On the Effect of Suboptimal Estimation of Mutual Information in Feature Selection and Classification

Kiran Karra, Member, IEEE, Lamine Mili, Fellow, IEEE,
E-mail: kiran.karra@vt.edu

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

Index Terms—Copula, Mutual Information, Skewed, Hybrid, Dependence.

1 INTRODUCTION

Many applications of data mining and machine learning utilize measures of the strength of association between random variables to reduce data redundancy and find interesting associations within datasets. To accomplish this, various measures of statistical association have been introduced in the literature, including the correlation coefficient \(\rho\) [1], MIC [2], the RDC [3], the dCor [4], the Ccor [5], CoS [6], and CIM [7]. In addition, many estimators of mutual information such as the kNN [8], the vME [9], and the AP [10] are used as measures of association, especially in machine learning. Although the theoretical forms of these measures of association are generally applicable, the estimators of these quantities often fall short with real world impediments. For example, characteristics such as whether the data are discrete or continuous, linear or nonlinear, skewed or balanced, monotonic or nonmonotonic, noisy or clean, and independent and identically distributed (i.i.d.) or serially dependent, to name a few, drive the performance of the estimator. Because the estimation of the strength of association is often abstracted from the algorithms which rely on them, more emphasis in machine learning research is currently placed on designing and developing new algorithms rather than more accurate estimation.

In this paper, we show the effect of suboptimal estimation of mutual information on feature selection and classification performance. We focus on the scenario where the strength of association needs to be measured between noisy i.i.d continuous and discrete random variables (henceforth referred to as hybrid random variables) that are skewed, where the number of unique outcomes of the discrete random variable are small and the dependence structures are nonlinear. This case represents an important subset of problems in machine learning, where real world datasets that often have nonlinear associations between them with skewed marginal distributions, need to be classified according to provided output labels. Additionally, we restrict ourselves to only compare estimators of the strength of statistical association that are proven to satisfy the data processing inequality (DPI); that is, all estimators of mutual information (kNN,vME,AP,H_{MI}) and the index CIM. Measures of association that are proven to satisfy the DPI are preferred in machine learning due to the relationship between the DPI and Markov chains [11]. Furthermore, the DPI assumption is implicit in many machine learning algorithms which utilize measures of the strength of association, such as the maximum-relevance minimum-redundancy (MRMR) algorithm for Markov network discovery and feature selection [12], [13].

The paper is organized as follows. We begin with an introduction to estimation of association between random variables and stochastic processes. Here, we discuss the difficulty in estimating the strength of association for the hybrid random variable scenario. In addition, we define a new metric for measuring the performance of an estimator that is easily computable and shown to be useful for assessing how well an estimator will perform under the real world impediments and scenarios considered in this paper. Next, we perform synthetic data simulations which measure the performance of the various estimators under consideration. To accomplish this, we utilize the copula framework to generate data with a wide range of dependence structures and marginal distributions to comprehensively test the performance of these estimators under various real world

- K. Karra and L. Mili are with the Department of Electrical and Computer Engineering, Virginia Polytechnic Institute and State University, Arlington, VA, 22203.

Manuscript received May 1, 2018.
impediments. We show that for the properties tested, the CIM estimator’s performance compares favorably to other DPI satisfying measures of mutual information. Next, we apply these various estimators to real-world datasets and characterize how suboptimal estimation of mutual information affects machine learning classification performance. We synthetically vary these real world datasets further in order to simulate more extreme scenarios that may be encountered in the real world, and characterize the performance through these variations. We show that according to the estimator response curve, the CIM estimator’s performance again compares favorably to other DPI satisfying measures of association. This corroborates our findings with synthetic data, which empirically proves the usefulness of the estimator response curve. Concluding remarks are then provided in the final section.

2 Estimating Dependence

Random variables \(X\) and \(Y\) are said to be dependent if \(f_{XY}(x,y) \neq f_X(x)f_Y(y)\), where \(f_{XY}\) is the joint distribution of \(X\) and \(Y\), and \(f_X\) and \(f_Y\) are the marginal distributions of \(X\) and \(Y\), respectively. The strength of that dependence can be viewed from two angles. From an information theory perspective, the strength of the dependence encompassed by the joint density, \(f_{XY}\), can loosely be stated as inversely proportional to the amount of disorder in the joint density. From a statistical perspective, the strength of the dependence can be viewed as a generalized distance measure between statistical independence, captured by the independence copula, \(\Pi\), and the copula of the joint distribution function \(F_{XY}\).

