copulas represent different kinds of dependence structures, with the Gaussian modeling linear dependence, and the remaining copulas modeling non-linear dependence patterns such as tail dependence. Each of the copulas has a single parameter, θ, which controls the strength of dependence, and corresponds directly to the mutual information contained within the dependence structure. Because the scale and support of θ varies between different copula families, our simulations modulate the strength of dependence between the random variables through the rank correlation measure, τ, which has a one-to-one correspondence to θ for every copula that was simulated. After picking a certain copula family and a value of τ, we generate random variates u and v. After generating these random variates u and v, we apply the inverse transform $F^{-1}(U) = X$ and $G^{-1}(V) = Y$ respectively to generate x and y. Here, we choose three different cases for X and Y to simulate real-world scenarios which may arise.

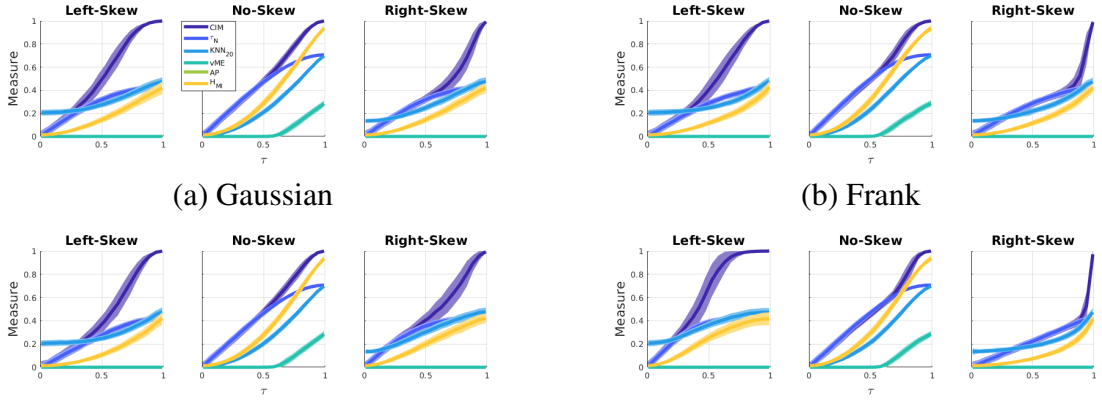
Connecting back to the machine learning perspective of continuous explanatory features, and discrete outcomes, we choose X to be a continuous random variable and Y to be a discrete random variable. The three scenarios considered are when X and Y are both skewed left, not skewed, and both skewed right. In the left skew situation, we choose X to be a Pearson distribution with mean of zero, standard deviation of one, and a skew of negative one. In the right skew situation, we choose X to be a Pearson distribution with mean of zero, standard deviation of one, and a skew of positive one. In the no skew situation, we choose X to be a Normal distribution with mean of zero and standard deviation of one. Similarly, in the left skew situation, we choose Y to be a discrete distribution, with a probability mass function taking on the vector $[0.9, 0.1]$. In the right skew situation, the probability mass function of Y is given by the vector $[0.1, 0.9]$. Finally, in the no skew situation, the probability mass function of Y is given by $[0.5, 0.5]$. In all these scenarios, the cardinality of the support set of Y is two. This corresponds to the binary classification scenario

discussed previously.

The results for these simulations, which are the response curves described in Fig. 5.2 for these estimators with continuous and discrete marginal distributions, are shown in Fig. 5.3. In them, the x-axis represents the strength of dependence, given by τ, and the y-axis represents the strength of dependence as measured by the various estimators. It can be seen for all scenarios tested, the state-of-the-art estimators *kNN*, *AP*, and *vME* all exhibit suboptimal estimator response curves. More concerningly, they exhibit less sensitivity when $\tau \geq 0.5$, but do not enhance sensitivity with weaker dependencies as would be hoped for from a more concave estimator response curve. In other words, as the strength of association between the random variables increases, the ability of these estimators to distinguish between them decreases in the hybrid random variable scenario! For the no-skew scenario, only the H_{MI} estimator seems to perform equivalently to the *CIM* estimator. This suggests that the *CIM* estimator should be used when measuring the strength of association between hybrid random variables.

5.4 Real World Data Simulations

In this section, we show how suboptimal estimation of mutual information affects feature selection and classification performance. To accomplish this, we take four real world datasets provided by the NIPS 2003 feature selection challenge Guyon et al. [2005], and apply the MRMR algorithm to select the most relevant features. The chosen datasets, Arcene, Dexter, Madelon, and Gisette, are binary classification datasets, where the input features are continuous variables, and the output is a discrete variable that can take on two values. We chose these datasets because from the perspective of measuring association between random variables, this presents the most "difficult" case. From a practicality perspective, this case is also highly relevant to machine learning problems, where predictive features are often continuous but the output class to be predicted has only a small



(a) Gaussian (b) Frank

(c) Gumbel (d) Clayton

Figure 5.3: Response curves for *kNN*,*vME*,*AP*,H_{MI}, and *CIM* for Skewed Hybrid Data. The x-axis shows the strength of dependence, captured by the rank correlation coefficient Kendall's τ, and the y-axis shows the estimated strength of dependence. Subplots titled Left-Skew have a continuous independent variable, distributed according to the Pearson distribution with a mean of zero, standard deviation of one, and a skew of negative one. No-Skew distributions have a continuous independent variable distributed according to a standard normal distribution, while right-skew distributions have a continuous independent variable distributed according to a Pearson distribution with a mean of zero, standard deviation of one, and a skew of positive one. Similarly, for the left-skew scenario, the dependent variable is a discrete random variable with a probability mass function (PMF) following the vector $[0.9, 0.1]$. The No-Skew dependent distribution is a discrete distribution following a PMF of $[0.5, 0.5]$, and the right-skew dependent distribution is a discrete distribution with PMF of $[0.1, 0.9]$

number of unique outcomes. Additionally, they represent datasets of various sample sizes and are from different fields in science. The Arcene dataset is 100 samples and contains 10000 features representing mass-spectrometric data from cancerous and non-cancerous cells. The Dexter dataset has 300 samples, and contains 20000 bag of words features for text classification. The Madelon dataset has 2000 samples and is an artificial dataset with 500 features to classify clusters in a five-dimensional hypercube. Finally, the Gisette dataset is 3000 samples of 5000 features representing text classification features to distinguish between the digits 4 and 9.

