Multivariate Maximal Correlation Analysis

Hoang Vu Nguyen[1] HOANG.NGUYEN@KIT.EDU
Emmanuel Müller[1,2] EMMANUEL.MUELLER@KIT.EDU
Jilles Vreeken[3] JILLES@MPI-INF.MPG.DE
Pavel Efros[1] PAVEL.EFROS@KIT.EDU
Klemens Böhm[1] KLEMENS.BOEHM@KIT.EDU

[1] Karlsruhe Institute of Technology, Germany
[2] University of Antwerp, Belgium
[3] Max-Planck Institute for Informatics & Saarland University, Germany

Abstract

Correlation analysis is one of the key elements of statistics, and has various applications in data analysis. Whereas most existing measures can only detect pairwise correlations between two dimensions, modern analysis aims at detecting correlations in multi-dimensional spaces.

We propose MAC, a novel *multivariate* correlation measure designed for discovering multi-dimensional patterns. It belongs to the powerful class of maximal correlation analysis, for which we propose a generalization to multivariate domains. We highlight the limitations of current methods in this class, and address these with MAC. Our experiments show that MAC outperforms existing solutions, is robust to noise, and discovers interesting and useful patterns.

1. Introduction

In data analysis we are concerned with analyzing large and complex data. One of the key aspects of this exercise is to be able to tell if a group of dimensions is mutually correlated. The ability to detect correlations is essential to very many tasks, e.g., feature selection (Brown et al., 2012), subspace search (Nguyen et al., 2013), multi-view acoustic feature learning for speech recognition (Arora & Livescu, 2013; Andrew et al., 2013), causal inference (Janzing et al., 2010), and subspace clustering (Müller et al., 2009).

In this paper, we specifically target at multivariate correlation analysis, i.e., the problem of *detecting correlations among two or more dimensions*. In particular, we want to detect complex interactions in high dimensional data. For example, genes may reveal only a weak correlation with a disease if each gene is considered individually, while when considered as a group of genes the correlation may be very strong (Zhang et al., 2008). In such applications pairwise correlation measures are not sufficient as they are unable to detect complex interactions of a group of genes.

Here we focus on maximal correlation analysis. It does not require assumptions on the data distribution, can detect non-linear correlations, is very efficient, and robust to noise. Maximal correlation analysis is our generalization of a number of powerful correlation measures that, in a nutshell, discover correlations hidden in data by (1) looking at various admissible transformations of the data (e.g., discretizations (Reshef et al., 2011), measurable mean-zero functions (Breiman & Friedman, 1985)), and (2) identifying the maximal correlation score (e.g., mutual information (Reshef et al., 2011), Pearson's correlation coefficient (Breiman & Friedman, 1985)) correspondingly.

The key reason these measures first transform the data is that otherwise only simple correlations can be detected: kernel transformations allow non-linear structures to be found that would go undetected in the original data space (Breiman & Friedman, 1985; Hardoon et al., 2004). In contrast, more complex measures such as mutual information can detect complex interactions without transformation, at the expense of having to assume and estimate the data distribution (Yin, 2004). Reshef et al. (2011) showed that instead of making assumptions, we should use the discretizations that yield the largest mutual information.

All these existing proposals, however, focus on pairwise correlations: Their solutions are specific for discovering correlations in two dimensions. While the search space of optimizing the transformations, e.g., discretizations (Reshef et al., 2011), is already large in this basic case, it grows exponentially with the number of dimen-

Proceedings of the 31^{st} International Conference on Machine Learning, Beijing, China, 2014. JMLR: W&CP volume 32.

sions. Hence, existing methods cannot straightforwardly be adapted to the multivariate setting, especially as their optimization heuristics are designed for pairwise analysis.

We address this problem by proposing **M**ultivariate M**a**ximal **C**orrelation Analysis (MAC), a novel approach to discovering correlations in multivariate data spaces. MAC employs a popular generalization of the mutual information called total correlation (Han, 1978), and discovers correlation patterns by identifying the transformations (e.g., discretizations in our case) of all dimensions that yield the maximal total correlation. While naive search for the optimal discretizations has issues regarding both efficiency and quality, we propose an efficient approximate algorithm that yields high quality. Our contributions are: (a) a generalization of Maximal Correlation Analysis to more than two dimensions, (b) a multivariate correlation measure for complex non-linear correlations without distribution assumptions, and (c) a simple and efficient method for estimating MAC that enables multivariate maximal correlation analysis. Note that MAC measures correlation for a given set of dimensions. To detect correlated sets one can use MAC in, e.g., subspace search frameworks (Nguyen et al., 2013) as used in our experiments.

2. Maximal Correlation Analysis

Definition 1 *The maximal correlation of real-valued random variables $\{X_i\}_{i=1}^d$ is defined as:*

$$\textsc{Corr}^*(X_1, \ldots, X_d) = \max_{f_1,\ldots,f_d} \textsc{Corr}(f_1(X_1), \ldots, f_d(X_d))$$

with Corr *being a correlation measure, $f_i : dom(X_i) \rightarrow \mathcal{A}_i$ being from a pre-specified class of functions, $\mathcal{A}_i \subseteq \mathbb{R}$.*

That is, maximal correlation analysis discovers correlations in the data by searching for the transformations of the X_i's that maximize their correlation (measured by Corr). Following Definition 1, to search for maximal correlation, we need to solve an optimization problem over a search space whose size is potentially exponential to the number of dimensions. The search space in general does not exhibit structure that we can exploit for an efficient search. Thus, it is infeasible to examine it exhaustively, which makes maximal correlation analysis on multivariate data very challenging. Avoiding this issue, existing work focuses on *pairwise* maximal correlations. More details are given below.

Instantiations of Def. 1. Breiman & Friedman (1985) defined the maximal correlation between two real-valued random variables X and Y as $\rho^*(X,Y) = \max_{f_1,f_2} \rho(f_1(X), f_2(Y))$ with $\textsc{Corr} = \rho$ being the Pearson's correlation coefficient, and $f_1 : \mathbb{R} \rightarrow \mathbb{R}$ and $f_2 : \mathbb{R} \rightarrow \mathbb{R}$ being two measurable mean-zero functions of X and Y, respectively. If f_1 and f_2 are non-linear functions, their method can find non-linear correlations.

Likewise, Rao et al. (2011) searched for $a, b \in \mathbb{R}$ that maximize $\textsc{Corr}(X,Y) = U(aX + b, Y)$, which equals to $\left|\int \kappa(ax + b - y)(p(x,y) - p(x)p(y))dxdy\right|$ where κ is a positive definite kernel function ($p(X)$, $p(Y)$, and $p(X,Y)$ are the pdfs of X, Y, and (X,Y), respectively). f_1 is a linear transformation function, and f_2 is the identity function. If κ is non-linear, they can find non-linear correlations.

Canonical correlation analysis (CCA) (Hotelling, 1936; Hardoon et al., 2004; Andrew et al., 2013; Chang et al., 2013), instead of analyzing two random variables, considers two data sets of the same size. That is, $X \in \mathbb{R}^A$ and $Y \in \mathbb{R}^B$ represent two groups of random variables. CCA looks for (non-)linear transformations of this data such that their correlation, measured by Corr, is maximized. In (Yin, 2004), Corr is the mutual information, $f_1 : \mathbb{R}^A \rightarrow \mathbb{R}$ and $f_2 : \mathbb{R}^B \rightarrow \mathbb{R}$ are linear transformations. $\textsc{Corr}(f_1(X), f_2(Y))$ is computed by density estimation. Along this line, Generalized CCA (Carroll, 1968; Kettenring, 1971) is an extension of CCA to multiple data sets. Its focus so far, however, is on linear correlations (van de Velden & Takane, 2012).

