

Using Correlation Dimension for Analysing Text Data*

Ilkka Kivimäki, Krista Lagus, Ilari T. Nieminen,
Jaakko J. Väyrynen, and Timo Honkela

Adaptive Informatics Research Centre,
Aalto University School of Science and Technology
firstname.lastname@tkk.fi
<http://www.cis.hut.fi/research/>

Abstract. In this article, we study the scale-dependent dimensionality properties and overall structure of text data with a method that measures correlation dimension in different scales. As experimental results, we present the analysis of text data sets with the Reuters and Europarl corpora, which are also compared to artificially generated point sets. A comparison is also made with speech data. The results reflect some of the typical properties of the data and the use of our method in improving various data analysis applications is discussed.

Keywords: Correlation dimension, dimensionality calculation, dimensionality reduction, statistical natural language processing.

1 Introduction

Knowing the *intrinsic dimensionality* of a data set can be a benefit, for instance, when deciding the parameters of a dimension reduction method. One popular technique for determining the intrinsic dimensionality of a finite data set is calculating its correlation dimension, which is a fractal dimension. This is usually done according to the method introduced by Grassberger and Procaccia in [1]. Usually the goal in these dimensionality calculations is to characterise a data set by a single statistic. Not much emphasis is always put to the notion of the dependence of correlation dimension on the scale of observation. However, as we will show, the scale-dependent dimensionality properties can vary between different data sets according to the nature of the data. Most neural network and statistical methods such as singular value decomposition or the self-organising map are usually applied without considering this fact. Even in papers studying dimensionality calculation methods (e.g. [2] and [3]) the scale-dependence of dimensionality is noted, but usually left without further discussion.

We focus on natural language data. It has been observed that the intrinsic dimensionality of text data, such as term-document matrices, is often much

* This work has been supported by the Academy of Finland and a grant from the Department of Mathematics and Statistics at the University of Helsinki (IK).

lower than the dimensionality of the original data space due to its sparseness and correlation in the data. In addition to term-document matrices, we also consider speech data and data about co-occurrences of words inside the same sentence, which is another approach of encoding semantic information in text.

In a research closely related to ours, [4] studies the local dimensionality of a word space concluding that the small-scale dimensionality is very low compared to the dimensionality of the data space. In that paper, the word space was built with the random indexing method, whereas we use a more standard setting. In [5] a method for calculating dimensionality is presented and the effect of different term-weighting methods on the dimensionality estimates is studied. Also in [6] dimensionality calculations were made for natural language data, in this case partly with the same data sets that we use. For the calculations, they used a modified version of a method originally proposed by [7] based on eigenvalue information of the autocorrelation matrix of the data.

2 The Scale-Dependent Correlation Dimension

Correlation dimension can be measured for a finite data set $\{x_1, \dots, x_N\} \subset \mathbb{R}^n$ by the Grassberger-Procaccia (GP) algorithm [1]. First we define the *correlation sum* $C(r)$ as the probability of a randomly chosen pair of data points being within distance r from each other:

$$C(r) = \frac{2}{N(N-1)} \sum_{i=1}^N \sum_{j=i}^N I(\rho(x_i, x_j) < r),$$

where $I(x)$ is the indicator function (i.e. 1, if the argument condition holds and 0 otherwise) and ρ is the metric. We will use the Euclidean distance as the metric. In the usual implementation of the GP-algorithm the correlation dimension is then defined as the slope of the linear segment in the plot of the correlation sum $C(r)$ against r in double logarithmic coordinates.

