Determining the number of clusters of high dimensional vectors via sine matrices and hierarchical agglomerative algorithms

G.F. Lima

1 Overview

Many natural language processing tools, such as Latent Semantic Analysis and word2vec, model a corpus of text as a high dimensional vector space. The semantic similarity between two words can be represented by the cosine of the angle between their respective vector representations. Conversely, one can represent the semantic *distance* by the sine of that angle. Given a list of words $\{w_1, \ldots, w_n\}$, one can obtain a similarity/distance matrix M such that the entry M_{ij} is the similarity/distance between words w_i and w_j.

In this paper we shall present a method for finding the number of clusters of a set of high dimensional vectors given their sine matrix.

2 Variation and distance matrices

A well known method for determining if two or more groups of data points are significantly distinguishable is to compare the sample variance inside the groups with the sample variance between the groups. In a normed vector space the sample variance is given by

$$\frac{1}{n-1}\sum_{i=1}^{i=n}(x_i - \bar{x})^2,$$

where $\bar{x}$ is the mean of the x_i. A sine matrix will only provide us with a pseudometric on the list of original points, and this might not include their mean, yet given the equation

$$\sum_{i=1}^{i=n}(x_i - \bar{x})^2 = \frac{1}{2n}\sum_{i=1}^{n}\sum_{j=1}^{n}(x_i - x_j)^2$$

the sample variance can also be expressed as

$$\frac{1}{2n(n-1)}\sum_{i=1}^{n}\sum_{j=1}^{n}(x_i - x_j)^2.$$

and the above formula can be generalised to a pseudometric d on the set of points as

$$\frac{1}{2n(n-1)}\sum_{i=1}^{n}\sum_{j=1}^{n} d(x_i, x_j)^2.$$

Further more, given the usual partition of the total sum of squares as

$$\sum_{i=1}^{m}\sum_{j=1}^{n_i}(x_{ij} - \bar{x})^2 = \sum_{i=1}^{m} n_i(\bar{x}_i - \bar{x})^2 + \sum_{i=1}^{m}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)^2,$$

or equivalently

$$\sum_{i=1}^{m} n_i(\bar{x}_i - \bar{x})^2 = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(x_{ij} - \bar{x})^2 - \sum_{i=1}^{m}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)^2$$

we may generalise the sample variance between groups given by

$$\frac{1}{m-1}\sum_{i=1}^{m} n_i(\bar{x}_i - \bar{x})^2$$

to

$$\frac{1}{m-1}\left(\frac{1}{2n}\sum_{i=1}^{m}\sum_{i'=1}^{m}\sum_{j=1}^{n_i}\sum_{j'=1}^{n_i} d(x_{ij}, x_{i'j'})^2 - \sum_{i}^{m}\frac{1}{2n_i}\sum_{j}^{n_i}\sum_{j'}^{n_i} d(x_{ij}, x_{ij'})^2\right).$$

From the previous formulas we may generalise the statistic

$$\frac{\frac{1}{m-1}\sum_{i=1}^{m} n_i(\bar{x}_i - \bar{x})^2}{\frac{1}{n-m}\sum_i\sum_j(x_{ij} - \bar{x}_i)}$$

which is commonly known as the ratio of the between-group variance and the within-group variance, to the following

$$\frac{\frac{1}{m-1}\left(\frac{1}{2n}\sum_{i=1}^{m}\sum_{i'=1}^{m}\sum_{j=1}^{n_i}\sum_{j'=1}^{n_i} d(x_{ij}, x_{i'j'})^2 - \sum_{i}^{m}\frac{1}{2n_i}\sum_{j}^{n_i}\sum_{j'}^{n_i} d(x_{ij}, x_{ij'})^2\right)}{\frac{1}{n-m}\left(\sum_{i=1}^{m}\frac{1}{2n_i}\sum_{j=1}^{n_i}\sum_{j'=1}^{n_i} d(x_{ij}, x_{ij'})\right)} \qquad (*)$$

Let S be a set of n points with metric d, and $\mathfrak{S} = \{S_1, \ldots, S_m\}$ a partition such that, if $i \neq j$, then $S_i \cap S_j = \emptyset$, and $\bigcup S_i = S$. We define the *variation ratio* of $(S, \mathfrak{S}, d)$, or $F(S, \mathfrak{S}, d)$, by the above formula $(*)$. In this context, given $1 \leq i \leq m$, and $1 \leq j \leq n_i$ the element x_{ij} belongs to the partition S_i.

The reader can probably see that, given a metric space (S, d), a hierarchical clustering algorithm will produce a sequence of partitions $\{\mathfrak{S}_i\}_i$ of the original space S.

