A Robust-Equitable Copula Dependence Measure for Feature Selection

Yale Chang
Department of ECE
Northeastern University

Yi Li
Department of Mathematics
Northeastern University

Adam(Aidong) Ding
Department of Mathematics
Northeastern University

Jennifer G. Dy
Department of ECE
Northeastern University

Abstract

Feature selection aims to select relevant features to improve the performance of predictors. Many feature selection methods depend on the choice of dependence measures. To select features that have complex nonlinear relationships with the response variable, the dependence measure should be equitable; i.e., it should treat linear and nonlinear relationships equally. In this paper, we introduce the concept of robust-equitability and identify a robust-equitable dependence measure robust copula dependence (RCD). This 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. In contrast, existing dependence measures cannot take these factors into account simultaneously. Experiments on synthetic and real-world datasets confirm our theoretical analysis, and illustrate its advantage in feature selection.

1 Introduction

In many applications, data samples are represented by features which domain experts assume to be important for the learning task. However not all of these features are useful: some may be irrelevant and some may be redundant. Therefore, feature selection is needed to improve the performance of learning tasks, decrease computational cost with fewer variables, and aid in understanding the factors that are important for prediction.

Feature selection algorithms [14, 8, 30] can be categorized as either filter, wrapper, or embedded methods. In filter methods [13, 30, 21, 10, 26], features are preselected without running the learning algorithm and are evaluated only through the intrinsic properties of the data. Wrapper methods [14, 9, 2] select features by "wrapping" the search around the learning algorithm and evaluate feature subsets based on classifier performance in each candidate feature subset. Embedded methods [28, 29] incorporate feature search and the learning algorithm (e.g., classifier) into a single optimization problem formulation. Contrary to filter methods, wrapper and embedded methods select features specific to the learning algorithm. Hence, they are most likely to be more accurate than filter methods on a particular classifier, but the features they choose may not be appropriate for other classifiers. Another limitation of wrapper methods is that wrappers are computationally expensive because they need to train and test the classifier for each feature subset candidate, which can be prohibitive when working with high-dimensional data.

Filter methods evaluate features based on some dependence criterion between features and the target variable. The goal is to select a subset of features that optimizes this criterion. An exhaustive search of 2^d possible feature subsets (where d is the number of features) is computationally impractical. Heuristic search strategies such as greedy approaches (e.g., sequential forward/backward search [23]) are commonly used but can lead to local optima. Random search methods, such as genetic algorithms, add some randomness to help escape from local optima. In some cases when the dimensionality is very high, one can only afford an individual search. Individual search methods evaluate each feature individually according to a criterion [8, 10]. They then select features, which either satisfy a condition or are top-ranked. The problem with individual search methods is that they ignore feature interaction and dependencies. One way of accounting for such interactions and dependencies is to select relevant features individually and then add a separate redundancy removal step to account for linear correlation between features [30]. Another way is to maxi-

Appearing in Proceedings of the 19^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2016, Cadiz, Spain. JMLR: W&CP volume 41.

mize relevance and minimize redundancy (mRMR) together [21].

Besides search strategies, the performances of filter methods also depend heavily on the choice of dependence measures. How to measure the dependence between random variables is a fundamental problem in statistics and machine learning. A commonly used dependence measure is the Pearson correlation coefficient (ρ_{lin}). However, this measure prefers linear relationships. Another popular measure is mutual information (MI). MI can handle nonlinear dependencies, but it is difficult to estimate [3, 24] (see Theorem 3 in Section 2). Kernel-based dependence measures [7, 6], such as 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 or NOCCO) [5, 6] is kernel-free, meaning it does not depend on the choice of kernels in the limit of infinite data. While HSNIC is kernel-free, both HSIC and HSNIC's values can vary when we use different scales. To make the kernel-based dependence measures invariant to strictly monotone transformation of the marginal variables, [22] applied Maximum Mean Discrepancy (MMD), which can be written in HSIC formulation, after empirical copula transformation (CMMD). Similarly, [11] applied HSNIC after empirical copula transformation (CHSNIC). However, CMMD and CHSNIC are only invariant to strictly monotone transformations; they fail to treat non-monotonic relations equally. [24] 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. [12] redefined 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. [12] proved that MI is self-equitable. The self-equitable dependence measure will treat all forms of relationships equally in the large data limit for the additive noise model.

