Journal of Machine Learning Research 18 (2017) 1-46 Submitted 2/16; Revised 3/17; Published 7/17

A Robust-Equitable Measure for Feature Ranking and Selection

A. Adam Ding A.DING@NEU.EDU
Department of Mathematics
Northeastern University
Boston, MA 02115, USA

Jennifer G. Dy JDY@ECE.NEU.EDU
Department of Electrical and Computer Engineering
Northeastern University
Boston, MA 02115, USA

Yi Li LI.YI3@HUSKY.NEU.EDU
Department of Mathematics
Northeastern University
Boston, MA 02115, USA

Yale Chang YCHANG@COE.NEU.EDU
Department of Electrical and Computer Engineering
Northeastern University
Boston, MA 02115, USA

Editor: Gal Elidan

Abstract

In many applications, not all the features used to represent data samples are important. Often only a few features are relevant for the prediction task. The choice of dependence measures often affect the final result of many feature selection methods. To select features that have complex nonlinear relationships with the response variable, the dependence measure should be *equitable*, a concept proposed by Reshef et al. (2011); that is, the dependence measure treats linear and nonlinear relationships equally. Recently, Kinney and Atwal (2014) gave a mathematical definition of *self-equitability*. In this paper, we introduce a new concept of *robust-equitability* and identify a robust-equitable copula dependence measure, the *robust copula dependence* (RCD) measure. RCD is based on the L_1-distance of the copula density from uniform and we show that it is equitable under both equitability definitions. We also prove theoretically that RCD is much easier to estimate than mutual information. Because of these theoretical properties, the RCD measure has the following advantages compared to existing dependence measures: it is robust to different relationship forms and robust to unequal sample sizes of different features. Experiments on both synthetic and real-world data sets confirm the theoretical analysis, and illustrate the advantage of using the dependence measure RCD for feature selection.

Keywords: dependence measure, feature selection, copula, equitability, mutual information

1. Introduction

The performance of machine learning algorithms is dependent on the input features representing each data sample. Often not all of these features are useful: some may be irrelevant and some may be redundant. Feature selection is thus needed to help improve the performance of learning tasks.

Moreover, feature selection can decrease the computational cost of algorithms, and provide domain experts with an increased understanding of which factors are important.

Feature selection algorithms can be categorized based on how the learning algorithms is incorporated into the selection algorithm: *filter*, *wrapper*, or *embedded* methods (Kohavi and John, 1997; Guyon and Elisseeff, 2003; Yu and Liu, 2004). Filter methods (Kira and Rendell, 1992; Yu and Liu, 2004; Peng et al., 2005; He et al., 2005; Song et al., 2007) pre-select features, without running the learning algorithm. Features are evaluated only through the intrinsic properties of the data. Wrapper methods (Kohavi and John, 1997; Guyon et al., 2002; Dy and Brodley, 2004) "wraps" the search around the learning algorithm and evaluate candidate feature subsets based on learning performance in each candidate feature subset. Embedded methods (Tibshirani, 1996; Vapnik and Vapnik, 1998) incorporate feature search and the learning algorithm into a single optimization problem formulation. Wrapper and embedded methods, contrary to filter methods, select features specific to the learning algorithm; thus, they are most likely to be more accurate than filter methods on a particular learning algorithm, but the features they choose may not be appropriate for other algorithms. Another limitation of wrapper methods is that they are computationally expensive because they need to train and test the learning algorithm for each feature subset candidate, which can be prohibitive when working with high-dimensional data.

Filter methods rely on measures based on intrinsic properties of the data. More specifically, they evaluate features based on some dependence measure criterion between features and the target variable and select the subset of features that optimizes this criterion. Let d be the number of original features. An exhaustive search, which involves 2^d possible feature subsets is computationally impractical. Thus, one commonly employs heuristic search strategies, such as greedy approaches (e.g., sequential forward/backward search (Pudil et al., 1994)). However, these strategies can lead to local optima. Random search methods, such as genetic algorithms, add some randomness to help escape from local optima. When the dimensionality is very high, one can only afford an individual search. Individual search methods (Guyon and Elisseeff, 2003; He et al., 2005) evaluate each feature individually according to a criterion and then select features, which either satisfies a condition or are top-ranked. The problem with individual search methods is that they ignore feature interaction and dependencies. To account for such interactions and dependencies, Yu and Liu (2004) selects relevant features individually and then add a separate redundancy removal step to account for linear correlation between features; Peng et al. (2005) suggests another way, by maximizing relevance and minimizing redundancy (mRMR) together.

In addition to search strategies, the performance of filter methods depend heavily on the choice of dependence measures. The ability to measure the dependence between random variables is a fundamental problem in statistics and machine learning. One of the simplest and most common dependence measure is the Pearson correlation coefficient (ρ_{lin}). However, this measure only captures linear relationships. Another popular measure is mutual information (MI). MI can capture nonlinear dependencies but is difficult to estimate (Fernandes and Gloor, 2010; Reshef et al., 2011) (see Theorem 3 in Section 4). Kernel-based dependence measures (Gretton et al., 2005a; Fukumizu et al., 2007) (e.g., the Hilbert-Schmidt Independence Criterion (HSIC)) have been introduced as an alternative to MI which does not require explicitly learning joint distributions. However, HSIC depends on the choice of kernels. Hilbert-Schmidt Normalized Information Criterion (HSNIC), also known as normalized conditional cross-covariance operator (NOCCO) (Fukumizu et al., 2007), is kernel-free, meaning it does not depend on the choice of kernels in the limit of infinite data. Even though HSNIC is kernel-free, both HSIC and HSNIC's values may vary when we use different

scales. Poczos et al. (2012) applied Maximum Mean Discrepancy (MMD) after empirical copula transformation to make the kernel-based dependence measure invariant to strictly monotone transformation of the marginal variables. The Copula-MMD (CMMD) can also be written in HSIC formulation after empirical copula transformation. Similarly, Reddi and Póczos (2013) applied HSNIC after empirical copula transformation (CHSNIC). Other dependence measures can also be applied after empirical copula transformation, resulting in measures that are also invariant to strictly monotone transformations. However, they (e.g., CMMD and CHSNIC) may fail to treat non-monotonic relations equally.

Reshef et al. (2011) proposed the concept of *equitability*, which states that a dependence measure should give equal importance to all relations: linear and nonlinear. For example, we expect a fair dependence measure to treat a perfectly linear relationship and a perfectly sinusoid relationship equally. Kinney and Atwal (2014) mathematically defined equitability by proposing *self-equitability*—under a nonlinear regression model with additive noise, a dependence measure should be invariant to any deterministic transformation of the marginal variables, under a nonlinear regression model with additive noise (a formal definition is provided in Definition 1, Section 2). A self-equitable dependence measure will treat all forms of relationships equally in the large data limit for the additive noise model. Kinney and Atwal (2014) proved that MI is self-equitable, and recommended its usage.

To choose among the many self-equitable dependence measures, we further propose a new *robust-equitability* concept such that the measure also treats all forms of relationships equally in the mixture noise model. That is, in a mixture distribution with p proportion of deterministic signal hidden in continuous independent background noise, the measure should reflect the signal strength p. The mixture noise model reflects real applications where measurements (features) are often corrupted with noise. For example, sensor data maybe corrupted by noise from hardware and environmental factors. Reshef et al. (2011, 2015b) considered equitability for a statistic. Our robust-equitability, as well as Kinney and Atwal (2014)'s self-equitability, is defined on the population quantity instead. Particularly, in the mixture distribution above, we define a dependence measure as *weakly-robust-equitable* if it is a monotone transformation of the proportion p, and is robust-equitable if it equals to p exactly.

In this paper, we show that among a class of self-equitable copula-based dependence measures, only *robust copula dependence* (RCD), defined as the total variational distance (the half of the L_1 distance) between copula density and uniform (independence) density, is also weakly-robust-equitable (and robust-equitable). Without referring to the copula density, RCD can be equivalently stated as the total variational distance between the probability distribution and the (independent) product of its marginal distributions, and is equivalent to the Silvey's Delta measure (Silvey, 1964). In the literature, the Silvey's Delta (RCD) was only cited as an abstract benchmark. Here, we propose a k-nearest-neighbor (KNN)-based estimator for RCD and prove its consistency. Besides the L_1 distance RCD, we also investigated properties of the L_2 distance between copula density and the uniform density (we call CD_2). CD_2 is the theoretical value of HSNIC in the large data limit (Fukumizu et al., 2007).

In addition, the robust-equitability study in this paper provides insights on the difficulty of estimating MI. Some authors studied the convergence of MI estimators by imposing the Hölder condition on the copula density. This Hölder condition, while being a standard condition for density estimations, does not hold for any commonly used copula (Omelka et al., 2009; Segers, 2012). Under a more realistic Hölder condition on the bounded region of copula density, we provide a the-

oretical proof that the mutual information (MI)'s minimax risk is infinite. This provides a theoretical explanation on the statistical difficulty of estimating MI observed by practitioners (Fernandes and Gloor, 2010; Reshef et al., 2011). Moreover, we prove that although both MI and CD_2 are self-equitable, they are not robust-equitable. Therefore, MI and CD_2 may not rank the features correctly by dependence strength in some cases, even in the large data limit. We confirm this phenomena on both synthetic and real-world data sets. In contrast, RCD is consistently estimable under the same condition. As for kernel-based dependence measures, HSIC and CMMD are neither self-equitable nor robust-equitable, HSNIC and CHSNIC are self-equitable but not robust-equitable and their estimators converge very slowly. Since RCD is the only measure that is both self-equitable and robust-equitable among these measures, it can be very useful for feature selection.

In summary, the contributions of this paper are: (1) the introduction of the concept of robust-equitability; (2) the identification of RCD as a dependence measure that is both self- and robust-equitable and the proposal of a practical consistent estimator for RCD; (3) theoretically proving that non-robust-equitable measures MI and CD_2 cannot be consistently estimated and showing that this can lead to incorrect selection of features when sample size is large or when sample sizes are unequal for different features; and finally, (4) demonstrating that the robust-equitable RCD is a better dependence measure for feature selection compared to existing dependence measures through experiments on synthetic and real-world data sets, in terms of robustness to function types, correctness in large sample size and correctness in unequal sample sizes. This paper is a substantially extended version of our conference version (Chang et al., 2016). In particular, this work includes the following additional materials: (1) a more complete treatment of the motivation and rationale of equitability definitions—we discuss the relationship of equitability to Renyi's theorems, to more copula-based dependence measures, and to independence tests; (2) a more complete theoretical treatment of the difficulty in estimation of mutual information (MI) versus RCD; in particular, we add a theorem showing that the difficulty of MI estimation is not due to the unboundedness of its definition, but is intrinsic due its being non-robust-equitable; and (3) more extensive empirical studies illustrating how equitability helps in feature selection.

The rest of this paper is organized as follows. In Section 2, we motivate the equitability concepts, discussing different equitability definitions and relationship to copula and Renyi's theorems. Particularly, we propose the concept of robust-equitability, and define a robust-equitable dependence measure called robust copula dependence. In Section 3, we prove MI and CD_2 are not consistently estimable. We also prove RCD can be consistently estimated and provide its estimators based on kernel density estimation (KDE) and k-nearest-neighbors (KNN). In Section 4, we provide feature selection experiments on synthetic and real data sets to demonstrate the advantage of RCD compared to existing dependence measures. We end with conclusions and discussions in Section 5.

2. A Robust-Equitable Dependence Measure

In this section, we investigate the theoretical properties of different dependence measures. In particular, we would like the dependence measures to have the following characteristics. We would like the dependence measures to rank a feature with less noise as having a stronger dependence with the response variable compared to features with more noise. We do not want measures that prefer a particular type of relationship (e.g., linear). Moreover, we do not want the measures to be too sensitive to sample size (i.e., when different features have unequal sample sizes, the dependence measure should not prefer a feature simply because it has more samples, but should still rank the features

based on the strength of the deterministic signal compared to noise). Note that it is becoming more common for databases to have features that have unequal sample sizes due to the prevalence of data collection from heterogeneous sources. For example, a clinical database may have more samples with clinical features compared to samples with genomic information. In such as setting, we would like to use all the data available to perform feature selection rather than to create equal sample sizes by throwing away samples from the larger set. In this paper, we formalize these ideas through the recently proposed equitability concept: We want to use dependence measures that reflect the noise level, regardless of relationship type.

2.1 Self-equitability, Rényi's Axioms and Copula-based Dependence Measures

Reshef et al. (2011) proposed that an equitable measure should "give similar scores to equally noisy relationships of different types." Kinney and Atwal (2014) mathematically defined self-equitability through invariance under all nonlinear relationships in the regression model

$$Y = f(X) + \epsilon, \tag{1}$$

where f is a deterministic function, ϵ is the random noise variable whose distribution may depend on $f(X)$ as long as ϵ has no additional dependence on X.

Definition 1 *A dependence measure $D[X;Y]$ is self-equitable if and only if $D[X;Y] = D[f(X);Y]$ whenever f is the function in model (1).*

Kinney and Atwal (2014) recommended usage of a self-equitable measure: mutual information (MI).

To understand self-equitability better, we notice that Definition 1 is very similar to Rényi's Sixth Axiom A6, both are defined through the invariance of the dependence measure under transformations. Rényi (1959) proposed seven axioms for dependence measures $D[X;Y]$. **(A1)** $D[X;Y]$ is defined for any random variables X and Y; **(A2)** symmetric $D[X;Y] = D[Y;X]$; **(A3)** $0 \leq D[X;Y] \leq 1$; **(A4)** $D(X;Y) = 0$ if and only if X and Y are statistically independent; **(A5)** $D(X;Y) = 1$ if either $X = f(Y)$ or $Y = g(X)$ for some Borel-measurable functions f and g; **(A6)** If f and g are Borel-measurable, one-one mappings of the real line into itself then $D[f(X);g(Y)] = D[X;Y]$; **(A7)** If the joint distribution of X and Y is bivariate Gaussian, with linear correlation coefficient ρ, then $D[X;Y] = |\rho|$.

For a symmetric dependence measure (satisfying Axiom A2), Axiom A6 can be rewritten as $D[f(X);Y] = D[X;Y]$ for all Borel-measurable f. Hence self-equitability is a weaker version requiring an extra assumption that f satisfies the model (1).

It is known that Rényi's maximum correlation coefficient is the only measure that satisfies all seven Rényi's Axioms. However, Rényi's maximum correlation coefficient has a number of major drawbacks, e.g., it equals 1 too often and is generally not effectively estimable (Schweizer and Wolff, 1981; Székely and Rizzo, 2009). Hence enforcing all seven axioms is often considered too strong a constraint on the dependence measure, while some axioms are often considered desirable. For example, HSIC is shown to satisfy the first four axioms by Gretton et al. (2005a) and Gretton et al. (2005b).

For comparison, another weakened version of Axiom A6 is to restrict the transformations to monotone functions (Schweizer and Wolff, 1981), but without imposing the regression model (1). We may call this version of Axiom (A6*) *weak-equitability*.

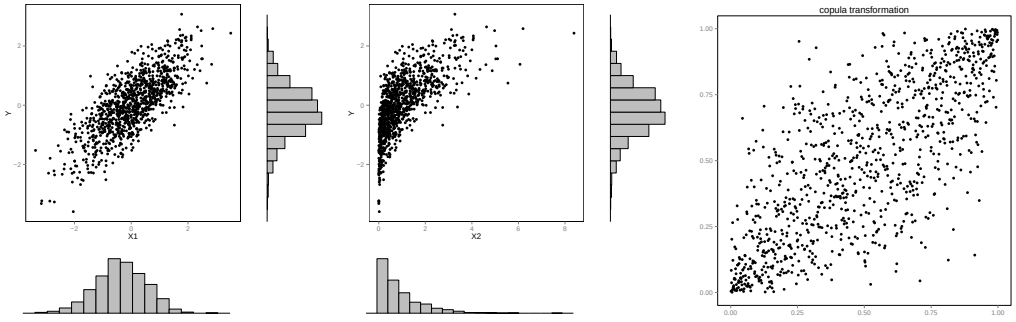


Figure 1: Left: Bivariate Gaussian with $\rho = 0.75$. Middle: Data with exponential marginal for X. Right: The Gaussian copula. The first two distributions have the same copula as in the right figure.

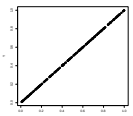
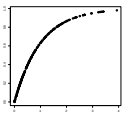
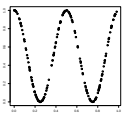
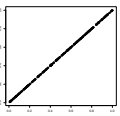
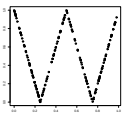
	raw data scale			copula transformation		
	A	B	C	A	B	C
						
	$\rho_{lin} = 1$	$\rho_{lin} = 0.866$	$\rho_{lin} = 0$	$\rho_{lin} = 1$	$\rho_{lin} = 1$	$\rho_{lin} = 0$

Table 1: Pearson correlation coefficient on three function relationships.

Definition 2 *A dependence measure $D[X;Y]$ is weak-equitable if and only if $D[X;Y] = D[f(X);Y]$ whenever f is a strictly monotone continuous deterministic function.*

The weak-equitable dependence measures treat all monotone (but not all nonlinear) relationships equally, and is a property shared by all copula-based dependence measures (e.g., CMMD and CHSNIC). Sklar's theorem ensures that, for any joint distribution function $F_{X,Y}(x,y) = Pr(X \leq x, Y \leq y)$, there exists a copula C—a probability distribution on the unit square $\mathcal{I}^2 = [0,1] \times [0,1]$—such that

$$F_{X,Y}(x,y) = C[F_X(x), F_Y(y)] \qquad \text{for all } x, y. \tag{2}$$

Here $F_X(x) = Pr(X \leq x)$ and $F_Y(y) = Pr(Y \leq y)$ are the marginal cumulative distribution functions (CDFs) of X and Y respectively. In other words, the copula C is the joint CDF of the two copula-transformed, uniformly distributed, variables $U = F_X(X)$ and $V = F_Y(Y)$. In this way, the copula decomposition separates the dependence from any marginal effects, and the copula C captures all the dependence between X and Y. Figure 1 shows the data from two distributions with different marginals but the same dependence structure.

Table 1 shows three simple examples and their respective copula transformations. We can see that the linear correlation ρ_{lin} prefers the linear relationship in (A). Applying on the copula-transformed variables on the right half of Table 1, ρ_{lin} (now equivalent to Spearman's ρ) becomes invariant to monotone transformation in (B), but still cannot capture the non-monotone nonlinear relationship in (C).

While copula-based dependence measures treats the monotone functions equally, equitability aims to also treat non-monotone functions equally. However, the original Rényi's Axiom A6 may be overly strong, and self-equitability aims to treat non-monotone functions equally only under the regression model (1).

We first consider some self-equitable copula-based dependence measures and further choose among these measures based on a new equitability definition in the next Subsection 2.2. Mutual information (MI), the recommended measure in Kinney and Atwal (2014), is self-equitable and is based on copula density $c(u, v)$,

$$\mathrm{MI} = \int_{I^2} \log[c(u, v)]c(u, v)dudv, \tag{3}$$

where I^2 is the unit square. We now consider a large class of self-equitable copula-based measures. Since the marginal variables X, Y are independent if and only if the corresponding copula distribution is uniform, we measure the dependence between X, Y through the distance between their copula distribution and the uniform distribution. Let the *Copula Distance* CD_α be the L_α distance between a copula density and the uniform copula density $\pi(u, v) = 1$.

$$\mathrm{CD}_\alpha = \int_{I^2} |c(u, v) - 1|^\alpha dudv, \quad \alpha > 0. \tag{4}$$

Combining Eq.4 in Fukumizu et al. (2007) and Eq.(4) here, CD_2 is the theoretical value of HSNIC in the large data limit. Our first result is that, the Copula Distance is self-equitable when $\alpha \geq 1$.

Lemma 1 *The Copula-Distance* CD_α *with* $\alpha \geq 1$ *is self-equitable.*

The proof follows from Theorems S3 and S4 of Kinney and Atwal (2014), since $g(x) = |x - 1|^\alpha$ is convex when $\alpha \geq 1$.

Remark: Schweizer and Wolff (1981) studied a class of dependence measures that are the L_α distance between a joint copula $C(u, v)$ and the uniform copula $\Pi(u, v) = uv$. The L_1, L_2 and L_∞ distance result in, the Wolf's σ, Hoeffding's Φ^2 and Wolf's κ respectively. Schweizer and Wolff (1981) showed that these measures satisfy a modified set of Rényi's Axioms, including Axiom (A6*) weak-equitability. In contrast to the Copula-Distance CD_α (L_α distance based on *copula densities*), these measures are based on the *cumulative distribution functions* and *are not self-equitable*. Since $C(u, v) = Pr(U \leq u, V \leq v)$ is the cumulative distribution function, such measures cumulate the deviation from independence from $u = 0$ to $u = 1$, and do not remain invariant for all nonlinear transformations f in model (1).

