Nikolaj Tatti. 2007. Distances between data sets based on summary statistics. Journal of Machine Learning Research, volume 8, pages 131-154.

Journal of Machine Learning Research 8 (2007) 131-154 Submitted 2/05; Revised 9/05; Published 1/07

Distances between Data Sets Based on Summary Statistics

Nikolaj Tatti NTATTI@CC.HUT.FI
HIIT Basic Research Unit
Laboratory of Computer and Information Science
Helsinki University of Technology, Finland

Editor: Dale Schuurmans

Abstract

The concepts of similarity and distance are crucial in data mining. We consider the problem of defining the distance between two data sets by comparing summary statistics computed from the data sets. The initial definition of our distance is based on geometrical notions of certain sets of distributions. We show that this distance can be computed in cubic time and that it has several intuitive properties. We also show that this distance is the unique Mahalanobis distance satisfying certain assumptions. We also demonstrate that if we are dealing with binary data sets, then the distance can be represented naturally by certain parity functions, and that it can be evaluated in linear time. Our empirical tests with real world data show that the distance works well.

Keywords: data mining theory, complex data, binary data, itemsets

1. Introduction

In this paper we will consider the following problem: Given two data sets D_1 and D_2 of dimension K, define a distance between D_1 and D_2. To be more precise, we consider the problem of defining the distance between two multisets of transactions, each set sampled from its own unknown distribution. We will define a dissimilarity measure between D_1 and D_2 and we will refer to this measure as *CM distance*.

Generally speaking, the notion of dissimilarity between two objects is one of the most fundamental concepts in data mining. If one is able to retrieve a distance matrix from a set of objects, then one is able to analyse data by using for example, clustering or visualisation techniques. Many real world data collections may be naturally divided into several data sets. For example, if a data collection consists of movies from different eras, then we may divide the movies into subcollections based on their release years. A distance between these data (sub)sets would provide means to analyse them as single objects. Such an approach may ease the task of understanding complex data collections.

Let us continue by considering the properties the CM distance should have. First of all, it should be a metric. The motivation behind this requirement is that the metric theory is a well-known area and metrics have many theoretical and practical virtues. Secondly, in our scenario the data sets have statistical nature and the CM distance should take this into account. For example, consider that both data sets are generated from the same distribution, then the CM distance should give small values and approach 0 as the number of data points in the data sets increases. The third requirement is that we should be able to evaluate the CM distance quickly. This requirement is crucial since we may have high dimensional data sets with a vast amount of data points.

The CM distance will be based on summary statistics, features. Let us give a simple example: Assume that we have data sets $D_1 = \{A,B,A,A\}$ and $D_2 = \{A,B,C,B\}$ and assume that the only feature we are interested in is the proportion of A in the data sets. Then we can suggest the distance between D_1 and D_2 to be $|3/4 - 1/4| = 1/2$. The CM distance is based on this idea; however, there is a subtle difficulty: If we calculate several features, then should we take into account the correlation of these features? We will do exactly that in defining the CM distance.

The rest of this paper is organised as follows. In Section 2 we give the definition of the CM distance by using some geometrical interpretations. We also study the properties of the distance and provide an alternative characterisation. In Section 3 we study the CM distance and binary data sets. In Section 4 we discuss how the CM distance can be used with event sequences and in Section 5 we comment about the feature selection. Section 6 is devoted for related work. The empirical tests are represented in Section 7 and we conclude our work with the discussion in Section 8.

2. The Constrained Minimum Distance

In the following subsection we will define our distance using geometrical intuition and show that the distance can be evaluated efficiently. In the second subsection we will discuss various properties of the distance, and in the last subsection we will provide an alternative justification to the distance. The aim of this justification is to provide more theoretical evidence for our distance.

2.1 The Definition

We begin by giving some basic definitions. By a *data set D* we mean a finite collection of samples lying in some finite space Ω. The set Ω is called *sample space*, and from now on we will denote this space by the letter Ω. The number of elements in Ω is denoted by $|\Omega|$. The number of samples in the data set D is denoted by $|D|$.

As we said in the introduction, our goal is not to define a distance directly on data sets but rather through some statistics evaluated from the data sets. In order to do so, we define a *feature function* $S : \Omega \to \mathbb{R}^N$ to map a point in the sample space to a real vector. Throughout this section S will indicate some given feature function and N will indicate the dimension of the range space of S. We will also denote the i^{th} component of S by S_i. Note that if we have several feature functions, then we can join them into one big feature function. A *frequency* $\theta \in \mathbb{R}^N$ of S taken with respect to a data set D is the average of values of S taken over the data set, that is, $\theta = \frac{1}{|D|}\sum_{\omega \in D} S(\omega)$. We denote this frequency by $S(D)$.

Although we do not make any assumptions concerning the size of Ω, some of our choices are motivated by thinking that $|\Omega|$ can be very large—so large that even the simplest operation, say, enumerating all the elements in Ω, is not tractable. On the other hand, we assume that N is such that an algorithm executable in, say, $O(N^3)$ time is feasible. In other words, we seek a distance whose evaluation time does not depend of the size of Ω but rather of N.

Let $\mathbb{P}$ be the set of all distributions defined on Ω. Given a feature function S and a frequency θ (calculated from some data set) we say that a distribution $p \in \mathbb{P}$ satisfies the frequency θ if $\mathrm{E}_p[S] = \theta$. We also define a *constrained set of distributions*

$$\mathcal{C}_+(S,\theta) = \{p \in \mathbb{P} \mid \mathrm{E}_p[S] = \theta\}$$

to be the set of the distributions satisfying θ. The idea behind this is as follows: From a given data set we calculate some statistics, and then we examine the distributions that can produce such frequencies.

We interpret the sets $\mathbb{P}$ and $\mathcal{C}_+(S,\theta)$ as *geometrical objects*. This is done by enumerating the points in Ω, that is, we think that $\Omega = \{1,2,,\ldots,|\Omega|\}$. We can now represent each distribution $p \in \mathbb{P}$ by a vector $u \in \mathbb{R}^{|\Omega|}$ by setting $u_i = p(i)$. Clearly, $\mathbb{P}$ can be represented by the vectors in $\mathbb{R}^{|\Omega|}$ having only non-negative elements and summing to one. In fact, $\mathbb{P}$ is a simplex in $\mathbb{R}^{|\Omega|}$. Similarly, we can give an alternative definition for $\mathcal{C}_+(S,\theta)$ by saying

$$\mathcal{C}_+(S,\theta) = \left\{ u \in \mathbb{R}^{|\Omega|} \mid \sum_{i\in\Omega} S(i)u_i = \theta, \sum_{i\in\Omega} u_i = 1, u \geq 0 \right\}. \quad (1)$$

Let us now study the set $\mathcal{C}_+(S,\theta)$. In order to do so, we define a *constrained space*

$$\mathcal{C}(S,\theta) = \left\{ u \in \mathbb{R}^{|\Omega|} \mid \sum_{i\in\Omega} S(i)u_i = \theta, \sum_{i\in\Omega} u_i = 1 \right\},$$

that is, we drop the last condition from Eq. 1. The set $\mathcal{C}_+(S,\theta)$ is included in $\mathcal{C}(S,\theta)$; the set $\mathcal{C}_+(S,\theta)$ consists of the non-negative vectors from $\mathcal{C}(S,\theta)$. Note that the constraints defining $\mathcal{C}(S,\theta)$ are vector products. This implies that $\mathcal{C}(S,\theta)$ is an affine space, and that, given two different frequencies θ_1 and θ_2, the spaces $\mathcal{C}(S,\theta_1)$ and $\mathcal{C}(S,\theta_2)$ are parallel.

Example 1 *Let us illustrate the discussion above with a simple example. Assume that* $\Omega = \{A,B,C\}$. *We can then imagine the distributions as vectors in* $\mathbb{R}^3$. *The set* $\mathbb{P}$ *is the triangle having* $(1,0,0)$, $(0,1,0)$, *and* $(0,0,1)$ *as corner points (see Figure 1). Define a feature function S to be*

$$S(\omega) = \begin{cases} 1 & \omega = C \\ 0 & \omega \neq C. \end{cases}$$

The frequency $S(D)$ *is the proportion of C in a data set D. Let* $D_1 = (C,C,C,A)$ *and* $D_2 = (C,A,B,A)$. *Then* $S(D_1) = 0.75$ *and* $S(D_2) = 0.25$. *The spaces* $\mathcal{C}(S,0.25)$ *and* $\mathcal{C}(S,0.75)$ *are parallel lines (see Figure 1). The distribution sets* $\mathcal{C}_+(S,0.25)$ *and* $\mathcal{C}_+(S,0.75)$ *are the segments of the lines* $\mathcal{C}(S,0.25)$ *and* $\mathcal{C}(S,0.75)$, *respectively.*

The idea of interpreting distributions as geometrical objects is not new. For example, a well-known boolean query problem is solved by applying linear programming to the constrained sets of distributions (Hailperin, 1965; Calders, 2003).

Let us revise some elementary Euclidean geometry: Assume that we are given two parallel affine spaces $\mathcal{A}_1$ and $\mathcal{A}_2$. There is a natural way of measuring the distance between these two spaces. This is done by taking the length of the shortest segment going from a point in $\mathcal{A}_1$ to a point in $\mathcal{A}_2$ (for example see the illustration in Figure 1). We know that the segment has the shortest length if and only if it is orthogonal with the affine spaces. We also know that if we select a point $a_1 \in \mathcal{A}_1$ having the shortest norm, and if we similarly select $a_2 \in \mathcal{A}_2$, then the segment going from a_1 to a_2 has the shortest length.

