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
Helsinki, Finland

Abstract

Many 0/1 datasets have a very large number of variables; however, 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 counts the number of independent columns needed to achieve the unnormalized fractal dimension of 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 PCA.

1. Introduction

Many binary datasets occurring in data mining applications are, on one hand, complex as they have a very large number of columns. On the other hand, some of those datasets could be potentially simple, as they are 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 straightforward. Many of the component methods, such as PCA and SVD are strongly based on the assumption that the data are real-valued. 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 $\log \mathbb{P}\left(Z < r\right)$ and $\log r$, for different values of the r, and fit a straight line to this; the slope of the line is known as the correlation dimension of the dataset.

Interpreting the correlation dimension of discrete data turns out to be a difficult task. To assist interpetation, we normalize the correlation dimension by considering what would be the number of variables in a certain random dataset with independent columns having the same correlation dimension. 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, irrespectively 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 and 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. The proofs are omitted due to the space limitations.

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 $0 \leq Z_D \leq K$. Informally, the correlation dimension is the slope of the line fitted in the log-log plot of $(r, \mathbb{P}(Z_D < r))$.

More formally, we first define the function $f : \mathbb{N} \rightarrow \mathbb{R}$ to be $f(r) = \mathbb{P}(Z_D < r)$. We extend this function to real numbers by linear interpolation.

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

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

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 $\mathrm{cd_R}(D; r_1, r_2)$ *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 α_1 and α_2 such that $0 \leq \alpha_1 < \alpha_2 \leq 1$. We define $\mathrm{cd_A}(D; \alpha_1, \alpha_2)$ to be $\mathrm{cd_R}(D; r_1, r_2)$, where the radii r_i are set to be $\max\left(f^{-1}(\alpha_i), 1\right)$. 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 $\mathrm{cd_R}(D; r_1, r_2)$ and $\mathrm{cd_A}(D; \alpha_1, \alpha_2)$ is that $\mathrm{cd_R}$ is defined by using the absolute bounds r_1 and r_2 for the radius r, whereas $\mathrm{cd_A}$ uses the parameters α_1 and α_2 to specify the sizes of the tail of the distribution. For instance, $\mathrm{cd_A}(D; 1/4, 3/4)$ 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 integers r can thus be done trivially in time $O(N^2 K)$, 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 $\mathbb{P}\left(Z < r\right)$ 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 $\mathrm{cd}_\mathrm{A}\left(D;\alpha_1,\alpha_2\right)$ 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 $\mathrm{cd}_\mathrm{A}\left(D;\alpha,1-\alpha\right)$ 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.

For the analysis we need to make some simplifying assumptions. One complication is caused by the fact that the definition of $\mathrm{cd}_\mathrm{R}\left(D;r_1,r_2\right)$ involves the slope of a set of points. However, note that $\mathcal{I}\left(D,r_1,r_2,1\right)$ contains only two points, and hence we have

$$\mathrm{cd}_\mathrm{R}\left(D;r_1,r_2,1\right) = \frac{\log f(r_2) - \log f(r_1)}{\log r_2 - \log r_1}.$$

Similarly, in the case of $\mathrm{cd}_\mathrm{A}\left(D;\alpha_1,\alpha_2,1\right)$ we have r_1 and r_2 such that $\alpha_i = f(r_i)$, and hence

$$\mathrm{cd}_\mathrm{A}\left(D;r_1,r_2,1\right) = \frac{\log \alpha_2 - \log \alpha_1}{\log r_2 - \log r_1}.$$

Throughout this section we assume that the parameter N in $\mathcal{I}\left(D,r_1,r_2,N\right)$ 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*

$$\mathrm{cd}_\mathrm{A}\left(D;\alpha,1-\alpha\right) \approx C(\alpha)\frac{\sum_i q_i}{\sqrt{\sum_i q_i(1-q_i)}},$$

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

$$\mathrm{cd}_\mathrm{A}\left(D;\alpha,1-\alpha\right) \approx C(\alpha)\sqrt{\frac{Kq}{1-q}}.$$

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 $\mathrm{cd}_\mathrm{A}\left(D;\alpha,1-\alpha\right)$ is maximized if the variables have frequency 0.5.*

Proposition 2 also tells that for a dataset with independent identically distributed columns, the dimension grows as a square root of the number of columns. If $\alpha = 1/4$, then the constant $C(\alpha)$ 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 $0 \leq c_1 < c_2 \leq 1$. Then we can approximate the correlation dimension as

$$\mathrm{cd}_\mathrm{R}\left(D;c_1\mu,c_2\mu\right) \approx C(c_1,c_2)\mu,$$

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 α 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 μ, not as a function of μ/σ.

