Unbiased Multivariate Correlation Analysis

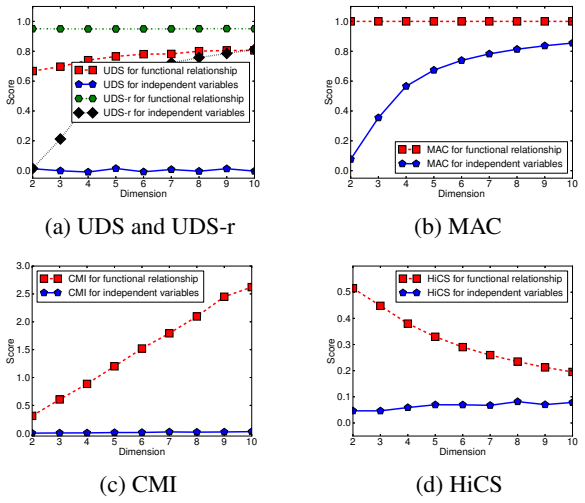
Yisen Wang[†,‡], **Simone Romano**[‡], **Vinh Nguyen**[‡], **James Bailey**[‡], **Xingjun Ma**[‡], **Shu-Tao Xia**[†]
[†] Dept. of Computer Science and Technology, Tsinghua University, China
[‡] Dept. of Computing and Information Systems, University of Melbourne, Australia
wangys14@mails.tsinghua.edu.cn
{simone.romano, vinh.nguyen, baileyj, xingjun.ma}@unimelb.edu.au; xiast@sz.tsinghua.edu.cn

Abstract

Correlation measures are a key element of statistics and machine learning, and essential for a wide range of data analysis tasks. Most existing correlation measures are for pairwise relationships, but real-world data can also exhibit complex *multivariate* correlations, involving three or more variables. We argue that multivariate correlation measures should be *comparable*, *interpretable*, *scalable* and *unbiased*. However, no existing measures satisfy all these requirements. In this paper, we propose an unbiased multivariate correlation measure, called UMC, which satisfies all the above criteria. UMC is a cumulative entropy based non-parametric multivariate correlation measure, which can capture both linear and non-linear correlations for groups of three or more variables. It employs a correction for chance using a statistical model of independence to address the issue of bias. UMC has high interpretability and we empirically show it outperforms state-of-the-art multivariate correlation measures in terms of statistical power, as well as for use in both subspace clustering and outlier detection tasks.

Introduction

Analysing correlations is a fundamental task in both statistics and machine learning. It has applications in many real-world learning tasks, *e.g.*, feature selection (Brown et al., 2012), subspace search (Nguyen et al., 2013), causal inference (Bareinboim, Tian, and Pearl, 2014) and subspace clustering (Kriegel, Kröger, and Zimek, 2009). For the setting of continuous variables (as opposed to discrete), most existing correlation measures focus on pairwise relationships. For example, Pearson's correlation coefficient detects bivariate linear correlations, and Maximal Information Coefficient (MIC) (Reshef et al., 2011) detects both linear and non-linear bivariate correlations. However, real-world data often contains three or more variables which can exhibit multivariate (higher-order) correlations. If bivariate based measures are used to identify multivariate correlations, through pairwise aggregation, multivariate correlations can potentially be overlooked. For example, it has been shown that genes may reveal only a weak correlation with a disease when considered individually, while the correlation for a group of genes may be very strong (Zhang et al., 2008).



(a) UDS and UDS-r (b) MAC

(c) CMI (d) HiCS

Figure 1: Raw scores for existing multivariate correlation measures. For interpretability and unbiasedness, a measure's score should be a horizontal line with value '1' for a functional relationship as the number of variables varies. It should be a horizontal line with value '0' for sets of independent variables of increasing size. *None of the four measures exhibits both behaviours.*

Identifying multivariate correlations is therefore an important activity in data analysis. Measures for multivariate correlation analysis have three important applications (Romano et al., 2016), (a) detection; (b) quantification; and (c) ranking. We argue that these applications require a multivariate correlation measure to meet the following criteria: (1) *Comparability* — The correlation score for a small set of variables can be meaningfully compared against the score for a large set of variables. The score should also lie within a predetermined range such as [0,1]; (2) *Interpretability* — The correlation score should be 0 for a set of independent variables and 1 for a set of variables having a strong relationship; (3) *Scalability* — The score should be efficient to compute as the number of variables or data size increases; (4) *Unbiasedness* — The score should be constant as the number of variables increases, if maintaining a fixed degree of relationship among them.

Recently proposed multivariate correlation measures, such as Universal Dependency Score (UDS) (Nguyen, Mandros, and Vreeken, 2016), Multivariate mAximal Correlation (MAC) (Nguyen et al., 2014), Cumulative Mutual Information (CMI) (Nguyen et al., 2013) and HiCS (Keller, Muller, and Bohm, 2012), only partially satisfy the above 4 criteria. For example, given a data set of 1000 data points with d variables $\{X_i\}_{i=1}^{d}$, we consider two simple correlations: a strong functional one (all X_i's are the same) and a weak independent one (all X_i's are random variables, drawn independently of each other). For good interpretability, we expect the measure score to be close to 1 for the functional relationship and to be 0 for a set of independent variables. More importantly, the measure should be stable for a fixed correlation as dimensionality (the number of variables) increases, *i.e.*, unbiased. However, Figure 1 shows that none of the above measures is able to satisfy these basic requirements. UDS, UDS-r (without regularization), MAC and CMI do not output a stable score, showing an increasing trend either for the functional relationship or independent variables, while HiCS demonstrates a decreasing trend for the functional relationship. We summarize the behaviour of these multivariate correlation measures in Table 1 and see that none of them satisfies all the above 4 criteria.