With these datasets, we perform feature selection with the maximum relevance, minimum redundancy (MRMR) feature selection algorithm Peng et al. [2005]. Briefly, MRMR is an approach to feature selection that utilizes mutual information to assess the relevance and redundancy of a set of features, given an output prediction class. The goal of MRMR is to solve

$$\arg\max_{|S|=k} I(X_S, Y),$$

where $X_S = \{X_i : i \in S\}$, k is the number of features to be selected, and $I(X, Y)$ is the mutual information between the random variables X and Y. However, measuring the mutual information of increasingly large dimensions of k is unfeasible due to the curse of dimensionality. MRMR attempts to overcome this by solving the optimization problem given by

$$\Phi(X_S, Y) = \frac{1}{|S|} \sum_{i \in S} I(X_i, Y) - \frac{1}{|S|^2} \sum_{i,j \in S} I(X_i, X_j). \tag{5.3}$$

To maximize this objective, the most important feature (the feature which has the maximum mutual information with the output) is chosen first. Then, additional features are added inductively using

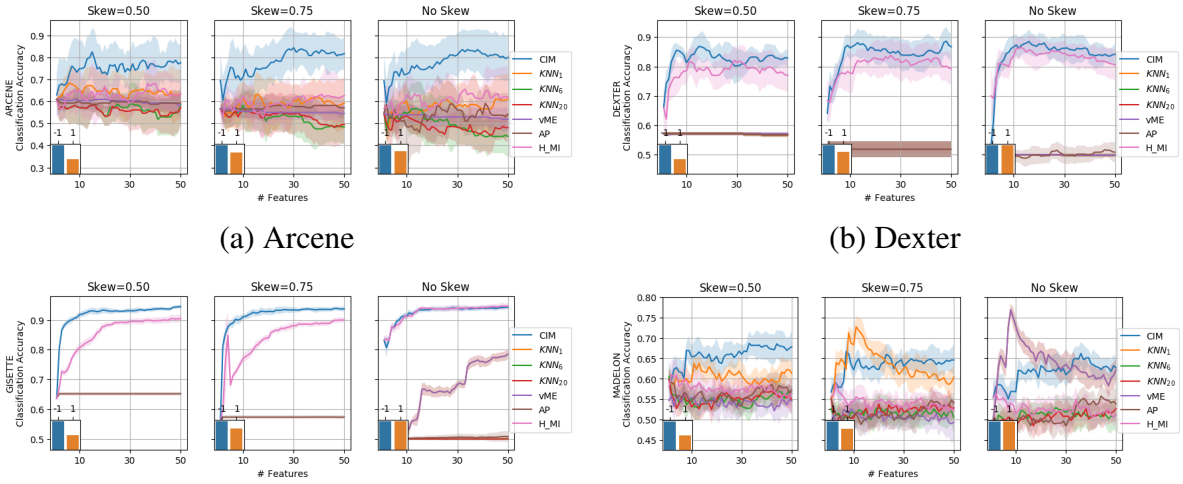
the function

$$\underset{X_j \in X \setminus S_m}{\arg\max} I(X_j, Y) - \frac{1}{m-1} \sum_{X_i \in S_m} I(X_i, X_j). \tag{5.4}$$

The first term in (5.4) represents the relevance of feature X_j to output Y, and the second term represents the redundancy between the selected features X_i and the current feature under consideration, X_j. Because the MRMR algorithm is based on measuring mutual information between input features (continuous or discrete) and the output class (typically discrete with small cardinality), our goal in these experiments is to understand how suboptimal estimation of mutual information affects MRMR. It is readily seen from (5.3) and (5.4) that more accurate estimation of mutual information should yield better feature selection results.

To test this hypothesis, we compare features selected by MRMR using different estimators of mutual information for the four datasets described above. To assess the performance of the feature selection, we apply classification algorithms on the selected features; higher classification performance implies a better estimator of mutual information because the same classification and feature selection algorithms are used across all tests. The estimators compared are *kNN-1*, *kNN-6*, *kNN-20*, *vME*, *AP*, *CIM*, and H_{MI}; these are chosen because they are proven to satisfy the DPI assumption required by MRMR. Using the selected features for each estimator, we then apply the k-nearest neighbors classification algorithm and score the classification performance using only the selected features on a validation dataset. This process is repeated when different amounts of data from the positive class are dropped, creating skewed output class distributions.

The results for these experiments are shown in Fig. 5.4. For each dataset, we show the 10-fold cross validation score of a *kNN* classifier as we increase the number of features that were selected, in order of importance as provided by the MRMR algorithm for each DPI satisfying estimator. We show the results for each dataset, where we skew the number of positive examples to be 50%, 75%,



(a) Arcene

(b) Dexter

(c) Gisette

(d) Madelon

Figure 5.4: Real World Data results for Feature Selection. For each dataset, we show the results of feature selection and subsequent classification by a *kNN* classification algorithm, for the output class balances shown in the inset plot. In the inset plot, the blue bar represents the number of negative class examples used for feature selection, and the orange bar represents the number of positive class examples used for feature selection.

and 100% of the number of negative examples. The output class distribution for each simulation is shown in the inset plot in Fig. 5.4. It is seen that for three of the four datasets tested, the *CIM* estimator compares favorably to all other estimators of mutual information considered. The results corroborate the findings in Section 5.3, where it was seen that with synthetic test vectors, the *CIM* estimator compares favorably to other DPI satisfying measures of the strength of association for hybrid random variables.

5.5 Conclusion

In this chapter, we have introduced a new concept for evaluating the performance of an estimator of the strength of association between random variables, the estimator response curve. We show that existing empirical properties for measuring the performance of an estimator are inadequate, and

explain how the estimator response curve fills this gap. We then explain a copula based methodology for measuring the response curve of an estimator, and apply this methodology to estimate the response curves of various estimators of the strength of association which satisfy the DPI criterion. Comparing the estimator response curves, we see that the *CIM* estimator performs best across the board in the hybrid random variable scenario, where data may be skewed. We then test these various estimators with real world data. The simulations show that the estimator response curves are a good indicator of how an estimator may perform in a scenario where the strengths of associations need to be ranked, as in feature selection and classification.

Chapter 6

Conclusions

In this dissertation, we have built off the pioneering work of Elidan, and attempted to fill a gap in copula research when applied to machine learning. More specifically, we have begun to address the issue of using copula based models with hybrid random variables. As stated in the previous chapters, hybrid random variables are an important case in machine learning, which often deals with both continuous and discrete data directly. We looked at three aspects of hybrid data: 1) modeling with Hybrid Copula Bayesian Networks, 2) analysis of hybrid data with *CIM*, and 3) assessing the performance of an estimator of the strength of association when applied to hybrid data using the copula framework. Although the work above highlights the importance of modeling and analyzing hybrid data, much work remains to be done in this realm.

The main thurst of improvement that can be made on all the work completed to date is to make it more scalable. Efficient, scalable structure learning and inference algorithms need to be developed in order to make the models viable. From a *CIM* perspective, parallelized algorithms that can work on chunks of the data, rather than having to load the entire dataset into memory, can prove useful in big data scenarios. These improvements to the current state-of-the-art have the potential to improve data analysis and machine learning in the upcoming age of data.

Bibliography

M. Ahsanullah, V. Nevzorov, and M. Shakil. *An Introduction to Order Statistics*, volume 3. Atlantis Press, 2013.

T. Bedford and R. Cooke. Vines–a new graphical model for dependent random variables. *The Annals of Statistics*, 30(4):1031–1068, 08 2002. doi: 10.1214/aos/1031689016.

P. Bellot, C. Olsen, P. Salembier, A. Oliveras-Vergés, and P. Meyer. Netbenchmark: a bioconductor package for reproducible benchmarks of gene regulatory network inference. *BMC Bioinformatics*, 16(1):312, 2015. doi: 10.1186/s12859-015-0728-4.