3 Simulating clusters

For the purposes of simulating clusters in high dimensions, we shall pick a number n of clusters of size s, a natural number d to represent the dimension of the subspace where each will be generated, and a (positive) real number v for the clusters variance. For each $1 \leq i \leq n$ we shall then generate a set C_i of s random vectors following a multivariate normal distribution of d dimensions with mean vector $(1, 1, \ldots, 1)$ and covariance matrix $v \cdot I_d$, where I_d is the identity $d \times d$ matrix.

To create our sample S of n clusters we shall embed the C_i in an m-dimensional space, where $m = n \times s$.

Let $\iota_k : \mathbb{R}^s \to \mathbb{R}^m$ be the embedding of $\mathbb{R}^s$ into $\mathbb{R}^m$ such that the vector $(x_1, \ldots, x_s)$ is mapped to

$$(\underbrace{0, \ldots, 0}_{k \text{ 0's}}, x_1, \ldots, x_s, 0, \ldots).$$

We shall then embed the cluster C_1 in to $\mathbb{R}^m$ as $\widehat{C_1} = \iota_0(C_1)$, the cluster C_2 as $\widehat{C_2} = \iota_s(C_2)$, and the cluster C_i as $\widehat{C_i} = \iota_{(i-1)s}(C_i)$. We obtain our sample S by taking the union $\bigcup \widehat{C_i}$.

4 Variance ratio of samples

Given a sample of clusters S and a metric d, using a hierarchical clustering algorithm we may obtain a finite sequence $\{\mathfrak{S}_i\}_{i=1}^N$ of partitions of S, and from that a finite sequence of real numbers $\{F(S, \mathfrak{S}_i, d)\}_{i=1}^N$.

We shall inspect the behaviour of such sequences by looking at 4 different cases, with the same number of clusters, the same sizes of clusters, and the same number of dimensions of the subspaces where they were generated. On all of them we will, for now, set the overlap factor to 0. For each case we will produce twenty different samples $S_1, \ldots, S_{20}$. For each sample S_j we will use an agglomerative single-linkage

hierarchical clustering algorithm to produce our sequence of N partitions $\{\mathfrak{S}_{ji}\}_{i=1}^{N}$, and, for each j, we will plot the function

$$\begin{array}{rccc} f_j: & \{1,\ldots,N\} & \to & \mathbb{R} \\ & i & \mapsto & F(S_j, \mathfrak{S}_{ji}, d) \end{array}$$

The four cases we will look at are:

1. Two clusters of size 15 generated in subspaces of dimension 10.
2. Three clusters of size 15 generated in subspaces of dimension 10.
3. Three clusters of size 15 generated in subspaces of dimension 20.
4. Four clusters of size 15 generated in subspaces of dimension 10.

For the sake of clarity, we have connected the consecutive points $(i, f_j(i))$ and $(i+1, f_j(i+1))$, and so each sample gives us a piece-wise linear function. The vertical dashed blue line indicates the index i where the hierarchical clustering algorithm produces a partition with the correct number of clusters and the correct sizes. We shall refer to this index as the *target* index.

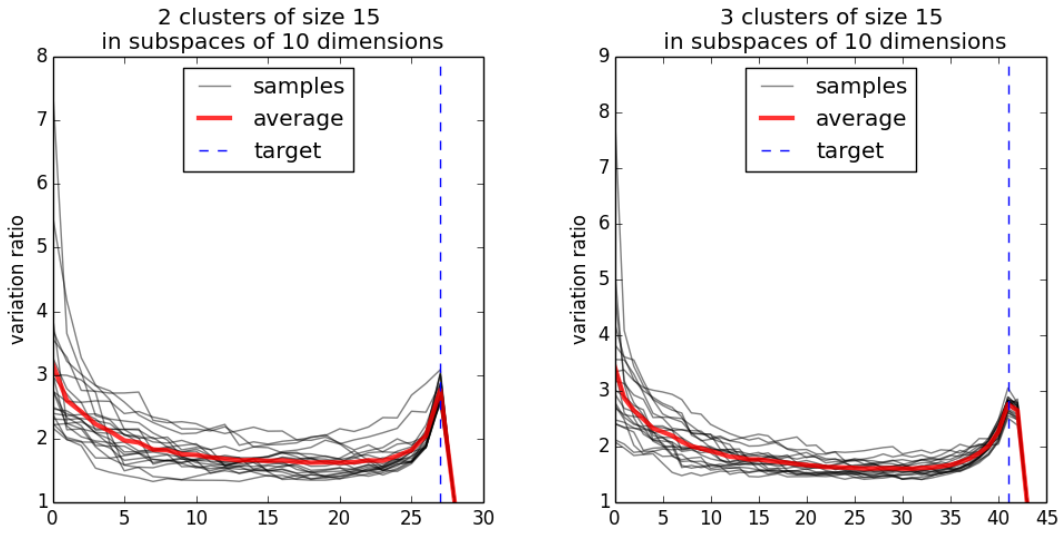

Figure 1: cases 1 and 2.