The first solution put forth to measure this strength of dependence was the correlation coefficient, \(\rho\). Although popular, the correlation coefficient has many drawbacks, with the most notable being that it can only measure linear dependence. Many real world datasets have nonlinear dependence structures, and the correlation coefficient does not fully capture the strength of association between the random variables in these scenarios. Nonlinear indices of association, including distance correlation, \(dCor\) \[4\], and the maximal information coefficient, \(MIC\) \[2\], were then introduced by researchers in order to overcome the linearity limitation of the correlation coefficient.

Another class of estimators of the strength of association between random variables and stochastic processes utilize the copula framework to overcome the linearity limitation. Copulas are multivariate joint probability distribution functions for which the marginal distributions are uniform \[14\]. The existence of a copula \(C(\cdot)\) associated with a collection of random variables, \(X_1, \ldots, X_n\), following a joint probability distribution function, \(F(\cdot)\), and marginals \(F_{X_i}(x_i)\) is ensured by Sklar’s theorem, which states that

\[
F(x_1, \ldots, x_n) = C(F_{X_1}(x_1), \ldots, F_{X_n}(x_n)).
\]

This theorem guarantees the unicity of the copula \(C\) for continuous random variables and it unveils its major property, which is its ability to capture the unique dependency structure between any continuous marginal distributions \(X_1, \ldots, X_n\). Thus, the copula \(C\) can be used to define a measure of dependence between continuous random variables. Because copulas capture the full dependence structure between random variables, they have been utilized by various indices, including copula correlation coefficient, \(Ccor\) \[5\], the copula statistic, \(CoS\) \[6\], the randomized dependence coefficient, \(RDC\) \[3\], and the copula index for monotonicity \(CIM\) \[7\].

The advantage of these aforementioned measures of association is that they are true indices; that is, that they take on values between zero and one, where one represents perfect association between the random variables or stochastic processes, and zero represents no association. For estimators which also satisfy Rényi’s properties of dependence measures, a value of zero also implies statistical independence \[15\]. The notable disadvantage of all of these indices is that all of them, except for the CIM, are not necessarily proven to satisfy certain desirable properties of estimators of association. These properties include Rényi’s properties \[15\], and the data processing inequality (DPI) \[11\]. For this reason, machine learning practitioners often use mutual information (MI) as an indicator of the strength of association. The mutual information between random variables \(X\) and \(Y\) is defined as

\[
I(X,Y) = \int_{\mathcal{X}} \int_{\mathcal{Y}} f_{XY}(x,y) \log \frac{f_{XY}(x,y)}{f_X(x)f_Y(y)} \, dx \, dy,
\]

where \(f_{XY}(x,y)\) is the joint distribution of \(X\) and \(Y\), and \(f_X(x)\) and \(f_Y(y)\) are the marginal distributions of \(X\) and \(Y\), respectively \[16\]. Four common estimators of mutual information include k-nearest neighbors, \(kNN\) \[3\], von Mises Expansion, \(vME\) \[2\], adaptive partitioning \(AP\) \[10\], and entropy based estimation, hereby denoted in this paper as \(H_{MI}\). Briefly, \(kNN\) based estimation of mutual information uses the k-nearest neighbors approach to estimate the univariate and multivariate densities in \[2\], and then applies the integral (or summation in the discrete case). The adaptive partitioning approach to estimating mutual information uses an algorithm to optimally partition the space spanned by the two random variables, such that mutual information can be accurately estimated \[10\]. Both the \(kNN\) and \(AP\) approaches attempt to accurately measure mutual information through partitioning of the space. Conversely, von Mises expansion utilizes influence functions to measure the mutual information between random variables. Finally, the entropy based estimator \(H_{MI}\) simply uses the relationship between mutual information and conditional entropy, given by

\[
I(X,Y) = H(Y) - H(Y|X),
\]

where

\[
H(X) = -\int_{-\infty}^{\infty} p(x) \log[p(x)] \, dx
\]

2.1 Where do these estimators fall short?

When measuring the strength of association between continuous and discrete random variables, most of the estimators previously mentioned fall short. In general, it becomes more difficult to measure association between continuous and discrete random variables as the number of unique discrete outcomes decreases \[17\]. The case of measuring the strength
of association between hybrid random variables, however, is extremely important in machine learning. From classification, clustering, and feature selection perspectives, features are typically amenable to be modeled as continuous random variables, while outcomes or clusters are better modeled as discrete random variables.

