Exploratory multivariate analysis

Hermann Moisl

Department of English Literature, Language, and Linguistics

University of Newcastle University Newcastle upon Tyne NE1 7RU

United Kingdom

Phone: (44) 0191 222 7781

Email: hermann.moisl@ncl.ac.uk

Keywords:

1. Exploratory multivariate analysis

2. Cluster analysis

3. Data creation

4. Data representation

5. Data sparsity

6. Intrinsic dimensionality

7. Data dimensionality reduction

8. Nonlinearity

1. Introduction

The proliferation of computational technology has generated an explosive production of electronically encoded information of all kinds. In the face of this, traditional paper-based methods for search and interpretation of data have been overwhelmed by sheer volume, and a variety of computational methods have been developed in an attempt to make the deluge tractable. As such methods have been refined and new ones introduced, something over and above tractability has emerged --new and unexpected ways of understanding the data. The fact that a computer can deal with vastly larger datasets than a human is an obvious factor, but there are two others of at least equal importance. One is the ease with which data can be manipulated and reanalyzed in interesting ways without the often prohibitive labour that this would involve using manual techniques, and the other is the extensive scope for visualization that computer graphics provide.

These developments have clear implications for corpus linguistics. On the one hand, large electronic corpora potentially exploitable by the linguist are being generated as a by-product of the many kinds of daily IT-based activity worldwide, and, on the other, more and more application-specific electronic linguistic corpora are being constructed. Effective analysis of such corpora will increasingly be tractable only by adapting the interpretative methods developed by the statistical, computational linguistics, information retrieval, data mining, and related communities.

The present chapter deals with one type of analytical tool: exploratory multivariate analysis. The discussion is in six main parts. The first part is the present introduction, the second explains what is meant by exploratory multivariate analysis, the third discusses the characteristics of data and the implications of these characteristics for generation and interpretation of analytical results, the fourth gives an overview of the various exploratory analytical methods currently available, the fifth reviews the application of exploratory multivariate analysis in corpus linguistics, and the sixth is a select bibliography. The material is presented in an intuitively accessible way, avoiding formalisms as much as possible. However, in order to work with multivariate analytical methods some background in mathematics and statistics is indispensable.

2. Exploratory multivariate analysis

Observation of nature plays a fundamental role in science. In current scientific method, a hypothesis about some natural phenomenon is proposed and its adequacy assessed using data obtained from observation of the domain of inquiry. But nature is dauntingly complex, and there is no practical or indeed theoretical hope of being able to observe even a small part of it exhaustively. Instead, the researcher selects particular aspects of the domain for observation. Each selected aspect is represented by a variable, and a series of observations is conducted in which, at each observation, the values for each variable are recorded. A body of data is thereby built up on the basis of which a hypothesis can be assessed. One might choose to observe only one aspect -the height of individuals in a population, say- in which case the data consists of more or less numerous values assigned to one variable; such data is univariate. If two values are observed -say height and weight- then the data is bivariate, if three trivariate, and so on up to some arbitrary number *n*; any data where *n* is greater than 1 is multivariate.

As the number of variables grows, so does the difficulty of understanding the data, that is, of conceptualizing the interrelationships of variables within a single data item on the one hand, and the interrelationships of complete data items on the other. Multivariate analysis is the computational use of mathematical and statistical tools for understanding these interrelationships in data.

Numerous techniques for multivariate analysis exist. They can be divided into two main categories which are often referred to as 'exploratory' and 'confirmatory'. Exploratory analysis aims to discover regularities in data which can serve as the basis for formulation of hypotheses about the domain of interest. Such techniques emphasize intuitively accessible, usually graphical representations of data structure. Confirmatory multivariate analysis attempts to determine whether or not there are significant relationships between some number of selected independent variables and one or more dependent ones. These two types are complementary in that the first generates hypotheses about data, and the second tries to determine whether or not such hypotheses are valid. Exploratory analysis is naturally prior to confirmatory; this chapter focuses on the former.

On multivariate analysis in general, see for example Everitt / Dunn (2001), Gordon (1999), Grimm / Yarnold (1995, 2000), Hair et al. (1998), Kachigan (1991), Tinsley / Brown (2000), Tabachnick / Fidell (2006).

3. Data

Data is ontologically different from the world. The world is as it is; data is an interpretation of it for the purpose of scientific study. The weather is not the meteorologist's data –measurements of such things as air temperature are. A text corpus is not the linguist's data – measurements of such things as average sentence length are. Data is constructed from observation of things in the world, and the process of construction raises a range of issues that determine the amenability of the data to analysis and the interpretability of the analytical results. The importance to exploratory multivariate analysis of understanding such data issues can hardly be overstated. On the one hand, 'however powerful the exploring tools, or aggressive the explorer, nothing can be discovered that is beyond the limits of the data itself' (Pyle (1999:46)). On the other, failure to understand relevant characteristics of data can lead to results and interpretations that are distorted or even worthless. For these reasons, an account of data issues is given before moving on to exploratory multivariate methods.

3.1 Variable selection

Given that data is an interpretation of some aspect of the world, what does such an interpretation look like? It is a description of the selected aspect in terms of variables. A variable is a symbol, and as such is a physical entity with a conventional semantics, where a conventional semantics is understood as one in which the designation of a physical thing as a symbol together with the connection between the symbol and what it represents are determined by agreement within a community. The symbol 'A', for example, represents the phoneme /a/ by common assent, not because there is any necessary connection between it and what it represents. Since each variable has a conventional semantics, the set of variables chosen to describe a domain of inquiry constitutes the template in terms of which the domain is interpreted. Selection of appropriate variables is, therefore, crucial to the success of any data analysis.

Which variables are appropriate in any given case? That depends on the nature of the research. Data can only be created in relation to a research question that provides an interpretative orientation in the domain of interest. Without such an orientation, how does one know what to observe, what is important, and what is not? The fundamental principle in variable selection is that the variables must describe all and only those aspects of the domain that are relevant to the research question. In general, this is an unattainable ideal. Any domain can be described by an essentially arbitrary number of finite sets of variables; selection of one particular set can only be done on the basis of personal knowledge of the domain and of the body of scientific theory associated with it, tempered by personal discretion. In other words, there is no algorithm for choosing an optimally relevant set of variables for a research question.

3.2 Data representation

If they are to be analyzed using mathematical methods, the selected variables need to be mathematically represented. A widely used way of doing this is vector space representation (Belew (2000:86-7), Lebart / Rajman (2000), Manning / Schütze (1999:539-44), Pyle (1999:202-22), Salton et al. (1975), Salton / McGill (1983:ch. 4). A vector is a sequence of scalars indexed by the positive integers 1, 2, ...*n*, where a scalar is a single number:

V =	1.6	2.4	7.5	0.6
	1	2	3	4

Figure 1: A vector

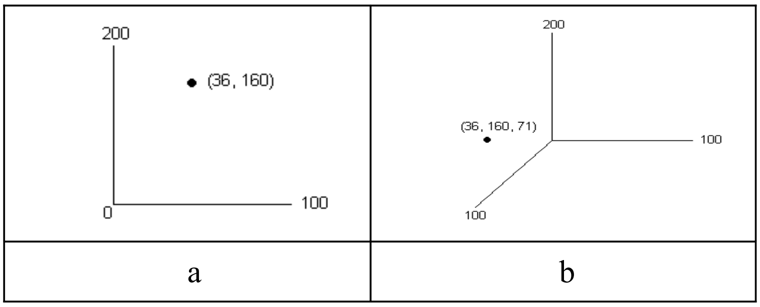
A vector space is a geometrical interpretation of a vector in which

i. the dimensionality of the vector, that is, its index length *n*, defines an *n*-dimensional space. There are various possible types of space, but for present purposes space is taken to be the Euclidean one familiar from elementary geometry, in which the axes are straight lines at right angles to one another.

ii. the sequence of scalars comprising the vector specifies coordinates in the space. These coordinates are relative to the scales of the axes.

iii. the vector itself is a point at the specified coordinates in the space.

For example, the two components of a vector *v* = [36 160] are coordinates of a point in a two-dimensional space, and those of *v* = [36 160 71] of a point in three-dimensional space:



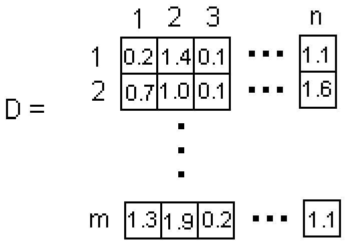
a

b

Figure 2: 2 and 3-dimensional vector spaces

A length-4 vector defines a point in 4-dimensional space, and so on to any dimensionality *n*. Mathematically there is no problem with spaces of dimension greater than 3. The only problem lies in the possibility of visualization and intuitive understanding. As the number of variables and thus dimensions grows beyond 3, graphical representation and intuitive comprehension of it become impossible: who can visualize points in a 4-dimensional space, not to speak of a 40-dimensional one? It often helps to keep in mind that mathematical dimension has no necessary connection with the three dimensions of the physical world.