Maximal Information Coefficient (MIC) (Reshef et al., 2011) analyzes the correlation of X and Y by identifying the discretizations of X and Y that maximize their mutual information, normalized according to their numbers of bins. Here, Corr is the normalized mutual information. f_1 and f_2 are functions mapping values of $dom(X)$ and $dom(Y)$ to $\mathcal{A}_1 = \mathbb{N}$ and $\mathcal{A}_2 = \mathbb{N}$ (with counting measures), respectively. Note that, MIC is applicable to CCA computation where mutual information is used (Yin, 2004).

Limitations of existing techniques. All of the above techniques are limited to either two dimensions or linear correlations. Regarding the first issue, we use MIC to illustrate our point. Consider a toy data set with three dimensions $\{A, B, C\}$. MIC can find two *separate* ways to discretize B to maximize its correlation with A and C, but it cannot find a discretization of B such that the correlation with regard to *both* A and C is maximized. Thus, MIC is not suited for calculating correlations over more than two dimensions. Further, adapting existing solutions to the multivariate setting is nontrivial due to the huge search space.

As an attempt towards enabling maximal correlation analysis for multivariate data without being constrained to specific types of correlations, we propose MAC, a generalization of MIC to more than two dimensions. We pick MIC since it explicitly handles different types of correlations. However, we show that a naive heuristic computation like MIC on multivariate data poses issues to both efficiency and quality. In contrast, MAC aims to address both aspects. We defer extending other types of correlation measures to the multivariate non-linear settings to future work.

3. Theory of MAC

In this section, we discuss the theoretical model of MAC. For brevity, we put the proofs of all theorems in the supplementary.[1] Consider a d-dimensional data set $\mathbf{D}$ with real-valued dimensions $\{X_i\}_{i=1}^d$ and N data points. We regard each dimension X_i as a random variable, distributed according to pdf $p(X_i)$. Mapping MAC to Def. 1, we have that CORR is the normalized total correlation (see below), and $f_i : dom(X_i) \rightarrow \mathbb{N}$ (with counting measure) is a discretization of X_i. The total correlation, also known as *multi-information*, is a popular measure of multivariate correlation and widely used in data analysis (Sridhar et al., 2010; Schietgat et al., 2011). The total correlation of $\{X_i\}_{i=1}^d$, i.e., of data set $\mathbf{D}$, written as $I(\mathbf{D})$, is $I(\mathbf{D}) = \sum_{i=1}^{d} H(X_i) - H(X_1, \ldots, X_d)$ where $H(.)$ is the Shannon entropy. We have:

Theorem 1 *$I(\mathbf{D}) \geq 0$ with equality iff $\{X_i\}_{i=1}^d$ are statistically independent.*

Thus, $I(\mathbf{D}) > 0$ when the dimensions of $\mathbf{D}$ exhibit any mutual correlation, regardless of the particular correlation types. However, in general the pdfs required for computing the entropies are unknown in practice. Estimating these is nontrivial, especially when the data available is finite and the dimensionality is high. One common practice is to discretize the data to obtain the probability mass functions. Yet, such existing methods are not designed towards optimizing correlation (Reshef et al., 2011). MAC in turn aims at addressing this problem.

Let g_i be a discretization of X_i into $n_i = |g_i|$ bins. We will refer to n_i as the *grid size* of X_i. We write $X_i^{g_i}$ as X_i discretized by g_i. We call $G = \{g_1, \ldots, g_d\}$ a d-dimensional grid of $\mathbf{D}$. For mathematical convenience, we focus only on grids G with $n_i \geq 2$. This has been shown to be effective in capturing complex patterns in the data, as well as detecting independence (Reshef et al., 2011). The product of grid sizes of G is $|G| = n_1 \times \ldots \times n_d$. We write $\mathbf{D}^G$ as $\mathbf{D}$ discretized by G. The grid G induces a probability mass function on $\mathbf{D}$, i.e., for each cell of G, its mass is the fraction of $\mathbf{D}$ falling into it. The total correlation of $\mathbf{D}$ given G becomes: $I(\mathbf{D}^G) = \sum_{i=1}^{d} H(X_i^{g_i}) - H(X_1^{g_1}, \ldots, X_d^{g_d})$.

For maximal correlation analysis, one could find an optimal grid G for $\mathbf{D}$ such that $I(\mathbf{D}^G)$ is maximized. However, the value of $I(\mathbf{D}^G)$ is dependent on $\{n_i\}_{i=1}^d$:

Theorem 2 $I(\mathbf{D}^G) \leq \sum_{i=1}^{d} \log n_i - \max(\{\log n_i\}_{i=1}^d)$.

[1] *http://www.ipd.kit.edu/~nguyenh/mac*

Thus, for unbiased optimization, we normalize $I(\mathbf{D}^G)$ according to the grid sizes. Hence, we maximize

$$I_n(\mathbf{D}^G) = \frac{I(\mathbf{D}^G)}{\sum_{i=1}^{d} \log n_i - \max(\{\log n_i\}_{i=1}^d)} \qquad (1)$$

which we name *the normalized total correlation*. From Theorems 1 and 2, we arrive at $I_n(\mathbf{D}^G) \in [0, 1]$. However, maximizing $I_n(\mathbf{D}^G)$ is not enough. Consider the case where each dimension has N distinct values. If we discretize each dimension into N bins, then $I_n(\mathbf{D}^G)$ becomes 1, i.e., maximal. To avoid this trivial binning, we need to impose the maximum product of grid sizes B of the grids G considered. For pairwise correlation ($d = 2$), Reshef et al. prove that $B = N^{1-\epsilon}$ with $\epsilon \in (0, 1)$. As generalizing this result to the multivariate case is beyond the scope of this paper, we adopt it and hence, restrict $n_i \times n_j < N^{1-\epsilon}$ for $i \neq j$. We define $\text{MAC}(\mathbf{D})$ as follows

$$\text{MAC}(\mathbf{D}) = \max_{\substack{G=\{g_1,\ldots,g_d\} \\ \forall i \neq j : n_i \times n_j < N^{1-\epsilon}}} I_n(\mathbf{D}^G) \quad . \qquad (2)$$

We will write $\text{MAC}(\mathbf{D})$ and $\text{MAC}(X_1, \ldots, X_d)$ interchangeably. We have $\text{MAC}(\mathbf{D}) \in [0, 1]$. When $\text{MAC}(\mathbf{D}) = 0$, we consider $\{X_i\}_{i=1}^d$ to be statistically independent. Due to insufficient sample sizes, the theoretical zero score might not happen in practice. Nevertheless, a low score always indicates a low mutual correlation of $\{X_i\}_{i=1}^d$, and vice versa. We will show that MAC performs very well in analyzing multivariate data (cf., Section 6).

4. Calculating MAC: Naive Approach

To use MAC in practice, we need to compute it efficiently. Let us consider naively extending the strategy that MIC uses. To approximate the optimal discretization of two dimensions, MIC employs a heuristic: for every equal-frequency discretization of a dimension, it searches for the discretization over the other dimension that maximizes the normalized mutual information.

Naively extending this to the multivariate case, for every set of grid sizes $\{n_i\}_{i=1}^d$ we would partition each set of $(d-1)$ dimensions into equal-frequency bins. We would then try to find the optimal discretization of the remaining dimension. For every set $\{n_i\}_{i=1}^d$, we repeat this per dimension, and report the maximum over these d values.

However, by using $n_i \times n_j < N^{(1-\epsilon)}$ for any $i \neq j$, one can prove that $n_1 \times \ldots \times n_d < N^{(1-\epsilon)d/2}$. Hence, we know the size of the search space is $O(N^d)$—which implies this scheme is infeasible for high dimensional data. In fact, even for two dimensions MIC already faces efficiency issues (Reshef et al., 2011).