Table 1: The behaviour of multivariate correlation measures

Measure	Compar.	Interpret.	Scal.	Unbias.
UDS	$\surd$	$\star\star$	$\surd$	$\times$
UDS-r	$\surd$	$\star\star$	$\surd$	$\times$
MAC	$\surd$	$\star\star$	$\times$	$\times$
CMI	$\times$	$\star$	$\surd$	$\times$
HiCS	$\surd$	$\star$	$\times$	$\times$
UMC	$\surd$	$\star\star\star$	$\surd$	$\surd$

Motivated by these observations, we propose an Unbiased Multivariate Correlation measure (UMC) that satisfies the above 4 criteria. UMC is designed by taking a unified view of existing correlation measures. It employs cumulative entropy that permits non-parametric computation on (continuous) empirical data (Rao et al., 2004; Di Crescenzo and Longobardi, 2009a). To address the interpretability and bias issues, we extend the framework proposed in (Romano et al., 2016) to multivariate correlation measures, and analytically derive the expected value of the conditional cumulative entropy under a statistical model of independence and use it for bias correction of UMC. Analytical computation (as opposed to Monte Carlo simulation) of this expected value ensures UMC can maintain a good degree of computational efficiency. Overall, UMC is purely non-parametric and suitable for capturing both linear and non-linear correlations amongst multiple variables, while achieving *comparability*, *interpretability*, *scalability* as well as *unbiasedness*.

Our contributions in this paper are fourfold: (1) We provide a unified view of existing multivariate correlation measures and the connections between them; (2) We theoretically and empirically analyse bias issues for existing multivariate correlation measures; (3) We propose an unbiased measure for multivariate correlation analysis — UMC; and (4) We empirically demonstrate that UMC achieves the best performance amongst existing multivariate correlation measures in terms of power and performance based on experiments with both synthetic and real-world data.

A Unified View of Correlation Measures

In this section, we review existing multivariate correlation measures in a unified way to highlight their connections. Consider a data set $\mathcal{D}$ with d real-valued variables $\{X_i\}_{i=1}^{d}$ and n data points. We also refer to d as the dimensionality of the data set. We regard each X_i as a random variable, characterized by its Probability Distribution Function (PDF) $p(X_i)$. A multivariate correlation measure M quantifies how much the relation of $\mathcal{D} = \{X_i\}_{i=1}^{d}$ deviates from statistical independence (Te Sun, 1980). That is to say, how much their joint probability distribution differs from the product of their marginal probability distributions,

$$M(\mathcal{D}) = \mathit{diff}\Big(p(X_1, \ldots, X_d), \prod_{i=1}^{d} p(X_i)\Big). \tag{1}$$

If $\mathit{diff}()$ is instantiated as the KL-divergence (Kullback and Leibler, 1951), Eq. (1) will become the Total Correlation (TC) measure (Te Sun, 1978),

$$\begin{aligned} TC(\mathcal{D}) &= KL\Big(p(X_1, \ldots, X_d) \parallel \prod_{i=1}^{d} p(X_i)\Big) \\ &= \Big[\sum_{i=1}^{d} H(X_i)\Big] - H(X_1, \ldots, X_d). \end{aligned} \tag{2}$$

However, the joint Shannon entropy $H(X_1, \ldots, X_d)$ requires estimation of the joint probability mass function $p(X_1, \ldots, X_d)$, which suffers from the empty space problem (Aggarwal and Philip, 2001) in high dimensional space. To avoid using the joint probability distribution, according to the factorization property of the joint probability, the KL divergence in Eq. (2) can be factorized as:

$$\begin{aligned} KL\Big(p(X_1, \ldots, X_d) \parallel \prod_{i=1}^{d} p(X_i)\Big) &= KL\Big(p(X_2|X_1) \parallel p(X_2)\Big) \\ + \cdots + KL\Big(p(X_d|X_1, \ldots, X_{d-1}) &\parallel p(X_d)\Big). \end{aligned} \tag{3}$$

This is just one possible factorization of the KL-divergence. In general, different variable orderings will lead to different factorization formulas. More importantly, this factorization inspires an approximation of Eq. (1) as:

$$M(\mathcal{D}) \sim \sum_{i=2}^{d} \mathit{diff}\Big(p(X_i), p(X_i|X_1, \ldots, X_{i-1})\Big), \tag{4}$$

where the conditional distributions are computed progressively with $i-1$ variables, partially mitigating the high dimensional empty space problem. Furthermore, this kind of progressive aggregation mechanism from small variable sets to large helps the measure scale well to high dimensionality.