Data typically consists of more or less numerous data items each of which is described in terms of the selected variables. Where vector space representation is used, each data item is described by a vector, and the data is consequently a collection of vectors. Such a collection is conveniently represented as a matrix in which the rows are the data items and the columns the variables. Thus, data consisting of *m* items each of which is described by *n* variables is represented by an *m* x *n* matrix D in which D_i (for $i = 1...m$) is the *i*'th data item, D_j (for $j = 1..n$) is the *j*'th variable, and D_{ij} the value of variable *j* for data item *i*.



$$D = \begin{array}{c|cccc|c} & 1 & 2 & 3 & & n \\ 1 & 0.2 & 1.4 & 0.1 & \cdots & 1.1 \\ 2 & 0.7 & 1.0 & 0.1 & \cdots & 1.6 \\ & & \vdots & & & \\ m & 1.3 & 1.9 & 0.2 & \cdots & 1.1 \end{array}$$

Figure 3: A matrix

3.3 Variable value assignment

The semantics of each variable determines a particular interpretation of the domain of inquiry; the domain is 'measured' in terms of the semantics, and that measurement constitutes the values of the variables. Measurement is fundamental in the creation of data because it makes the link between data and the world, and thus allows the results of data analysis to be applied to understanding of the world (Pyle (1999:ch.2)).

Measurement is only possible in terms of some scale. There are various types of measurement scale, and these are discussed in the relevant textbooks (Hair et al. (1998:6-9), Pyle (1999:ch.2)), but for present purposes the main dichotomy is between numeric and non-numeric. The multivariate methods referred to in due course assume numeric measurement, and for that reason the same is assumed in what follows.

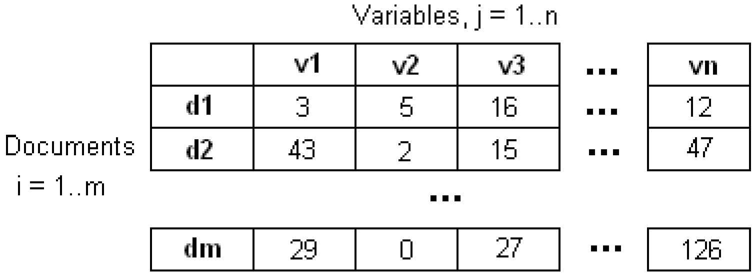
3.4 Data transformation

Once the data has been constructed, it may be necessary to transform it in various ways prior to analysis. Discussion of these in the abstract can quickly become intangible. For forestall this, a specific case is assumed: that the corpus being analyzed is a collection D of some number *m* of documents --of Middle English texts, say-- and the research aim is to classify them on the basis of relative frequency of lexical types that they contain. The data abstracted from D is an *m* x *n* matrix Q in which

i. each of the rows Q_i, for *i* = 1..*m*, represents a single data item --in this case, a document D_i.

ii. each column *j*, for *j* = 1..*n*, is one of *n* variables, each representing a single lexical type that occurs at least once in D. 'Lexical type' is here defined as an abstraction over a set of identical lexical tokens, 'lexical token' as a string of alphanumeric symbols, and 'abstraction' as a set label; the lexical type CAT = {*x* | *x* = 'cat'}, for example. On the type-token distinction see Manning / Schütze (1999:21-3,124-130), Palmer (2000), and the discussion by Baroni in chapter 39 of this Handbook.

iii. the matrix elements Q_{ij} contain integers representing the frequency of lexical type *j* in document *i*. Each of the *m* rows in Q is therefore a frequency profile for a single text.



Variables, j = 1..n

Documents i = 1..m

	v1	v2	v3	...	vn
d1	3	5	16	...	12
d2	43	2	15	...	47
		...			
dm	29	0	27	...	126

Figure 4: Lexical frequency data matrix

Obviously, this example is only one of many possibilities. The data items / matrix rows might be informants in a sociolinguistic or dialectological survey and the variables / matrix columns phonetic segments, or the rows might be phonetic segments and the columns phonetic features like voicing, and so on. The lexical frequency example was selected because it is generic with respect to a wide range of possible applications.

3.4.1 Adjustment for variation in document length

Documents in collections often vary in length. If the variation is substantial, the data abstracted from the collection must be adjusted to avoid distorted results. To see why, assume that all the documents are in the same language, and that, in this language, a given lexical type *j* has probability p_j of occurring. Then, the longer the document, the more likely it is that *j* will occur one or more times: if p_j is 0.01, then on average *j* will occur once every 100 words, twice every 200, and so on. Now, say that *j* occurs 10 times in two documents in D, d_i and d_k. Knowing only this, one would naturally judge that, in terms of their usage of *j*, the two documents are identical and that *j* is consequently of no use in distinguishing d_i from d_j. If, however, one also knows that d_i is 1000 words long, and d_k only 500, this is no longer the case. The frequency for d_i is what one would expect given p_j, but d_k uses *j* with higher-than-expected frequency, and this disparity can be used in distinguishing d_i from d_k: d_k, unlike d_i, is especially interested in what *j* denotes.

What is required is some way of adjusting the data matrix so that not just frequency but its significance relative to document length can be represented and thus incorporated into subsequent analysis. One approach is to transform the rows of the matrix into vectors of length 1:

$$Q_i = \frac{Q_i}{|Q_i|}$$

where $|Q_i|$ is the norm or length of row vector Q_i, defined as:

$$|Q_i| = \sqrt[2]{Q_{i1}{}^2 + Q_{i2}{}^2 + \ldots + Q_{in}{}^2}$$

The effect is to adjust the vector values representing document Q_i in proportion to the length of Q_i: the greater the length, the smaller the result of the division and thus the smaller the post-adjustment values in Q_i, and vice versa as the Q_i vector length shortens. This adjustment is part of a method for measuring relative proximity of document vectors in Information Retrieval called cosine normalization. Another approach is to transform the row vectors in relation to the average length of documents in the collection D:

$$Q_i = Q_i\ (\mu\ /\ l_{Qi})$$

where (i) Q_i is the *i*'th document's lexical frequency profile in the data matrix Q, (ii) l_{Qi} is the total number of lexical tokens in document Q_i, and (iii) μ is the mean document length in terms of lexical tokens across all documents d ε D, so that:

$$\mu = \sum_{i = 1..m}(l_{Qi})\ /\ m$$

Thus, the values in each lexical frequency profile vector Q_i are multiplied by the ratio of the average number of lexical tokens per document across the collection D to the number of tokens in Q_i. The longer the document the numerically smaller the ratio, and vice versa; the effect is therefore to decrease the values in the vectors that represent long documents, and increase them in vectors that represent short ones, relative to average document length.

On transformation of data relative to document length see Belew (2000:89-92), Lebart / Rajman (2000:477-505), Singhal et al. (1996).

3.4.2 Sparsity minimization

Sparsity is a major issue in data analysis generally. The concept of the manifold is central to understanding of why this is so. It comes from mathematical topology (Munkres (2000)), a branch of pure mathematics concerned with geometrical properties; for present purposes it can be understood as the shape of data in *n*-dimensional space. What is the 'shape' of data (Pyle (1999:84-6))? Consider a reasonably large data set of, say, 1000 3-dimensional real-valued vectors, no two of which are identical. If these vectors are plotted in 3-dimensional space, they form a cloud of points with an identifiable shape within the general space, as in Figure 5:

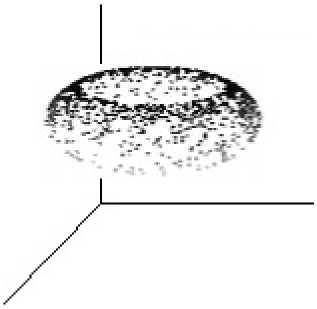


Figure 5: A manifold in 3-dimensional space

That shape is a manifold. The idea extends directly to any dimensionality, though such general spaces cannot be shown graphically. 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, it is intuitively clear that there have to be enough data points to give it adequate definition. If, as in Figure 6a, there are just two points, the only reasonable manifold to propose is a line; any number of alternative manifolds are, of course, possible --the two points could come from a far more complex manifold like Figure 6c-- but to propose this on the basis of just two points would clearly be unjustified.

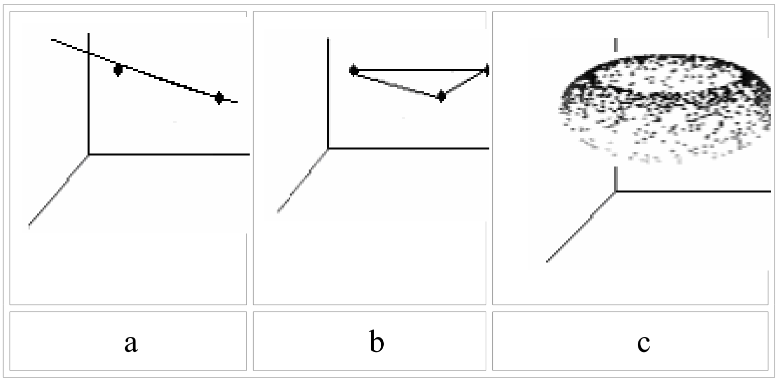


Figure 6: Degrees of manifold definition

Where there are 3 points a plane would be reasonable, as in Figure 6b. But it is only as the number of data points grows that the true shape of the manifold emerges, as in 6c. The general rule, therefore, is: the more data the better. After a certain point, increasing the amount of data becomes redundant in the sense that it simply confirms an already-clear manifold shape, but it doesn't do any harm.