5. Calculating MAC: Our Approach

We propose a simple and efficient greedy method for estimating MAC. To compute $\text{MAC}(\mathbf{D})$, one typically has to find concurrently the discretizations of all dimensions that maximize their normalized total correlation $I_n(\mathbf{D}^G)$ (see Eq. (1) and (2)), which is the source of the computational intractability. Our intuition is to *serialize* this search. That is, step-by-step we find the dimension and its discretization that maximizes its normalized total correlation with all the dimensions already selected and discretized. In particular, we first identify two dimensions X'_1 and X'_2 such that $\text{MAC}(X'_1, X'_2)$ is maximal among all pairs of dimensions. Then, at each subsequent step $k \in [2, d-1]$, let $C_k = \{X'_1, \ldots, X'_k\}$ be the set of dimensions already picked and discretized. We aim to (a) identify the dimension X'_{k+1} that is most likely correlated with C_k *without* having to pre-discretize X'_{k+1}, and (b) find the discretization of X'_{k+1} yielding the MAC score of X'_{k+1} and C_k. Finally, we approximate $\text{MAC}(\mathbf{D})$ using the grid G obtained. From now on, when using Shannon entropy, we imply the involved dimensions have been discretized, e.g., we leave the superscript and write X_i for $X_i^{g_i}$. The details of our method are as follows.

5.1. Identifying and discretizing X'_1 and X'_2

We compute $\text{MAC}(X, Y)$ for every pair of dimensions (X, Y) and pick (X'_1, X'_2) with the largest MAC score.

To compute $\text{MAC}(X, Y)$, for each pair of grid sizes (n_X, n_Y) with $n_X n_Y < N^{1-\epsilon}$, we maximize $H(X) - H(X|Y) = H(Y) - H(Y|X)$. Note that, one could solve this through MIC. However, since MIC fixes one dimension to equal-frequency bins before discretizing the other dimension, we conjecture that the solution of MIC is suboptimal. Instead, we compute $\text{MAC}(X, Y)$ by cumulative entropy (Nguyen et al., 2013).

The cumulative entropy of X, denoted $h(X)$, is defined as

$$h(X) = -\int_{dom(X)} P(X \leq x) \log P(X \leq x) dx \quad (3)$$

where $P(X \leq x)$ is the probability that $X \leq x$.

The conditional cumulative entropy is given as

$$h(X|Y) = \int h(X|y) p(y) dy \quad (4)$$

where $p(Y)$ is the pdf of Y.

It holds $h(X|Y) \geq 0$ with equality iff X is a function of Y. Also, $h(X) \geq h(X|Y)$ with equality iff X is independent of Y. Thus, the larger $h(X) - h(X|Y)$, the more correlated X and Y are. Therefore, by maximizing $h(X) - h(X|Y)$, we maximize the correlation between X and Y, and hence, intuitively maximizes $H(X) - H(X|Y)$.

W.l.o.g., let $X(1) \leq \ldots \leq X(N)$ be realizations of X. We have (Nguyen et al., 2013):

$$h(X) = -\sum_{j=1}^{N-1} (X(j+1) - X(j)) \frac{j}{N} \log \frac{j}{N} \quad . \quad (5)$$

Computing $h(X|Y)$ is more problematic since the pdf of Y is unknown in practice. We solve the issue as follows.

Since we want to maximize $h(X) - h(X|Y)$, or, as $h(X)$ is constant, minimize $h(X|Y)$, we formulate a novel problem: *Given a grid size of Y, find the respective discretization g of Y that minimizes $h(X|Y^g)$*. Solving this problem, we essentially find the optimal discretization of Y at the given grid size that intuitively maximizes $H(X) - H(X|Y)$ without having to discretize X at the same time. In a nutshell, this is our key improvement over MIC.

We prove that our new optimization problem can be solved at multiple grid sizes simultaneously by dynamic programming. In particular, w.l.o.g., let $Y(1) \leq \ldots \leq Y(N)$ be realizations of Y. Further, let

$$Y(j, m) = \{Y(j), Y(j+1), \ldots, Y(m)\}$$

where $j \leq m$. Slightly abusing notation, we write $Y(1, N)$ as Y. We use $h(X|\langle Y(j, m)\rangle)$ to denote $h(X)$ computed using the $(m - j + 1)$ points of $\mathbf{D}$ corresponding to $Y(j)$ to $Y(m)$, projected onto X. For $1 \leq l \leq m \leq N$, we write

$$f(m, l) = \min_{g:|g|=l} h(X|Y^g(1, m))$$

where g is a discretization of $Y(1, m)$ in to l bins, and $Y^g(1, m)$ is the discretized version of $Y(1, m)$ by g. For $1 < l \leq m \leq N$, we have

Theorem 3 *We have:* $f(m, l) = \min_{j \in [l-1, m)} \frac{j}{m} f(j, l-1) + \frac{m-j}{m} h(X|\langle Y(j+1, m)\rangle)$.

Theorem 3 shows that the optimal discretization of $Y(1, m)$ can be derived from that of $Y(1, j)$ with $j < m$. This allows us to design a dynamic programming algorithm to find optimal discretizations of Y at different grid sizes n_Y. As $n_X \geq 2$, we only consider $n_Y < N^{1-\epsilon}/2$. Note that, we search for multiple grid sizes since we need them for normalization.

We apply the same optimization process for $h(Y) - h(Y|X)$. Then, we combine the optimal discretizations of both X and Y where $n_X n_Y < N^{1-\epsilon}$. We compute $\text{MAC}(X, Y)$ accordingly. We identify (X'_1, X'_2) as the pair of dimensions with the largest MAC score.

5.2. Efficient heuristic to identify X'_{k+1}

In practice, one could identify X'_{k+1} and its optimal discretization concurrently by computing $\text{MAC}(X, C_k)$ for

every dimension X left, and select the dimension with the best MAC score as X'_{k+1}. Note that since all dimensions in C_k have already been discretized, we do not discretize them again. We prove in Section 5.3 that $\text{MAC}(X, C_k)$ can be solved by dynamic programming. Yet, doing this for each and every dimension X left may become inefficient for high dimensional data. Thus, our intuition here is to first heuristically identify X'_{k+1} and then find its optimal discretization.

In particular, let n'_i be the grid size of $X'_i \in C_k$. From Eq. (2), it follows that to compute $\text{MAC}(X, C_k)$, we need to maximize

$$\frac{\left(\sum_{i=1}^{k} H(X'_i)\right) + H(X) - H(X|C_k) - H(C_k)}{\log n + \sum_{i=1}^{k} \log n'_i - \max(\{\log n\} \cup \{\log n'_i\}_{i=1}^{k})} \quad (6)$$

where $n < \frac{N^{(1-\epsilon)}}{\max(\{n'_i\}_{i=1}^{k})}$ and X is discretized into n bins. Let us consider the following term

$$\frac{\left(\sum_{i=1}^{k} H(X'_i)\right) + h(X) - h(X|C_k) - H(C_k)}{h(X) + \sum_{i=1}^{k} \log n'_i - \max(\{\log n'_i\}_{i=1}^{k})} \quad . \quad (7)$$

Informally speaking, we can regard both Eq. (6) and (7) to represent the normalized mutual correlation of X and all dimensions in C_k. They have very similar properties. First, their values are in $[0, 1]$. Second, they are both equal to 0 iff (discretized) X and all dimensions in C_k are statistically independent. Third, they are both maximal when there exists $X'_i \in C_k$ such that (discretized) X and all dimensions in $C_k \setminus \{X'_i\}$, each is a function of X'_i. The detailed explanation of all three properties is skipped for brevity.