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

$$\mathrm{cd}_\mathrm{R}\left(D;r_1,r_2\right) = \mathrm{cd}_\mathrm{R}\left(D';Nr_1,Nr_2\right).$$

Given a dataset D with K columns, we denote by $\mathrm{ind}\left(D\right)$ a random binary dataset having K independent variables such that the probability of ith variable being 1 is equal to the probability of ith column of a random transaction sampled from D being 1. Alternatively, $\mathrm{ind}\left(D\right)$ 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 $\mathrm{ind}\left(D\right)$, 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*

$$\mathrm{cd_A}\left(D; \alpha, 1-\alpha\right) \leq \mathrm{cd_A}\left(\mathrm{ind}\left(D\right); \alpha, 1-\alpha\right),$$

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 $\mathrm{Var}\left[Z_D\right]$ of the variable Z_D can be shown to be no more than the variance $\mathrm{Var}\left[Z_{\mathrm{ind}(D)}\right]$; 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 seems to shrink 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 $\mathrm{ind}\left(H, p\right)$ 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 $\mathrm{cd_A}\left(\mathrm{ind}\left(H,p\right); \alpha, 1-\alpha\right)$: setting $q = 2p(1-p)$ we have

$$\mathrm{cd_A}\left(\mathrm{ind}\left(H,p\right); \alpha, 1-\alpha\right) \approx C(\alpha)\sqrt{\frac{Hq}{1-q}}.$$

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

$$\mathrm{cd_A}\left(D; \alpha, 1-\alpha\right) \text{ and } \mathrm{cd_A}\left(\mathrm{ind}\left(H,s\right); \alpha, 1-\alpha\right)$$

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

$$\mathrm{cd_A}\left(\mathrm{ind}\left(K,s\right); \alpha, 1-\alpha\right) = \mathrm{cd_A}\left(\mathrm{ind}\left(D\right); \alpha, 1-\alpha\right),$$

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 $\mathrm{ncd_A}\left(D; \alpha, 1-\alpha\right)$, to be an integer H such that

$$\left|\mathrm{cd_A}\left(\mathrm{ind}\left(H,s\right); \alpha, 1-\alpha\right) - \mathrm{cd_A}\left(D; \alpha, 1-\alpha\right)\right|$$

is minimized. Proposition 2 implies the following statement.

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

$$\mathrm{ncd_A}\left(D; \alpha, 1-\alpha\right) \approx \left(\frac{\mathrm{cd_A}\left(D\right)\alpha, 1-\alpha}{\mathrm{cd_A}\left(\mathrm{ind}\left(D\right)\right)\alpha, 1-\alpha}\right)^2 K.$$

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 $\mathrm{cd_A}\left(D; 1/4, 3/4, 50\right)$.

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

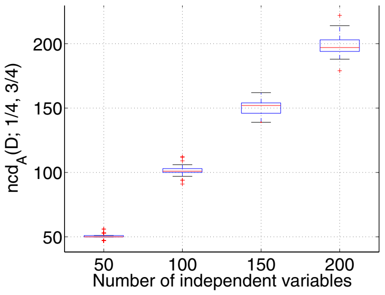

Figure 1. Normalized correlation dimension for data having K independent dimensions for $K \in \{50, 100, 150, 200\}$.

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 $p_{\max}$ uniformly at random from $[0, 1]$. Then, the probability p_i was picked uniformly 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.

In the second experiment 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 μ/σ, where $\mu = \mathrm{E}\left[Z_D\right]$ and $\sigma^2 = \mathrm{Var}\left[Z_D\right]$. The figure shows the behavior predicted by Proposition 2: the normalized fractal dimension is a linear function of μ/σ, and the slope is very close to $C(1/4) = 0.815$.