In dealing with high-dimensional data, however, having too much is rarely a problem. Quite the opposite --the usual situation with high-dimensional data is that there is far too little. High-dimensional spaces are inherently sparse, and, to achieve adequate definition of the data manifold, the amount of data required very rapidly becomes intractably large; this phenomenon is known as the 'curse of dimensionality'. To see the problem, consider three data sets each of which contains 10 items, no two of which are identical:

i. Set 1 is univariate, and the single variable can take integer values in the range 1..10. The ratio of data points to possible values is 10/10 = 1, that is, the data points completely fill the data space.

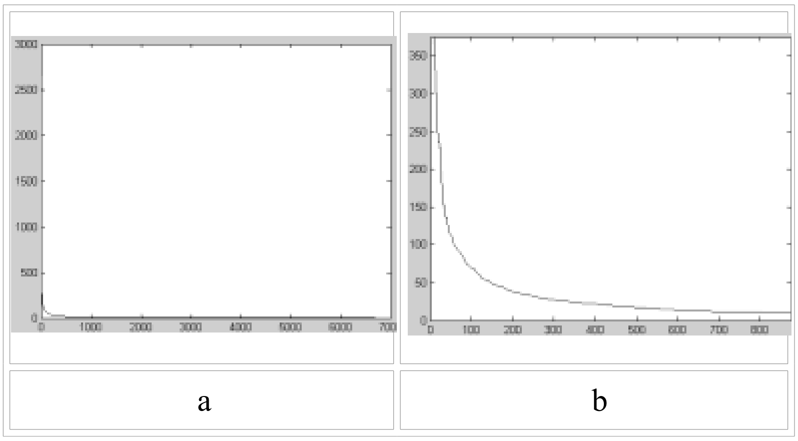
ii. Set 2 is bivariate, and each of the two variables can take integer values in the range 1..10. The ratio of data points to possible value pairs is 10 / (10 x 10) = 0.1, that is, the data points occupy 10% of the data space.

iii. Set 3 is trivariate, and each of the three variables can take integer values in the range 1..10. The ratio of data points to possible value triples is 10 / (10 x 10 x 10) = 0.01, that is, the data points occupy 1% of the data space.

And so on for increasing dimensionality: for a data set of fixed size d, the ratio of actual to possible points in the data space is d / r^n, where r is the number of different values that each variable can take (assuming for simplicity that all variables are identical in this respect). In other words, as dimensionality increases, the ratio of actual to possible points in the data space decreases at an exponential rate. In principle, therefore, a manifold consisting of some fixed number of vectors very rapidly becomes sparser as the dimensionality of the space in which it is embedded grows; to maintain its resolution at any preferred ratio, the number of vectors required must therefore grow exponentially with the dimensionality. Getting enough data becomes a serious problem even at relatively low dimensionalities, and an insuperable one soon thereafter. In practice the problem is not as severe as all this might suggest, since a typical real-world data set is not in general evenly or randomly spread around its vector space, but rather tends to be concentrated in one or more distinct regions of the space. Dimensionality nevertheless remains a potential problem for data analysis in any given application, and the moral is that dimensionality should be kept as low as possible consistent with the need to describe the domain of inquiry adequately.

For discussion of issues relating to high-dimensional data see Bishop (1995:chs.1,8), Pyle (1999:ch.2,355-60,424-34), Verleysen (2003), Verleysen et al. (2003).

Data sparsity has a particular relevance in corpus linguistics because the object of study is spoken or written natural language, and lexical distribution in samples of natural language have a characteristic shape. This shape is exemplified in a plot of lexical types in the Qur'an. The frequencies of these types were calculated, sorted into descending order of magnitude, and plotted:



a

b

Figure 7: Frequencies of lexical types in the Qur'an

Figure 7a is the full plot, and 7b is a zoomed-in region near the origin to display the shape of that region more clearly. There is a relatively small number of very frequent types, a moderate number of moderately frequent types, and a large number of very infrequent ones. This distribution is characteristic of lexical frequency distributions in natural language text generally (Baayen (2001); see also Manning / Schütze (1999:20-29) and Baroni's discussion in chapter 39 of this Handbook), and its shape remains pretty much constant even for natural language text corpora many orders of magnitude larger than the Qur'an: the number of the few very frequent types continues to grow quickly as the corpus size grows and the number of moderately frequent types continues to grow moderately quickly, but the frequencies of the very infrequent types change hardly at all --instead, more and more types are added to the list. It is therefore clear that, in corpus linguistics studies where lexical frequency plays a role, the data will in general be very sparse on account of the large number of infrequent lexical type variables.

The obvious solution to sparsity is to select an optimal set of variables at the data design stage, but this is more easily said than done. As already noted, there is no algorithm for choosing an optimally relevant and therefore minimal set of variables for a research question, and therefore no way of knowing a priori whether the dimensionality of a given data set is as low as it can be. Because of this, a range of methods for transforming data matrices so as to reduce their dimensionality has been developed, and, in cases where the data is sparse, application of one or more or these methods can very substantially improve analytical results by giving the manifold better definition.

a) Stemming

Fundamental to the morphologies of many languages is the process whereby prefixes and suffixes are attached to lexical stems, and/or the lexical stems themselves mutated in some way, in order to mark syntactic function or some modification to the primitive semantic denotation of the stem. Document collections written in such languages typically contain more or less numerous morphological variants of primitive lexical stems. Such variants can be considered to be equivalent for purposes of text analysis and information retrieval; stemming is the reduction of morphological variants to their common primitive stem.

Where lexical frequency plays a role in data creation, stemming offers scope for substantial dimensionality reduction. If a lexical type is defined as the set of identical alphabetic strings, then each variant of a given stem is treated as a distinct lexical type and assigned a column in the data matrix. If, however, all the morphological variants of a stem are collapsed into an equivalence class which then constitutes the lexical type, so that, for example, the type CAT = {*x* | *x* = a morphological variant of 'cat'} such as 'cats', 'catty', 'cattery' and so on, the number of types and thus columns of the frequency matrix can be more or less substantially reduced, depending on the morphological characteristics of the language in question. The frequency of a lexical type so defined in the frequency matrix is then the sum of the frequencies of the aggregated variants.

At first glance it might seem that creation of such equivalence classes loses information, and that this loss is bound adversely to affect the validity of analyses based on the data. Just the opposite is true, however. If lexical types are regarded as sets of identical tokens, then each type is represented as a separate variable column in the data matrix, and all columns are treated equally in the analytical methods cited later. The implication is that morphologically related tokens are treated exactly the same as unrelated ones. In other words, there is no distinction between the semantic distances among morphological variants of a single stem on the one hand, and those between unrelated stems on the other --the semantic difference between 'administer' and 'administration' is taken to be the same as that between 'administer' and 'cow'. If, as here, the aim is to classify documents on the basis of their lexical semantics, this is bound to distort the data and thus the analytical results based on it. Creation of equivalence classes based on morphological relatedness eliminates this distortion.

On stemming algorithms and their application in computational text processing see Frakes (1992) and Hull (1996).

b) Variable selection

A seminal principle in Information Retrieval, extensively confirmed by empirical results, is that not all lexical types in a document collection are equally useful in document classification (for example, Belew (2000:ch.3), van Rijsbergen (1979:ch.2), Salton / McGill (1983:ch.3)). Various ways of identifying relatively more useful variables exist, and this section gives an overview of some of the most often used ones. The focus is on lexical frequency in document collections, but the techniques are straightforwardly applicable to other kinds of data, and thus to a wide range of analyses in corpus linguistics.

i. Dimensionality reduction based on lexical type frequency

Luhn, one of the founders of modern Information Retrieval, proposed that the relative frequency of lexical types in a document collection is a fundamental criterion for classifying documents relative to one another (discussed in Belew (2000:76ff), van Rijsbergen

(1979:15ff), Salton / McGill (1983:60-63)). The intuition underlying this is simple: if an author uses a word repeatedly in a text, then the text is more likely to be about what the word denotes than it is to be about the denotation of a word that is infrequently used; documents with similar lexical frequency profiles are classified together and distinguished from those whose profiles are different. Luhn also observed, however, that the usefulness of a lexical type for document classification does not increase monotonically with frequency, and more specifically that very frequent types on the one hand and very infrequent ones on the other are less useful for the purpose than medium frequency ones. He therefore proposed that both very infrequent and very frequent words be discarded. Substantial dimensionality reduction can be achieved in this way, but Luhn did not provide any clear criteria for determining upper and lower frequency thresholds, and there is consequently the ever-present danger that too many or too few types will be eliminated, thus compromising classification based on the set of retained variables.

ii. Dimensionality reduction based on variance

As we saw in the foregoing discussion of data, any variable *x* is an interpretation of some aspect of the world, and a value assigned to *x* is a measurement of the world in terms of that interpretation. If *x* is to describe more than one object --the heights of 1000 people, say-- then it must take values characteristic of each person.

