Using electronic corpora to study language variation: the problem of data sparsity

Hermann Moisl

1. Introduction

The proliferation of computational technology has generated an explosive production of electronically encoded information. Traditional philological methods for search and interpretation of this information have been overwhelmed by sheer volume, and computational methods have been developed to make the deluge tractable. These developments have implications for corpus-based study of language variation. As more and larger electronic corpora appear, effective analysis of them will increasingly be tractable only by adapting the interpretative methods developed by the statistical, information retrieval, and related communities. To use such analytical methods effectively, however, issues that arise with respect to the abstraction of data from corpora have to be understood.

This paper addresses an issue that has a fundamental bearing on the validity of analytical results based on such data: sparsity. The discussion is in three main parts. The first part shows how a particular class of computational methods, exploratory multivariate analysis, can be used in language variation research, the second explains why data sparsity can be a problem in such analysis, and the third outlines a solution.

2. Exploratory multivariate analysis in the study of language variation
A typical research question in the study of language variation is: given a corpus comprising a collection of documents each of which represents linguistic characteristics of a single speaker, can the documents and thus the speakers be interestingly classified on the basis of those characteristics? This kind of question can be answered using an empirical methodology known as exploratory multivariate analysis (Moisl 2008).

2.1 The nature of exploratory multivariate analysis

In describing a domain of interest, the researcher selects particular aspects of the domain which seem salient to the research question, and each selected aspect is represented by a variable. If only one aspect of the domain is observed the data is said to be univariate, if two aspects are observed the data is bivariate, if three trivariate, and so on up to some number \( n \). Any data where \( n \) is greater than 1 is multivariate.

The larger the number of variables, the more difficult data is to interpret. Where, say, 100 people are described in terms of a single variable 'age', visual inspection of the data is usually sufficient to identify age patterns. For two variables 'age' and 'height' identification of patterns by direct inspection becomes more difficult. If, however, these people were described by 50 variables ('age, 'height', 'eye colour', income'...), the data would be incomprehensible to most people. In general, as the number of variables grows, so does the difficulty of conceptualizing the interrelationships between and among objects on the basis of those variables. Exploratory multivariate analysis provides mathematically-based
methods for understanding data when it has too many variables for it to be interpretable by direct inspection.

2.2 Application to historical dialectology

Classification is one of the main applications of exploratory multivariate analysis, and as such is applicable in language variation research. To exemplify this, we consider the *Newcastle Electronic Corpus of Tyneside English* (NECTE), a corpus of dialect speech from North-East England (Allen *et al.* 2006) that includes phonetic transcriptions of 63 speaker interviews and associated social data. We have carried out exploratory analysis of the transcriptions with the aim of generating hypotheses about phonetic variation among the speakers (Moisl *et al.* 2006). The analysis was based on comparison of phonetic profiles associated with each of the NECTE speakers, where a profile was the number of times a given speaker used each of the phonetic segments in the NECTE transcription scheme. There are 156 segments, so a profile comprised 156 variables. The 63 profiles were represented as a 63 x 156 matrix $N_{63,156}$, a fragment of which is shown in Figure 1. The aim was to classify the speakers in accordance with the frequency values in their profiles.

@@ Insert Figure 1 here

$N_{63,156}$ is an example of data that is simply too large and complex to be interpretable by direct inspection. It was therefore analyzed using hierarchical cluster analysis (*Everitt et al.* 2001), a widely used exploratory
method that represents relative similarity among data items as a
constituency tree. The result is shown in Figure 2.

@@ Insert Figure 2 here

Trees like this are familiar to linguists as representations of phrase
structure, but differ from linguistic trees in the following respects:

- The leaves are not lexical tokens but labels for the data items; the
column in Figure 2 headed 'Speakers' contains the NECTE speaker
labels, and the other three columns social data to be discussed shortly.
- They represent not grammatical constituency but relativities of
similarity between clusters. The lengths of the branches linking the
clusters represent degrees of similarity: the shorter the branch, the more
similar the clusters --NG1 and NG2 in Figure 2 are very dissimilar,
NG1a(i) and NG1a(ii) very similar, and so on.