2.2 Robust Equitability

To select among the many self-equitable dependence measures, we want to consider additional equitability conditions. Some self-equitable dependence measures may not perform well in practice. For example, Rényi's maximum correlation coefficient (Rcor) satisfies the stronger Rényi's Axiom A6, thus it is also self-equitable. $Rcor(X; Y) = \sup_{f,g} \rho[f(X); g(Y)]$, where ρ is the linear correlation coefficient and the supremum is taken over all Borel-measurable functions f and g. Rcor has a number of major drawbacks, e.g., it equals 1 too often and is generally not effectively estimable (Schweizer and Wolff, 1981; Székely and Rizzo, 2009).

We observe the deficiencies more clearly in another self-equitable measure, the ideal dependence coefficient (IDC): $IDC(X; Y) = 0$ if X and Y are independent and $IDC(X; Y) = 1$ otherwise. IDC satisfies the first six Rényi's Axioms, and is self-equitable. It equals one for all dependent X and Y, providing no distinction of the dependence strength. IDC is only an abstract measure, the estimation of IDC is equivalent to *testing independence* between X and Y. However,

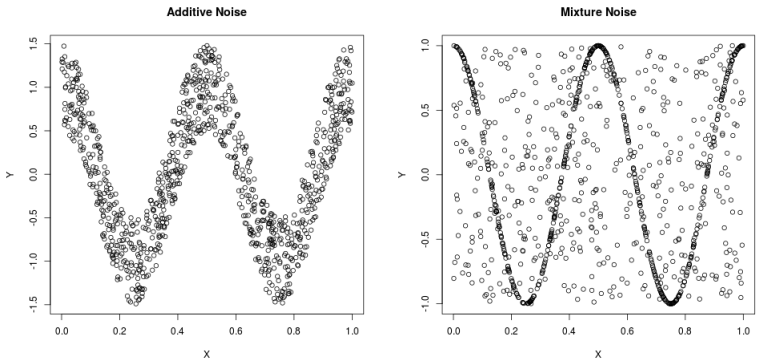


Figure 2: Left: Additive noise for self-equitability. Right: Mixture noise for robust-equitability.

for feature selection, IDC is not helpful at all because it provides no distinction of the dependence strength for different features. It is also hard to estimate. We wish for a new equitability criterion to exclude trivial dependence measures like IDC.

The self-equitability definition focuses on the regression $Y = f(X) + \epsilon$. However, in practice, this additive noise model does not capture all data types. In some cases, for example in sensor measurements, the deterministic signal is hidden in continuous background noise. Figure 2 illustrates these two types of noise. The Left subfigure shows additive noise on a deterministic sinusoidal function. The Right subfigure is the same deterministic signal on a uniform background noise. Mathematically, after the copula transformation, the second mixture noise model is described by a mixture copula: a continuous copula on the unit square I^2 is added to a deterministic signal C_s, which is a singular copula. Any copula can always be separated into a singular component and an absolutely continuous component (Nelsen, 2006, page 27). Independent background noise is represented by taking the absolutely continuous component as the independence copula $\Pi(u, v) = uv$ on I^2. Therefore, with p proportion of hidden deterministic relationship, the copula $C = pC_s + (1-p)\Pi$. Here C_s is a singular copula representing the deterministic relationship, so that its support $\mathcal{S}$ has Lebesgue measure zero. The equitability in this mixture noise model means that the dependence measure should give the same value for all types of deterministic signal C_s.

Definition 3 *A dependence measure $D[X;Y]$ is robust-equitable if and only if $D[X;Y] = p$ whenever (X, Y) follows a distribution whose copula is $C = pC_s + (1-p)\Pi$, for a singular copula C_s.*

Among the self-equitable Copula-Distances ($\alpha \geq 1$), L_1 distance is the special case that does reflect the proportion of deterministic relationship in the mixture copula.

$$CD_1 = p \int_{\mathcal{S}} C(du, dv) + \int_{\mathcal{I}^2 \setminus \mathcal{S}} |(1-p) - 1| dudv = p(1) + p = 2p.$$

Therefore we define the scaled version of CD_1 as *robust copula dependence (RCD)*

$$RCD = \frac{1}{2} CD_1 = \frac{1}{2} \int_{\mathcal{I}^2} |c(u, v) - 1| dudv. \tag{5}$$

Lemma 2 *The robust copula dependence RCD is robust-equitable.*

Mathematically, RCD is the same as Silvey's Delta (Silvey, 1964):

$$\Delta = \int_{\phi>1} [p(x,y) - p_X(x)p_Y(y)]dxdy,$$

where p_X and p_Y are the marginal probability densities for X and Y, p is the joint probability density for X and Y, and $\phi(x,y) = p(x,y)/[p_X(x)p_Y(y)]$. We write equation (5) in terms of the absolutely continuous copula density for ease of understanding. When part of the copula is singular, the RCD in (5) can be defined as in Silvey (1964), interpreting ϕ as the Radon-Nikodym derivative of the joint distribution with respect to a dominating probability measure which does cover the possibility of singularity. Alternatively, for a mixture copula C, the RCD can be defined as the limit of $\lim_{m\to\infty} RCD(C_m)$ for equation (5) on any sequence of continuous copulas $\{C_1, C_2, ...\}$ that converges to C. The convergence means that $\lim_{m\to\infty} \|C_m - C\|_1 := 2 \lim_{m\to\infty} \sup_A |C_m(A) - C(A)| = 0$, where the supremum is taken over all Borel sets A. A second way of interpretation is helpful in thinking about why CD_α, when $\alpha > 1$, can not be made robust-equitable and why this leads to statistical difficulties in estimation which we will discuss in detail in the next section.

Roughly speaking, for the mixture copula $C = pC_s + (1-p)\Pi$, the copula density for the absolutely continuous component is $c_c(u,v) = 1-p$, while we can imagine $c_s(u,v)$ as an abstract copula density for the singular component such that $\int_{\mathcal{B}} c_s(u,v)dudv := \int_{\mathcal{B}} C(du,dv) = C_s(\mathcal{B})$ for any subset $\mathcal{B} \subset \mathcal{S}$. Since $\mathcal{S}$ has Lebesgue measure zero, $c_s(u,v) = \infty$ for $(u,v) \in \mathcal{S}$ so that c_s is not a proper density, but rather an abstract limit of the sequence $\lim_{m\to\infty} c_{m,s}$. Here for any convergent sequence of continuous copulas $\{C_1, C_2, ...\}$ above, $c_{m,s}(u,v) = c_m(u,v) - c_c(u,v)$ is the continuous copula density that approaches the abstract $c_s(u,v)$. For any open set $\mathcal{B}_\mathcal{O} \supset \mathcal{B}$, $C_s(\mathcal{B}_\mathcal{O}) = \int_{\mathcal{B}_\mathcal{O}} c_s(u,v)dudv := \lim_{m\to\infty} \int_{\mathcal{B}_\mathcal{O}} c_{m,s}(u,v)dudv$, and $C_s(\mathcal{B}) = \lim_{\mathcal{B}_\mathcal{O}\to\mathcal{B}} C_s(\mathcal{B}_\mathcal{O})$. Hence for any $\alpha > 1$, $\int_{\mathcal{B}}[c_s(u,v)]^\alpha dudv = \int_{\mathcal{S}}[c_s(u,v)]^{\alpha-1}c_s(u,v)dudv = \int_{\mathcal{S}} \infty c_s(u,v)dudv = \infty$. So that $CD_\alpha = \infty$ whenever $\alpha > 1$ and $p > 0$. Similarly, $MI = \infty$ for all $p > 0$. They do not distinguish the dependence strength in the mixture distribution according to the signal proportion p, and can not be transformed to be robust-equitable as they over-emphasize the singular component (high copula density region).

If the dependence measure does not equal p exactly but is a monotone function of p, then we can scale it to get a robust-equitable version, and call it weakly-robust-equitable.

Definition 4 *A dependence measure $D[X;Y]$ is weakly-robust-equitable if and only if $D[X;Y]$ is a strictly monotone function of p whenever (X,Y) follows a distribution whose copula is $C = pC_s + (1-p)\Pi$, for a singular copula C_s.*

Lemma 3 *The Copula-Distance CD_α is weakly-robust-equitable if and only if $\alpha \leq 1$.*

When $\alpha > 1$, since $CD_\alpha = \infty$ whenever $p > 0$, those are not weakly-robust-equitable. When $\alpha < 1$, $\int_{\mathcal{S}}[c_s(u,v)]^\alpha dudv = \int_{\mathcal{B}}[c_s(u,v)]^{\alpha-1}c_s(u,v)dudv = 0$ so that the contribution from the singular region $\mathcal{S}$ is zero. In these cases,

$$CD_\alpha = 0 + \int_{\mathcal{I}^2\setminus\mathcal{S}} |(1-p) - 1|^\alpha dudv = p^\alpha$$

is weakly-robust-equitable. And CD_1 is weakly-robust-equitable from Lemma 2.

The self-equitability for additive noise model requires that $\alpha \geq 1$, while the weakly-robust-equitablity in the mixture noise model requires that $\alpha \leq 1$. Hence only $\alpha = 1$ satisfies the equitability condition in both noise models, and resulting in the robust-equitable RCD.

2.2.1 RCD, OTHER EQUITABILITY DEFINITIONS AND RÉNYI'S AXIOM

Having introduced our measure, RCD, and the robust-equitability definition, we can further compare them to the other equitability definitions in the literature. Reshef et al. (2011) considers equitability as the ability of a statistic $\hat{D}$ to approximately reflect the nonlinear R^2 over different relationships. They proposed a statistic MIC and demonstrate numerically such equitability through simulated examples. Kinney and Atwal (2014) propose to formalize such concept for the population parameter $D[X;Y]$ such that an R^2-*equitable* measure equals $g(R^2[f(X),Y])$ in the additive noise model (1) $Y = f(X) + \epsilon$, and showed that no nontrivial dependence measure can satisfy this R^2-equitability. The self-equitability definition is proposed as an alternative. Murrell et al. (2014) pointed out that such impossibility results are due to the non-identifiability due to the specification of ϵ term, allowing ϵ to possibly depend on $f(X)$. Under such specification, for example, a noiseless parabola can be realized as a noisy version of a noiseless linear relationship (Murrell et al., 2014). Reshef et al. (2015b) propose another formal equitability framework through interpretable intervals of a statistic under additive homoscedastic noises for both X and Y.

Our robust-equitability definition shares some common characteristics with both Kinney and Atwal (2014)'s and Reshef et al. (2015b)'s approach respectively. Similar to Kinney and Atwal (2014), our robust-equitability focuses on the population quantity $D[X;Y]$ instead of a statistic $\hat{D}[X;Y]$. This allows proof of theoretical equitability properties for specific dependence measures, as the statistical estimation error $\hat{D} - D$ can be kept as a separate issue. Robust-equitability does have implications on the statistical estimation error bounds which will be discussed in the next section. Self-equitability and robust-equitability focus on the invariance of $D[X;Y]$ as in Rényi's Axiom A6, but for different noise models. Reshef et al. (2015b) and our robust-equitability definition each focuses on a noise model to avoid the non-identifiability issue in Kinney and Atwal (2014)'s model: additive homoscedastic noise and mixture uniform noise respectively. Under each model, there is a clearly identifiable quantity of interest: the nonlinear R^2 and the mixture proportion p respectively.

Furthermore, our RCD satisfies the first five Rényi's Axioms. Particularly, $RCD = 0$ if and only if X and Y are statistically independent, $RCD = 1$ if X and Y are related through a deterministic relationship. And RCD is symmetric in that $RCD[X;Y] = RCD[Y;X]$. Notice that this symmetric property is an appropriate requirement for feature selection with the filtering method mRMR (Peng et al., 2005). Given a data set $\mathbf{X} \in \mathbb{R}^{n\times d}$, where n is the number of samples, d is the number of features, and target variable $Y \in \mathbb{R}^{n\times 1}$, let X_i denote the i-th feature, mRMR finds from the d-dimensional feature space, $\mathbb{R}^d$, a subspace of m features that optimally characterize Y, by solving the following optimization problem:

$$\max_S \frac{1}{|S|} \sum_{X_i \in S} D[X_i;Y] - \frac{1}{|S|^2} \sum_{X_i, X_j \in S} D[X_i;X_j], \tag{6}$$

where S is the optimal feature set, the first term maximizes relevance and the second term minimizes redundancy. Notice that this method only requires bivariate dependence between different features $D[X_i;X_j]$ in addition to the bivariate dependence between each feature and response variable $D[X_i;Y]$. The feature selection results of mRMR with an asymmetric dependence measure would depend on how the features are ordered, which is an undesirable characteristic.

RCD satisfies self-equitability, which is a weakened version of the sixth Rényi's Axiom. The Rényi's Axiom A7 requires the dependence measure to agree with the natural quantity of $|\rho|$ for bivariate Gaussian distributions (which corresponds to a linear regression model). Our RCD does not satisfy that, but instead agree with $|\rho|$ for the mixture noise setting with a linear deterministic relationship, since in that case $p = |\rho|$. Our robust-equitability definition requires the measure equals p exactly, which provides an easy interpretation in that it is an equitable extension of Pearson's correlation $|\rho|$ to all forms of hidden nonlinear deterministic relationships. It is not essential to require the exact equality to p, as equaling to a monotone function of p (weakly-robust-equitability) would enable a robust-equitable version of the measure through a transformation. However, precisely equaling to p is nice due to the above easier interpretation. Notice that $R^2 = \rho^2$ for the additive noise regression model with a linear relationship (bivariate Gaussian distribution). Hence the nonlinear R^2 can be similarly considered as an equitable extension of Pearson's correlation in the additive noise model to all forms of nonlinear regression relationships. However, unlike p, R^2 does not satisfy the symmetric property since regressing Y on X and regressing X on Y do not give the same value.

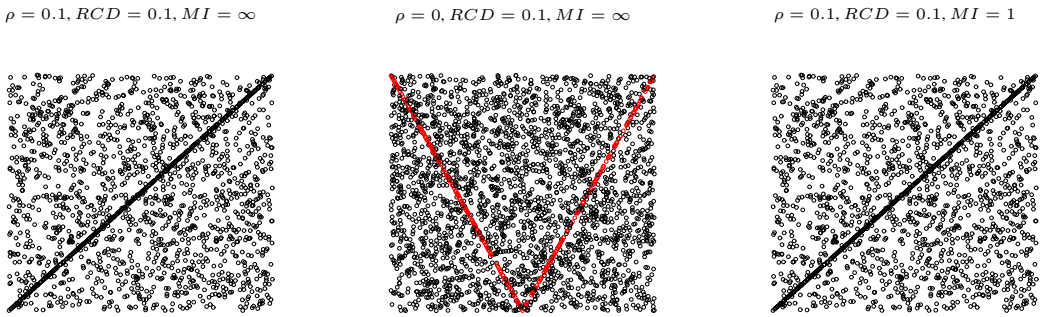
For discrete random variables, Equation (5) corresponds to the Kolmogorov dependence measure in the pattern recognition literature (Vilmansen, 1972, 1973; Ekdahl and Koski, 2006) and also known as the Mortara dependence index (Bagnato et al., 2013). In the discrete case, the measure has a maximum value less than 1. In contrast, $RCD = 1$ when X and Y are deterministically related. For continuous random variables X and Y, Silvey's Delta has been cited only as an abstract concept and no practical estimator was used in the literature for data analysis. The new name, ***Robust Copula Dependence*** (RCD), emphasizes the fact that it is a robust-equitable copula-based dependence measure.

2.3 Testing Independence Versus Estimation Errors of Dependence Measures

In practice, feature selection is based on an estimator $\hat{D}(X;Y)$ on the data set, since the exact value of the dependence measure $D(X;Y)$ is unknown to the user. Hence the feature selection results are affected by the estimation error $\hat{D} - D$. The estimation error also needs to be studied.

An estimator $\hat{D}[X;Y]$ is often used to test the independence between X and Y. Some studies compare different dependence measures $D[X;Y]$s by the power of independence testing using their corresponding estimators $\hat{D}[X;Y]$s (Reshef et al., 2011; Simon and Tibshirani, 2011; Reshef et al., 2015a). However, while independence testing is related to the estimation of dependence measures $D[X;Y]$, the power of the independence test is not a proper way of comparing dependence measures $D[X;Y]$. In fact, as mentioned in Section 2.2, the independence test corresponds to an estimation of the trivial measure IDC: $IDC(X;Y) = 0$ if X and Y are independent and $IDC(X;Y) = 1$ otherwise. Thus the power comparison is comparing the performance of $\hat{D}[X;Y]$ in estimating $IDC(X;Y)$ rather than estimating its corresponding parameter $D[X;Y]$. Also, good independence test statistics may not have corresponding interpretable dependence measures (Sun and Zhao, 2014).

As Reshef et al. (2015b) pointed out, the estimator $\hat{D}[X;Y]$ for equitable $D[X;Y]$ is most powerful at testing the hypothesis if the signal strength exceeding a threshold $D[X;Y] \geq D_0$, rather than being most powerful at testing independence $D[X;Y] = 0$. Besides simply testing for independence, dependence measures serve another important purpose: ranking the strength of the dependence relationships. For example, in the World Health Organization (WHO) data set in Reshef et al. (2011) the vast majority of the hundreds of the variables show dependence with other variables.



(a) (b) (c)

Figure 3: Hidden in background noise are 10% (red colored) data on a deterministic curve in (a) and (b), on a narrow strip around the line in (c).

So the independence tests do not provide much information there. To achieve sparse representation (feature selection), it is important to pick out the strongest dependence relationships. Equitable dependence measure ensures that the signal strength is reflected, rather than the functional form.

The robust-equitability definitions also have implications on the estimation errors. Some self-equitable measures can equal one too often, the extreme case being the IDC. When the dependence measure $D[X;Y]$ equals one for too many type of distributions, it does not distinguish dependence strength among them, and also makes its estimation difficult. Robust-equitability excludes such measures. In the next section, we theoretically show that robust-equitable RCD is intrinsically much easier to estimate than non-robust-equitable (but only self-equitable) MI and CD_2. This is due to the instability in the theoretical values of MI and CD_2. The following examples illustrate the difference between self-equitability (Figure 3: a versus b) and robust-equitability (Figure 3: a versus c). The self-equitable measures (RCD, MI), unlike ρ, have the same value in Figures 3(a) and 3(b) where each has 10% deterministic data on two different curves. In Figure 3(c), 10% of the data fall around (rather than exactly on) the line in a very small strip of area $0.1/exp(10) = 4.5 \times 10^{-6}$, which is very close to the distribution in Figure 3(a). The robust-equitable RCD values are very similar (differ only in 10^{-6} order) in Figures 3(a) and 3(c), but the MI values changes from ∞ to 1. Since these two almost indistinguishable distributions result in very different theoretical MI values (∞ and 1), no estimator can do well.

3. Statistical Estimation Errors

In this section, we study the estimation errors theoretically, and provide a practical estimator for RCD. In particular, we analyze the statistical estimation error $\hat{D} - D$.

3.1 The Inconsistency Results on Estimation of Mutual Information and CD_2

We theoretically show that MI and CD_2 are much harder to estimate compared to the robust-equitable RCD. We formally quantify the estimation difficulty through the minimax convergence rate over a family $\mathfrak{C}$.

Denote $\mathbf{z} = (u, v)$. Let $\mathfrak{C}$ be the family of continuous copulas with the density satisfying the following Hölder condition on the region where $c(\mathbf{z})$ is bounded above by some constant $M > 1$, denoted as A_M:

$$|c(\mathbf{z}_1) - c(\mathbf{z}_2)| \leq M_1 \|\mathbf{z}_1 - \mathbf{z}_2\|_{l_1}, \tag{7}$$

for a constant M_1 and for all $\mathbf{z}_1 \in A_M$, $\mathbf{z}_2 \in A_M$, and $\|\cdot\|_{l_1}$ denotes the l_1 norm.

The estimation of MI has been studied extensively in the literature. Over all distributions, even discrete ones, no uniform rate of convergence is possible for MI (Antos and Kontoyiannis, 2001; Paninski, 2003). On the other hand, many estimators were shown to converge to MI for every distribution. These results are not contradictory, but rather common phenomena for many parameters. The first result is about the uniform convergence over all distributions, while the second result is about the pointwise convergence for each distribution. The first restriction is too strong, while the second restriction is too weak. The difficulty of estimating a parameter needs to be studied for uniform convergence over a properly chosen family.

As MI is defined through the copula density, it is natural to consider the families generally used in density estimation literature. Starting from Farrell (1972), it is standard to study the minimax rate of convergence for density estimation over the class of functions whose m-th derivatives satisfy the Hölder condition. Since the minimax convergence rate usually is achieved by the kernel density estimator, it is also the optimal convergence rate of density estimation under those Hölder classes. Generally, with the Hölder condition imposed on the m-th derivatives, the optimal rate of convergence for two-dimensional kernel density estimator is $n^{-(m+1)/(2m+4)}$ (Silverman, 1986; Scott, 1992).

