# What is the dimension of your binary data?

Nikolaj Tatti Taneli Mielikäinen Aristides Gionis Heikki Mannila HIIT Basic Research Unit, Department of Computer Science University of Helsinki and Helsinki University of Technology

### Abstract

Many 0/1 datasets have a very large number of variables; on the other hand, they are sparse and the dependency structure of the variables is simpler than the number of variables would suggest. Defining the effective dimensionality of such a dataset is a nontrivial problem. We consider the problem of defining a robust measure of dimension for 0/1 datasets, and show that the basic idea of fractal dimension can be adapted for binary data. However, as such the fractal dimension is difficult to interpret. Hence we introduce the concept of normalized fractal dimension. For a dataset D, its normalized fractal dimension is the number of columns in a dataset D' with independent columns and having the same (unnormalized) fractal dimension as D. The normalized fractal dimension measures the degree of dependency structure of the data. We study the properties of the normalized fractal dimension and discuss its computation. We give empirical results on the normalized fractal dimension, comparing it against baseline measures such as PCA. We also study the relationship of the dimension of the whole dataset and the dimensions of subgroups formed by clustering. The results indicate interesting differences between and within datasets.

## 1 Introduction

Many 0/1-datasets occurring in data mining are on one hand complex, as they have a very high number of columns. On the other hand, the datasets can be simple, as they might be very sparse or have lots of structure. In this paper we consider the problem of defining a notion of *effective dimension* for a binary dataset. We study ways of defining a concept of dimension that would somehow capture the complexity or simplicity of the dataset. Such a notion of effective dimension can be used as a general score describing the complexity or simplicity of the dataset; Some potential applications of the intrinsic dimensionality of a dataset include model selection problems in data analysis; it can also be used in speeding up certain computations (see, e.g., [9]). For continuous data there are many ways of defining the dimension of a dataset. One approach is to use decomposition methods such as SVD, PCA, or NMF (nonnegative matrix factorization) [14, 19] and to count how many components are needed to express, say, 90% of the variance in the data. This number of components can be viewed as the number of effective dimensions in the data.

In the aforementioned methods it is assumed that the dataset is embedded into a higher-dimensional space by some (smooth) mapping. The other main approach is to use a different concept, that of fractal dimensions [3, 9, 15, 23]. Very roughly, the concept of fractal dimension is based on the idea of counting the number of observations in a ball of radius r and looking what the rate of growth of the number is as a function of r. If the number grows as  $r^k$ , then the dimensionality of the data can be considered to be k. Note that this approach does not provide any mapping that can be used for the dimension reduction. Such mapping does not even make sense because the dimension can be non-integral.

Applying these approaches to binary data is not easy. Many of the component methods, such as PCA and SVD are strongly based on the assumption that the data are realvalued. NMF looks for a matrix decomposition with nonnegative entries and hence is somewhat better suited for binary data. However, the factor matrices may have continuous values, which makes them difficult to interpret. The component techniques aimed at discrete data (such as multinomial PCA [6] or latent Dirichlet allocation (LDA) [4]) are possible alternatives, but interpreting the results is hard.

In this paper we explore the notion of effective dimension for binary datasets by using the basic ideas from fractal dimensions. Essentially, we consider the distribution of the pairwise distances between random points in the dataset. Denoting by Z this random variable, we study the ratio of  $\log \mathbb{P}(Z < r)$  and  $\log r$ , for different values of the r, and fit a straight line to this; the slope of the line is the correlation dimension of the dataset.

Interpreting the correlation dimension of discrete data turns out to be quite difficult too, because the values of the correlation dimension tend to very small. To relieve this problem, we normalize them by considering what would be the number of variables in a dataset with the same correlation dimension but with independent columns. This *normalized correlation dimension* is our main concept.

We study the behavior of the correlation dimension and the normalized correlation dimension, both theoretically and empirically. We give approximations for correlation dimension, in the case of independent variables, showing that it decreases when the data becomes more sparse. We also give theoretical evidence indicating that positive correlations between the variables lead to smaller correlation dimensions.

Our empirical results for generated data show that the normalized correlation dimension of a dataset with K independent variables is very close to K, irrespective of the sparsity of the attributes. We demonstrate that adding positive correlation decreases the dimension. For real datasets, we show that different datasets have quite different normalized correlation dimensions, and that the ratio of the number of variables to the normalized correlation dimension varies a lot. This indicates that the amount of structure in the datasets is highly variable. We also compare the normalized correlation dimension against the number of PCA components needed to explain 90% of the variance in the data, showing interesting differences among the datasets.

The rest of this paper is organized as follows. In Section 2 we define the correlation dimension for binary datasets. we analyze the correlation dimension in Section 3. The correlation dimension produces too small values and hence in Section 4 we provide means for scaling the dimension. In Section 5 we represent our tests with real world datasets. In Section 6 we review the related literature, and Section 7 is a short conclusion.

## 2 Correlation Dimension

There are several possible definitions of the fractal dimension of a subset of the Euclidean space; see, e.g., [3, 23] for a survey; the *Rényi dimensions* [23] form a fairly general family. The standard definitions of the fractal dimension are not directly applicable in the discrete case, but they can be modified to fit in.

The basic idea in the fractal dimensions is to study the distance between two random data points.

We focus on the correlation dimension. Consider a 0/1 dataset D with K variables. Denote by  $Z_D$  the random variable whose value is the  $L_1$  distance between two randomly chosen points from D; thus  $0 \le Z_D \le K$ . Informally, the correlation dimension is the slope of the line fitted in the log-log plot of  $(r, \mathbb{P}(Z_D < r))$ .

The more formal definition is more complex because the non-continuity of  $\mathbb{P}(Z_D < r)$  causes misbehavior in our later definitions. To remedy these problems we first define function  $f : \mathbb{N} \to \mathbb{R}$  to be  $f(r) = \mathbb{P}(Z_D < r)$ . We extend

this function to real numbers by linear interpolation. Thus f(r) is a continuous function being equal to  $\mathbb{P}(Z_D < r)$  when r is an integer.

Let  $0 \le r_1 < r_2 \le K$ . Then the different radii r and the function f for a given dataset D determine the point set

$$\mathcal{I}(D, r_1, r_2, N) = \{ (\log r, \log f(r)) \mid r = r_1 + \frac{i(r_2 - r_1)}{N}, i = 0 \dots N \}.$$

We usually omit the parameter N for the sake of brevity.

For example, assume that  $\mathbb{P}(Z_D \leq r) \propto r^d$  for some d, that is, the number of pairs of points within distance d grows as  $r^d$ . Then  $\mathcal{I}(D, r_1, r_2)$  is a straight line and the correlation dimension is equal to d.