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. We define a dependence measure *robust copula dependence* (RCD) as half the L_1 distance between copula density and uniform (independence) density. RCD is equivalent to the Silvey's Delta measure [25]. We show that, among a class of self-equitable copula-based dependence measures, only RCD (i.e. Silvey's Delta) is also robust-equitable. In the literature, Silvey's Delta (RCD) was only cited as an abstract benchmark. We propose a k-nearest-neighbor (KNN)-based estimator for RCD and prove its consistency. Corresponding to the RCD definition, we define CD_2 as the L_2 distance between copula density and uniform density. CD_2 is the theoretical value of HSNIC in large data limit [6]. We also prove that although both MI and CD_2 are self-equitable, they are not robust-equitable. Therefore, MI and CD_2 may not rank features correctly by dependence strength in some cases, even in the large data limit. We confirm this phenomena on both synthetic and real-world datasets. 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 all dependence measures that we have found in the literature, it can be very useful for feature selection.

The contributions of this work are: (1) the introduction of the concept of robust-equitability and the proposal of a practical consistent estimator for the robust-equitable dependence measure RCD; (2) 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, (3) 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 datasets, in terms of robustness to function types, correctness in large sample size and correctness in unequal sample sizes.

2 A Robust-Equitable Dependence Measure

As the feature selection results depend critically on the dependence measures used, we investigate the theoretical properties of different dependence measures. Particularly, we would like the dependence measures to rank features with less noise as having stronger dependence with the response variable. We want measures that do not prefer a particular type of relationship (such as linear). Also we do not want the con-

clusions to be too sensitive to sample size. That is, when we have unequal sample sizes on different features, the dependence measure should not prefer the feature with more samples, but still rank the features by the strength of the deterministic signal compared to noise. Note that databases with unequal sample sizes on different features are now becoming more common due to the prevalence of collecting data from heterogeneous sources (for example, often there will be more samples with clinical features compared to samples with genomic information). Rather than throwing away samples from the larger set to get equal sample sizes for all features, we would like to use all data available to perform feature selection. We formalize these ideas through the recently proposed equitability concept. That is, we want to use dependence measures that reflect the noise level, regardless of types of relationships.

2.1 Equitability

[24] first proposed the concept of *equitability*: a dependence measure should treat all types of (linear and nonlinear) relationships equally. This property extends beyond copula-based dependence measures (e.g., CMMD and CHSNIC) which treat all monotone relationships equally. Sklar's theorem decomposes the joint distribution function (CDF), $F_{X,Y}(x,y)$, into the marginal distributions and the copula function

$$F_{X,Y}(x,y) = Pr(X \leq x, Y \leq y) = C[F_X(x), F_Y(y)], \quad (1)$$

for all x, y. Here $F_X(x) = Pr(X \leq x)$ and $F_Y(y) = Pr(Y \leq y)$ are the marginal CDFs of X and Y respectively. The copula $C(u,v) = Pr(U \leq u, V \leq v)$ is the CDF of copula-transformed, uniformly distributed, variables $U = F_X(X)$ and $V = F_Y(Y)$. In this way, dependence measures on the copula transformed variables are invariant to strictly monotone transformations. Moreover, the copula decomposition separates the dependence (copula) from any marginal effects. Figure 1 shows the data from two distributions with different marginals but the same dependence structure.

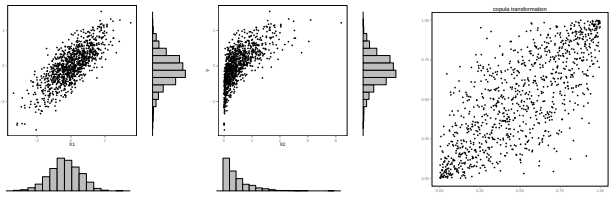

Figure 1: Left: Bivariate Gaussian with $\rho = 0.75$. Middle: Data with exponential marginal for X. Right: The Gaussian copula. The first two distributions have the same copula as in the right figure.

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).

