1

Introduction

Our aim is to provide an introduction to symbolic data and how such data can be analyzed. Classical data on p random variables are represented by a single point in p-dimensional space $\Re^p$. In contrast, symbolic data with measurements on p random variables are p-dimensional hypercubes (or hyperrectangles) in $\Re^p$, or a Cartesian product of p distributions, broadly defined. The 'hypercube' would be a familiar four-sided rectangle if, for example, $p = 2$ and the two random variables take values over an interval $[a_1, b_1]$ and $[a_2, b_2]$, say, respectively. In this case, the observed data value is the rectangle $R = [a_1, b_1] \times [a_2, b_2]$ with vertices $(a_1, a_2),\ (a_2, b_2),\ (b_1, a_2)$, and (b_1, b_2). However, the $p = 2$ dimensional hypercube need not be a rectangle; it is simply a space in the plane. A classical value as a single point is a special case. Instead of an interval, observations can take values that are lists, e.g., {good, fair} with one or more different values in the list. Or, the observation can be a histogram. Indeed, there are many possible formats for symbolic data. Basic descriptions of symbolic data, the types of symbolic data, how they contrast with classical data, how they arise, and their inherent internal properties are covered in Chapter 2, before going on to specific analytic methodologies in the chapters that follow.

At the outset, however, it is observed that a symbolic observation in general has an internal variation. For example, an individual whose observed value of a random variable is $[a, b],\ a \neq b$, is interpreted as taking (several) values across that interval. (This is not to be confused with uncertainty or impression when the variable takes a (single) value in that interval with some level of uncertainty.) Unless stated otherwise, this interpretation applies throughout this volume. Analogous arguments apply for other types of symbolic data. A classical observation with its single point value perforce has no internal variation, and so analyses deal with variation between observations only. In contrast, symbolic data deal with the internal variation of each observation plus the variation between observations.

Symbolic Data Analysis: Conceptual Statistics and Data Mining   L. Billard and E. Diday

Symbolic data arise in a variety of different ways. Some data are inherently symbolic. For example, if the random variable of interest is 'Color' and the population is 'Species of birds', then any one realization for a given species can take several possible colors from the set of all colors, e.g., a magpie has colors {white, black}. An all-black bird is not a magpie, it is a different kind of bird. A different kind of species, not a magpie, may be such that some birds are all white, some all black, or some both black and white. In that case, the possible colors are {white, black, white and black}. Or, it may not be possible to give the exact cost of an apple (or shirt, or product, or . . .) but only its cost that takes values in the range [16, 24] cents (say). We note also that an interval cost of [16, 24] differs from that of [18, 22] even though these two intervals both have the same midpoint value of 20. A classical analysis using the same midpoint (20) would lose the fact that these are two differently valued realizations with different internal variations.

In another direction, an insurance company may have a database of hundreds (or millions) of entries each relating to one transaction for an individual, with each entry recording a variety of demographic, family history, medical measurements, and the like. However, the insurer may not be interested in any one entry per se but rather is interested in a given individual (Colin, say). In this case, all those single entries relating to Colin are aggregated to produce the collective data values for Colin. The new database relating to Colin, Mary, etc., will perforce contain symbolic-valued observations. For example, it is extremely unlikely that Mary always weighed 125 pounds, but rather that her weight took values over the interval [123, 129], say. Or, in a different scenario, when a supermarket keeps track of the sales of its products, it is not really interested in the sale of any one item (e.g., my purchase of bread yesterday) but more interested in the variation of its sales of bread (or cheese, or meat, or . . .) in a given period. It may also be interested in patterns of purchases, e.g., are the two products bread and butter often bought at the same time? Likewise, a drug company is interested in the overall patterns of purchases of certain medications rather than Frank's particular use; a hospital wants to track different types of patient ailments; an automotive insurance company is less interested in a single car accident but more interested in car accidents of different driver demographic categories, or of different types of vehicles. Rather than focusing on a single DNA sequence, a geneticist may be interested in tracking the occurrences of a particular amino acid in a protein chain instead of one single DNA sequence. For example, CCU, CCC, CCA, and CCG all code for proline; similarly, GCU, GCC, GCA, and GCG code for alanine.

In all these and similar situations, there is a basic concept of interest, in effect a second level of observations, whose values are obtained by aggregation over the observations for those individuals (at a first level) who constitute the descriptor of that concept. By 'aggregation', it is meant the collection of individuals who satisfy the descriptor of interest. For example, suppose the variable is farm acreage, and Drew and Calvin have acreages of 7.1 and 8.4, respectively, both in the northeast region. Then, if the concept of interest is acreage of farms in the northeast, aggregating Drew and Calvin produces an acreage of [7.1,8.4]. How aggregation is implemented is discussed in Chapter 2. We note that the original dataset may itself contain symbolic data. The aggregated dataset, however, will almost certainly contain symbolic data regardless of whether the first-level entries are classical or symbolic values.