M. Ben Hassine, L. Mili, and K. Karra. A Copula Statistic for Measuring Nonlinear Multivariate Dependence, 2016.

D. Bonett and T. Wright. Sample Size Requirements for Estimating Pearson, Kendall and Spearman Correlations. *Psychometrika*, 65(1):23–28, 2000.

Y. Chang, Y. Li, A. Ding, and J. Dy. A Robust-Equitable Copula Dependence Measure for Feature Selection. *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, 2016.

A. Charpentier, J.-D. Fermanian, and O. Scaillet. The Estimation of Copulas: Theory and Practice. *Copulas: from theory to applications in finance*, pages 35–62, 2007.

R. Chotimodum, T. Santiwipanont, and S. Sumetkijakan. Asymptotic Dependence of Markov Chains joined by a Patched Fréchet Copula. *The Annual Pure and Applied Mathematics Conference*, 2014. doi: 10.13140/2.1.3638.5921.

T. Cover and J. Thomas. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, 2006.

G. Darbellay and I. Vajda. Estimation of the Information by an Adaptive Partitioning of the Observation Space. *IEEE Transactions on Information Theory*, 45(4):1315–1321, May 1999. doi: 10.1109/18.761290.

W. Darsow, B. Nguyen, and E. Olsen. Copulas and Markov Processes. *Illinois J. Math.*, 36(4): 600–642, 12 1992.

A. de Leon and B. Wu. Copula-based Regression Models for a Bivariate Mixed Discrete and Continuous Outcome. *Statistics in Medicine*, 2011.

P. Deheuvels. La fonction de dépendance empirique et ses propriétés. Académie Royale de Belgique. *Bulletin de la Classe des Sciences*, 65(5):274–292, 1979.

M. Denuit and P. Lambert. Constraints on Concordance Measures in Bivariate Discrete Data. *Journal of Multivariate Analysis*, 93(1):40 – 57, 2005. doi: http://dx.doi.org/10.1016/j.jmva.2004.01.004.

G. Elidan. Copula Bayesian Networks. In *Advances in Neural Information Processing Systems 23*. Curran Associates, Inc., 2010.

G. Elidan. Copula Network Classifiers. In *15th International Conference on Artificial Intelligence and Statistics*. JMLR Workshop and Conference Proceedings, 2012.

G. Elidan. Copulas in machine learning. In *Copulae in mathematical and quantitative finance*, pages 39–60. Springer, 2013.

P. Embrechts, L. F., and M. A. Modelling Dependence with Copulas and Applications to Risk Management, 2001.

S. García, J. Luengo, S. J., L. V., and H. F. A Survey of Discretization Techniques: Taxonomy and Empirical Analysis in Supervised Learning. *IEEE Transactions on Knowledge and Data Engineering*, 25:734–750, 2013.

C. Genest and J. Nešlehová. A Primer on Copulas for Count Data. *ASTIN Bulletin*, 2007.

I. Guyon, S. Gunn, A. Ben-Hur, and G. Dror. Result analysis of the nips 2003 feature selection challenge. In *Advances in neural information processing systems*, pages 545–552, 2005.

D. Helsel and R. Hirsch. *Statistical Methods in Water Resources Techniques of Water Resources Investigations, Book 4, chapter A3*. U.S. Geological Survey, 2002. URL `http://pubs.water.usgs.gov/twri4a3`.

H. Joe. Multivariate concordance. *Journal of Multivariate Analysis*, 35(1):12 – 30, 1990. doi: http://dx.doi.org/10.1016/0047-259X(90)90013-8.

K. Kandasamy, A. Krishnamurthy, B. Poczos, L. Wasserman, and J. Robins. Nonparametric von Mises Estimators for Entropies, Divergences and Mutual Informations. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 397–405. Curran Associates, Inc., 2015.

K. Karra and L. Mili. Hybrid Copula Bayesian Networks. *Probabilistic Graphical Models 2016 - JMLR: Workshop and Conference Proceedings*, 2016.

K. Karra and L. Mili. Copula Index for Detecting Strength and Monotonicity between Stochastic Signals, 2018a.

K. Karra and L. Mili. On the Effect of Suboptimal Estimation of Mutual Information in Feature Selection and Classification, 2018b.

M. Kendall. A New Measure of Rank Correlation. *Biometrika*, 30(1-2):81–93, 1938.

M. Kendall. The Treatment of Ties in Ranking Problems. *Biometrika*, 33(3):239–251, 1945.

J. Kinney and G. Atwal. Equitability, Mutual Information, and the Maximal Information Coefficient. *Proceedings of the National Academy of Sciences*, 111(9):3354–3359, 2014.

A. Kolesarova, R. Mesiar, J. Mordelova, and C. Sempi. Discrete Copulas. *Fuzzy Systems, IEEE Transactions on*, 2006.

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

A. Kraskov, H. Stögbauer, and P. Grassberger. Estimating Mutual Information. *Phys. Rev. E*, 69, 2004. doi: 10.1103/PhysRevE.69.066138.

A. Lagerås. Copulas for Markovian Dependence. *Bernoulli*, 16(2):331–342, 05 2010.

M. Lichman. UCI machine learning repository - census+income dataset, 2013. URL `http://archive.ics.uci.edu/ml/datasets/Census+Income`.

E. Linfoot. An informational measure of correlation. *Information and Control*, 1957.

D. Lopez-Paz, P. Henning, and B. Schölkopf. The Randomized Dependence Coefficient. In *Advances in Neural Information Processing Systems 26*. Curran Associates, Inc., 2013.

L. Madsen and D. Birkes. Simulating dependent discrete data. *Journal of Statistical Computation and Simulation*, 83(4):677–691, 2013. URL `http://dx.doi.org/10.1080/00949655.2011.632774`.

A. Margolin, I. Nemenman, K. Basso, C. Wiggins, G. Stolovitzky, R. Favera, and A. Califano. ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context. *BMC Bioinformatics*, 7(1):S7, 2006.

M. Mesfioui and J.-F. Quessy. Concordance Measures for Multivariate Non-Continuous Random Vectors. *Journal of Multivariate Analysis*, 2010.

P. Meyer, K. Kontos, F. Lafitte, and G. Bontempi. Information-Theoretic Inference of Large Transcriptional Regulatory Networks. *EURASIP Journal on Bioinformatics and Systems Biology*, 2007.

S. Moral, R. Rumi, and A. Salmern. Mixtures of Truncated Exponentials in Hybrid Bayesian Networks. In *Symbolic and Quantitative Approaches to Reasoning with Uncertainty*. Springer Berlin Heidelberg, 2001.

R. Nelsen. *An Introduction to Copulas*. Springer-Verlag New York, 2006.

J. Nešlehová. On Rank Correlation Measures for Non-Continuous Random Variables. *Journal of Multivariate Analysis*, 2007.

A. Panagiotelis, C. Czado, and H. Joe. Pair Copula Constructions for Multivariate Discrete Data. *Journal of the American Statistical Association*, 107(499):1063–1072, 2012.