Therefore, when studying the convergence of MI estimators, it is very tempting to impose the Hölder condition on the copula density. In fact, imposing the Hölder condition (7) on the whole I^2, Liu et al. (2012) showed that the kernel density estimation (KDE) based MI estimator converges at the parametric rate of $n^{-1/2}$. Pál et al. (2010) also considered similar Hölder condition when they studied the convergence of k-nearest-neighbor (KNN) based MI estimator. However, such a condition is usually too strong for copula density, thus these results cannot fully reflect the true difficulty of MI estimation. When $c(u, v)$ is unbounded, the Hölder condition (7) cannot hold for the region where $c(u, v)$ is big. Hence imposing this Hölder condition (7) on the whole I^2 would exclude many commonly used continuous copula densities (e.g., Gaussian, student-T, etc.) since their densities are unbounded (Omelka et al., 2009; Segers, 2012). Therefore, we impose it only on the region where the copula density is small. Specifically, we assume that the Hölder condition holds only on the region $A_M = \{(u, v) : c(u, v) < M\}$ for a constant $M > 1$. Then this condition is satisfied by all common continuous copulas in the book by Nelsen (2006). For example, all Gaussian copulas satisfy the Hölder condition (7) on A_M for some constants $M > 1$ and $M_1 > 0$. But no Gaussian copulas, except the independence copula Π, satisfy the Hölder condition (7) over the whole $\mathcal{I}^2$.

Theorem 4 *Let $\widehat{MI}$ be any estimator of the mutual information* MI *based on the observations $\mathbf{Z}_1 = (U_1, V_1)$, ..., $\mathbf{Z}_n = (U_n, V_n)$ from a copula distribution $C \in \mathfrak{C}$. And let $\widehat{CD}_\alpha$ be any estimator of the CD_α in equation (4). Then*

$$\begin{aligned} \sup_{C \in \mathfrak{C}} E[|\widehat{MI}(C) - \mathrm{MI}(C)|] &= \infty, \textit{ and} \\ \sup_{C \in \mathfrak{C}} E[|\widehat{CD}_\alpha(C) - CD_\alpha(C)|] &= \infty, \textit{ for any } \alpha > 1. \end{aligned} \tag{8}$$

The detailed proof is provided in AppendixA. This theorem states that MI and CD_2 cannot be consistently estimated over the family $\mathfrak{C}$. This result does not depend on the estimation method used, as it reflects the theoretical instability of these quantities. There are many estimators for MI: kernel density estimation (KDE) (Moon et al., 1995), the k-nearest-neighbor (KNN) (Kraskov et al., 2004), maximum likelihood estimation of density ratio (Suzuki et al., 2009). However, practitioners are often frustrated by the unreliability of these estimation (Fernandes and Gloor, 2010; Reshef et al., 2011). This theorem provides a theoretical explanation.

Notice that the inconsistency results over this family $\mathfrak{C}$ is not due to the unboundedness of MI and CD_2. They can be transformed into correlation measures with values between 0 and 1 (Joe, 1989): $MIcor = \sqrt{1 - e^{-2MI}}$ and $\phi cor = \sqrt{CD_2/(1 + CD_2)}$. The $MIcor$ is known as the Linfoot correlation in the literature (Speed, 2011). We use the name $MIcor$ to indicate it as the scaled version of MI. The next theorem showed that $MIcor$ cannot be consistently estimated over the family $\mathfrak{C}$ also.

Theorem 5 *Let $\widehat{MIcor}$ be any estimator of $MIcor$ based on the observations $\mathbf{Z}_1 = (U_1, V_1)$, ..., $\mathbf{Z}_n = (U_n, V_n)$ from a copula distribution $C \in \mathfrak{C}$. Then*

$$\sup_{C \in \mathfrak{C}} E[|\widehat{MIcor}(C) - MIcor(C)|] \quad \geq \quad a_2 > 0, \tag{9}$$

for a positive constant a_2.

The detailed proof is provided in Appendix B.

The estimation difficulty of these dependence measures is due to their lack of smoothness related to being not weakly-robust-equitable. Reshef et al. (2015a, Section 4) proved a similar lack of smoothness of MI and $MIcor$, while their proposed statistic MIC may be considered a smoothed version which shows equitable behavior under their framework. Our results are stronger in that: (a) our results establish the statistical difficulty of estimation via minimax rate, and (b) our results apply to a broader class of dependence measures.

3.2 The Consistent Estimation of RCD

The equitability definitions and error analysis above assume a bivariate dependence measure. In this section, we will state the estimation results for a general d-dimensional RCD in equation (5): $RCD = \frac{1}{2}\int_{I^d} |c(\mathbf{z}) - 1| d\mathbf{z}$ for d-dimensional $\mathbf{z}$. That is, for $\mathbf{z} = (z_1, ..., z_d)$, the copula transformation changes each dimension to uniformly distributed variables $u_j = F_u^{-1}(z_j)$, $j = 1, ..., d$. Then $\int |c(\mathbf{z}) - 1| d\mathbf{z} := \int ... \int |c(u_1, ..., u_d) - 1| du_1 ... du_d$. The robust-equitability definition can be easily changed to the d-dimensional mixture copula with a singular component and the d-dimensional uniform distribution. And, in the d-dimensional case, the calculation above equation (5) still holds so that RCD is robust-equitable. Notice that other equitability definitions such as self-equitability is only defined for the bivariate case, and the filtering feature selection method mRMR also uses only the bivariate dependence.

Mathematically, MI (and $MIcor$) is unstable because it overweighs the region with large density $c(\mathbf{z})$ values. From equation (3), MI is the expectation of $\log[c(\mathbf{z})]$ under the true copula distribution $c(\mathbf{z})$. In contrast, RCD in (5) takes the expectation at the independence case Π instead. Even if $c(\mathbf{z})$ cannot be consistently estimated in the region A_M^c (the complement of A_M), its error contribution to $\widehat{RCD}$ can be bounded. The following theorem, which is proved in Appendix C, shows the result for the KDE estimator for RCD.

Theorem 6 *Let the KDE estimator of the d-dimensional copula density based on observations* $\mathbf{Z}_1, \cdots, \mathbf{Z}_n$ *be*

$$\hat{c}_{kde}(\mathbf{Z}) = \frac{1}{nh^d} \sum_{i=1}^{n} K(\frac{\mathbf{Z}_i - \mathbf{Z}}{h}). \tag{10}$$

We assume the following conditions:

- *The bandwidth* $h \to 0$ *and* $nh^d \to \infty$.
- *The kernel* K *is non-negative and has a compact support in,* $\mathbb{B}_0 = \{\mathbf{Z} : \|\mathbf{Z}\|_{l_2} \leq 1\}$, *the d-dimensional unit ball centered at* 0.
- *The kernel* K *is bounded.* $M_K = \max_{s \in \mathbb{B}_0} K(s)$, $\int_{\mathbb{B}_0} K(s)ds = 1$, $\mu_2^2 = \int_{\mathbb{B}_0} K^2(s)ds < \infty$

Then the plugged-in estimator $\widehat{RCD} = RCD(\hat{c}_{kde})$ *has a risk bound*

$$\sup_{C \in \mathfrak{C}} E[|\widehat{RCD} - RCD|] \leq M_1 h + \frac{\sqrt{2}\mu_2}{\sqrt{n}h^{\frac{d}{2}}} + O(\frac{1}{nh^d}). \tag{11}$$

In addition to the KDE based RCD estimator, we can estimate RCD consistently by plugging in the KNN estimator (Loftsgaarden and Quesenberry, 1965) of the copula density: $\hat{c}(\mathbf{z}) = k/n/A_{r(k,n)}$ using copula based observations $\mathbf{Z}_1, \mathbf{Z}_2, \cdots, \mathbf{Z}_n$. Here $r(k, n)$ is the distance from (d-dimensional) $\mathbf{z}$ to the k-th closest of $\mathbf{Z}_1, \mathbf{Z}_2, \cdots, \mathbf{Z}_n$, and A_r is the volume of the d-dimensional hyper-ball with radius r. Then $\widehat{RCD} = RCD(\hat{c}) = \sum_{\hat{c}(\mathbf{Z}_i)>1}[1 - 1/\hat{c}(\mathbf{Z}_i)]/n$.

Theorem 7 *Assuming* c *in* $\mathfrak{C}$ *has bounded continuous second order derivative in* A_M, $k \to \infty$ *and* $(k/n) \to 0$ *when* $n \to \infty$. *Then the plugged-in KNN estimator* $\widehat{RCD} = RCD(\hat{c})$ *has a risk bound*

$$\sup_{C \in \mathfrak{C}} E[|\widehat{RCD} - RCD|] \leq \tilde{c_1}(\frac{k}{n\epsilon})^{\frac{2}{d}} + \frac{\tilde{c_2}}{\sqrt{k}} + 2\epsilon, \tag{12}$$

for some finite constants $\tilde{c_1}$ *and* $\tilde{c_2}$, *and* $\epsilon = \epsilon(n)$ *is any sequence converging to* 0 *slower than* k/n.

Here, the extra technical assumption on the second order derivative allows a simpler proof (provided in Appendix D) by citing formulas in Mack and Rosenblatt (1979). Without it, RCD can still be estimated consistently as in Theorem 6. The error bound (12) is minimized by $\epsilon = (k/n)^{2/(d+2)}$ and $k = n^{4/(d+6)}$. Hence, in the bivariate ($d = 2$) case, we have $k = O(\sqrt{n})$. Simulations in Appendix E suggests a practical estimator with $k = 0.25\sqrt{n}$.

When RCD is estimated well under a sample size, further increasing the sample size does not change its estimation value much. In contrast, the estimated MI and CD_2 values can continuously change by a large margin as sample size increases, altering the ranking of features, sometimes to the wrong order.

Computational Complexity of KNN-based RCD Estimator The computation of KNN-based RCD is dominated by empirical ranking for each dimension and k-nearest neigbhor search for each sample. The ranking can be solved by mergesort, which costs $\mathcal{O}(dn \log n)$ for d dimensions (Cormen, 2009). The k-nearest neighbor search can be solved using k-d tree construction, which has $\mathcal{O}(kdn \log n)$ complexity when d is small (no larger than 20) (Bentley, 1975). However, the complexity can increase to $\mathcal{O}(kdn^2)$ if d becomes large. Therefore, the overall complexity is $\mathcal{O}(kdn \log n)$ for low dimension data and $\mathcal{O}(kdn^2)$ for high dimension data.

4. Experimental Results

In this section, we empirically verify the properties of RCD in our theoretical analysis.

We first check the estimation errors for RCD in synthetic experiments with additive noise and mixture noise respectively. For each type of noise, we simulate data with several different relationships so as to show the effect of self-equitability and robust-equitability respectively. In particular, we compare the RCD estimator with an MI estimator based on the same density estimation. Due to the non-robust-equitability of MI, in the mixture noise cases, the MI estimator varies widely with the sample sizes. In contrast, RCD converges as sample sizes increases. Therefore, MI may provide misleading ranking of features with unequal sample sizes. Also, the ranking between relationships with the two different noise types are greatly affected by the sample sizes under MI, while ranking under RCD remains relatively stable.

We then conduct several synthetic experiments to illustrate the properties in feature selection, and then show that similar patterns exist on real-world data sets. In Section 4.3, we show that: (1) Non-self-equitable dependence measures may provide misleading ranking under additive noise; (2) Non-robust-equitable dependence measures may provide misleading ranking under mixture noise when features have unequal sample sizes; (3) The ranking by non-robust-equitable dependence measures between the two types of noises are sensitive to sample size. Section 4.4 shows that similar behavior occurs in three real data examples. This confirms that the advantages of self-equitable and robust-equitable dependence measures are not just theoretical, but are real in some practical situations.

Furthermore, we compare the performance of feature selection by the filter method mRMR (Peng et al., 2005) as a feature search strategy and using various dependence measures as measures of relevance and redundancy (refer to Equation (6)). We conducted the mRMR method first on synthetic examples in Section 4.5, to illustrate why non-self-equitability or non-robust-equitability could lead to misleading results. We then perform mRMR on nine benchmark data sets from the UCI Data Repository (Lichman, 2013) in Section 4.6. Notice that the feature selection performance on a particular data set is affected by the type of existing relationships and the type of predictors used. For example, Pearson's correlation with a linear regression predictor should perform best if linear relationship is dominant in a data set. For a fair comparison, we measure performance by 10-fold cross-validated MSE of spline regression, a general nonlinear predictor (Friedman, 1991), using the selected features. Self-equitability and robust-equitability lead to equitable and robust feature selection. Hence RCD should provide stable performance across different types of data, not necessarily best in each situation. However, over many data sets with different types of nonlinear relationships, robust-equitable RCD would provide best average performance as confirmed on these nine benchmark data sets.

There are some parameters to be set for computing various dependence measures. For kernel based measures, we follow the settings used by Fukumizu et al. (2007). For HSNIC, we set the regularization parameter $\epsilon_n = 10^{-5}n^{-3.1}$ to satisfy the convergence guarantee given by Theorem 5 from Fukumizu et al. (2007). As discussed in the previous section, we set $k = 0.25\sqrt{n}$ for the k-NN estimator of MI, RCD and CD_2.

4.1 Estimation Errors and Equitability

In this section, we study the estimation errors of our RCD estimates through synthetic experiments, and examine the equitability effect in ranking features. We first generate data from four different

type	level	Linear		Square Root		Cubic		Quadratic	
		1k	10k	1k	10k	1k	10k	1k	10k
add	0.4	0.35(0.01)	0.38(0.00)	0.36(0.01)	0.38(0.00)	0.36(0.01)	0.38(0.00)	0.35(0.01)	0.37(0.00)
	0.6	0.54(0.01)	0.58(0.00)	0.54(0.01)	0.58(0.00)	0.55(0.01)	0.58(0.00)	0.53(0.01)	0.57(0.00)
	0.8	0.76(0.00)	0.79(0.00)	0.76(0.01)	0.79(0.00)	0.76(0.01)	0.79(0.00)	0.75(0.01)	0.78(0.00)
mix	0.4	0.43(0.02)	0.43(0.01)	0.42(0.02)	0.42(0.01)	0.42(0.02)	0.42(0.01)	0.39(0.02)	0.42(0.01)
	0.6	0.62(0.02)	0.62(0.00)	0.61(0.02)	0.62(0.00)	0.61(0.02)	0.62(0.00)	0.59(0.01)	0.61(0.01)
	0.8	0.81(0.01)	0.81(0.00)	0.80(0.01)	0.81(0.00)	0.80(0.01)	0.81(0.00)	0.78(0.01)	0.80(0.00)

Table 2: The expected values of RCD estimates (with standard deviation in parenthesis) based on 100 simulations, under various functional types, sample sizes, noise types and noise levels.

type	level	Linear		Square Root		Cubic		Quadratic	
		1k	10k	1k	10k	1k	10k	1k	10k
add	0.4	0.77(0.01)	0.77(0.00)	0.77(0.01)	0.77(0.00)	0.75(0.01)	0.75(0.00)	0.00(0.04)	0.00(0.01)
	0.6	0.90(0.00)	0.90(0.00)	0.90(0.01)	0.90(0.00)	0.87(0.01)	0.87(0.00)	0.00(0.04)	0.00(0.01)
	0.8	0.98(0.00)	0.98(0.00)	0.96(0.00)	0.96(0.00)	0.91(0.00)	0.91(0.00)	0.00(0.04)	0.00(0.01)
mix	0.4	0.40(0.03)	0.40(0.01)	0.33(0.03)	0.33(0.01)	0.34(0.03)	0.33(0.01)	-0.01(0.03)	0.00(0.01)
	0.6	0.60(0.03)	0.60(0.01)	0.52(0.03)	0.51(0.01)	0.51(0.03)	0.5(0.01)	0.00(0.03)	0.00(0.01)
	0.8	0.80(0.02)	0.80(0.01)	0.73(0.02)	0.72(0.01)	0.69(0.03)	0.69(0.01)	0.00(0.04)	0.00(0.01)

Table 3: The expected values of ρ estimates (with standard deviation in parenthesis) based on 100 simulations.

functional types: linear, square root, cubic, and quadratic (cases A, C, D, F in Table 11 of Appendix E) with two sample sizes of $n = 1000$ and $n = 10000$ respectively. Also, data with two different noise paradigms and three noise levels are tested with three measures, RCD, (non-self-equitable) ρ and (non-robust-equitable) CD_2.

As we can see from Table 2, the standard deviation of the RCD estimates are small, and the expected value of RCD converges to the true values as sample size increases. The expected values of RCD was similar for the different functional relationships. They are closer to the true values under the mixture noise than under the additive noise. For either noise type, the estimates are very accurate for sample size $n = 10,000$. Although there are some random estimation errors, the RCD estimates would not miss-rank features with moderate difference in true RCD values. That is, it never ranks features with real RCD = 0.2 as more dependent than features with real RCD = 0.4, under samples $n = 1000$ or $n = 10,000$. So the RCD estimates can be used to provide reliable dependence ranking that do not change dramatically under sample sizes $n = 1000$ versus $n = 10,000$.

In contrast, the non-self-equitable Pearson's ρ values depend on the functional relationship. Under the mixture noise, the estimated ρ values are close to the mixture proportion only for the linear relationship. Hence its ranking of features is heavily influenced by the functional relationships. Particularly, it fails to detect the quadratic relationship in the last column of Table-3. Even when 80% of data follows the deterministic quadratic relationship, it is still ranked as less dependent than the features with other functional relationships (even if the other features has only 40% deterministic mixture proportion).

For the non-robust-equitable CD_2 in Table 4, its value changes dramatically under sample sizes $n = 1000$ versus $n = 10,000$, especially under mixture noise. This demonstrates that sample size affects the ranking by CD_2 in contrast to the ranking by the robust-equitable RCD. If the

type	level	Linear		Square Root		Cubic		Quadratic	
		1k	10k	1k	10k	1k	10k	1k	10k
add	0.4	0.94(0.04)	1.23(0.02)	0.96(0.04)	1.28(0.02)	1.08(0.04)	1.45(0.02)	0.88(0.03)	1.19(0.01)
	0.6	1.7(0.04)	2.25(0.02)	1.81(0.05)	2.46(0.02)	2.19(0.06)	3.11(0.03)	1.6(0.04)	2.27(0.02)
	0.8	3.62(0.06)	5.06(0.03)	3.72(0.07)	5.47(0.03)	4.35(0.10)	7.52(0.05)	2.93(0.06)	4.96(0.03)
mix	0.4	1.29(0.09)	2.62(0.07)	1.30(0.09)	2.69(0.072)	1.24(0.10)	2.59(0.06)	0.83(0.06)	1.71(0.04)
	0.6	2.70(0.15)	5.78(0.10)	2.73(0.15)	5.89(0.09)	2.66(0.17)	5.76(0.10)	1.69(0.08)	3.63(0.06)
	0.8	4.75(0.15)	10.24(0.11)	4.79(0.17)	10.34(0.11)	4.75(0.16)	10.24(0.11)	2.87(0.08)	6.27(0.07)

Table 4: The expected values of CD_2 estimates (with standard deviation in parenthesis) based on 100 simulations.

two features have unequal sample sizes, then the feature with larger sample size has a built-in preference by CD_2. For each of the functional types, CD_2 ranks a feature with mixture proportion (of deterministic data) 0.6 but large sample size $n = 10000$ as more dependent than a strongly dependent feature with mixture proportion 0.8 but smaller sample size $n = 1000$. Also, for features with different noise types, their ranking are inconsistent when sample size changes. For example, for the linear relationship under sample size $n = 1000$, CD_2 ranks the feature with RCD = 0.6 mixture noise as less dependent than the feature with RCD = 0.8 additive noise. But when sample size is increased to $n = 10000$, the ranking between these two features reverses. While it is not necessary for other dependence measures to rank features across different noise types in the same order as RCD, the stability of the ranking under different sample sizes is desirable. The non-robust-equitable dependence measures may not provide consistent ranking.

In summary, we observe three advantages of RCD for feature selection, in comparison to other dependence measures. (1) Non-equitable measures such as ρ may prefer certain functional relationships (say, linear), while RCD treat them equitably. (2) Self-equitable but non-robust-equitable measure such as CD_2 prefer features with larger sample size. (3) Self-equitable but non-robust-equitable measure such as CD_2 does not provide stable ranking among features when sample size changes.

4.2 Multivariate Equitability Analysis

We now perform equitability analysis on simulated multivariate data. Following the framework of Reshef et al. (2014); Murrell et al. (2016), we generate noisy data from various functional forms, and plot the estimated dependence measure values against the signal level in Figure 4. Our robust-equitability definition extends the natural signal level to higher dimensional case in the mixture noise model. Here we study three dimensional ($d = 3$) cases of six different nonlinear function relationships, generated with different portion of uniform noise from 0 to 0.9 (with signal portion from 0.1 to 1). The sample size $n = 1000$ is used in this experiment. Figure 5 plots the six three dimensional different nonlinear function relationships at signal level 0.8 (0.2 proportion of uniform noise).