Each estimator class above presents a different opportunity for why the hybrid random variable scenario is difficult for estimation. The correlation coefficient, $\rho$, is actually the standardized linear regression coefficient between random variables $X$ and $Y$. If $Y$ takes on a small number of unique outcomes, the MMSE objective for solving the regression coefficient does not properly capture the dynamics of the data, and in fact violates an implicit assumption of linear regression, that the dependent variable, $Y$, be continuous. This is illustrated in Fig. 1. In it, the independent random variable, $X$, and the dependent random variable, $Y$, are perfectly associated; the rule

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

describes the functional relationship in Fig. 1. However, the correlation coefficient is 0.86.

As for the Maximal Information Coefficient, $MIC$ and other mutual information based methods such as $kNN$, $AP$, and $H_{MI}$, discretization of the continuous random variable is required in order to apply formulas such as (2) and (3). However, discretization of a random variable cannot be performed optimally without taking into account the end goals of the discretized data, and in addition, there is information loss in the discretization process [18].

The copula based estimators mentioned above are also not immune to the hybrid scenario. When modeling discrete random variables with copulas, Sklar’s theorem does not guarantee the unicity of the copula $C$ and many copulas satisfy [1] due to ties in the data [17]. This ambiguity is the reason why copula based estimators also have difficulty measuring association between continuous and discrete random variables. The exception to this is $CIM$, which is based on a bias-corrected measure of concordance, $\tau_{KL}$, that can account for continuous and discrete data simultaneously. This is explained in further detail in Section 2.2. of Karra and Mili’s manuscript: Copula Index for Detecting Dependence and Monotonicity between Stochastic Signals [7].

### 3 Estimator Response Curve

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

There are several properties of estimators of the strength of association which are important in characterizing its performance. Theoretically provable properties include Renyi’s seven properties of a dependence measure and the data processing inequality (DPI) for dependence measures [11], [15]. Briefly, Renyi’s seven properties of a measure of dependence are important because they establish the space over which the estimator can be used, the exchangeability of the random variables under consideration, and the ranges of possible outputs of the estimators. Similarly, DPI is important for an estimator because it ensures that when measuring the strength of association between random variables in a causal chain, that indirect causes measure weaker than more direct causes [11]. It is interesting to note that most estimators of mutual information, such as $kNN$, $CME$, $AP$, and $H_{MI}$ do not satisfy Renyi’s seven properties. Conversely, most indices of dependence, including $dCor$, $Ccor$, $CoS$, and $RDC$ are not proven to satisfy the DPI. The notable exception here is the $CIM$, which is proven to satisfy both Renyi’s properties and the DPI [7].

Important empirical properties of an estimator include the statistical power, bias, and consistency. Statistical power is the likelihood that the estimator measures an association between the random variables, when the random variables are statistically dependent. Usually, the power of an estimator is characterized across a range of dependencies and noise levels to fully assess the estimator’s ability to detect different types of dependencies [7]. The importance of power, especially under linear dependence, was originally outlined in Simon and Tibshirani’s work [19]. Statistical bias is the difference between the true value and the estimated value of the quantity to be measured, and can be defined mathematically as $Bias_{\theta}[\hat{\theta}] = E_{x|\theta}[\hat{\theta}] - \theta$, where $\theta$ is the true value of the quantity to be estimated, $\hat{\theta}$ is the estimated value, and $E_{x|\theta}$ is the expected value over the conditional distribution $P(x|\theta)$. However, bias is typically only computed under the scenario of independence, where it is known that the value of the $\theta$ should be 0. Finally, statistical consistency measures the asymptotic properties of the estimator; an estimator is said to be consistent if the estimated value approaches the true value of the estimator as the sample size grows. Stated mathematically, an estimator of $T_n$ of $\theta$ is said to be consistent if $\lim_{n\to\infty} T_n = \theta$.