K. Pearson. Note on Regression and Inheritance in the Case of Two Parents. *Proceedings of the Royal Society of London*, 58:240–242, 1895.

H. Peng, F. Long, and C. Ding. Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27(8):1226–1238, 2005.

A. Rényi. On Measures of Dependence. *Acta Mathematica Academiae Scientiarum Hungarica*, 10(3):441–451, 1959.

D. Reshef, Y. Reshef, H. Finucane, S. Grossman, G. McVean, P. Turnbaugh, E. Lander, M. Mitzenmacher, and P. Sabeti. Detecting Novel Associations in Large Data Sets. *Science*, 334(6062): 1518–1524, 2011.

Y. Reshef, D. Reshef, P. Sabeti, and M. Mitzenmacher. Equitability, interval estimation, and statistical power. *arXiv preprint arXiv:1505.02212*, 2015.

M. Rey and V. Roth. Copula mixture model for dependency-seeking clustering. In *Proceedings of The 29th International Conference on Machine Learning*, 2012.

M. Scarsini. On Measures of Concordance. *Stochastica*, 8(3):201–218, 1984.

U. Schmock. *On the Asymptotic Behaviour of the Estimator of Kendalls Tau*. PhD dissertation, T.U. Munich, 2010.

B. Schweizer and A. Sklar. On nonparametric measures of dependence for random variables. *Studia Mathematica*, 1974.

N. Simon and R. Tibshirani. Comment on "detecting novel associations in large data sets" by reshef et al, science dec 16, 2011, 2014.

M. Sklar. *Fonctions de répartition à n dimensions et leurs marges*. Université Paris 8, 1959.

M. S. Smith and M. A. Khaled. Estimation of Copula Models with Discrete Margins via Bayesian Data Augmentation. *Journal of the American Statistical Association*, 2012.

C. Spearman. The Proof and Measurement of Association between Two Things. *The American Journal of Psychology*, 15(1):72–101, 1904.

Z. Szabó. Information Theoretical Estimators Toolbox. *Journal of Machine Learning Research*, 15:283–287, 2014.

G. Székely, M. Rizzo, and N. Bakirov. Measuring and Testing Dependence by Correlation of Distances. *The Annals of Statistics*, 35(6):2769–2794, 12 2007.

F. Vandenhende and P. Lambert. Improved Rank-Based Dependence Measures for Categorical Data. *Statistics and Probability Letters*, 63(2):157–163, 2003.

J. Yang, S. Cheng, and L. Zhang. Bivariate Copula Decomposition in Terms of Comonotonicity, Countermonotonicity and Independence. *Insurance: Mathematics and Economics*, 39(2):267 – 284, 2006. doi: http://dx.doi.org/10.1016/j.insmatheco.2006.02.015.

Y. Zheng, J. Yang, and J. Huang. Approximation of Bivariate Copulas by Patched Bivariate Fréchet Copulas. *Insurance: Mathematics and Economics*, 48(2):246 – 256, 2011. doi: http://dx.doi.org/10.1016/j.insmatheco.2010.11.002.

X.-W. Zhu, S.-S. Liu, L.-T. Qin, F. Chen, and H.-L. Liu. Modeling non-monotonic doseresponse relationships: Model evaluation and hormetic quantities exploration. *Ecotoxicology and Environmental Safety*, 89:130 – 136, 2013.

Appendices

Appendix A

Proof of Proposition 1

Proof As in Nešlehová [2007], let $w \in [0,1]$ and set

$$x(w) := F_X^{(-1)}(w+) \qquad u(w) = \begin{cases} 1 & \text{if } P[X = x(w)] = 0 \\ \frac{w - P[X < x(w)]}{P[X = x(w)]} & \text{else} \end{cases},$$

where $F_X^{(-1)}(x+)$ denotes the right hand side limit of the generalized inverse of F_X. For $(w_1, \ldots, w_n) \in [0,1]^n$ one can construct points $(x_1(w_1), \ldots, x_n(w_n))$ such that

$$\begin{aligned} &P[\Psi(X_1,U_1) \leq w_1, \ldots, \Psi(X_n,U_n) \leq w_n] \\ &\quad = \sum_{b \in S} C_{\mathbf{U}}(u_1(w_1), \ldots, u_n(w_n)) \times \\ &\quad P[X_1 \square x_1(w_1), \ldots, X_n \square x_n(w_n)], \end{aligned} \tag{A.1}$$

where S is the set of all possible combinations of binary vectors of length n,

$$u_i(w_i) = \begin{cases} 1 & \text{if } S_b(i) = 1 \\ u_i(w_i) & \text{else} \end{cases}, \qquad \square = \begin{cases} = & \text{if } S_b(i) = 1 \\ < & \text{else} . \end{cases},$$

$\square$ stands for the mathematical symbol defined above, and $S_b(i)$ is the i^{th} element of the b^{th} binary vector in the set S. Now suppose that $(w_1, \ldots, w_n) \in RANGE\{F_1\} \times \cdots \times RANGE\{F_n\}$. As in Nešlehová [2007, Proof of Proposition 4], we have that $x_i(w_i) := F_{X_i}^{(-1)}(w_i)$ or $x_i > F_{X_i}^{(-1)}(w_i)$, which implies that

$$u_i(w_i) = \begin{cases} 0 & \text{if } P[X_i = x_i(w_i)] > 0 \\ 1 & \text{if } P[X_i = x_i(w_i)] = 0. \end{cases}$$

Substituting the appropriate values of $u_i(w_i)$ depending on $P[X_i = x_i(w_i)$, we arrive at

$$\begin{aligned} C_{\Psi(\mathbf{X},\mathbf{U})}(w_1, \ldots, w_n) &= P[\psi(X_1, U_1), \ldots, \psi(X_n, U_n)] \\ &= P[X_1 \leq F_{X_1}^{(-1)}(w_1), \ldots, X_1 \leq F_{X_n}^{(-1)}(w_n)] \\ &= C_{\mathbf{X}}(w_1, \ldots, w_n). \end{aligned}$$

■

Appendix B

Proof of Theorem 3

Proof Let X and Y be two random variables such that X and Y are associated through the mapping $g(\cdot)$. Additionally, let Ω be the range-space of $(X, Y) \in \mathcal{R}^2$. Partition the sample space Ω into subspaces Ω_i, where each Ω_i corresponds to a monotonic section of the range space of (X, Y) . The partition of the sample space implies that $\bigcup_i \Omega_i = \Omega$, $\Omega_i \cap \Omega_j \in \emptyset \; \forall i \neq j$. Define the random variable $X_i = X(\Omega_i)$. Additionally, define

$$\forall i, \qquad g_i(x) = \begin{cases} g(x) & \forall x \in X_i \\ 0 & \text{otherwise} \end{cases}$$

$\therefore g(x) = \sum_{i=1}^{n} g_i(x)$. We can then write

$$\begin{aligned} F_Y(y) &= P(Y \leq y) \\ &= P(g(X) \leq y) && \textit{(by substitution)} \\ &= \sum_i P(g_i(X_i) \leq y) P(X_i|X) && \textit{(by Law of Total Probability)} \end{aligned}$$

