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

Hermann Moisl

University of Newcastle upon Tyne

Introduction

The proliferation of computational technology has generated an explosive production of electronically encoded information of all kinds. In the face of this, traditional philological methods for search and interpretation of data have been overwhelmed by volume, and computational methods have been developed in an attempt to make the deluge tractable. These developments have clear implications for corpus-based linguistics in general, and for corpus-based study of language variation in particular. As more and larger electronic corpora become available, effective analysis of them will increasingly be tractable only by adapting the interpretative methods developed by the statistical, information retrieval, pattern recognition, 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 some solutions.

1. ***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 the linguistic characteristics of a single speaker --phonetic, phonological, morphological, lexical, or syntactic-- can the documents and thus the speakers be classified on the basis of those characteristics? This kind of question can be answered using an empirical methodology known as exploratory multivariate analysis (Andrienko & Andrieko 2005).

1.1 *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. Take, for example, data in which 100 people are described in terms of a single variable ‘age’. Visual inspection would suffice to classify the people on that variable. If these people are described by two variables ‘age’ and ‘height’, classification becomes more difficult, but visual inspection is probably still sufficient. If, however, they are described by, say, 50 variables (‘income’, ‘eye colour’, etc), classification by visual

inspection becomes intractable for most people. In general, as the number of variables grows, so does the difficulty of conceptualizing the interrelationships of variables on the one hand, and the interrelationships of objects –here people– described by those variables on the other. Exploratory multivariate analysis is a general term for mathematically-based methods for understanding data when it has too many variables for it to be comprehensible via direct inspection.

1.2 *Application to historical dialectology*

Exploratory multivariate analysis methods are intended to classify any given set of objects described by more or less numerous variables. Because this is the kind of research question with which language variation research is often concerned, their extension to corpus analysis is a natural step. To exemplify this extension, we consider the Newcastle Electronic Corpus of Tyneside English (NECTE), a corpus of dialect speech from North-East England (Allen et al. 2006). It includes phonetic transcriptions of 63 interviews together with social data about the speakers, and as such offers an opportunity to study the phonetic dialectology of Tyneside speech of the late 1960s. We have begun that study using exploratory analysis of the transcriptions with the aim of generating hypotheses about phonetic variation among speakers and speaker groups (Moisl et al. 2006). These studies were based on comparison of phonetic profiles associated with each of the NECTE speakers, where a profile is the number of times a given speaker uses each of the phonetic segments in the NECTE transcription

scheme. There are 156 segments, so a speaker profile is described by 156 variables. The 63 speaker profiles are represented as a 63 x 156 matrix N, a fragment of which is shown in Figure 1. The aim is to classify the speakers in accordance with the frequency values in their profiles.

	v1: i	v2: iː	...	v156: ɔ
Speaker 1	23	4	...	7
Speaker 2	3	56	...	4
...	...	...	...	...
Speaker 63	18	35	...	8

Figure 1: NECTE phonetic segment frequency data matrix N

N 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 nested tree.

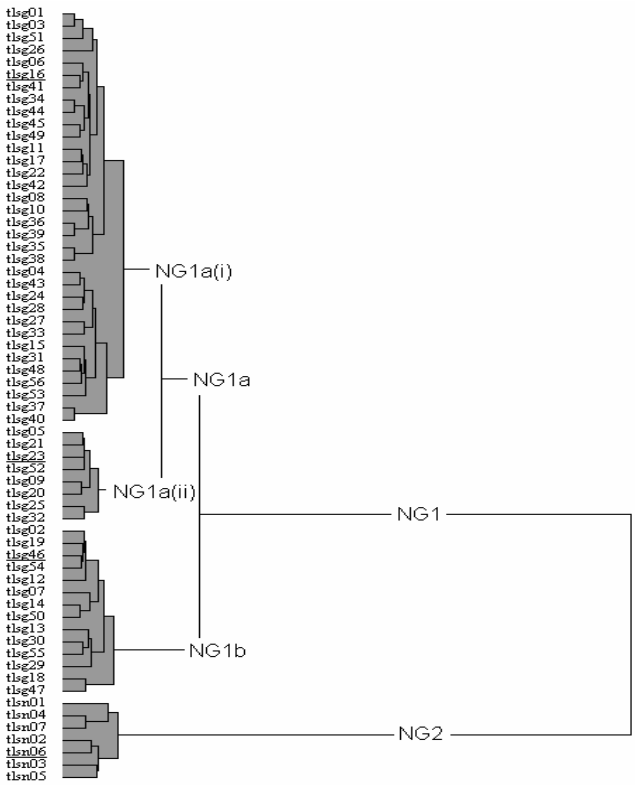

Figure 2: Cluster analysis of the NECTE data matrix N

The hierarchical analysis of N in Figure 2 partitions the NECTE speakers into groups on the basis of their phonetic usage. The main distinction is between middle class speakers from Newcastle on the north side of the river Tyne (NG2) and working class speakers from Gateshead on the south (NG1). The Gateshead speakers are further categorized into NG1b (exclusively male) and NG1b (mainly through not exclusively female). and NG1a is subcategorized into NG1a(i) (working class females) and NG1a(ii) (males and females with relatively higher socioeconomic status).