While these three empirical properties are important to assess an estimator’s performance, none of them capture the notion of the rate of change of an estimated value, as the strength of dependence between the random variables...
changes. If the rate of increase (or decrease) of an estimated value is not proportional to the rate of increase (or decrease) in the dependence strength between the random variables, then in noisy small sample scenarios, there is a nonzero likelihood that the estimator will incorrectly rank the strength of associations between features and output classes in supervised learning. This becomes especially important in the hybrid random variable scenario, where it is already more difficult to measure the strength of association between two random variables \cite{17}. These rates of increase determine the estimators ability to distinguish stronger from weaker relationships, when both relationships are statistically significant. We term the relationship between the actual strength of association between the random variables, \(X\) and \(Y\), and the estimated strength of association between \(X\) and \(Y\) over the entire range of possible strengths of statistical association to be the response curve of an estimator. The response curve can help explain how an estimator will perform when multiple strengths’ of associations need to be measured and ranked, as in mutual information based Markov network discovery and feature selection \cite{12, 15}.

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

Various response curves are shown in Fig. 2. The linear response is shown in purple; in it, the estimator attempts to distinguish between all strengths of dependence equally, while in the convex curves shown with \(o\) markings in green and blue, stronger dependencies are have a higher likelihood of being ranked correctly. Conversely, in the concave response curves denoted with marks in teal and yellow, the estimator has a higher likelihood of ranking weaker dependencies correctly. The curve is scale-invariant, because it examines the rates of change of an estimator, rather than absolute values. It also shows that nonlinear rescaling of an estimators output may affect its ability to correctly rank strengths of association. An example of this is Linfoots informational coefficient of correlation \cite{20}. Here, the mutual information between random variables \(X\) and \(Y\) is rescaled according to the the relationship

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

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

The curves in Fig. 2 also show the variance of the estimated quantity as a function of the strength of dependence. The variance, along with the concavity/convexity of the estimator determines the probability of correctly ranking dependencies between different pairs of random variables. More specifically, the probability of correctly ranking two different pairs of random variables according to their strength of association is inversely proportional to the area encompassed by the rectangle covering the space between the maximum possible value the estimator can take on for the weaker dependency, denoted by \(\hat{\theta}_{\text{weaker}}\), and the minimum possible value the estimator can take on for the stronger dependency, denoted by \(\hat{\theta}_{\text{stronger}}\). For example, if Fig. 2 suppose that the true value of the strength of association between \(X_1\) and \(Y\) is 0.6, and the true value of the strength of association between \(X_2\) and \(Y\) is 0.8, and our goal is to rank them according to their strengths, using an estimator. If the estimator had a response curve similar to the blue curve, then the probability of misranking these dependencies is zero here, because \(\hat{\theta}_{\text{weaker}}\) denoted by the green \(\times\) symbol is less than \(\hat{\theta}_{\text{stronger}}\) denoted by the green four pointed star. Conversely, if the estimator had a response curve similar to the yellow curve, then \(\hat{\theta}_{\text{weaker}}\) denoted by the red \(\times\) symbol is greater than \(\hat{\theta}_{\text{stronger}}\) denoted by the red four pointed star. The probability of misranking these dependencies is nonzero and is proportional to the area given by the area in the red shaded rectangle in Fig. 2. These probabilities do not need to be exactly computed, but identifying them helps to understand how an estimator may perform for these applications. In summary, the estimator response curve is a quick, marginal distribution invariant method to assess how the estimator will perform under various dependencies.