According to Reshef et al. (2014), a measure is more equitable if the length of a band, when cutting each plot in Figure 4 with a horizontal line, is smaller. That is, the measure with smaller bandwidth could capture the dependence purely based on the noise level, and is robust to different (linear and nonlinear) relationships. On the other hand, if the band is very wide, it will give the same score for data with a wide range of noise levels, and hence could not identify strong relationships correctly. Figure 4 shows that RCD is the most equitable, since it has the narrowest bandwidth and

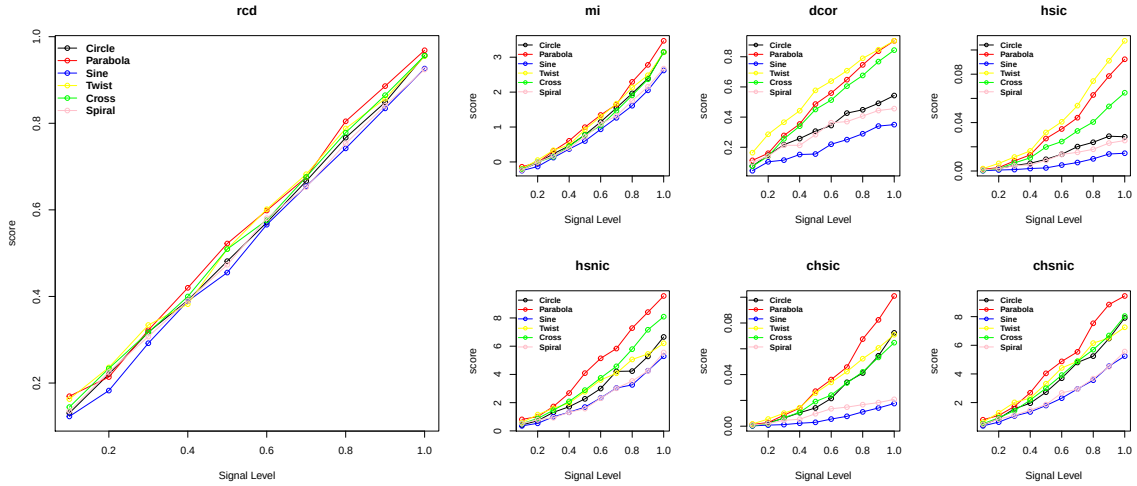


Figure 4: Seven dependence measure values versus signal levels.

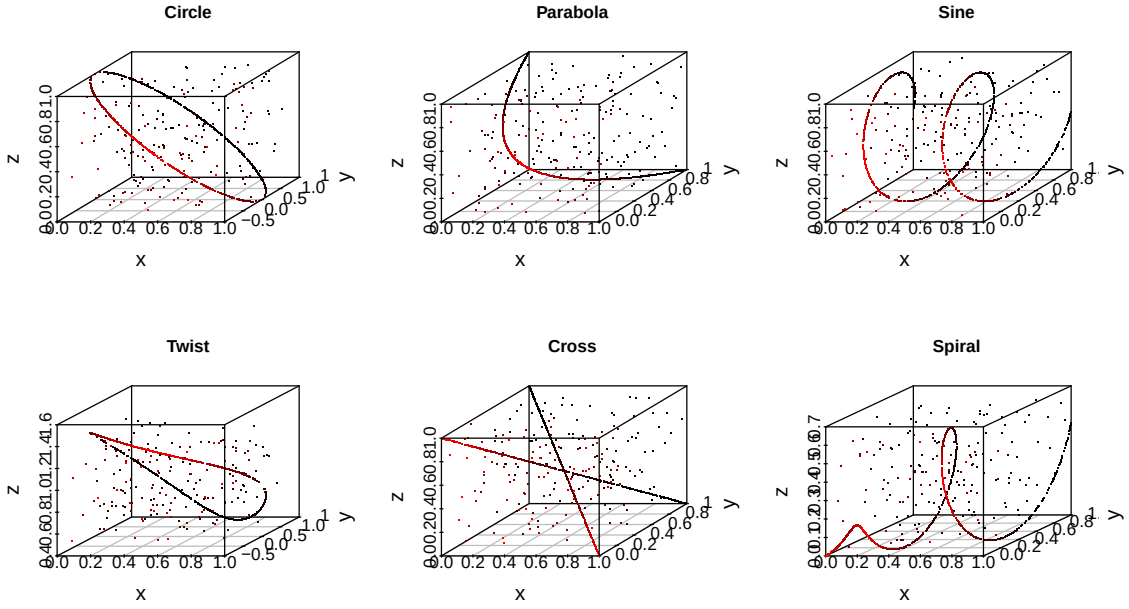


Figure 5: Six function types, circle, parabola, sine/cosine, twisted curve, two cross lines, and a spiral curve, with 20% uniform noise from 0 to 0.9 (with signal portion from 0.1 to 1). The examples in this figure is of signal level 0.8.

lies on the signal level line. The self-equitable mutual information is second best, has relatively narrower bandwidth compared to other measures except RCD.

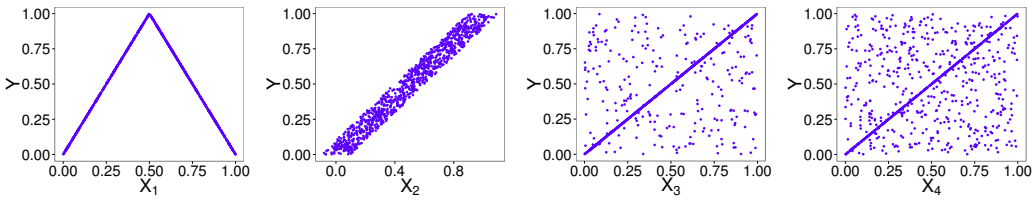


Figure 6: (X_1, Y) has a nonmonotonic relationship with deterministic signal. (X_2, Y) has linear relationship with uniform additive noise with width $d = 0.2$. (X_3, Y) has background noise with a 0.75 linear signal portion. (X_4, Y) is similar to (X_3, Y) but with a 0.5 signal portion.

	X_1		X_2		X_3		X_4	
n	300	10k	300	10k	300	10k	300	10k
ρ_{lin}	0.021	0.00043	0.98	0.98	0.75	0.75	0.51	0.51
HSIC	0.036	0.033	0.095	0.093	0.061	0.057	0.025	0.025
CMMD	0.034	0.034	0.095	0.095	0.060	0.056	0.026	0.025
HSNIC	3.40	3.57	3.63	3.98	3.55	4.08	1.84	1.81
CHSNIC	3.30	3.60	3.59	3.95	3.53	4.07	1.84	1.81
MI	5.06	7.62	2.28	2.37	3.65	5.46	1.82	2.97
CD_2	21.37	102.09	3.75	3.74	24.10	146.73	7.86	44.15
RCD	0.93	0.99	0.77	0.80	0.75	0.76	0.52	0.52
MSE	0.00098	0.0023	0.0033	0.0031	0.037	0.037	0.064	0.063

Table 5: Dependence measure values for synthetic data. Sample sizes $n = 300$ and $n = 10,000$ are considered. Each row corresponds to one type of measure. The MSE is presented in the last row.

4.3 Synthetic Data Sets I: Ranking Features

To compare the performance of each dependence measure in feature selection, we consider four features X_1, X_2, X_3, X_4 and target variable Y as shown in Figure 6. Y has a nonmonotonic but deterministic relationship with X_1 and a linear relationship with X_2 plus some additive noise. In addition, Y has linear relationships with both X_3 and X_4 corrupted by increasing level of continuous background noise. For each feature X_i, we calculate its dependence measure with Y for different sample sizes $n = 300$ and $10,000$. Results are presented in Table 5.

Since X_1 has a deterministic relationship with Y, it should be ranked as more dependent than the other features. X_3 and X_4 has mixture noise with the mixture proportions of 0.75 and 0.5 respectively. We can see that the values learned by RCD are close to those values and correctly ranks X_3 as more dependent than X_4. The last row of Table 5 reports the 10-fold cross-validated mean-squared-error (MSE) of a nonlinear predictor using each feature. Here RCD results are consistent with the MSE results, providing higher scores for those with lower MSE values (more predictive of Y). Now, we inspect the other dependence measures and observe the three issues mentioned in Section 4.1.

Self-equitability. We expect the self-equitable measures to treat linear and nonlinear models equally (i.e., they should prefer X_1 over X_2 because X_1 is purely deterministic while X_2 has some noise). As we can see from Table 5, Pearson correlation coefficient ρ_{lin}, and kernel-based measures prefer

X_2 more than X_1. On the other hand, self-equitable measures MI, CD_2 and RCD were able to rank the features correctly. Although HSNIC and CHSNIC have the same value as CD_2 in the large data limit, empirically they behave similarly to other non-self-equitable kernel-based measures due to slow convergence of their estimators (Reddi and Póczos, 2013).

Selection Correctness in Unequal Sample Sizes. In real applications, some features may have some missing measurements, resulting in unequal number of samples among the various features. In this setting, we still want to compare feature relevance. An ideal dependence measure should not be influenced greatly by unequal sample sizes. Let us take a closer look at MI and CD_2 and on how they rank features X_3 and X_4. Note that X_3 has a stronger signal-to-noise proportion than X_4 ($p = 0.5$ versus $p = 0.75$); thus, ideally, one would like the measure to rank X_3 higher than X_4 as empirically confirmed by the MSE results. The ranking provided by MI and CD_2 is correct when both features have the same sample size, $n = 10,000$. However, if the stronger feature X_3 has missing measurements so that $n = 300$ for X_3, then X_3 is ranked lower than X_4 by CD_2, which would mislead feature selection algorithms. MI will make the same mistake if the sample size for X_4 further increases.

Selection Stability in Different Sample Sizes. Ideally, a measure should not vary too much as the sample size changes. However, we observe that MI, CD_2, and HSNIC's ranking of features X_2, X_3 and X_4 is affected when the sample size is increased. With fixed sample size $n = 300$, MI ranks X_2 as having higher deterministic relationship with Y compared to X_4. However, when the sample size is increased to $n = 10,000$, it reverses the ranking of these features. This is due to its non-robust-equitability and resulting estimation difficulty, as proved in Theorems 4 and 5. Additionally, similar phenomenon appears for CHSNIC and HSNIC on features X_2 and X_3. We observe that when $n = 300$, they rank X_2 as having higher dependence with Y compared to X_3. However, when $n = 10,000$, these rankings are reversed. These inconsistencies may mislead feature selection algorithms.

4.4 Real Data Sets I: Ranking Features

In this subsection, we verify that the equitability properties in Subsection 4.3 are also observed on real data.

Self-equitability. Consider the stock data set from StatLib[1]. This data set provides daily stock prices for ten aerospace companies. Our task is to determine the relative relevance of the stock price of the first two companies (X_1, X_2) in predicting that of the fifth company (Y). The scatter plots of Y against X_1, X_2 are presented in Figure 7. Ideally, self-equitable measures should prefer X_1 over X_2 because the MSE associated with X_1 is lower even though it has a more complex function form. As we can see from Table 8, self-equitable measures MI, CD_2, and RCD all correctly select X_1. While measures that are not self-equitable fail to select the right feature.

Selection Correctness in Unequal Sample Sizes. Consider the KEGG metabolic reaction network data set (Lichman, 2013). Our task is to select the most relevant features in predicting target variable 'Characteristic path length' (Y). The 'Average shortest path' (X_1), 'Eccentricity' (X_2) and 'Closeness centrality' (X_3) are used as candidate features. Observe the ranking of X_1 and X_3 from Table 6, when they have equal sample sizes (either 1000 or 20,000), MI, CD_2 and RCD all rank X_1 as being more relevant than X_3. The MSE values also confirmed that X_1 is more predictive of Y than X_3. However, if there are missing measurements of X_1, then we may need to compare X_1

1. `http://lib.stat.cmu.edu/`

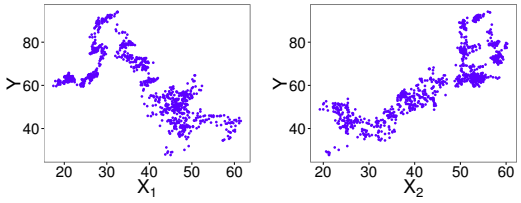


Figure 7: $(X_i, Y), i = 1, 2$ in stock data set

Measures	$X1$	$X2$
ρ_{lin}	-0.68	0.83
HSIC	0.053	0.068
CMMD	0.062	0.073
HSNIC	1.95	2.16
CHSNIC	1.90	1.99
MI	2.06	1.92
CD_2	3.88	3.13
RCD	0.68	0.67
MSE	0.18	0.23

Figure 8: Measures for X_1, X_2 in stock data set

	X_1		X_2		X_3	
n	1k	20k	1k	20k	1k	20k
MI	3.39	3.95	3.23	3.66	2.94	3.54
CD_2	12.05	31.65	10.67	22.44	9.77	28.30
RCD	0.85	0.86	0.82	0.84	0.77	0.80
MSE	0.030	0.028	0.032	0.032	0.14	0.14

Table 6: Dependence measure for three features in metabolic reaction network data set

with 1000 samples and X_3 with 20,000 samples. The feature X_3 with less signal strength but larger sample size is given higher ranking by MI and CD_2, degrading the performance of feature selection algorithms. In contrast, RCD correctly identifies X_1 as being more relevant than X_3 even with the unequal sample size.

Selection Stability in Different Sample Sizes. Ideally, a measure should not vary too much as the sample size changes. However, in Table 6, CD_2's ranking of features X_2 and X_3 is affected by the increase in sample size. If we fix sample size $n = 1000$, CD_2 ranks X_2 as more relevant than X_3 in predicting Y, agreeing with the MSE ranking. However, when the sample size increases to $n = 20,000$, CD_2 will prefer X_3. CD_2 will select the feature X_3 when the sample size is large even though it is less relevant to Y. RCD has the same ranking under both sample sizes.

4.5 Synthetic Data Sets II: mRMR Feature Selection

In this part, we investigate the performance of feature selection for each dependence measure with mRMR (Peng et al., 2005). We generate data from the following additive regression model $Y = 1.5\cos(3\pi X_1) + (1 - 2|2X_2 - 1|)^2 + \epsilon$, where X_1 and X_2 are uniformly distributed on $[0, 1]$, and $\epsilon \sim N(0, 0.05)$. We consider feature selection from twenty features. The first two features are X_1 and X_2. The next six features are noisy versions of X_1 and X_2, with some as mixtures. They are also related to Y. A good feature selection method should select X_1 and X_2 before these more

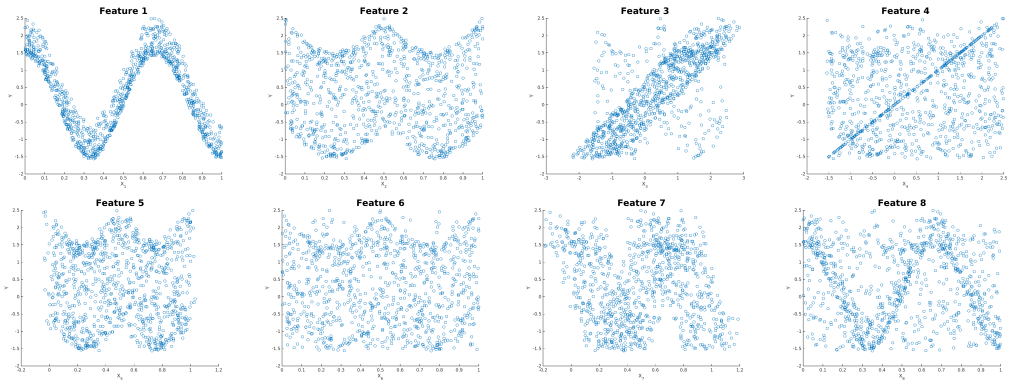


Figure 9: Scatter plots of the features $X_1, \cdots, X_8$ versus Y.

Feature X_1	X_1
Feature X_2	X_2
Feature X_3	25% X_2 and 75% Y with additive noise $U(-0.75, 0.75)$
Feature X_4	20% X_2, 20% Y and 60% background noise
Feature X_5	X_2 with additive noise $U(-0.05, 0.05)$
Feature X_6	50% X_2 and 50% background noise
Feature X_7	X_1 with additive noise $U(-0.2, 0.2)$
Feature X_8	50% X_1 and 50% background noise
Feature X_9,...,X_{20}	pure random noise

Table 7: Features that are used in the additive regression model.

noisy features X_3 to X_8. The rest 12 features X_9 to X_{20} are simply random noise not related to X_1, X_2 or Y. The features are listed in Table 7. And we plot the first eight features versus Y in one simulation run in Figure 9.

We generate data sets with size $n = 1000$, and select the features with mRMR based on the eight dependence measures. We repeat this experiment fifty times and record the order of each feature being selected in each data sets. We then apply spline regression model using the top one to ten selected features and report their respective cross-validated mean square error (MSE). The cross-validated MSE averaged over 50 runs are plotted in Figure 10.

In this synthetic example, RCD yields the lowest MSE. We can see why by looking at the feature selection result in more details. Table 8 reports the features that are most frequently selected as the top one to five features using mRMR. Inversely, Table 9 shows the median order of being selected for each relevant feature X_1, ..., X_8.

From the tables, RCD correctly selects the two true features X_1 and X_2 as the top two features, thus resulting in the lowest MSE curve in Figure 10. The non-self-equitable measures, for this data distribution, incorrectly ranks first the feature X_3 which has linear relationship with Y in parts of the data (75% mixture). Then Pearson's correlation ρ does not rank X_1 and X_2 high because the relationships are nonlinear. Some nonlinear non-self-equitable measures (HSNIC, CMMD, CHSNIC) are able to rank the feature X_1 second, but cannot select feature X_2 which has a nonlinear relationship with Y. That is due to X_2, which also has some dependence with X_3, was penalized when X_3 was selected first. The self-equitable measures MI, CD_2 and RCD, in contrast, correctly rank X_1 first. However, only the robust-equitable RCD ranks X_2 second. The non-robust-equitable MI and CD_2 incorrectly ranks the noisy X_4 second, instead of X_2, due to mishandling of the mixture noise.

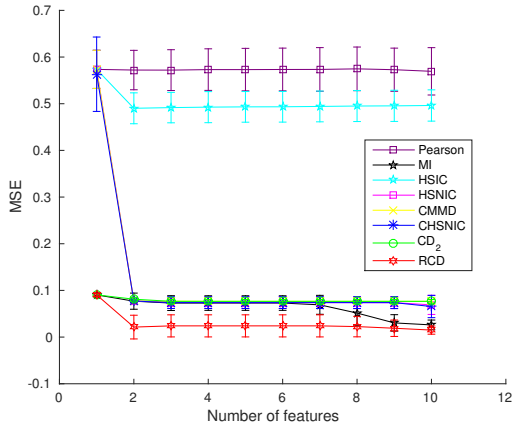


Figure 10: Cross-validated MSE plot with error bar based on 50 repeated runs.

Order of selection	1	2	3	4	5
ρ	$X_3(100\%)$	*	$X_4(50\%)$	*	*
HSIC	$X_3(100\%)$	$X_7(96\%)$	*	*	*
HSNIC	$X_3(100\%)$	$X_1(100\%)$	$X_4(78\%)$	*	*
CMMD	$X_3(100\%)$	$X_1(100\%)$	*	$X_4(52\%)$	*
CHSNIC	$X_3(100\%)$	$X_1(100\%)$	$X_4(98\%)$	*	*
MI	$X_1(100\%)$	$X_4(78\%)$	$X_3(90\%)$	*	*
CD_2	$X_1(100\%)$	$X_4(80\%)$	$X_3(80\%)$	*	*
RCD	$X_1(100\%)$	$X_2(82\%)$	$X_3(90\%)$	*	*

Table 8: The most frequently selected top five features, with the relative frequency in parenthesis. Asterisk indicates one of the random noise features (X_9, ..., X_{20}).

Feature	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
ρ	20	19	1	3.5	18	15	18	16
HSIC	18	20	1	4	19	16	2	17
HSNIC	2	17.5	1	3	20	19	18	10
CMMD	2	20	1	4	18	16	19	17
CHSNIC	2	15.5	1	3	20	19	17	10
MI	1	20	3	2	9	19	13	12
CD_2	1	20	3	2	17	18	16	19
RCD	1	2	3	7	20	19	13	6

Table 9: Median of the order of selection for the first eight features in each dependence measure experiment among fifty repeated runs.

