CMI: An Information-Theoretic Contrast Measure for Enhancing Subspace Cluster and Outlier Detection

Hoang Vu Nguyen[○] Emmanuel Müller[○●] Jilles Vreeken[●] Fabian Keller[○] Klemens Böhm[○]

[○] Karlsruhe Institute of Technology, Germany
{hoang.nguyen, emmanuel.mueller, fabian.keller, klemens.boehm}@kit.edu

[●] University of Antwerp, Belgium
{emmanuel.mueller, jilles.vreeken}@ua.ac.be

Abstract

In many real world applications data is collected in multi-dimensional spaces, with the knowledge hidden in subspaces (i.e., subsets of the dimensions). It is an open research issue to select meaningful subspaces without any prior knowledge about such hidden patterns. Standard approaches, such as pairwise correlation measures, or statistical approaches based on entropy, do not solve this problem; due to their restrictive pairwise analysis and loss of information in discretization they are bound to miss subspaces with potential clusters and outliers.

In this paper, we focus on finding subspaces with strong mutual dependency in the selected dimension set. Chosen subspaces should provide a high discrepancy between clusters and outliers and enhance detection of these patterns. To measure this, we propose a novel contrast score that quantifies mutual correlations in subspaces by considering their cumulative distributions—without having to discretize the data. In our experiments, we show that these high contrast subspaces provide enhanced quality in cluster and outlier detection for both synthetic and real world data.

1 Introduction

Clustering and outlier detection are two key data mining tasks. They are widely used, such as in bioinformatics, for detecting functionally dependent genes, in marketing, for customer segmentation, in health surveillance, for anomaly detection, and so on. For these techniques to work well, some kind of dependency between the objects in a given data space is required, i.e., high similarity between clustered objects and high deviation between outliers and the residual data distribution.

Obviously, detecting clusters and outliers in uniformly random distributed spaces, e.g., considering a data space with independent dimensions, does not make sense at all. With more and more dimensions such effects tend to hinder data mining tasks, which is widely known as the "curse of dimensionality" [6]. Not just a fringe theoretical case, we observe this effect in practice, for example, in gene expression analysis where each gene is described with very many expression values under different medical treatments. In general, we observe a loss of contrast between clusters and outliers in the full space (all given dimensions) of the data, while the meaningful knowledge is hidden in subspaces (i.e., subsets of the dimensions) that show a high dependency between the selected dimensions.

Recently, more attention has been placed on *subspace clustering* [3, 1, 20, 21] and *subspace outlier detection* [2, 16, 22]. Both of these paradigms detect a set of relevant dimensions for each individual cluster or oultier. Hence, they are able to detect meaningful patterns even if only few dimensions are relevant for the individual pattern. However, they all face a common problem in the selection of subspaces. Each of the techniques re-invents a very specific subspace selection scheme according to the underlying cluster or outlier model. Only few techniques have focused on general solutions to the problem of *subspace search* designed for clustering [9, 5] or outlier mining [27, 15]. In this work, we follow this general idea of subspace search. We aim at a further generalization for the selection of relevant subspaces.

More specifically, we aim at selecting *high contrast subspaces* that potentially provide high contrast between clustered and outlying objects. Due to its generality this problem statement poses several open questions. First, it is unclear how to measure the contrast of a given set of dimensions. Solutions based on correlation analysis and entropy measures seem promising but show major drawbacks w.r.t. pairwise analysis, discretization, and the empty space problem, as we will explain later. Second, one requires robust statistics to capture the mutual dependence of dimensions. Existing solutions performing a pairwise analysis miss important higher-order dependencies that can only be identified when multiple dimensions are considered together. Finally, a subspace selection has to be performed in an efficient manner in order to scale with the increasing number of dimensions, i.e., an exponential search space.

We tackle all three of these challenges by our contrast measure. It is independent of any cluster or outlier model and purely based on the statistical dependence of data observed in a multi-dimensional subspace. Furthermore, it is directly applicable to continuous data and does not fall prey to the information loss by previous discretization techniques. It is designed to capture mutual dependencies, and thus, quantifies the subspace deviation from the condition of uncorrelated and independent dimensions: "The larger the deviation from the mutual independence assumption, the higher the contrast of a subspace." Hence, we instantiate our measure based on the analysis of cumulative distributions in different subspaces. Cumulative distributions have the advantage that they can be computed directly on empirical data. Furthermore, we propose a scalable processing scheme to select high contrast subspaces. Due to the exponential search space we rely on an approximative solution based on beam search.

Overall, our contributions are as follows: (a) a set of abstract quality criteria for subspace search based on contrast analysis, (b) our multi-variate contrast measure based on cumulative distributions for continuous data, (c) a scalable subspace search method applying our contrast measure for subspace selection, and (d) quality enhancement for both subspace clustering and outlier mining as a result of high contrast.

2 Related Work

Pairwise measures and space transformations. First, we discuss approaches that assess dependencies between dimensions. Spearman correlation and modern variants [24] are aimed at pairwise correlations. However, higher order interactions (i.e., mutual dependence) among several dimensions can be missed. Similarly, dimensionality reduction techniques [19], including PCA, are not aware of locally clustered projections; they only measure the (non-)linear dependence between dimensions, meaning that they consider one (global) projection, and may hence miss interesting local projections containing subspace clusters and outliers. Our method, on the other hand, is not limited to a pairwise assessment and provides multiple projections for clustering and outlier mining. It can cope with mutual dependencies in arbitrary subspace projections.

Feature selection. Next, we consider methods for unsupervised feature selection. Recent methods [11, 17] perform iteratively a partitioning and feature selection. They first partition the data (e.g., by EM clustering), and then they evaluate feature subsets based on the obtained clusters. Another approach [25] aims at different feature subsets for different clusters. However, it focuses on disjoint clusters and does neither allow overlapping clusters nor outliers. Our method is more general and is aware of outliers and overlap of clusters. In general, feature selection differs from our approach in major aspects. Current feature selection methods are specifically bound to clustering. In contrast, our method is more general and suitable for both cluster and outlier mining in multiple subspaces. Most approaches [11, 17] select a single projection of the data space, which uncovers some certain cluster structure in the data. These methods are limited to one subspace, while we mine multiple possibly overlapping subspaces. Yet keeping only one subspace may miss local projections containing different subspace clusters [21].