Therefore, instead of solving Eq. (6) for every n and every X to obtain X'_{k+1}, we propose to use Eq. (7) as a surrogate indicator of how likely a dimension X is indeed X'_{k+1} (the larger the indicator, the better). This indicator has three advantages: (a) it does not require us to discretize X, (b) it is independent of grid size n, and (c) it can be computed much more efficiently (see Eq. (5)). Note that we are not restricted to this heuristic: If there are enough computational resources, one can just skip this step and run the solution in Section 5.3 for every dimension X not yet processed.

5.3. Discretizing X'_{k+1}

For readability, we use X to denote X'_{k+1} in this section. To find the optimal discretization of X, for each grid size n, we find the respective discretization of X that maximizes $H(X) - H(X, C_k)$; we ignore $\left(\sum_{i=1}^{k} H(X'_i)\right)$ as it has a fixed value. We prove that this can be solved at multiple grid sizes simultaneously by dynamic programming.

In particular, w.l.o.g., let $X(1) \leq \ldots \leq X(N)$ be realizations of X. Further, let

$$X(j, m) = \{X(j), X(j+1), \ldots, X(m)\}$$

where $j \leq m$. As before, we write $X(1, N)$ as X. We use $H(C_k|\langle X(j, m)\rangle)$ to denote $H(C_k)$ computed using the $(m - j + 1)$ points of $\mathbf{D}$ corresponding to $X(j)$ to $X(m)$, projected onto the dimensions of C_k. Note that the bins of each dimension in C_k are intact. For $1 \leq l \leq m \leq N$, we write

$$F(m, l) = \max_{g:|g|=l} H(X^g(1, m)) - H(X^g(1, m), C_k)$$

where g is a discretization of $X(1, m)$ in to l bins, and $X^g(1, m)$ is the discretized version of $X(1, m)$ by g. For $1 < l \leq m \leq N$, we have

Theorem 4 *We have:* $F(m, l) = \max_{j \in [l-1, m)} \frac{j}{m} F(j, l-1) - \frac{m-j}{m} H(C_k|\langle X(j+1, m)\rangle)$.

We design a dynamic programming search following Theorem 4, and identify the best discretizations of X at different grid sizes $n < \frac{N^{(1-\epsilon)}}{\max(\{n'_i\}_{i=1}^{k})}$. Then, we use Eq. (6) to identify the optimal discretization of X.

5.4. Complexity analysis

Using the original set of cut points per dimension, the time complexity of using dynamic programming for each dimension is $O(N^3)$, which would be restrictive for large data. To address this, when discretizing a dimension with maximum grid size max_grid, we limit its number of cut points to $c \times max\_grid$ with $c > 1$. Similar to MIC, we do this using equal-frequency binning on the dimension with the number of bins equal to $(c \times max\_grid + 1)$. More elaborate pre-processing, such as (Mehta et al., 2005), can be considered, yet is beyond the scope of this work. Regarding c, the larger it is, the more candidate discretizations we consider, and hence, the better the result. However, setting c too high causes computational issues. Our preliminary empirical analysis shows that $c = 2$ offers a good balance between quality and efficiency, and we will use this as the default value in the experiments.

The time complexity of MAC includes (a) the cost of presorting the values of all dimensions $O(dN \log N)$, (b) the cost of finding and discretizing X'_1 and X'_2 $O(d^2 N^{3(1-\epsilon)})$, and (c) the cost of finding and discretizing subsequent dimensions $O(d^2 N + dN^{3(1-\epsilon)})$. The overall complexity is $O(d^2 N^{3(1-\epsilon)})$. As we fix ϵ to 0.5 in our implementation, the complexity of MAC is $O(d^2 N^{1.5})$.

6. Experiments

For assessing the performance of MAC in detecting pairwise correlations, we compare against MIC (Reshef et al., 2011) and DCOR (Székely & Rizzo, 2009), two state-of-the-art correlation measures. However, neither MIC nor DCOR are directly applicable in the multivariate setting. In order to make a meaningful comparison, we consider two approaches for extending these methods: (a) taking the sum of pairwise correlation scores and normalizing it by the total number of pairs, and (b) taking the maximum of these scores. Empirically, we found the second option to yield best performance, and hence, we use this as the multivariate extension for both MIC and DCOR.

For comparability and repeatability of our experiments we provide data, code, and parameter settings on our project website.[2]

6.1. Synthetic data

To evaluate how MAC performs in different settings, we first use synthetic data. We aim to show MAC can successfully detect both pairwise and multivariate correlations.

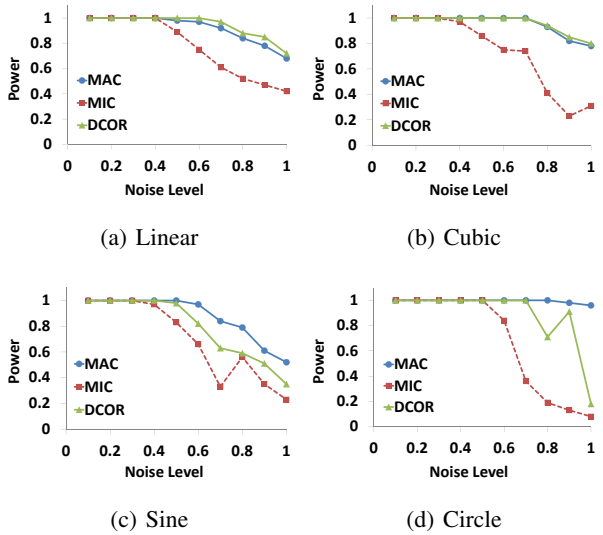
Assessing functional correlations. As a first experiment, we investigate whether MAC can detect linear and non-linear functional correlations. To this end, we create data sets simulating four different functions.

As performance metric we use the power of the measures, as in (Reshef et al., 2011): For each function, the null hypothesis is that the data dimensions are statistically independent. For each correlation measure, we determine the cutoff for testing the independence hypothesis by (a) generating 100 data sets of a fixed size, (b) computing the correlation score of each data set, and (c) setting the cutoff according to the significance level $\alpha = 0.05$. We then generate 100 data sets *with* correlations, adding Gaussian noise. The power of the measure is the proportion of the new 100 data sets whose correlation scores exceed the cutoff.

Results on pairwise functional correlations. We create data sets of 1000 data points, using respectively a linear, cubic, sine, and circle as generating functions. Recall that for pairwise cases, we search for the optimal discretization of one dimension at a time (Section 5.1). We claim this leads to better quality than MIC, which heuristically fixes a discretization on the remaining dimension.

We report the results in Fig. 1. Overall, we find that MAC outperforms MIC on all four functions. Further, we see that MAC and DCOR have about the same power in detecting linear and cubic correlations. For the more complex correlations, the performance of DCOR starts to drop. This suggests MAC is better suited than DCOR for measuring and detecting strongly non-linear correlations.

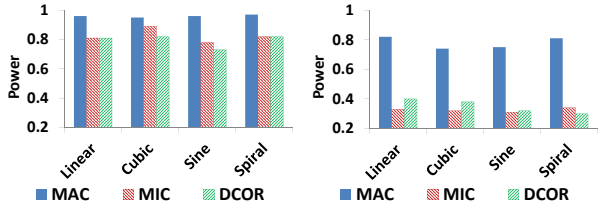
[2] *http://www.ipd.kit.edu/~nguyenh/mac*



(a) Linear (b) Cubic

(c) Sine (d) Circle

Figure 1. [Higher is better] Baseline results for 2-dimensional functions, statistical power vs. noise.