**Definition 1.** *The* correlation dimension  $cd_R(D; r_1, r_2)$  for a binary dataset D and radii  $r_1$  and  $r_2$  is the slope of the least-squares linear approximation  $\mathcal{I}(Z, r_1, r_2)$ .

Assume that we are given  $\alpha_1$  and  $\alpha_2$  such that  $0 \le \alpha_1 < \alpha_2 \le 1$ . We define  $\operatorname{cd}_A(D; \alpha_1, \alpha_2)$  to be  $\operatorname{cd}_R(D; r_1, r_2)$ , where the radii  $r_i$  are set to be  $\max(f^{-1}(\alpha_i), 1)$ . The reason for truncating  $r_i$  is to avoid some misbehavior occurring with extremely sparse datasets.

That is,  $\mathcal{I}(D, r_1, r_2)$  is the set of points containing the logarithm of the radius r and the logarithm of the fraction of pairs of points from D that have  $L_1$  distance less than or equal to r. The correlation dimension is the slope of the line that fits these points best. The difference between  $\operatorname{cd}_R(D; r_1, r_2)$  and  $\operatorname{cd}_A(D; \alpha_1, \alpha_2)$  is that  $\operatorname{cd}_R$  is defined by using the absolute bounds  $r_1$  and  $r_2$  for the radius r, whereas  $cd_A$  uses the parameters  $\alpha_1$  and  $\alpha_2$  to specify the sizes of the tail of the distribution. For instance,  $cd_A(D; 1/4, 3/4)$  is the correlation dimension obtained by first computing the values  $r_1$  and  $r_2$  such that one quarter of the pairs of points have distance below  $r_1$ , and one quarter of the pairs have distance above  $r_2$ . The dimension is then obtained by computing N + 1 points  $(\log r, \log f(r))$  with  $r_1 \leq r \leq r_2$ , and by fitting a line to these points, in the least-squares sense.

How can we compute the correlation dimension of a binary dataset D? The probability  $\mathbb{P}(Z_D < r)$  can be computed

$$\frac{1}{|D|^2} \sum_{x \in D} \sum_{y \in D} I(|x - y| < r),$$

where I(|x - y| < r) is the indicator function having value 1 if |x - y| < r, and value 0 otherwise. Computing the values  $\mathbb{P}(Z_D < r)$  for all r can thus be done trivially in time  $O(N^2K)$ , where N is the number of points in D and K is the number of variables. A sparse matrix representation yields to a running time of O(NM), where M is the total number of 1's in the data: If point i has  $m_i$  1's, then  $\sum_{i} m_i = M$ , and computing the all pairwise distances takes time

$$\sum_{i=1}^{N} \sum_{j=1}^{N} (m_i + m_j) = 2NM.$$

If the number of points in a dataset is so large that quadratic computation time in the number of points is too slow, we can take a random subset  $D_s$  from D and estimate the probability  $\mathbb{P}(Z < r)$  by

$$\frac{1}{|D| |D_s|} \sum_{x \in D} \sum_{y \in D_s} I(|x - y| < r)$$

or by

$$\frac{1}{|D_s|^2} \sum_{x \in D_s} \sum_{y \in D_s} I(|x - y| < r)$$

## **3** Properties of binary correlation dimension

In this section we analyze the properties of the correlation dimension  $cd_R(D; r_1, r_2)$  for binary datasets. We show the following results under some simplifying assumptions. First, we prove that if the original data has independent columns, then the correlation dimension grows as the probabilities of the individual variables get closer to 0.5. Second, we show that in the independent case  $cd_A(D; \alpha, 1 - \alpha)$  grows as  $\sqrt{K}$ , where K is the number of attributes (columns) in the dataset. Third, we prove that if the variables are not independent, then the correlation dimension is smaller than for a dataset with the same margins but independent variables.

The analysis is not easy, and we need to make some simplifying assumptions. One complication is caused by the fact that the definition of  $\operatorname{cd}_R(D; r_1, r_2)$  involves computing the slope of a set of points. However, note that  $\mathcal{I}(D, r_1, r_2, 1)$  contains only two points, and hence we have

$$\operatorname{cd}_{R}(D; r_{1}, r_{2}, 1) = \frac{\log f(r_{2}) - \log f(r_{1})}{\log r_{2} - \log r_{1}}$$

Similarly, in the case of  $cd_A(D; \alpha_1, \alpha_2, 1)$  we have  $r_1$ and  $r_2$  such that  $\alpha_i = f(r_i)$ , and hence

$$\operatorname{cd}_{A}(D; r_{1}, r_{2}, 1) = \frac{\log \alpha_{2} - \log \alpha_{1}}{\log r_{2} - \log r_{1}}$$

Throughout this section we will assume that the parameter N in  $\mathcal{I}(D, r_1, r_2, N)$  is equal to 1.

**Proposition 2.** Assume that the dataset D has K independent variables, and that the probability of the variable i being 1 is  $p_i$  for each i, and let  $q_i = 2p_i(1 - p_i)$ . Assuming that K is large enough, we have

$$\operatorname{cd}_A(D; \alpha, 1 - \alpha) \approx C(\alpha) \frac{\sum_i q_i}{\sqrt{\sum_i q_i(1 - q_i)}},$$

where  $C(\alpha)$  is a constant depending only on  $\alpha$ . In particular, if all probabilities  $p_i$  are equal to p, then for q = 2p(1-p) we have

$$\operatorname{cd}_A(D; \alpha, 1 - \alpha) = C(\alpha) \sqrt{\frac{Kq}{1 - q}}.$$

The proposition indicates that the correlation dimension is maximized for variables as close to 0.5 as possible.

**Corollary 3.** Assume the dataset D has independent columns. The correlation dimension  $cd_A(D; \alpha, 1 - \alpha)$  is maximized if the variables have frequency 0.5.

The proposition also tells that for a dataset with independent identically distributed columns, the dimension grows as a square root of the number of columns.

Proof of Proposition 2. Recall that

$$\operatorname{cd}_{A}(D; \alpha, 1 - \alpha) = \frac{\log(1 - \alpha) - \log \alpha}{\log r_{2} - \log r_{1}},$$

where  $r_1$  and  $r_2$  are such that  $\alpha = f(r_1)$  and  $1 - \alpha = f(r_2)$ . The numerator is  $\log((1 - \alpha)/\alpha)$ . Assume that K is large enough that we can estimate f(r) by  $\mathbb{P}(Z_D < r)$ .

We next study the denominator  $\log r_2 - \log r_1$ . We have to analyze the distribution of the random variable  $Z_D$ , the  $L_1$  distance between two randomly chosen points from D. For simplicity, we denote  $Z_D$  by Z in the sequel. Let  $Z_i$ be the indicator variable having value 1 if two randomly chosen elements from D disagree in variable i; then  $Z = \sum_{i=1}^{K} Z_i$ .