Despite the advantages of Eq. (4), it is sensitive to the ordering of the variables $\{X_1, \ldots, X_d\}$. The maximal correlation analysis approach (Breiman and Friedman, 1985; Rao

et al., 2011) suggests an order-free technique as follows. Let $\mathcal{F}_d$ be the set of bijective functions σ: $\{1,\ldots,d\} \rightarrow \{1,\ldots,d\}$, we have:

$$M(\mathcal{D}) \sim \max_{\sigma \in \mathcal{F}_d} \sum_{i=2}^{d} diff\left(p(X_{\sigma(i)}), p(X_{\sigma(i)}|X_{\sigma(1)},\ldots,X_{\sigma(i-1)})\right). \quad (5)$$

Existing multivariate correlation measures can now be interpreted within this framework. HiCS is based on a specific marginalized form of Eq. (1), *i.e.*, $diff(p(X_i), p(X_i|\{X_1,\ldots,X_d\}\backslash\{X_i\}))$, and computed as an average across multiple iterations where X_i is randomly picked; MAC is based on Eq. (2) using Shannon entropy estimated through data discretization; CMI and UDS are based on Eq. (5), each instantiating $diff()$ with the cumulative entropy $h(X_i)$ (defined below), $CMI = \max_{\sigma \in \mathcal{F}_d} \sum_{i=2}^{d} [h(X_{\sigma(i)}) - h(X_{\sigma(i)}|X_{\sigma(1)},\ldots,X_{\sigma(i-1)})]$, and $UDS = CMI / \sum_{i=2}^{d} h(X_{\sigma(i)})$, which is a normalized version of CMI. Our proposed measure, UMC, is also based on Eq. (5) using cumulative entropy to instantiate $diff()$. However, different from UDS, we adopt a tight upper bound as the normalization factor, and employ a correction for chance using a statistical model of independence. These assist in addressing the drawbacks of the existing measures mentioned in Table 1.

Cumulative Entropy

It is common practice to discretize real-valued data to estimate Shannon entropy, but this can lead to loss of information. On the other hand, differential entropy, which works directly on continuous variables, has been shown to be problematic for assessing correlations (Rao et al., 2004). In this paper, we employ cumulative entropy (Di Crescenzo and Longobardi, 2009a), which can be estimated on real-valued data without data discretization. The CMI and UDS measures are also based on cumulative entropy.

The cumulative entropy of a continuous random variable Z, denoted as $h(Z)$, is defined as:

$$h(Z) = -\int P(Z \leq z) \log P(Z \leq z) dz, \quad (6)$$

where $P(Z \leq z)$ is the Cumulative Distribution Function (CDF) of Z. Similar to Shannon entropy, $h(Z)$ captures the degree of uncertainty in Z. The only difference is that the cumulative entropy is defined on the CDF. Since $0 \leq P(Z \leq z) \leq 1$, it follows that $h(Z) \geq 0$.

The conditional cumulative entropy of any real-valued random variable Z given a random variable Y is defined as:

$$h(Z|Y) = \int h(Z|y) p(y) dy, \; y \in \text{domain}(Y), \quad (7)$$

where

$$h(Z|y) = -\int P(Z \leq z|y) \log P(Z \leq z|y) dz. \quad (8)$$

It has two important properties (Rao et al., 2004; Di Crescenzo and Longobardi, 2009a).

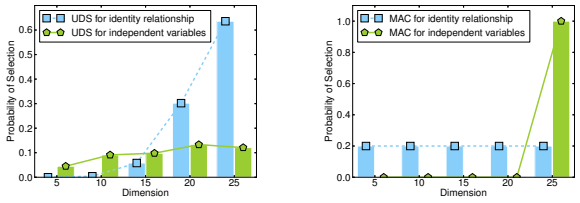
Theorem 1. *$h(Z|Y) \geq 0$ with equality iff Z is a function of Y.*

Theorem 2. *$h(Z|Y) \leq h(Z)$ with equality iff Z is statistically independent of Y.*

Dimensionality Bias Analysis

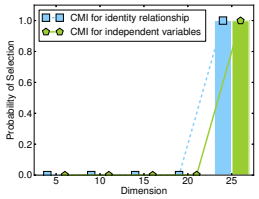
The presence of bias is harmful for a multivariate correlation measure, since it can make it hard to interpret and unreliable in real-world learning tasks, especially for quantification and ranking related applications. In this section, we identify and analyse the issue of dimensionality bias empirically and theoretically.

Empirically, to illustrate the dimensionality bias issue, we generate 5 data sets with $d = [5, 10, 15, 20, 25]$ variables and $n = 1000$ data points for two correlations: i) identity relationship (all variables are identical) and ii) independent set of variables (all variables are randomly and independently drawn). For each existing correlation measure, we select the data set which achieves the highest correlation score, and then repeat this process 10,000 times. The probability of selecting a data set according to each possible dimensionality is shown in Figure 2.

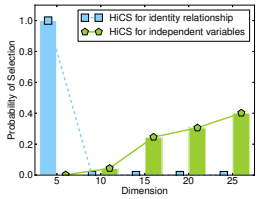


(a) UDS

(b) MAC



(c) CMI



(d) HiCS

Figure 2: Dimensionality bias of correlation measures for identity relationship and independent variables. The ideal case would be an equal 20% probability for each possible dimensionality. The figure is best viewed in color.