### 4 Synthetic Simulations

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

In our synthetic data simulations, we generate data according to the procedure outlined in Nelsen for generating dependent random variables with arbitrary dependence structures \cite{14}. We begin by generating data from the Gaussian, Frank, Gumbel, and Clayton copulas. These copulas represent different kinds of dependence structures, with the Gaussian modeling linear dependence, and the remaining copulas modeling non-linear dependence patterns such as tail dependence. Each of the copulas has a single parameter,
The results for these simulations, which are the response curves described in Fig. 2, for these estimators with continuous and discrete marginal distributions, are shown in Fig. 3. In them, the x-axis represents the strength of dependence, given by \( \tau \), and the y-axis represents the strength of dependence as measured by the various estimators. It can be seen for all scenarios tested, the state-of-the-art estimators \( kNN, AP \), and \( cME \) all exhibit suboptimal estimator response curves. More concerning, they exhibit less sensitivity when \( \tau \geq 0.5 \), but do not enhance sensitivity with weaker dependencies as would be hoped for from a more concave estimator response curve. In other words, as the strength of association between the random variables increases, the ability of these estimators to distinguish between them decreases in the hybrid random variable scenario! For the no-skew scenario, only the \( H_{MI} \) estimator seems to perform equivalently to the \( C\dot{I}M \) estimator. This suggests that the \( C\dot{I}M \) estimator should be used when measuring the strength of association between hybrid random variables.

5 Real World Data Simulations

In this section, we show how suboptimal estimation of mutual information affects feature selection and classification performance. To accomplish this, we take four real world datasets provided by the NIPS 2003 feature selection challenge [21], and apply the MRMR algorithm to select the most relevant features. The chosen datasets, Arcene, Dexter, Madelon, and Gisette, are binary classification datasets, where the input features are continuous variables, and the output is a discrete variable that can take on two values. We chose these datasets because from the perspective of measuring association between random variables, this presents the most “difficult” case. From a practicality perspective, this case is also highly relevant to machine learning problems, where predictive features are often continuous but the output class to be predicted has only a small number of unique outcomes. Additionally, they represent datasets of various sample sizes and are from different fields in science. The Arcene dataset is 100 samples and contains 10000 features representing mass-spectrometric data from cancerous and non-cancerous cells. The Dexter dataset has 300 samples, and contains 20000 bag of words features for text classification. The Madelon dataset has 2000 samples and is an artificial dataset with 500 features to classify clusters in a five-dimensional hypercube. Finally, the Gisette dataset is 3000 samples of 5000 features representing text classification features to distinguish between the digits 4 and 9.

With these datasets, we perform feature selection with the maximum relevance, minimum redundancy (MRMR) feature selection algorithm [13]. Briefly, MRMR is an approach to feature selection that utilizes mutual information to assess the relevance and redundancy of a set of features, given an output prediction class. The goal of MRMR is to solve

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

where \( X_S = \{X_i : i \in S\} \), \( k \) is the number of features to be selected, and \( I(X, Y) \) is the mutual information between

\( \theta \), which controls the strength of dependence, and corresponds directly to the mutual information contained within the dependence structure. Because the scale and support of \( \theta \) varies between different copula families, our simulations modulate the strength of dependence between the random variables through the rank correlation measure, \( \tau \), which has a one-to-one correspondence to \( \theta \) for every copula that was simulated. After picking a certain copula family and a value of \( \tau \), we generate random variates \( u \) and \( v \). After generating these random variates \( u \) and \( v \), we apply the inverse transform \( F^{-1}(U) = X \) and \( G^{-1}(V) = Y \) respectively to generate \( x \) and \( y \). Here, we choose three different cases for \( X \) and \( Y \) to simulate real-world scenarios which may arise.

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

**Fig. 2. Illustration of the various estimator response curves.** The linear response is shown in purple; in it, the estimator attempts to distinguish between all strengths of dependence equally, while in the convex curves shown with \( o \) markings in green and blue, stronger dependencies are have a higher likelihood of being ranked correctly. Conversely, in the concave response curves denoted with marks in teal and yellow, the estimator has a higher likelihood of ranking weaker dependencies correctly. The red shaded rectangle shows a region of non-zero probability that an estimator with the yellow response curve would have in misranking dependence strengths. The red shaded rectangle shows a region of non-zero probability mass function of \( 0 \) by the vector \( u \).
the random variables $X$ and $Y$. However, measuring the mutual information of increasingly large dimensions of $k$ is unfeasible due to the curse of dimensionality. MRMR attempts to overcome this by solving the optimization problem given by

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

(5)

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

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

(6)

The first term in (6) represents the relevance of feature $X_j$ to output $Y$, and the second term represents the redundancy between the selected features $X_i$ and the current feature under consideration, $X_j$. Because the MRMR algorithm is based on measuring mutual information between input features (continuous or discrete) and the output class (typically discrete with small cardinality), our goal in these experiments is to understand how suboptimal estimation of mutual information affects MRMR. It is readily seen from (5) and (6) that more accurate estimation of mutual information should yield better feature selection results.