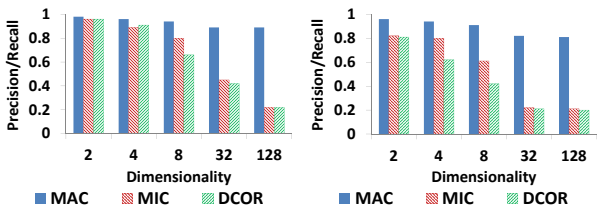


(a) Noise = 20% (b) Noise = 80%

Figure 2. [Higher is better] Statistical power vs. noise for 32-dimensional functions.

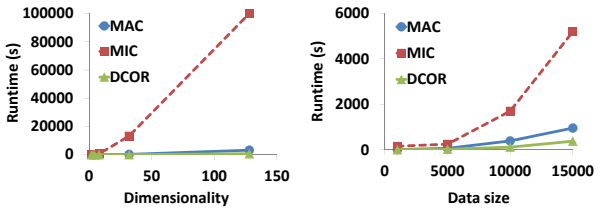
Results on multivariate functional correlations. Next we consider multivariate correlations. To this end we again create data sets with 1000 data points, but of differing dimensionality. Among the functions is a multi-dimensional spiral. As a representative, we show the results for 32-variate functions in Fig. 2. We see that MAC outperforms both MIC and DCOR in all cases. We also see that MAC is well suited for detecting multivariate correlations.

Assessing non-functional correlations. Finally, we consider multivariate non-functional correlations. To this end we generate data with density-based subspace clusters as in (Müller et al., 2009). For each dimensionality $r < d$, we select w subspaces $\mathcal{S}_c$ having r dimensions (called correlated subspaces), and embed two density-based clusters representing correlation patterns. Since density-based clusters can have arbitrarily complex shapes and forms, we can simulate non-functional correlations of arbitrary com-



(a) Noise = 20% (b) Noise = 80%

Figure 3. [Higher is better] Precision/Recall vs. noise for non-functional correlations (i.e., clusters).



(a) Runtime vs. dimensionality (b) Runtime vs. data size

Figure 4. [Lower is better] Scalability of correlation measures with regard to dimensionality and data size.

plexity. For each correlated subspace, we create another subspace by substituting one of its dimensions by a randomly sampled noisy dimension. Thus, in total, we have $2w$ subspaces. We compute the correlation score for each of these subspaces, and pick the top-w subspaces $\mathcal{S}_t$ with highest scores. The power of the correlation measure is identified as its precision and recall, i.e., $\frac{|\mathcal{S}_c \cap \mathcal{S}_t|}{w}$ since $|\mathcal{S}_c| = |\mathcal{S}_t| = w$. We add noise as above.

Results on non-functional correlations. We create data sets with 1000 data points, of varying dimensionality. For each value of r and w, we repeat the above process 10 times and consider the average results, noting that the standard deviations are very small. As a representative, we present the results with $w = 10$ in Fig. 3. We see that compared to both MIC and DCOR, MAC identifies these correlations best. Notably, its performance is consistent across different dimensionalities. In addition, MAC is robust against noise.

Scalability. Finally, we examine scalability of measures with regard to dimensionality and data size. For the former, we generate data sets with 1000 data points and dimensionality varied. For the latter, we generate data sets with dimensionality 4 and data size varied. We show the results in Fig. 4. Each result is the average of 10 runs. Overall, in both dimensionality and data size, we find that MAC scales much better than MIC and is close to DCOR.

The experiments so far show that MAC is a very efficient and highly accurate multivariate correlation measure.

6.2. Real-world data

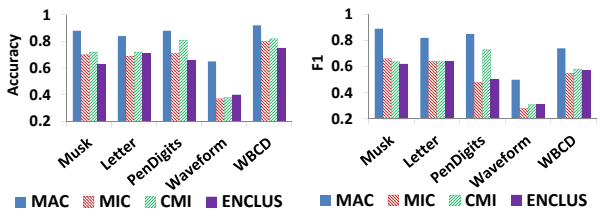
Next, we consider real-world data. We apply MAC in two typical applications of correlations measures in data analysis: cluster analysis and data exploration.

Cluster analysis. For cluster analysis, it has been shown that mining subspace clusters is particularly useful when the subspaces show high correlation, i.e., include few or no irrelevant dimensions (Müller et al., 2009). Thus, in this experiment, we plug MAC and MIC into the Apriori subspace search framework to assess their performance. Here, we omit DCOR as we saw above that MIC and DCOR perform similarly on multivariate data. Instead, we consider ENCLUS (Cheng et al., 1999) and CMI (Nguyen et al., 2013) as baselines. Both are subspace search methods using respectively total correlation and cumulative mutual information as selection criterion. They internally use these basic correlation measures, e.g., ENCLUS computes total correlation using a density estimation rather than maximal correlation analysis. We show an enhanced performance by using MAC instead.

Our setup follows existing literature (Müller et al., 2009; Nguyen et al., 2013): We use each measure for subspace search, and apply DBSCAN (Ester et al., 1996) to the top 100 subspaces with highest correlation scores. Using these we calculate Accuracy and F1 scores. To do so, we use 7 labeled data sets from different domains ($N \times d$): Musk (6598×166), Letter Recognition (20000×16), PenDigits (7494×16), Waveform (5000×40), WBCD (569×30), Diabetes (768×8), and Glass (214×9), taken from the UCI ML repository. For each data set, we regard the class labels as the ground truth clusters.

Fig. 5 shows the results for the 5 most high dimensional datasets; the remainder is reported in the supplementary material. Overall, MAC achieves the highest clustering quality. It consistently outperforms MIC, CMI, and ENCLUS. Notably, it discovers higher dimensional subspaces. Recall that Apriori imposes the requirement that each subspace is only considered if all of its child subspaces show high correlation. Whereas MAC correctly identifies correlations in these lower-order projections, the other methods assign inaccurate correlation scores more often, which prevents them from finding the larger correlated subspaces. As a result, MAC detects correlations in multivariate real-world data sets better than its competitors.

By applying a Friedman test (Demsar, 2006) at significance level $\alpha = 0.05$, we find that the observed differences in Accuracy and F1 are significant. By performing a post-hoc Nemenyi test we learn that MAC significantly outperforms MIC and ENCLUS. We also perform a Wilcoxon signed rank test with $\alpha = 0.05$ to compare MAC and CMI. The result shows MAC to significantly outperform CMI.



(a) Accuracy (b) F1

Figure 5. [Higher is better] Clustering results on real-world data sets taken from the UCI ML repository.

Discovering novel correlations. To evaluate the efficacy of MAC in data exploration, we apply MAC on a real-world data set containing climate and energy consumption measures of an office building in Frankfurt, Germany (Wagner et al., 2014). After data pre-processing to handle missing values, our final data set contains 35601 records and 251 dimensions. Some example dimensions are room CO_2 concentration, indoor temperature, temperature produced by heating systems, drinking water consumption, and electricity consumption. Since this data set is unlabeled, we cannot calculate clustering quality as above. Instead, we perform subspace mining to detect correlated subspaces, and investigate the discovered correlations. In particular, our objective is to study how climate and energy consumption indicators interact with each other.

Below we present interesting correlations we discovered using MAC, that were *not* discovered using the other measures. All reported correlations are significant at $\alpha = 0.05$. We verified all findings with a domain expert, resulting in some already known correlations, and others that are novel.

An example of a known multivariate correlation discovered using MAC is between the temperatures inside different office rooms located in the same section of the building. Another example is the correlation between the air temperature supplied to the heating system, the temperature of the heating boiler, and the amount of heating it produces. This relation is rather intuitive and expected. The most interesting point is the interaction between the temperature of the heating boiler and the amount of heating produced. Intuitively, the higher the former, the larger the latter. However, the correlation is not linear. Instead, it seems to be a combination of two quadratic functions (Fig. 6).