Subspace search. We now discuss methods for selecting relevant subspaces. They avoid the limitations of the above paradigms, and focus on multiple projections with arbitrary dimensionality. Existing methods, however, rely on discretization of continuous dimensions [9, 27] or only work with binary data [28] and/or discrete data [8].

ENCLUS [9] and PODM [27] detect subspaces with low entropy and high interest, discretizing continuous dimensions into equi-width bins in order to compute the entropy measure. By requiring discretization, these methods have unintuitive parameters, and are hence inherently susceptible to knowledge loss and to the curse of dimensionality. To some extent, these limitations have been tackled by HiCS [15], which works directly on continuous data. It quantifies the differences between the marginal and conditional distribution in a random dimension of the considered subspace; by its random nature it may hence might miss relevant subspaces. Further, it is exposed to the curse of dimensionality w.r.t. conditional distributions in high-dimensional spaces.

Our method, on the other hand, can reliably score contrast, regardless of subspace dimensionality. Furthermore, for each subspace we aim to find that permutation of dimensions that yields optimal contrast.

Cluster and outlier detection in subspaces. Specific methods for clustering and outlier detection have been proposed. However, they do not provide a general notion of subspace selection. They select subspaces very specifically to the underlying cluster [3, 1, 26, 20, 21] or outlier [2, 16, 22] definitions. In contrast to all these solutions, our goal is to design a contrast measure that is applicable to subspace selection for different mining paradigms. We show its instantiations to clustering and outlier detection and evaluate its quality.

3 Basic Notions for Contrast Assessment

Given a database DB of size N and dimensionality D, we want to measure the contrast of any lower dimensional subspace S with dimensionality $1 \leq d \leq$

D. Our assessment is based on the full space of all dimensions given by $F = \{X_1, \ldots, X_D\}$. Each dimension i is associated with a random variable X_i that has a continuous value domain $dom(X_i) = \mathbb{R}$. We use the notion of density distribution $p_{X_i}(x_i)$ for the projected database on dimension i. We write $p_{X_i}(x_i)$ as $p(x_i)$ when the context is clear. Any non-empty subset $S \in \mathcal{P}(F)$ is called a subspace of DB. The dimensionality of S is denoted as $dim(S)$. W.l.o.g., $\{X_1, \ldots, X_d\}$ is used as representative for any d-dimensional subspace S in our analysis.

3.1 Contrast Assessment. As our general notion of a contrast measure we have the following formalization:

DEFINITION 3.1. ***Contrast Measure of Subspaces:***

$$C : \mathcal{P}(F) \setminus \{\emptyset\} \to \mathbb{R}$$

In general, the contrast score $C(S)$ quantifies the difference of S w.r.t. the baseline of d independent and randomly distributed dimensions. In the following we provide different instantiation of this contrast measure and discuss formal properties of the instantiations. Let us first formalize the independence baseline. For d random variables $X_1, \ldots, X_d$, there are two types of independence we are interested in.

DEFINITION 3.2. ***Mutual Independence:***
$X_1, \ldots, X_d$ are mutually independent iff

$$p(x_1, \ldots, x_d) = p(x_1) \cdots p(x_d)$$

DEFINITION 3.3. ***m-wise Independence:***
$X_1, \ldots, X_d$ are m-wise independent with $m \leq d$ iff any subset $\{X_{i_1}, \ldots, X_{i_m}\} \subseteq \{X_1, \ldots, X_d\}$ is mutually independent.

Please note that pairwise independence is modeled as a special case of m-wise independence when $m = 2$. However, pairwise analysis misses important higher-order dependencies that can only be identified when multiple dimensions are considered altogether. Therefore, we focus on higher-order dependencies and their contrast assessment. A subspace is referred to as uncorrelated if its dimensions are mutually independent. Our goal is to design a contrast measure C that quantifies as closely as possible the deviation of subspaces from uncorrelated ones. In other words, for a d-dimensional subspace S with dimensions $\{X_1, \ldots, X_d\}$, its contrast depends on how much the difference between $p(x_1, \ldots, x_d)$ and $p(x_1) \cdots p(x_d)$ is:

$$C(S) \sim \mathit{diff}\,(p(x_1, \ldots, x_d), p(x_1) \cdots p(x_d))$$

Contrast of one-dimensional subspaces is undefined. Thus, we restrict the contrast measure C to two- or higher-dimensional subspaces. In the following, we propose three properties for a meaningful contrast assessment based on the idea "deviating from uncorrelated subspaces":

Property 1 (Discriminative contrast scores): For subspaces S_1 and S_2 such that $dim(S_1) = dim(S_2)$, if S_1 is more correlated than S_2 then $C(S_1) > C(S_2)$.

Property 2 (Zero contrast score): $C(S) = 0$ if and only if the dimensions of S are mutually independent.

Property 3 (Awareness of m-wise independence): If the dimensions of S are m-wise independent but not mutually independent then $C(S)$ is small but not zero. This is because m-wise independence does not guarantee mutual independence.

Furthermore, C should be directly applicable to continuous data, i.e., we do not require discretization to obtain the probability mass functions. Since discretization causes knowledge loss, this property is mandatory.

3.2 Discussion of Properties. Looking at existing techniques, ENCLUS [9] instantiates the *diff* function by the well-known total correlation $\sum_{i=1}^{d} H(X_i) - H(X_1, \ldots, X_d)$ where $X_1, \ldots, X_d$ are *discretized* versions of the original dimensions. PODM [27] also discretizes data and instantiates the *diff* function as $\sum \frac{1}{p(x_1, \ldots, x_d)}$ where $p(x_1, \ldots, x_d) \neq 0$. The instantiation of HiCS [15] is done by averaging over multiple random runs of the form $\mathit{diff}\,(p(x_i), p(x_i | \{x_1, \ldots, x_d\} \setminus \{x_i\}))$ where X_i is picked randomly.