Given that the correlation scores should be the same for every data set of all dimensionalities, they should each have an equal chance of being selected (*i.e.*, 20%). However, Figure 2 clearly demonstrates that all these measures are biased to data sets with high dimensionalities for the independent variables. On the other hand, for the identity relationship, except MAC which almost shows no bias, all the other measures suffer from different biases, *i.e.*, UDS and CMI are biased to high dimensionalities, while HiCS is biased to low dimensionalities.

Theoretically, CMI has been proved to be biased to high dimensionality in (Nguyen, Mandros, and Vreeken, 2016). Here, we prove that UDS is also biased to high dimensionality in some cases (proof in the *Supplementary Material A* [1]).

[1] https://sites.google.com/site/umcsupplementary/home

Theorem 3. *$UDS(X_1, \ldots, X_{d+1}) \geq UDS(X_1, \ldots, X_d)$ whenever $X_{d+1} = X_i, i \in \{1, 2, \ldots, d\}$. That is to say, UDS scores a set of variables higher when there are repeated (redundant) variables.*

Unbiased Multivariate Correlation Measure

In this section, we propose UMC, an Unbiased Multivariate Correlation measure, which satisfies *comparability*, *interpretability*, *scalability* and *unbiasedness*. Specifically, a statistical model of independence is employed to address the dimensionality bias issue, providing a constant baseline and good interpretability. The factorization property from Eq. (5) is adopted to guarantee good scalability to high dimensionality, and a tight upper bound is used for normalization to achieve good comparability. Furthermore, an analytical computation formula for the expected value of the conditional cumulative entropy under the statistical independence model is derived.

Statistical Model of Independence

The technique of correcting bias using a statistical model of independence has already been successfully used in several pairwise dependency measures, *e.g*, Adjusted Mutual Information (AMI) (Vinh, Epps, and Bailey, 2009) and Adjusted Rand Index (ARI) (Hubert and Arabie, 1985) in the context of comparing two clusterings. The common statistical model of independence they use is the permutation model (Hubert and Arabie, 1985). Herein, we adopt the same permutation model of independence, extending it to multivariate correlation measures.

Definition 1 (The permutation model of independence)**.** *Samples are generated by permuting the n data points $\{X_i(k)\}_{k=1}^n$ of each variable X_i in the data set $\mathcal{D} = \{X_i\}_{i=1}^d$ independently. Let τ_i denote an independent random permutation of n indices for each X_i, then $\mathcal{D}^{\text{perm}} = \{\big(X_1(\tau_1(k)), \ldots, X_d(\tau_d(k))\big)\}_{k=1}^n$. In other words, if $\mathcal{D}$ is in a $n \times d$ dimensional matrix format and we independently permute the elements of each column, then the distributions of the columns (the marginal distribution of each X_i) remain the same, but columns become independent of each other.*

The expected value $\mathbb{E}_0$ of a correlation measure under this model can be regarded as its "correlation-by-chance" value, *i.e.*, the bias, so it can be used for adjusting the measure to achieve a constant baseline. This is the strategy we adopt for UMC, explained below.

Definition of UMC

Based on Eq. (5), using the cumulative entropy, the tight upper bound and the expected value under the permutation model of independence, we obtain:

Definition 2 (Unbiased Multivariate Correlation, UMC)**.**

$$UMC(\mathcal{D}) = \max_{\sigma \in \mathcal{F}_d} \frac{\sum_{i=2}^d [h(X_{\sigma(i)}) - h(X_{\sigma(i)}|X_{\sigma(1)}, \ldots, X_{\sigma(i-1)}) - \mathbb{E}_0]}{\sum_{i=2}^d [A - \mathbb{E}_0]} \tag{9}$$

where σ is an ordering of variables, A denotes a tight upper bound $A = h(X_{\sigma(i)}) - \min(\{h(X_{\sigma(i)}|X_{\sigma(1)}, \ldots, X_{\sigma(i-1)})\}_{i=2}^d)$ and $\mathbb{E}_0$ represents the expected value under the permutation model $\mathbb{E}_0[h(X_{\sigma(i)}) - h(X_{\sigma(i)}|X_{\sigma(1)}, \ldots, X_{\sigma(i-1)})]$.

UMC belongs to the class of maximal correlation analysis. By considering the maximum value, we aim at uncovering the best correlation score of the variables involved, and making UMC variable-order invariant. In addition, UMC possesses several important properties required from a good multivariate correlation measure (proofs in *Supplementary Material B*): 1) $UMC(X_1, \ldots, X_d)$ is greater than or equal to 0 on average, and attains 1 as maximum value; 2) $UMC(X_1, \ldots, X_d)$ is equal to 0 on average when $\{X_1, \ldots, X_d\}$ are statistically independent; 3) $UMC(X_1, \ldots, X_d) = 1$ if there exists X_i such that each in $\{X_1, \ldots, X_d\} \backslash \{X_i\}$ is a function of X_i.