MAC also finds an interesting correlation between drinking water consumption, the outgoing temperature of the air conditioning (cooling) system, and the room CO_2 concentration. There is a clear tendency: the more water consumed, the higher the CO_2 concentration (Fig. 7(a)). Besides, there is a sinusoidal-like correlation between the drinking water consumption and the outgoing temperature

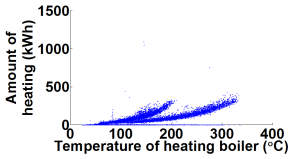
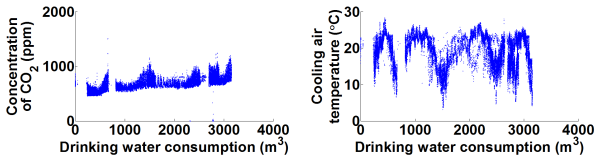


Figure 6. Temperature of boiler and amount of heating.



(a) Water vs. CO_2 (b) Water vs. cooling air

Figure 7. Correlation of indoor climate and energy consumption.

of the cooling system (Fig. 7(b)). These correlations, novel to our domain expert, offer a view on how human behavior interacts with indoor climate and energy consumption.

7. Conclusions

We introduced MAC, a maximal correlation measure for multivariate data. It discovers correlation patterns by identifying the discretizations of all dimensions that maximize their normalized total correlation. We proposed an efficient estimation of MAC that also ensures high quality. Experiments showed that MAC successfully discovered interesting complex correlations in real-world data sets.

The research proposed here gives way to computing the total correlation on empirical data, which has wide applications in various fields. In addition, it demonstrates the potential of multivariate maximal correlation analysis to data analytics. Through MAC, we have shown that searching for the optimal transformations of all dimensions concurrently is impractical. Instead, we conjecture that: *To efficiently solve the optimization problem in Definition 1, one needs to find an order of* $\{X_i\}_{i=1}^d$ *to process as well*. Solving this conjecture for other general cases is part of our future work on maximal correlation analysis.

Acknowledgments

We thank the anonymous reviewers for helpful comments. HVN is supported by the German Research Foundation (DFG) within GRK 1194. EM is supported by the YIG program of KIT as part of the German Excellence Initiative. JV is supported by the Cluster of Excellence "Multimodal Computing and Interaction" within the Excellence Initiative of the German Federal Government. EM and JV are supported by Post-Doctoral Fellowships of the Research Foundation – Flanders (FWO).

References

Andrew, Galen, Arora, Raman, Bilmes, Jeff, and Livescu, Karen. Deep canonical correlation analysis. In *ICML (3)*, pp. 1247–1255, 2013.

Arora, Raman and Livescu, Karen. Multi-view cca-based acoustic features for phonetic recognition across speakers and domains. In *ICASSP*, pp. 7135–7139, 2013.

Breiman, Leo and Friedman, Jerome H. Estimating optimal transformations for multiple regression and correlation. *J. Am. Stat. Assoc.*, 80(391):580–598, 1985.

Brown, Gavin, Pocock, Adam, Zhao, Ming-Jie, and Luján, Mikel. Conditional likelihood maximisation: A unifying framework for information theoretic feature selection. *JMLR*, 13:27–66, 2012.

Carroll, John D. Generalization of canonical correlation analysis to three or more sets of variables. In *Proceedings of the American Psychological Association*, pp. 227–228, 1968.

Chang, Billy, Krüger, Uwe, Kustra, Rafal, and Zhang, Junping. Canonical correlation analysis based on hilbert-schmidt independence criterion and centered kernel target alignment. In *ICML (2)*, pp. 316–324, 2013.

Cheng, Chun Hung, Fu, Ada Wai-Chee, and Zhang, Yi. Entropy-based subspace clustering for mining numerical data. In *KDD*, pp. 84–93, 1999.

Demsar, Janez. Statistical comparisons of classifiers over multiple data sets. *JMLR*, 7:1–30, 2006.

Ester, Martin, Kriegel, Hans-Peter, Sander, Jörg, and Xu, Xiaowei. A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD*, pp. 226–231, 1996.

Han, Te Sun. Nonnegative entropy measures of multivariate symmetric correlations. *Information and Control*, 36 (2):133–156, 1978.

Hardoon, David R., Szedmak, Sandor, and Shawe-Taylor, John. Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*, 16(12):2639–2664, 2004.

Hotelling, Harold. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377, 1936.

Janzing, Dominik, Hoyer, Patrik O., and Schölkopf, Bernhard. Telling cause from effect based on high-dimensional observations. In *ICML*, pp. 479–486, 2010.

Kettenring, Jon R. Canonical analysis of several sets of variables. *Biometrika*, 58(3):433–451, 1971.

Mehta, Sameep, Parthasarathy, Srinivasan, and Yang, Hui. Toward unsupervised correlation preserving discretization. *IEEE Transactions on Knowledge and Data Engineering*, 17(9):1174–1185, 2005.

Müller, Emmanuel, Günnemann, Stephan, Assent, Ira, and Seidl, Thomas. Evaluating clustering in subspace projections of high dimensional data. *PVLDB*, 2(1):1270–1281, 2009.

Nguyen, Hoang Vu, Müller, Emmanuel, Vreeken, Jilles, Keller, Fabian, and Böhm, Klemens. CMI: An information-theoretic contrast measure for enhancing subspace cluster and outlier detection. In *SDM*, pp. 198–206, 2013.

Rao, Murali, Seth, Sohan, Xu, Jian-Wu, Chen, Yunmei, Tagare, Hemant, and Príncipe, José C. A test of independence based on a generalized correlation function. *Signal Processing*, 91(1):15–27, 2011.

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

Schietgat, Leander, Costa, Fabrizio, Ramon, Jan, and De Raedt, Luc. Effective feature construction by maximum common subgraph sampling. *Machine Learning*, 83(2): 137–161, 2011.

Sridhar, Muralikrishna, Cohn, Anthony G., and Hogg, David C. Unsupervised learning of event classes from video. In *AAAI*, 2010.

Székely, Gábor J. and Rizzo, Maria L. Brownian distance covariance. *Annals of Applied Statistics*, 3(4):1236–1265, 2009.

van de Velden, Michel and Takane, Yoshio. Generalized canonical correlation analysis with missing values. *Computational Statistics*, 27(3):551–571, 2012.

Wagner, Andreas, Lützkendorf, Thomas, Voss, Karsten, Spars, Guido, Maas, Anton, and Herkel, Sebastian. Performance analysis of commercial buildings—results and experiences from the german demonstration program 'energy optimized building (EnOB)'. *Energy and Buildings*, 68:634–638, 2014.

Yin, Xiangrong. Canonical correlation analysis based on information theory. *Journal of Multivariate Analysis*, 91 (2):161–176, 2004.

Zhang, Xiang, Pan, Feng, Wang, Wei, and Nobel, Andrew B. Mining non-redundant high order correlations in binary data. *PVLDB*, 1(1):1178–1188, 2008.