Additionally,

$$P(g_i(X_i) \leq y) = P(X_i \leq g_i^{-1}(y))$$

$$= \begin{cases} F_{X_i}(g_i^{-1}(y)) & \text{if } g_i^{-1}(y) \text{ is increasing} \\ 1 - F_{X_i}(g_i^{-1}(y)) & \text{if } g_i^{-1}(y) \text{ is decreasing} \\ \begin{cases} 1 & \text{if } x_i \geq K_i \\ 0 & \text{else} \end{cases} & \text{if } X_i = K_i \end{cases}$$

where K_i is a constant.

$$g_i^{-1}(y) = g_i^{-1}(g(x)) = x_i \implies F_{X_i}(g_i^{-1}(y)) = F_{X_i}(x_i)$$

where $x_i \in X_i$. From the definition of conditional probability, we have that $P(g_i(X_i) \leq y) = \frac{F_X(x)}{P(X_i|X)} \quad \forall x \in X_i$ due to $P(X_i \cap X_j) = 0$. Because $F_{X_i}(x_i)$ and $F_X(x)$ are proportional to each other by the constant $P(X_i|X)$, and $F_{X_i}(x_i) = F_{X_i}(g_i^{-1}(y)) = F_Y(y)$, $F_Y(y)$ is a piecewise linear function of $F_X(x)$.

■

Appendix C

Proof of Theorem 4

Proof

We begin by proving that although τ does not satisfy DPI, $|\tau|$ does. To show this, we exploit the relationship between copulas and Markov chains initially presented by Darsow et al. [1992], and the ability to express any copula as a convex combination of its Fréchet-Hoeffding bounds on patches of the unit-square [Yang et al., 2006, Zheng et al., 2011]. Another crucial property of copulas that we rely on for this proof comes from the 5^{th} property of concordance, which states that if $C_1 \preceq C_2$, then $\kappa(C_1) \leq \kappa(C_2)$, where κ is a measure of concordance [Scarsini, 1984]. Using these three elements, we show that $C_{XY} \succeq C_{XZ}$ when $C_{XZ} \succeq \Pi$, which completes the proof that $|\tau|$ satisfies the DPI. Note that considering the set of copulas $\{C : C \succeq \Pi\}$ is equivalent to computing the absolute value of any measure of concordance. This is due to the fact that if a copula C_1 has a negative concordance τ_1, (i.e. $C_1 \preceq \Pi$), then we can define $C_2(u, v) = C_1(v, u)$, where $\tau_2 = |\tau_1|$, implying that $C_2 \succeq \Pi$.

From Darsow et al. [1992], if X, Y, and Z follow a Markov Chain $X \to Y \to Z$, then $C_{XZ}(u, v) = C_{XY}(u, v) * C_{YZ}(u, v) = \int_0^1 \frac{\partial C_{XY}(u,t)}{\partial t} \frac{\partial C_{YZ}(t,v)}{\partial t} dt$. Using this, we would like to show that $|\tau(C_{XZ})| \leq$

$|\tau(C_{XY})|$. From the 5^{th} property of concordance, this is equivalent to showing that $C_{XZ} \preceq C_{XY}$, with the condition that $C_{XY}, C_{XZ} \succeq \Pi$. From Zheng et al. [2011], we can decompose a copula C as the convex combination of the Fréchet-Hoeffding lower and upper bounds, W and M respectively, and the independence copula Π, on patches of the unit-square. Decomposing C_{YZ}, we have

$$C_{YZ}(u,v) = \sum_{i=0}^{m-1}\sum_{j=0}^{m-1} p_{ij}\left[\alpha_{ij}M^{ij}(u,v) + \beta_{ij}\Pi^{ij}(u,v) + \gamma_{ij}W^{ij}(u,v)\right]$$

$$\implies \frac{\partial C_{YZ}(t,v)}{\partial t} = \sum_{i=0}^{m-1}\sum_{j=0}^{m-1} p_{ij}\left[\alpha_{ij}\frac{\partial M^{ij}(t,v)}{\partial t} + \beta_{ij}\frac{\partial \Pi^{ij}(t,v)}{\partial t} + \gamma_{ij}\frac{\partial W^{ij}(t,v)}{\partial t}\right]$$

where $\alpha_{ij} + \beta_{ij} + \gamma_{ij} = 1$ and $\sum_{i=0}^{m-1}\sum_{j=0}^{m-1} p_{ij} = 1$. Substituting, we get

$$\begin{aligned}
C_{XZ}(u,v) = C_{XY} * C_{YZ}(u,v) &= \int_0^1 \frac{\partial C_{XY}(u,t)}{\partial t}\frac{\partial C_{YZ}(t,v)}{\partial t}dt \\
&= \sum_{i=0}^{m-1}\sum_{j=0}^{m-1}\left[\alpha_{ij}\int_0^1 \frac{\partial C_{XY}^{ij}(u,t)}{\partial t}\frac{\partial M^{ij}(t,v)}{\partial t}dt + \beta_{ij}\int_0^1 \frac{\partial C_{XY}^{ij}(u,t)}{\partial t}\frac{\partial \Pi^{ij}(t,v)}{\partial t}dt\right. \\
&\left. + \gamma_{ij}\int_0^1 \frac{\partial C_{XY}^{ij}(u,t)}{\partial t}\frac{\partial W(t,v)}{\partial t}dt\right] \\
&= \sum_{i=0}^{m-1}\sum_{j=0}^{m-1}\left[\alpha_{ij}[C_{XY}^{ij} * M^{ij}] + \beta_{ij}[C_{XY}^{ij} * \Pi^{ij}] + \gamma_{ij}[C_{XY}^{ij} * W^{ij}]\right]
\end{aligned}$$

From Darsow et al. [1992], we have that $C*M = C, C*\Pi = \Pi$, and $C*W(u,v) = u - C(u, 1-v)$. From here on, for clarity of exposition, we omit the indices ij and the notation below is implied for each patch. Substituting these relations, we get

$$C_{XZ} = \alpha C_{XY} + \beta\Pi + \gamma[u - C_{XY}(u, 1-v)]$$

Due to the 2-increasing property of copulas and $\alpha + \beta + \gamma = 1$, we can say that for each patch, $C_{XY} \succeq C_{XY} \implies C_{XY} \succeq \alpha C_{XY}$, and $C_{XY} \succeq \Pi \implies C_{XY} \succeq \beta\Pi$. Additionally, by assumption, we have

$$\begin{aligned}
C_{XY} \succeq \Pi &\implies C_{XY}(u,v) \geq \Pi(u,v)\ \forall u, v \in \mathbf{I} \\
&\implies C_{XY}(u, 1-v) \geq \Pi(u, 1-v) \\
&\implies C_{XY}(u, 1-v) \geq u - \Pi(u,v) \\
&\implies C_{XY}(u, 1-v) \geq u + (-\Pi(u,v)) \\
&\implies C_{XY}(u, 1-v) \geq u + (-C_{XY}(u,v)) \\
&(\textit{because } C_{XY}(u,v) \geq \Pi(u,v)\ \textit{implies} - C_{XY}(u,v) \leq -\Pi(u,v)) \\
&\implies C_{XY}(u,v) \geq u - C_{XY}(u, 1-v) \\
&\implies C_{XY} \succeq C_{XY} * W \\
&\implies C_{XY} \succeq \gamma(C_{XY} * W)
\end{aligned}$$

Thus, $C_{XY} \succeq C_{XZ}$ with the constraint that $C_{XZ} \succeq \Pi$ for every patch, and because we have a convex combination of patches and increasing the number of patches, m, decreases the approximation error to an arbitrarily small amount, it follows that $|\tau|$ satisfies DPI. We now use this result to show that *CIM* satisfies the DPI.