2. ***The problem of data sparsity***

Sparsity is a major issue in data analysis generally (Verleysen 2003; Verleysen et al. 2003). Why this is so is best explained in terms of a widely used way of representing data: vector space representation. A vector is a sequence of numbers indexed by the positive integers 1, 2, 3...*n*.

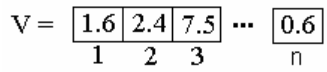

Figure 3: A vector

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:

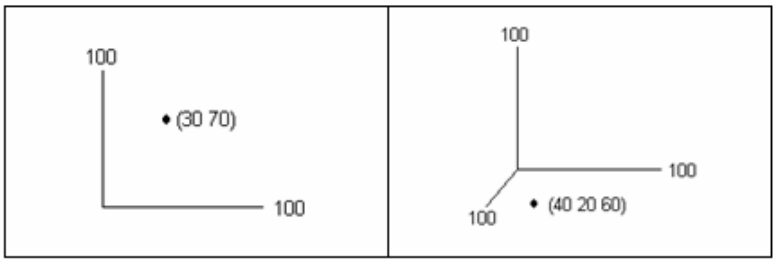

Figure 4: Vectors in 2- and 3-dimensional space

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

Given a data matrix in which the rows are the data items and the columns the variables, that matrix defines a manifold in *n*-dimensional space. The concept 'manifold' comes from mathematical topology (Munkres 2000); for present purposes it can be understood as the shape of data in space. What is the 'shape' of data? Assume a data matrix with 1000 3-dimensional vectors. If these vectors are plotted in 3-dimensional space, they form a cloud of points. Depending on the nature of the interrelationships of the objects that the vectors describe, that cloud might be completely random, or might have some nonrandom structure (ie, Figure 5).

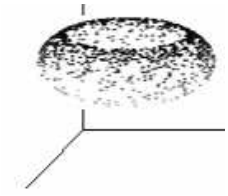

Figure 5: A manifold in 3-dimensional space

The shape defined by 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.

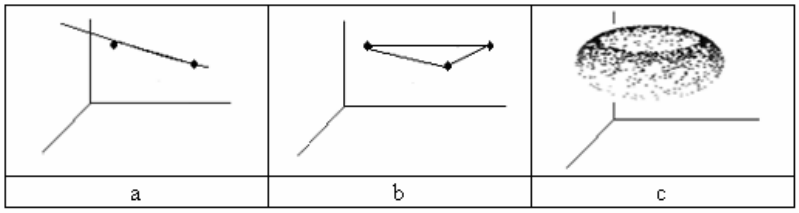

Figure 6: Manifolds in 3-dimensional space

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 of the manifold emerges, as in Figure 6c. The general rule, therefore, is: the more data the better for manifold definition.

Getting enough data can be, and with high-dimensional multivariate data usually is, difficult or even intractable (Bishop 2006:33-8**;** Verleysen 2003; Verleysen et al. 2003). The problem is that the space in which the manifold is embedded grows very quickly with dimensionality and, to retain a reasonable degree of manifold definition, more and more data is required until, equally quickly, getting enough becomes impossible.

Assume an application in which the frequency of each variable is determined for each data item, and that, for simplicity, frequency is always in the range 0..9. Where there are 2 variables, the number of possible 2-dimensional vectors such as (0,9), (3,4), and so on is 10 x 10 = 100. This is the data space. Where there are 3 variables, the number of possible 3-dimensional vectors (0,9,2), (3,4,7) and so on is 10 x 10 x 10 = 1000. For 4 variables the data space is 10 x 10 x 10 x 10 = 10000. In general, the size of

a data space is r^d , where r is the measurement range of the variables (here 0..9) and d the dimensionality. The r^d function generates an extremely rapid increase of data space size with dimensionality: even the modest d = 8 allows for 100,000,000 possible vectors. This is a problem because, the larger the data 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 frequency of usage 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, or, put another way, the vectors occupy 24% of the data space. If we now want to analyze those 24 speakers in terms of their usage 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 = 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 above 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 for any given corpus. And what are the prospects for dimensionalities higher than 8?