Multivariate Maximal Correlation Analysis
Supplementary Material

Hoang Vu Nguyen[1] HOANG.NGUYEN@KIT.EDU
Emmanuel Müller[1,2] EMMANUEL.MUELLER@KIT.EDU
Jilles Vreeken[3] JILLES@MPI-INF.MPG.DE
Klemens Böhm[1] KLEMENS.BOEHM@KIT.EDU

[1] Karlsruhe Institute of Technology, Germany
[2] University of Antwerp, Belgium
[3] Max-Planck Institute for Informatics & Saarland University, Germany

1. Proofs

Proof of Theorem 1. We have

$$I(\mathbf{D}) = \sum_{i=2}^{d} H(X_i) - H(X_i|X_1,\ldots,X_{i-1}) \quad .$$

Since conditioning reduces entropy, it holds that

$$H(X_i) \geq H(X_i|X_1,\ldots,X_{i-1}) \quad .$$

Equality holds iff X_i is independent of $\{X_1,\ldots,X_{i-1}\}$ with $2 \leq i \leq d$. This in turn is equivalent to the fact that $\{X_i\}_{i=1}^{d}$ are statistically independent. □

Proof of Theorem 2. We have:

$$I(\mathbf{D}^G) = \sum_{i=2}^{d} H(X_i^{g_i}) - H(X_i^{g_i}|X_1^{g_1},\ldots,X_{i-1}^{g_{i-1}}) \quad .$$

By not considering the subtracted terms $H(X_i^{g_i}|X_1^{g_1},\ldots)$, we have $I(\mathbf{D}^G) \leq \sum_{i=2}^{d} H(X_i^{g_i})$. Since $H(X_i^G) \leq \log(n_i)$, we arrive at $I(\mathbf{D}^G) \leq \sum_{i=2}^{d} \log(n_i)$. Considering all permutations of $\{X_1\ \ldots, X_d\}$, it holds that

$$I(\mathbf{D}^G) \leq \sum_{i=1}^{d} \log(n_i) - \max_{1\leq i\leq d} \log(n_i) \quad .$$

□

Proof of Theorem 3. Let $g^* = \arg\min_{g:|g|=l} h(X|Y^g(1,m))$. We denote l bins that g^* generates on Y as $b_1,\ldots,b_l$. We write $|b_t|$ as the number of values of Y in b_t. Further, let $c_z = \sum_{t=1}^{z} |b_t|$. Note that each bin of Y is non-empty, i.e., $c_z \geq z$. We use $h(X|b_t)$ to denote $h(X)$ computed using the points of $\mathbf{D}$ corresponding to the realizations of Y in b_t, projected onto X. We have: $f(m,l)$

$$\begin{aligned}
&= \sum_{t=1}^{l} \frac{|b_t|}{m} h(X|b_t) \\
&= \sum_{t=1}^{l-1} \frac{|b_t|}{m} h(X|b_t) + \frac{|b_l|}{m} h(X|b_l) \\
&= \frac{c_{l-1}}{m} \sum_{t=1}^{l-1} \frac{|b_t|}{c_{l-1}} h(X|b_t) + \frac{|b_l|}{m} h(X|b_l) \\
&= \frac{c_{l-1}}{m} f(c_{l-1}, l-1) \\
&\quad + \frac{m - c_{l-1}}{m} h(X|\langle Y(c_{l-1}+1, m)\rangle) \quad .
\end{aligned}$$

In the last line, $\sum_{t=1}^{l-1} \frac{|b_t|}{c_{l-1}} h(X|b_t)$ is equal to $f(c_{l-1}, l-1)$ because otherwise, we could decrease $f(m,l)$ by choosing a different discretization of $Y(1, c_{l-1})$ into $l-1$ bins. This in turn contradicts our definition of $f(m,l)$. Since $c_{l-1} \in [l-1, m)$ and $f(m,l)$ is minimal over all $j \in [l-1, m)$, we arrive at the final result. □

Proof of Theorem 4. Let

$$g^* = \arg\max_{g:|g|=l} H(X^g(1,m)) - H(X^g(1,m), C_k) \quad .$$

We denote l bins that g^* generates on X as $b(X)_1,\ldots,b(X)_l$. We write $|b(X)_t|$ as the number of values of X in $b(X)_t$. For each $X_i' \in C_k$, we denote its bins as $b(X_i')_1,\ldots,b(X_i')_{n_i'}$.

Let $c_z = \sum_{t=1}^{z} |b(X)_t|$. Note that each bin of X is non-

empty, i.e., $c_z \geq z$. We use $H(C_k|b_t)$ to denote $H(C_k)$ computed using the points of $\mathbf{D}$ corresponding to the realizations of X in b_t, projected onto C_k.

We write $(t, t_1, \ldots, t_k)$ as the number of points in the cell made up by bins $b(X)_t, b(X'_1)_{t_1}, \ldots, b(X'_k)_{t_k}$. We use $(t, *, \ldots, *)$ to also denote $b(X)_t$. We note that

$$|(t, *, \ldots, *)| = \sum_{t_1=1}^{n'_1} \cdots \sum_{t_k=1}^{n'_k} |(t, t_1, \ldots, t_k)|$$

We have: $F(m, l)$

$$\begin{aligned}
&= \sum_{t=1}^{l} \frac{|(t, *, \ldots, *)|}{m} \log \frac{m}{|(t, *, \ldots, *)|} \\
&- \sum_{t=1}^{l} \sum_{t_1=1}^{n'_1} \cdots \sum_{t_k=1}^{n'_k} \frac{|(t, t_1, \ldots, t_k)|}{m} \log \frac{m}{|(t, t_1, \ldots, t_k)|} \\
&= \sum_{t=1}^{l} \sum_{t_1=1}^{n'_1} \cdots \sum_{t_k=1}^{n'_k} \frac{|(t, t_1, \ldots, t_k)|}{m} \log \frac{|(t, t_1, \ldots, t_k)|}{|(t, *, \ldots, *)|} \\
&= \sum_{t=1}^{l-1} \sum_{t_1=1}^{n'_1} \cdots \sum_{t_k=1}^{n'_k} \frac{|(t, t_1, \ldots, t_k)|}{m} \log \frac{|(t, t_1, \ldots, t_k)|}{|(t, *, \ldots, *)|} \\
&\quad + \sum_{t_1=1}^{n'_1} \cdots \sum_{t_k=1}^{n'_k} \frac{|(l, t_1, \ldots, t_k)|}{m} \log \frac{|(l, t_1, \ldots, t_k)|}{|(l, *, \ldots, *)|} \\
&= \frac{c_{l-1}}{m} \times \\
&\sum_{t=1}^{l-1} \sum_{t_1=1}^{n'_1} \cdots \sum_{t_k=1}^{n'_k} \frac{|(t, t_1, \ldots, t_k)|}{c_{l-1}} \log \frac{|(t, t_1, \ldots, t_k)|}{|(t, *, \ldots, *)|} \\
&\quad + \frac{|(l, *, \ldots, *)|}{m} \times \\
&\sum_{t_1=1}^{n'_1} \cdots \sum_{t_k=1}^{n'_k} \frac{|(l, t_1, \ldots, t_k)|}{|(l, *, \ldots, *)|} \log \frac{|(l, t_1, \ldots, t_k)|}{|(l, *, \ldots, *)|} \\
&= \frac{c_{l-1}}{m} F(c_{l-1}, l-1) \\
&\quad - \frac{m - c_{l-1}}{m} H(C_k|b(X)_l) \\
&= \frac{c_{l-1}}{m} F(c_{l-1}, l-1) \\
&\quad - \frac{m - c_{l-1}}{m} H(C_k|\langle X(c_{l-1}+1, m)\rangle) \quad .
\end{aligned}$$

In the last line,

$$\sum_{t=1}^{l-1} \sum_{t_1=1}^{n'_1} \cdots \sum_{t_k=1}^{n'_k} \frac{|(t, t_1, \ldots, t_k)|}{c_{l-1}} \log \frac{|(t, t_1, \ldots, t_k)|}{|(t, *, \ldots, *)|}$$

is equal to $F(c_{l-1}, l-1)$ because otherwise, we could increase $F(m, l)$ by choosing a different discretization of $X(1, c_{l-1})$ into $l-1$ bins. This in turn contradicts our definition of $F(m, l)$. Since $c_{l-1} \in [l-1, m)$ and $F(m, l)$ is maximal over all $j \in [l-1, m)$, we arrive at the final result. □