Denote by  $q_i = E[Z_i]$  the probability that two randomly chosen points from D differ in coordinate i. If  $p_i$ is the probability that variable i in D has value 1, then  $q_i = 2p_i(1-p_i)$ , and it is easy to see that  $q_i \leq 1/2$ .

As  $Z = \sum_{i=1}^{K} Z_i$ , the variable Z has a binomial distribution. For simplicity we use the normal approximation: Z is distributed as  $N(\mu, \sigma)$ , where  $\mu = \sum_i q_i$  and  $\sigma^2 = \sum_i q_i(1-q_i)$ . If K is large enough, this approximation is accurate.

By the symmetry of the normal distribution there is a constant c such that  $r_1 = \mu - c\sigma$  and  $r_2 = \mu + c\sigma$ . Actually, c is the inverse of the cumulative distribution function of the normal distribution with parameters 0 and 1, i.e.,  $c = \Phi^{-1}(\alpha) = \sqrt{2} \operatorname{erf}^{-1}(2\alpha - 1)$  The denominator is

$$\log r_2 - \log r_1 = \log \frac{\mu + c\sigma}{\mu - c\sigma} = \log \sum_{n=0}^{\infty} \left(\frac{2c\sigma}{\mu}\right)^n.$$

Dropping all but the two first terms and using the series for logarithm we obtain that the numerate is

$$\log r_2 - \log r_1 \approx \frac{2c\sigma}{\mu}.$$

By setting

$$C(\alpha) = \frac{\log((1-\alpha)/\alpha)}{2c} = \frac{\log((1-\alpha)/\alpha)}{2\sqrt{2} \operatorname{erf}^{-1}(2\alpha-1)}$$

we have the desired result.

If  $\alpha = 1/4$ , then the constant  $C(\alpha)$  in Proposition 2 is about 0.815.

The correlation dimension has an interesting connection to the average distance in randomly picked point pairs.

**Proposition 4.** Assume that the dataset D has K independent variables, and that the probability of variable i being 1 is  $p_i$ . Let  $q_i = \sum_i 2p_i(1 - p_i)$ . Let  $\mu = \sum_i q_i$  be the average distance of two randomly picked points.

Assume that we are given two constants  $c_1$  and  $c_2$  such that  $0 \le c_1 < c_2 \le 1$ . Then we can approximate the correlation dimension as

$$\operatorname{cd}_R(D;c_1\mu,c_2\mu)\approx C(c_1,c_2)\mu,$$

where  $C(c_1, c_2)$  depends only of  $c_1$  and  $c_2$ .

Note that Proposition 4 gives an approximation for the quantity  $cd_R$ , while Proposition 2 is about  $cd_A$ ; this, however, is a superficial difference. More important is the fact that in Proposition 4 we look at the case where the bounds  $r_1$  and  $r_2$  are on the same side of the mean, whereas the bounds corresponding to  $\alpha$  and  $1 - \alpha$  from Proposition 2 are on the two sides of the mean. This implies that Proposition 4 gives a stronger bound: the dimension grows as a function of the mean  $\mu$ , not as a function of  $\mu/\sigma$ .

**Example 5.** Let D be a dataset with K dimensions, and consider the set D' obtained by copying each variable in D to N new variables. Then

$$\mathbb{P}\left(Z_D < r\right) = \mathbb{P}\left(Z_{D'} < Nr\right),$$

and hence

$$\operatorname{cd}_{R}(D;r_{1},r_{2})=\operatorname{cd}_{R}(D';Nr_{1},Nr_{2}).$$

Given a dataset D with K columns, we denote by ind (D) a random binary variable having K independent components such that the probability of *i*th component being 1 is equal to the probability of *i*th column of D being 1. Alternatively, ind (D) can be considered as a dataset obtained by permuting each column of D independently. We conjecture that the correlation dimension of D is always smaller than the correlation dimension of ind (D), given that the original variables are all positively correlated. **Conjecture 6.** Assume the marginal probability of all original variables are less than 0.5, and that all pairs of original variables are positively correlated. Then

$$\operatorname{cd}_A(D; \alpha, 1 - \alpha) \leq \operatorname{cd}_A(\operatorname{ind}(D); \alpha, 1 - \alpha)$$

*i.e., the correlation dimension of the original data is not larger than the correlation dimension of the data with each column permuted randomly.* 

Support for this conjecture is provided by the fact that the variance  $\operatorname{Var}[Z_D]$  of the variable  $Z_D$  can be shown to be no more than the variance  $\operatorname{Var}[Z_{\operatorname{ind}(D)}]$ ; this does not, however, suffice for the proof. The intuition behind the above conjecture is similar to what one observes in other types of definitions of dimension: if we randomly permute each column of a dataset, we expect to see the rank of the matrix to grow, and also explain an increase the number of PCA components needed to explain, say, 90% of the variance. In the experimental section we show the empirical evidence for Conjecture 6.

## 4 Normalized correlation dimension

The definition of correlation dimension (Definition 1) is based on the definition of correlation dimension for continuous data. We have argued that the definition has some simple intuitive properties: for a dataset with independent variables the dimension is smaller if the variables are sparse, and the dimension shrinks if we add structure to the data by making variables positively correlated.

However, the scale of the correlation dimension is not very intuitive: the dimension of a dataset with K independent variables is not K, although this would be the most natural value. The correlation dimension gives much smaller values and hence we need some kind of normalization.

We showed Section 3 that under some conditions independent variables maximize the correlation dimension. Informally, we define the *normalized correlation dimension* of a dataset D to be the number of variables that a dataset with independent variables must have in order to have the same correlation dimension as D does.

More formally, let  $\operatorname{ind}(H, p)$  be a dataset with H independent variables, each of which is equal to 1 with probability p. From Proposition 1 we have an explicit formula for  $\operatorname{cd}_A(\operatorname{ind}(H, p); \alpha, 1 - \alpha)$ : setting q = 2p(1 - p) we have

$$\operatorname{cd}_A(\operatorname{ind}(H,p);\alpha,1-\alpha) \approx C(\alpha)\sqrt{\frac{Hq}{1-q}}.$$

If the dataset would have the same marginal frequency, say *s*, for each variable, the normalized correlation dimension

of a dataset D could be defined to be the number H, such that

$$\operatorname{cd}_A(D; \alpha, 1 - \alpha)$$
 and  $\operatorname{cd}_A(\operatorname{ind}(H, s); \alpha, 1 - \alpha)$ 

are as close to each other as possible.