Unless all 1000 people are exactly the same height, these values will vary. This possibility of variation gives *x* its descriptive utility: a constant value for *x* says that what *x* represents in the world does not change, moderate variation in the value says that that aspect of the world changes only a little, and widely differing values that it changes substantially. In general, therefore, the possibility of variation in the values assigned to variables is fundamental to the ability of variables to represent reality.

Classification of documents or of anything else therefore depends on there being variation in their characteristics. When the objects to be classified are described by variables, then the variables are only useful for the purpose if there is significant variation in the values that it takes. If, for example, a large random collection of people was described by variables like height, weight, and income, there would be substantial variation in values for each of them, and they could legitimately be used to classify the people in the sample. On the other hand, a variable like 'has nose' would be effectively useless, since, with very few exceptions, everyone has a nose --there would be almost no variation in the boolean value 1 for this variable. In any classification exercize, therefore, one is looking for variables with substantial variation in their values, and can disregard variables with little or no variation.

Mathematically, the degree of variation in the values of a variable is described by its variance. The variance of a set of variable values is the average deviation of those values from their mean. Assume a set of *n* values $\{x_1, x_2 ... x_n\}$ assigned to a variable *x*. The mean of these values μ is $(x_1 + x_2 + ... + x_n) / n$. The amount by which any given value x_i differs from μ is then $x_i - \mu$. The average difference from μ across all values is therefore $\Sigma_{i=1..n} (x_i - \mu) / n$. This average difference of variable values from their mean almost but not quite corresponds to the definition of variance. One more step is necessary, and it is technical rather than conceptual. Because μ is an average, some of the variable values will be greater than μ, and some will be less. Consequently, some of the differences $(x_i - \mu)$ will be positive and some negative. When all the $(x_i - \mu)$ are added up, as above, they will cancel each other out. To prevent this, the $(x_i - \mu)$ are squared. The standard definition of variance for *n* values $\{x_1, x_2 ... x_n\}$ assigned to a variable *x*, therefore, is:

$$v = (\sum_{i=1..n} (x_i - \mu)^2) / n$$

Given a data matrix Q in which the rows are cases and the columns are lexical type variables describing the cases, and also that the aim is to classify the cases on the basis of the differences among them, the application of variance to dimensionality reduction is straightforward: eliminate all variables with low variance, that is, variables whose values do not vary enough for them to be useful in document classification. As with the upper and lower thresholds discussed in the preceding section, this begs the question of how low is too low, that is, of selecting a threshold.

iii. Lexical frequency distribution

Spärck Jones (1972) proposed what was to become a standard principle in Information Retrieval: that a lexical type's usefulness is determined not by its absolute frequency across a collection, but by the pattern of variation in its frequency across the documents. To gain an intuition for this, assume a collection of documents related to the computer industry. At one end of the range are very low frequency words that, as expected, are of little or no use for document classification: a word like 'coffee' that occurs a few times in one or two documents that caution against spills into keyboards is insignificant in relation to the semantic content of the collection as a whole, and a word like 'bicycle' that occurs only once tells us only that the document in which it appears is unique on that criterion. At the other end of the range, a word like 'computer' and its morphological variants is likely to be both very frequent across the collection and to occur in most if not all the documents, and as such is a poor criterion for classifying documents despite its high absolute frequency: if all the documents are about computers, being about computers is not a useful distinguishing criterion. In short, lexical frequency on its own is not a reliable classification criterion. The most useful lexical types are those whose occurrences are both relatively frequent and not, like 'computer', uniformly spread across all collection documents but rather occur in clumps, such that a relatively few documents contain most or all the occurrences, and the rest of the collection few or none; 'debug', for example, can be expected to occur frequently in documents that are primarily about computer programming and compiler design, but only infrequently if at all in those about, say, word processing. On this criterion, lexical types are selected in accordance with their 'clumpiness' of occurrence across documents in a collection.

Three methods used in Information Retrieval for determining clumpiness in data are:

i. TF.IDF ('Term Frequency x Inverse Document Frequency'): Belew (2000:84-5), Buckley (1993), Robertson (2004), Roberston / Spärck Jones (2004), Salton / McGill (1983:63), Spärck-Jones (1972).

ii. Signal-noise ratio: Belew (2000:83-4), Salton / McGill (1983:63-6).

iii. Poisson term distribution: Belew (2000:73 ff), van Rijsbergen (1979:27-9); Church / Gayle (1995a, 1995b).

Space constraints do not permit these to be described here, and the reader is referred to the cited references.

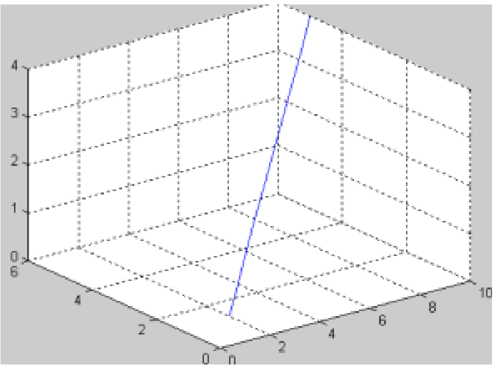
c) Variable redefinition

Dimensionality reduction can be achieved by replacing the variables that have been chosen to describe the domain of interest with different variables that describe the domain as well as, or almost as well as, the originals, but are fewer in number.

We have seen that a data set of *n*-dimensional vectors defines a manifold in *n*-dimensional space. In such a space, it is possible in principle to have manifolds whose dimensionality is *k*, where $k < n$. Consider the 3-dimensional data set in Figure 8a:

v1	**v2**	**v3**
1	0.5	0.4
2	1	0.8
3	1.5	1.2
4	2	1.6
5	2.5	2
6	3	2.4
7	3.5	2.8
8	4	3.2
9	4.5	3.6
10	5	4

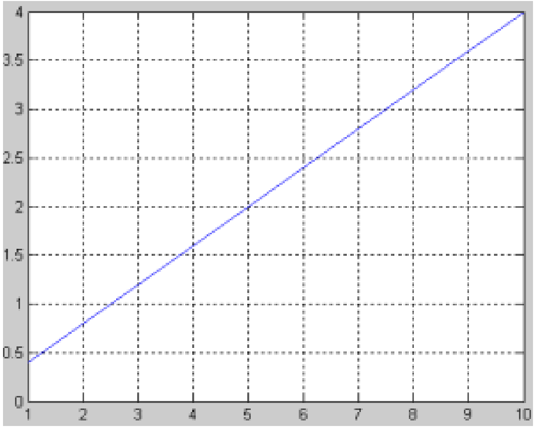
a



b

Figure 8: A one-dimensional manifold in 3-dimensional space

Plotting this data in 3-dimensional space (Figure 8b) shows it to describe a line. But that line can be redescribed in 2 dimensions:

<table>
<tr>
<td>
<table>
<tr><th>v1</th><th>v3</th></tr>
<tr><td>1</td><td>0.4</td></tr>
<tr><td>2</td><td>0.8</td></tr>
<tr><td>3</td><td>1.2</td></tr>
<tr><td>4</td><td>1.6</td></tr>
<tr><td>5</td><td>2</td></tr>
<tr><td>6</td><td>2.4</td></tr>
<tr><td>7</td><td>2.8</td></tr>
<tr><td>8</td><td>3.2</td></tr>
<tr><td>9</td><td>3.6</td></tr>
<tr><td>10</td><td>4</td></tr>
</table>
</td>
<td>


</td>
</tr>
<tr><td>a</td><td>b</td></tr>
</table>

Figure 9: A one-dimensional manifold in 2-dimensional space

In fact, the line can be redescribed in 1 dimension -- its length, 10.63-- by its distance from 0 on the real-number line:

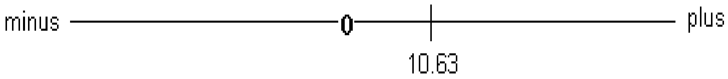


Figure 10: A one-dimensional manifold in 1-dimensional space

Consider another example --a plane in 3-dimensional space:

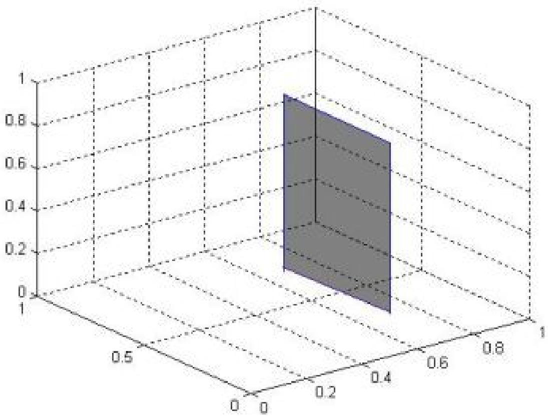


Figure 11: A two-dimensional manifold in 3-dimensional space

This plane can be redescribed in 2-dimensional space

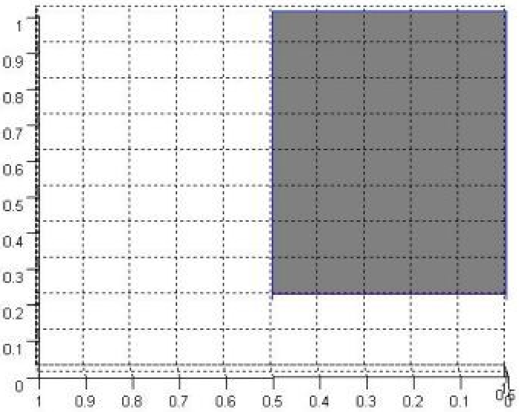


Figure 12: A two-dimensional manifold in 2-dimensional space

And, as usual, this concept extends straightforwardly to any dimensionality.