As seen in (4.3), *CIM* constructs a copula for each region of concordance and discordance, and computes the absolute value of Kendall's τ for each of these regions. Thus, to show that *CIM* satisfies DPI, we first show that for random variables X, Y, and Z that follow a Markov Chain

$X \to Y \to Z$, if the domain of X is subdivided into disjoint sets X_i, then the associated Y_i and Z_i random variables (according to the Markov Chain, X is the cause of Y and Z, and hence Y_i and Z_i are associated with X_i) also satisfy DPI.

Define subsets of X as X_i, such that $\bigcup_i X_i \in X$ and $X_i \cap X_j \in \emptyset \forall\ i \neq j$, then

$$f_{X_i,Y_i}(x_i, y_i) = \frac{f_{XY}(x,y)}{f_{X_i|X}(x_i|x)}$$
$$f_{X_i,Z_i}(x_i, z_i) = \frac{f_{XZ}(x,z)}{f_{X_i|X}(x_i|x)}$$

Recall that because X, Y, and Z satisfy DPI, the relation

$$\int_Y \int_X f_{XY}(x,y)\mathbf{log}\left(\frac{f_{XY}(x,y)}{f_X(x)f_Y(y)}\right)dxdy \geq \int_Z \int_X f_{XZ}(x,z)\mathbf{log}\left(\frac{f_{XZ}(x,z)}{f_X(x)f_Z(z)}\right)dxdy$$

holds. Additionally, $f_{X_i|X}(x_i|x)$ is a constant. Hence, the following must hold:

$$\int_Y \int_X \frac{f_{XY}(x,y)}{f_{X_i|X}(x_i|x)}\mathbf{log}\left(\frac{f_{XY}(x,y)}{f_{X_i|X}(x_i|x)f_X(x)f_Y(y)}\right)dxdy \geq$$
$$\int_Z \int_X \frac{f_{XZ}(x,z)}{f_{X_i|X}(x_i|x)}\mathbf{log}\left(\frac{f_{XZ}(x,z)}{f_{X_i|X}(x_i|x)f_X(x)f_Z(z)}\right)dxdz$$
$$\implies \int_{Y_i} \int_{X_i} f_{X_iY_i}(x_i,y_i)\mathbf{log}\left(\frac{f_{X_iY_i}(x_i,y_i)}{f_{X_i}(x_i)f_{Y_i}(y_i)}\right)dx_idy_i \geq$$
$$\int_{Z_i} \int_{X_i} f_{X_iZ_i}(x_i,z_i)\mathbf{log}\left(\frac{f_{X_iZ_i}(x_i,z_i)}{f_{X_i}(x_i)f_{Z_i}(z_i)}\right)dx_idz_i$$
$$\implies X_i \to Y_i \to Z_i$$
$$\implies C_{X_i,Y_i} \succeq C_{X_i,Z_i}$$
$$\implies |\tau(X_i,Y_i)| \geq |\tau(X_i,Z_i)|$$

From the above, because X_i, Y_i, and Z_i is shown to satisfy the DPI, we can say that

$$\sum_i w_i |\tau(X_i, Y_i)| \geq \sum_i w_i |\tau(X_i, Z_i)|$$
$$\implies CIM(X, Y) \geq CIM(X, Z)$$

because $\sum_i w_i = 1$. ■

We briefly note some interesting observations about the relationship between Markov Chains and copulas. From Lagerås [2010], it is known that if the Markov chain $X_0 \to X_1 \to \cdots \to X_n$ is time-homegenous and C is the copula for (X_0, X_1), then the copula for (X_0, X_n) is given by C^{*n}, which is defined as the n-fold Markov product of C with itself. It is then shown in Chotimodum et al. [2014] that if $\alpha + \gamma < 1$ in the decomposition specified above, $\lim_{n\to\infty} C^{*n} = \Pi$. We have shown above that even in the non time-homogenous case, that the copula between variables that are further apart in the Markov chain are "smaller" than copulas closer together. Returning to the time-homogenous case, it is interesting to also note that the W is an unstable copula, as $\lim_{n\to\infty} W^{*n} = W$ if n is odd, and $\lim_{n\to\infty} W^{*n} = M$ if n is even. Finally, M is somewhat of an "apex" copula, and the only way to maintain this perfect dependence is to compute the Markov product with M.

Appendix D

Proof of Theorem 5

Proof With $\alpha \to 0, n \to \infty$, from (4.9), *CIM* detection criterion given by (4.8) reduces to

$$|\tau_{KL}| < |\tau'_{KL}|. \tag{D.1}$$

Under the assumption that the noise distribution is stationary over the data being analyzed, in the limit as $n \to \infty$, if points belong to the same region, then $|\tau_{KL}| \geq |\tau'_{KL}|$, and $|\tau_{KL}| < |\tau'_{KL}|$ if newly added points belong to a different region. Thus, as $n \to \infty$, the region detection criterion given by (4.8) will detect any region boundary with probability of 1.

■

Appendix E

Rényi's Properties

CIM satisfies the first property since it operates on copula transformed data (pseudo-observations, which exist for any random variable) rather than the raw data. Because of the following two identities: $\Sigma_i w_i = 1$, $\mathbf{min}(|\tau_N^i|) = 0$ and $\mathbf{max}(|\tau_N^i|) = 1$, the value of *CIM* given by (4.3) takes values between 0 and 1, and thus the third property is satisfied. In the independence case, because there are no regions of concordance or discordance, (4.3) reduces to $|\tau_N| = 0$. From Scarsini [1984], any measure of concordance is equal to zero when X and Y are independent; because *CIM* reduces to $|\tau_N|$, which is an absolute value of the concordance measure τ for independent random variables, we can state that if $X \perp\!\!\!\perp Y$, then $CIM = 0$. Conversely, the proof given below, shows that if $CIM = 0$, then $X \perp\!\!\!\perp Y$. Thus, the fourth property is satisfied. The fifth property is also satisfied because Kendall's τ is invariant to increasing or decreasing transforms [see Nelsen, 2006, Theorem 5.1.8], so the convex sum of Kendall's τ must also be invariant to increasing or decreasing transforms. Note that we cannot guarantee that τ_N be invariant to bijective transforms because the scaling in (4.2) depends upon the marginal distributions. Thus, the fifth property is only valid for *CIM* for continuous random variables. We later introduce the CIM^S in (4.5), which satisfies the fifth property for discrete and hybrid random variables also. The second and sixth properties are

satisfied by virtue of Theorem 3. Finally, the seventh property is satisfied because ***CIM*** metric is the absolute value of Kendall's τ for a Gaussian copula and can be converted to the correlation coefficient θ, with the relation $\theta = \sin(\frac{CIM\pi}{2})$. This works because the Gaussian copula captures monotonic linear dependence, and hence there is only one region.