The preceding discussion and the fact that the constrained spaces are affine motivates us to give the following definition: Assume that we are given two data sets, namely D_1 and D_2 and a

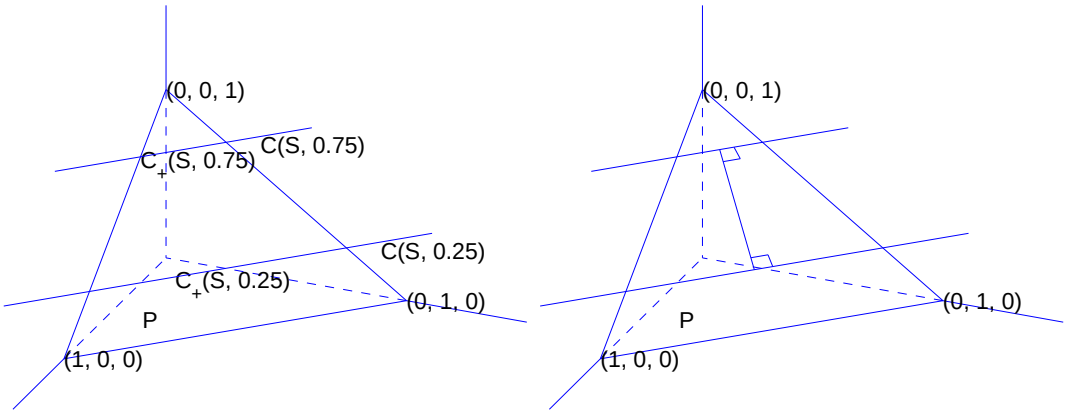


Figure 1: A geometrical interpretation of the distribution sets for $|\Omega| = 3$. In the left figure, the set $\mathbb{P}$, that is, the set of all distributions, is a triangle. The constrained spaces $\mathcal{C}(S, 0.25)$ and $\mathcal{C}(S, 0.75)$ are parallel lines and the distribution sets $\mathcal{C}_+(S, 0.25)$ and $\mathcal{C}_+(S, 0.75)$ are segments of the constrained spaces. In the right figure we added a segment perpendicular to the constraint spaces. This segment has the shortest length among the segments connecting the constrained spaces.

feature function S. Let us shorten the notation $\mathcal{C}(S, S(D_i))$ by $\mathcal{C}(S, D_i)$. We pick a vector from each constrained space having the shortest norm

$$u_i = \operatorname*{argmin}_{u \in \mathcal{C}(S, D_i)} \|u\|_2, \quad i = 1, 2.$$

We define the distance between D_1 and D_2 to be

$$d_{CM}(D_1, D_2 \mid S) = \sqrt{|\Omega|}\, \|u_1 - u_2\|_2. \tag{2}$$

The reasons for having the factor $\sqrt{|\Omega|}$ will be given later. We will refer to this distance as *Constrained Minimum (CM) distance*. We should emphasise that u_1 or u_2 may have negative elements. Thus the CM distance is *not* a distance between two distributions; it is rather a distance based on the frequencies of a given feature function and is motivated by the geometrical interpretation of the distribution sets.

The main reason why we define the CM distance using the constrained spaces $\mathcal{C}(S, D_i)$ and not the distribution sets $\mathcal{C}_+(S, D_i)$ is that we can evaluate the CM distance efficiently. We discussed earlier that Ω may be very large so it is crucial that the evaluation time of a distance does not depend on $|\Omega|$. The following theorem says that the CM distance can be represented using the frequencies and a covariance matrix

$$\operatorname{Cov}[S] = \frac{1}{|\Omega|} \sum_{\omega \in \Omega} S(\omega) S(\omega)^T - \left(\frac{1}{|\Omega|} \sum_{\omega \in \Omega} S(\omega) \right) \left(\frac{1}{|\Omega|} \sum_{\omega \in \Omega} S(\omega) \right)^T.$$

Theorem 1 *Assume that* $\operatorname{Cov}[S]$ *is invertible. For the CM distance between two data sets* D_1 *and* D_2 *we have*

$$d_{CM}(D_1, D_2 \mid S)^2 = (\theta_1 - \theta_2)^T \operatorname{Cov}^{-1}[S] (\theta_1 - \theta_2),$$

where $\theta_i = S(D_i)$.

The proofs for the theorems are given in Appendix.

The preceding theorem shows that we can evaluate the distance using the covariance matrix and frequencies. If we assume that evaluating a single component of the feature function S is a unit operation, then the frequencies can be calculated in $O(N|D_1| + N|D_2|)$ time. The evaluation time of the covariance matrix is $O(|\Omega| N^2)$ but we assume that S is such that we know a closed form for the covariance matrix (such cases will be discussed in Section 3), that is, we assume that we can evaluate the covariance matrix in $O(N^2)$ time. Inverting the matrix takes $O(N^3)$ time and evaluating the distance itself is $O(N^2)$ operation. Note that calculating frequencies and inverting the covariance matrix needs to be done only once: for example, assume that we have k data sets, then calculating the distances between every data set pair can be done in $O\left(N\sum_i^k |D_i| + N^3 + k^2N^2\right)$ time.

Example 2 *Let us evaluate the distance between the data sets given in Example 1 using both the definition of the CM distance and Theorem 1. We see that the shortest vector in* $\mathcal{C}(S, 0.25)$ *is* $u_1 = \left(\frac{3}{8}, \frac{3}{8}, \frac{1}{4}\right)$. *Similarly, the shortest vector in* $\mathcal{C}(S, 0.75)$ *is* $u_2 = \left(\frac{1}{8}, \frac{1}{8}, \frac{3}{4}\right)$. *Thus the CM distance is equal to*

$$d_{CM}(D_1, D_2 \mid S) = \sqrt{3}\,\|u_1 - u_2\|_2 = \sqrt{3}\left[\frac{2^2}{8^2} + \frac{2^2}{8^2} + \frac{2^2}{4^2}\right]^{1/2} = \frac{3}{\sqrt{8}}.$$

The covariance of S is equal to $\mathrm{Cov}[S] = \frac{1}{3} - \frac{1}{3}\frac{1}{3} = \frac{2}{9}$. *Thus Theorem 1 gives us*

$$d_{CM}(D_1, D_2 \mid S) = \left[\mathrm{Cov}^{-1}[S]\left(\frac{3}{4} - \frac{1}{4}\right)^2\right]^{1/2} = \left[\frac{9}{2}\left(\frac{2}{4}\right)^2\right]^{1/2} = \frac{3}{\sqrt{8}}.$$

From Theorem 1 we see a reason to have the factor $\sqrt{|\Omega|}$ in Eq. 2: Assume that we have two data sets D_1 and D_2 and a feature function S. We define a new sample space $\Omega' = \{(\omega, b) \mid \omega \in \Omega, b = 0, 1\}$ and transform the original data sets into new ones by setting $D_i' = \{(\omega, 0) \mid \omega \in D_i\}$. We also expand S into Ω' by setting $S'(\omega, 1) = S'(\omega, 0) = S(\omega)$. Note that $S(D_i) = S'(D_i')$ and that $\mathrm{Cov}[S] = \mathrm{Cov}[S']$ so Theorem 1 says that the CM distance has not changed during this transformation. This is very reasonable since we did not actually change anything essential: We simply added a bogus variable into the sample space, and we ignored this variable during the feature extraction. The size of the new sample space is $|\Omega'| = 2|\Omega|$. This means that the difference $\|u_1 - u_2\|_2$ in Eq. 2 is smaller by the factor $\sqrt{2}$. The factor $\sqrt{|\Omega|}$ is needed to negate this effect.

2.2 Properties

We will now list some important properties of $d_{CM}(D_1, D_2 \mid S)$.

Theorem 2 $d_{CM}(D_1, D_2 \mid S)$ *is a pseudo metric.*

The following theorem says that adding external data set to the original data sets makes the distance smaller which is very reasonable property.

Theorem 3 *Assume three data sets* D_1, D_2, *and* D_3 *over the same set of items. Assume further that* D_1 *and* D_2 *have the same number of data points and let* $\varepsilon = \frac{|D_3|}{|D_1| + |D_3|}$. *Then*

$$d_{CM}(D_1 \cup D_3, D_2 \cup D_3 \mid S) = (1 - \varepsilon) d_{CM}(D_1, D_2 \mid S).$$

Theorem 4 *Let A be a $M \times N$ matrix and b a vector of length M. Define $T(\omega) = AS(\omega) + b$. It follows that $d_{CM}(D_1, D_2 \mid T) \leq d_{CM}(D_1, D_2 \mid S)$ for any D_1 and D_2.*

Corollary 5 *Adding extra feature functions cannot decrease the distance.*

Corollary 6 *Let A be an invertible $N \times N$ matrix and b a vector of length N. Define $T(\omega) = AS(\omega) + b$. It follows that $d_{CM}(D_1, D_2 \mid T) = d_{CM}(D_1, D_2 \mid S)$ for any D_1 and D_2.*

Corollary 6 has an interesting interpretation. Note that $T(D) = AS(D) + b$ and that $S(D) = A^{-1}(T(D) - b)$. This means that if we know the frequencies $S(D)$, then we can infer the frequencies $T(D)$ without a new data scan. Similarly, we can infer $S(D)$ from $T(D)$. We can interpret this relation by thinking that $S(D)$ and $T(D)$ are merely different representations of the same feature information. Corollary 6 says that the CM distance is equal for any such representation.

2.3 Alternative Characterisation of the CM Distance

We derived our distance using geometrical interpretation of the distribution sets. In this section we will provide an alternative way for deriving the CM distance. Namely, we will show that if some distance is of Mahalanobis type and satisfies some mild assumptions, then this distance is proportional to the CM distance. The purpose of this theorem is to provide more theoretical evidence to our distance.

We say that a distance d is of Mahalanobis type if

$$d(D_1, D_2 \mid S)^2 = (\theta_1 - \theta_2)^T C(S)^{-1} (\theta_1 - \theta_2),$$

where $\theta_1 = S(D_1)$ and $\theta_2 = S(D_2)$ and $C(S)$ maps a feature function S to a symmetric $N \times N$ matrix. Note that if $C(S) = \mathrm{Cov}[S]$, then the distance d is the CM distance. We set $\mathbb{M}$ to be the collection of all distances of Mahalanobis type. Can we justify the decision that we examine only the distances included in $\mathbb{M}$? One reason is that a distance belonging to $\mathbb{M}$ is guaranteed to be a metric. The most important reason, however, is the fact that we can evaluate the distance belonging to $\mathbb{M}$ efficiently (assuming, of course, that we can evaluate $C(S)$).

Let $d \in \mathbb{M}$ and assume that it satisfies two additional assumptions:

1. If A is an $M \times N$ matrix and b is a vector of length M and if we set $T(\omega) = AS(\omega) + b$, then $C(T) = AC(S)A^T$.
2. Fix two points ω_1 and ω_2. Let $\sigma : \Omega \to \Omega$ be a function swapping ω_1 and ω_2 and mapping everything else to itself. Define $U(\omega) = S(\sigma(\omega))$. Then $d(\sigma(D_1), \sigma(D_2) \mid U) = d(D_1, D_2 \mid S)$.

The first assumption can be partially justified if we consider that A is an invertible square matrix. In this case the assumption is identical to $d(\cdot, \cdot \mid AS + b) = d(\cdot, \cdot \mid S)$. This is to say that the distance is independent of the representation of the frequency information. This is similar to Corollary 6 given in Section 2.2. We can construct a distance that would satisfy Assumption 1 in the invertible case but fail in a general case. We consider such distances pathological and exclude them by making a broader assumption. To justify Assumption 2 note that the frequencies have not changed, that is, $U(\sigma(D)) = S(D)$. Only the representation of single data points have changed. Our argument is that the distance should be based on the frequencies and not on the values of the data points.