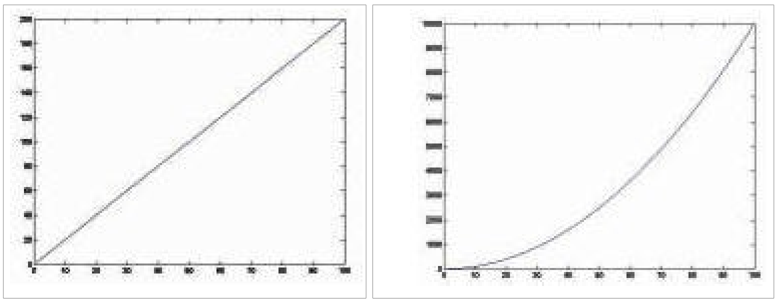
In general, therefore, a line can be described in one dimension, two dimensions, three dimensions, or any number of dimensions one likes. Essentially, though, it is a 1-dimensional object; its 'intrinsic dimensionality' (Verleysen (2003)) is 1. The minimum number of dimensions required to describe a line is 1; higher-dimensional descriptions are possible but redundant. A plane was described in two and three dimensions. Could it also, like a line, be described in one dimension? No: the intrinsic dimensionality of a plane is 2 --the corresponding data set must be 2-dimensional at least, giving the coordinates of the points that describe it.

The concept of intrinsic dimensionality applies straightforwardly to dimensionality reduction. The informational content of data is conceptualized as a *k*-dimensional manifold in the *n*-dimensional space defined by the data variables. Where $k = n$, that is, where the intrinsic dimensionality of the data corresponds to the number of data variables, no dimensionality reduction is possible without significant loss of information. However, the foregoing discussion of data creation noted that, when describing a domain of interest, selection of variables is at the discretion of the researcher. It is therefore possible that the selection of variables in any given application will be suboptimal in the sense that there is redundancy among the variables, that is, that they overlap with one another in terms of the information they represent about the domain; where there is a significant amount of redundancy, it is possible in principle to represent this information using a smaller number of variables, thus reducing the dimensionality of the data. In such a case, the aim of dimensionality reduction of data is to discover its intrinsic dimensionality *k*, for $k < n$, and to redescribe its informational content in terms of those *k* dimensions.

The most often used variable redefinition method is principal component analysis (PCA), on which see Jolliffe (2002); briefer accounts are in Bishop (1995:310-14), Everitt / Dunn (2001:ch.3), Grimm / Yarnold (1995:99-134), Hair et al. (1998:87-138), Oakes (1998:96-108), Tabachnick / Fidell (2006), Webb (2002:319-44), Woods et al. (1986:ch.15). PCA is a particular case of Singular Value Decomposition (SVD) on which see Lebart / Rajman (2000) and Manning / Schütze (1999:554-66). Both PCA and SVD are linear methods; nonlinear variable definition methods are described in Diamantaras / Kung (1996), Bishop (1995:314-19), Pyle (1999:355-83).

3.4.3 Data linearization

In physical systems there is a fundamental distinction between linear and nonlinear behaviour. To get an intuition for what is involved, and why the distinction is important, here is an experiment. Kick a ball and note how far it goes. Kick it again, but this time twice as hard, and once again note how far it goes. The natural expectation is that it will go twice as far, and this expectation is fulfilled. This is linear behaviour: the effect is proportional to the cause. But take the experiment further. Kick the ball in a series, each time twice as hard as the time before: k, 2k, 4k, 8k and so on. If it goes 10 metres for k, and 20 metres for 2k, will it also go 40 metres for 4k, and 80 metres for 8k? No. As it is kicked harder and harder, it goes faster and further. Air resistance becomes a factor at higher speeds, and so does rolling resistance. The ball might only go 78 metres for an 8k kick, and 150 metres for a 16k kick, etc. Eventually, the kick will be so hard that the ball bursts and goes hardly any distance at all. This is nonlinear behaviour: it is the breakdown of proportionality between cause and effect in physical systems, and it can generate a variety of complex and often unexpected --including chaotic-- behaviours. In nature there are few truly linear systems. Nonlinearity pervades the physical world (Bertuglia (2005)), and, because it does, data manifolds that describe the world are likely to contain nonlinearities. Figure 13a shows a linear relationship between two variables *x* and *y*, and figure 13b a nonlinear one:



a

b

Figure 13: Linear and nonlinear 1-dimensional manifolds

In the linear case there is an invariant proportionality between *x* and *y*, and that invariance is represented by a straight line; in the nonlinear case, the relationship between *x* and *y* varies with different values of *x*, and that variance is represented by a curved line. In three dimensions, linear data might generate a plane (Figure 14a) and nonlinear data a curved surface (Figure 14b):

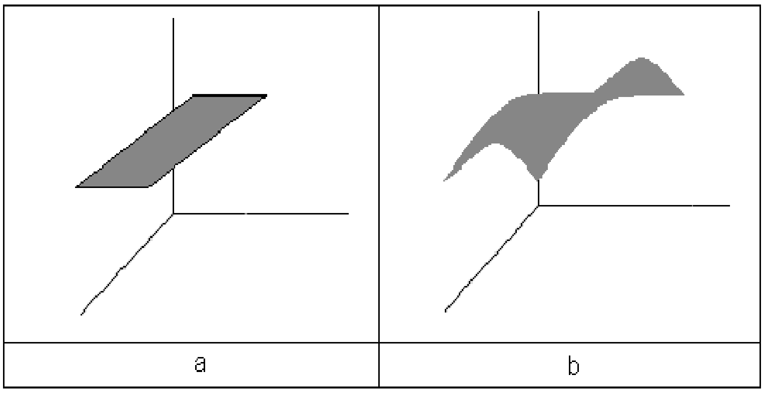


Figure 14: Linear and nonlinear 2-dimensional manifolds

In general, linear manifolds are lines and planes, and nonlinear ones curves and curved surfaces; these cannot be shown graphically for higher dimensionalities. Nonlinear manifolds can range from fairly simple curves, as above, to highly complex ones.

The first step is to determine whether or not a given matrix in fact contains significant nonlinearity. This seems obvious, but, for high-dimensional data, it is not always or even usually straightforward. In the light of the foregoing observation that nonlinearity pervades the natural world, the strong suspicion must be that the generating process is nonlinear, but this is not certain. Even if the generating process is known to be nonlinear, moreover, there is no guarantee that every data set it generates will contain nonlinearities. This sounds paradoxical, but consider the shape of the familiar nonlinear logistic function, which models a range of natural processes:

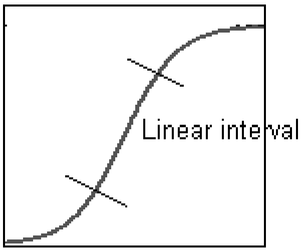
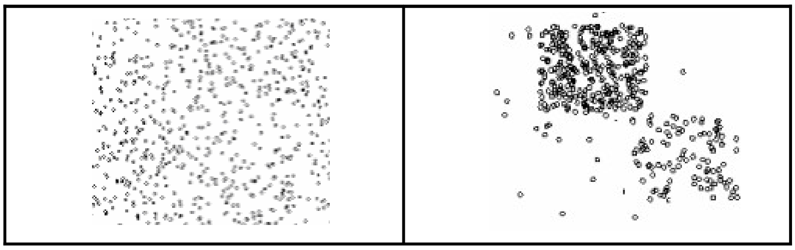


Figure 15: Graph of logistic function

Though it is nonlinear globally, there is a relatively large interval that is linear or near-linear; if the data of interest happens to come from that interval of output values, then it is linear even though it was generated by a nonlinear process. A priori reasoning cannot, in short, establish whether or not a data set contains significant nonlinearities. Only direct examination of the data will establish this. The usual method is to plot pairs of variables and then examine the plots for deviation from linearity, but where the number of variables is large this quickly becomes burdensome (Hair et al. (1998:75- 83), Tabachnick / Fidell (2006)). One alternative is to linearize the data matrix (for example Croft et al. (1992:350-64)). The nonlinearities may, however, themselves be of interest, and linearization throws the baby out with the bath water. Another alternative is to use an analytical method that can accommodate nonlinearities, on which more below.

4. Exploratory multivariate methods

Exploratory methods are essentially variations on a theme: cluster analysis. Cluster analysis aims to identify and graphically to represent nonrandomness in the distribution of vectors in *n*-dimensional space. Spatial regularities in the graphical representations are interpreted as reflecting regularities in the natural process that generated the data, and support hypotheses about the characteristics of the process. In Figure 16a, for example, the vectors are spread more or less uniformly in two-dimensional space; there are some local concentrations, but these are not clearly defined and it is difficult to infer anything about the process that generated the data other than that it appears to be broadly random. In Figure 16b, on the other hand, there are clearly defined concentrations of vectors such that two groups of points are spatially relatively close in each group, and spatially relatively far from each other, which suggests that the generating process is strongly nonrandom.



a	b

Figure 16: Random and nonrandom data

In two or three dimensions, such distributions can be plotted and interpreted by eye. In higher dimensions this is no longer possible, however; the various cluster analysis methods are just different ways of representing nonrandom structure in higher-dimensional data graphically in two or three dimensional space.