3. ***Solutions***

Given that provision of additional data to improve the definition of sparse manifolds is not always a tractable prospect, the remaining alternatives are: (i) to use sparse manifolds for exploratory analysis and to live with the consequent unreliability, or (ii) to attempt to reduce the sparsity. The remainder of the discussion addresses (ii).

Various methods have been developed to reduce sparsity, 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 relatively low-variance variables.

Classification of documents depends on there being variation in the characteristics of interest to the research question --if there is no variation, the documents are identical and cannot be classified relative to one another. Variables describing the characteristics of interest are thus only useful for classification if there is significant variation in the values they take. In any classification exercize, 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 Q 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, then the application of variance to dimensionality reduction is straightforward: eliminate from Q all columns with low variance. The 63 x 156 NECTE matrix N 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 was calculated, sorted by decreasing magnitude, and plotted; the result is shown in figure 7:

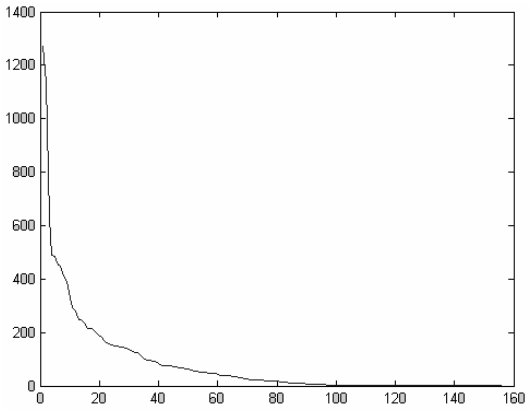

Figure 7: Sorted column variances of the NECTE data matrix N

The variables to the right of –generously—the 80th have such low variance that they can be eliminated from consideration. They were, therefore, removed from N, resulting in a reduced-dimensionality 63 x 80 matrix. The analysis of this reduced matrix gave the cluster tree shown in Figure 2.

Conclusion

The discussion began by observing that (i) as more and larger electronic corpora become available for the study of language variation, effective analysis of them will increasingly be tractable only by using

mathematically and statistically based interpretative methods, and (ii) to use such methods effectively, issues that arise with respect to the abstraction of data from corpora have to be understood. Data sparsity is such an issue. The discussion was in three main parts. The first part showed how a particular class of computational methods, exploratory multivariate analysis, can be used in language variation research, the second explained why data sparsity can be a problem in such analysis, and the third outlined some solutions. The conclusion is that exploratory analysis of any linguistic corpus in which the data is high-dimensional must reduce the data matrix dimensionality as much as possible consistent with the need to describe the corpus adequately.

References

Allen W., Beal J., Corrigan K., Maguire W., Moisl, H. 2006. "A linguistic 'time capsule': the Newcastle Electronic Corpus of Tyneside English". *Creating and Digitizing Language Corpora, Volume. 2: Diachronic Databases*, ed. J. Beal, K. Corrigan, H. Moisl, 16-48. Basingstoke, UK: Palgrave Macmillan.

Andrienko, N., Andrienko, G. 2005. *Exploratory Analysis of Spatial and Temporal Data: A Systematic Approach*. Berlin and Heidelberg: Springer Verlag.

Bishop, C. 2006. *Pattern Recognition and Machine Learning*. Berlin and Heidelberg: Springer Verlag.

Church, K., Gale, W. 1995. "Poisson mixtures". *Natural Language Engineering* 1. 163-90.

Everitt, B., Landau, S., Leese, M. 2001. *Cluster Analysis*, 4th ed. London: Arnold.

Jolliffe, I. 2002. *Principal Component Analysis*, 2nd ed. Berlin and Heidelberg: Springer Verlag.

Moisl, H., Maguire W., Allen W. 2006. "Phonetic variation in Tyneside: exploratory multivariate analysis of the Newcastle Electronic Corpus of Tyneside English". *Language Variation. European Perspectives*, ed. F. Hinskens, 127-141. Amsterdam: John Benjamins.

Munkres, J. 2000. *Topology*, 2nd ed. New Jersey: Pearson Education International.

Robertson, S. 2004. "Understanding inverse document frequency: on theoretical arguments for IDF". *Journal of Documentation* 60. 503-520.

Verleysen, M. 2003. "Learning high-dimensional data". *Limitations and future trends in neural computation*, ed. S. Ablameyko, L. Goras, M. Gori, V. Piuri, V., 141-162. Amsterdam: IOS Press.

Verleysen, M., François, D., Simon, G., Wertz, V. 2003. "On the effects of dimensionality on data analysis with neural networks". *International Work-Conference on Artificial and Natural Neural Networks*, ed. J. Mira, 105-112.

.