Theorem 7 *Let $d \in M$ satisfying Assumptions 1 and 2. If $C(S)$ is invertible, then there is a constant $c > 0$, not depending on S, such that $d(\cdot, \cdot \mid S) = c d_{CM}(\cdot, \cdot \mid S)$.*

3. The CM distance and Binary Data Sets

In this section we will concentrate on the distances between binary data sets. We will consider the CM distance based on itemset frequencies, a very popular statistics in the literature concerning binary data mining. In the first subsection we will show that a more natural way of representing the CM distance is to use parity frequencies. We also show that we can evaluate the distance in linear time. In the second subsection we will provide more theoretical evidence why the CM distance is a good distance between binary data sets.

3.1 The CM Distance and Itemsets

We begin this section by giving some definitions. We set the sample space Ω to be

$$\Omega = \{\omega \mid \omega = (\omega_1, \ldots, \omega_K), \omega_i = 0, 1\},$$

that is, Ω is the set of all binary vectors of length K. Note that $|\Omega| = 2^K$. It is custom that each dimension in Ω is identified with some symbol. We do this by assigning the symbol a_i to the i^{th} dimension. These symbols are called *attributes* or *items*. Thus when we speak of the attribute a_i we refer to the i^{th} dimension. We denote the set of all items by $\mathbb{A} = \{a_1, \ldots, a_K\}$. A non-empty subset of $\mathbb{A}$ is called *itemset*.

A *boolean formula* $S : \Omega \to \{0, 1\}$ is a feature function mapping a binary vector to a binary value. We are interested in two particular boolean formulae: Assume that we are given an itemset $B = \{a_{i_1}, \ldots, a_{i_L}\}$. We define a *conjunction function* S_B to be

$$S_B(\omega) = \omega_{i_1} \wedge \omega_{i_2} \wedge \cdots \wedge \omega_{i_K},$$

that is, S_B results 1 if and only if all the variables corresponding the itemset B are on. Given a data set D the frequency $S_B(D)$ is called the frequency of the itemset B. Conjunction functions are popular and there are a lot of studies in the literature concerning finding itemsets that have large frequency (see e.g., Agrawal et al., 1993; Hand et al., 2001). We also define a *parity function* T_B to be

$$T_B(\omega) = \omega_{i_1} \oplus \omega_{i_2} \oplus \cdots \oplus \omega_{i_K},$$

where $\oplus$ is the binary operator XOR. The function T_B results 1 if and only if the number of active variables included in B are odd.

A collection $\mathcal{F}$ of itemsets is said to be *antimonotonic* or *downwardly closed* if each non-empty subset of an itemset included in $\mathcal{F}$ is also included in $\mathcal{F}$. Given a collection of itemsets $\mathcal{F} = \{B_1, \ldots, B_N\}$ we extend our definition of the conjunction function by setting $S_{\mathcal{F}} = [S_{B_1}, \ldots, S_{B_N}]^T$. We also define $T_{\mathcal{F}} = [T_{B_1}, \ldots, T_{B_N}]^T$.

Assume that we are given an antimonotonic family $\mathcal{F}$ of itemsets. We can show that there is an invertible matrix A such that $T_{\mathcal{F}} = AS_{\mathcal{F}}$. In other words, we can get the parity function $T_{\mathcal{F}}$ from the conjunction function $T_{\mathcal{F}}$ by an invertible linear transformation. Corollary 6 now implies that

$$d_{CM}(D_1, D_2 \mid S_{\mathcal{F}}) = d_{CM}(D_1, D_2 \mid T_{\mathcal{F}}), \tag{3}$$

for any D_1 and D_2. The following lemma shows that the covariance matrix $\mathrm{Cov}[T_{\mathcal{F}}]$ of the parity function is very simple.

Lemma 8 *Let $T_{\mathcal{F}}$ be a parity function for a family of itemsets $\mathcal{F}$, then* $\mathrm{Cov}\left[T_{\mathcal{F}}\right] = 0.5I$*, that is, the covariance matrix is a diagonal matrix having* 0.5 *at the diagonal.*

Theorem 1, Lemma 8, and Eq. 3 imply that

$$d_{CM}\left(D_1, D_2 \mid S_{\mathcal{F}}\right) = \sqrt{2}\left\|\theta_1 - \theta_2\right\|_2, \tag{4}$$

where $\theta_1 = T_{\mathcal{F}}\left(D_1\right)$ and $\theta_2 = T_{\mathcal{F}}\left(D_2\right)$. This identity says that the CM distance can be calculated in $O(N)$ time (assuming that we know the frequencies θ_1 and θ_2). This is better than $O(N^3)$ time implied by Theorem 1.

Example 3 *Let $I = \left\{\left\{a_j\right\} \mid j = 1 \ldots K\right\}$ be a family of itemsets having only one item. Note that $T_{\left\{a_j\right\}} = S_{\left\{a_j\right\}}$. Eq. 4 implies that*

$$d_{CM}\left(D_1, D_2 \mid S_I\right) = \sqrt{2}\left\|\theta_1 - \theta_2\right\|_2,$$

where θ_1 and θ_2 consists of the marginal frequencies of each a_j calculated from D_1 and D_2, respectively. In this case the CM distance is simply the L_2 distance between the marginal frequencies of the individual attributes. The frequencies θ_1 and θ_2 resemble term frequencies (TF) used in text mining (see e.g., Baldi et al., 2003).

Example 4 *We consider now a case with a larger set of features. Our motivation for this is that using only the feature functions S_I is sometimes inadequate. For example, consider data sets with two items having the same individual frequencies but different correlations. In this case the data sets may look very different but according to our distance they are equal.*

Let $\mathcal{C} = I \cup \left\{a_j a_k \mid j,k = 1 \ldots K, j < k\right\}$ be a family of itemsets such that each set contains at most two items. The corresponding frequencies contain the individual means and the pairwise correlation for all items. Let $S_{a_j a_k}$ be the conjunction function for the itemset $a_j a_k$. Let $\gamma_{jk} = S_{a_j a_k}\left(D_1\right) - S_{a_j a_k}\left(D_2\right)$ be the difference between the correlation frequencies. Also, let $\gamma_j = S_{a_j}\left(D_1\right) - S_{a_j}\left(D_2\right)$. Since

$$T_{a_j a_k} = S_{a_j} + S_{a_k} - 2S_{a_j a_k}$$

it follows from Eq. 4 that

$$d_{CM}\left(D_1, D_2 \mid S_C\right)^2 = 2\sum_{j<k}\left(\gamma_j + \gamma_k - 2\gamma_{jk}\right)^2 + 2\sum_{j=1}^{K}\gamma_j^2. \tag{5}$$

3.2 Characterisation of the CM Distance for Itemsets

The identity given in Eq. 4 is somewhat surprising and seems less intuitive. A question arises: why this distance is more natural than some other, say, a simple L_2 distance between the itemset frequencies. Certainly, parity functions are less intuitive than conjunction functions. One answer is that the parity frequencies are decorrelated version of the traditional itemset frequencies.

However, we can clarify this situation from another point of view: Let $\mathcal{A}$ be the set of all itemsets. Assume that we are given two data sets D_1 and D_2 and define *empirical distributions* p_1 and p_2 by setting

$$p_i(\omega) = \frac{\text{number of samples in } D_i \text{ equal to } \omega}{|D_i|}.$$

The constrained spaces of $S_{\mathcal{A}}$ are singular points containing only p_i, that is, $\mathcal{C}(S_{\mathcal{A}}, D_i) = \{p_i\}$. This implies that

$$d_{CM}(D_1, D_2 \mid S_{\mathcal{A}}) = \sqrt{2^K}\,\|p_1 - p_2\|_2 . \tag{6}$$

In other words, the CM distance is proportional to the L_2 distance between the empirical distributions. This identity seems very reasonable. At least, it is more natural than, say, taking L_2 distance between the traditional itemset frequencies.

The identity in Eq. 6 holds only when we use the features $S_{\mathcal{A}}$. However, we can prove that a distance of the Mahalanobis type satisfying the identity in Eq. 6 and some additional conditions is equal to the CM distance. Let us explain this in more detail. We assume that we have a distance d having the form

$$d(D_1, D_2 \mid S_{\mathcal{F}})^2 = (\theta_1 - \theta_2)^T C(S_{\mathcal{F}})^{-1} (\theta_1 - \theta_2),$$

where $\theta_1 = S_{\mathcal{F}}(D_1)$ and $\theta_2 = S_{\mathcal{F}}(D_2)$ and $C(S_{\mathcal{F}})$ maps a conjunction function $S_{\mathcal{F}}$ to a symmetric $N \times N$ matrix. The distance d should satisfy the following mild assumptions.

1. Assume two antimonotonic families of itemsets $\mathcal{F}$ and $\mathcal{H}$ such that $\mathcal{F} \subset \mathcal{H}$. It follows that $d(\cdot,\cdot \mid S_{\mathcal{F}}) \leq d(\cdot,\cdot \mid S_{\mathcal{H}})$.

2. Adding extra dimensions (but not changing the features) does not change the distance.

The following theorem says that the assumptions and the identity in Eq. 6 are sufficient to prove that d is actually the CM distance.

Theorem 9 *Assume that a Mahalanobis distance d satisfies Assumptions 1 and 2. Assume also that there is a constant c_1 such that*

$$d(D_1, D_2 \mid S_{\mathcal{A}}) = c_1 \|p_1 - p_2\|_2 .$$

Then it follows that for any antimonotonic family $\mathcal{F}$ we have

$$d(D_1, D_2 \mid S_{\mathcal{F}}) = c_2 d_{CM}(D_1, D_2 \mid S_{\mathcal{F}}),$$

for some constant c_2.

4. The CM distance and Event Sequences

In the previous section we discussed about the CM distance between the binary data sets. We will use similar approach to define the CM distance between sequences.

An *event sequence* s is a finite sequence whose symbols belong to a finite alphabet Σ. We denote the length of the event sequence s by $|s|$, and by $s(i, j)$ we mean a subsequence starting from i and ending at j. The subsequence $s(i, j)$ is also known as *window*. A popular choice for statistics of event sequences are *episodes* (Hand et al., 2001). A *parallel episode* is represented by a subset of the alphabet Σ. A window of s satisfies a parallel episode if all the symbols given in the episode occur in the window. Assume that we are given an integer k. Let W be a collection of windows of s having the length k. A *frequency* of a parallel episode is the proportion of windows in W satisfying the episode. We should point out that this mapping destroys the exact ordering of the sequence. On the other hand, if some symbols occur often close to each other, then the episode consisting of these symbols will have a high frequency.