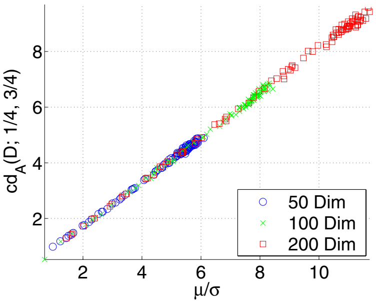

Figure 2. Correlation dimension as a function of μ/σ for data with independent columns (see Proposition 2). The y-axis is $\mathrm{cd}_\mathrm{A}\left(D; 1/4, 3/4\right)$ and the x-axis is μ/σ, where $\mu = \mathrm{E}\left[Z_D\right]$ and $\sigma^2 = \mathrm{Var}\left[Z_D\right]$. The slope of the line 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 $\mathrm{cd}_\mathrm{A}\left(D; 1/4, 3/4, 50\right)$ against $\mathrm{cd}_\mathrm{A}\left(D; 1/4, 3/4, 1\right)$. We see from Figure 3 that the correlation dimension has little dependency of N.

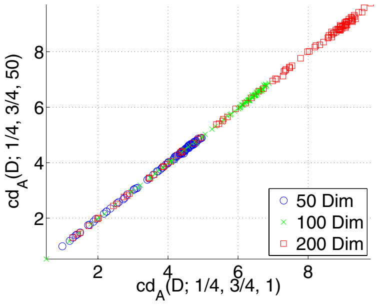

Figure 3. Correlation dimension $\mathrm{cd}_\mathrm{A}\left(D; 1/4, 3/4, 50\right)$ as a function of $\mathrm{cd}_\mathrm{A}\left(D; 1/4, 3/4, 1\right)$ for data having K independent dimensions for $K \in \{50, 100, 200\}$.

In the fourth experiment we verified the quality of the approximation of Proposition 4. We used the same data in the previous experiment. Figure 4 shows the correlation dimension against $\mu = \mathrm{E}\left[Z_D\right]$, 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 μ. However, the slope becomes more gentle as the number of columns increases.

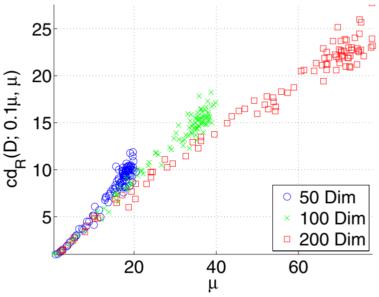

Figure 4. Correlation dimension as a function of μ for data with independent columns (see Proposition 4). The y-axis is $\mathrm{cd}_\mathrm{A}\left(D; 1/4, 3/4\right)$ and the x-axis is $\mu = \mathrm{E}\left[Z_D\right]$, the average distance between two random points.

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:

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

and

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

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_{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

$$\begin{aligned}\mathbb{P}\left(Z_i = 1 \mid Z_{i-1} = 0\right) &= \mathbb{P}\left(Z_i = 0 \mid Z_{i-1} = 1\right)\\ &= 2t_i\left(1 - t_i\right).\end{aligned}$$

A rough measure of the amount of correlation in the data is $t = \sum 2t_i\left(1-t_i\right)$. 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.

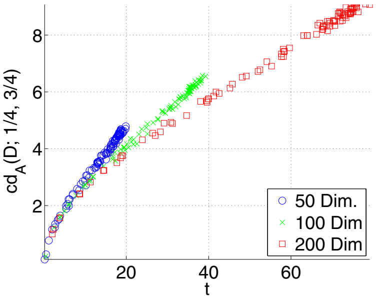

Figure 5. Correlation dimension as a function of t, a rough measure of correlation in a dataset. The y-axis is $\mathrm{cd}_{\mathrm{A}}\left(D; 1/4, 3/4\right)$ and the x-axis is the quantity $t = \sum 2t_i\left(1-t_i\right)$, where t_i is the reversal probability between columns i and $i-1$.

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 $\mu = \mathrm{E}\left[Z_D\right]$. Figure 6 shows the normalized correlation dimension as a function of μ for the datasets used in Figure 1. We see that the normalized dimension does not depend of sparsity, as expected.