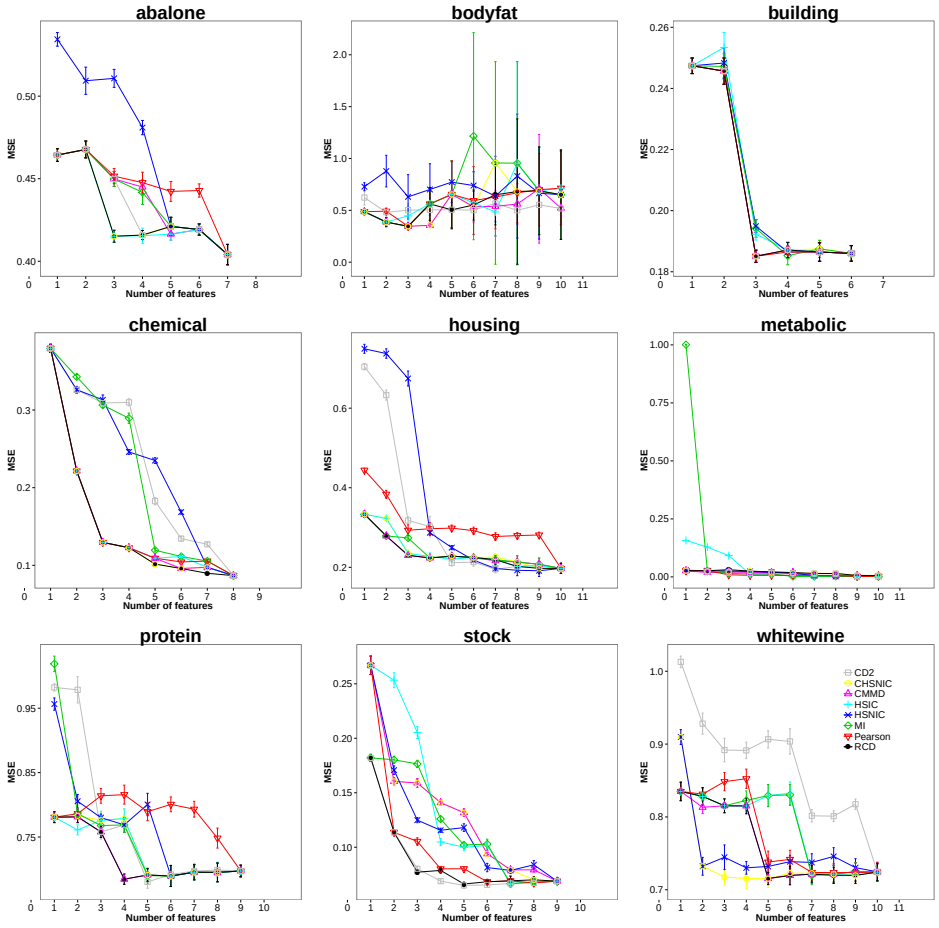


Figure 11: MSE of mRMR-based Feature Selection on nine real-world data sets

4.6 Real Data Sets II: mRMR Feature Selection

Here we used the various dependence measures as measures of relevance and redundancy in the mRMR-based search strategy, and compare the feature selection results on nine real-world data sets. Due to the cubic computational cost of kernel-based measures (HSNIC, CHSNIC), up to 1000 samples are used for each data set. We compare the results of RCD versus other dependence measures by showing plots of 10-fold cross-validated MSE using spline regressor with the features selected by these measures versus the number of selected features in Figure 11.

We used the Kruskal-Wallis test to compare the MSE between different measures. Table 10 lists the top dependence measures in order of their MSE. For each data set, we only include the dependence measures resulting in MSE equivalent to the best MSE (p-value > 0.05 for Kruskal-Wallis test) in the table. We can see that RCD performs as one of the best in 8 out of 9 data sets. Most other measures are worse off in more than half of the data sets. Only CHSNIC and CMMD are close in performance. In particular, CMMD is among the best measures in 5 data sets. CHSNIC performs as one of the best in 6 data sets and actually beats RCD in one data set, *whitewine*. RCD in fact find the best top feature in every data set including *whitewine*. However, the best second feature

Data set	Top Dependence Measures
abalone	**RCD** CHSNIC HSIC
bodyfat	**RCD** CMMD CHSNIC
building	**RCD** CHSNIC Pearson CMMD CD_2
chemical	**RCD** CHSNIC
housing	**RCD** CMMD
metabolic	**RCD** CHSNIC CD_2 CMMD HSNIC Pearson
protein	**RCD** CMMD
stock	**RCD** CD_2
whitewine	CHSNIC

Table 10: Measures ranked by predictive MSE

in *whitewine* was not selected by RCD. Overall, RCD has the best performance in mRMR-based feature selection compared to competing dependence measures.

5. Conclusions and Discussions

As the data size explodes, researchers are studying increasingly complex relationships among features. Restricting the focus on simple linear relationship can miss very informative features. Therefore, how to measure the dependence strength equitably for various functional relationships has attracted recent interest from researchers (Reshef et al., 2011; Kinney and Atwal, 2014; Murrell et al., 2014; Reshef et al., 2015b). This paper provides a theoretical treatment of various equitability definitions, including our proposal of the robust-equitability concept. The robust copula dependence (RCD) is proven to be both self-equitable and robust-equitable. Theoretically we show that RCD is intrinsically easier to estimate than some other self-equitable dependence measures (such as MI and CD_2). Particularly, through minimax rate of convergence, we provide a theoretical explanation for the difficulty of accurately estimating MI which is noted by practitioners. A practical estimator is provided for RCD, which enables its usage in feature selection.

Through theoretical and empirical studies, we have shown that RCD does better in ranking the features according to deterministic signal strengths compared to other dependence measures. The non-self-equitable measures may prefer noisy features with certain types of relationships (e.g., monotonic) over less noisy features with more complex relationships. Self-equitable but non-robust-equitable measures (such as MI and CD_2) overcome this deficiency but have estimation problems, leading to non-robust feature selection particularly when comparing features with unequal sample sizes. RCD can be used in feature selection to overcome these limitations.

Using nonlinear dependence measures, rather than the Pearson's correlation, in high-dimensional data analysis (e.g. independent component analysis) is becoming more popular to deal with possibly non-Gaussian noises. The equitability properties of RCD make it an ideal choice of dependence measure in such applications. Replacing measures such as MI by RCD may lead to more robust results as shown in the examples of Section 4. RCD estimation in high-dimensional case however, similar to MI, may be inaccurate as it involves high-dimensional density estimation. The improvement on nonparametric RCD estimation remains an ongoing research effort, and can lead to wider applications.

Acknowledgments

We would like to acknowledge support for this project from NSF grant CCF-1442728.

Appendix A. Proof of Theorem 4

For simplicity, we focus on the bivariate case (X and Y are each one-dimensional variables). The extension of the proof to the multivariate case is straight forward. We first work on mutual information, then show the similar arguments on the copula distances. To prove the theorem, we use Le Cam (1973)'s method to find the lower bound on the minimax risk of the estimating mutual information MI. To do this, we will use a more convenient form of Le Cam's method developed by Donoho and Liu (1991). Define the module of continuity of a functional T over the class $\mathbf{F}$ with respect to Hellinger distance as in equation (1.1) of Donoho and Liu (1991):

$$w(\varepsilon) = \sup\{|T(F_1) - T(F_2)| : H(F_1, F_2) \leq \varepsilon, F_i \in \mathbf{F}\}. \tag{13}$$

Here $H(F_1, F_2)$ denotes the Hellinger distance between F_1 and F_2. Then the minimax rate of convergence for estimating $T(F)$ over the class $\mathbf{F}$ is bounded below by $w(n^{-1/2})$.

We now look for a pair of density functions $c_1(u, v)$ and $c_2(u, v)$ on the unit square for distributions that are close in Hellinger distance but far away in their mutual information. This provides a lower bound on the module of continuity for mutual information MI over the class $\mathfrak{C}$, and hence leads to a lower bound on the minimax risk. We outline the proof next.

We first divide the unit square into three disjoint regions R_1, R_2 and R_3 with $R_1 \cup R_2 \cup R_3 = [0, 1]\times[0, 1]$. The first density function $c_1(u, v)$ puts probability masses δ, a and $1-a-\delta$ respectively on the regions R_1, R_2 and R_3 each uniformly. The a is an arbitrary small fixed value, for example, $a = 0.01$. For now, we take δ to be another small fixed value. The area of the region is chosen so that $c_1(u, v) = M$ on region R_2 and $c_1(u, v) = M^*$ on region R_1 for a very big M^*. The second density function $c_2(u, v)$, compared to $c_1(u, v)$, moves a small probability mass ε from R_1 to R_2. We will see that the Hellinger distance between c_1 and c_2 is of the same order as ε, but the change in MI is unbounded for big M^*. Hence module of continuity $w(\varepsilon)$ is unbounded for mutual information MI. Therefore the MI can not be consistently estimated over the class $\mathfrak{C}$.

Specifically, the region R_1 is chosen to be a narrow strip immediately above the diagonal, $R_1 = \{(u, v) : -\delta_1 < u - v < 0\}$; and R_2 is chosen to be a narrow strip immediately below the diagonal, $R_2 = \{(u, v) : 0 \leq u - v < \delta_2\}$. The remaining region is $R_3 = [0, 1]\times[0, 1]\setminus(R_1 \cup R_2)$. The values of δ_1 and δ_2 are chosen so that the areas of regions R_1 and R_2 are δ/M^* and a/M respectively. Then clearly $c_1(u, v) = M^*$ on R_1; $c_1(u, v) = M$ on R_2; $c_1(u, v) = (1-a-\delta)/(1-a/M-\delta/M^*)$ on R_3. And $c_2(u, v) = M^* - \varepsilon(M^*/\delta)$ on R_1; $c_2(u, v) = M + \varepsilon(M/a)$ on R_2; $c_2(u, v) = c_1(u, v)$ on R_3. See the Figure 12.

Then we have

$$\begin{aligned}
2H^2(c_1, c_2) &= \textstyle\int(\sqrt{c_2(u, v)} - \sqrt{c_1(u, v)})^2 du dv \\
&= (\sqrt{M^* - \varepsilon(M^*/\delta)} - \sqrt{M^*})^2\delta/M^* + (\sqrt{M + \varepsilon(M/a)} - \sqrt{M})^2 a/M \\
&= \delta(\sqrt{1 - \varepsilon/\delta} - 1)^2 + a(\sqrt{1 + \varepsilon/a} - 1)^2 \\
&= \delta(\varepsilon/2\delta)^2 + a(\varepsilon/2a)^2 + o(\varepsilon^2) \\
&= \varepsilon^2(\tfrac{1}{4\delta} + \tfrac{1}{4a}) + o(\varepsilon^2).
\end{aligned}$$

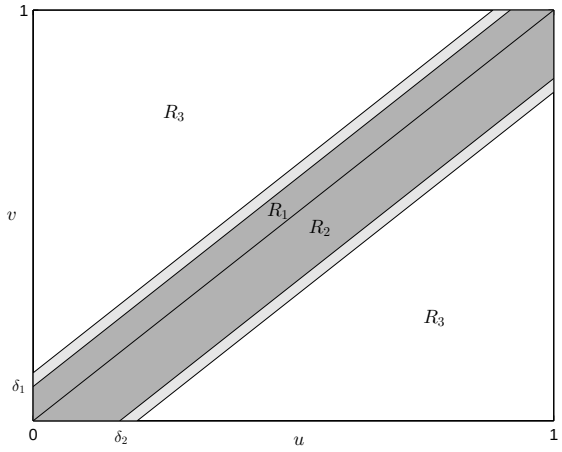


Figure 12: The plot shows the regions R_1, R_2 and R_3. The other two narrow strips neighboring R_1 and R_2 are for the continuity correction mentioned at the end of the proof.

Hence the Hellinger distance is of the same order as ε:

$$H(c_1, c_2) = \varepsilon\sqrt{\frac{1}{8\delta} + \frac{1}{8a}} + o(\varepsilon).$$

On the other hand, the difference in the mutual information is

$$\begin{aligned} & \mathrm{MI}(c_1) - \mathrm{MI}(c_2) \\ = \;& \delta\log(M^*) + a\log(M) - (\delta-\varepsilon)\log[M^* - \varepsilon(M^*/\delta)] - (a+\varepsilon)\log[M + \varepsilon(M/a)] \\ = \;& \varepsilon\log(M^*) - \varepsilon\log(M) - (\delta-\varepsilon)\log(1-\varepsilon/\delta) - (a+\varepsilon)\log(1+\varepsilon/a). \end{aligned} \tag{14}$$

Here M, δ and a are fixed constants. Hence when $M^* \to \infty$, this difference in MI also goes to ∞. For example, if we let $M^* = e^{1/(\varepsilon)^2}$, then the module of continuity $w(\varepsilon) \geq O(1/\varepsilon)$. That means, the rate of convergence is at least $O(w(n^{-1/2})) = O(n^{1/2}) \to \infty$. In other words, MI can not be consistently estimated.

Now, let us consider the $\mathrm{CD}_\alpha = \int_{I^2} |c(u,v) - 1|^\alpha du dv$, for $\alpha > 1$, where I^2 is the unit square.

$$\begin{aligned} & \mathrm{CD}_\alpha(c_1) - \mathrm{CD}_\alpha(c_2) \\ = \;& |M^*-1|^\alpha\delta/M^* + |M^{-}1|^\alpha a/M - |M^* - 1 - \varepsilon(M^*/\delta)|^\alpha\delta/M^* + |M - 1 + \varepsilon(M/a)|^\alpha a/M \\ = \;& [|M^*-1|^\alpha - |M^* - 1 - \varepsilon(M^*/\delta)|^\alpha]\delta/M^* + [|M^{-}1|^\alpha - |M - 1 + \varepsilon(M/a)|^\alpha]a/M \\ = \;& \alpha[(M^*-1)^{\alpha-1}M^*/\delta - (M-1)^{\alpha-1}M/\alpha]\varepsilon + o(\varepsilon^2). \end{aligned} \tag{15}$$

Again, M, δ and a are fixed constants. Hence when $M^* \to \infty$, this difference in $\mathrm{CD}_\alpha, \alpha > 1$ also goes to ∞. For example, if we let $M^* = (\varepsilon^{-2} + M^\alpha)^{\frac{1}{\alpha-1}} + 1$, then the module of continuity $w(\varepsilon) \geq O(1/\varepsilon)$. Note that $\alpha > 1$ is essential here. That means, the rate of convergence is at least $O(w(n^{-1/2})) = O(n^{1/2}) \to \infty$. In other words, $\mathrm{CD}_\alpha, \alpha > 1$ can not be consistently estimated.

The above outlines the main idea of the proof, ignoring some mathematical subtleties. One is that the example densities c_1 and c_2 are only piecewise continuous on the three regions, but not truly continuous as required for the class $\mathfrak{C}$. This can be easily remedied by connecting the three pieces linearly. Specifically we set the densities $c_i(u,v) = M$, $i = 1, 2$, on the boundary between R_1 and R_3, $\{(u,v) : u - v = -\delta_1\}$, and on the boundary between R_2 and R_3, $\{(u,v) : u - v = \delta_2\}$. Then we use two narrow strips within R_3, $\{(u,v) : -\delta_3 \le u - v \le -\delta_1\}$ and $\{(u,v) : \delta_2 \le u - v \le \delta_4\}$ to connect the constant $c_i(u,v)$ values on the rest of region R_3 with the boundary value $c_i(u,v) = M$ continuously through linear (in $u - v$) $c_i(u,v)$'s on the two strips that satisfies the Hölder condition (7) of the main text. By the Hölder condition, the connection can be made with strips of width at most $(M - 1 + a + \delta)/M_1$. This continuity modification does not affect the calculation of the difference $\text{MI}(c_1) - \text{MI}(c_2)$ or $\text{CD}_\alpha(c_1) - \text{CD}_\alpha(c_2)$ above as c_1 and c_2 only differ on regions R_1 and R_2. Within regions R_1 and R_2, the densities c_1 and c_2 can be further similarly connected continuously linearly in $u - v$. As there is no Hölder condition on A_M^c, the connection within R_1 and R_2 can be as steep as we want. Clearly the order obtained through above calculations will not change if we make these connections very steep so that their effect is negligible.

Another technical subtlety is that the c_1 and c_2 defined above are only densities on the unit square but not copula densities which require uniform marginal distributions. However, it is clear that the marginal densities for c_is are uniform over the interval $(\delta_3, 1 - \delta_4)$ and linear in the rest of interval near the two end points 0 and 1. The copulas densities c_i^*'s corresponding to c_i's can be calculated directly through Sklar's decomposition (1) in the main text. It is easy to see that the order for the module of continuity $w(\varepsilon)$ remains the same for using the corresponding copula densities c_i^*'s.

Appendix B. Proof of Theorem 5

The proof is almost the same as the proof for MI, but need some modification of the pair of least favorable c_1 and c_2 above. The small difference in Hellinger distance of c_1 and c_2 can lead to unbounded difference in $MI(c_1)$ and $MI(c_2)$ since MI is unbounded. After the transformation $MIcor = \sqrt{1 - e^{-2MI}}$ is bounded. The difference between $MIcor(c_1)$ and $MIcor(c_2)$ in the above example is actually small since the MI are big for both c_1 and c_2 (leading to corresponding $MIcor$s close to zero). However, $MIcor$ is also very hard to estimate over the class $\mathfrak{C}$. To see this, we follow the same reasoning above but modify the example of c_1 and c_2. First, we notice that for any pair of densities c_1 and c_2,

$$\begin{aligned} |MIcor(c_1) - MIcor(c_2)| &= |\sqrt{1 - e^{-2MI(c_1)}} - \sqrt{1 - e^{-2MI(c_2)}}| \\ &= |\tfrac{[1-e^{-2MI(c_1)}]-[1-e^{-2MI(c_2)}]}{\sqrt{1-e^{-2MI(c_1)}}+\sqrt{1-e^{-2MI(c_2)}}}| \\ &\ge \tfrac{1}{2}|e^{-2MI(c_1)} - e^{-2MI(c_2)}| \\ &= \tfrac{1}{2}e^{-2MI(c_1)}|1 - e^{-2[MI(c_1)-MI(c_2)]}|. \end{aligned}$$

For the difference $MIcor(c_1) - MIcor(c_2)$ to be the same order of the difference $MI(c_1) - MI(c_2)$, we need to set $MI(c_1)$ at constant order when $\varepsilon \to 0$.

Therefore, we modify the above c_1 to have probability mass $\delta = 2\varepsilon$ in region R_1, varying with the ε value instead of fixed as before. And we set $M^* = e^{1/\varepsilon}$, leading to

$$\begin{aligned} & MI(c_1) \\ = \; & \delta \log(M^*) + a\log(M) + (1-a-\delta)\log[(1-a-\delta)/(1-a/M-\delta/M^*)] \\ = \; & 2 + a\log(M) + (1-a-2\varepsilon)\log[(1-a-2\varepsilon)/(1-a/M-2\varepsilon e^{-1/\varepsilon})], \end{aligned}$$

which converges to a fixed constant $a_1 = 2 + a\log(M) + (1-a)\log[(1-a)/(1-a/M)]$ as $\varepsilon \to 0$. Using (14), recall that $\delta = 2\varepsilon$ and $M^* = e^{1/\varepsilon}$, we have

$$\begin{aligned} & MI(c_1) - MI(c_2) \\ = \; & \varepsilon\log(M^*) - \varepsilon\log(M) - (\delta-\varepsilon)\log(1-\varepsilon/\delta) - (a+\varepsilon)\log(1+\varepsilon/a) \\ = \; & 1 - \varepsilon\log(M) - \varepsilon\log(1/2) - (a+\varepsilon)\log(1+\varepsilon/a), \end{aligned}$$

which converges to 1 as $\varepsilon \to 0$. Hence we have

$$\lim_{\varepsilon\to 0} w(\varepsilon) \geq \lim_{\varepsilon\to 0} \frac{1}{2} e^{-2MI(c_1)} |1 - e^{-2[MI(c_1)-MI(c_2)]}| = \frac{1}{2} e^{-2a_1}(1 - e^{-2(1)}),$$

a positive constant $a_2 = e^{-2a_1}(1-e^{-2})/2$. Therefore, $MIcor$ can not be estimated consistently over the class $\mathfrak{C}$ either.

Appendix C. Proof of Theorem 6

The first two terms in (11) corresponds to bias and standard deviation of kernel density estimation when the copula density is bounded. When the copula density is unbounded, the kernel density estimation $\hat{c}(\mathbf{Z})$ is not consistent. However, a smaller order $O(\frac{1}{nh^d})$ term bounds the overall error contribution to $\widehat{RCD}$ resulting from $\hat{c}(\mathbf{Z})$ in the unbounded copula density region.