In order to apply the CM distance we will now describe how we can transform a sequence s to a binary data set. Assume that we are given a window length k. We transform a window of length k into a binary vector of length $|\Sigma|$ by setting 1 if the corresponding symbol occurs in the window, and 0 otherwise. Let D be the collection of these binary vectors. We have now transformed the sequence s to the binary data set D. Note that parallel episodes of s are represented by itemsets of D.

This transformation enables us to use the CM distance. Assume that we are given two sequences s_1 and s_2, a collection of parallel episodes $\mathcal{F}$, and a window length k. First, we transform the sequences into data sets D_1 and D_2. We set the CM distance between the sequences s_1 and s_2 to be $d_{CM}(D_1, D_2 \mid S_{\mathcal{F}})$.

5. Feature Selection

We will now discuss briefly about feature selection—a subject that we have taken for granted so far. The CM distance depends on a feature function S. How can we choose a good set of features?

Assume for simplicity that we are dealing with binary data sets. Eq. 6 tells us that if we use all itemsets, then the CM distance is L_2 distance between empirical distributions. However, to get a reliable empirical distribution we need an exponential number of data points. Hence we can use only some subset of itemsets as features. The first approach is to make an expert choice without seeing data. For example, we could decide that the feature function is S_I, the means of the individual attributes, or S_C, the means of individual attributes and the pairwise correlation.

The other approach is to infer a feature function from the data sets. At first glimpse this seems an application of feature selection. However, traditional feature selection fails: Let S_I be the feature function representing the means of the individual attributes and let $S_{\mathcal{A}}$ be the feature function containing all itemsets. Let ω be a binary vector. Note that if we know $S_I(\omega)$, then we can deduce $S_{\mathcal{A}}(\omega)$. This means that S_I is a *Markov blanket* (Pearl, 1988) for $S_{\mathcal{A}}$. Hence we cannot use the Markov blanket approach to select features. The essential part is that the traditional feature selection algorithms deal with the *individual* points. We try to select features for whole data sets.

Note that feature selection algorithms for singular points are based on training data, that is, we have data points divided into clusters. In other words, when we are making traditional feature selection we *know* which points are close and which are not. In order to make the same ideas work for data sets we need to have similar information, that is, we need to know which data sets are close to each other, and which are not. Such an information is rarely provided and hence we are forced to seek some other approach.

We suggest a simple approach for selecting itemsets by assuming that frequently occurring itemsets are interesting. Assume that we are given a collection of data sets D_i and a threshold σ. Let I be the itemsets of order one. We define $\mathcal{F}$ such that $B \in \mathcal{F}$ if and only if $B \in I$ or that B is a σ-frequent itemset for some D_i.

6. Related Work

In this section we discuss some existing methods for comparing data sets and compare the evaluation algorithms. The execution times are summarised in Table 1.

Distance	Time
CM distance (general case)	$O(NM + N^2\,\lvert\Omega\rvert + N^3)$
CM distance (known cov. matrix)	$O(NM + N^3)$
CM distance (binary case)	$O(NM + N)$
Set distances	$O(M^3)$
Kullback-Leibler	$O(NM + N\,\lvert\Omega\rvert)$
Fischer's Information	$O(NM + N^2\,\lvert D_2\rvert + N^3)$

Table 1: Comparison of the execution times of various distances. The number $M = |D_1| + |D_2|$ is the number of data points in total. The $O(NM)$ term refers to the time needed to evaluate the frequencies $S(D_1)$ and $S(D_2)$. Kullback-Leibler distance is solved using Iterative Scaling algorithm in which one round has N steps and one step is executed in $O(|\Omega|)$ time.

6.1 Set Distances

One approach to define a data set distance is to use some natural distance between single data points and apply some known set distance. Eiter and Mannila (1997) show that some data set distances defined in this way can be evaluated in cubic time. However, this is too slow for us since we may have a vast amount of data points. The other downsides are that these distances may not take into account the statistical nature of data which may lead into problems.

6.2 Edit Distances

We discuss in Section 4 of using the CM distance for event sequences. Traditionally, edit distances are used for comparing event sequences. The most famous edit distance is Levenshtein distance (Levenshtein, 1966). However, edit distances do not take into account the statistical nature of data. For example, assume that we have two sequences generated such that the events are sampled from the uniform distribution independently of the previous event (a zero-order Markov chain). In this case the CM distance is close to 0 whereas the edit distance may be large. Roughly put, the CM distance measures the dissimilarity between the statistical characteristics whereas the edit distances operate at the symbol level.

6.3 Minimum Discrimination Approach

There are many distances for distributions (see Baseville, 1989, for a nice review). From these distances the CM distance resembles the statistical tests involved with Minimum Discrimination Theorem (see Kullback, 1968; Csiszár, 1975). In this framework we are given a feature function S and two data sets D_1 and D_2. From the set of distributions $\mathcal{C}_+(S, D_i)$ we select a distribution maximising the entropy and denote it by p_i^{ME}. The distance itself is the Kullback-Leibler divergence between p_1^{ME} and p_2^{ME}. It has been empirically shown that p_i^{ME} represents well the distribution from which D_i is generated (see Mannila et al., 1999). The downsides are that this distance is not a metric (it is not even symmetric), and that the evaluation time of the distance is infeasible: Solving p_i^{ME} is **NP**-hard (Cooper, 1990). We can approximate the Kullback-Leibler distance by Fischer's

information, that is,

$$D\left(p_1^{ME} \| p_2^{ME}\right) \approx \frac{1}{2}\left(\theta_1 - \theta_2\right)^T \mathrm{Cov}^{-1}\left[S \mid p_2^{ME}\right]\left(\theta_1 - \theta_2\right),$$

where $\theta_i = S(D_i)$ and $\mathrm{Cov}\left[S \mid p_2^{ME}\right]$ is the covariance matrix of S taken with respect to p_2^{ME} (see Kullback, 1968). This resembles greatly the equation in Theorem 1. However, in this case the covariance matrix depends on data sets and thus generally this approximation is not a metric. In addition, we do not know p_2^{ME} and hence we cannot evaluate the covariance matrix. We can, however, estimate the covariance matrix from D_2, that is,

$$\mathrm{Cov}\left[S \mid p_2^{ME}\right] \approx \frac{1}{|D_2|}\sum_{\omega \in D_2} S(\omega)S(\omega)^T - \frac{1}{|D_2|^2}\left[\sum_{\omega \in D_2} S(\omega)\right]\left[\sum_{\omega \in D_2} S(\omega)^T\right].$$

The execution time of this operation is $O(N^2\,|D_2|)$.

7. Empirical Tests

In this section we describe our experiments with the CM distance. We begin by examining the effect of different feature functions. We continue studying the distance by applying clustering algorithms, and finally we represent some interpretations to the results.

In many experiments we use a base distance d_U defined as the L_2 distance between the itemset frequencies, that is,

$$d_U\left(D_1, D_2 \mid S\right) = \sqrt{2}\,\|\theta_1 - \theta_2\|_2, \tag{7}$$

where θ_i are the itemset frequencies $\theta_i = S(D_i)$. This type of distance was used in Hollmén et al. (2003). Note that $d_U\left(D_1, D_2 \mid ind\right) = d_{CM}\left(D_1, D_2 \mid ind\right)$, where *ind* is the feature set containing only individual means.

7.1 Real World Data Sets

We examined the CM distance with several real world data sets and several feature sets. We had 7 data sets: *Bible*, a collection of 73 books from the Bible,[1] *Addresses*, a collection of 55 inaugural addresses given by the presidents of the U.S.,[2] *Beatles*, a set of lyrics from 13 studio albums made by the Beatles, *20Newsgroups*, a collection of 20 newsgroups,[3] *TopGenres*, plot summaries for top rated movies of 8 different genres, and *TopDecades*, plot summaries for top rated movies from 8 different decades.[4] *20Newsgroups* contained (in that order) 3 religion groups, 3 of politics, 5 of computers, 4 of science, 4 recreational, and *misc.forsale*. *TopGenres* consisted (in that order) of *Action*, *Adventure*, *SciFi*, *Drama*, *Crime*, *Horror*, *Comedy*, and *Romance*. The decades for *TopDecades* were 1930–2000. Our final data set, *Abstract*, was composed of abstracts describing NSF awards from 1990–1999.[5]

1. The books were taken from `http://www.gutenberg.org/etext/8300` on July 20, 2005.
2. The addresses were taken from `http://www.bartleby.com/124/` on August 17, 2005.
3. The data set was taken from `http://www.ai.mit.edu/~jrennie/20Newsgroups/`, a site hosted by Jason Rennie, on June 8, 2001.
4. The movie data sets were taken from `http://www.imdb.com/Top/` on January 1, 2006.
5. The data set was taken from `http://kdd.ics.uci.edu/databases/nsfabs/nsfawards.data.html` on January 13, 2006.

Bible and *Addresses* were converted into binary data sets by taking subwindows of length 6 (see the discussion in Section 4). We reduced the number of attributes to 1000 by using the mutual information gain. *Beatles* was preprocessed differently: We transformed each song to its binary bag-of-words representation and selected 100 most informative words. In *20Newsgroups* a transaction was a binary bag-of-words representation of a single article. Similarly, In *TopGenres* and in *TopDecades* a transaction corresponded to a single plot summary. We reduced the number of attributes in these three data sets to 200 by using the mutual information gain. In *Abstract* a data set represented one year and a transaction was a bag-of-words representation of a single abstract. We reduced the dimension of *Abstract* to 1000.

7.2 The Effect of Different Feature Functions

We begin our experiments by studying how the CM distance (and the base distance) changes as we change features.

We used 3 different sets of features: *ind*, the independent means, *cov*, the independent means along with the pairwise correlation, and *freq*, a family of frequent itemsets obtained by using APRIORI (Agrawal et al., 1996). We adjusted the threshold so that *freq* contained $10K$ itemsets, where K is the number of attributes.

We plotted the CM distances and the base distances as functions of $d_{CM}(ind)$. The results are shown in Figure 2. Since the number of constraints varies, we normalised the distances by dividing them with $\sqrt{N}$, where N is the number of constraints. In addition, we computed the correlation of each pair of distances. These correlations are shown in Table 2.

	d_{CM} vs. d_{CM}			d_U vs. d_U			d_{CM} vs. d_U	
Data set	*cov* *ind*	*freq* *ind*	*freq* *cov*	*cov* *ind*	*freq* *ind*	*freq* *cov*	*cov* *cov*	*freq* *freq*
20Newsgroups	0.996	0.725	0.733	0.902	0.760	0.941	0.874	0.571
Addresses	1.000	0.897	0.897	0.974	0.927	0.982	0.974	0.743
Bible	1.000	0.895	0.895	0.978	0.946	0.989	0.978	0.802
Beatles	0.982	0.764	0.780	0.951	0.858	0.855	0.920	0.827
TopGenres	0.996	0.817	0.833	0.916	0.776	0.934	0.927	0.931
TopDecades	0.998	0.735	0.744	0.897	0.551	0.682	0.895	0.346
Abstract	1.000	0.985	0.985	0.996	0.993	0.995	0.996	0.994
Total	0.998	0.702	0.709	0.934	0.894	0.938	0.910	0.607

Table 2: Correlations for various pairs of distances. A column represents a pair of distances and a row represents a single data set. For example, the correlation between $d_{CM}(ind)$ and $d_{CM}(cov)$ in *20Newsgroups* is 0.996. The last row is the correlation obtained by using the distances from all data sets simultaneously. Scatterplots for the columns 1–2 and 4–5 are given in Fig. 2.