We tested Proposition 7 by plotting the normalized dimension as a function of $K\mathrm{cd}_{\mathrm{A}}\left(D\right)^2/\mathrm{cd}_{\mathrm{A}}\left(\mathrm{ind}\left(D\right)\right)^2$. We used the generated datasets from the previous experiment and from our fifth experiment, as well. Figure 7 reveals that the approximation is good for the used datasets.

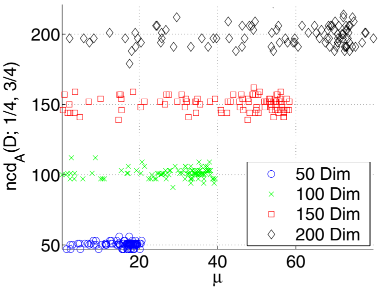

Figure 6. Normalized correlation dimension as a function of μ, the average distance between two random points. The x-axis is $\mu = \mathrm{E}\left[Z_D\right]$ and the y-axis is $\mathrm{ncd}_{\mathrm{A}}\left(D; 1/4, 3/4\right)$.

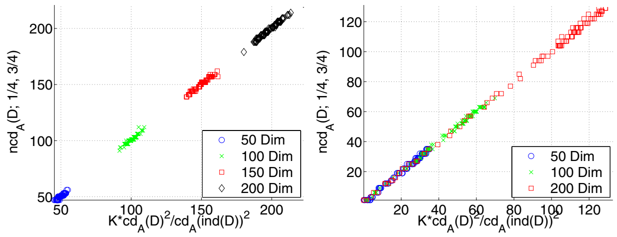

Figure 7. Normalized correlation dimension as a function of $K\mathrm{cd}_{\mathrm{A}}\left(D\right)^2/\mathrm{cd}_{\mathrm{A}}\left(\mathrm{ind}\left(D\right)\right)^2$. The left figure contains datasets with independent columns and in the right figure adjacent columns of the datasets depend on each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

5.2. Real datasets

In this section we investigate how our dimensions behave with 9 real-world datasets: *Accidents*, *Courses*, *Kosarak*, *Paleo*, *POS*, *Retail*, *WebView-1*, *WebView-2* and *20 Newsgroups*. The basic information about the datasets is summarized in Table 1.

The datasets are as follows. *20 Newsgroups*[1] is a collection of approximately 20 000 newsgroup documents across 20 different newsgroups [18]. Data in *Accidents*[2] 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].

[1] http://people.csail.mit.edu/jrennie/20Newsgroups/

[2] http://fimi.cs.helsinki.fi/data/accidents.dat.gz

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	340 183	11 500 870	7.21
Courses	5 021	2 405	64 743	0.54
Kosarak	41 271	990 002	8 019 015	0.02
Paleo	139	501	3 537	5.08
POS	1 657	515 597	3 367 020	0.39
Retail	16 470	88 162	908 576	0.06
WebView-1	497	59 602	149 639	0.51
WebView-2	3 340	77 512	358 278	0.14

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-of-sale 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*[6] consists of (anonymized) click-stream data of a Hungarian on-line news portal. *Retail*[7] is a retail market basket data supplied by an anonymous Belgian retail supermarket store [5]. The dataset *Paleo*[8] 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 $\mathrm{cd_A}\left(D; 1/4, 3/4\right)$ for each dataset. In order to do that, we needed to estimate the probabilities $\mathbb{P}\left(Z_D < r\right)$. Since some of the datasets had a very large amount of rows (see Table 1), we estimate the probabilities $\mathbb{P}\left(Z_D < r\right)$ by

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

where $I\left(|x-y| < r\right)$ 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 $\mathrm{cd_A}\left(\mathrm{ind}\left(D\right); 1/4, 3/4\right)$, 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 $\mathbb{P}\left(Z_{\mathrm{ind}(D)} < r\right)$. The estimation was done by generating 10 000 points from the distribution of $Z_{\mathrm{ind}(D)}$.

The dimensions $\mathrm{cd_A}\left(D\right)$ and $\mathrm{cd_A}\left(\mathrm{ind}\left(D\right)\right)$ are given in Table 2. We see that the dimensions are very small. The reason is that the datasets are quite sparse. We also observe that $\mathrm{cd_A}\left(\mathrm{ind}\left(D\right)\right)$ is always larger than $\mathrm{cd_A}\left(D\right)$, which suggests that there is at least some structure in the datasets.