There is an extensive range of cluster analysis methods together with a large associated literature: for example Arabie et al. (1992), Duda et al. (2001:ch.10), Everitt / Dunn (2001), Everitt et al. (2001), Gordon (1999), Gore (2000), Grimm / Yarnold (1995, 2000), Hair et al. (1998:ch.9), Jain / Dubes (1988), Jain et al. (1999), Kachigan (1991), Manning / Schütze (1999:ch.14), Oakes (1998:ch.3), Tan et al. (2006:ch.8,9), Tinsley / Brown (2000), Tabachnick / Fidell (2006), Webb (2002:ch.10), Woods et al. (1986:ch.14). There is no hope of describing individual methods in detail here, so what follows gives an overview in three parts. The first part introduces basic issues in cluster analysis, the second cites some commonly used methods, and the third issues a caution about using those methods.

a) Basic issues

The most important thing to realize about cluster analysis is that there is no single 'best' method; see for example Everitt et al. (2001:ch.8), Tan et al. (2006:639-

42). In any particular application, selection of one or more methods must be informed by a variety of considerations, three of the most important of which are:

i. How much is known about the cluster structure of the data?

In cluster analysis there is a distinction between methods which make no a priori assumptions about the structure of given data and attempt to discover clusters purely on the basis of the data's characteristics, and those which presuppose that the data has a cluster structure and require specification of the number of clusters in advance of analysis. If little or nothing is known about the cluster structure, then one of the former methods is appropriate, but if there is a reasonable degree of certainty about its structure then one of the latter type of method, such as k-means clustering or kernel-based adaptive algorithms, can be used (for survey of these methods see Webb (2002)). The present discussion is concerned with exploratory analysis, and as such is henceforth concerned only with methods that make no a priori assumptions about data.

ii. Is the data linear or nonlinear?

The selected method or methods must be compatible with the data being analyzed. For continuous-valued numerical data such as that being discussed here, the main criterion for compatibility is whether the data manifold is linear or not. Data that contains significant nonlinearity must be analyzed using a nonlinear clustering method; use of a linear method in such a case misrepresents the structure of the data to greater or lesser degrees, depending on the nature of the nonlinearity. What does it means for a method to be linear or nonlinear? Assume a curved manifold in *n*-dimensional space. What is the distance d_{ij} between any two points *i* and *j* on that manifold? A linear method measures that distance as a straight line joining the points, ignoring the manifold's curvature, whereas a nonlinear method measures the distance along the surface of the manifold, thereby taking account of the curvature. Depending on the amount of curvature, the difference between the two measures can be significant and can therefore significantly affect analysis based on it. An example is the distance between two points A and B on the perimeter of the circle in Figure 17: the linear distance between them is a chord drawn through the interior, and the nonlinear one the length of the perimeter segment between the points as indicated by the arc in the figure:

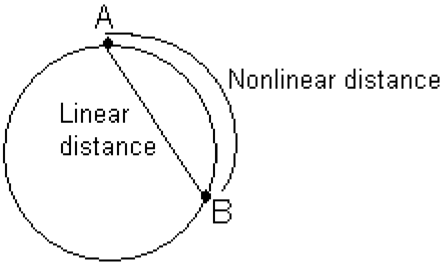


Figure 17: Linear and nonlinear distance

Consider, for example, the problem of discovering a classification for the data in Figure 18a. The data space must be partitioned such that all the points in the left-hand cluster fall into one partition, and all the points in the right-hand cluster into another. Linear methods are, by definition, limited to doing this using straight lines or surfaces; in this case, that is sufficient.

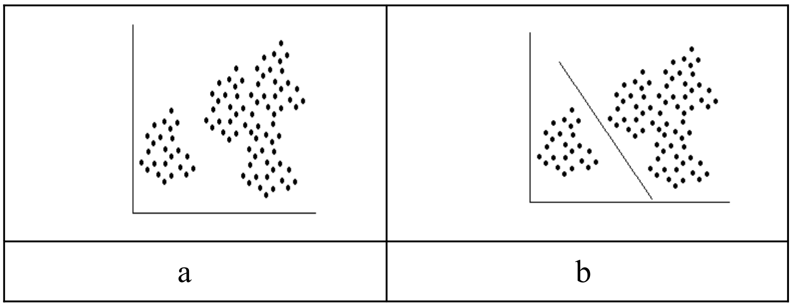


Figure 18: Linearly separable clusters

For the data in figure 19a, however, there is no straight line that can separate the two clusters without misclassifying some of the points, as in 19b. What is required for correction classification is a method for finding a nonlinear partition, as in 19c.

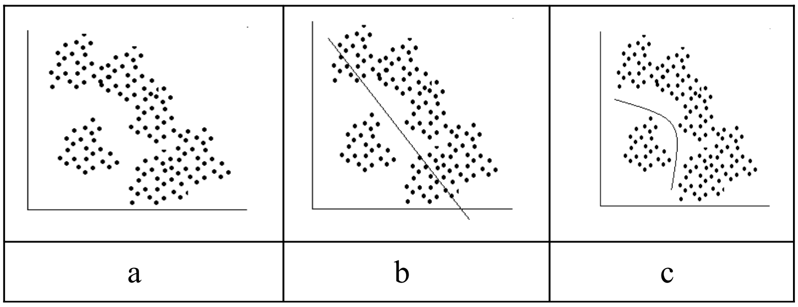


Figure 19: Nonlinearly separable clusters

iii. Is a hierarchical or nonhierarchical analysis required?

The fundamental aim of exploratory analysis is to generate hypotheses about some domain of inquiry, and it may be that, in any particular case, some methods provide representations of structure that do this more usefully than others. The main distinction among methods in this regard is between those that generate hierarchically ordered clusters, and those that do not and are therefore described as nonhierarchical. Nonhierarchical methods generate graphical representations in two or three dimensional space such that, given a suitable measure of proximity, vectors which are spatially or topologically relatively close to one another in high-dimensional space are spatially or topologically close to one another in their two or three dimensional representation, and vectors which are relatively far from one another in high-dimensional space are clearly separated, either by relative spatial distance or by some other graphical means, resulting --in the case of nonrandom data-- in a configuration of well defined clusters. Figures 18 and 19 above are a two-dimensional example; a three-dimensional one might look like Figure 20:

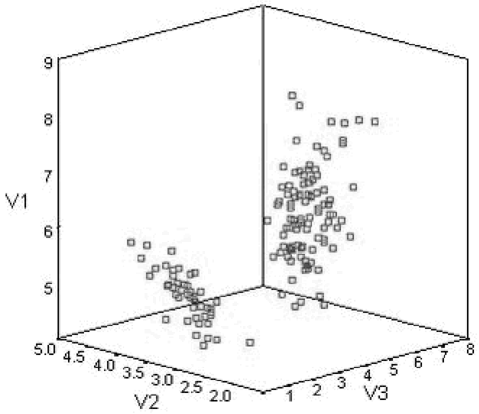


Figure 20: Clusters in 3-dimensional space

Hierarchical methods, on the other hand, represent proximity structure in high-dimensional data not as spatial clusters but as 'dendrograms':

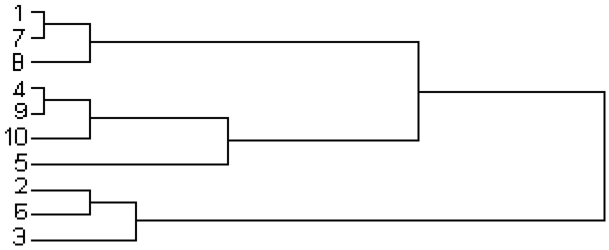


Figure 21: A cluster dendrogram

A dendrogram is simply a tree of the kind linguists are familiar with from sentence structure analysis. It is shown horizontally rather in the vertical orientation that is more usual in linguistics in order to make it more readily representable on a page, and the labels at the 'leaves' are not lexical tokens but labels for the vectors in the data set --'1' is the first vector, '2' the second, and so on. Like a linguistic phrase structure tree, a dendrogram shows constituency structure: in this tree, vectors 1 and 7 constitute a 'phrase' that combined with vector 8 so form a superordinate 'phrase', which itself combines with (4,9,10,5) to form an even higher-level 'phrase', and so on. Unlike a linguistic phrase structure tree, however, this one represents not grammatical constituency but vector proximity in *n*-dimensional space: vectors 1 and 7 are relatively very close and both of them are quite close to 8; vectors 4 and 9 are relatively close and both are quite close to 10; and so on. And, again unlike grammatical phrase structure trees, the lengths of the branches linking 'phrases' represents relative degrees of proximity: that the lines for linking 4 and 9 are relatively very short indicates that the corresponding vectors are close in *n*-dimensional space, but the relatively long lines between (1,7,8) and (4,9,10,5) indicate considerable distance. In the light of this, the cluster interpretation of the above tree is straightforward: there are two main clusters: (1,7,8,4,9,10,5) and (2,6,3); within each of the two main clusters there are subclusters (1,7,8) and (4,9,10,5), and (2,6) and (3); and so on.

b) Cluster analysis methods

This section lists some widely-used exploratory cluster analysis methods. It is not even nearly exhaustive; an extensive range of methods is available, for which the reader is referred to the literature on cluster analysis cited earlier in this section.

i. Linear methods

Hierarchical linear methods comprise a group of closely related algorithms which define proximity in *n*-dimensional space, the nature of a cluster, and clustering algorithms in a variety of ways: see for example Duda et al. (2001:ch.10), Everitt et al. (2001), Everitt / Dunn (2001:ch.6), Gordon (1999:69-109), Gore (2000), Hair et al. (1998:469-518), Jain et al. (1999:275-9), Kachigan (1991:261-70), Oakes (1998:110-120).