To treat non-monotone functions equally, we further choose among copula-based measures by their equitability. [12] defined self-equitability by extending the invariance property to all nonlinear relationships in the regression model $Y = f(X) + \epsilon$, 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 $Y = f(X) + \epsilon$.*

A self-equitable dependence measure will have the same values in all three examples (A), (B) and (C) in Table 1. Using a self-equitable dependence measure in feature selection avoids (in the large data limit) preference for certain types of relationships such as linear or monotone relationship, in the additive noise model.

Mutual information (MI) is self-equitable and is based on copula density $c(u,v)$, MI $= \iint_{I^2} \log[c(u,v)]c(u,v)dudv$, 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 $\pi(u,v) = 1$.

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

Combining Eq.4 in [6] and Eq.(2) 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 [12], since $g(x) = |x-1|^\alpha$ is convex when $\alpha \geq 1$.

However, in practice, regression $Y = f(X) + \epsilon$ with additive noise does not capture all types of noise. 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 deter-

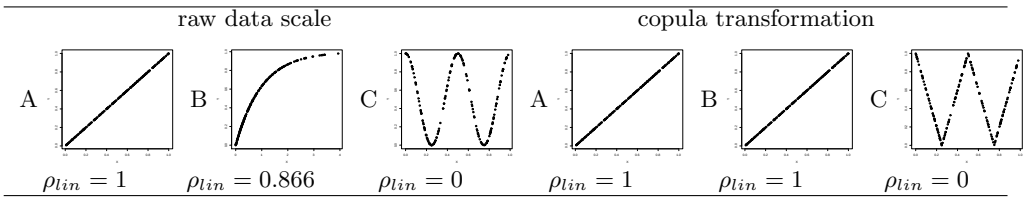

Table 1: Pearson correlation coefficient on three function relationships.

ministic sinusoidal function. The Right subfigure is the same deterministic signal on a uniform background noise. This is mathematically described by a mixture copula: the uniform copula $\Pi(u, v) = uv$ on a unit square I^2 is added to a deterministic signal C_s, which is a singular copula [20]. An equitable dependence measure should give the same value for all types of deterministic signal C_s.

Definition 2 *A dependence measure $D[X;Y]$ is robust-equitable if and only if $D[X;Y] = p$ whenever (X, Y) follows a distribution with copula $C = pC_s + (1-p)\Pi$, for a singular copula C_s.*

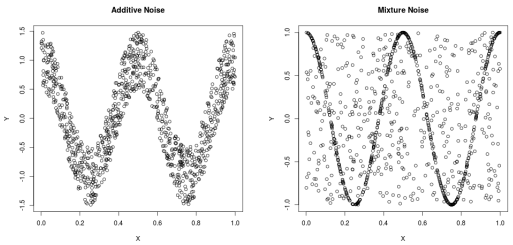

Figure 2: Left: Additive noise for self-equitability. Right: Mixture noise for robust-equitability.

Both MI and CD_2 are not robust-equitable since their theoretical values are always ∞ for this setting when $p > 0$. Thus in practice, their ranking of features under this type of noise is mostly affected by the estimation errors in MI and CD_2, which is very sensitive to sample size. They are very difficult to estimate in this setting, which cause inconsistencies in feature selection, and can result in large prediction errors.

2.2 Robust Copula Dependence

Robust copula dependence is defined as $RCD = \frac{1}{2}\mathrm{CD}_1 = \frac{1}{2}\iint_{I^2} |c(u, v) - 1| dudv$. RCD is a copula density-based measure that agrees with the Pearson correlation coefficient when the deterministic signal in Definition 2 is linear. RCD is self-equitable by Lemma 1. Moreover, RCD is also robust-equitable, which means it can deal with both noise types in Figure 2.

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

Proof. By the robust copula dependence definition and robust-equitability definition 2: $RCD = \frac{1}{2}[p\int_{\mathcal{S}} C(du, dv) + \int_{I^2\setminus\mathcal{S}} |(1-p)-1| dudv] = \frac{1}{2}[p+p] = p$, where $\mathcal{S}$ is the support of the singular copula C_s.

Mathematically, RCD is equivalent to Silvey's Delta [25]: $\Delta = \iint_{\phi>1}[p(x, y) - p_X(x)p_Y(y)]dxdy$, 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 $\phi(x, y) = p(x, y)/[p_X(x)p_Y(y)]$. [25] interprets ϕ as the Radon-Nikodym derivative of the joint distribution with respect to the product of marginal distributions and can cover the possibility of singularity. Also the definition holds for X and Y with higher dimensions $d_X \geq 1$ and $d_Y \geq 1$. We study the estimation errors next, and provide a practical estimator for RCD.