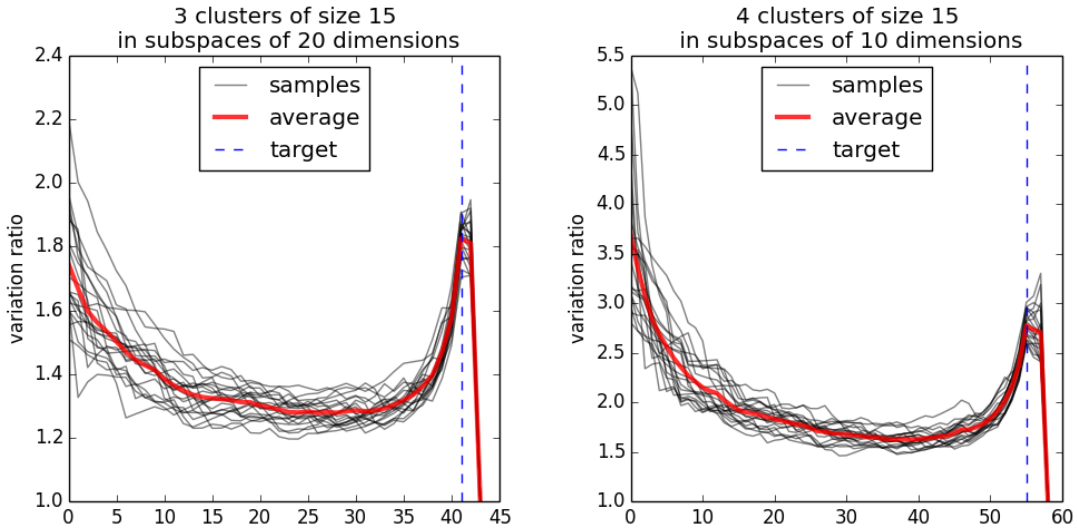

Figure 2: cases 3 and 4.

From the figures we see that the correct partitions coincide with the highest slopes, or more precisely, the target index is where the function

$$\begin{array}{rccc} \Delta(f_j): & \{2,\ldots,N\} & \to & \mathbb{R} \\ & i & \mapsto & f_j(i) - f_j(i-1) \end{array}$$

obtains its maximum.

5 Tests

5.1 Multiple clusters

We fixed the number of clusters, their sizes, and the dimension of the subspaces where they were generated, and then generated 100 different random samples – as described in section 3. The following tables sumarize the prediction accuracy of our algorithm. Each attempt was labeled as correct if it successfully predicted the number of clusters and all their sizes for samples with two, three, four, and five clusters.

		Size of clusters		
		5	10	15
subspace dimension	5	76%	87%	97%
	10	85%	90%	97%
	15	83%	99%	99%
	20	87%	99%	98%

Table 1: Percentage of correct predictions for samples with two clusters.

		Size of clusters		
		5	10	15
subspace dimension	5	80%	92%	89%
	10	72%	90%	97%
	15	82%	90%	100%
	20	77%	96%	96%

Table 2: Percentage of correct predictions for samples with three clusters.

		Size of clusters		
		5	10	15
subspace dimension	5	71%	87%	87%
	10	72%	83%	99%
	15	76%	90%	95%
	20	60%	91%	98%

Table 3: Percentage of correct predictions for samples with four clusters.

		Size of clusters		
		5	10	15
subspace dimension	5	68%	82%	90%
	10	50%	84%	98%
	15	61%	88%	99%
	20	48%	90%	95%

Table 4: Percentage of correct predictions for samples with five clusters.

5.2 Single cluster samples

For samples with only a single cluster, the algorithm exhibits a significantly lower accuracy, as the following table can show.

		Size of clusters		
		5	10	15
subspace dimension	5	14%	0%	0%
	10	7%	0%	0%
	15	5%	0%	0%
	20	4%	0%	0%

Table 5: Percentage of correct predictions for samples with a single cluster.

This is due to the fact that the between-group variance is zero when there is only one partition of S, yet this partition represents the correct number of clusters. One could say that this case is outside the scope of our algorithm, since our variation ratio presuposes more than one group of points.

In practice, what happens is that our algorithm indicates a cut off index smaller than the correct one, and, since we are using an aglomerative clustering algorithm, it yields the single cluster split into multiple parts which haven't yet been aglomerated.

6 Conclusion

The algorithm presents good predictive power for samples of 2 clusters or more where the clusters occur in distinct dimensions.