None of these techniques fulfills all properties mentioned. Considering Property 1, the measure of ENCLUS is unreliable because of the knowledge loss caused by data discretization. Further, the use of the joint probability mass function $p(x_1, \cdots, x_d)$ also is problematic. In particular, $H(X_1, \ldots, X_d) = -\sum p(x_1, \ldots, x_d) \log p(x_1, \ldots, x_d)$ with $p(x_1, \ldots, x_d)$ measured by the relative number of points in the respective hypercube. For increasing d, most hypercubes are empty and the non-empty ones most likely contain only one data point each [2, 19]. Taking into account that $\lim_{x \to 0} x \log x = 0$, $H(X_1, \ldots, X_d)$ approaches $-\sum_{i=1}^{N} \frac{1}{N} \log \frac{1}{N} = \log N$. Hence, when d is large enough and all X_i have a similar distribution (e.g., uniformly dense), any d-dimensional subspaces S_1 and S_2 have very similar contrast: $C(S_1) \approx C(S_2)$. In other words, the measure of ENCLUS produces indifferent contrasts for high-dimensional subspaces. Thus, it fails to satisfy Property 1, i.e., the most basic property. PODM relying on data discretization and the joint probability $p(x_1, \ldots, x_d)$ suffers the same issue. As for HiCS, the random choice of X_i causes potential loss of

contrast as some attribute may not be tested against the remaining ones. In addition, HiCS uses conditional probability distributions with $(d-1)$ conditions and exposes itself to the same problem of empty space.

Considering Properties 2 and 3, since ENCLUS works with discretized data that causes loss of knowledge, it only satisfies these properties with a proper grid resolution. Such a resolution is data-dependent. PODM misses both Properties 2 and 3 since its measure just relies on the joint probability, i.e., it does not measure dependency. A zero contrast assigned by HiCS does not imply uncorrelated spaces since there is no guarantee that all dimensions are assessed against the others at least once. Thus, HiCS does not meet Property 2. Furthermore, HiCS does not aim at m-wise independence and thus does not address Property 3.

4 Methodology

In order to address all three properties, we first introduce a novel notion of mutual information, called Cumulative Mutual Information (*CMI*), which is instantiated based on a new notion of entropy, called Cumulative Entropy (*CE*). We then verify that *CMI* addresses Properties 1 to 3. Since *CMI* is dependent on the order of subspace dimensions, we then devise an approach to select a dimension permutation that approximates the optimal *CMI* value for a given subspace. Due to space limitation, all proofs for the following theorems will be provided as an extended version of this paper.

4.1 Cumulative mutual information.

Given continuous random variables $X_1, \ldots, X_d$, their cumulative mutual information $CMI(X_1, \ldots, X_d)$ is defined as:

$$\sum_{i=2}^{d} \mathit{diff}\,(p(x_i), p(x_i|x_1, \ldots, x_{i-1}))$$

Intuitively, $CMI(X_1, \ldots, X_d)$ measures the mutual information of $X_1, \ldots, X_d$ by aggregating the difference between $p(x_i)$ and $p(x_i|x_1, \ldots, x_{i-1})$ for $2 \leq i \leq d$. Loosely speaking, it is the sum of the contrasts of subspaces $(X_1, X_2), \ldots, (X_1, \ldots, X_i), \ldots, (X_1, \ldots, X_d)$ if we consider $\mathit{diff}\,(p(x_i), p(x_i|x_1, \ldots, x_{i-1}))$ to be the contrast of the subspace $(X_1, \ldots, X_i)$. The reason for using lower-dimensional subspace projections is to avoid the empty space phenomenon. Since probability functions are not available at hand and can only be estimated, e.g., by data discretization, we aim at implementing $\mathit{diff}\,(p(x_i), p(x_i|x_1, \ldots, x_{i-1}))$ using cumulative distributions. In this paper, we instantiate *CMI* by means of *CE* and conditional *CE* that are based on cumulative distributions. We demonstrate in Section 5 how these allow efficient contrast calculation without discretizing data. Their definitions are given below:

Definition 4.1. *The cumulative entropy for a continuous random variable X, denoted $h_{CE}(X)$, is defined as:*

$$h_{CE}(X) = -\int_{dom(X)} P(X \leq x) \log P(X \leq x) dx$$

Our notion of cumulative entropy is based on [10]. However, it is more general since it is not restricted to non-negative random variables. Furthermore, we extend the notion of *CE* to conditional cumulative entropy and prove that it maintains some important properties of traditional conditional entropy as follows:

Definition 4.2. *The conditional CE of any continuous random variable X knowing that some random vector $V \in \mathbb{R}^B$ (with B being a positive integer) takes the value v is defined as:*

$$h_{CE}(X|v) = -\int_{dom(X)} P(X \leq x|v) \log P(X \leq x|v) dx$$

The CE of X conditioned by V is:

$$E_V[h_{CE}(X|V)] = \int_{dom(V)} h_{CE}(X|v) p(v) dv$$

Just like the usual conditional entropy, we denote $E_V[h_{CE}(X|V)]$ as $h_{CE}(X|V)$ for notational convenience. The conditional *CE* has two important properties given by the following theorems:

Theorem 4.1. *$E_V[h_{CE}(X|V)] \geq 0$ with equality iff there exists a function $f : dom(V) \to dom(X)$ such that $X = f(V)$.*