Our first observation from the results is that $d_{CM}(cov)$ resembles $d_{CM}(ind)$ whereas $d_{CM}(freq)$ produces somewhat different results.

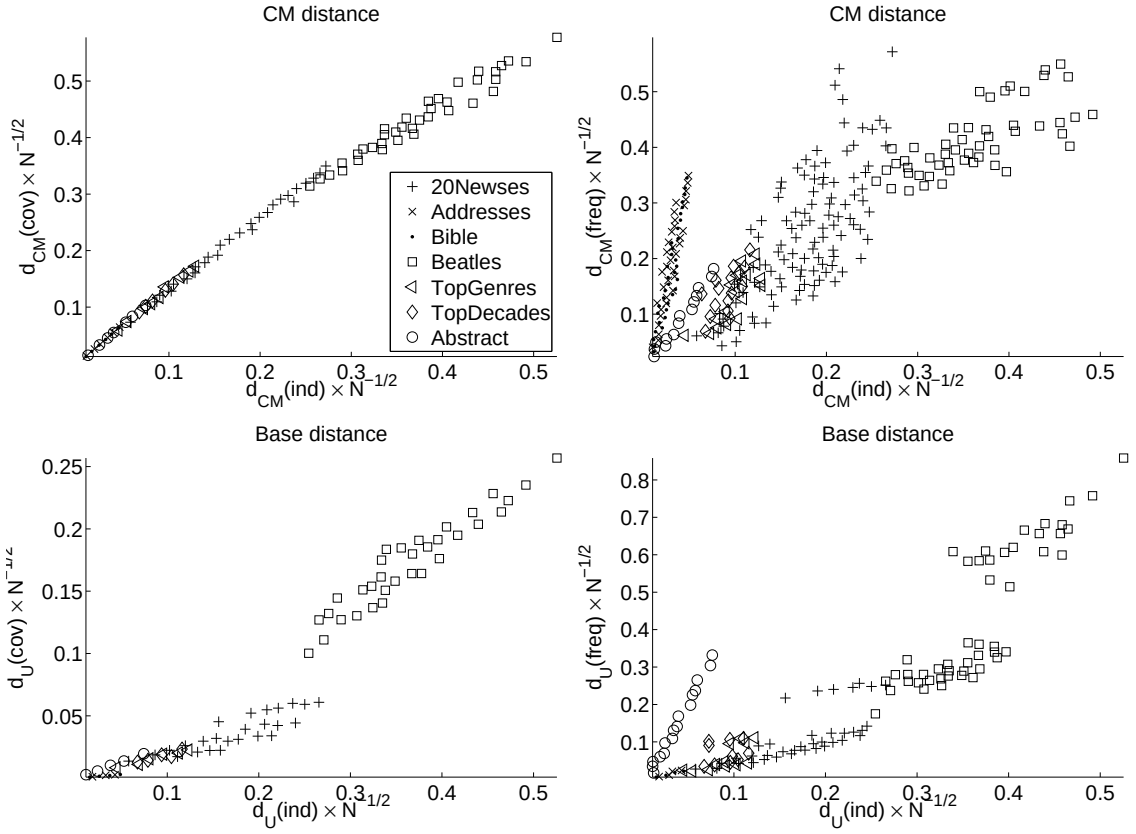


Figure 2: CM and base distances as functions of $d_{CM}(ind)$. A point represents a distance between two data sets. The upper two figures contain the CM distances while the lower two contain the base distance. The distances were normalised by dividing $\sqrt{N}$, where N is the number of constraints. The corresponding correlations are given in Table 2. Note that x-axis in the left (right) two figures are equivalent.

The correlations between $d_{CM}(cov)$ and $d_{CM}(ind)$ are stronger than the correlations between $d_U(cov)$ and $d_U(ind)$. This can be explained by examining Eq. 5 in Example 4. If the dimension is K, then the itemsets of size 1, according to Eq. 5, involve $\frac{1}{2}K(K-1)+K$ times in computing $d_{CM}(cov)$, whereas in computing $d_U(cov)$ they involve only K times. Hence, the itemsets of size 2 have smaller impact in $d_{CM}(cov)$ than in $d_U(cov)$.

On the other hand, the correlations between $d_{CM}(freq)$ and $d_{CM}(ind)$ are weaker than the correlations between $d_U(freq)$ and $d_U(ind)$, implying that the itemsets of higher order have stronger impact on the CM distance.

7.3 Clustering Experiments

In this section we continue our experiments by applying clustering algorithms to the distances. Our goal is to compare the clusterings obtained from the CM distance to those obtained from the base distance (given in Eq. 7).

We used 3 different clustering algorithms: a hierarchical clustering with complete linkage, a standard K-median, and a spectral algorithm by Ng et al. (2002). Since each algorithm takes a number of clusters as an input parameter, we varied the number of clusters between 3 and 5. We applied clustering algorithms to the distances $d_{CM}(cov)$, $d_{CM}(freq)$, $d_U(cov)$, and $d_U(freq)$, and compare the clusterings obtained from $d_{CM}(cov)$ against the clusterings obtained from $d_U(cov)$, and similarly compare the clusterings obtained from $d_{CM}(freq)$ against the clusterings obtained from $d_U(freq)$.

We measured the performance using 3 different clustering indices: a ratio r of the mean of the intra-cluster distances and the mean of the inter-cluster distances, Davies-Bouldin (DB) index (Davies and Bouldin, 1979), and Calinski-Harabasz (CH) index (Calinski and Harabasz, 1974).

The obtained results were studied in the following way: Given a data set and a performance index, we calculated the number of algorithms in which $d_{CM}(cov)$ outperformed $d_U(cov)$. The distances $d_{CM}(freq)$ and $d_U(freq)$ were handled similarly. The results are given in Table 3. We also calculated the number of data sets in which $d_{CM}(cov)$ outperformed $d_U(cov)$, given an algorithm and an index. These results are given in Table 4.

		$d_{CM}(cov)$ vs. $d_U(cov)$			$d_{CM}(freq)$ vs. $d_U(freq)$				
	Data set	*r*	*DB*	*CH*	*r*	*DB*	*CH*	Total	*P*
1.	*20Newsgroups*	0/9	2/9	7/9	8/9	5/9	9/9	31/54	0.22
2.	*Speeches*	9/9	6/9	3/9	9/9	9/9	9/9	**45**/**54**	0.00
3.	*Bible*	9/9	7/9	2/9	9/9	7/9	9/9	**43**/**54**	0.00
4.	*Beatles*	0/9	3/9	6/9	0/9	1/9	0/9	**10**/**54**	0.00
5.	*TopGenres*	0/9	4/9	5/9	0/9	1/9	0/9	**10**/**54**	0.00
6.	*TopDecades*	3/9	7/9	2/9	7/9	7/9	9/9	**35**/**54**	0.02
7.	*Abstract*	9/9	8/9	1/9	0/9	2/9	1/9	21/54	0.08
	Total	30/63	37/63	26/63	33/63	32/63	37/63	195/378	0.50
	P	0.61	0.13	0.13	0.61	0.80	0.13		

Table 3: Summary of the performance results of the CM distance versus the base distance. A single entry contains the number of clustering algorithm configurations (see Column 1 in Table 4) in which the CM distance was better than the base distance. The *P*-value is the standard Fisher's sign test.

We see from Table 3 that the performance of CM distance against the base distance depends on the data set. For example, the CM distance has tighter clusterings in *Speeches*, *Bible*, and *TopDecade* whereas the base distance outperforms the CM distance in *Beatles* and *TopGenre*.

Table 4 suggests that the overall performance of the CM distance is as good as the base distance. The CM distance obtains a better index 195 times out of 378. The statistical test suggests that this is a tie. The same observation is true if we compare the distances algorithmic-wise or index-wise.

7.4 Distance Matrices

In this section we will investigate the CM distance matrices for real-world data sets. Our goal is to demonstrate that the CM distance produces interesting and interpretable results.