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

The results for these experiments are shown in Fig. 4. For each dataset, we show the 10-fold cross validation score of a kNN classifier as we increase the number of features that were selected, in order of importance as provided by the MRMR algorithm for each DPI satisfying estimator. We show the results for each dataset, where we skew the number of positive examples to be 50%, 75%, and 100% of the number of negative examples. The output class distribution for each simulation is shown in the inset plot in Fig. 4. It is seen that for three of the four datasets tested, the CIM estimator compares favorably to all other estimators of mutual information considered. The results corroborate the findings in Section 4, where it was seen that with synthetic test vectors, the CIM estimator compares favorably to other DPI satisfying measures of the strength of association for hybrid random variables.

6 Conclusion

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

ACKNOWLEDGMENTS

The authors would like to thank the Hume Center at Virginia Tech for its support.
References

[1] K. Pearson, “Note on Regression and Inheritance in the Case of Two Parents,” Proceedings of the Royal Society of London, vol. 58, pp. 240–242, 1895.

[2] D. Reshef, Y. Reshef, H. Finucane, S. Grossman, G. McVean, P. Turnbaugh, E. Lander, M. Mitzenmacher, and P. Sabeti, “Detecting Novel Associations in Large Data Sets,” Science, vol. 334, no. 6062, pp. 1518–1524, 2011.

[3] D. Lopez-Paz, P. Henning, and B. Schölkopf, “The Randomized Dependence Coefficient,” in Advances in Neural Information Processing Systems 26. Curran Associates, Inc., 2013.

[4] G. Székely, M. Rizzo, and N. Bakirov, “Measuring and Testing Dependence by Correlation of Distances,” The Annals of Statistics, vol. 35, no. 6, pp. 2769–2794, 12 2007.

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

[6] M. Ben Hassine, L. Mili, and K. Karra, “A Copula Statistic for Measuring Nonlinear Multivariate Dependence,” 2016.

[7] K. Karra and L. Mili, “Copula Index for Detecting Dependence and Monotonicity between Stochastic Signals,” Arxiv PrePrint, 2018.

[8] A. Kraskov, H. Stögbauer, and P. Grassberger, “Estimating Mutual Information,” Phys. Rev. E, vol. 69, 2004.

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

[10] G. Darbellay and I. Vajda, “Estimation of the Information by an Adaptive Partitioning of the Observation Space,” IEEE Transactions on Information Theory, vol. 45, no. 4, pp. 1315–1321, May 1999.

[11] J. Kinney and G. Atwal, “Equitability, Mutual Information, and the Maximal Information Coefficient,” Proceedings of the National Academy of Sciences, vol. 111, no. 9, pp. 3354–3359, 2014.

[12] A. Margolin, I. Nemenman, K. Basso, C. Wiggins, G. Stolovitzky, R. Favera, and A. Califano, “ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context,” BMC Bioinformatics, vol. 7, no. 1, p. S7, 2006.

[13] H. Peng, F. Long, and C. Ding, “Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 27, no. 8, pp. 1226–1238, 2005.

[14] R. Nelsen, An Introduction to Copulas. Springer-Verlag New York, 2006.

[15] A. Rényi, “On Measures of Dependence,” Acta Mathematica Academiae Scientiarum Hungarica, vol. 10, no. 3, pp. 441–451, 1959.

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

[17] C. Genest and J. Nešlehová, “A Primer on Copulas for Count Data,” ASTIN Bulletin, 2007.

[18] S. Garcia, J. Luengo, S. J., L. V., and H. F., “A Survey of Discretization Techniques: Taxonomy and Empirical Analysis in Supervised Learning,” IEEE Transactions on Knowledge and Data Engineering, vol. 25, pp. 734–750, 2013.

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

[20] E. Linfoot, “An informational measure of correlation,” Information and Control, 1957.

[21] I. Guyon, S. Gunn, A. Ben-Hur, and G. Dror, “Result analysis of the nips 2003 feature selection challenge,” in Advances in neural information processing systems, 2005, pp. 545–552.