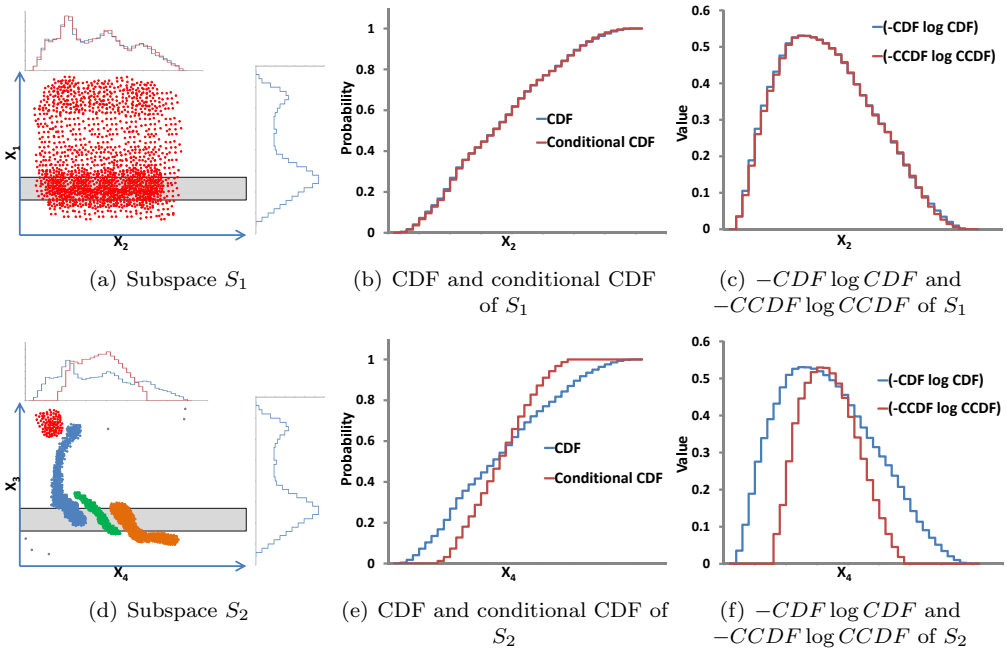
Theorem 4.2. *$E_V[h_{CE}(X|V)] \leq h_{CE}(X)$ with equality iff X is independent of V.*

Under *CE*, $\mathit{diff}\,(p(x), p(x|\ldots))$ is set to $h_{CE}(X) - h_{CE}(X|\ldots)$. Therefore, $CMI(X_1, \ldots, X_d)$ becomes:

$$\sum_{i=2}^{d} h_{CE}(X_i) - \sum_{i=2}^{d} h_{CE}(X_i|X_1, \ldots, X_{i-1})$$

where $h_{CE}(X_i|X_1, \ldots, X_{i-1})$ is $h_{CE}(X_i|V)$ with $V = (X_1, \ldots, X_{i-1})$ being a random vector in $dom(X_1) \times \cdots \times dom(X_{i-1})$.

Regarding the three properties, similar to traditional mutual information, the more correlated $X_1, \ldots, X_d$ are, the smaller the conditional *CE*s are, i.e., the larger is *CMI*. Thus *CMI* is able to capture subspace correlation (Property 1). To illustrate this property, we use the toy example in Figure 1. It depicts the scatter plots, CDF plots, and plots of the function $-P(X \leq x) \log P(X \leq x)$, namely $-CDF \log CDF$,



(a) Subspace S_1 (b) CDF and conditional CDF of S_1 (c) $-CDF \log CDF$ and $-CCDF \log CCDF$ of S_1

(d) Subspace S_2 (e) CDF and conditional CDF of S_2 (f) $-CDF \log CDF$ and $-CCDF \log CCDF$ of S_2

Figure 1: Example of low and high contrast subspaces with different *CMI*s

of two subspaces S_1 and S_2 (CCDF means conditional CDF). The blue lines stand for the marginal distribution of the corresponding dimension. The red lines feature the conditional distribution of one dimension obtained by selecting a range of the remaining dimension (gray strips). One can see that S_2 has higher contrast than S_1 and hence, $CMI(X_3, X_4)_{\text{selected range}} = 4.344 > CMI(X_1, X_2)_{\text{selected range}} = 0.113$. Further, even when high-order conditional *CE*s may be impacted by the curse of dimensionality, *CMI* still yields distinguishable contrast for high-dimensional subspaces due to its member low-order conditional *CE*s. If $X_1, \ldots, X_d$ are m-wise independent, then $CMI(X_1, \ldots, X_d)$ is low as $h_{CE}(X_i) - h_{CE}(X_i|\ldots)$ vanishes for $i \leq m$ (Property 3). However, we have proved that $CMI = 0$ iff $X_1, \ldots, X_d$ are mutually independent (Property 2).

THEOREM 4.3. *$CMI(X_1, \ldots, X_d) \geq 0$ with equality iff $X_1, \ldots, X_d$ are mutually independent.*

4.2 Choice of permutation. *CMI* can be used as our contrast measure. However, *CMI* changes with dimension permutations. In order to make our contrast measure permutation-independent we investigate a heuristic search of the maximal contrast.

Our goal is to find a permutation that maximizes the contrast of a given subspace $S = \{X_1, \ldots, X_d\}$. Since *CMI* is permutation variant, there are $d!$ possible cases in total. Together with the exponential number of subspaces, a brute-force approach is impractical. We therefore apply a heuristic to obtain a permutation that approximates the optimal one. In particular, we first pick a pair of dimensions X_a and X_b ($1 \leq a \neq b \leq d$) such that $h_{CE}(X_b) - h_{CE}(X_b|X_a)$ is maximal among the possible pairs. We then continue selecting the next dimension X_c ($c \neq a$ and $c \neq b$) such that $h_{CE}(X_c) - h_{CE}(X_c|X_a, X_b)$ is maximal among the remaining dimensions. Likewise, at each step, assuming $I = \{X_{p_1}, \ldots, X_{p_k}\}$ is the set of dimensions already picked and $R = \{X_{r_1}, \ldots, X_{r_{d-k}}\}$ is the set of remaining ones, we select the dimension $X_{r_i} \in R$ such that $h_{CE}(X_{r_i}) - h_{CE}(X_{r_i}|I)$ is maximal. The process goes on until no dimension is left. Denoting the permutation obtained by our strategy as π_{opt}, the contrast of S is defined as $CMI(\pi_{opt}(X_1, \ldots, X_d))$.