The problem with this way of normalizing the dimension is that it takes as the point of comparison a dataset where all the variables have the same marginal frequency. This is very far from being true in real data. Thus we modify the definition slightly.

We first find a value s such that

$$\operatorname{cd}_{A}\left(\operatorname{ind}\left(K,s\right);\alpha,1-\alpha\right)=\operatorname{cd}_{A}\left(\operatorname{ind}\left(D\right);\alpha,1-\alpha\right),$$

i.e., a summary of the marginal frequencies of the columns of D: s is the frequency that variables of an independent dataset should have in order that it has the same correlation dimension as D has when the columns of D have been randomized. We define the *normalized correlation dimension*, denoted by  $ncd_A(D; \alpha, 1 - \alpha)$ , to be an integer H such that

$$\operatorname{cd}_A(\operatorname{ind}(H,s);\alpha,1-\alpha) = \operatorname{cd}_A(D;\alpha,1-\alpha).$$

Proposition 2 implies the following statement.

**Proposition 7.** *Given a dataset* D *with* K *columns, the dimension*  $\operatorname{ncd}_A(D; \alpha, 1 - \alpha)$  *can be approximated by* 

$$\operatorname{ncd}_{A}(D; \alpha, 1 - \alpha) \approx \left(\frac{\operatorname{cd}_{A}(D; \alpha, 1 - \alpha)}{\operatorname{cd}_{A}(\operatorname{ind}(D); \alpha, 1 - \alpha)}\right)^{2} K$$

For examples, see the beginning of the next section.

## **5** Experimental results

In this section we describe our experimental results. We first describe some results on synthetic data, and then discuss real datasets and compare the normalized correlation dimension against PCA.

Unless otherwise mentioned, the dimension used in our experiments was  $\operatorname{cd}_A(D; \alpha_1, \alpha_2, N)$  such that  $\alpha_1 = 1/4$ ,  $\alpha_1 = 3/4$ , and N = 50.

#### 5.1 Synthetic datasets

In this section we provide empirical evidence to support the analysis in Sections 3 and 4. In the first experiment we generated 100 datasets with K independent columns and random margins  $p_i$ . For each dataset, the margins  $p_i$  were randomly picked by first picking  $p_{\text{max}}$  uniformly at random from [0, 1]. Then, the probability  $p_i$  was picked uniformly



Figure 1. Normalized correlation dimension for data having K independent dimensions for  $K \in \{50, 100, 150, 200\}$ .

from  $[0, p_{max}]$ ; this method results in datasets with different densities. The box plot in Figure 1 shows that the normalized dimension is very close to K, the number of variables in the data. This shows that for independent data the normalized correlation dimension is equal to the number of variables, and that the sparsity of the data does not influence the results.

Next we tested Proposition 2 with synthetic data. We generated 100 datasets having independent columns and random margins, generated as described above. Figure 2 shows the correlation dimension as a function of  $\mu/\sigma$ , where  $\mu = E[Z_D]$  and  $\sigma^2 = \text{Var}[Z_D]$ . The figure shows the behavior predicted by Proposition 2: the normalized fractal dimension is a linear function of  $\mu/\sigma$ , and the slope is very close to C(1/4) = 0.815.



Figure 2. Correlation dimension as a function of  $\mu/\sigma$  for data with independent columns (see Proposition 2). The *y*-axis is  $cd_A(D; 1/4, 3/4)$  and the *x*-axis is  $\mu/\sigma$ , where  $\mu = E[Z_D]$  and  $\sigma^2 = Var[Z_D]$ . The slope of the line is about C(1/4) = 0.815.

The theoretical section analyzes only the simplest form of the correlation dimension, that is, the case where N = 1. We tested how the dimension behaves for different N. In order to do that, we used generated datasets from the previous experiments and plotted  $cd_A(D; 1/4, 3/4, 50)$  against  $cd_A(D; 1/4, 3/4, 1)$ . We see from Figure 3 that the correlation dimension has little dependency of N.



Figure 3. Correlation dimension  $cd_A(D; 1/4, 3/4, 50)$  as a function of  $cd_A(D; 1/4, 3/4, 1)$ .

Next we verified the quality of the approximation of Proposition 4. We used the same data from the previous experiment. Figure 4 shows the correlation dimension against  $\mu = E[Z_D]$ , the average distance of two random points. From the figure we see that Proposition 4 is partly supported: the correlation dimension behaves as a linear function of  $\mu$ . However, the slope becomes more gentle as the number of columns increases.



Figure 4. Correlation dimension as a function of  $\mu$  for data with independent columns (see Proposition 4). The *y*-axis is  $cd_A(D; 1/4, 3/4)$  and the *x*-axis is  $\mu = E[Z_D]$ , the average distance between two random points.

Our fifth experiment tested how positive correlation affects the correlation dimension. Conjecture 6 predicts that positive correlation should decrease the correlation dimension. We tested this conjecture by creating random datasets D such that column i depends on column i - 1. Let  $X_i$  be variable number i in the generated dataset. We generated data by a Markov process between the variables:

$$\mathbb{P}(X_i = 1 \mid X_{i-1} = 0) = \mathbb{P}(X_i = 0 \mid X_{i-1} = 1) = t_i$$

and

$$\mathbb{P}(X_1 = 1) = \mathbb{P}(X_1 = 0) = 0.5$$

where  $X = [X_1, \ldots, X_k]$  is the random element of D.

The reversal probabilities  $t_i$  were randomly picked as follows: For each dataset we picked uniformly a random number  $t_{\text{max}}$  from the interval [0, 1]. We picked  $t_i$  uniformly from the interval  $[0, t_{\text{max}}]$ . Note that if the reversal probabilities were 0.5, then the dataset would have independent columns. Denoting  $Z = Z_D$ , we have

$$\mathbb{P}(Z_i = 1 \mid Z_{i-1} = 0) = \mathbb{P}(Z_i = 0 \mid Z_{i-1} = 1)$$
  
= 2t<sub>i</sub> (1 - t<sub>i</sub>).

A rough measure of the amount of correlation in the data is  $t = \sum 2t_i (1 - t_i)$ . Figure 5 shows the correlation dimension as a function of the quantity t. We see that the datasets with strong correlations tend to have small dimensions, as the theory predicts.



Figure 5. Correlation dimension as a function of t, a rough measure of correlation in a dataset. The *y*-axis is  $\operatorname{cd}_A(D; 1/4, 3/4)$  and the *x*-axis is the quantity  $t = \sum 2t_i (1 - t_i)$ , where  $t_i$  is the reversal probability between columns *i* and i - 1.