In other words, one first finds a *scale* $[r_i, r_j]$, where the loglog-plot of the correlation sum appears linear and then computes the correlation dimension $\hat{\nu}$ as the logarithmic derivative on this interval:

$$\hat{\nu} = \hat{\nu}(r_i, r_j) = \frac{\log(C(r_j)/C(r_i))}{\log(r_j/r_i)}.$$

However, instead of defining only one segment or *scale* $[r_i, r_j]$, the dimensionality of a data set can actually be measured scale-dependently by studying the local derivatives with different r . For estimating these dimensionality curves, we decided throughout the paper (after experimenting with different values) to use 100 measuring points r_1, \dots, r_{100} , distributed logarithmically on the interval $[r_1 = \min \rho(x_i, x_j), r_{100} = \max \rho(x_i, x_j)]$. For illustration purposes, we use an additional smoothing method by a simple triangular kernel of window length w .

The window length then defines the width of the scale of observation. Thus we can finally define the scale-dependent dimensionality of the data set as

$$\nu(r) = \frac{1}{2w} \sum_{\substack{j=i-w, \\ j \neq i}}^{i+w} \hat{\nu}(r_i, r_j), \text{ when } r \in [r_i, r_{i+1}]. \tag{1}$$

Again, after experimenting with different window lengths, we fixed the value at $w = 6$ throughout the paper.

3 Experiments

3.1 Reuters

The first data set for experiments with the method presented above is a document collection gathered from the Reuters corpus of news articles [8]. We took a subset of the corpus consisting of 10 000 articles. The documents have been preprocessed by removing stop words and reducing all words into their stems. The frequencies of the subset’s 300 most frequent terms were counted for each article resulting in a matrix of 10 000 vectors with dimensionality 300. As a common preprocessing method, a term frequency–inverse document frequency or *tf/idf*–weighting [9] was performed for the raw frequencies.

The dimensionality curve for the Reuters data set is shown in Figure 1(a). The curve indicates some essential properties of a term–document matrix. Starting from the large scale on the right end of the curve, the dimensionality starts increasing from zero as the scale narrows down from the diameter of the data set. It soon reaches a maximum value which would traditionally be interpreted as the dimensionality of the set. For our 10 000 article subset of the Reuters collection the dimensionality would thus be approximately $\nu(r) = 7.5$. Continuing on to smaller scales on the left of the maximum value, the dimensionality decreases significantly.

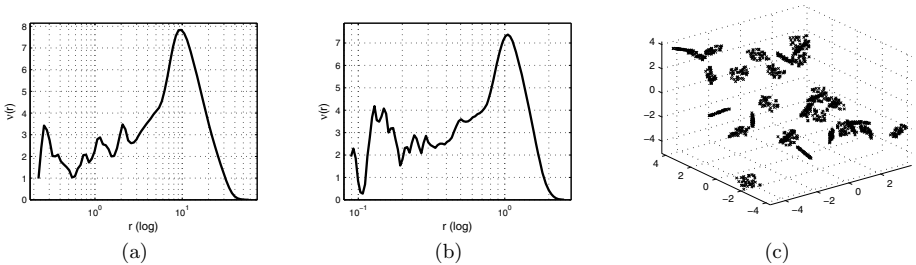


Fig. 1. The dimensionality curve of $\nu(r)$ for the Reuters data set (a), the data set consisting of 200 3-D clusters with 40 points scattered in 20-D space (b) and the illustration of 2-D clusters scattered in 3-D space

The low values of $\nu(r)$ with small r mean that the data shows high auto-correlation locally. We will simulate this kind of behaviour with an artificially generated set of points that consists of low-dimensional clusters scattered in a high-dimensional embedding space. This idea is illustrated in Figure 1(c), which shows 2-dimensional clusters scattered in 3-dimensional space. Our example data set contains 200 subsets of 40 points from uniform distribution in a 3-dimensional unit cube. These 200 clusters are each rotated by a random rotating matrix and then shifted by a randomly generated vector in a 20-dimensional embedding space. The dimensionality curve for this data set is shown in Figure 1(b).