Let $M_2 = \frac{M+1}{2}$, $A_{M_2} = \{\mathbf{Z} | c(\mathbf{Z}) \leq M_2\}$, $T_1(c) = \int_{A_{M_2}} (1 - c(\mathbf{Z}))_+ d\mathbf{Z}$, $T_2(c) = \int_{A^c_{M_2}} (1 - c(\mathbf{Z}))_+ d\mathbf{Z}$, $RCD = T_1(c) + T_2(c)$, and $\widehat{RCD} = T_1(\hat{c}) + T_2(\hat{c})$

Firstly, we consider the region A_{M_2} with bounded copula density. Here we calculate the bias and variance of the kernel density estimator using standard methods first.

$$\bar{c}_n(\mathbf{Z}) = E[\hat{c}_{kde}(\mathbf{Z})] = \frac{1}{h^d} \int K(\frac{\mathbf{z}-\mathbf{Z}}{h}) c(\mathbf{z}) d\mathbf{z} = \int K(s) c(\mathbf{Z} + sh) ds.$$

Hence

$$\begin{aligned} |Bias(\mathbf{Z})| = |\textstyle\int K(s)c(\mathbf{Z}+sh)ds - c(\mathbf{Z})| \; & \leq \textstyle\int_{\mathbb{B}_0} K(s)|c(\mathbf{Z}+sh) - c(\mathbf{Z})|ds \\ & \leq \textstyle\int_{\mathbb{B}_0} K(s) M_1 h ds \\ & = M_1 h. \end{aligned} \tag{16}$$

$$\begin{aligned} |Var(\mathbf{Z})| \; & = \tfrac{1}{n} Var[\tfrac{1}{h^d} K(\tfrac{\mathbf{Z}_1 - \mathbf{Z}}{h})] \; \leq \tfrac{1}{n} E[\tfrac{1}{h^{2d}} K^2(\tfrac{\mathbf{Z}_1-\mathbf{Z}}{h})] \\ & = \tfrac{1}{nh^d} \textstyle\int_{\mathbb{B}_0} K^2(s) c(\mathbf{Z}+sh) ds \\ & \leq \tfrac{1}{nh^d} \textstyle\int_{\mathbb{B}_0} K^2(s)[c(\mathbf{Z}) + M_1 h] ds \\ & = \tfrac{\mu_2^2}{nh^d}[c(\mathbf{Z}) + M_1 h]. \end{aligned} \tag{17}$$

Hence the integrated mean square error of the density estimator $\hat{c}_n(\mathbf{Z})$ over regions A_{M_2} is

$$\begin{aligned} IMSE(\mathbf{Z}) &= \int_{A_{M_2}} [Bias^2(\mathbf{Z}) + Var(\mathbf{Z})] d\mathbf{Z} \\ &\leq \int_{A_{M_2}} [M_1^2 h^2 + \tfrac{\mu_2^2}{nh^d}[c(\mathbf{Z}) + M_1 h] d\mathbf{Z} \ \leq M_1^2 h^2 + \tfrac{\mu_2^2}{nh^d}[1 + M_1 h] \\ &\leq M_1^2 h^2 + \tfrac{2\mu_2^2}{nh^d} \end{aligned} \tag{18}$$

Hence the error of $\widehat{RCD}$ on A_{M_2} is bounded by

$$\begin{aligned} E|T_1(\hat{c}) - T_1(c)| &\leq E \int_{A_{M_2}} |(1 - \hat{c}_n(\mathbf{Z}))_+ - (1 - c(\mathbf{Z}))_+| d\mathbf{Z} \\ &\leq E \int_{A_{M_2}} |\hat{c}_n(\mathbf{Z}) - c(\mathbf{Z})| d\mathbf{Z} \\ &\leq \sqrt{E \int_{A_{M_2}} (\hat{c}_n(\mathbf{Z}) - c(\mathbf{Z}))^2 d\mathbf{Z}} \\ &\leq \sqrt{d^2 M_1^2 h^2 + \tfrac{2\mu_2^2}{nh^d}} \\ &\leq dM_1 h + \sqrt{2}\mu_2 (\tfrac{1}{nh^d})^{1/2}. \end{aligned}$$

Now we consider the region $A_{M_2}^c$ with unbounded copula density. For $\mathbf{Z} \in A_{M_2}^c$, $\hat{c}(\mathbf{Z})$ does not have a finite variance bound in (17). But we can bound the variance by the expectation $\bar{c}_n(\mathbf{Z}) = E[\hat{c}_n(\mathbf{Z})]$. Let $M_3 = \frac{M_2+1}{2}$, when h small, $\mathbf{Z} \in A_{M_2}^c$ implies $\mathbf{Z} + sh \in A_{M_3}^c$. Hence

$$|Var(\mathbf{Z})| \leq \frac{1}{nh^d} \int_{\mathbb{B}_0} K^2(s) c(\mathbf{Z} + sh) ds \leq \frac{M_K}{nh^d} \int_{\mathbb{B}_0} K(s) c(\mathbf{Z} + sh) ds = \frac{M_K}{nh^d} \bar{c}_n(\mathbf{Z})$$

Using Chebyshev's inequality,

$$\begin{aligned} E[1_{\{\hat{c}_n(\mathbf{Z})<1\}}] = P(\hat{c}_n(\mathbf{Z}) < 1) &\leq P(|\bar{c}_n(\mathbf{Z}) - \hat{c}_n(\mathbf{Z})| > \bar{c}_n(\mathbf{Z}) - 1) \\ &\leq \frac{Var[\hat{c}_n(\mathbf{Z})]}{[\bar{c}_n(\mathbf{Z}) - 1]^2} \\ &\leq \frac{M_K}{nh^d} \frac{\bar{c}_n(\mathbf{Z})}{[\bar{c}_n(\mathbf{Z}) - 1]^2} \ \leq \frac{M_K M_4}{nh^d} \end{aligned}$$

where $M_4 = \frac{M_3}{(M_3-1)^2}$.

Hence the error of $\widehat{RCD}$ on $A_{M_2}^c$ is bounded by

$$E|T_2(\hat{c}) - T_2(c)| = E[T_2(\hat{c})] \leq \int_{A_{M_2}^c} E[1_{\{\hat{c}_n(\mathbf{Z})<1\}}] d\mathbf{Z} \leq \frac{M_K M_4}{nh^d}$$

Combining the above results:

$$E[|\widehat{RCD} - RCD|] \leq M_1 h + \frac{\sqrt{2}\mu_2}{\sqrt{n} h^{d/2}} + \frac{M_K M_4}{nh^d}. \tag{19}$$

This finishes the proof.

Note that we can use any L_p norm ($1 \leq p \leq \infty$) in the Hölder condition: equation (7). The kernel K is then assumed to have support in the unit ball $\mathbb{B}_0$ corresponding to that L_p norm. The proof remains exactly the same. We in fact will use L_∞ norm in our estimator for computational simplicity. In that case, the unit ball $\mathbb{B}_0 = \{\mathbf{Z} : \|\mathbf{Z}\|_{l_\infty} \leq 1\}$ is in fact the d-dimensional cube.

Appendix D. Proof of Theorem 7

Here for $\widehat{RCD} = RCD(\hat{c})$ we use the k-NN estimator (Loftsgaarden and Quesenberry, 1965) of the copula density

$$\hat{c}_{knn}(\mathbf{Z}) = \frac{\frac{k(n)}{n}}{A_{r(k(n),n),\mathbf{Z}}}, \tag{20}$$

where $\mathbf{Z}_1, \mathbf{Z}_2, \cdots, \mathbf{Z}_n$ are the copula based observations, $r(k(n), n)$ is the distance from $\mathbf{Z}$ to the k^{th} closest of $\mathbf{Z}_1, \mathbf{Z}_2, \cdots, \mathbf{Z}_n$ and $A_{r(k(n),n),\mathbf{Z}}$ is the volume of the d-dimensional hyper-ball with radius $r(k(n), n)$.

In the following, without ambiguity, we denote $r(k(n), n)$ by r, and $k(n)$ by k. Hence the volume $A_{r,\mathbf{Z}}$ is $v_d \cdot r^d$, where v_d is the volume of the d-dimensional unit ball $\mathbb{B}_0$. And $\hat{c}_{knn}(\mathbf{Z}) = k/(v_d n r^d)$. For l_2 norm, $v_d = \pi^{d/2}/\Gamma[(d+2)/2]$ where $\Gamma(\cdot)$ denotes the Gamma function.

Moore and Yackel (1977a) showed that, for bounded densities, there is equivalence between the consistency of the KDE density estimator and the consistency of the k-NN estimator. To cite the results of (Moore and Yackel, 1977a), we assume a slightly stronger version of the Hölder condition than (7). That is, we assume that c also has bounded continuous *second order* derivative in A_M. Let $Q(\mathbf{Z}) = tr[\frac{\partial^2 c(\mathbf{Z})}{\partial \mathbf{Z}^2}]$ denote the trace of the Hessian matrix of copula density $c(\mathbf{Z})$. For the d-dimensional vector $\mathbf{Z} = (z_1, ..., z_d)$, the Hessian matrix $\partial^2 c(\mathbf{Z})/\partial \mathbf{Z}^2$ has entries

$$[\frac{\partial^2 c(\mathbf{Z})}{\partial \mathbf{Z}^2}]_{ij} = \frac{\partial^2 c(\mathbf{Z})}{\partial z_i \partial z_j}.$$

Then we rewrite the error bound in Theorem 7 explicitly as

$$\sup_{C \in \mathfrak{C}} E[|\widehat{RCD} - RCD|] \leq 2\bar{Q}(\frac{k}{n\epsilon})^{\frac{2}{d}} + \frac{2M}{\sqrt{k}} + 2\epsilon,$$

where $\bar{Q} = \frac{1}{2(d+2)\pi}\Gamma^{2/d}(\frac{d+2}{2}) \sup_{\mathbf{Z} \in A_M} Q(\mathbf{Z})$, and $\epsilon = \epsilon(n)$ is any sequence converging to 0 slower than k/n. We suppress the n from the notation in ϵ without ambiguity as in k and r above.

We shall use the following asymptotic results on k-NN density estimator in Mack and Rosenblatt (1979). Denote $\tilde{Q}(\mathbf{Z}) = \frac{1}{2(d+2)\pi}\Gamma^{2/d}(\frac{d+2}{2})Q(\mathbf{Z})$. Then

$$\begin{aligned} Bias[\hat{c}_{knn}(\mathbf{Z})] &= \frac{\tilde{Q}(\mathbf{Z})}{c(\mathbf{Z})^{2/d}}(\frac{k}{n})^{2/d} + O(\frac{c(\mathbf{Z})}{k}) + o((\frac{k}{n})^{2/d}), \\ Var[\hat{c}_{knn}(\mathbf{Z})] &= \frac{c^2(\mathbf{Z})}{k} + o(\frac{1}{k}). \end{aligned} \tag{21}$$

These expressions provide control on the error contribution of $\hat{c}(\mathbf{Z})$ to $\widehat{RCD}$ when $c(\mathbf{Z})$ is bounded both from above and from below. Similar to the proof of KDE-based $\widehat{RCD}$, we prove that the error contribution to $\widehat{RCD}$ from the big copula density region is of a smaller order $O(1/k)$. Different from the KDE, the k-NN density estimator also does not have finite bias bound in (21) when the copula density $c(\mathbf{Z})$ is not bounded below. Therefore, we also need to control the error contribution to $\widehat{RCD}$ from the small ($< \epsilon$) copula density region separately.

As before, let M_2 be a constant between 1 and M, say, $M_2 = \frac{M+1}{2}$. We now separate the three regions by copula density: $A^c_{M_2} = \{\mathbf{Z} : c(\mathbf{Z}) > M_2\}$ (big), $A_{M_2,\epsilon} = \{\mathbf{Z} : \epsilon \leq c(\mathbf{Z}) \leq M_2\}$ (middle) and $A_\epsilon = \{\mathbf{Z} : c(\mathbf{Z}) < \epsilon\}$ (small). Then we can separate RCD into three components

$RCD = T_1(c) + T_2(c) + T_3(c)$: $T_1(c) = \int_{A_{M_2^c}} [1 - c(\mathbf{Z})]_+ d\mathbf{Z}$, $T_2(c) = \int_{A_{M_2,\epsilon}} [1 - c(\mathbf{Z})]_+ d\mathbf{Z}$ and $T_3(c) = \int_{A_\epsilon} [1 - c(\mathbf{Z})]_+ d\mathbf{Z}$.

Firstly, we look at the error bound on $A^c_{M_2}$, the region of big copula density. Similar to the KDE, the error in $\hat{c}_{knn}(\mathbf{Z})$ can be arbitrarily large for $\mathbf{Z} \in A^c_{M_2}$. However, the error only leads to the error in $\widehat{RCD}$ if $\hat{c}_{knn}(\mathbf{Z}) < 1$. From equation (20), $\hat{c}_{knn}(\mathbf{Z}) < 1$ if and only if

$$r > (\frac{k}{nv_d})^{1/d} \stackrel{\text{def}}{=} \bar{r}.$$

This occurs when at most $k-1$ of observations $\mathbf{Z}_1, \mathbf{Z}_2, \cdots, \mathbf{Z}_n$ fall into the ball $\mathbb{B}(\mathbf{Z}; \bar{r})$ which is centered at $\mathbf{Z}$ with radius $\bar{r}$.

Let $\bar{N}(\mathbf{Z})$ denotes the number of observations falling into $\mathbb{B}(\mathbf{Z}; \bar{r})$. Then $\bar{N}(\mathbf{Z})$ follows a binomial distribution with mean $n\bar{p}$, where $\bar{p} = \int_{\mathbb{B}(\mathbf{Z};\bar{r})} c(\mathbf{z}) d\mathbf{z}$. Since $k/n \to 0$, $\bar{r} \to 0$. Hence $M_1\bar{r} < (M_2 - 1)/2$ when n is large enough. Then the whole ball $\mathbb{B}(\mathbf{Z}; \bar{r})$ is contained in $A^c_{M_3}$ with $M_3 = (M_2 + 1)/2$ as before. Hence $\bar{p} = \int_{\mathbb{B}(\mathbf{Z};\bar{r})} c(\mathbf{z}) d\mathbf{z} \geq M_3 v_d \bar{r}^d = M_3 k/n$. Using Chebyshev's inequality,

$$\begin{aligned} Pr[\hat{c}_{knn}(\mathbf{Z}) < 1] &= E[\mathbb{1}\{\bar{N}(\mathbf{Z}) < k\}] \leq \tfrac{Var[\bar{N}(\mathbf{Z})]}{\{E[\bar{N}(\mathbf{Z})] - k\}^2} = \tfrac{n\bar{p}(1-\bar{p})}{(n\bar{p}-k)^2} \\ &\leq \tfrac{1}{n\bar{p}[1-k/(n\bar{p})]^2} \leq \tfrac{1}{M_3 k[1-1/M_3]^2} \\ &= O(\tfrac{1}{k}). \end{aligned}$$

Hence

$$E|T_1(c) - T_1(\hat{c}_{knn})| = \int_{A^c_{M_2}} E[1_{\{\hat{c}_{knn}(\mathbf{Z})<1\}}] d\mathbf{Z} \leq \frac{1}{M_3 k[1 - 1/M_3]^2} = O(\frac{1}{k}).$$

Secondly, we look at the error bound on $A_{M_2,\epsilon}$, the region of middle copula density. Using (21), for $\mathbf{Z} \in A_{M_2,\epsilon}$, the mean squared error of $\hat{c}_{knn}(\mathbf{Z})$ is

$$\begin{aligned} E[(\hat{c}_n(\mathbf{Z}) - c(\mathbf{Z}))^2] &= bias^2(\mathbf{Z}) + Var(\mathbf{Z}) \\ &= [\frac{\tilde{Q}(\mathbf{Z})}{c(\mathbf{Z})^{2/d}} (\frac{k}{n})^{2/d}]^2 + \frac{c^2(\mathbf{Z})}{k} + o((\frac{k}{n})^{4/d} + \frac{1}{k}) \\ &\leq (\frac{\bar{Q}^2}{\epsilon^{4/d}})(\frac{k}{n})^{4/d} + \frac{M_2^2}{k} + o((\frac{k}{n})^{4/d} + \frac{1}{k}). \end{aligned}$$

Hence

$$E|\hat{c}_{knn}(\mathbf{Z}) - c(\mathbf{Z})| \leq \sqrt{E[(\hat{c}_n(\mathbf{Z}) - c(\mathbf{Z}))^2]} \leq \sqrt{2}[\bar{Q}(\frac{k}{n\epsilon})^{2/d} + \frac{M_2}{\sqrt{k}}][1 + o(1)].$$

We get

$$E[T_2(\hat{c}_{knn}) - T_2(c)] \leq E[\int_{A_{M_2,\epsilon}} |\hat{c}_{knn}(\mathbf{Z}) - c(\mathbf{Z})| d\mathbf{Z}] \leq \sqrt{2}[\bar{Q}(\frac{k}{n\epsilon})^{2/d} + \frac{M_2}{\sqrt{k}}][1 + o(1)]. \quad (22)$$

Thirdly, we look at the error bound on A_ϵ, the region of small copula density. From equation (20), $\hat{c}_{knn}(\mathbf{Z}) \geq 2\epsilon$ if and only if

$$r \leq (\frac{k}{n 2\epsilon v_d})^{1/d} \stackrel{\text{def}}{=} r^*.$$

This occurs when at least k of observations $\mathbf{Z}_1, \mathbf{Z}_2, \cdots, \mathbf{Z}_n$ fall into the ball $\mathbb{B}(\mathbf{Z}; r^*)$. Since $k/(n\epsilon) \to 0$, $r^* \to 0$.

Let $N^*(\mathbf{Z})$ denotes the number of observations falling into $\mathbb{B}(\mathbf{Z}; r^*)$. Then $\bar{N}(\mathbf{Z})$ follows a binomial distribution with mean np^*, where $p^* = \int_{\mathbb{B}(\mathbf{Z};r^*)} c(\mathbf{z})d\mathbf{z}$.

Using Taylor expansion, we have (from last line page 228 in Biau et al. (2011))

$$\int_{\mathbb{B}(\mathbf{Z};r)} c(\mathbf{z})d\mathbf{z} = c(\mathbf{Z})v_d r^d + \tilde{Q}(\mathbf{Z})v_d r^{d+2} + o(r^{d+2}).$$

Therefore, using $r^* \to 0$, we have $p^* = c(\mathbf{Z})v_d(r^*)^d[1 + o(1)] \leq \epsilon v_d(r^*)^d[1 + o(1)]$ which converges to $k/(2n)$. Hence for n big, $p^* < 0.6k/n$. Using Chebyshev's inequality,

$$\begin{aligned} Pr[\hat{c}_{knn}(\mathbf{Z}) > 2\epsilon] &= E[\mathbb{1}\{N^*(\mathbf{Z}) < k\}] \leq \tfrac{Var[N^*(\mathbf{Z})]}{[k - E(N^*(\mathbf{Z}))]^2} = \tfrac{np^*(1-p^*)}{(k-np^*)^2} \\ &\leq \tfrac{0.6k}{(0.4k)^2} < \tfrac{3}{k} \\ &= O(\tfrac{1}{k}). \end{aligned}$$

Since $c(\mathbf{Z}) \leq \epsilon$, if $\hat{c}_{knn}(\mathbf{Z}) \leq 2\epsilon$, then $|\hat{c}_{knn}(\mathbf{Z}) - c(\mathbf{Z})| \leq 2\epsilon$. Hence

$$\begin{aligned} E|T_3(c) - T_3(\hat{c}_{knn})| &\leq \textstyle\int_{A_\epsilon} E|\hat{c}_{knn}(\mathbf{Z}) - c(\mathbf{Z})|d\mathbf{Z} \leq \int_{A_\epsilon}\{2\epsilon + Pr[\hat{c}_{knn}(\mathbf{Z}) > 2\epsilon]\}d\mathbf{Z} \\ &< 2\epsilon + \tfrac{3}{k} = O(\epsilon + \tfrac{1}{k}). \end{aligned}$$

Finally, when combining the three parts, the terms $O(1/k) = o(1/\sqrt{k}) < (2-\sqrt{2})/\sqrt{k}$. Hence we arrive at

$$\sup_{C \in \mathfrak{C}} E[|\widehat{RCD} - RCD|] \leq 2[\bar{Q}(\frac{k}{n\epsilon})^{2/d} + \frac{M_2}{\sqrt{k}} + \epsilon], \tag{23}$$

which finished the proof.

Note that we can use other l_p norms, which changes the v_d in the proof to the volume of the unit ball under the corresponding norm. The rate does not change.

We can also prove the consistency under Hölder condition without assuming continuous second derivatives. However, that involve tedious derivation of bias and variance bounds similar to (21) for k-NN density estimators. We provide the simple proof here by citing (21) from Mack and Rosenblatt (1979).