Proof of the properties of Eq. (7) and (8) of the main paper. For the first property, Eq. (7) is the normalized total correlation of discretized X and all dimensions in C_k, so its value is in $[0, 1]$. Considering Eq. (8), from $h(X) \geq h(X|C_k)$, we have

$$\left(\sum_{i=1}^{k} H(X'_i)\right) + h(X) - h(X|C_k) - H(C_k) \geq 0 \quad .$$

Further, from $h(X|C_k) \geq 0$

$$\begin{aligned}
&\left(\sum_{i=1}^{k} H(X'_i)\right) + h(X) - h(X|C_k) - H(C_k) \\
&\leq h(X) + \sum_{i=1}^{k} \log n'_i - \max(\{\log n'_i\}_{i=1}^{k}) \quad .
\end{aligned}$$

Therefore, the value of Eq. (8) is in $[0, 1]$.

For the second property, from Theorem 1, it holds that Eq. (7) is equal to zero iff discretized X and all dimensions in C_k are statistically independent. Also, Eq. (8) is equal to zero iff all the dimensions in C_k are statistically independent, *and* X is independent of C_k; hence,

$$p(X, C_k) = p(X)p(C_k) = p(X)p(X'_1) \cdots p(X'_k)$$

Thus, X and all dimensions in C_k are statistically independent. We therefore have proven the second property.

For the third property, following (Han, 1978), it holds that Eq. (7) is maximal when there exists $X'_i \in C_k$ such that discretized X and all dimensions in $C_k \setminus \{X'_i\}$, each is a function of X'_i. Considering Eq. (8), it is maximal when (a) there exists $X'_i \in C_k$ such that all dimensions in $C_k \setminus \{X'_i\}$, each is a function of X'_i, and (b) X is a function of C_k. These two conditions imply X is also a function of X'_i. Thus, we have the third property proven. □

2. Extended Related Work

Canonical correlation analysis (CCA) was first proposed in (Hotelling, 1936) for analyzing two data sets of the same size. This seminal work focuses on linear correlations. Prominent subsequent developments could be categorized into two lines. The first line is to extend CCA to capture non-linear correlations, for instance, Kernel CCA (Hardoon et al., 2004), CCA based on mutual information (Yin, 2004), Deep CCA (Andrew et al., 2013), and

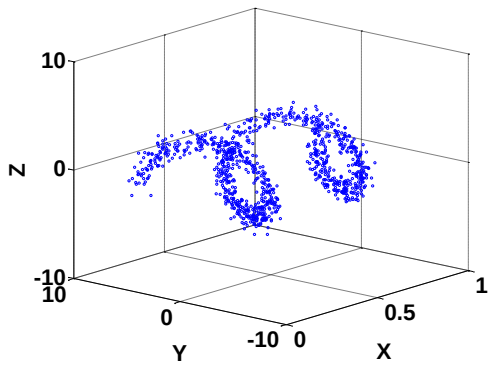
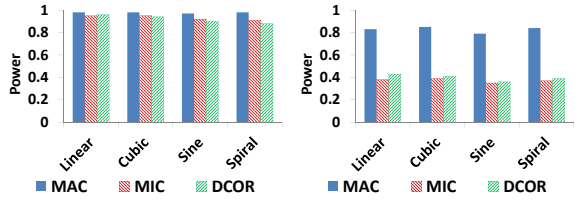


Figure 1. 3-d spiral function.



(a) Noise = 20% (b) Noise = 80%

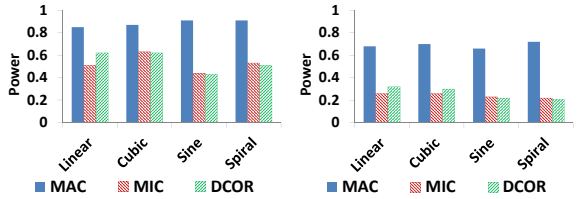
Figure 2. [Higher is better] Statistical power vs. noise for 4-dimensional functions.

CCA based on Hilbert-Schmidt independence criterion and the centered kernel target alignment (Chang et al., 2013). All these powerful developments mainly target the pairwise setting. The second line extends CCA to the multivariate setting (Carroll, 1968; Kettenring, 1971; van de Velden & Takane, 2012). Related methods are designed to handle multiple data sets with the focus so far on linear correlations. It is an interesting research direction to find a merge of the two lines of CCA extensions to enable CCA to detect non-linear correlations in the multivariate setting.

3. Extended Set of Experiments

Results on multivariate functional correlations. Figure 1 shows an example 3-d spiral function. Figure 2 displays the statistical power against different noise levels of MAC, MIC, and DCOR on 4-variate functions. The results of 128-variate functions are in Figure 3. We can see that MAC outperforms the other two methods as noise increases.

Cluster analysis. The clustering Accuracy and F1 of all methods are displayed in Table 1. One can see that MAC obtains the best clustering results on all datasets given the same subspace search scheme.



(a) Noise = 20% (b) Noise = 80%

Figure 3. [Higher is better] Statistical power vs. noise for 128-dimensional functions.

References

Andrew, Galen, Arora, Raman, Bilmes, Jeff, and Livescu, Karen. Deep canonical correlation analysis. In *ICML (3)*, pp. 1247–1255, 2013.

Carroll, J. D. Generalization of canonical correlation analysis to three or more sets of variables. In *Proceedings of the American Psychological Association*, pp. 227–228, 1968.

Chang, Billy, Krüger, Uwe, Kustra, Rafal, and Zhang, Junping. Canonical correlation analysis based on hilbert-schmidt independence criterion and centered kernel target alignment. In *ICML (2)*, pp. 316–324, 2013.

Hardoon, David R., Szedmak, Sandor, and Shawe-Taylor, John. Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*, 16(12):2639–2664, 2004.

Hotelling, Harold. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377, 1936.

Kettenring, J. R. Canonical analysis of several sets of variables. *Biometrika*, 58(3):433–451, 1971.

van de Velden, Michel and Takane, Yoshio. Generalized canonical correlation analysis with missing values. *Computational Statistics*, 27(3):551–571, 2012.

Yin, Xiangrong. Canonical correlation analysis based on information theory. *Journal of Multivariate Analysis*, 91 (2):161–176, 2004.

	MAC	**MIC**	**CMI**	**ENCLUS**
Musk (6598×166)				
F1	**0.89**	0.66	0.64	0.62
Accuracy	**0.88**	0.70	0.72	0.63
Letter (20000×16)				
F1	**0.82**	0.64	0.64	0.64
Accuracy	**0.84**	0.69	0.72	0.71
PenDigits (7494×16)				
F1	**0.85**	0.48	0.73	0.50
Accuracy	**0.88**	0.71	0.81	0.66
Waveform (5000×40)				
F1	**0.50**	0.28	0.31	0.31
Accuracy	**0.65**	0.37	0.38	0.40
WBCD (569×30)				
F1	**0.74**	0.55	0.58	0.57
Accuracy	**0.92**	0.80	0.82	0.75
Diabetes (768×8)				
F1	**0.72**	0.54	0.71	0.25
Accuracy	**0.78**	0.53	0.76	0.67
Glass (214×9)				
F1	**0.70**	0.32	0.59	0.26
Accuracy	**0.70**	0.47	0.68	0.52

Table 1. Clustering results on real-world data sets.