The high autocorrelation at small distances suggests that the documents differ only in the frequencies of a few different key terms from each other. This phenomenon can cause problems with large document collections because there is not enough information for making a detailed analysis of the collection. Thus the dimensionality curve could possibly be used to evaluate the differentiability of a word-document data set and to develop a better feature selection method through a hierarchical clustering. For instance, expanding the feature space within proper clusters with terms that are significant to the documents in that cluster making the differentiation of the document vectors easier. Related work has already been done in [10] and also in [11].

3.2 ISOLET

To contrast the shape of the dimensionality curve for text data we will use the ISOLET speech data set [12]. It consists of 7797 samples of spoken English letters of which 617 acoustic features were measured resulting in 7797 vectors in 617-dimensional space. The dimensionality curve is shown in Figure 2(a)¹. One can interpret a part with linear behaviour in the correlation integral in the scale between $r = 4.7$ and $r = 6.8$, where we get a rough estimate of $\nu(r) = 13.3$.

An interesting feature of the ISOLET data set dimensionality curve is in the small scale where the dimensionality value seems to explode. This explosion is caused by the noise in the acoustic speech data which produces variance in small scale in all the noisy feature components of the embedding space. Again we use an artificially generated data set to support this observation. Figure 2(b) shows the dimensionality curve of a data set consisting of 8000 random points from the 3-dimensional unit hypercube to which 20-dimensional Gaussian white noise with a variance of 10^{-3} has been added.

Figures 1(b) and 2(b) can now be compared to each other. The sets represented by the curves differ quite a lot in their scale-dependent dimensionalities, but show also similarity in dimensionalities in certain scales. In the extreme case, a bad method or a careless examination could lead to an interpretation of the sets having the same dimensionality. This again supports the significance of investigating dimensionality properties of data sets in several scales.

¹ Here we omitted the smallest point-pair distance from further analysis as it was several magnitudes smaller than the rest of the distances forcing the curve to drop to zero at the left end.

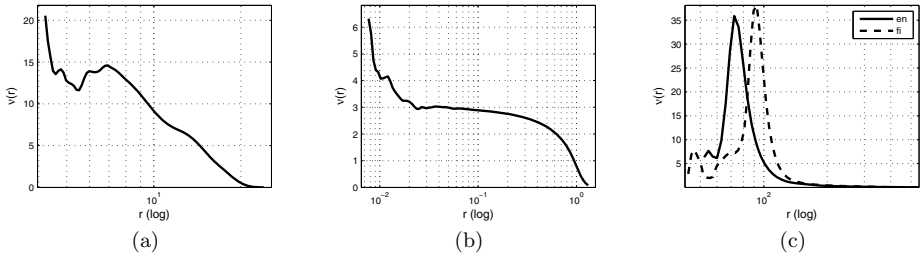


Fig. 2. $\nu(r)$ for the ISOLET data set (a), the 3-D data set with 20-D Gaussian white noise (b) and for the English (solid line) and Finnish (dashed line) Europarl data sets (c)

3.3 Europarl

The Europarl corpus version 3 contains the minutes from the sessions of the EU parliament from the years 1997-2006 in 11 languages [13]. We used it to study dimensionality properties of word co-occurrence data. Only the results of experiments conducted with the English and Finnish parts of the corpus are presented here. We used a subset of the corpus containing the sessions of 591 days that were recorded for both languages. Both of the subsets were preprocessed by first removing the XML-tagging used in the files, then applying sentence boundary detection, tokenization and removal of punctuation and special characters.

The data matrix had the 1000 most frequent words as the term vectors and the 20000 most frequent words as the context features. For the elements of the matrix, we counted the frequencies of each term and feature word occurring in the same sentence. Finally, we took the logarithm of the frequencies increased by one as is done in the *tf/idf*-method, however preserving the frequency rank information in this case.