2.3 Statistical Estimation Errors

Feature selection with a non-self-equitable dependence measure prefers certain types of relationships. Feature selection with a non-robust-equitable dependence measure (such as MI and CD_2) is often sensitive to sample size. Even for a tiny proportion of deterministic signal, $\mathrm{MI} = \infty$ and $\mathrm{CD}_2 = \infty$ in the large data limit. Their ranking of features in finite sample size is very much determined by the estimation errors in $\widehat{\mathrm{MI}}$ and $\widehat{\mathrm{CD}_2}$.

We theoretically show that MI and CD_2 are much harder to estimate compared to the robust-equitable RCD. This is due to the instability in the theoretical values of MI and CD_2. For example, $\mathrm{MI} = \infty$ when 10% of linear deterministic data $Y = X$ is mixed with continuous uniform background noise on I^2. In contrast, $\mathrm{MI} = 1$ when these 10% of data instead fall around the line $Y = X$ in a very small strip of area $0.1/exp(10) = 4.5 \times 10^{-6}$. No estimator can do well in these two almost indistinguishable distributions with very different MI values (∞ and 1). 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:

$$|c(\mathbf{z}_1) - c(\mathbf{z}_2)| \leq M_1 \|\mathbf{z}_1 - \mathbf{z}_2\|_{l_1}, \qquad (3)$$

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 Hölder condition (only possible on bounded density regions) is a standard condition for studying density estimation errors. However, all common copula densities, except the independent copula [20], are unbounded and cannot satisfy the Hölder condition on the whole I^2. Thus the estimation errors should be studied over the family $\mathfrak{C}$ instead. MI has infinite minimax risk over the family $\mathfrak{C}$, thus it cannot be consistently estimated.

Theorem 3 *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 (2). Then*

$$\begin{aligned} \sup_{C\in\mathfrak{C}} E[|\widehat{MI}(C) - \mathrm{MI}(C)|] &= \infty, \text{ and} \\ \sup_{C\in\mathfrak{C}} E[|\widehat{CD}_\alpha(C) - CD_\alpha(C)|] &= \infty, \text{ for any } \alpha > 1. \end{aligned} \tag{4}$$

The detailed proof is provided in the supplementary material. 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) [19], the k-nearest-neighbor (KNN) [15], maximum likelihood estimation of density ratio [27]. However, practitioners are often frustrated by the unreliability of these estimation [3, 24]. This theorem provides a theoretical explanation.

In contrast to MI, the contribution from the unbounded copula density region A_M^c to RCD is very small. Even if $c(\mathbf{z})$ cannot be consistently estimated, its error can be bounded. The following theorem shows the result for the KDE estimator for RCD.

Theorem 4 *Let the KDE estimator of the d-dimensional copula density based on observations $\mathbf{Z}_1, \cdots, \mathbf{Z}_n$ be*

$$\hat{c}_{kde}(\mathbf{Z}) = \frac{1}{nh^d} \sum_{i=1}^{n} K(\frac{\mathbf{Z}_i - \mathbf{Z}}{h}). \tag{5}$$

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

$$\sup_{C\in\mathfrak{C}} E[|\widehat{RCD} - RCD|] \leq M_1 h + \frac{\sqrt{2}\mu_2}{\sqrt{n}h^{\frac{d}{2}}} + O(\frac{1}{nh^d}). \tag{6}$$

In addition to the KDE based RCD estimator, we can estimate RCD consistently by plugging in the KNN estimator [17] 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 $\widehat{RCD} = RCD(\hat{c}) = \sum_{\hat{c}(\mathbf{Z}_i)>1}[1 - 1/\hat{c}(\mathbf{Z}_i)]/n$. The computational complexity of this estimator remains $O(n^2 \log n)$ for multivariate X and Y (when $d_X > 1, d_Y > 1$).

Theorem 5 *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 estimator $\widehat{RCD} = RCD(\hat{c})$ has a risk bound*

$$\sup_{C\in\mathfrak{C}} E[|\widehat{RCD} - RCD|] \leq \tilde{c_1}(\frac{k}{n\epsilon})^{\frac{2}{d}} + \frac{\tilde{c_2}}{\sqrt{k}} + 2\epsilon, \tag{7}$$

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 the supplement) by citing formulas in [18]. Without it, RCD can still be estimated consistently as in Theorem 4. The error bound (7) 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 the supplement 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, 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.