According to UMC's definition, a search is performed over all variable orderings, *i.e.*, $d!$ possible ones, for the maximal value. However, such a search is impractical. Therefore, we adopt the same heuristics as the UDS measure (Nguyen, Mandros, and Vreeken, 2016) by specifying an approximation of the optimal ordering, avoiding the exhaustive search. More specifically, the approximate optimal ordering $\sigma^* \in \mathcal{F}_d$ is $h(X_{\sigma^*(1)}) \geq \cdots \geq h(X_{\sigma^*(d)})$, *i.e.*, orders variables in descending order of cumulative entropy value. In the following, to simplify, we will assume the optimal ordering of $\mathcal{D}$ is $\{X_1, \ldots, X_d\}$ satisfying $h(X_1) \geq \cdots \geq h(X_d)$. Since $h(X_i)$'s are in decreasing order and $h(X_i|X_1, \ldots, X_{i-1}) \leq h(X_i)$, we can conclude that $\min(\{h(X_i|X_1, \ldots, X_{i-1})\}_{i=2}^d)$ is the last conditional cumulative entropy $h(X_d|X_1, \ldots, X_{d-1})$. Therefore we do not need to search for the minimum. These techniques make UMC more computationally efficient.

Computing the Cumulative Entropy

The unconditional cumulative entropy can be calculated in closed-form (Di Crescenzo and Longobardi, 2009b). Let $X_i(1) \leq \cdots \leq X_i(n)$ be the ordered data points of X_i, we have

$$h(X_i) = -\sum_{j=1}^{n-1} (X_i(j+1) - X_i(j)) \frac{j}{n} \log \frac{j}{n}. \tag{10}$$

For computing conditional cumulative entropy, an optimal correlation-aware discretization method has been proposed, which does not break the correlation structures in the data (Reshef et al., 2011; Nguyen, Mandros, and Vreeken, 2016). The idea is as follows: since the ordering of variables has been fixed, firstly, we search for the discretization of X_1 that maximizes $h(X_2) - h(X_2|X_1)$. Then, we fix the previous discretization of X_1, and only search for the optimal discretization of X_2 that maximizes $h(X_3) - h(X_3|X_1, X_2)$. We follow this rationale till we obtain $h(X_d) - h(X_d|X_1, \ldots, X_{d-1})$. Without re-discretizing any previous variables, a substantial amount of computation time can be saved. The optimal discretization at each step can be efficiently found by dynamic programming (see Reshef et al. (2011); Nguyen et al. (2014); Nguyen, Mandros, and

Vreeken (2016) for proofs). In addition, to avoid *overfitting*, we set the minimum number of data points falling into each bin cell to n^ϵ, with $0 < \epsilon < 1$. This guarantees a sufficient number of points for computing the conditional cumulative entropy, similar to the technique in (Wang et al., 2016).

Analytical Derivation of the Expected Value $\mathbb{E}_0$

In this section, we derive an analytical formula for the expected value $\mathbb{E}_0$ under the permutation model of independence. As shown in Eq. (9), $\mathbb{E}_0$ consists of two parts: $\mathbb{E}_0[h(X_i)]$ and $\mathbb{E}_0[h(X_i|X_1,\ldots,X_{i-1})]$. Under the permutation model, $h(X_i)$ is invariant, so only $\mathbb{E}_0[h(X_i|X_1,\ldots,X_{i-1})]$ needs to be computed. To simplify, let $X = X_i$ and $V = \{X_1,\ldots,X_{i-1}\}$. $h(X|V)$ is computed using the m bin cells $\{v_1,\ldots,v_m\}$ obtained with dynamic programming. Therefore,

$$h(X|V) = \sum_{\lambda=1}^{m} h(X|v_\lambda)\frac{|v_\lambda|}{n}, \tag{11}$$

where $|v_\lambda|$ is the number of points in the bin cell v_λ. The expected value under the permutation model is equal to:

$$\mathbb{E}_0[h(X|V)] = \mathbb{E}_0\Big[\sum_{\lambda=1}^{m} h(X|v_\lambda)\tfrac{|v_\lambda|}{n}\Big] = \tfrac{1}{n}\sum_{\lambda=1}^{m}|v_\lambda|\mathbb{E}_0[h(X|v_\lambda)], \tag{12}$$

where the conditional cumulative entropy $\mathbb{E}_0[h(X|v_\lambda)]$ for each bin cell v_λ is computed according to Eq. (10):

$$\mathbb{E}_0[h(X|v_\lambda)] = \mathbb{E}_0\Big[-\sum_{j=1}^{|v_\lambda|-1}(X(j+1)-X(j))\frac{j}{|v_\lambda|}\log\frac{j}{|v_\lambda|}\Big]. \tag{13}$$

The key task is to compute $\mathbb{E}_0[X(j+1)-X(j)]$. In the absence of any information about the distribution of X, and considering that $|v_\lambda|$ is small because of the dynamic programming optimization, it is natural to assume that $X|v_\lambda$ follows a maximum entropy distribution, *i.e.*, the uniform distribution. Thus, if $max = \max(X)$ and $min = \min(X)$, we get $\mathbb{E}_0[X(j+1)-X(j)] = \frac{max-min}{|v_\lambda|+1}$. We estimate the range $max - min$ using the interquartile range IQR: $Q_3 - Q_1$. Based on it, we propose the following approximation: $\mathbb{E}_0[X(j+1)-X(j)] \approx \frac{1.5(Q3-Q1)}{|v_\lambda|+1}$. This approximation makes the estimate more accurate when X is not uniform. Indeed $\frac{Q3-Q1}{max-min} = \frac{1}{2}$ for the uniform distribution, and $\frac{Q3-Q1}{max-min} \approx \frac{1}{1}$ for an extremely skewed distribution.