To minimize the error bound in (12), we get $\epsilon = (k/n)^{2/(d+2)}$ and $k = n^{4/(d+6)}$. So in bivariate ($d = 2$) case, we take $k = O(n^{1/2})$. Taking k below the $n^{4/(d+6)}$ rate will make the $O(1/\sqrt{k})$ term dominant in the error bound. In that case, the asymptotic results on the k-NN density estimation states that $\sqrt{k}[\hat{c}(\mathbf{Z}) - c(\mathbf{Z})]/c(\mathbf{Z})$ converge to a standard Gaussian distribution. Then $\sqrt{k}[\widehat{RCD} - RCD]$ converges to an integral of a Gaussian process.

Appendix E. Selection of Tuning Parameter in the Practical Estimator

For a practical estimator for $\widehat{RCD}$, we need to decide the bandwidth in KDE-based estimator or the number of neighbors k in KNN-based estimator. Theorem 4 and Theorem 5 provides the rates. For bivariate case, $h = O(n^{-1/4})$ and $k = O(\sqrt{n})$. To decide the constant coefficient, we used empirical simulations.

First, for KDE estimators, we tested $\widehat{RCD}$ on nine functions (listed in Table 11) with various levels of additive noises. Four sample sizes of $n = 10^2$, 10^3, 10^4 and 10^5 are used.Figure 13 plots

the simulation results using $h = 0.25n^{-1/4}$. We can see that the performance of $\widehat{RCD}$ improves as sample size increases, and gives very accurate estimates for RCD under big sample sizes. For illustration, we showed the plots with bandwidth $h = 0.1n^{-1/4}$ and $h = 0.5n^{-1/4}$ in Figure 14 and Figure 15 respectively. Those bandwidth choices are clearly either too small ($h = 0.1n^{-1/4}$ estimator overshoot in several cases when RCD is small) or too big ($h = 0.5n^{-1/4}$ estimator converges slowly when RCD is large). Hence the bandwidth $h = 0.25n^{-1/4}$ is a good choice.

A	Linear	$y = x$
B	Quadratic	$y = x^2$
C	Square Root	$y = \sqrt{x}$
D	Cubic	$y = x^3$
E	Centered Cubic	$y = 4(x - 1/2)^3$
F	Centered Quadratic	$y = 4x(1 - x)$
G	Cosine (Period 1)	$y = [\cos(2\pi x) + 1]/2$
H	Circle	$(x - 1/2)^2 + y^2 = 1/4$
I	Cross	$y = \pm(x - 1/2)$

Table 11: The function relationships used in Figures 13 - 18.

According to the equivalence results between the KDE and the KNN estimator by (Moore and Yackel, 1977b), the k in the KNN density estimation corresponds to the bandwidth in KDE estimator as $c(\mathbf{z})(2h)^2 = k/n$. As the mean of copula density $c(\alpha)$ is one, $h = 0.25n^{-1/4}$ corresponds to $k = n(2h)^2 = 0.25\sqrt{n}$. The simulation results for KNN-based $\widehat{RCD}$ with $k = 0.25\sqrt{n}$, $k = 0.1\sqrt{n}$ and $k = 0.5\sqrt{n}$ are plotted in Figures 16 - 18. Similar pattern as in KDE-based estimator are observed. Hence we propose the practical KNN-based $\widehat{RCD}$ to use $k = 0.25\sqrt{n}$.

Furthermore, we also checked the KNN-based $\widehat{RCD}$ on the mixture noise setting used in definition 2: a proportion (p) of deterministic function is hidden in independent continuous noise. Six types of deterministic function are used, as listed in Table 12. When $n = 5000$, the $\widehat{RCD}$ is close to the true value p in the simulations. And compared to the two choices of $k = 0.1\sqrt{n}$ and $k = 0.5\sqrt{n}$, $k = 0.25\sqrt{n}$ provide a good balance of approximating the true values when RCD is small or large.

A	Linear	$y = x$
B	Centered Quadratic	$y = 4(x - 1/2)^2$
C	Cosine	$y = \cos(4\pi x)$
D	Cross	$y = \pm x 1_{\{0 \le x \le 1\}}$
E	Circle	$(2x - 1)^2 + y^2 = 1$
F	Cross 2	$y = \pm(x - 1/2)1_{\{0 \le x \le 1\}}$

Table 12: The function relationships used in Figures 19.

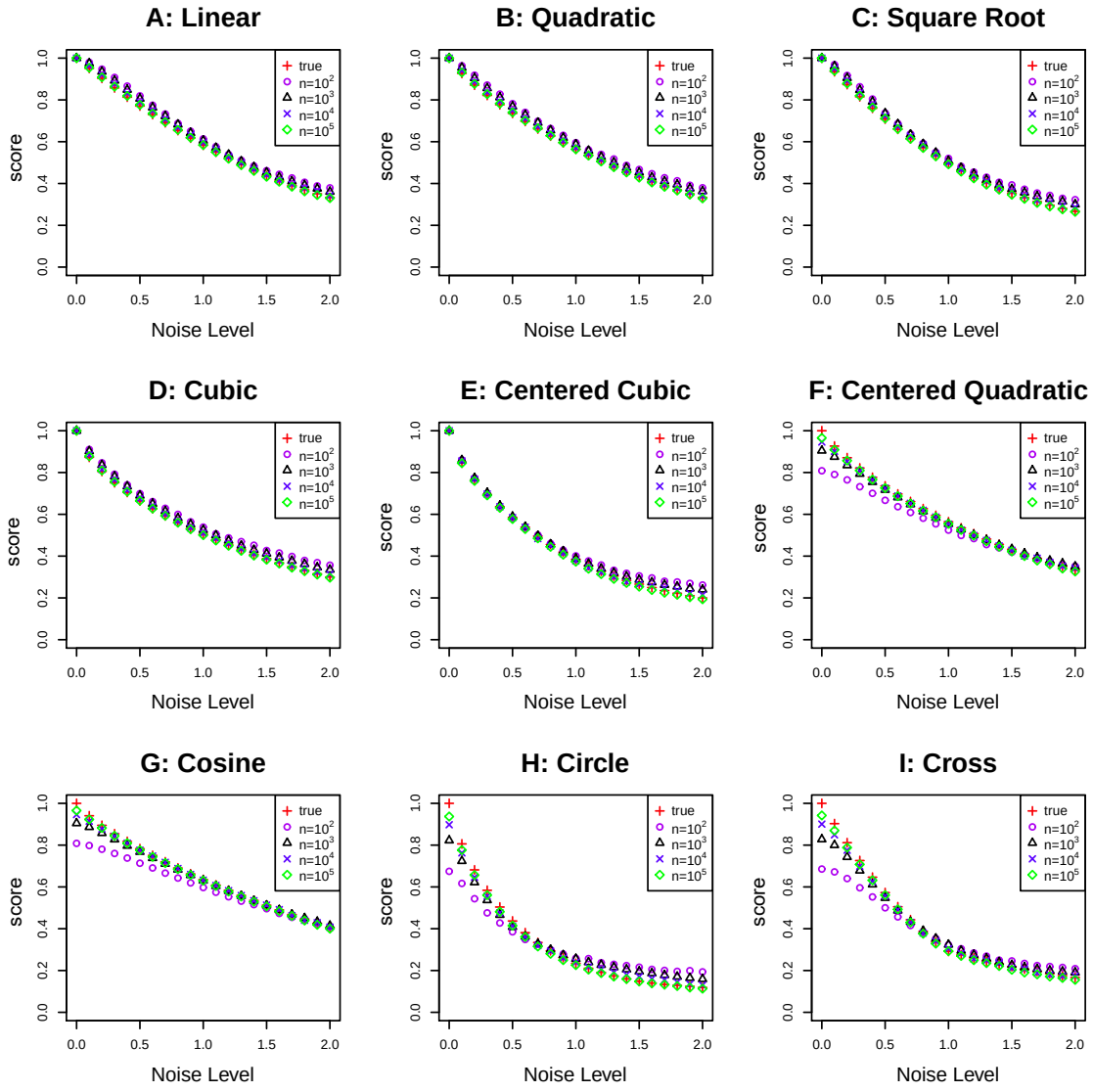


Figure 13: The comparison of RCD with its estimated values under different sample sizes. This estimator uses the square kernel density estimator with bandwidth $h = 0.25n^{-1/4}$.

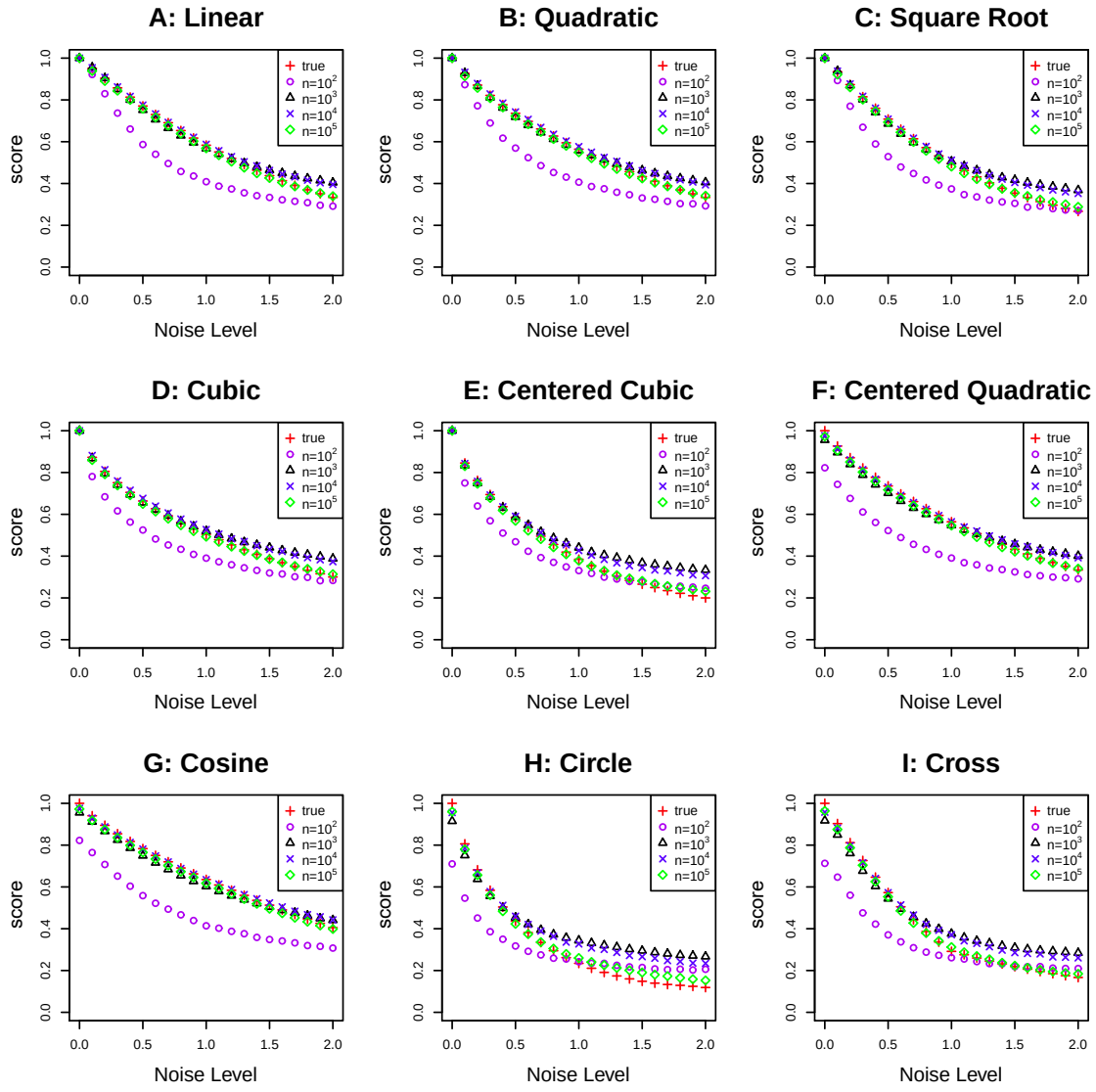


Figure 14: The comparison of RCD with its estimated values under different sample sizes. This estimator uses the square kernel density estimator with bandwidth $h = 0.1n^{-1/4}$.

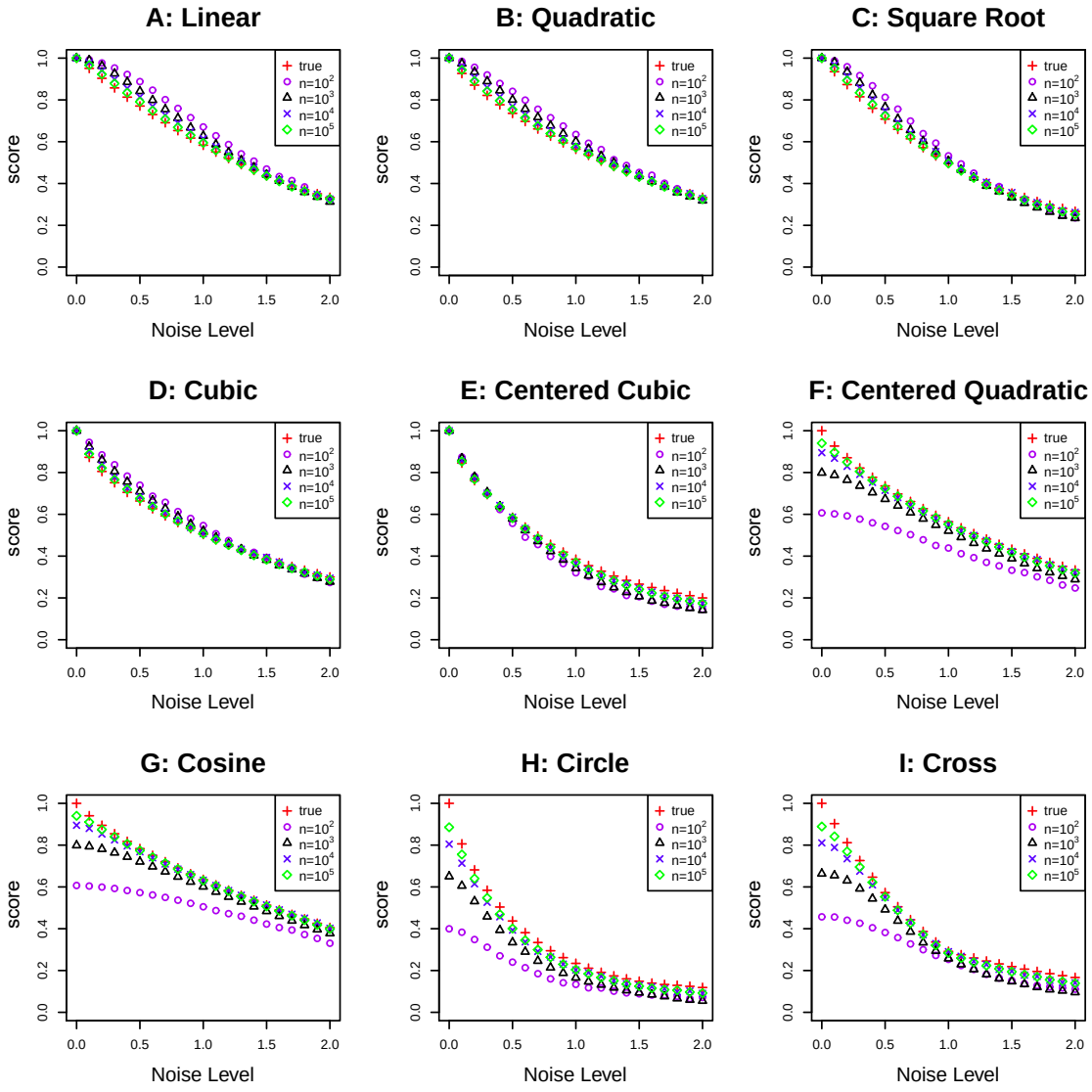


Figure 15: The comparison of RCD with its estimated values under different sample sizes. This estimator uses the square kernel density estimator with bandwidth $h = 0.5n^{-1/4}$.

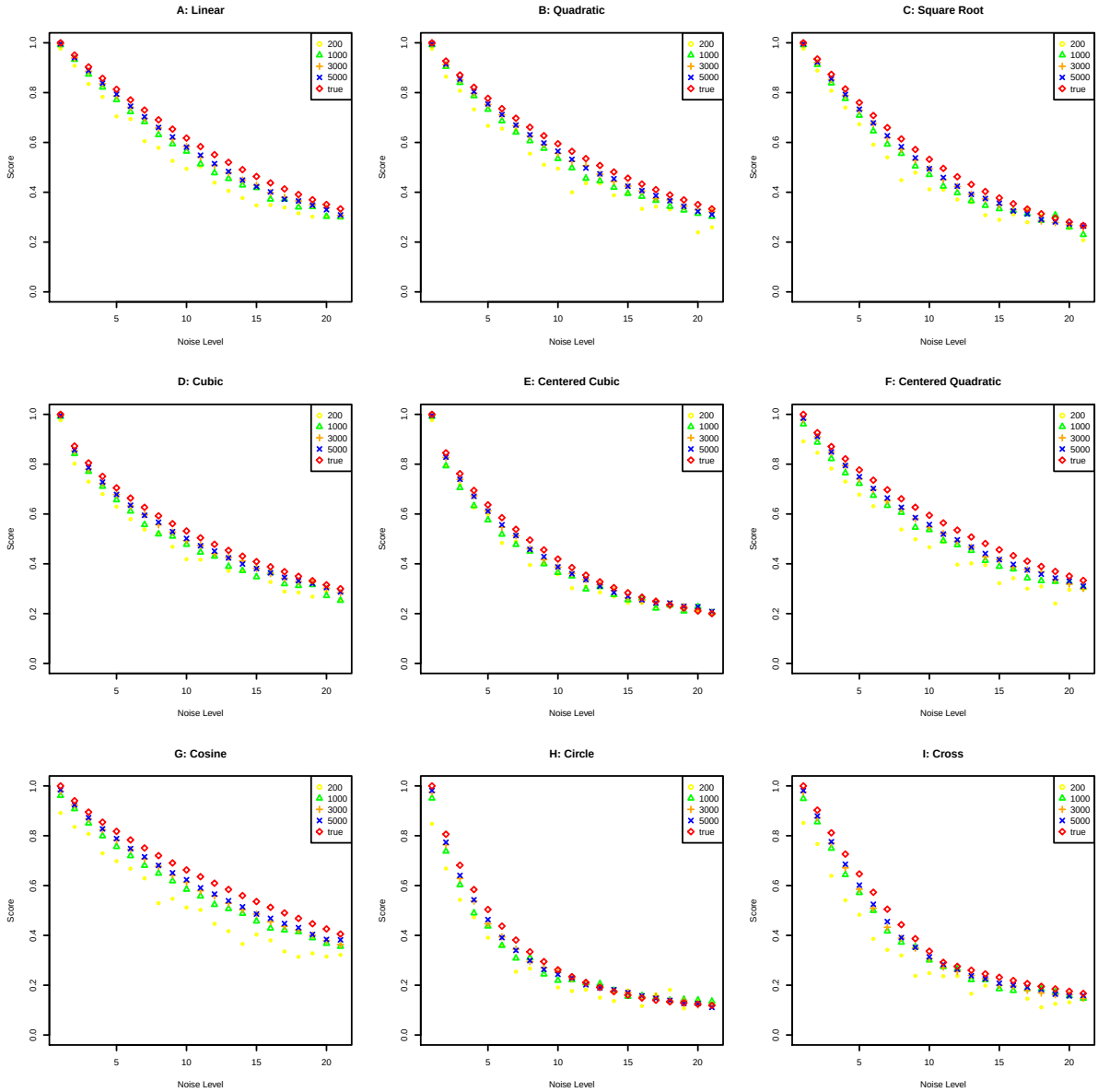


Figure 16: Additive noise with $k = 0.25\sqrt{n}$.