3 Experimental Results

To empirically verify the properties of RCD in our theoretical analysis, we design several synthetic experiments and then work with real-world datasets. To measure the performance of feature selection using a dependence measure, we calculate the 10-fold cross-validated mean-squared-error (MSE) of a nonlinear predictor using the selected features. A good dependence measure should rank the features according to MSE. We first use three synthetic examples to show the robustness of RCD. It should rank features correctly (in MSE) for different types of functions, in the large data limit and for unequal sample sizes. The other measures should provide wrong ranks in at least one of the three examples due to either non-self-equitability or non-robust-equitability. For synthetic data, we also provide the theoretical mean-squared-error (TMSE) of the best prediction $E(Y|X)$ for each

feature X. This provides assurance that, for the three examples, the ranking in MSE is consistent with ranking in TMSE. Therefore, the conclusions are not due to training error. Then we also investigate whether this behavior is similarly true 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.

Finally, we compare the performance of feature selection by the filter method mRMR [21], using various dependence measures as measures of relevance and redundancy. 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, the Pearson's correlation with a linear regression predictor should do best if linear relationship is dominant in a data set. For a fair comparison, we measure the performance by the 10-fold cross-validated MSE of spline regression, a general nonlinear predictor [4], using selected features. We used six benchmark datasets, for nonlinear predictions, from the UCI Data Repository [16]. 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 six benchmark datasets.

There are some parameters to be set for computing various dependence measures. For kernel based measures, we follow the settings used by [6]. For HSNIC, we set the regularization parameter $\epsilon_n = 10^{-5}n^{-3.1}$ to satisfy the convergence guarantee given by Theorem 5 from [6]. We set $k = 0.25\sqrt{n}$ for the k-NN estimator of MI, RCD and CD_2.

3.1 Synthetic Datasets

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 3. 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 2. In particular, we are interested in comparing the performance of each dependence measure from the following perspectives: (1) how the equitability property affects the predictive performance; and (2) how the sample size affects the results.

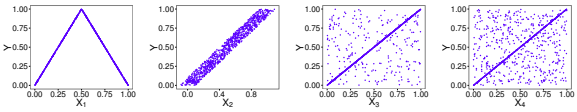

Figure 3: (X_1, Y) has a nonmonotonic relationship with deterministic signal. (X_2, Y) has linear relationship with uniform additive noise with width $d = 0.2$. (X_3, Y) has background noise with linear signal portion 0.75. (X_4, Y) is similar to (X_3, Y) but with signal portion 0.5.

Ideal Result. We report the theoretical mean-squared-error (TMSE) and the 10-fold cross-validated MSE values in the last two rows of Table 2. They both increase with X_i $(i = 1, \cdots, 4)$, which means our features are arranged in deceasing importance. Note that our RCD results are consistent with the MSE results, providing higher scores for those with lower MSE values and that the ranking is not affected by varying sample size. In addition, note that the proportion of deterministic signal, p, to noise (as defined in Definition 2) for the synthetic data X_1 is 1, X_3 is 0.75 and X_4 is 0.5 which are close to the values learned by RCD.

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 2, 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 [11], see supplement.

Selection Correctness in Large Sample Size. 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 correctly ranks X_2 as having higher dependence 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 Theorem 3. 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 $n = 10,000$, they fail to provide the correct ranking as they did when $n = 300$. CD_2 also incorrectly ranks X_3 and X_4 higher than X_2. These inconsistencies may mislead

	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
TMSE	0	0	0.0033	0.0033	0.010	0.010	0.042	0.042
MSE	0.00098	0.0023	0.0033	0.0031	0.037	0.037	0.064	0.063

Table 2: 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.

feature selection algorithms since they provide higher scores for the features with less signal strength when the sample size is large.

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 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; 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.

3.2 Real Datasets

We first verify that the equitability properties in subsection 3.1 are also observed on real data. Then we perform a comparison of the various dependence measures in feature selection on six real-world datasets.