Figure 2 partitions the NECTE speakers into a hierarchy of clusters on the
basis of their phonetic usage. It can be given a sociolinguistic interpretation
by taking the social data that the NECTE corpus associates with the
speakers into account. There is, for example, a close correlation between the
cluster structure on the one hand, and the gender, educational, and
occupational attributes of the speakers on the other. The main phonetic
distinction is between clusters NG1 and NG2: NG2 corresponds to a small
group of speakers from Newcastle on the north shore of the river Tyne for
whom no detailed social data is available, but who are known to have been
male and female academics, and NG1 comprises mainly but not exclusively
working class speakers from Gateshead on the south shore of the Tyne. The
Gateshead speakers are subclustered into NG1a, which contains a mix of male and female manual workers with minimal education and of male and female administrative workers with additional education., and NG1b, which consists of male manual workers and a single female manual worker with minimal education; NG1a further subclusters the manual and the administrative workers into NG1a(i) and NG1a(ii) respectively; and so on.

3. The problem of data sparsity

Sparsity is a major issue in data analysis generally (Lee & Verleysen 2007; Verleysen 2003). Why this is so can be explained in terms of vector space representation. A vector is a sequence of numbers indexed by the positive integers 1, 2, 3...n.

A vector space is a geometrical interpretation of a vector in which the dimensionality $n$ of the vector defines an $n$-dimensional space, the sequence of numerical values comprising the vector specifies coordinates in the space, and the vector itself is a point at the specified coordinates. For example, the two components of a vector $v = (30 \ 70)$ in Figure 4 are coordinates of a point in a two-dimensional space, and those of $v = (40 \ 20 \ 60)$ of a point in three-dimensional space:

A length-4 vector defines a point in 4-dimensional space, and so on to any dimensionality $n$. 

@@ Insert Figure 3 here

@@ Insert Figure 4 here
A matrix in which the rows are data items and the \( n \) columns variables defines a manifold in \( n \)-dimensional space, where 'manifold' is understood as the shape of data in space (Munkres 2000). What is the 'shape' of data? Assume a matrix with 1000 3-dimensional vectors. If these vectors are plotted they form a cloud of points. Depending on the interrelationships of the objects that the vectors describe, that cloud might have some nonrandom structure; an example is shown in Figure 5.

@@ Insert Figure 5 here

The shape of the vector cloud is a manifold, and the idea extends directly to any dimensionality. For the purposes of this discussion, therefore, a manifold is a set of vectors in \( n \)-dimensional space.

To discern the shape of a manifold, there must be enough data points to give it adequate definition. If, as in the Figure 6a, there are just two points, the only reasonable manifold to propose is a line.

@@ Insert Figure 6 here

Where there are 3 points, a plane as in Figure 6b is reasonable. But it is only as the number of data points grows that the true shape emerges (Figure 6c). The general rule is: the more data the better for manifold definition.

Getting enough high-dimensional multivariate data is usually difficult or even intractable (Bishop 2006:33-8; Lee & Verleysen 2007; Verleysen 2003). The problem is that the space in which the manifold is embedded grows very quickly with dimensionality and, to retain a reasonable manifold definition, more and more data is required until, equally quickly, getting enough becomes impossible.
Assume some bivariate data in which both variables record frequency in the range 0..9: the number of possible vectors like (0,9), (3,4), and so on is $10 \times 10 = 100$. For trivariate frequency data the number of possible vectors like (0,9,2) and (3,4,7) is $10 \times 10 \times 10 = 1000$. In general, the number of possible vectors is $r^d$, where $r$ is the measurement range (here 0..9) and $d$ the dimensionality. The $r^d$ function generates an extremely rapid increase in data space size with dimensionality: even a modest $d = 8$ for a 0..9 range allows for 100,000,000 vectors. This is a problem because, the larger the dimensionality, the more difficult it becomes to define the manifold sufficiently well to achieve reliable analytical results.

To see why, assume that we want to analyze, say, 24 speakers in terms of their usage frequency of two phonetic segments; these segments are rare, so a range of 0..9 is sufficient. The ratio of actual to possible vectors in the space is $24/100 = 0.24$, that is, the vectors occupy 24% of the data space. If one analyzes the 24 speakers in terms of three phonetic segments, the ratio of actual to possible vectors is $24/1000 = 0.024$ or 2.4 % of the data space. In the eight-dimensional case, it is $24/100000000$, or 0.00000024 %. A fixed number of vectors occupies proportionately less and less of the data space with increasing dimensionality. In other words, the data space becomes so sparsely inhabited by vectors that the shape of the manifold cannot, in general, be reliably determined.