Nonhierarchical linear methods include PCA and SVD in that, if dimensionality is reduced to 2 or 3, the data vectors can be displayed and any clusters visually identified using conventional plotting tools. Another widely-used linear nonhierarchical method is Multidimensional Scaling: Borg / Groenen (2005), Davison / Sireci (2000), Everitt / Dunn (2001:ch.5), Gordon (1999:157-67), Grimm / Yarnold (1995:137-68), Hair et al. (1998:519-74), Kachigan (1991:271-78), Woods et al. (1986:262-65).

ii. Nonlinear methods

In parallel with linear PCA and SVD, nonlinear variable redefinition methods can be used for cluster analysis if dimensionality is reduced to 2 or 3 (Bishop (1995:314-19), Pyle (1999:358-83), Diamantaras / Kung (1996), Duda et al. (2001:569-70), Grimes (2006)). Beyond these, there is a good range of methods, such as Isomap (Tenenbaum et al. (2000), Locally Linear Embedding (Roweis / Saul (2000), and the very widely used (Kaski et al. 1998; Oja et al. 2001) Self-Organizing Map (Kohonen (2001)).

c) Caution

Different methods can and often do generate different results when applied to the same data. This is partly because the methods make explicit or implicit assumptions about what constitutes a cluster and how clusters so defined can be algorithmically identified, and partly because they depend to greater or lesser degrees on parameter values that are user-specified, often on a heuristic basis. It is not obvious which method and/or combination of parameter values is to be preferred in any specific application, or why. This leads to an obvious question: what are these clustering methods really telling us about the structure of the data they describe --how reliable, in other words, are they, and are they in fact of any use at all if they cannot be relied on to reveal the true structure of the data?

In the literature there are two main approaches to an answer. One is to attempt to establish the validity of cluster results using numerical measures (Everitt et al. (2001:ch.8), Duda et al. (2001:557-9), Tan et al. (2006:532-55)). The other approach is to apply a variety of different clustering methods to the same data and to compare the results: a clear convergence on one particular cluster structure is held to support the validity of that structure with respect to the data. And, of course, the two approaches can be used in combination.

6. Exploratory multivariate analysis in corpus linguistics

Any collection or written or spoken language potentially comes within the remit of corpus linguistics, and as such 'corpus linguistics' can include not only the traditional subdisciplines of linguistics proper such as phonology, morphology, and so on, but also the philological work that comes under the heading 'humanities computing' as well as information retrieval and data mining from full-text collections. To keep the length of this section within reasonable bounds, therefore, there is no attempt at exhaustiveness. The aim is rather to provide a selection of references that is representative of the applications of exploratory multivariate methods to language corpora.

- Language classification: Kita (1999).

- Phonetics & phonology: Berdan (1978), Miller / Nicely (1955), Shepard (1972, Jassem / Lobacz (1995).
- Morphology: Oakes / Taylor (1994).
- Syntax: Gries (2001), Gamallo et al. (2005).
- Lexical semantics & word-sense disambiguation: Yarowsky (2000), Stevenson / Wilks (2003), Pedersen (2006); Watters (2002), Landauer et al. (1998), Zernik (1991).
- Dialectology: Babitch / LeBrun (1989), Chambers / Trudgill (1998:135-48), Herringa / Nerbonne (2001), Kessler (1995), Kleiweg et al. (2004), Nerbonne / Heeringa (2001).
- Sociolinguistics: Chambers (1995), Horvath (1985), Jones-Sargent (1983), Moisl / Jones (2005), Moisl / Maguire / Allen (2006), Sankoff et al. (1989).
- Language register / textual genre variation: Biber and his co-workers have published extensively on this topic. See Biber's discussion in Chapter 40 of this Handbook, which also contains a full bibliography.

- Text classification: Lebart / Rajman (2000), Willett (1988), Manning / Schütze (1999:ch.16). Included here also is the large amount of work in Information Retrieval and Data Mining: see for example Belew (2000), Salton / McGill (1983), van Rijsbergen (1979), Strzalkowski (1999), Tan et al. (2006), Webb (2002).
- Stylometry: Hoover (2003), Ledger (1995), Linmans (1998), McEnery / Oakes (2000), Mealand (1995), Temple (1996). See also Oakes' discussion in chapter 52 of this Handbook.

7. Literature

Arabie, P. / Hubert, L. / de Soete, G. (eds) (1992), *Clustering and Classification*. River Edge, New Jersey: World Scientific Press.

Baayen, R. (2001), *Word Frequency Distributions*. Dordrecht: Kluwer.

Babitch, R. / LeBrun, E. (1989), Dialectometry as computerized agglomerative hierarchical classification analysis. In: *Journal of English Linguistics* 22, 83-90.

Belew, R. (2000), *Finding Out About: A cognitive perspective in search engine technology and the WWW*. Cambridge: Cambridge University Press.

Berdan, R. (1978), Multidimensional analysis of vowel variation. In: Sankoff, D. (ed) *Linguistic Variation. Models and Methods*. New York: Academic Press, 149-60.

Bertuglia, C. (2005), *Nonlinearity, Chaos, and Complexity: The dynamics of natural and social systems*. Oxford: Oxford University Press.

Bishop, C. (1995), *Neural Networks for Pattern Recognition*. Oxford: Clarendon Press.

Borg, I. / Groenen, P. (2005), *Modern Multidimensional Scaling: Theory and Applications*, 2nd ed. Berlin: Springer.

Buckley, C. (1993), The importance of proper weighting methods. In: Bates, M. (ed) *Human Language Technology.* San Mateo, CA: Morgan Kaufmann.

Chambers, J. (2003), *Sociolinguistic Theory. Linguistic variation and its social significance*, 2nd ed. Oxford: Blackwell Publishers.

Chambers, J. / Trudgill, P. (1998), *Dialectology*, 2nd ed. Cambridge: Cambridge University Press.

Church, K. / Gale, W. (1995a), Poisson mixtures. In: *Natural Language Engineering* 1, 163-90.

Church, K. / Gale, W. (1995b), Inverse Document Frequency (IDF): A Measure of Deviation from Poisson. In: *Proceedings of the Third Workshop on Very Large Corpora*, 121-130.

Croft, A. / Davison, R. / Hargreaves, M. (1992), *Engineering Mathematics*. Wokingham: Addison-Wesley.

Dale, R. / Moisl, H. / Somers, H. (eds) (2000), *Handbook of Natural Language Processing*. New York: Marcel Dekker.

Davison, M. / Sireci, S. (2000), Multidimensional scaling. In: Tinsley / Brown 2000, 323-52.

Diamantaras, K. / Kung, S. (1996), *Principal Component Neural Networks: Theory and Applications*. New Jersey: Wiley Interscience.

Duda, R. / Hart, P. / Stork, D. (2001) *Pattern Classification*, 2nd ed. New York: Wiley Interscience.

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

Everitt, B. / Dunn, G. (2001), *Applied Multivariate Data Analysis*, 2nd ed. London: Arnold.

Frakes, W. (1992), Stemming Algorithms. In: Frakes, W. / Baeza-Yates, R. (eds), *Information Retrieval: Data Structures & Algorithms*. Englewood Cliffs, NJ: Prentice-Hall.

Gamallo, P. / Agustini, A. / Lopes, G. (2005), Clustering Syntactic Positions with Similar Semantic Requirements. In: *Computational Linguistics* 31, 107-146.

Gordon, A. (1999), *Classification*, 2nd ed. London:

Chapman & Hall.

Gore, P. (2000), Cluster Analysis. In: Tinlsey / Brown 2000, 297-321.

Gries, A. (2001), Multifactorial Analysis of Syntactic Variation: Particle Movement Revisited. In: *Journal of Quantitative Linguistics* 8, 33-50.

Grimes, C. (2006), *Nonlinear Dimensionality Reduction*, Chapman & Hall / CRC.

Grimm, L. / Yarnold, P. (eds) (1995), *Reading and Understanding Multivariate Statistics*. American Psychological Association.

Grimm, L. / Yarnold, P. (eds) (2000), *Reading and Understanding More Multivariate Statistics*. American Psychological Association.