Self-equitability. Consider the stock dataset from StatLib [1]. This dataset 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 4. 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 5, self-equitable measures MI, CD_2, and RCD all correctly select X_1. While the non-self-equitable measures fail to select the right feature.

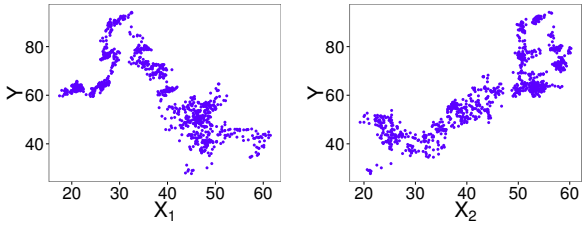

Figure 4: $(X_i, Y), i = 1, 2$ in stock dataset

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 5: Measures for X_1, X_2 in stock dataset

Selection Correctness in Large Sample Size. Consider the KEGG metabolic reaction network dataset [16]. 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. Ideally, a measure should not vary too much as the sample size changes. However, in Table 3, 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 gives the correct ranking, i.e., X_2 is more relevant than X_3 in predicting Y. 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.

Selection Correctness in Unequal Sample Sizes. We would like to be able to compare feature relevance even when some features have missing measurements.

	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 3: Dependence measure for three features in metabolic reaction network dataset

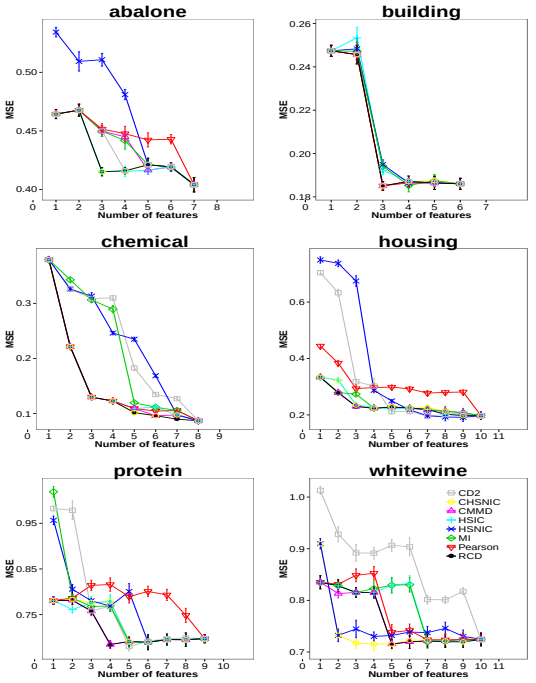

Figure 6: MSE of mRMR-based Feature Selection on six real-world datasets

An ideal dependence measure should not be influenced by unequal sample sizes. Observe the ranking of X_1 and X_3 from Table 3, when they have equal sample sizes (either 1000 or 20,000), MI, CD_2 and RCD all give the correct ranking of X_1 as being more relevant than X_3. However, if there are missing measurements of X_1, then we may need to compare X_1 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, our new dependence measure RCD is robust to relationship forms, consistently selects the correct features for large sample size and unequal sample sizes.

Feature Selection Experiments.

Here we compare the performance of various dependence measures used in the computationally efficient filter method, mRMR [21], for feature selection.

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 six real-world datasets. Due to the quadratic computational cost of kernel-based measures, up to 1000 samples are used for each dataset. 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 6. We used the Kruskal-Wallis test to compare the MSE between different measures. Table 4 list 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). We can see that RCD performs the best in 5 out of 6 data sets. Most other measures are worse off in more than half of the data sets. Only CHSNIC is close in performance. RCD is best in *housing* and *protein*, CHSNIC is best in *whitewine*, and they are close in performance for the other 4 data sets.