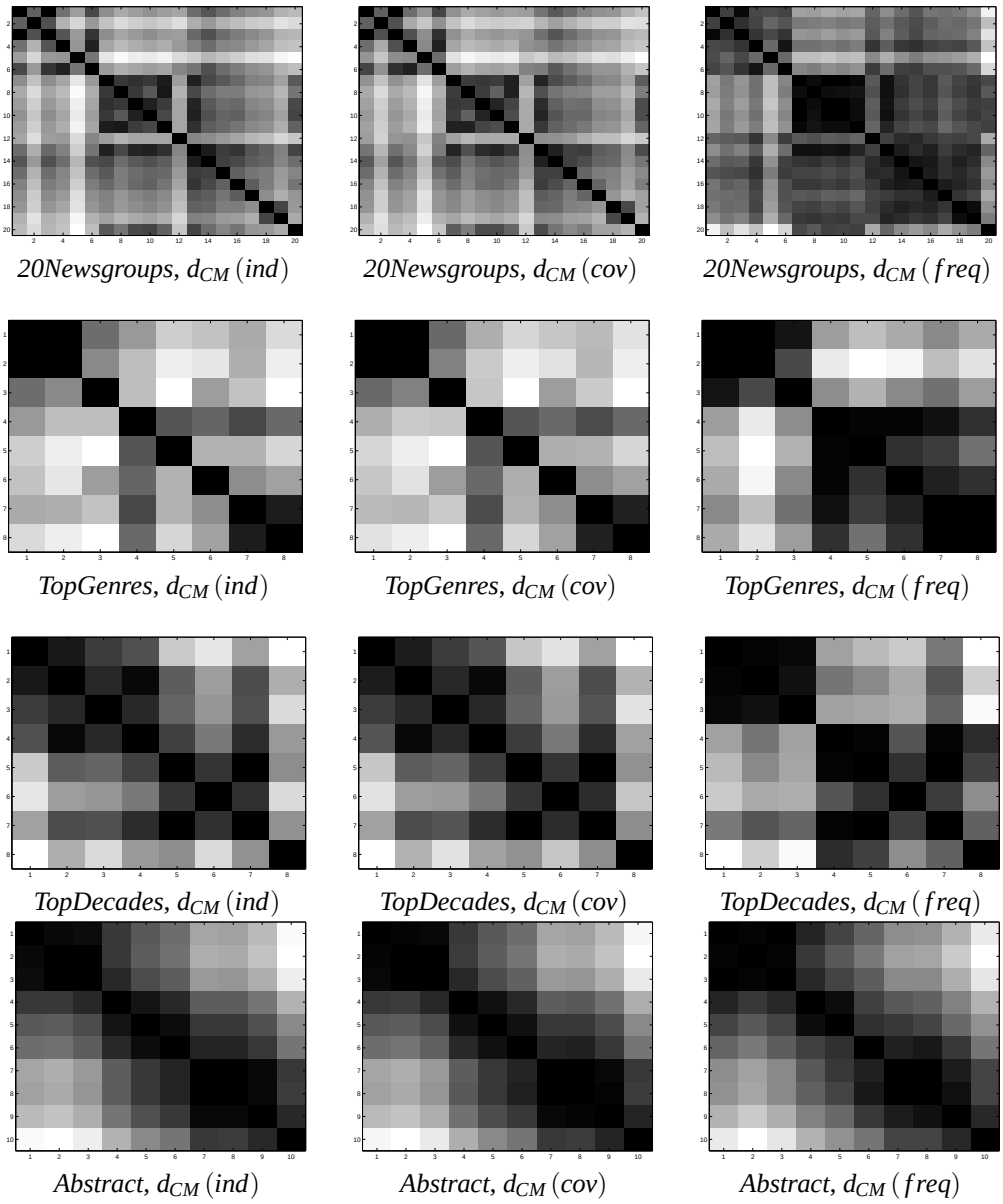
		$d_{CM}(cov)$ vs. $d_U(cov)$			$d_{CM}(freq)$ vs. $d_U(freq)$				
	Algorithm	*r*	*DB*	*CH*	*r*	*DB*	*CH*	Total	*P*
1.	K-MED(3)	4/7	2/7	5/7	4/7	4/7	4/7	23/42	0.44
2.	K-MED(4)	4/7	4/7	3/7	4/7	4/7	4/7	23/42	0.44
3.	K-MED(5)	4/7	4/7	3/7	4/7	4/7	4/7	23/42	0.44
4.	LINK(3)	3/7	4/7	3/7	2/7	3/7	4/7	19/42	0.44
5.	LINK(4)	3/7	4/7	3/7	4/7	3/7	4/7	21/42	0.88
6.	LINK(5)	3/7	3/7	4/7	4/7	2/7	4/7	20/42	0.64
7.	SPECT(3)	3/7	6/7	1/7	3/7	4/7	4/7	21/42	0.88
8.	SPECT(4)	3/7	4/7	3/7	4/7	4/7	4/7	22/42	0.64
9.	SPECT(5)	3/7	6/7	1/7	4/7	4/7	5/7	23/42	0.44
	Total	30/63	37/63	26/63	33/63	32/63	37/63	195/378	0.50
	P	0.61	0.13	0.13	0.61	0.80	0.13		

Table 4: Summary of the performance results of the CM distance versus the base distance. A single entry contains the number of data sets (see Column 1 in Table 3) in which the CM distance was better than the base distance. The *P*-value is the standard Fisher's sign test.

We calculated the distance matrices using the feature sets *ind*, *cov*, and *freq*. The matrices are given in Figures 4 and 3. In addition, we computed performance indices, a ratio of the mean of the intra-cluster distances and the mean of the inter-cluster distances, for various clusterings and compare these indices to the ones obtained from the base distances. The results are given in Table 5.

			cov		*freq*	
Data	Clustering	ind	d_{CM}	d_U	d_{CM}	d_U
Bible	Old Test. | New Test.	0.79	0.79	0.82	0.73	0.81
	Old Test. | Gospels | Epistles	0.79	0.79	0.81	0.73	0.81
Addresses	1–32 | 33–55	0.79	0.80	0.85	0.70	0.84
	1–11 | 12–22 | 23–33 | 34–44 | 45–55	0.83	0.83	0.87	0.75	0.87
Beatles	1,2,4–6 | 7–10,12–13 | 3 | 11	0.83	0.86	0.83	0.88	0.61
	1,2,4,12,13 | 5–10 | 3 | 11	0.84	0.85	0.84	0.89	0.63
20Newsgroups	Rel.,Pol. | Rest	0.76	0.77	0.67	0.56	0.62
	Rel.,Pol. | Comp., misc | Rest	0.78	0.78	0.79	0.53	0.79
TopGenres	Act.,Adv., SciFi | Rest	0.74	0.73	0.64	0.50	0.32
TopDecades	1930–1960 | 1970–2000	0.84	0.83	0.88	0.75	0.88
	1930–1950 | 1960–2000	0.88	0.88	0.98	0.57	1.06

Table 5: Statistics of various interpretable clusterings. The proportions are the averages of the intra-cluster distances divided by the averages of the inter-cluster distances. Hence small fractions imply tight clusterings.



20Newsgroups, $d_{CM}(ind)$ | *20Newsgroups*, $d_{CM}(cov)$ | *20Newsgroups*, $d_{CM}(freq)$

TopGenres, $d_{CM}(ind)$ | *TopGenres*, $d_{CM}(cov)$ | *TopGenres*, $d_{CM}(freq)$

TopDecades, $d_{CM}(ind)$ | *TopDecades*, $d_{CM}(cov)$ | *TopDecades*, $d_{CM}(freq)$

Abstract, $d_{CM}(ind)$ | *Abstract*, $d_{CM}(cov)$ | *Abstract*, $d_{CM}(freq)$

Figure 3: Distance matrices for *20Newsgroups*, *TopGenres*, *TopDecades*, and *Abstract*. In the first column the feature set *ind* contains the independent means, in the second feature set *cov* the pairwise correlation is added, and in the third column the feature set *freq* consists of $10K$ most frequent itemsets, where K is the number of attributes. Darker colours indicate smaller distances.

We should stress that standard edit distances would not work in these data setups. For example, the sequences have different lengths and hence Levenshtein distance cannot work.

The imperative observation is that, according to the CM distance, the data sets have structure. We can also provide some interpretations to the results: In *Bible* we see a cluster starting from the

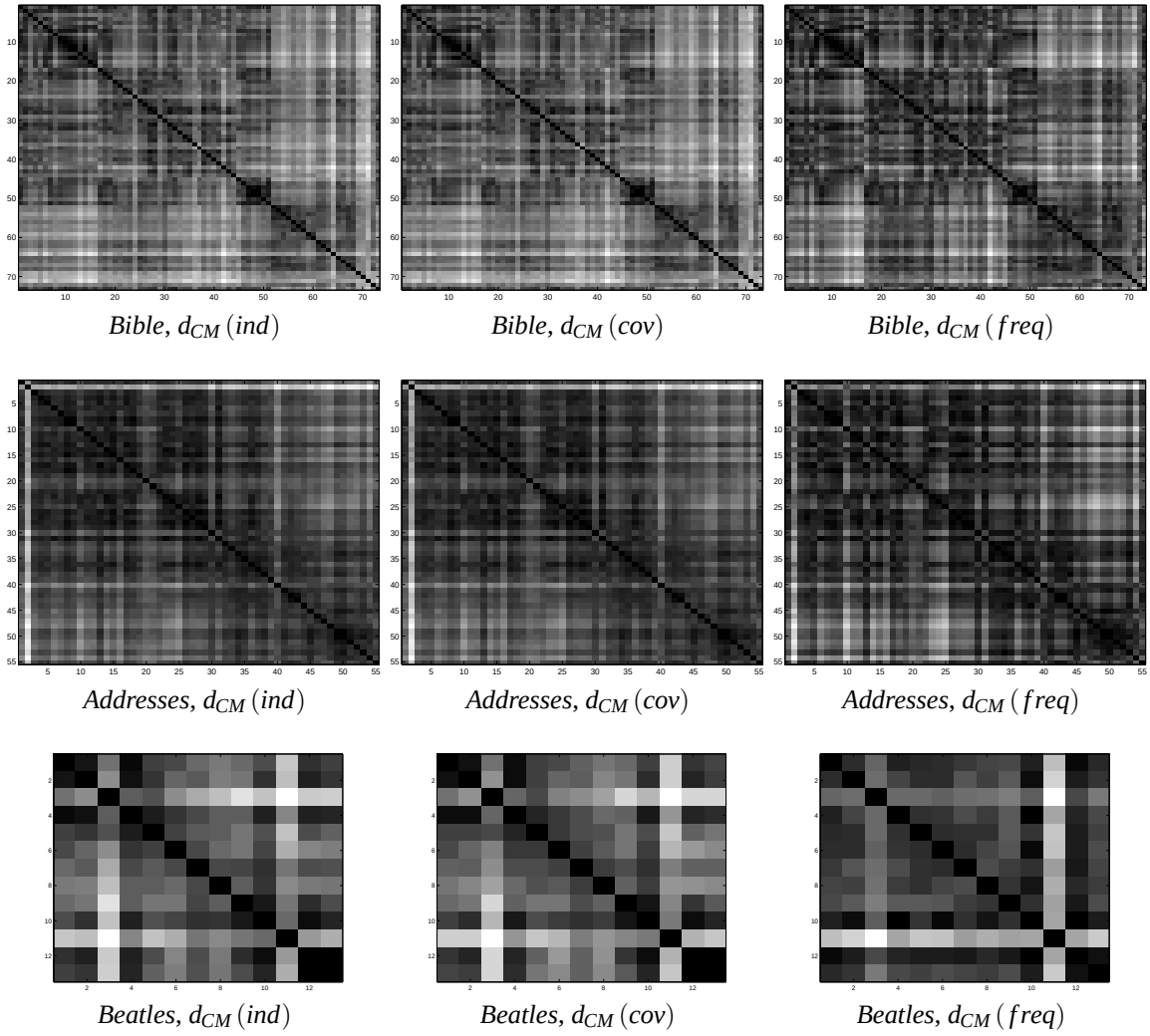


Figure 4: Distance matrices for *Bible*, *Addresses*, and *Beatles*. In the first column the feature set *ind* contains the independent means, in the second feature set *cov* the pairwise correlation is added, and in the third column the feature set *freq* consists of $10K$ most frequent itemsets, where K is the number of attributes. Darker colours indicate smaller distances.