5 Algorithmic Approach

For a D-dimensional data set, there are $2^D - 1$ candidate subspaces to examine. The exponential number of subspaces makes a brute-force search impractical. A scalable subspace exploration framework is required. Moreover, the contrast measure must also permit efficient computation. In this section, we first introduce an approximate yet scalable levelwise subspace search framework. We then proceed to discuss how to compute our measure efficiently.

5.1 Scalable subspace exploration. Our aim is to mine high contrast subspaces upon which subspace clus-

tering and outlier detection techniques are applied. To tackle the exponential search space, we target at a processing scheme that trades off accuracy for efficiency. More specifically, we rely on the intuition that a high contrast high-dimensional subspace likely has its high contrast reflected in its lower-dimensional projections. In the field of subspace clustering, there is an analogous observation: Subspace clusters tend to have their data points clustered in all of their lower-dimensional projections [3, 21]. One can then apply a levelwise scheme to mine subspaces of contrast larger than a pre-specified value. However, to facilitate parameterization of our method, we avoid imposing direct thresholds on contrast scores produced by *CMI*.

Instead, we design a beam search strategy to obtain efficiency. Starting with two-dimensional subspaces, in each step we use top M subspaces of high contrast to generate new candidates in a levelwise manner. A newly generated candidate is only considered if all of its child subspaces have high contrast. First, this permits tractable time complexity. Second, interaction among different subspace dimensionality is taken into account and selected subspaces are ensured to have high contrast. Third, we avoid redundancy, if $T \subseteq S$ and S has higher contrast than T then T is excluded from the final result.

5.2 Efficient contrast computation. To compute *CMI*, we need to compute *CE* and conditional *CE*.

Let $X_1 \leq \ldots \leq X_n$ be i.i.d. random samples of the continuous random variable X. Then $h_{CE}(X)$ can be calculated as follows:

$$h_{CE}(X) = -\sum_{i=1}^{n-1}(X_{i+1} - X_i)\frac{i}{n}\log\frac{i}{n}$$

In contrast to this straightforward computation, it is not as simple to calculate the conditional *CE* in an accurate and efficient way. In the following, we first point out that due to limited data, sticking to the exact formula of conditional *CE* may lead to inaccurate results. We then propose a strategy to resolve this while ensuring that data discretization is not required.

First, w.l.o.g., consider the space $[-1/2, 1/2]^d$ containing N limited data points. The d dimensions are $X_1, \ldots, X_d$. Our goal is to compute $h_{CE}(X_1|X_2, \ldots, X_d)$ using limited available data. From Definition 4.2: $h_{CE}(X_1|X_2, \ldots, X_d) = \int_{-1/2}^{1/2} \cdots \int_{-1/2}^{1/2} h(X_1|x_2, \ldots, x_d)p(x_2, \ldots, x_d)dx_2 \cdots dx_d$. Further:

$h_{CE}(X_1|x_2, \ldots, x_d) = \lim_{\varepsilon \to 0^+} h_{CE}(X_1|x_2 - \varepsilon \leq X_2 \leq x_2 + \varepsilon, \ldots, x_d - \varepsilon \leq X_d \leq x_d + \varepsilon)$

Taking into account that the total number of data points N is limited, the expected number of points contained in the hypercube $[x_2 - \varepsilon, x_2 + \varepsilon] \times \cdots \times [x_d - \varepsilon, x_d + \varepsilon]$, which is $N(2\varepsilon)^{d-1}$, approaches 0 as $\varepsilon \to 0^+$. For high-dimensional spaces, the problem is exacerbated as one faces the empty space phenomenon. With empty hypercubes (or even hypercubes of one data point), $h_{CE}(X_1|x_2, \ldots, x_d)$ vanishes. Hence, $h_{CE}(X_1|X_2, \ldots, X_d)$ becomes 0. We thus encounter a paradox: *If sticking to the exact formula of conditional CE, we may end up with an inaccurate result.* To alleviate this problem, we must ensure to have enough points for meaningful calculation. Therefore, we propose data summarization by clustering.

Clustering summarizes the data by means of clusters. Since the number of clusters is generally much less than the original data size, we may have more data points in each cluster. Hence, the issue of limited data is mitigated. Assuming that a clustering algorithm A is used on DS projected to $\{X_2, \ldots, X_d\}$ resulting in Q clusters $\{C_1, \ldots, C_Q\}$ (the support of C_i is $|C_i|$), we propose to estimate $h_{CE}(X_1|X_2, \ldots, X_d)$ by:

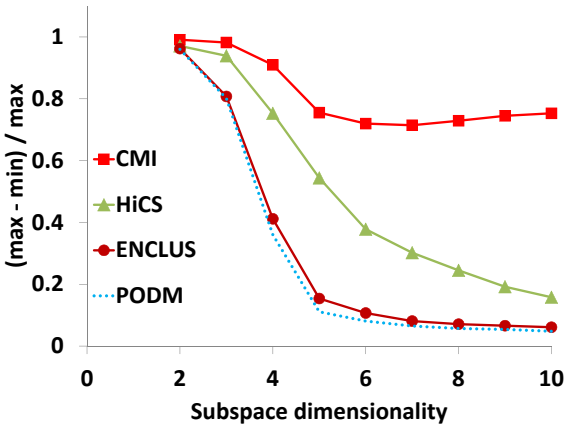
$$\sum_{i=1}^{Q} \frac{|C_i|}{N} h_{CE}(X_1|C_i)$$

If Q is kept small enough, we will have enough points for a meaningful computation of $h_{CE}(X_1|C_i)$ regardless of the dimensionality d. As our cluster-based approach does not rely on any specific cluster notion, it can be instantiated by any method. To ensure efficient computation of the contrast measure, we use the one-pass k-means clustering strategy introduced in [23] with $k = Q$. We obtain Q clusters summarizing the data. For the parameter Q, if it is set too high, we may end up with high runtime and not enough data in each cluster for a reliable estimation of conditional *CE*. If it is instead set to 1, i.e., no clustering at all, $h_{CE}(X_1|\cdots)$ becomes $h_{CE}(X_1)$, i.e., there is a loss of information. In all of our experiments, we set $Q = 10$. Using clustering, one can verify that the conditional *CE* is less than or equal to its respective unconditional one.