Next, we go back to the first experiment to see whether the normalized correlation dimension depends on the sparsity of data. Note that sparse datasets have small  $\mu =$  $E[Z_D]$ . Figure 6 shows the normalized correlation dimension as a function of  $\mu$  for the datasets used in Figure 1. We see that the normalized dimension does not depend of sparsity, as expected.

Finally, we tested Proposition 7 by plotting the normalized dimension as a function of  $\frac{K \operatorname{cd}_A(D)^2}{\operatorname{cd}_A(\operatorname{ind}(D))^2}$ . We used the generated datasets from the previous experiment and from our fifth experiment, as well. Results given in Figure 7 reveal that the approximation is good for the used datasets.

### 5.2 Real-world datasets

In this section we investigate how our dimensions behave with 9 real-world datasets: *Accidents, Courses, Kosarak*,



Figure 6. Normalized correlation dimension as a function of  $\mu$ , the average distance between two random points. The *x*-axis is  $\mu = E[Z_D]$  and the *y*-axis is  $ncd_A(D; 1/4, 3/4)$ .



Figure 7. Normalized correlation dimension as a function of  $K \operatorname{cd}_A(D)^2 / \operatorname{cd}_A(\operatorname{ind}(D))^2$ . The top figure contains datasets with independent columns and in the bottom figure adjacent columns of the datasets depend on each other.

*Paleo, POS, Retail, WebView-1, WebView-2* and *20 News-groups*. The basic information about the datasets is summarized in Table 1.

Table 1. The basic statistics of the datasets. The column K corresponds to the the number of columns and the column N to the number of rows. The last column is the density of 1's in percentages.

Data	K	N	# of 1s	Dens.
Accidents	469	340183	11500870	7.21
Courses	5021	2405	64743	0.54
Kosarak	41271	990002	8019015	0.02
Paleo	139	501	3537	5.08
POS	1657	515597	3367020	0.39
Retail	16470	88162	908576	0.06
WebView-1	497	59602	149639	0.51
WebView-2	3340	77512	358278	0.14

The datasets are as follows. 20 Newsgroups<sup>1</sup> is a collection of approximately 20 000 newsgroup documents across 20 different newsgroups [18]. Data in Accidents<sup>2</sup> were obtained from the Belgian "Analysis Form for Traffic Accidents" forms that is filled out by a police officer for each traffic accident that occurs with injured or deadly wounded casualties on a public road in Belgium. In total, 340 183 traffic accident records are included in the dataset [12]. The datasets  $POS^3$ ,  $WebView-1^4$  and  $WebView-2^5$  were contributed by Blue Martini Software as the KDD Cup 2000 data [16]. POS contains several years worth of point-ofsale data from a large electronics retailer. WebView-1 and WebView-2 contain several months worth of click-stream data from two e-commerce web sites. Kosarak<sup>6</sup> consists of (anonymized) click-stream data of a Hungarian on-line news portal. Retail<sup>7</sup> is a retail market basket data supplied by an anonymous Belgian retail supermarket store [5]. The dataset Paleo<sup>8</sup> contains information of species fossils found in specific paleontological sites in Europe [10]. Courses is a student-course dataset of courses completed by the Computer Science students of the University of Helsinki.

We began our experiments by computing the correlation dimension  $\operatorname{cd}_A(D; 1/4, 3/4)$  for each dataset. In order to do that, we needed to estimate the probabilities  $\mathbb{P}(Z_D < r)$ . Since some of the datasets had a very large amount of rows (see Table 1), we estimate the probabilities  $\mathbb{P}(Z_D < r)$  by

$$\frac{1}{|D| |D_s|} \sum_{x \in D} \sum_{y \in D_s} I(|x - y| < r), \qquad (1)$$

where I(|x - y| < r) is 1 if |x - y| < r, and 0 otherwise. The set  $D_s$  was a random subset of D containing 10 000 points. Since *Paleo* and *Courses* have small number of rows, no sampling is used and  $D_s$  was set to D for these datasets. The evaluation times are discussed in the end of the section.

We also computed  $cd_A$  (ind (D); 1/4, 3/4), the correlation dimension for the datasets with the same column margins but independent columns. Our goal was to use these numbers to provide empirical evidence for the theoretical sections. To calculate the dimensions we need to estimate the probabilities  $\mathbb{P}(Z_{ind(D)} < r)$ . The estimation was done by generating 10 000 points from the distribution of  $Z_{ind(D)}$ .

The dimensions  $\operatorname{cd}_A(D)$  and  $\operatorname{cd}_A(\operatorname{ind}(D))$  are given in Table 2. We see that the dimensions are very small. The

<sup>&</sup>lt;sup>1</sup>http://people.csail.mit.edu/jrennie/20Newsgroups/ <sup>2</sup>http://fimi.cs.helsinki.fi/data/accidents.dat.gz <sup>3</sup>http://www.ecn.purdue.edu/KDDCUP/data/BMS-POS.dat.gz <sup>4</sup>http://www.ecn.purdue.edu/KDDCUP/data/BMS-WebView-1.dat.gz <sup>5</sup>http://www.ecn.purdue.edu/KDDCUP/data/BMS-WebView-2.dat.gz

<sup>6</sup>http://fimi.cs.helsinki.fi/data/kosarak.dat.gz

<sup>&</sup>lt;sup>7</sup>http://fimi.cs.helsinki.fi/data/retail.dat.gz

<sup>&</sup>lt;sup>8</sup>NOW public release 030717 available from [10].

reason is that the datasets are quite sparse. We also observe that  $cd_A (ind (D))$  is always larger than  $cd_A (D)$ , which suggests that there is at least some structure in the datasets.

In addition, we used  $\operatorname{cd}_A(\operatorname{ind}(D))$  to verify Proposition 2. This was done by computing  $\mu/\sigma$ , where  $\mu = \operatorname{E}[Z_{\operatorname{ind}(D)}]$  and  $\sigma^2 = \operatorname{Var}[Z_{\operatorname{ind}(D)}]$ . We also computed

$$\hat{C}(1/4) = \operatorname{cd}_A(\operatorname{ind}(D); 1/4, 3/4) \frac{\sigma}{\mu}.$$

Note that Proposition 2 suggests that  $\hat{C}(1/4) \approx 0.8$ . Table 2 shows us that this is indeed the case.

Table 2. Correlation dimensions of the datasets. In the second column,  $D' = \operatorname{ind} (D)$ . The third column is the fraction  $\mu/\sigma$ , where  $\mu = \operatorname{E} [Z_{D'}]$  and  $\sigma^2 = \operatorname{Var} [Z_{D'}]$ . The fourth column is an estimate of the coefficient C(1/4) obtained by dividing  $\operatorname{cd}_A(D')$  with  $\mu/\sigma$ .