46th book. The New Testament starts from the 47th book. An alternative clustering is obtained by separating the Epistles, starting from the 52th book, from the Gospels. In *Addresses* we some temporal dependence. Early speeches are different than the modern speeches. In *Beatles* we see that the early albums are linked together and the last two albums are also linked together. The third album, *Help!*, is peculiar. It is not linked to the early albums but rather to the later work. One explanation may be that, unlike the other early albums, this album does not contain any cover songs. In *20Newsgroups* the groups of politics and of religions are close to each other and so are the computer-related groups. The group *misc.forsale* is close to the computer-related groups. In *TopGenres Action* and *Adventure* are close to each other. Also *Comedy* and *Romance* are linked. In

TopDecades and in *Abstract* we see temporal behaviour. In Table 5 the CM distance outperforms the base distance, except for *Beatles* and *TopGenres*.

8. Conclusions and Discussion

Our task was to find a versatile distance that has nice statistical properties and that can be evaluated efficiently. The CM distance fulfils our goals. In theoretical sections we proved that this distance takes properly into account the correlation between features, and that it is the only (Mahalanobis) distance that does so. Even though our theoretical justifications are complex, the CM distance itself is rather simple. In its simplest form, it is the L_2 distance between the means of the individual attributes. On the other hand, the CM distance has a surprising form when the features are itemsets.

In general, the computation time of the CM distance depends of the size of sample space that can be exponentially large. Still, there are many types of feature functions for which the distance can be solved. For instance, if the features are itemsets, then the distance can be solved in polynomial time. In addition, if the itemsets form an antimonotonic family, then the distance can be solved in linear time.

In empirical tests the CM distance implied that the used data sets have structure, as expected. The performance of the CM distance compared to the base distance depended heavily on the data set. We also showed that the feature sets *ind* and *cov* produced almost equivalent distances, whereas using frequent itemsets produced very different distances.

Sophisticated feature selection methods were not compared in this paper. Instead, we either decided explicitly the set of features or deduced them using APRIORI. We argued that we cannot use the traditional approaches for selecting features of data sets, unless we are provided some additional information.

Acknowledgments

The author would like to thank Heikki Mannila and Kai Puolamäki for their extremely helpful comments.

Appendix A.

In this section we will prove the theorems given in this paper.

A.1 Proof of Theorem 1

To simplify the notation denote $S_0(x) = 1$, $\theta_1^* = [1, \theta_{11}, \ldots, \theta_{1N}]^T$ and $\theta_2^* = [1, \theta_{21}, \ldots, \theta_{2N}]^T$. The norm function restricted to the affine space has one minimum and it can be found using Lagrange multipliers. Thus we can express the vectors u_i in Eq. 2

$$u_{ij} = \lambda_i^T S(j),$$

where $j \in \Omega$ and λ_i is the column vector of length $N+1$ consisting of the corresponding Lagrange multipliers. The distance is equal to

$$\begin{aligned} d_{CM}(D_1, D_2 \mid S)^2 &= |\Omega| \, \|u_1 - u_2\|_2^2, \\ &= |\Omega| \sum_{j \in \Omega} (u_{1j} - u_{2j})(u_{1j} - u_{2j}), \\ &= |\Omega| \sum_{j \in \Omega} (u_{1j} - u_{2j}) \left(\lambda_1^T S(j) - \lambda_2^T S(j)\right), \\ &= |\Omega| \, (\lambda_1 - \lambda_2)^T \sum_{j \in \Omega} (u_{1j} - u_{2j}) S(j), \\ &= |\Omega| \, (\lambda_1 - \lambda_2)^T (\theta_1^* - \theta_2^*). \end{aligned}$$

Since $u_i \in \mathcal{C}(S, \theta_i)$, the multipliers λ_i can be solved from the equation

$$\theta_i^* = \sum_{j \in \Omega} S(j) u_{ij} = \sum_{j \in \Omega} S(j) \lambda_i^T S(j) = \left(\sum_{j \in \Omega} S(j) S(j)^T \right) \lambda_i,$$

that is, $\theta_i^* = A\lambda_i$, where A is an $(N+1) \times (N+1)$ matrix $A_{xy} = \sum_j S_x(j) S_y(j)$. It is straightforward to prove that the existence of $\mathrm{Cov}^{-1}[S]$ implies that A is also invertible. Let B be an $N \times N$ matrix formed from A^{-1} by removing the first row and the first column. We have

$$\begin{aligned} |\Omega| \, \|u_1 - u_2\|_2^2 &= |\Omega| \, (\theta_1^* - \theta_2^*)^T A^{-1} (\theta_1^* - \theta_2^*), \\ &= |\Omega| \, (\theta_1 - \theta_2)^T B (\theta_1 - \theta_2). \end{aligned}$$

The last equality is true since $\theta_{10}^* = \theta_{20}^*$.

We need to prove that $|\Omega| B = \mathrm{Cov}^{-1}[S]$. Let $[c; B]$ be the matrix obtained from A^{-1} by removing the first row. Let $\gamma = \mathrm{E}[S]$ taken with respect to the uniform distribution. Since the first column of A is equal to $|\Omega| [1, \gamma]$, it follows that $c = -B\gamma$. From the identity

$$c_x A_{(0,y)} + \sum_{z=1}^{N} B_{(x,z)} A_{(z,y)} = \delta_{xy}$$

we have

$$\sum_{z=1}^{N} B_{(x,z)} \left(A_{(z,y)} - A_{(0,y)} \gamma_z \right) = \sum_{z=1}^{N} |\Omega| B_{(x,z)} \left(|\Omega|^{-1} A_{(z,y)} - \gamma_y \gamma_z \right) = \delta_{xy}.$$

Since $|\Omega|^{-1} A_{(z,y)} - \gamma_z \gamma_y$ is equal to the (z,y) entry of $\mathrm{Cov}[S]$, the theorem follows.

A.2 Proofs of Theorems given in Section 2.2

Proof [Theorem 2] The covariance matrix $\mathrm{Cov}[S]$ in Theorem 1 depends only on S and is positive definite. Therefore, the CM distance is a Mahalanobis distance. ■

Proof [Theorem 3] Let $\theta_i = S(D_i)$ for $i = 1, 2, 3$. The frequencies for $D_1 \cup D_3$ and $D_2 \cup D_3$ are $(1-\varepsilon)\theta_1 + \varepsilon\theta_3$ and $(1-\varepsilon)\theta_2 + \varepsilon\theta_3$, respectively. The theorem follows from Theorem 1. ■

The following lemma proves Theorem 4.

Lemma 10 *Let $A : \mathbb{R}^N \to \mathbb{R}^M$ and define a function $T(\omega) = A(S(\omega))$. Let $\phi = T(D)$ and $\theta = S(D)$ be the frequencies for some data set D. Assume further that there is no two data sets D_1 and D_2 such that $S(D_1) = S(D_2)$ and $T(D_1) \neq T(D_2)$. Then $d_{CM}(D_1, D_2 \mid T) \leq d_{CM}(D_1, D_2 \mid S)$. The equality holds if for a fixed ϕ the frequency θ is unique.*

Before proving this lemma, let us explain why the uniqueness requirement is needed: Assume that the sample space Ω consists of two-dimensional binary vectors, that is,

$$\Omega = \{(0,0), (1,0), (0,1), (1,1)\}.$$

We set the features to be $S(\omega) = [\omega_1, \omega_2]^T$. Define a function $T(x) = [\omega_1, \omega_2, \omega_1\omega_2]^T = [S_1(\omega), S_2(\omega), S_1(\omega)S_2(\omega)]^T$. Note that uniqueness assumption is now violated. Without this assumption the lemma would imply that $d_{CM}(D_1, D_2 \mid T) \leq d_{CM}(D_1, D_2 \mid S)$ which is in general false.

Proof Let $\theta_1 = S(D_1)$ and $\phi_1 = T(D_1)$. Pick $u \in \mathcal{C}(S, \theta_1)$. The frequency of S taken with the respect to u is θ_1 and because of the assumption the corresponding frequency of T is ϕ_1. It follows that $\mathcal{C}(S, \theta_i) \subseteq \mathcal{C}(T, \phi_i)$. The theorem follows from the fact that the CM distance is the shortest distance between the affine spaces $\mathcal{C}(S, \theta_1)$ and $\mathcal{C}(S, \theta_2)$. ■

A.3 Proof of Theorem 7

It suffices to prove that the matrix $C(S)$ is proportional to the covariance matrix $\mathrm{Cov}[S]$. The notation $\delta(\omega_1 \mid \omega_2)$ used in the proof represents a feature function $\delta : \Omega \to \{0,1\}$ which returns 1 if $\omega_1 = \omega_2$ and 0 otherwise.

Before proving the theorem we should point one technical detail. In general, $C(S)$ may be singular, especially in Assumption 1. In our proof we will show that $C(S) \propto \mathrm{Cov}[S]$ and this does not require $C(S)$ to be invertible. However, if one wants to evaluate the distance d, then one must assume that $C(S)$ is invertible.

Fix indices i and j such that $i \neq j$. Let $T(\omega) = [S_i(\omega), S_j(\omega)]^T$. If follows from Assumption 1 that

$$C(T) = \begin{bmatrix} C_{ii}(S) & C_{ij}(S) \\ C_{ji}(S) & C_{jj}(S) \end{bmatrix}.$$

This implies that $C_{ij}(S)$ depends only on S_i and S_j. In other words, we can say $C_{ij}(S) = C_{ij}(S_i, S_j)$. Let $\rho : \{1, \ldots, N\} \to \{1, \ldots, N\}$ be some permutation function and define $U(x) = \left[S_{\rho(1)}(x), \ldots, S_{\rho(N)}(x)\right]^T$. Assumption 1 implies that

$$C_{\rho(i)\rho(j)}(S) = C_{ij}(U) = C_{ij}(U_i, U_j) = C_{ij}(S_{\rho(i)}, S_{\rho(j)}).$$

This is possible only if all non-diagonal entries of C have the same form or, in other words, $C_{ij}(S) = C_{ij}(S_i, S_j) = C(S_i, S_j)$. Similarly, the diagonal entry S_{ii} depends only on S_i and all the diagonal entries have the same form $C_{ii}(S) = C(S_i)$. To see the connection between $C(S_i)$ and $C(S_i, S_j)$ let $V(\omega) = [S_i(\omega), S_i(\omega)]^T$ and let $W(\omega) = [2S_i(\omega)]^T$. We can represent $W(\omega) = V_1(\omega) + V_2(\omega)$. Now Assumption 1 implies

$$\begin{aligned} 4C(S_i) = C(W) &= C(V_{11}) + 2C(V_{12}, V_{21}) + C(V_{22}), \\ &= 2C(S_i) + 2C(S_i, S_i) \end{aligned}$$

which shows that $C(S_i) = C(S_i, S_i)$. Fix S_j and note that Assumption 1 implies that $C(S_i, S_j)$ is a linear function of S_i. Thus C has a form

$$C(S_i, S_j) = \sum_{\omega \in \Omega} S_i(\omega) h(S_j, \omega)$$

for some specific map h. Let $\alpha \in \Omega$. Then $C(\delta(\omega \mid \alpha), S_j) = h(S_j, \alpha)$ is a linear function of S_j. Thus C has a form

$$C(S_i, S_j) = \sum_{\omega_1, \omega_2 \in \Omega} S_i(\omega_1) S_j(\omega_2) g(\omega_1, \omega_2)$$

for some specific g.