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

$$\hat{C}(1/4) = \mathrm{cd_A}\left(\mathrm{ind}\left(D\right); 1/4, 3/4\right)\frac{\sigma}{\mu}.$$

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, $D' = \mathrm{ind}\left(D\right)$. The third column is the fraction μ/σ, where $\mu = \mathrm{E}\left[Z_{D'}\right]$ and $\sigma^2 = \mathrm{Var}\left[Z_{D'}\right]$. The fourth column is an estimate of the coefficient $C(1/4)$ obtained by dividing $\mathrm{cd_A}\left(D'\right)$ with μ/σ.

Data	$\mathrm{cd_A}\left(D\right)$	$\mathrm{cd_A}\left(D'\right)$	μ/σ	$\hat{C}\left(1/4\right)$
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 $\mathrm{ncd_A}\left(D; 1/4, 3/4\right)$. For this we computed the probability s such that

$$\mathrm{cd_A}\left(\mathrm{ind}\left(K, s\right); \alpha, 1-\alpha\right) = \mathrm{cd_A}\left(\mathrm{ind}\left(D\right); \alpha, 1-\alpha\right)$$

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

[3] http://www.ecn.purdue.edu/KDDCUP/data/BMS-POS.dat.gz

[4] http://www.ecn.purdue.edu/KDDCUP/data/BMS-WebView-1.dat.gz

[5] http://www.ecn.purdue.edu/KDDCUP/data/BMS-WebView-2.dat.gz

[6] http://fimi.cs.helsinki.fi/data/kosarak.dat.gz

[7] http://fimi.cs.helsinki.fi/data/retail.dat.gz

[8] NOW public release 030717 available from [10].

Table 3. Normalized correlation dimensions of the datasets.

Data	K	ncd_A	$\frac{\mathrm{ncd}_\mathrm{A}(D)}{K}$	$\frac{K\mathrm{cd}_\mathrm{A}(D)^2}{\mathrm{cd}_\mathrm{A}(\mathrm{ind}(D))^2}$
Accidents	469	220	0.47	222.91
Courses	5 021	304	0.06	344.24
Kosarak	41 271	2 378	0.06	3 684.78
Paleo	139	15	0.11	19.90
POS	1 657	181	0.11	242.91
Retail	16 470	1 791	0.11	2 107.52
WebView-1	497	190	0.38	214.33
WebView-2	3 340	359	0.11	512.97

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.

We also tested the accuracy of Proposition 7 with *20 Newsgroups* dataset[9]. 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.

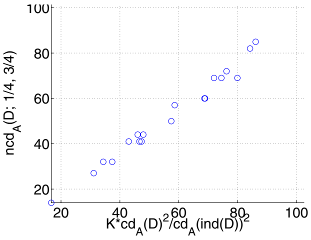

Figure 8. Normalized correlation dimension as a function of $K\mathrm{cd}_\mathrm{A}\left(D\right)^2/\mathrm{cd}_\mathrm{A}\left(\mathrm{ind}\left(D\right)\right)^2$. Each point represents one newsgroup in *20 Newsgroups* dataset.

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.

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	11 500 870	973	0.085
Courses	64 743	9	0.141
Paleo	3 537	0.1	0.039
Kosarak	8 019 015	793	0.099
POS	3 367 020	447	0.133
Retail	908 576	103	0.113
WebView-1	149 639	17	0.114
WebView-2	358 278	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 PCA.

We performed PCA to our datasets and computed the percentage of the variance explained by the M first PCA variables, where $M = \mathrm{ncd}_\mathrm{A}\left(D\right)$. 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 $\mathrm{ncd}_\mathrm{A}\left(D\right)$ PCA components explain relatively large portion of the variance for *Accidents*, *POS*, and *WebView-1*, but explains less for *Paleo* and *WebView-2*.

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. 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.

[9] The messages were converted into bag-of-words representations and 200 most informative variables were kept.

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

Data	$\mathrm{ncd}_{\mathrm{A}}(D)$	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	1 394

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 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 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 $d << K$. Furthermore, computing the VC-dimension of a given dataset is a difficult problem [26].

Also the work on random projections and dimensionality reductions, such as in [1], is related but that 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

[1] 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.