Finally, following Eq. (13), we get

$$\mathbb{E}_0[h(X|v_\lambda)] = -\frac{1.5(Q3-Q1)}{|v_\lambda|+1}\sum_{j=1}^{|v_\lambda|-1}\frac{j}{|v_\lambda|}\log\frac{j}{|v_\lambda|} = \frac{1.5(Q3-Q1)}{|v_\lambda|+1}\left(\frac{\zeta'(-1)-\zeta'(-1,|v_\lambda|)}{|v_\lambda|}+\frac{(|v_\lambda|-1)\log|v_\lambda|}{2}\right) \tag{14}$$

where $\zeta'(s)$ is the derivative of the Riemann zeta function, and $\zeta'(s,a)$ is the partial derivative (Elizalde, 1986) of the generalized Riemann zeta function with respect to the first argument (Di Crescenzo and Longobardi, 2009a).

By plugging Eq. (14) into Eq. (12), we derive the expectation value $\mathbb{E}_0[h(X|V)]$. Eventually, we obtain the analytical formula $\mathbb{E}_0$:

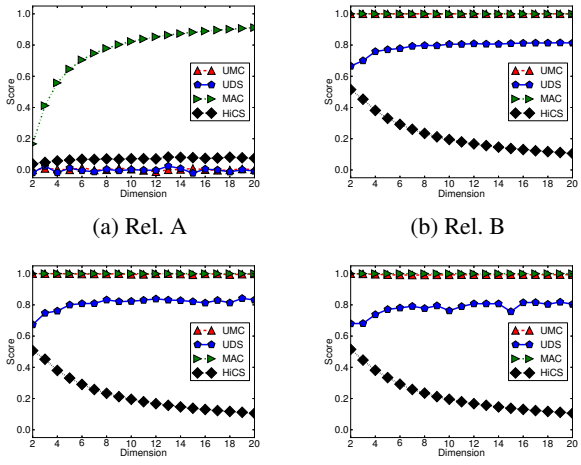
$$\mathbb{E}_0[h(X)-h(X|V)] = h(X) - \frac{1}{n}\sum_{\lambda=1}^{m}|v_\lambda|\times \frac{1.5(Q3-Q1)}{|v_\lambda|+1}\left(\frac{\zeta'(-1)-\zeta'(-1,|v_\lambda|)}{|v_\lambda|}+\frac{(|v_\lambda|-1)\log|v_\lambda|}{2}\right) \tag{15}$$

Experiments

In this section, we empirically assess UMC. Firstly, we study UMC's performance on synthetic data. Secondly, we incorporate UMC into a state-of-the-art subspace beam search algorithm (Keller, Muller, and Bohm, 2012) to mine correlated subspaces for subspace clustering and outlier detection on real data. Our baselines are the existing multivariate correlation measures: UDS (Nguyen, Mandros, and Vreeken, 2016), MAC (Nguyen et al., 2014), CMI (Nguyen et al., 2013) and HiCS (Keller, Muller, and Bohm, 2012). For each baseline, the parameters are optimized following its respective original paper. For UMC, we set $\epsilon = 0.3$. Note that all the results are the average of 100 trials.

Performance on Synthetic Data

Raw Score. We craft different correlations among the d variables of each data set $\mathcal{D} = \{X_i\}_{i=1}^{d}$, including: (1) **Rel. A**: independent variables. Each X_i is randomly and independently sampled from $[0,1]$. (2) **Rel. B**: identity relationship. X_1 is uniformly sampled from $[0,1]$ and $X_i = X_1$ for all $i = 2,3,...,d$. (3) **Rel. C**: power law relationship. X_1 is uniformly sampled from $[0,1]$ and $X_i = (X_1)^i$ for all $i = 2,3,...,d$. (4) **Rel. D**: sin relationship. X_1 is uniformly sampled from $[0,1]$ and $X_i = sin(X_{i-1})$ for all $i = 2,3,...,d$. We vary $d \in [2,20]$, and set $n = 1000$.



(a) Rel. A (b) Rel. B

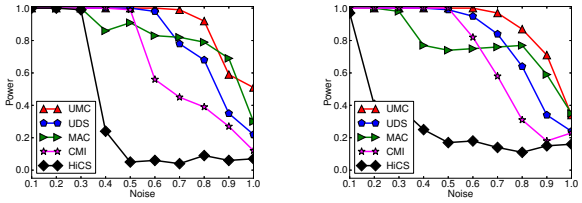
(c) Rel. C (d) Rel. D

Figure 3: Raw measure scores on synthetic data with different relationships. UMC shows a flat line at 0 for (a) and flat line at 1 for each of (b-d), which is the desired behavior. The figure is best viewed in color.