What about using more data, as proposed earlier? Let's say that 24% occupancy of the data space is judged to be adequate for manifold resolution. To achieve that for the 3-dimensional case one would need 240
vectors, for the 4-dimensional case 2400, and for the 8-dimensional one 24,000,000. This may or may not be possible. And what are the prospects for dimensionalities higher than 8?

4. Solutions

Given that provision of additional data to improve the definition of a sparse manifold is not always possible, the alternatives are either to use it as is and to live with the consequent unreliability, or to attempt to reduce the sparsity. The remainder of the discussion addresses (ii).

Various ways of reducing sparsity exist, such as tf/idf (Robertson 2004), Poisson distribution (Church & Gale 1995), and principal component analysis (Jolliffe 2002). We look at a method that is conceptually simpler than any of these: elimination of low-variance variables.

Classification of documents depends on there being variation in the characteristics of interest --if there is no variation, the documents are identical and cannot be classified. In any classification exercise, therefore, variables with little or no variation can be disregarded. Mathematically, the degree of variation in the values of a variable is described by its variance, that is, by the average deviation of the variable values from their mean. Given, on the one hand, a matrix in which the rows are the data objects and the columns are variables describing those objects, and on the other that the aim is to classify the objects on the basis of the differences among them, the application of variance to dimensionality reduction is straightforward: eliminate low-variance columns from the matrix.
The NECTE matrix $N_{63,156}$ is very sparse, since there are only 63 vectors in a 156-dimensional space, but many of the 156 variables are superfluous and can be eliminated, greatly reducing dimensionality and thus sparsity. The variance for each of the columns of $N_{63,156}$ was calculated, sorted by decreasing magnitude, and plotted; the result is shown in figure 7:

@@ Insert Figure 7 here

The variables to the right of the 80th have such low variance that they can be eliminated. They were, therefore, removed from $N_{63,156}$, resulting in a reduced-dimensionality 63 x 80 matrix $N_{63,80}$. The analysis of this reduced matrix gave the cluster tree shown in Figure 8.

@@ Insert Figure 8 here

Comparison of this tree to the one in Figure 2 shows that the basic cluster structure has remained the same in the sense that the four main clusters and their hierarchical relationship in Figure 2 are replicated in Figure 8. There are, however, some differences of detail.

- There has been considerable rearrangement of speakers in NG1a(i). In addition, one of the two males tlsg54 has moved out of this cluster to NG1b, and several females (tlsg34, tlsg42, tlsg45, tlsg51) have moved in. The net effect is that gender-based subclusters have emerged.

- NG1a(ii) in Figure 8 continues to consist of male and female speakers with additional education or administrative occupations or both. There are, however, a few speaker reassignments: tlsg34 and tlsg53 have moved to NG1a(i), and tlsg52 has come into NG1a(ii) from NG1a(i).
NG1b is now entirely male: the single female tlsg51 in Figure 2 has here moved to NG1a(i). In addition, tlsg56 has moved to NG1a(i), and tlsg50 and tlsg54 have moved into NG1b from NG1a(i).

The overall effect of the dimensionality reduction has been twofold: to reassign a relatively small number of speakers to different clusters, and to tidy up the cluster tree based on the full 156-dimensional data in the sense that gender-based subclusters have emerged in NG1a(i) and NG1b is now entirely male. For further discussion of Figure 8 see Moisl et al. (2006).

Should the reduced-dimensionality analysis be preferred to the full-dimensional one? Tidiness is not an argument in favour of the reduced-dimensionality analysis: nature is not compelled to respect the human predisposition to regularity, and it may be that the fit between phonetic usage and speaker social characteristics in the NECTE corpus is in fact untidy. The argument from data sparsity does, however, provide a principled basis for favouring of the reduced-dimensionality analysis. The discussion in section 3 above has shown that, relative to a fixed number of data items, reducing dimensionality in general improves the definition of the data manifold and thus the quality of any analysis based on it.

5. Conclusion

Exploratory multivariate analysis is a useful tool in corpus-based linguistics research, but sparsity can be a problem on account of poor definition of the data manifold in its vector space when the data is high-dimensional and the number of data items is relatively small. In such a
situation the data dimensionality should be reduced as much as possible consistent with the need to describe the domain of interest adequately, since dimensionality reduction can be expected to improve the manifold definition and hence the reliability of analytical results.

6. References