Data	$\mathrm{cd}_{A}\left(D\right)$	$\mathrm{cd}_{A}\left(D'\right)$	$\mu/\sigma$	$\hat{C}\left(1/4\right)$
Accidents	3.79	5.50	6.67	0.83
Courses	1.56	5.94	7.29	0.82
Kosarak	0.96	3.21	3.96	0.81
Paleo	1.21	3.20	3.87	0.83
POS	1.14	2.98	3.62	0.82
Retail	1.33	3.73	4.49	0.83
WebView-1	1.27	1.93	2.26	0.86
WebView-2	1.01	2.58	3.05	0.85

We continued our experiments by calculating the normalized correlation dimension  $ncd_A(D; 1/4, 3/4)$ . For this we computed the probability p such that

$$\operatorname{cd}_A(\operatorname{ind}(K, p); \alpha, 1 - \alpha) = \operatorname{cd}_A(\operatorname{ind}(D); \alpha, 1 - \alpha)$$

using binary search. Also, the normalized dimension itself was computed by using binary search. The normalized dimensions are given in Table 3.

Recall that the normalized correlation dimension of data D indicates how many variables a dataset D' with independent columns should have so that the distributional behavior of the pairwise distances between points would be about the same in D and D'. Thus we note, for example, that for the *Paleo* data the dimensionality is about 15, a fraction of 11% of the number of columns in the original data.

The last column in Table 3 is the estimate predicted by Proposition 7. Unlike with the synthetic datasets (see Section 5.1), the estimate is poor in some cases. A probable reason is that the examined datasets are extremely sparse, and hence the techniques used to obtain Proposition 7 are no longer accurate. This is supported by the observation that *Accident* has the best estimate and the largest density.

 Table 3. Normalized correlation dimensions of the datasets.

Data	K	$\mathrm{ncd}_A$	$\frac{\operatorname{ncd}_A(D)}{K}$	$\frac{K \mathrm{cd}_A(D)^2}{\mathrm{cd}_A(\mathrm{ind}(D))^2}$
Accidents	469	220	0.47	222.91
Courses	5021	304	0.06	344.24
Kosarak	41271	2378	0.06	3684.78
Paleo	139	15	0.11	19.90
POS	1657	181	0.11	242.91
Retail	16470	1791	0.11	2107.52
WebView-1	497	190	0.38	214.33
WebView-2	3340	359	0.11	512.97

We also tested the accuracy of Proposition 7 with 20 *Newsgroups* dataset<sup>9</sup>. In Figure 8 we plotted the normalized correlation dimension as a function of the estimate. We see that the approximation overestimates the dimension but the accuracy is better than in Table 3.



Figure 8. Normalized correlation dimension as a function of  $K \operatorname{cd}_A(D)^2 / \operatorname{cd}_A(\operatorname{ind}(D))^2$ . Each point represents one newsgroup in 20 *Newsgroups* dataset.

We will compare the normalized correlation dimensions against PCA in the next subsection.

Next we studied the running times of the computation of the correlation dimension. Computing the distance of two binary vectors can be done in O(M) time, where M is the number of 1's in the two vectors. Hence, estimating the probabilities using Equation 1 can be done in  $O(|D_s|L)$ , where L is the number of 1's in D. We need also to fit the slope to get the actual dimension, but the time needed for this operation is negligible compared to the time needed for estimating the probabilities. Note that in our setup, the size of  $D_s$  was fixed to  $10\,000$  (except for *Paleo* and *Courses*). Hence, the running time is proportional to the number of 1's in a dataset. The running times are given in Table 4.

<sup>&</sup>lt;sup>9</sup>The messages were converted into bag-of-words representations and 200 most informative variables were kept.

Table 4. The running times of the correlation dimension in seconds for various datasets. Time/# of 1's: time in milliseconds divided by the number of 1's in the data.

Data	# of 1's	Time	Time/# of 1's
Accidents	11500870	973	0.085
Courses	64743	9	0.141
Paleo	3537	0.1	0.039
Kosarak	8019015	793	0.099
POS	3367020	447	0.133
Retail	908576	103	0.113
WebView-1	149639	17	0.114
WebView-2	358278	40	0.112

## 5.3 Correlation Dimension vs. other methods

There are different approaches for measuring the structure of a dataset. In this section we study how the normalized dimension compares with other methods. Namely, we compared the normalized fractal dimension against the PCA approach and the average correlation coefficient.

We performed PCA to our datasets and computed the percentage of the variance explained by the M first PCA variables, where  $M = \text{ncd}_A(D)$ . Additionally, we calculated how many PCA components are needed to explain 90% of the variance. The results are given in Table 5. We observe that  $\text{ncd}_A(D)$  PCA components explain relatively large portion of the variance for *Accidents*, *POS*, and *WebView-1*, but explains less for *Paleo* and *WebView-2*.

Table 5. Normalized correlation dimensions versus PCA for various datasets. The second column is the percentage of variance explained by  $ncd_A(D)$  variables and the third column is the number of variables needed to explain 90% of the variance.

Data	$\operatorname{ncd}_{A}\left(D\right)$	PCA (%)	90% PCA Dim.
Accidents	220	99.83	81
Paleo	15	48.50	79
POS	181	84.48	246
WebView-1	190	87.89	208
WebView-2	359	59.73	1394

We next tested how robust the normalized correlation dimension is with respect to the selection of variables.

Let us first explain the setup of our study. Since especially PCA is time-consuming, we created subsets of the data by taking randomly 1000 transactions<sup>10</sup>. Let  $\pi_M(D)$  be the dataset obtained from D by selecting M columns at random. We used different numbers of variables M for different datasets. For each dataset D we took 50 random subsets  $\pi_M(D)$  and use them for our analysis.

We first performed PCA to each  $\pi_M(D)$  and computed the number of variables explaining 90% of the variance. We also computed the average correlation coefficient for each dataset. To be more precise, let  $c_{ij}$  be the correlation coefficient between columns *i* and *j* in  $\pi_M(D)$ . We define the average correlation coefficient to be

$$\operatorname{corr}(D, M) = \frac{1}{M(M-1)} \sum_{i < j} |c_{ij}|$$

Since structure in a dataset is seen as a small normalized fractal dimension, we expect that  $ncd_A(\pi_N(D))$  will correlate positively with the PCA approach and negatively with the average correlation coefficient corr  $(\pi_N(D))$ . The results are given in Figure 9.

We see from Figure 9 that there is a large degree of dependency between these methods: The normalized dimension correlates positively with PCA dimension and negatively with the average correlation, as expected. The most interesting behavior is observed in the *Paleo* dataset. We see that whereas PCA dimension says that *Paleo* should have relatively high dimension, the normalized dimension suggests a very small value. The average correlation agrees with the normalized dimension. Also, we know that *Paleo* has a very strong structure (by looking at the data) so this suggests that the PCA approach overestimates the intrinsic dimension for *Paleo*. This behavior can perhaps be partly explained also by considering the margins of the datasets. The margins of *Paleo* are relatively homogeneous whereas the margins of the rest datasets are skewed.