Proof From (4.3), it is seen that CIM can take on the value of 0 if and only if $\tau_N^i = 0 \; \forall i$, and from Definition 9 of Nešlehová [2007] that τ_N can only take a value of 0 if the $4 \int C_{XY}^S dC_{XY}^S - 1 = 0$. Thus, our proof will show that under the constraint of monotonicity, if $4 \int C_{XY}^S dC_{XY}^S - 1 = 0$, then $C_{XY}^S = \Pi$.

We begin by recognizing that from Nešlehová [2007], because C_{XY}^S follows the same concordance properties as C_{XY}, we show the proof for C_{XY} for clarity of exposition. Recall that CIM divides the dependency into regions of monotonicity, where in each region, the dependency is either comonotonic or countermonotonic. Mathematically, we can write this as: $C_{XY}^i \geq \Pi$ or $C_{XY}^i \leq \Pi \; \forall u, v \in \mathbb{I}^2$ and i represents the i^{th} region. From the properties of concordance ordering Nelsen [2006], it is known that if $C_1 \succ C_2$, then $\kappa(C_1) \geq \kappa(C_2)$. Because $\tau(C_{XY}^i, C_{XY}^i) = 0 \iff C_{XY}^i = \Pi$ under the condition that either $C_{XY}^i(u, v) \geq \Pi(u, v) \; \forall u, v \in \mathbb{I}^2$ or $C_{XY}^i(u, v) \leq \Pi(u, v) \; \forall u, v \in \mathbb{I}^2$, then $C_{XY}^i = \Pi$.

■

Appendix F

CIM estimation Algorithm

Algorithm 1 *CIM*

function COMPUTE-CIM(msi, α)

  si $\leftarrow [1, \frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \ldots, \frac{1}{msi}]$ ▷ Scanning increments to be tested

  $\mathbf{uv_{cfg}} \leftarrow$ [u-v, v-u] ▷ Orientations of data to be tested

  $m_{max} \leftarrow 0, \mathbf{RR} \leftarrow []$

  for uv_{cfg} in $\mathbf{uv_{cfg}}$ **do**

    for si in **si** **do**

      $\boldsymbol{\tau}, \mathbf{R} \leftarrow$ SCAN-UNIT-SQ(si, uv_{cfg}, α)

      $m \leftarrow 0$

      for all $\boldsymbol{\tau}, \mathbf{R}$ **do** ▷ Compute (4.3) for detected regions

        $n_R \leftarrow$ GETNUMPOINTS$(\mathbf{R})$ [1]

        $m \leftarrow m + \frac{n_R}{n}\tau_R$

      end for

      if $m > m_{max}$ **then** ▷ Maximize (4.3) over all scanning increments

        $m_{max} \leftarrow m$

[1] gets the number of points encompassed by the region $\mathbf{R}$

$RR \leftarrow R$

end if

end for

end for

return $m_{max}, \mathbf{RR}$

end function

function SCAN-UNIT-SQ(si, uv_{cfg}, α)

$\mathbf{R} \leftarrow$ CREATENEWREGION [2]

$\mathbf{RR} \leftarrow []$

while uniqSqNotCovered **do** [3]

$\mathbf{R} \leftarrow$ EXPANDREGION(si, uv_{cfg}) [4]

$m \leftarrow |\hat{\tau}_{KL}(\mathbf{R})|$ ▷ $|\hat{\tau}_{KL}|$ of the points encompassed by $\mathbf{R}$

$n_R \leftarrow$ GETNUMPOINTS$(\mathbf{R})$

$\sigma_C \leftarrow 4(1 - \hat{\tau}_{KL}(\mathbf{R})^2)$ ▷ Hypothesis test detection threshold

if $\neg$NEWREGION$(\mathbf{R})$ **then** [5]

if $m < (m_{prev} - \frac{\sigma_C}{\sqrt{n_R}} u_{1-\frac{\alpha}{2}})$ **then**

$\mathbf{RR} \leftarrow$ STOREREGION$(\mathbf{R})$ [6]

$\boldsymbol{m} \leftarrow m$

$\mathbf{R} \leftarrow$ CREATENEWREGION(uv_{cfg})

end if

end if

$m_{prev} \leftarrow m$

[2] creates a new region of monotonicity from the boundary where the previous region was determined to end

[3] a variable which flags when the expansion of **R** is covering the entire unit square.

[4] expands the region by the scanning increment amount, si, as depicted in Fig. 4.6 in the orientation specified by the uu_{cfg}

[5] determines if the region **R** was created in the last loop iteration or not

[6] called when a boundary between regions is detected; stores the region **R**'s boundaries and the value of $|\tau_{KL}|$ for this region.

37: **end while**

38: **return** $m, \mathbf{RR}$

39: **end function**

Appendix G

Streaming τ_{KL} Algorithm

Algorithm 2 τ_{KL}^{S}

function CONSUME

$ii_{end} \leftarrow ii_{end} + 1$

$mm \leftarrow mm + 1$ ▷ Increment number of samples, m, we have processed

$mmc2 \leftarrow mmc2 + mm - 1$ ▷ Increment running value of $\binom{m}{2}$ for denominator

▷ Get the subset of **u** and **v**

$\mathbf{u}' \leftarrow \mathbf{u}(ii_{begin} : ii_{end})$, $\mathbf{v}' \leftarrow \mathbf{v}(ii_{begin} : ii_{end})$

▷ Compute ordering of new sample, in relation to processed samples

$\Delta\mathbf{u} \leftarrow \mathbf{u}'(end) - \mathbf{u}'(end - 1 : -1 : 1)$

$\Delta\mathbf{v} \leftarrow \mathbf{v}'(end) - \mathbf{v}'(end - 1 : -1 : 1)$

$u^+ \leftarrow \Sigma\left[\mathbb{1}(\Delta\mathbf{u} > 0 \cap \Delta\mathbf{v} \neq 0)\right], u^- \leftarrow \Sigma\left[\mathbb{1}(\Delta\mathbf{u} < 0 \cap \Delta\mathbf{v} \neq 0)\right]$

$v^+ \leftarrow \Sigma\left[\mathbb{1}(\Delta\mathbf{v} > 0 \cap \Delta\mathbf{u} \neq 0)\right], v^- \leftarrow \Sigma\left[\mathbb{1}(\Delta\mathbf{v} < 0 \cap \Delta\mathbf{u} \neq 0)\right]$