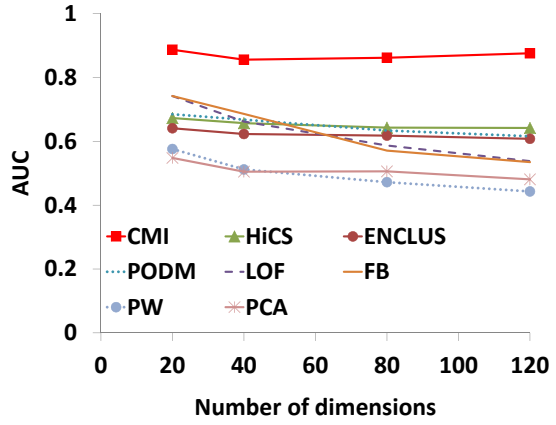
6 Experiments

We compare *CMI*, to three subspace search methods: ENCLUS [9], HiCS [15], and PODM [27]. As further baselines we include random selection (FB) [18], PCA [19], and pairwise correlation (PW) [24]. For *CMI* we use $M = 400$ and $Q = 10$, unless stated otherwise. In order to assist comparability and future research in this area, we provide our algorithm, all datasets, parameters and further material on our website.[1]

[1] `http://www.ipd.kit.edu/~muellere/CMI/`



(a) Contrast score vs. dimensionality



(b) AUC vs. dimensionality

Figure 2: Subspace quality w.r.t. dimensionality

We evaluate how mining of high contrast subspaces improves the result quality of outlier detection and clustering techniques. Therefore, LOF [7] and DBSCAN [12], two well-established methods, are used on top of the tested approaches. For fair comparison, we use the same parameter settings for both LOF and DBSCAN.

To ensure succinct sets of subspaces that allow for post-analysis, only the best 100 subspaces of each technique are utilized for clustering and outlier detection. Outlier detection results are assessed by the Area Under the ROC Curve (AUC) as in [18, 15, 22]. Clustering results are evaluated by means of F1, Accuracy, and E4SC as in [21, 14].

6.1 Impact of dimensionality. To illustrate that *CMI* is robust w.r.t. increasing dimensionality of subspaces, we evaluate it on a synthetic data set of 20 dimensions and 5120 instances, generated according to [22]. In this data, subspace clusters are embedded in randomly selected 2–10 dimensional subspaces. Additionally, 120 outliers are created deviating from these clusters. Please note that in this experiment, we perform an exhaustive search without any pruning. Because of the large total number of subspaces $(2^{20} - 1)$, we only experiment up to $d = 10$ to avoid excessive runtime. We record $\frac{\max A_d - \min A_d}{\max A_d}$ where A_d is the set of contrast scores of all d-dimensional subspaces. For $2 \leq d \leq 10$, $\min A_d \approx 0$ (as there are uncorrelated d-dimensional subspaces) and $\max A_d \neq 0$ (as there are correlated d-dimensional subspaces with clusters and outliers). Hence, ideally $\frac{\max A_d - \min A_d}{\max A_d} = 1$ for $2 \leq d \leq 10$. The results, plotted in Figure 2(a), show that HiCS, ENCLUS, and PODM do not scale well with higher dimensionality. In contrast, *CMI* is more robust to dimensionality and yields discriminative contrast scores even for high-dimensional subspaces.

6.2 Synthetic data: cluster and outlier mining. Based on the method described in [22], we generate synthetic data sets with 5120 data points and 20, 40, 80, and 120 dimensions. Each data set contains subspace clusters embedded in randomly chosen 2-6 dimensional subspaces and 120 outliers deviating from these clusters.

Quality for outlier mining. The quality of subspaces is evaluated by inspecting how the selected subspaces enhance outlier detection compared to LOF in the full space. The results are shown as Figure 2(b). Overall, *CMI* outperforms the competing techniques and is stable with increasing dimensionality. The performance of LOF degrades with increasing dimensionality of data. Similarly, FB [18] is affected by random choice of low contrast projections. The pairwise method PW [24] and PCA show worst performance, due to their inability to measure contrast in multi-dimensional subspaces. As subsequent evaluation confirmed this trend, we exclude PW and PCA in the experiments below.

Quality for clustering. Here, subspace quality is assessed by clustering results. DBSCAN is used as the baseline method. Furthermore, for all methods tested, we reduced redundancy in clustering output [4]. The results in Table 1 show that *CMI* achieves best quality and best scalability for increasing dimensionality. High E4SC values of *CMI* indicate that it performs well in selecting subspaces containing clusters and outliers.

Runtime vs. Dimensionality. Besides accuracy, we are also interested in scalability w.r.t. runtime. In this experiment, previous synthetic data sets are reused. Since the tendency of all methods is similar in both outlier detection and clustering, we only present the runtime for outlier detection. We display in Figure 3(a) the total time for completing the task, i.e., time for mining subspaces (cf., Figure 3(b)) and time for outliers mining. We can see that *CMI* scales better than our competitors.