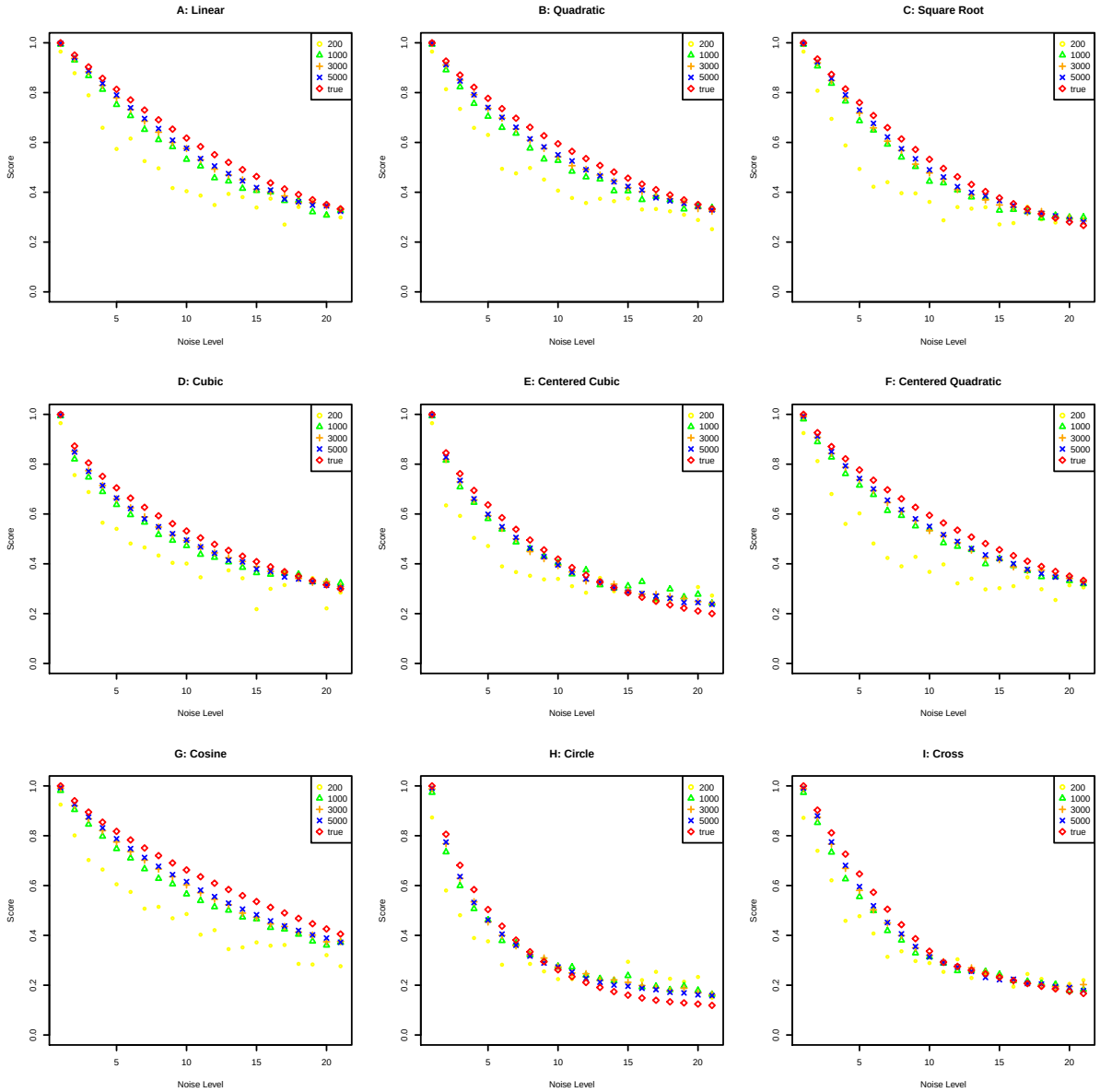


Figure 17: Additive noise with $k = 0.1\sqrt{n}$.

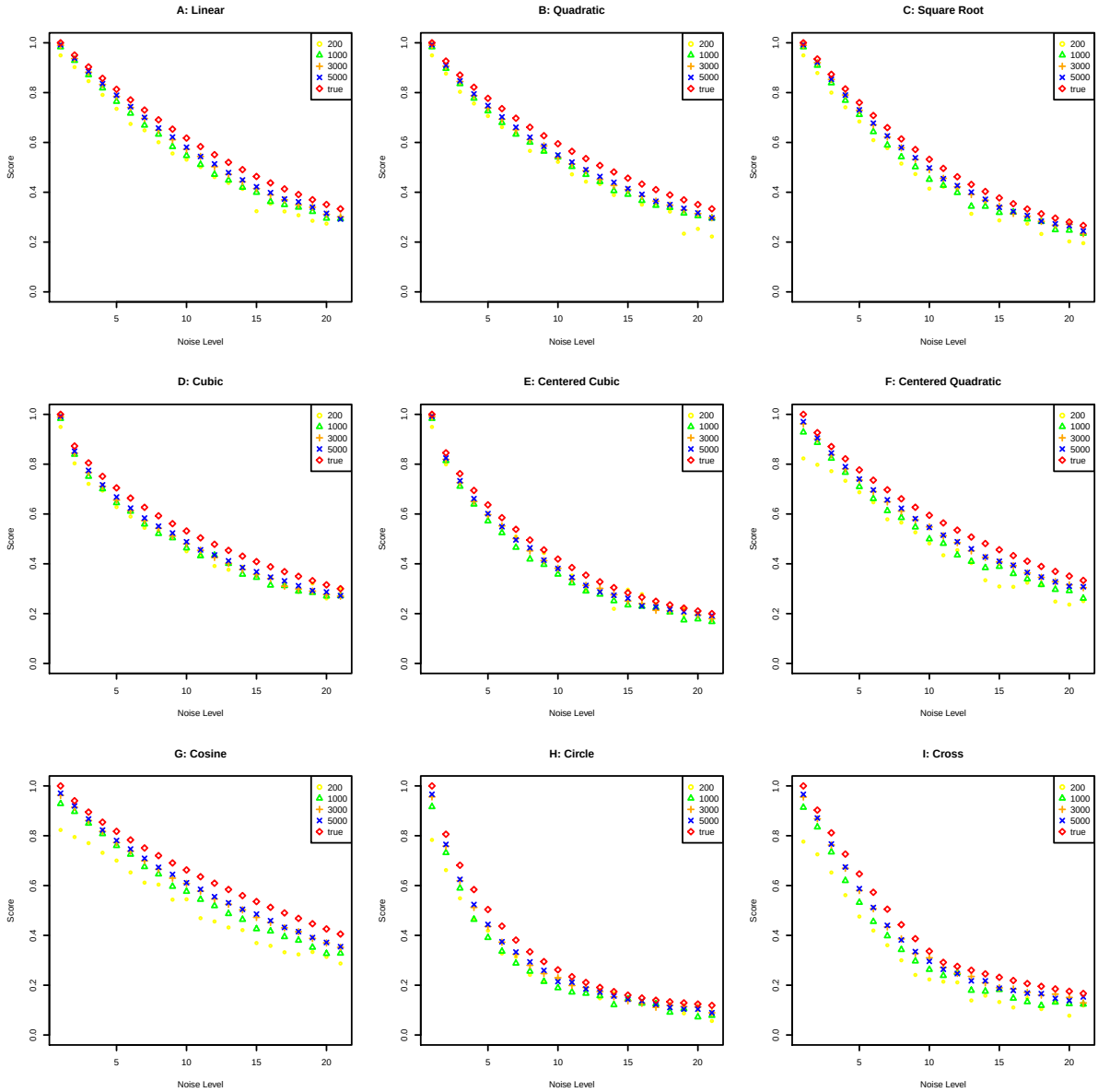


Figure 18: Additive noise with $k = 0.5\sqrt{n}$.

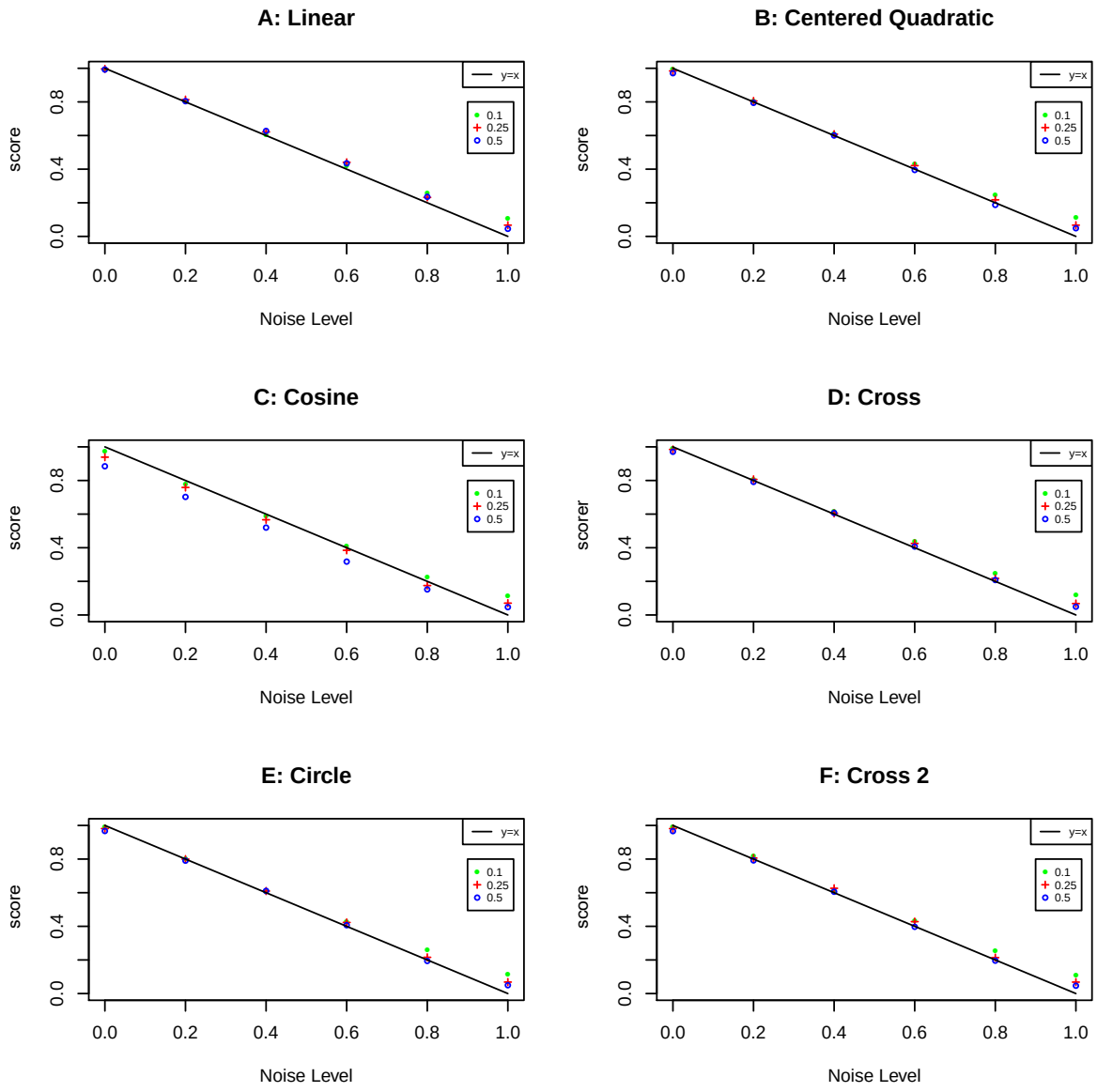


Figure 19: Mixture noise with $k = c\sqrt{n}$, where $c = 0.1, 0.25, 0.5$.

References

András Antos and Ioannis Kontoyiannis. Convergence properties of functional estimates for discrete distributions. *Random Structures & Algorithms*, 19(3-4):163–193, 2001.

Luca Bagnato, Lucio De Capitani, and Antonio Punzo. Testing serial independence via density-based measures of divergence. *Methodology and Computing in Applied Probability*, pages 1–15, 2013. ISSN 1387-5841. doi: 10.1007/s11009-013-9320-4.

Jon Louis Bentley. Multidimensional binary search trees used for associative searching. *Communications of the ACM*, 18(9):509–517, 1975.

Gérard Biau, Frédéric Chazal, David Cohen-Steiner, Luc Devroye, and Carlos Rodríguez. A weighted k-nearest neighbor density estimate for geometric inference. *Electron. J. Statist.*, 5: 204–237, 2011. doi: 10.1214/11-EJS606.

Yale Chang, Yi Li, Adam Ding, and Jennifer Dy. A robust-equitable copula dependence measure for feature selection. In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Statistics*. Citeseer, 2016.

Thomas H Cormen. *Introduction to algorithms*. MIT press, 2009.

David L. Donoho and Richard C. Liu. Geometrizing rates of convergence, ii. *The Annals of Statistics*, 19(2):pp. 633–667, 1991. ISSN 00905364.

Jennifer G Dy and Carla E Brodley. Feature selection for unsupervised learning. *The Journal of Machine Learning Research*, 5:845–889, 2004.

Magnus Ekdahl and Timo Koski. Bounds for the loss in probability of correct classification under model based approximation. *J. Mach. Learn. Res.*, 7:2449–2480, December 2006. ISSN 1532-4435.

R. H. Farrell. On the best obtainable asymptotic rates of convergence in estimation of a density function at a point. *The Annals of Mathematical Statistics*, 43(1):pp. 170–180, 1972. ISSN 00034851.

Andrew D Fernandes and Gregory B Gloor. Mutual information is critically dependent on prior assumptions: would the correct estimate of mutual information please identify itself? *Bioinformatics*, 26(9):1135–1139, 2010.

Jerome H Friedman. Multivariate adaptive regression splines. *The annals of statistics*, pages 1–67, 1991.

Kenji Fukumizu, Arthur Gretton, Xiaohai Sun, and Bernhard Schölkopf. Kernel measures of conditional dependence. In *NIPS*, volume 20, pages 489–496, 2007.

Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In *Algorithmic learning theory*, pages 63–77. Springer, 2005a.

Arthur Gretton, Ralf Herbrich, Alexander Smola, Olivier Bousquet, and Bernhard Schölkopf. Kernel methods for measuring independence. *J. Mach. Learn. Res.*, 6:2075–2129, December 2005b. ISSN 1532-4435.

Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *The Journal of Machine Learning Research*, 3:1157–1182, 2003.

Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. Gene selection for cancer classification using support vector machines. *Machine learning*, 46(1-3):389–422, 2002.

Xiaofei He, Deng Cai, and Partha Niyogi. Laplacian score for feature selection. In *Advances in neural information processing systems*, pages 507–514, 2005.

Harry Joe. Relative entropy measures of multivariate dependence. *Journal of the American Statistical Association*, 84(405):157–164, 1989. doi: 10.1080/01621459.1989.10478751.

Justin B Kinney and Gurinder S Atwal. Equitability, mutual information, and the maximal information coefficient. *Proceedings of the National Academy of Sciences*, 111(9):3354–3359, 2014.

Kenji Kira and Larry A Rendell. The feature selection problem: Traditional methods and a new algorithm. In *AAAI*, volume 2, pages 129–134, 1992.

Ron Kohavi and George H John. Wrappers for feature subset selection. *Artificial intelligence*, 97 (1):273–324, 1997.

Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Physical review E*, 69(6):066138, 2004.

Lucien Le Cam. Convergence of estimates under dimensionality restrictions. *The Annals of Statistics*, pages 38–53, 1973.

M. Lichman. UCI machine learning repository. `http://archive.ics.uci.edu/ml`, 2013.

Han Liu, John D Lafferty, and Larry A Wasserman. Exponential concentration for mutual information estimation with application to forests. In *NIPS*, pages 2546–2554, 2012.

D. O. Loftsgaarden and C. P. Quesenberry. A nonparametric estimate of a multivariate density function. *Ann. Math. Statist.*, 36(3):1049–1051, 06 1965. doi: 10.1214/aoms/1177700079.

Y. P. Mack and M. Rosenblatt. Multivariate k-nearest neighbor density estimates. *Journal of Multivariate Analysis*, 9(1):1–15, 1979.

Young-Il Moon, Balaji Rajagopalan, and Upmanu Lall. Estimation of mutual information using kernel density estimators. *Physical Review E*, 52(3):2318, 1995.

David S. Moore and James W. Yackel. Consistency properties of nearest neighbor density function estimators. *Ann. Statist.*, 5(1):143–154, 01 1977a. doi: 10.1214/aos/1176343747.

David S. Moore and James W. Yackel. Large sample properties of nearest neighbor density function estimators. *Statistical Decision Theory and Related Topics*, II:269–279, 1977b.

Ben Murrell, Daniel Murrell, and Hugh Murrell. R2-equitability is satisfiable. *Proceedings of the National Academy of Sciences*, 111(21):E2160–E2160, 2014.

Ben Murrell, Daniel Murrell, and Hugh Murrell. Discovering general multidimensional associations. *PloS one*, 11(3):e0151551, 2016.

R. B. Nelsen. *An introduction to copulas (Springer series in statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006. ISBN 0387286594.

Marek Omelka, Irène Gijbels, and Noël Veraverbeke. Improved kernel estimation of copulas: weak convergence and goodness-of-fit testing. *The Annals of Statistics*, 37(5B):3023–3058, 2009.

Dávid Pál, Barnabás Póczos, and Csaba Szepesvári. Estimation of renyi entropy and mutual information based on generalized nearest-neighbor graphs. In *NIPS*, pages 1849–1857, 2010.

Liam Paninski. Estimation of entropy and mutual information. *Neural Computation*, 15(6):1191–1253, 2003.

Hanchuan Peng, Fulmi Long, and Chris Ding. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(8):1226–1238, 2005.

B Poczos, Z Ghahramani, and J Schneider. Copula-based kernel dependency measures. In *International Conference on Machine Learning*, 2012.

Pavel Pudil, Jana Novovičová, and Josef Kittler. Floating search methods in feature selection. *Pattern recognition letters*, 15(11):1119–1125, 1994.

Sashank J Reddi and Barnabás Póczos. Scale invariant conditional dependence measures. In *Proceedings of The 30th International Conference on Machine Learning*, pages 1355–1363, 2013.

A. Rényi. On measures of dependence. *Acta Mathematica Academiae Scientiarum Hungarica*, 10 (3-4):441–451, 1959. ISSN 0001-5954. doi: 10.1007/BF02024507.

David N Reshef, Yakir A Reshef, Hilary K Finucane, Sharon R Grossman, Gilean McVean, Peter J Turnbaugh, Eric S Lander, Michael Mitzenmacher, and Pardis C Sabeti. Detecting novel associations in large data sets. *science*, 334(6062):1518–1524, 2011.

Yakir A. Reshef, David N. Reshef, Pardis C. Sabeti, and Michael Mitzenmacher. Theoretical foundations of equitability and the maximal information coefficient, 2014.

Yakir A Reshef, David N Reshef, Hilary K Finucane, Pardis C Sabeti, and Michael M Mitzenmacher. Measuring dependence powerfully and equitably. *arXiv preprint arXiv:1505.02213*, 2015a.

Yakir A Reshef, David N Reshef, Pardis C Sabeti, and Michael M Mitzenmacher. Equitability, interval estimation, and statistical power. *arXiv preprint arXiv:1505.02212*, 2015b.

B. Schweizer and E. F. Wolff. On nonparametric measures of dependence for random variables. *The Annals of Statistics*, 9(4):pp. 879–885, 1981. ISSN 00905364.

D.W. Scott. *Multivariate density estimation: theory, practice, and visualization*. Wiley Series in Probability and Statistics. Wiley, 1992. ISBN 9780471547709.

Johan Segers. Asymptotics of empirical copula processes under non-restrictive smoothness assumptions. *Bernoulli*, 18(3):764–782, 2012.

Bernard W Silverman. *Density estimation for statistics and data analysis*, volume 26. CRC press, 1986.

S. D. Silvey. On a measure of association. *Ann. Math. Statist.*, 35(3):1157–1166, 09 1964. doi: 10.1214/aoms/1177703273.

Noah Simon and Robert Tibshirani. Comment on detecting novel associations in large data sets by reshef et al, science dec 16, 2011. *Science*, 2011.

Le Song, Alex Smola, Arthur Gretton, Karsten M Borgwardt, and Justin Bedo. Supervised feature selection via dependence estimation. In *Proceedings of the 24th international conference on Machine learning*, pages 823–830. ACM, 2007.

Terry Speed. A correlation for the 21st century. *Science*, 334(6062):1502–1503, 2011. doi: 10.1126/science.1215894. URL `http://www.sciencemag.org/content/334/6062/1502.short`.

Ning Sun and Hongyu Zhao. Putting things in order. *Proceedings of the National Academy of Sciences*, 111(46):16236–16237, 2014. doi: 10.1073/pnas.1418862111. URL `http://www.pnas.org/content/111/46/16236.short`.

Taiji Suzuki, Masashi Sugiyama, and Toshiyuki Tanaka. Mutual information approximation via maximum likelihood estimation of density ratio. In *ISIT*, volume 9, pages 463–467, 2009.

Gábor J Székely and Maria L Rizzo. Brownian distance covariance. *The annals of applied statistics*, pages 1236–1265, 2009.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

Vladimir Naumovich Vapnik and Vlamimir Vapnik. *Statistical learning theory*, volume 1. Wiley New York, 1998.

Toomas R. Vilmansen. On dependence and discrimination in pattern recognition. *IEEE Transactions on Computers*, 21(9):1029–1031, September 1972. ISSN 0018-9340. doi: 10.1109/TC.1972.5009090.

T.R. Vilmansen. Feature evaluation with measures of probabilistic dependence. *IEEE Transactions on Computers*, 22(4):381–388, April 1973. ISSN 0018-9340. doi: 10.1109/T-C.1973.223725.

Lei Yu and Huan Liu. Efficient feature selection via analysis of relevance and redundancy. *The Journal of Machine Learning Research*, 5:1205–1224, 2004.