The dimensionality curves for the English and Finnish Europarl data sets are shown in Figure 2(c). The overall shape of both curves looks the same. The long and thin right tails imply that some of the 1000 data points in both sets are spread very far from the others and also that there are large scale correlations in the data. However, the sharp peak in both curves shows that the interpoint distances are concentrated on a narrow scale causing rather high dimensionality estimates in both cases. The peak dimensionality values are 35.9 for English and 38.1 for Finnish. On the left of the peak value, the dimensionality curves both descend showing low-dimensionality in small scales for both data sets. However, this effect is not as pronounced as with the Reuters data set because of the shortness of the left tail in both curves. One more thing worth noting is the different positioning of the curves on the horizontal axis. Further experiments, not reported here, suggest that the location of the peak value on the horizontal axis correlates highly with the average sentence length in words of the language, but also the word type/token ratio may have an impact on this phenomenon.

4 Conclusions and Discussion

We have presented a method for observing the scale-dependent dimensionality of a finite data set based on the Grassberger-Procaccia algorithm. The dimensionality curves obtained with our method give interesting information about the structure of the data set and show some typical characteristics of the phenomenon causing the data. We illustrated the method with natural language data and artificially generated data discussing its indications and benefits.

The relevance of scale-dependent dimensionality for dimensionality reduction, clustering and other data analysis methods seems to be an interesting topic for future research, which, according to our knowledge, has not received much attention before. Also the reliability of the GP-algorithm, as used in our study, needs to be studied more. An additional interesting question is how different data analysis methods respond to scale invariance or self-similarity (or the lack of them) in a data set. These ideas and questions will get the authors' attention in future investigations and also serve as the motivation to the whole content of this article.

References

1. Grassberger, P., Procaccia, I.: Characterization of strange attractors. *Phys. Rev. Lett.* 50(5), 346–349 (1983)
2. Camastra, F.: Data dimensionality estimation methods: a survey. *Pattern Recognition* 36(12), 2945–2954 (2003)
3. Theiler, J.: Estimating fractal dimension. *Journal of the Optical Society of America A* 7, 1055–1073 (1990)
4. Karlgren, J., Holst, A., Sahlgren, M.: Filaments of meaning in word space. *Advances in Information Retrieval*, pp. 531–538 (2008)
5. Kumar, C.A., Srinivas, S.: A note on effect of term weighting on selecting intrinsic dimensionality of data. *Journal of Cybernetics and Information Technologies* 9(1), 5–12 (2009)
6. Kohonen, T., Nieminen, I.T., Honkela, T.: On the quantization error in SOM vs. VQ: A critical and systematic study. In: *Proceedings of WSOM 2009*, pp. 133–144 (2009)
7. Fukunaga, K., Olsen, D.R.: An algorithm for finding intrinsic dimensionality of data. *IEEE Trans. Comput.* 20, 176–183 (1971)
8. Lewis, D.D., Yang, Y., Rose, T.G., Li, F.: Rcv1: A new benchmark collection for text categorization research. *J. Mach. Learn. Res.* 5, 361–397 (2004)
9. Manning, C.D., Schütze, H.: *Foundations of statistical natural language processing*. MIT Press, Cambridge (1999)
10. Vinay, V., Cox, I.J., Milic-Frayling, N., Wood, K.R.: Measuring the complexity of a collection of documents. In: Lalmas, M., MacFarlane, A., Rüger, S.M., Tombros, A., Tsikrika, T., Yavlinsky, A. (eds.) *ECIR 2006*. LNCS, vol. 3936, pp. 107–118. Springer, Heidelberg (2006)
11. Cai, D., He, X., Han, J.: Document clustering using locality preserving indexing. *IEEE Transactions on Knowledge and Data Engineering* 17(12), 1624–1637 (2005)
12. Cole, R., Fauty, M.: Spoken letter recognition. In: *HLT 1990: Proceedings of the Workshop on Speech and Natural Language*, pp. 385–390 (1990)
13. Koehn, P.: Europarl: A parallel corpus for statistical machine translation. In: *Machine Translation Summit X*, pp. 79–86 (2005)