	CMI	HiCS	Enclus	Podm	DBScan	FB
			20 dimensions			
F1	**0.96**	**0.96**	0.72	0.75	0.65	0.67
Acc.	**0.98**	0.96	0.75	0.82	0.67	0.68
E4SC	**0.92**	0.75	0.42	0.36	0.19	0.27
			40 dimensions			
F1	**0.93**	0.88	0.65	0.72	0.54	0.61
Acc.	**0.93**	0.74	0.68	0.76	0.61	0.66
E4SC	**0.89**	0.73	0.27	0.34	0.21	0.23
			80 dimensions			
F1	**0.94**	0.83	0.62	0.68	0.57	0.61
Acc.	**0.95**	0.74	0.66	0.81	0.62	0.69
E4SC	**0.86**	0.57	0.22	0.34	0.24	0.25
			120 dimensions			
F1	**0.94**	0.86	0.52	0.61	0.55	0.63
Acc.	**0.94**	0.72	0.68	0.71	0.58	0.62
E4SC	**0.87**	0.64	0.18	0.23	0.21	0.19

Table 1: Clustering results on synthetic data sets

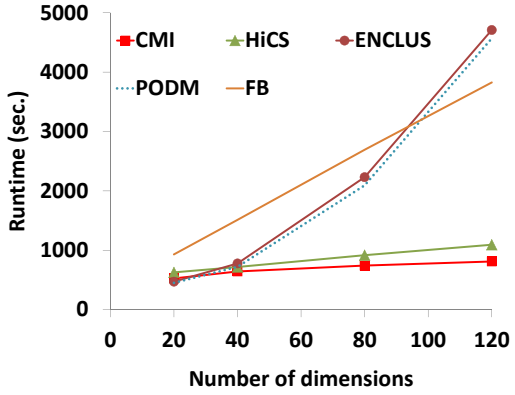
Dataset	CMI	HiCS	Enclus	Podm	LOF	FB
Thyroid	**0.96**	0.95	0.94	0.91	0.86	0.93
WBCD	**0.95**	0.94	0.94	0.87	0.87	0.87
Diabetes	**0.73**	0.72	0.71	0.69	0.71	0.72
Glass	**0.82**	0.80	0.80	0.78	0.77	0.78
Ion	**0.83**	0.82	0.82	0.78	0.78	0.79
Pendigits	**0.98**	0.95	0.94	0.86	0.94	0.93
Segment	**0.94**	0.84	0.88	0.89	0.76	0.86
Lympho	**0.95**	0.86	0.67	0.67	**0.95**	**0.95**
Madelon	**0.60**	0.59	0.51	0.56	0.59	0.59

Table 2: Outlier mining: AUC on real world data

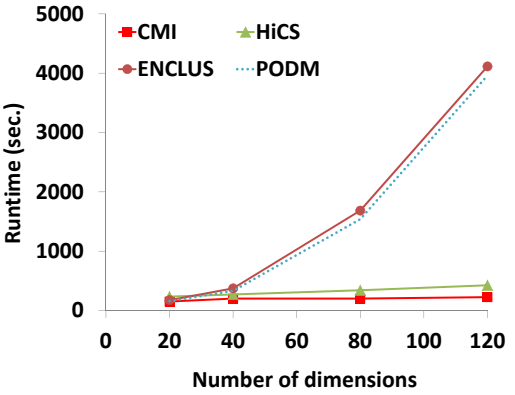
Although *FB* does not spend much time for mining high contrast subspaces, it clearly suffers from high overall runtimes. This is due to high-dimensional subspaces that very likely have low contrast, and hence, induce the costly detection of many false alarms. ENCLUS and PODM also scale badly as their contrast measures are inefficient in terms of time complexity. Since *CMI* prunes low contrast subspaces much better than HiCS, it can avoid exploring many high-dimensional subspaces. The inclusion of such subspaces on the other hand causes the outlier score computation phase of HiCS to be longer. In conclusion, *CMI* is faster than all tested approaches and yields higher accuracy.

6.3 Evaluation on real world data. All real world databases used in our experiments are from the UCI Machine Learning Repository [13] and have been used as benchmarks in recent publications [18, 20, 21, 15].

Quality for outlier mining. We evaluate the performance of all subspace search methods with outlier detection on real world data. We perform experiments on 9 benchmark datasets, using the minority class as ground truth for the evaluation of the detected outliers. In some of these data sets, e.g., *Pendigits*, all classes have identical support and we down-sample one class to 10% of its original size—a commonly used procedure in outlier evaluation [18, 15, 22]. The results in Table 2 show that *CMI* achieves the best AUC in all data sets. In addition, we show the wall-clock runtimes in Table 3. The overall conclusion is that our method provides the best quality enhancement for LOF.



(a) total runtime



(b) subspace search only

Figure 3: Runtime w.r.t. dimensionality

Dataset	CMI	HiCS	Enclus	Podm	FB
Thyroid	**17.33**	27.54	49.32	48.11	53.60
WBCD	**16.42**	17.11	33.63	34.55	24.49
Diabetes	**1.74**	1.80	4.74	4.63	5.56
Glass	**0.24**	**0.24**	0.27	0.26	0.27
Ion	**6.01**	6.19	7.31	7.19	8.07
Pendigits	**1368.23**	1616.96	2153.09	2094.36	1854.56
Segment	**101.23**	107.99	225.46	218.34	150.80
Lympho	**4.10**	6.08	6.37	5.79	5.31
Madelon	**23.45**	25.82	315.22	304.57	232.48

Table 3: Runtime (in seconds) for outlier detection

Quality for clustering. As we show in Table 4, *CMI* provides also the best quality improvement w.r.t. clustering. It outperforms traditional full space DBSCAN and existing subspace search methods that fail to identify clusters due to scattered subspace projections. In contrast to the competing approaches, we achieve a clear quality enhancement for both subspace clustering and subspace outlier detection.