The results are shown in Figure 3. We did not include CMI here, as CMI is not normalized. Figure 3 shows that UMC has excellent interpretability, outputting 0 for independent variables as dimensionality increases. Furthermore, UMC outputs 1 for functional relationships as dimensionality varies. This is in contrast to the other measures, none of which behaves correctly across all four cases. From this perspective, UMC successfully addresses the dimensionality bias issue compared to the existing measures.

Dimensionality Bias. UMC is almost unbiased, showing almost equal probability of selecting a data set regardless of dimensionality (more details are in the *Supplementary Material C*).

Statistical Power. We test the power of the measures with similar methodology to (Reshef et al., 2011) by adding Gaussian noise within the range $[0.1, 1.0]$. Figure 4 reports the results for Rel. B and Rel. C on the data set with 1000 data points and 25 variables. The power for HiCS degenerates very quickly because it cannot work in high dimensional spaces (see Figure 1). UMC, UDS and CMI all make use of the factorization property in Eq. (5), so they can cope well with high dimensionality. UMC outperforms UDS and CMI on noisier data sets. MAC does not work well because it outputs a high score for independent variables, which makes it hard to identify when there exists a real relationship.



(a) Rel. B (b) Rel. C

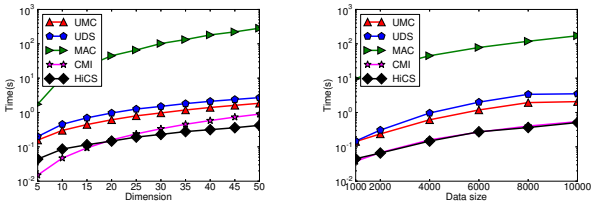
Figure 4: Power of the measures with regards to noise. The higher the better. The figure is best viewed in color.

Scalability. We examine the scalability of the measures with regards to dimensionality and data size on a PC platform with Intel Core i7-3770 CPU and 32GB RAM. To measure the running time versus the data dimensionality, we generate data sets with 4000 data points and varying dimensionality. To quantify the efficiency with regards to the data size, we generate data sets with 20 variables and varying data size. Figure 5 demonstrates that UMC is very computationally efficient. In particular, it is even faster than UDS.

Performance on Real-World Data

We evaluate the performance of UMC on real-world data sets involving more complex correlations, considering two typical applications: subspace clustering and outlier detection. Due to space limitations, detailed results for the latter task are presented in the *Supplementary Material D*.

Subspace Clustering. Müller et al. (2009) showed that mining clusters from the subspaces with high correlation usually produces more meaningful results. Thus, measuring the correlation of a subspace is a critical aspect for subspace clustering performance.



(a) Runtime to dimensionality (b) Runtime to data size

Figure 5: Runtime with varying dimensionality and data size (Time axis is in log scale).

Following the existing literature (Müller et al., 2009), we plug the correlation measures into a state-of-the-art subspace search method (Keller, Muller, and Bohm, 2012) to find highly correlated subspaces, and then evaluate the quality of these subspaces through DBSCAN (Ester et al., 1996) clustering on the top 50 subspaces with highest measure score. The F1 score (Müller et al., 2009) is adopted as the performance metric to compare clustering results against the ground truth.

We test 11 real UCI data sets[2] widely used for benchmarking in the clustering community and previous work on correlation measures, using their class labels as ground truth. Table 2 shows that UMC achieves the highest subspace quality compared to all other measures. We believe the reason lies in the unbiasedness of UMC, which enables it to correctly find the truly correlated subspaces, compared to other measures that may assign inflated correlation scores. Specifically, UDS and CMI are biased to subspaces with higher dimensionality regardless of whether they possess true correlations. MAC and HiCS output very close scores for subspaces with correlations or without correlations. By applying a Friedman test (Demšar, 2006) at the 0.05 significance level, we find that the observed differences in F1 value are significant. Under a Wilcoxon signed rank test (Demšar, 2006) with 0.05 significance level, the difference between UMC and UDS is statistically significant.

Table 2: Clustering results (F1 score) on real data

Data ($n \times d$)	UMC	UDS	MAC	CMI	HiCS
WBC (198×33)	**0.89**	0.82	0.81	0.79	0.75
Shape (160×17)	**0.89**	**0.89**	0.85	0.82	0.77
Glass (214×9)	**0.63**	0.60	0.60	0.59	0.37
WBCD (569×30)	**0.82**	0.77	0.74	0.58	0.60
Diabetes (768×8)	**0.84**	0.79	0.52	0.71	0.53
Leaves (1600×64)	**0.81**	0.70	0.61	0.52	0.45
Pendigits (7494×16)	**0.87**	0.83	0.81	0.73	0.55
Waveform (5000×40)	**0.64**	0.58	0.46	0.32	0.21
Optical (5620×64)	**0.67**	0.61	0.48	0.40	0.36
Musk (6598×166)	**0.95**	0.92	0.88	0.61	0.58
Letter (20000×16)	**0.84**	0.82	0.82	0.64	0.49

[2]http://archive.ics.uci.edu/ml/index.html

Outlier Detection. UMC also outperforms other measures significantly at the 0.05 significance level (see *Supplementary Material D*).