▷ Compute the running numerator, K, of τ_{KL}

if $u^+ < u^-$ **then**

$kk \leftarrow v^- - v^+$

else

$kk \leftarrow v^{+} - v^{-}$

end if

$K \leftarrow K + kk$

▷ Count number of times values in u and v repeat

uMap$(\mathbf{u}'(end)) \leftarrow$ **uMap**$(\mathbf{u}'(end)) + 1, uu \leftarrow uu +$ **uMap**$(\mathbf{u}'(end)) - 1$

vMap$(\mathbf{v}'(end)) \leftarrow$ **vMap**$(\mathbf{v}'(end)) + 1, vv \leftarrow vv +$ **vMap**$(\mathbf{v}'(end)) - 1$

▷ Compute threshold for determining if data is hybrid via a threshold heuristic

if $\neg \bmod (mm, OOCTZT)$ **then**

$mmG \leftarrow mmG + 1$, $ctzt \leftarrow ctzt + mmG - 1$

end if

$uuCtz \leftarrow (uu \leq ctzt)$, $vvCtz \leftarrow (vv \leq ctzt)$

▷ Compute the denominator of τ_{KL} depending on whether data was hybrid or not

if $(uuCtz \cap vv > 0) \cup (vvCtz \cap uu > 0)$ **then**

$tt \leftarrow$ **max**(uu, vv)

$den \leftarrow \sqrt{mmc2 - tt}\sqrt{mmc2 - tt}$

else

$den \leftarrow \sqrt{mmc2 - uu}\sqrt{mmc2 - vv}$

end if

if $K == 0 \cap den == 0$ **then**

$\tau_{KL} = 0$

else

$\tau_{KL} = \frac{K}{den}$

end ifreturn τ_{KL}

end function

Appendix H

Real World Data Processing Methodology

Real-world data analyzed for the monotonicity results shown above in Table 4.1 was derived from online sources.

Gene Expression Data

The gene expression related data was downloaded from the Broad Institute at the URL:

`http://portals.broadinstitute.org/cgi-bin/cancer/datasets.cgi`

The enumeration below lists the specific files which were downloaded from the URL provided above (all with the .gct extension).

1. ALL
2. beer_lung_for_p53
3. Breast_A
4. Breast_B
5. Children_NE
6. Common_miRNA
7. crash_and_burn
8. DLBCL_A
9. DLBCL_B
10. DLBCL_C
11. DLBCL_D
12. Erythroid

13. GCM_All
14. glioma_classic_hist
15. glioma_nutt_combo
16. hep_japan
17. HL60
18. HSC_FDR002
19. Iressa_Patient1_ams
20. leuGMP
21. leukemia.top1000
22. lung_datasetB_outcome
23. LungA_1000genes
24. met
25. miGCM_218
26. MLL_AF9
27. mLung
28. Multi_A
29. Multi_B
30. Normals_Leu
31. Novartis_BPLC.top1000
32. PDT_miRNA
33. Rap3hour_control
34. Rap24hour_control
35. Res_p0005
36. Sens_p001
37. Sens_p0005
38. medullo_datasetC_outcome
39. lung_annarbor_outcome_only
40. med_macdonald_from_childrens
41. megamiR_data.normalized.log2.th6
42. Myeloid_Screen1_newData_021203_ams.AML_poly_mono
43. Sanger_Cell_Line_project_Affymetrix_QCed_Data_n798

The files, natively in GCT format, were stripped of metadata and converted to CSV files, and scanned for significant dependencies ($\alpha < 0.05$). The number of regions for the significant dependencies were then counted to determine the number of monotonic regions. The script to perform the conversion from GCT to CSV is provided at:

`https://github.com/stochasticresearch/depmeas/tree/master/test/python/gcttocsv.py`

Additionally, the Matlab scripts to process the pairwise dependencies and produce the monotonicity results is provided at:

`https://github.com/stochasticresearch/depmeas/tree/master/test/analyze_cancerdata.m`

Financial Returns Data

The financial returns related data was downloaded from both

`finance.yahoo.com`

and

`Investing.com`

We query the web API of these websites to download all available historical data (from Jan 1985 - Jan 2017) for the following indices:

1. A50
2. AEX
3. AXJO
4. BFX
5. BSESN
6. BVSP
7. CSE
8. DJI
9. FCHI
10. FTSE
11. GDAXI
12. GSPC
13. HSI
14. IBEX
15. ITMIB40
16. IXIC
17. JKSE
18. KOSPI
19. KSE
20. MICEX
21. MXX
22. NK225
23. NSEI
24. OMXC20
25. OMXS
26. PSI20
27. SETI
28. SPTSX
29. SSEC
30. SSMI
31. STOXX50E
32. TA25
33. TRC50
34. TWII
35. US2000
36. XU100

Less than 1% of the downloaded data was missing. In order to ease processing, missing data fields were imputed with the last known index price. The first difference of the stock prices was calculated in order to derive the returns data. The returns data was first determined to be stationary by the Dickey-Fuller test. After these procedures, pairwise dependencies between the time series were computed. Because different amounts of historical data were available for the various indices, only the subset of data which belonged to both time series was tested for a significant dependency.

The script to perform the missing data imputation and raw data normalization is provided at:

`https://github.com/stochasticresearch/depmeas/tree/master/test/python/normalizeStocksFiles.py`

Additionally, the Matlab scripts to process the pairwise dependencies and produce the monotonicity results is provided at:

`https://github.com/stochasticresearch/depmeas/tree/master/test/analyze_stocksdata.m`

Finally, the raw stocks data is provided at

`https://figshare.com/articles/Stocks_Data/4620325`

Climate Data

The climate data was downloaded from the following links:

1. `https://www.kaggle.com/uciml/el-nino-dataset`
2. `https://www.kaggle.com/sogun3/uspollution`
3. `https://tinyurl.com/berkeleyearth`

The El-Nino data was normalized by extracting the zonal winds, meridional winds, humidity, air temperature, and sea surface temperature data from the dataset. The code to extract these features, and all to be described features from other climate related datasets is provided at:

`https://github.com/stochasticresearch/depmeas/tree/master/test/python/normalizeClimateFiles.py`

Because these datapoints were collected over multiple decades and large chunks of missing data existed, each chunk of contiguous data (with respect to time) was analyzed separately. The script to identify these chunks and coherently compute pairwise dependencies after checking for stationarity is provided at:

`https://github.com/stochasticresearch/depmeas/tree/master/test/analyze_elnino.m`

The global land temperatures data was normalized by extracting the land temperature for each country over the available date ranges. Again, due to significant chunks of missing data, each chunk of contiguous data was analyzed separately. The script to identify these chunks and coherently compute pairwise dependencies after checking for stationarity is provided at:

`https://github.com/stochasticresearch/depmeas/tree/master/test/analyze_landtemperatures.m`

US pollution data was normalized by extracting NO_2 Air Quality Indicators (AQI), O_3 AQI, SO_2 AQI, and CO AQI for each location over the available date ranges. Again, due to significant chunks of missing data, each chunk of contiguous data was analyzed separately. The script to identify these chunks and coherently compute pairwise dependencies after checking for stationarity is:

`https://github.com/stochasticresearch/depmeas/tree/master/test/analyze_pollution.m`