Let α, β, and γ be distinct points in Ω. An application of Assumption 2 shows that $g(\alpha, \beta) = C(\delta(\omega \mid \alpha), \delta(\omega \mid \beta)) = C(\delta(\omega \mid \alpha), \delta(\omega \mid \gamma)) = g(\alpha, \gamma)$. Thus g has a form $g(\omega_1, \omega_2) = a\delta(\omega_1 \mid \omega_2) + b$ for some constants a and b.

To complete the proof note that Assumption 1 implies that $C(S + b) = C(S)$ which in turns implies that $\sum_x g(\omega_1, \omega_2) = 0$ for all y. Thus $b = -a|\Omega|^{-1}$. This leads us to

$$\begin{aligned} C(S_i, S_j) &= \sum_{\omega_1, \omega_2 \in \Omega} S_i(\omega_1) S_j(\omega_2) \left(a\delta(\omega_1 \mid \omega_2) - a|\Omega|^{-1} \right), \\ &= a \sum_{\omega \in \Omega} S_i(\omega) S_j(\omega) - a \left(\sum_{\omega \in \Omega} S_i(\omega) \right) \left(\sum_{\omega \in \Omega} |\Omega|^{-1} S_j(\omega) \right), \\ &\propto \mathrm{E}[S_i S_j] - \mathrm{E}[S_i]\,\mathrm{E}[S_j], \end{aligned}$$

where the means are taken with respect to the uniform distribution. This identity proves the theorem.

A.4 Proof for Lemma 8

Let us prove that $\mathrm{Cov}[T_{\mathcal{F}}] = 0.5I$. Let A be an itemset. There are odd number of ones in A in exactly half of the transactions. Hence, $\mathrm{E}\left[T_A^2\right] = \mathrm{E}[T_A] = 0.5$. Let $B \neq A$ be an itemset. We wish to have $T_B(\omega) = T_A(\omega) = 1$. This means that ω must have odd number of ones in A and in B. Assume that the number of ones in $A \cap B$ is even. This means that $A - B$ and $B - A$ have odd number of ones. There is only a quarter of all the transactions that fulfil this condition. If $A \cap B$ is odd, then we must an even number of ones in $A - B$ and $B - A$. Again, there is only a quarter of all the transactions for which this holds. This implies that $\mathrm{E}[T_A T_B] = 0.25 = \mathrm{E}[T_A]\,\mathrm{E}[T_B]$. This proves that $\mathrm{Cov}[T_{\mathcal{F}}] = 0.5I$.

A.5 Proof of Theorem 9

Before proving this theorem let us rephrase it. First, note even though $d(\cdot, \cdot \mid \cdot)$ is defined only on the conjunction functions $S_{\mathcal{F}}$, we can operate with the parity function $T_{\mathcal{F}}$. As we stated before there is an invertible matrix A such that $T_{\mathcal{F}} = AS_{\mathcal{F}}$. We can write the distance as

$$d(D_1, D_2 \mid S_{\mathcal{F}})^2 = (A\theta_1 - A\theta_2)^T \left(A^{-1}\right)^T C(S_{\mathcal{F}})^{-1} A^{-1} (A\theta_1 - A\theta_2).$$

Thus we define $C(T_{\mathcal{F}}) = AC(S_{\mathcal{F}})A^T$. Note that the following lemma implies that the condition stated in Theorem 9 is equivalent to $C(T_{\mathcal{A}}) = cI$, for some constant c. Theorem 9 is equivalent to stating that $C(T_{\mathcal{F}}) = cI$.

The following lemma deals with some difficulties due the fact that the frequencies should arise from some valid distributions

Lemma 11 *Let $\mathcal{A}$ be the family of all itemsets. There exists $\varepsilon > 0$ such that for each real vector γ of length $2^K - 1$ that satisfies $\|\gamma\|_2 < \varepsilon$ there exist distributions p and q such that $\gamma = \mathrm{E}_p[T_\mathcal{A}] - \mathrm{E}_q[T_\mathcal{A}]$.*

Proof To ease the notation, add $T_0(x) = 1$ to $T_\mathcal{A}$ and denote the end result by T^*. We can consider T^* as a $2^K \times 2^K$ matrix, say A. Let p be a distribution and let u be the vector of length 2^K representing the distribution. Note that we have $Au = \mathrm{E}_p[T^*]$. We can show that A is invertible. Let U some $2^K - 1$ dimensional open ball of distributions. Since A is invertible, the set $V^* = \{Ax \mid x \in U\}$ is a $2^K - 1$ dimensional open ellipsoid. Define also V by removing the first coordinate from the vectors of V^*. Note that the first coordinate of elements of V^* is equal to 1. This implies that V is also a $2^K - 1$ dimensional open ellipsoid. Hence we can pick an open ball $N(\theta, \varepsilon) \subset V$. The lemma follows. ■

We are now ready to prove Theorem 9:

Abbreviate the matrix $C(T_\mathcal{F})$ by C. We will first prove that the diagonal entries of C are equal to c. Let $\mathcal{A}$ be the family of all itemsets. Select $G \in \mathcal{F}$ and define $\mathcal{R} = \{H \in \mathcal{F} \mid H \subseteq G\}$. As we stated above, $C(T_\mathcal{A}) = cI$ and Assumption 2 imply that $C(T_\mathcal{R}) = cI$. Assumption 1 implies that

$$d(\cdot,\cdot \mid S_\mathcal{R})^2 \leq d(\cdot,\cdot \mid S_\mathcal{F})^2 \leq d(\cdot,\cdot \mid S_\mathcal{A})^2. \tag{8}$$

Select ε corresponding to Lemma 11 and let $\gamma_\mathcal{A} = [0,\ldots,\varepsilon/2,\ldots,0]^T$, that is, $\gamma_\mathcal{A}$ is a vector whose entries are all 0 except the entry corresponding to G. Lemma 11 guarantees that there exist distributions p and q such that $d(p,q \mid S_\mathcal{A})^2 = c\|\gamma_\mathcal{A}\|_2^2$. Let $\gamma_\mathcal{F} = \mathrm{E}_p[T_\mathcal{F}] - \mathrm{E}_q[T_\mathcal{F}]$ and $\gamma_\mathcal{R} = \mathrm{E}_p[T_\mathcal{R}] - \mathrm{E}_q[T_\mathcal{R}]$. Note that $\gamma_\mathcal{R}$ and $\gamma_\mathcal{F}$ has the same form as γ_A. It follows from Eq. 8 that

$$c\varepsilon^2/4 \leq C_{G,G}\varepsilon^2/4 \leq c\varepsilon^2/4,$$

where $C_{G,G}$ is the diagonal entry of C corresponding to G. It follows that $C_{G,G} = c$.

To complete the proof we need to show that $C_{G,H} = 0$ for $G, H \in \mathcal{F}, G \neq H$. Assume that $C_{X,Y} \neq 0$ and let s be the sign of $C_{G,H}$. Apply Lemma 11 again and select $\gamma_\mathcal{A} = [0,\ldots,\varepsilon/4,0,\ldots,0,s\varepsilon/4,\ldots,0]^T$, that is, $\gamma_\mathcal{A}$ has $\varepsilon/4$ and $s\varepsilon/4$ in the entries corresponding to G and H, respectively, and 0 elsewhere. The right side of Eq. 8 implies that

$$2c\varepsilon^2/16 + 2|C_{G,H}|\varepsilon^2/16 \leq 2c\varepsilon^2/16$$

which is a contradiction and it follows that $C_{G,H} = 0$. This completes the theorem.

References

Rakesh Agrawal, Tomasz Imielinski, and Arun N. Swami. Mining association rules between sets of items in large databases. In Peter Buneman and Sushil Jajodia, editors, *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, pages 207–216, Washington, D.C., 26–28 1993.

Rakesh Agrawal, Heikki Mannila, Ramakrishnan Srikant, Hannu Toivonen, and Aino I. Verkamo. Fast discovery of association rules. In Usama M. Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, and Ramasamy Uthurusamy, editors, *Advances in Knowledge Discovery and Data Mining*, pages 307–328. AAAI Press/The MIT Press, 1996.

Pierre Baldi, Paolo Frasconi, and Padhraic Smyth. *Modeling the Internet and the Web*. John Wiley & Sons, 2003.

Michéle Baseville. Distance measures for signal processing and pattern recognition. *Signal Processing*, 18(4):349–369, 1989.

Toon Calders. *Axiomatization and Deduction Rules for the Frequency of Itemsets*. PhD thesis, University of Antwerp, Belgium, 2003.

Tadeusz Calinski and Jerzy Harabasz. A dendrite method for cluster analysis. *Communications in Statistics*, 3:1–27, 1974.

Gregory Cooper. The computational complexity of probabilistic inference using Bayesian belief networks. *Artificial Intelligence*, 42(2–3):393–405, Mar. 1990.

Imre Csiszár. I-divergence geometry of probability distributions and minimization problems. *The Annals of Probability*, 3(1):146–158, Feb. 1975.

David L. Davies and Donald W. Bouldin. A cluster separation measure. *IEEE Transactions of Pattern Analysis and Machine Intelligence*, 1(2):224–227, April 1979.

Thomas Eiter and Heikki Mannila. Distance measures for point sets and their computation. *Acta Informatica*, 34(2):109–133, 1997.

Theodore Hailperin. Best possible inequalities for the probability of a logical function of events. *The American Mathematical Monthly*, 72(4):343–359, Apr. 1965.

David Hand, Heikki Mannila, and Padhraic Smyth. *Principles of Data Mining*. The MIT Press, 2001.

Jaakko Hollmén, Jouni K Seppänen, and Heikki Mannila. Mixture models and frequent sets: Combining global and local methods for 0-1 data. In *Proceedings of the SIAM Conference on Data Mining (2003)*, 2003.

Solomon Kullback. *Information Theory and Statistics*. Dover Publications, Inc., 1968.

Vladimir I. Levenshtein. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady.*, 10(8):707–710, February 1966.

Heikki Mannila, Dmitry Pavlov, and Padhraic Smyth. Prediction with local patterns using cross-entropy. In *Knowledge Discovery and Data Mining*, pages 357–361, 1999.

Andrew Y. Ng, Michael I. Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems 14*, 2002.

Judea Pearl. *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1988.