We computed the correlation coefficients between the normalized correlation dimension and the number of PCA components needed. We also computed the correlation for the normalized correlation dimension and the average correlations. These correlations coefficients were computed for each dataset D separately (recall that there were 50 random subsets for each D). Also, we calculated the correlations for the case when all the datasets were considered simultaneously. In addition, since *Paleo* behaved like an outlier, we computed the coefficients for the case where all datasets except *Paleo* were present. The results are given in Table 6 and they support the conclusions we draw from Figure 9.

## 5.4 Correlation dimension for subgroups generated by clustering

In this section we study how the correlation dimension of a dataset is related to the dimensions of its subsets. We

<sup>&</sup>lt;sup>10</sup>except for *Paleo* which had only 501 rows.



Figure 9. Normalized correlation dimension for random subsets of the data. The *y*-axis is the normalized correlation dimension (divided by the number of columns). In the upper panel the *x*-axis is number of PCA components needed to explain 90% of the variance, divided by the number of columns. In the lower panel the *x*-axis is the average correlation. A single point represent one random subset of the particular dataset. The number of variables *M* for the subset is shown in parentheses in the legend.

consider the case where the subsets are generated by clustering. The connection of the dimensions of the clusters and the dataset itself is not trivial.

We first studied the subject empirically using the *Paleo* dataset. There is a cluster structure in *Paleo*, and hence we used k-means to find 3 clusters and computed the dimensions for these clusters. The dimensions are given in Table 7.

We also conducted experiments with 20 Newsgroups. First, we calculated the normalized correlation dimension for each separate newsgroup. Then we created mixed datasets from 4 newsgroups, one of religious, one about computers, one recreational, and one science newsgroup. There were 240 such datasets in total. We computed the dimensions for each mix and compare them to the average

Table 6. Correlations between normalized dimension against PCA and average correlation. Each row represents 50 random subsets of the particular dataset (see Figure 9). The second last row contains the correlations obtained by using the subsets from all the datasets simultaneously. The last row is similar to the second last row except *Paleo* dataset was omitted.

	$ncd_A(D; 1/4, 3/4)$ vs.		
Data	PCA (90%)	$\operatorname{corr}\left(D\right)$	
Accident	0.44	-0.23	
Courses	-0.51	-0.01	
Kosarak	-0.21	-0.02	
Paleo	0.10	-0.31	
POS	0.27	-0.54	
Retail	-0.48	-0.18	
WebView-1	0.06	-0.33	
WebView-2	0.70	-0.49	
Total	0.09	-0.44	
Total without Paleo	0.60	0.13	

Table 7. Correlation dimension and normalized correlation dimension for *Paleo* data and its clusters. The clusters were obtained using the *k*-means algorithm.

Data	# of rows	$\mathrm{cd}_{A}\left(D\right)$	$\operatorname{ncd}_{A}\left(D\right)$
Cluster 1	51	2.56	37
Cluster 2	378	1.60	50
Cluster 3	72	2.53	46
Average	-	2.23	44.33
Whole data	501	1.21	15

dimensions of the newsgroups contained in the mixing. The scatterplot of the dimensions is given in Figure 10.

From the results we see that for our datasets the clusters tend to have higher dimensions than the whole dataset. We also see from Figure 10 that there is a positive correlation between the dimension of a cluster and the dimension of the whole dataset.

# 6 Related work

There has been a significant amount of work in defining the concept of dimensionality in datasets. Even though most of the methods can be adapted to the case of binary



Figure 10. Dimensions for cluster-structured data. Each point represents a mixture of 4 different newsgroups. The left figure contains the correlation dimension and the right figure contains the normalized correlation dimension. The *x*-axis is the average dimension of the components used in a mixture and the *y*-axis is the dimension of the mixture itself.

data, they are not specifically tailored for it. For instance, many methods assume real-valued numbers and they compute vectors/components that have negative or continuous values that are difficult to interpret. Such methods include, PCA, SVD, and non-negative matrix factorization (NMF) [14, 19]. Other methods such as multinomial PCA (mPCA) [6], and latent Dirichlet allocation (LDA) [4] assume specific probabilistic models of generating the data and the task is to discover latent components in the data rather than reasoning about the intrinsic dimensionality of the data. Methods for exact and approximate decompositions of binary matrices into binary matrices in Boolean semiring have also been proposed [11, 21, 22], but similarly to mPCA and LDA, they focus on finding components instead of the intrinsic dimensionality.

The concept of fractal dimension has found many applications in the database and data mining communities, such as, making nearest neighbor computations more efficient [24], speeding up feature selection methods [29], outlier detection [27], and performing clustering tasks based on the local dimensionality of the data points [13].

Many different notions of complexity of binary datasets have been proposed and used in various contexts, for instance VC-dimension [2], discrepancy [7], Kolmogorov complexity [20] and entropy-based concepts [8, 25]. In some of the above cases, such as Kolmogorov complexity and entropy methods, there is no direct interpretation of the measures as a notion of dimensionality of the data as they are measures of compressibility. VC-dimension measures the dimensionality of discrete data, but it is rather conservative as a binary dataset having VC-dimension *d* means that there are *d* columns such that the projection of the dataset on those coordinates results all possible bit vectors of length *d*. Hence, VC-dimension does not make any difference between datasets  $\{0,1\}^d$  and  $\{x \in \{0,1\}^K : \sum_{i=1}^K x_i \leq d\}$ , although there is a great difference when  $d \ll K$ . Furthermore, computing the VC-dimension of a given dataset is a difficult problem [26].

Related is also the work on random projections and dimensionality reductions, such as in [1], but this line of research has different goals than ours. Finally, methods such as multidimensional scaling (MDS) [17] and Isomap [28] focus on embedding the data (not necessarily binary) in low-dimensional spaces with small distortion, mainly for visualization purposes.

## 7 Concluding remarks

We have given a definition of the effective dimension of a binary dataset. The definition is based on ideas from fractal dimensions: We studied how the distribution of the distances between two random data points from the dataset behaves, and fit a slope to the log-log set of points. We defined the notion of normalized correlation dimension. It measures the number of dimensions of the appropriate density that a dataset with independent variables should have to have the same correlation dimension as the original dataset.

We studied the behavior of correlation dimension and normalized correlation dimension, both theoretically and empirically. Under certain simplifying assumptions, we were able to prove approximations for correlation dimension, and we verified these results using synthetic data.