Hair, J. / Anderson, R. / Tatham, R. / Black, W. (1998), *Multivariate Data Analysis*, 5th ed. London: Prentice-Hall International.

Heeringa, W. / Nerbonne, J. (2001), Dialect areas and dialect continua. In: *Language Variation and Change* 13, 375-400.

Hoover, D. (2003), Multivariate Analysis and the Study of Style Variation. In: *Literary and Linguistic Computing* 18, 341-360.

Horvath, B. (1985), *Variation in Australian English*. Cambridge: Cambridge University Press.

Hull, D. (1996), Stemming algorithms: a case study for detailed evaluation. In: *Journal of the American Society for Information Science*, 47(1), 70-84.

Jain, A. / Dubes, R. (1988), *Algorithms for clustering data*. Englewood Cliffs, NJ: Prentice Hall.

Jain, A. / Murty, M. / Flynn, P. (1999), Data clustering: A review. In: *ACM Computing Surveys* 31, 264–323.

Jassem, W., Lobacz, P. (1995), Multidimensional Scaling and its Applications in a Perceptual Analysis of Polish Consonants. In: *Journal of Quantitative Linguistics* 2, 105-24.

Jolliffe, I. (2002), *Principal Component Analysis*, 2nd ed. Berlin: Springer.

Jones-Sargent V. (1983) *Tyne Bytes: a computerized sociolinguistic study of Tyneside English*. Frankfurt: P. Lang.

Kachigan, S. (1991), *Multivariate Statistical Analysis. A conceptual introduction*. New York: Radius Press.

Kaski, S. / Kangas, J., / Kohonen, T. (1998), Bibliography of Self-Organizing Map (SOM) Papers: 1981-1997. In: *Neural Computing Surveys* 1, 102-350.

Kessler, B. (1995), Computational dialectology in Irish Gaelic. In: *Seventh Conference of the European Chapter of the Association for Computational Linguistics*, University College Dublin, March 1995.

Kita, K. (1999), Automatic Clustering of Languages Based on Probabilistic Models. In: *Journal of Quantitative Linguistics* 6, 167-171.

Kleiweg, P., / Nerbonne, J. / Bosveld, L. (2004), Geographic projection of cluster composites. In: Blackwell, A. / Marriott, K. / Shimojima, A. (eds) *Diagrammatic Representation and Inference. Third International Conference, Diagrams 2004. Cambridge, UK, March 2004*. Berlin: Springer, 392-394.

Kohonen, T. (2001), *Self-Organizing Maps*, 3rd ed. Berlin: Springer.

Landauer, T. / Foltz, P. / Laham, D. (1998), Introduction to Latent Semantic Analysis. In: *Discourse Processes* 25, 259-284.

Lebart, L. / Rajman, M. (2000), Computing similarity. In: Dale et al. 2000, 477- 505.

Ledger, G. (1995), An Exploration of Differences in the Pauline Epistles using Multivariate Statistical Analysis. In: *Literary and Linguistic Computing* 10, 85-97.

Linmans, A. (1998), Correspondence Analysis of the Synoptic Gospels. In: *Literary and Linguistic Computing* 13, 1-13.

Manning, C. / Schütze, H. (1999), *Foundations of Statistical Natural Language Processing*. Cambridge, Mass.: MIT Press.

Mealand, D. (1997), Measuring genre differences in Mark with correspondence analysis. In: *Literary and Linguistic Computing* 12, 227-245.

McEnery, T. / Oakes, M. (2000,) Authorship identification and computational stylometry, in Dale et al. 2000, 545-62.

Miller, G. / Nicely, P. (1955), An analysis of perceptual confusion among English consonants. In: *Journal of the Acoustic Society of America* 27, 338-52.

Moisl, H. / Jones, V. (2005), Cluster analysis of the Newcastle Electronic Corpus of Tyneside English: a comparison of methods. In: *Literary and Linguistic Computing* 20, 125-46.

Moisl, H. / Maguire, W. / Allen, W. (2006), Phonetic variation in Tyneside: exploratory multivariate analysis of the Newcastle Electronic Corpus of Tyneside English. In: Hinskens, F. (ed) *Language Variation. European Perspectives*. Amsterdam: Meertens Institute.

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

Nerbonne J. / Heeringa W. (2001), Computational comparison and classification of dialects. In: *Dialectologia et Geolinguistica* 9, 69-83.

Oakes, M. (1998), *Statistics for Corpus Linguistics.* Edinburgh: Edinburgh University Press.

Oakes, M. / Taylor, M. (1994), Morphological analysis in vocabulary selection for the Ringdoc pharmacological database. In Barahona, P. / Veloso, M. / Bryant, J. (eds) *Proceedings of the 12th International Congress of the European Federation for Medical Informatics (1994),* 523-8.

Oja, M. / Kaski, S. / and Kohonen, T. (2001), Bibliography of self-organizing map (SOM) papers: 1998-2001. In: *Neural Computing Surveys* 3, 1-156.

Palmer, D. (2000), Tokenisation and sentence segmentation. In: Dale et al. 2000*,* 11-35.

Pedersen, T. (2006), Unsupervised Corpus Based Methods for WSD. In: Agirre, E. / Edmonds, P. (eds), *Word Sense Disambiguation : Algorithms and Applications*. Berlin: Springer.

Pyle, D. (1999), *Data Preparation for Data Mining.* San Francisco: Morgan Kaufmann.

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

Robertson, S. / Spärck Jones, K. (2004), IDF term weighting and IR research lessons. In: *Journal of Documentation* 60, 521-523.

Roweis, S. / Saul, L. (2000), Nonlinear dimensionality reduction by locally linear embedding. In: *Science* 290, 2323-2326.

Salton, G. / Wong, A. / Yang, C. (1975), A vector space model for automatic indexing. *Comunications of the ACM* 18, 613-20.

Salton, G. / McGill, M. (1983), *Introduction to Modern Information retrieval*. Auckland: McGraw-Hill.

Sankoff, D. / Cedergren, H. / Kemp, W. / Thibault, P. / Vincent, D. (1989), Montreal French: language, class, and ideology. In: Fasold, R. / Schiffrin, D. (eds) *Language Change and Variation*. Amsterdam: John Benjamins, 107-118.

Shepard, R. (1972), Psychological representation of speech sounds. In: David, E. / Denes, P. (eds) *Human Communication. A Unified View*. London: McGraw-Hill.

Singhal, A. / Salton, G. / Mitra, M. / Buckley, C. (1996), Document length normalization. In: *Information Processing & Management* 32(5), 619-633.

Spärck Jones, K. (1972), Exhaustivity and specificity. In: *Journal of Documentation* 28, 11-21.

Stevenson, M. / Wilks, Y. (2003), Word-sense disambiguation. In: Mitkov, R. (ed) (2003), *The Oxford Handbook of Computational Linguistics*, Oxford: Oxford University Press, 249-65.

Strzalkowski, T. (1999), *Natural Language Information Retrieval*. Dordrecht: Kluwer.

Tabachnick, B. / Fidell, L. (2006), *Using Multivariate*

Statistics, 5th ed. Boston: Allyn and Bacon.

Tan, P. / Steinbach, M. / Kumar, V. (2006), *Introduction to Data Mining*. Boston: Pearson Addison-Wesley.

Temple, J. (1996), A multivariate synthesis of published Platonic stylometric data. In: *Literary and Linguistic Computing* 11, 67-75.

Tenenbaum, J. / de Silva V. / Langford J. (2000), A global geometric framework for nonlinear dimensionality reduction. In: *Science* 290, 2319—2323.

Tinsley, H. / Brown, S. (2000), *Handbook of Applied Multivariate Statistics and Mathematical Modelling.* New York: Academic Press.

van Rijsbergen, C. (1979), *Information Retrieval*, 2nd ed. London: Butterworths.

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

Verleysen, M. / François, D. / Simon, G. / Wertz, V. (2003),. On the effects of dimensionality on data analysis with neural networks. In Mira, J. (ed), *International Work-Conference on Artificial and Natural Neural Networks*, Mao, Menorca (Spain), June 3-6, 2003, 105-112.

Vesanto, J. / Alhoniemi, E. (2000), Clustering of the self-organizing map. In: *IEEE Transactions on Neural Networks* 11(3), 586-600.

Watters, P. (2002), Discriminating English Word Senses Using Cluster Analysis**.** In: *Journal of Quantitative Linguistics* 9, 77-86.

Webb, A. (2002), *Statistical Pattern Recognition*, 2nd ed. New Jersey: John Wiley. & Sons.

Willett, P. (1988), Recent trends in hierarchic document clustering: a critical review. In: *Information Processing and Management* 24, 577-97.

Woods, A. / Fletcher, P. / Hughes, A. (1986), *Statistics in Language Studies*. Cambridge: Cambridge University Press.

Yarowsky, D. (2000), Word-sense disambiguation. In: Dale et al. 2000, 629-54.

Zernik, U. (1991), Train 1 vs Train 2: tagging word sense in a corpus. In: Zernik, U. (ed) *Lexical acquisition: exploiting on-line resources to build a lexicon*. Hillsdale NJ: Lawrence Erlbaum Associates.