	CMI	HiCS	Enclus	Podm	DBScan	FB
			Wisconsin Breast Cancer			
F1	**0.79**	0.75	0.44	0.40	0.73	0.60
Acc.	**0.77**	0.72	0.69	0.67	0.71	0.69
E4SC	**0.76**	0.70	0.53	0.49	0.67	0.59
			Shape			
F1	**0.82**	0.77	0.76	0.74	0.55	0.76
Acc.	**0.84**	0.78	0.66	0.69	0.34	0.41
E4SC	**0.71**	0.64	0.58	0.63	0.38	0.44
			Pendigits			
F1	**0.73**	0.55	0.50	0.51	0.52	0.63
Acc.	**0.81**	0.75	0.66	0.64	0.68	0.77
E4SC	**0.68**	0.54	0.56	0.55	0.52	0.53
			Diabetes			
F1	**0.71**	0.53	0.25	0.15	0.52	0.58
Acc.	**0.76**	0.66	0.67	0.63	0.68	0.70
E4SC	**0.65**	0.34	0.11	0.07	0.52	0.52
			Glass			
F1	**0.59**	0.37	0.26	0.29	0.32	0.42
Acc.	**0.68**	0.54	0.52	0.55	0.32	0.44
E4SC	**0.52**	0.40	0.35	0.38	0.24	0.28

Table 4: Clustering: Quality on real world data

7 Conclusions

We proposed *CMI*, a new contrast measure for multi-dimensional data. It is based on cumulative entropy of subspaces and does not require data discretization. Furthermore, it is not restricted to pairwise analysis, captures mutual dependency among dimensions, and scales well with increasing subspace dimensionality. Overall, it is more accurate and more efficient than previous subspace search methods. Experiments on various real world databases show that *CMI* provides improvement for both cluster and outlier detection.

Acknowledgments

This work is supported by the German Research Foundation (DFG) within GRK 1194 (Hoang Vu Nguyen), by the YIG program of KIT as part of the German Excellence Initiative (Emmanuel Müller). Emmanuel Müller and Jilles Vreeken are supported by Post-Doctoral Fellowships of the Research Foundation – Flanders (FWO).

References

[1] C. C. Aggarwal, C. M. Procopiuc, J. L. Wolf, P. S. Yu, and J. S. Park. Fast algorithms for projected clustering. In *SIGMOD*, 1999.

[2] C. C. Aggarwal and P. S. Yu. Outlier detection for high dimensional data. In *SIGMOD*, 2001.

[3] R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan. Automatic subspace clustering of high dimensional data for data mining applications. In *SIGMOD*, 1998.

[4] I. Assent, R. Krieger, E. Müller, and T. Seidl. DUSC: Dimensionality unbiased subspace clustering. In *ICDM*, 2007.

[5] C. Baumgartner, C. Plant, K. Kailing, H.-P. Kriegel, and P. Kröger. Subspace selection for clustering high-dimensional data. In *ICDM*, 2004.

[6] K. S. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft. When is "nearest neighbor" meaningful? In *ICDT*, 1999.

[7] M. M. Breunig, H.-P. Kriegel, and J. S. Raymond T. Ng. LOF: Identifying density-based local outliers. In *SIGMOD*, 2000.

[8] P. Chanda, J. Yang, A. Zhang, and M. Ramanathan. On mining statistically significant attribute association information. In *SDM*, 2010.

[9] C. H. Cheng, A. W.-C. Fu, and Y. Zhang. Entropy-based subspace clustering for mining numerical data. In *KDD*, 1999.

[10] A. D. Crescenzo and M. Longobardi. On cumulative entropies. *J. Statist. Plann. Inference*, 139, 2009.

[11] J. G. Dy and C. E. Brodley. Feature selection for unsupervised learning. *JMLR*, 5, 2004.

[12] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD*, 1996.

[13] A. Frank and A. Asuncion. UCI machine learning repository [http://archive.ics.uci.edu/ml], 2010.

[14] S. Günnemann, I. Färber, E. Müller, I. Assent, and T. Seidl. External evaluation measures for subspace clustering. In *CIKM*, 2011.

[15] F. Keller, E. Müller, and K. Böhm. HiCS: High contrast subspaces for density-based outlier ranking. In *ICDE*, 2012.

[16] H.-P. Kriegel, E. Schubert, A. Zimek, and P. Kröger. Outlier detection in axis-parallel subspaces of high dimensional data. In *PAKDD*, 2009.

[17] M. H. C. Law, M. A. T. Figueiredo, and A. K. Jain. Simultaneous feature selection and clustering using mixture models. *IEEE Trans. Pattern Anal. Mach. Intell.*, 26(9), 2004.

[18] A. Lazarevic and V. Kumar. Feature bagging for outlier detection. In *KDD*, 2005.

[19] J. Lee and M. Verleysen. *Nonlinear Dimensionality Reduction*. Springer, New York, 2007.

[20] E. Müller, I. Assent, S. Günnemann, R. Krieger, and T. Seidl. Relevant subspace clustering: Mining the most interesting non-redundant concepts in high dimensional data. In *ICDM*, 2009.

[21] E. Müller, S. Günnemann, I. Assent, and T. Seidl. Evaluating clustering in subspace projections of high dimensional data. *PVLDB*, 2(1), 2009.

[22] E. Müller, M. Schiffer, and T. Seidl. Statistical selection of relevant subspace projections for outlier ranking. In *ICDE*, 2011.

[23] C. Ordonez and E. Omiecinski. Efficient disk-based K-means clustering for relational databases. *IEEE Trans. Knowl. Data Eng.*, 16(8), 2004.

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

[25] V. Roth and T. Lange. Feature selection in clustering problems. In *NIPS*, 2003.

[26] K. Sequeira and M. J. Zaki. SCHISM: A new approach for interesting subspace mining. In *ICDM*, 2004.

[27] M. Ye, X. Li, and M. E. Orlowska. Projected outlier detection in high-dimensional mixed-attributes data set. *Expert Syst. Appl.*, 36(3), 2009.

[28] X. Zhang, F. Pan, W. Wang, and A. B. Nobel. Mining non-redundant high order correlations in binary data. *PVLDB*, 1(1), 2008.