Conclusions

In this paper, we proposed an unbiased multivariate correlation measure (UMC) fulfilling *comparability*, *interpretability*, *scalability* and *unbiasedness*. We argued that existing multivariate correlation measures only partially satisfy these criteria, and we theoretically and empirically identified dimensionality bias issues for them. In our proposed UMC measure, we employed a statistical model of independence, *i.e.*, the permutation model, under which an analytical formula of the expected value was derived, helping UMC improve interpretability and avoid dimensionality bias. Experimental evaluation convincingly demonstrated that UMC outperforms existing measures for a range of tasks.

Acknowledgments

This research is supported by the National Natural Science Foundation of China (No. 61371078).

References

Aggarwal, C., and Philip, S. Y. 2001. Outlier detection for high dimensional data. In *Proceedings of the 2001 ACM SIGMOD International Conference on Management of Data (SIGMOD-01)*, 37–46.

Bareinboim, E.; Tian, J.; and Pearl, J. 2014. Recovering from selection bias in causal and statistical inference. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence (AAAI-14)*, 2410–2416.

Breiman, L., and Friedman, J. H. 1985. Estimating optimal transformations for multiple regression and correlation. *Journal of the American statistical Association* 80(391):580–598.

Brown, G.; Pocock, A.; Zhao, M.-J.; and Luján, M. 2012. Conditional likelihood maximisation: a unifying framework for information theoretic feature selection. *Journal of Machine Learning Research* 13(Jan):27–66.

Demšar, J. 2006. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research* 7(Jan):1–30.

Di Crescenzo, A., and Longobardi, M. 2009a. On cumulative entropies. *Journal of Statistical Planning and Inference* 139(12):4072–4087.

Di Crescenzo, A., and Longobardi, M. 2009b. On cumulative entropies and lifetime estimations. In *International Work-Conference on the Interplay Between Natural and Artificial Computation*, 132–141.

Elizalde, E. 1986. An asymptotic expansion for the first derivative of the generalized riemann zeta function. *Mathematics of Computation* 47(175):347–350.

Ester, M.; Kriegel, H.-P.; Sander, J.; Xu, X.; et al. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the 2nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-96)*, 226–231.

Hubert, L., and Arabie, P. 1985. Comparing partitions. *Journal of Classification* 2(1):193–218.

Keller, F.; Muller, E.; and Bohm, K. 2012. Hics: high contrast subspaces for density-based outlier ranking. In *Proceedings of the 2012 IEEE 28th International Conference on Data Engineering (ICDE-12)*, 1037–1048.

Kriegel, H.-P.; Kröger, P.; and Zimek, A. 2009. Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 3(1):1.

Kullback, S., and Leibler, R. A. 1951. On information and sufficiency. *The Annals of Mathematical Statistics* 22(1):79–86.

Müller, E.; Günnemann, S.; Assent, I.; and Seidl, T. 2009. Evaluating clustering in subspace projections of high dimensional data. *Proceedings of the VLDB Endowment* 2(1):1270–1281.

Nguyen, H. V.; Müller, E.; Vreeken, J.; Keller, F.; and Böhm, K. 2013. Cmi: An information-theoretic contrast measure for enhancing subspace cluster and outlier detection. In *Proceedings of the 13th SIAM International Conference on Data Mining (SDM-13)*, 198–206.

Nguyen, H. V.; Müller, E.; Vreeken, J.; Efros, P.; and Böhm, K. 2014. Multivariate maximal correlation analysis. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, 775–783.

Nguyen, H.-V.; Mandros, P.; and Vreeken, J. 2016. Universal dependency analysis. In *Proceedings of the 16th SIAM International Conference on Data Mining (SDM-16)*, 792–800.

Rao, M.; Chen, Y.; Vemuri, B. C.; and Wang, F. 2004. Cumulative residual entropy: a new measure of information. *IEEE Transactions on Information Theory* 50(6):1220–1228.

Rao, M.; Seth, S.; Xu, J.; Chen, Y.; Tagare, H.; and Príncipe, J. C. 2011. A test of independence based on a generalized correlation function. *Signal Processing* 91(1):15–27.

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

Romano, S.; Vinh, N. X.; Bailey, J.; and Verspoor, K. 2016. A framework to adjust dependency measure estimates for chance. In *Proceedings of the 16th SIAM International Conference on Data Mining (SDM-16)*, 423–431.

Te Sun, H. 1978. Nonnegative entropy measures of multivariate symmetric correlations. *Information and Control* 36:133–156.

Te Sun, H. 1980. Multiple mutual informations and multiple interactions in frequency data. *Information and Control* 46:26–45.

Vinh, N. X.; Epps, J.; and Bailey, J. 2009. Information theoretic measures for clusterings comparison: is a correction for chance necessary? In *Proceedings of the 26th International Conference on Machine Learning (ICML-09)*, 1073–1080.

Wang, Y.; Tang, Q.; Xia, S.-T.; Wu, J.; and Zhu, X. 2016. Bernoulli random forests: Closing the gap between theoretical consistency and empirical soundness. In *Proceedings of the Twenty-fifth International Joint Conference on Artificial Intelligence (IJCAI-16)*, 2167–2173.

Zhang, X.; Pan, F.; Wang, W.; and Nobel, A. 2008. Mining non-redundant high order correlations in binary data. *Proceedings of the VLDB Endowment* 1(1):1178–1188.