Dataset	Top Dependence Measures
abalone	**RCD** CHSNIC HSIC
building	**RCD** CHSNIC Pearson CMMD CD_2
chemical	**RCD** CHSNIC
housing	**RCD** CMMD
protein	**RCD** CMMD
whitewine	CHSNIC

Table 4: Measures ranked by predictive MSE

4 Conclusions

The performance of filter-based feature selection algorithms depends on the choice of dependence measures. In this paper, we introduced a new concept of robust-equitability. We proved that robust copula dependence (RCD) is a robust-equitable and self-equitable dependence measure, and provided a practical estimator. The non-self-equitable HSIC and CMMD prefer features with monotonic relationships over less noisy features with more complex relationships. MI and CD_2 overcome that deficiency but have estimation problems, leading to non-robust feature selection in large sample sizes and when comparing features of unequal sample sizes. The theoretical value of HSNIC/CHSNIC in the large data limit is equal to CD_2. However, in practice HSNIC and CHSNIC converge very slowly and exhibit similar deficiencies as the non-self-equitable HSIC and CMMD. In contrast, our dependence measure RCD can overcome these limitations and rank features according to deterministic signal strengths, making RCD suitable for feature selection as confirmed in our empirical study on synthetic and real-world datasets.

Acknowledgement

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

References

[1] http://lib.stat.cmu.edu/.

[2] J. G. Dy and C. E. Brodley. Feature selection for unsupervised learning. *The Journal of Machine Learning Research*, 5:845–889, 2004.

[3] A. D. Fernandes and G. 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.

[4] J. H. Friedman. Multivariate adaptive regression splines. *The annals of statistics*, pages 1–67, 1991.

[5] K. Fukumizu, F. R. Bach, and A. Gretton. Statistical consistency of kernel canonical correlation analysis. *J. Mach. Learn. Res.*, 8:361–383, May 2007.

[6] K. Fukumizu, A. Gretton, X. Sun, and B. Schölkopf. Kernel measures of conditional dependence. In *NIPS*, volume 20, pages 489–496, 2007.

[7] A. Gretton, O. Bousquet, A. Smola, and B. Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In *Algorithmic learning theory*, pages 63–77. Springer, 2005.

[8] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *The Journal of Machine Learning Research*, 3:1157–1182, 2003.

[9] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene selection for cancer classification using support vector machines. *Machine learning*, 46(1-3):389–422, 2002.

[10] X. He, D. Cai, and P. Niyogi. Laplacian score for feature selection. In *Advances in neural information processing systems*, pages 507–514, 2005.

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

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

[13] K. Kira and L. A. Rendell. The feature selection problem: Traditional methods and a new algorithm. In *AAAI*, volume 2, pages 129–134, 1992.

[14] R. Kohavi and G. H. John. Wrappers for feature subset selection. *Artificial intelligence*, 97(1):273–324, 1997.

[15] A. Kraskov, H. Stögbauer, and P. Grassberger. Estimating mutual information. *Physical review E*, 69(6):066138, 2004.

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

[17] D. O. Loftsgaarden and C. P. Quesenberry. A nonparametric estimate of a multivariate density function. *Ann. Math. Statist.*, 36(3):1049–1051, 06 1965.

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

[19] Y.-I. Moon, B. Rajagopalan, and U. Lall. Estimation of mutual information using kernel density estimators. *Physical Review E*, 52(3):2318, 1995.

[20] R. B. Nelsen. *An Introduction to Copulas (Springer Series in Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[21] H. Peng, F. Long, and C. 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.

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

[23] P. Pudil, J. Novovičová, and J. Kittler. Floating search methods in feature selection. *Pattern recognition letters*, 15(11):1119–1125, 1994.

[24] D. N. Reshef, Y. A. Reshef, H. K. Finucane, S. R. Grossman, G. McVean, P. J. Turnbaugh, E. S. Lander, M. Mitzenmacher, and P. C. Sabeti. Detecting novel associations in large data sets. *science*, 334(6062):1518–1524, 2011.

[25] S. D. Silvey. On a measure of association. *Ann. Math. Statist.*, 35(3):1157–1166, 09 1964.

[26] L. Song, A. Smola, A. Gretton, K. M. Borgwardt, and J. Bedo. Supervised feature selection via dependence estimation. In *Proceedings of the 24th international conference on Machine learning*, pages 823–830. ACM, 2007.

[27] T. Suzuki, M. Sugiyama, and T. Tanaka. Mutual information approximation via maximum likelihood estimation of density ratio. In *ISIT*, volume 9, pages 463–467, 2009.

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

[29] V. N. Vapnik and V. Vapnik. *Statistical learning theory*, volume 1. Wiley New York, 1998.

[30] L. Yu and H. Liu. Efficient feature selection via analysis of relevance and redundancy. *The Journal of Machine Learning Research*, 5:1205–1224, 2004.