Our empirical results for real data show that different datasets have clearly very different normalized correlation dimensions. In general, the normalized correlation dimension correlates with the number of PCA components that are needed to explain 90% of the variance in the data, but there are also intriguing differences.

Traditionally, dimension means the degrees of freedom in the dataset. One can consider a dataset embedded into a high-dimensional space by some (smooth) embedding map. Traditional methods such as PCA try to negate this embedding. Fractal dimensions, however, are based on different notion, the behavior of the volume of data as a function of neighborhoods. This means that the methods in this paper do not provide a mapping to a lower-dimensional space, and hence traditional applications, such as feature reduction, are not (directly) possible. However, our study shows that fractal dimensions have promising properties and we believe that these dimensions are important as such.

A fundamental difference between the normalized correlation dimension and PCA is the following. For a dataset with independent columns PCA has no effect and selects the columns that have the highest variance until some selected percentage of the variance is explained. Thus, the number of PCA components needed depends on the margins of the columns. On the other hand, the normalized correlation dimension is always equal to the number of variables for data with independent columns.

Obviously, several open problems remain. It would be interesting to have more general results about the theoretical behavior of the normalized correlation dimension. In the empirical side the study of the correlation dimensions of the data and its subsets seems to be a promising direction.

## References

- D. Achlioptas. Database-friendly random projections: Johnson-lindenstrauss with binary coins. *Journal of Computer and System Sciences*, 66(4):671–687, 2003.
- [2] M. Anthony and N. Biggs. Computational Learning Theory: An Introduction. Cambridge Tracts in Theoretical Computer Science. Cambridge University Press, 1997.
- [3] M. Barnsley. Fractals Everywhere. Academic Press, 1988.
- [4] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993– 1022, 2003.
- [5] T. Brijs, G. Swinnen, K. Vanhoof, and G. Wets. Using association rules for product assortment decisions: A case study. In *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 15-18, 1999, San Diego, CA, USA*, pages 254– 260. ACM, 1999.
- [6] W. Buntine and S. Perttu. Is multinomial PCA multi-faceted clustering or dimensionality reduction? In C. Bishop and B. Frey, editors, *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics*, pages 300–307, 2003.
- [7] B. Chazelle. *The Discrepancy Method*. Cambridge University Press, 2000.
- [8] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley, 1991.
- [9] C. Faloutsos and I. Kamel. Beyond uniformity and independence: Analysis of R-trees using the concept of fractal dimension. In *PODS*, pages 4–13. ACM Press, 1994.
- [10] M. Fortelius. Neogene of the old world database of fossil mammals (NOW). University of Helsinki, http://www.helsinki.fi/science/now/, 2005.
- [11] F. Geerts, B. Goethals, and T. Mielikäinen. Tiling databases. In E. Suzuki and S. Arikawa, editors, *Discovery Science*, volume 3245 of *Lecture Notes in Computer Science*, pages 278–289. Springer, 2004.
- [12] K. Geurts, G. Wets, T. Brijs, and K. Vanhoof. Profiling high frequency accident locations using association rules. In *Proceedings of the 82nd Annual Transportation Research Board, Washington DC. (USA), January 12-16*, 2003.
- [13] A. Gionis, A. Hinneburg, S. Papadimitriou, and P. Tsaparas. Dimension induced clustering. In R. Grossman, R. Bayardo, and K. P. Bennett, editors, *KDD*, pages 51–60. ACM, 2005.
- [14] I. Jolliffe. Principal Component Analysis. Springer Series in Statistics. Springer, 2nd edition, 2002.
- [15] B. Kégl. Intrinsic dimension estimation using packing numbers. In S. T. S. Becker and K. Obermayer, editors, Advances in Neural Information Processing Systems 15, pages 681–688, 2003.

- [16] R. Kohavi, C. Brodley, B. Frasca, L. Mason, and Z. Zheng. KDD-Cup 2000 organizers' report: Peeling the onion. *SIGKDD Explorations*, 2(2):86–98, 2000.
- [17] J. B. Kruskal. Multidimensional scaling by optimizing goodness of t to a nonmetric hypothesis. *Psychometrica*, 29:1–26, 1964.
- [18] K. Lang. Newsweeder: Learning to filter netnews. In Proceedings of the Twelfth International Conference on Machine Learning, pages 331–339, 1995.
- [19] D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. In T. K. Leen, T. G. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems 13*, pages 556–562, 2001.
- [20] M. Li and P. Vitányi. An Introduction to Kolmogorov Complexity and Its Applications. Texts in Computer Science. Springer-Verlag, 3rd edition, 1997.
- [21] P. Miettinen, T. Mielikäinen, A. Gionis, G. Das, and H. Mannila. The discrete basis problem. In J. Fürnkranz, T. Scheffer, and M. Spiliopoulou, editors, *Knowledge Discovery in Databases: PKDD 2006 – 10th European Conference on Principles and Practice of Knowledge Discovery in Databases, Berlin, Germany, Croatia, September 18–22, 2006, Proceedings*, Lecture Notes in Computer Science. Springer, 2006.
- [22] S. D. Monson, N. J. Pullman, and R. Rees. A survey of clique and biclique coverings and factorizations of (0, 1)matrices. *Bulletin of the ICA*, 14:17–86, 1995.
- [23] E. Ott. *Chaos in Dynamical Systems*. Cambridge University Press, 1997.
- [24] B.-U. Pagel, F. Korn, and C. Faloutsos. Deflating the dimensionality curse using multiple fractal dimensions. In *ICDE*, pages 589–598. IEEE Computer Society, 2000.
- [25] P. Palmerini, S. Orlando, and R. Perego. Statistical properties of transactional databases. In H. Haddad, A. Omicini, R. L. Wainwright, and L. M. Liebrock, editors, *SAC*, pages 515–519. ACM, 2004.
- [26] C. H. Papadimitriou and M. Yannakakis. On limited nondeterminism and the complexity of the V-C dimension. *Journal of Computer and System Sciences*, 53(2):161–170, 1996.
- [27] S. Papadimitriou, H. Kitagawa, P. B. Gibbons, and C. Faloutsos. LOCI: fast outlier detection using the local correlation integral. In U. Dayal, K. Ramamritham, and T. M. Vijayaraman, editors, *ICDE*, pages 315–326. IEEE Computer Society, 2003.
- [28] J. B. Tenenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.
- [29] C. Traina Jr., A. J. M. Traina, L. Wu, and C. Faloutsos. Fast feature selection using fractal dimension. In K. Becker, A. A. de Souza, D. Y. de Souza Fernandes, and D. C. F. Batista, editors, *SBBD*, pages 158–171. CEFET-PB, 2000.