In these two types of settings, the original database can be small or large. A third setting is when the original database is large, very large, such as can be generated by contemporary computers. Yet these same computers may not have the capacity to execute even reasonably elementary statistical analyses. For example, a computer requires more memory to invert a matrix than is needed to store that matrix. In these cases, aggregation of some kind is necessary even if only to reduce the dataset to a more manageable size for subsequent analysis. There are innumerable ways to aggregate such datasets. Clearly, it makes sense to seek answers to reasonable scientific questions (such as those illustrated in the previous paragraphs) and to aggregate accordingly. As before, any such aggregation will perforce produce a dataset of symbolic values.

Finally, we note that the aggregation procedures adopted and their emergent symbolic datasets are not necessarily obtained from clustering procedures. However, standard (classical, and indeed also symbolic) clustering methods do produce clusters of observations whose values when summarized for the respective random variables are symbolic in nature (unless the analyst chooses to use classical summaries such as average values, which values as we shall see do not retain as much information as do the symbolic counterparts).

Chapter 3 presents methodologies for obtaining basic descriptive statistics for one random variable whose values are symbolic valued such as intervals, namely, a histogram and its empirical probability distribution relative, along with the empirical mean and variance. Descriptive statistics are extended to $p = 2$ dimensions in Chapter 4. A key feature in these calculations is the need to recognize that each observation has its own internal variation, which has to be incorporated into the respective methodologies. Thus we see, for example, that a single observation (a sample of size 1) whose value is the interval $[a, b] = [2, 8]$, say, has a sample variance of 3, whereas a single classical observation (the midpoint 5, say) has sample variance of 0. That is, the classical analysis loses some of the information contained in the datum. In both Chapter 3 and Chapter 4, methodologies are developed for multi-valued, interval-valued, modal multi-valued and modal interval-valued (i.e., histogram-valued) observations.

A second distinctive feature is the presence of so-called rules. In addition to the internal variation mentioned in the previous paragraph, there can be structural variations especially after aggregation of a larger dataset. For example, in studying batting performances of teams in a baseball league, the numbers of hits and at-bats are aggregated over the players of the respective teams. While for each individual player the number of hits would not exceed the number of at-bats, it is possible that when aggregated by team the number of hits could exceed the number of at-bats for some regions of the resulting hypercube. Clearly, it is necessary to add a rule to the team data so as to retain the fact that, for the individual players, the number of hits does not exceed the number of at-bats (see Section 4.6 for a more detailed discussion of this example). This type of rule is a logical dependency rule and is designed to maintain the data integrity of the original observations. Such rules are often necessary for symbolic data but not for classical data. On the other hand, there are numerous rules (the list is endless) that apply equally to classical and symbolic data, e.g., structural rules such as taxonomy and hierarchy trees that can prevail. Special attention is paid to logical dependency rules when calculating the basic descriptive statistics of Chapters 3 and 4. Subsequent methodologies will by and large be described without specific detailed attention to these rules.

Chapters 5–7 deal with principal components, regression and clustering techniques, respectively. These methods are extensions of well-known classical theory applied or extended to symbolic data. Our approach has been to assume the reader is knowledgeable about the classical results, with the present focus on the adaptation to the symbolic data setting. Therefore, only minimal classical theory is provided. Regression methodologies are provided for multi-valued data, interval-valued data and histogram-valued data, along with methods for handling taxonomy tree structures and hierarchy tree structures.

The work on principal components considers two methods, the vertices method and the centers method. Both deal with interval-valued data, the former as its name suggests focusing on the vertices of the hypercube associated with each data value, and the latter using the midpoint center values. Principal component methods for multi-valued and modal-valued data currently do not exist.

Clustering techniques in Chapter 7 look at partitioning, hierarchies, divisive clustering, and pyramid clustering in particular. Since these methods depend on dissimilarity and distance measures, the chapter begins with developing these measures for symbolic data. Many of these measures and techniques are also extensions of their classical counterparts. Where available, methodologies are presented for multi-valued, interval-valued, and histogram-valued variables. In some instances, a mixed variable case with some variables multi-valued and other variables interval-valued is considered (usually through the well-known oils dataset).

Perhaps the most striking aspect of this catalogue of techniques is the paucity of available methodologies for symbolic data. It is very apparent that contemporary datasets will be increasingly symbolic in content. Indeed, Schweizer (1984) has declared that ‘distributions are the numbers of the future’, and ‘distributions’ are examples of what are symbolic data. Therefore, it is imperative that suitable techniques be developed to analyze the resulting data. In that sense the field is wide open providing lots of opportunities for new research developments. In another direction, while there are some methodologies presently existing, most seem intuitively correct, especially since results for the particular case of classical data do indeed emerge from the symbolic analyses. However, except in rare isolated instances, there is next to no work yet being done on establishing the mathematical underpinning necessary to achieve rigor in these results. Again, opportunity knocks for those so inclined.

The attempt has been made to provide many illustrative examples throughout. In several places, datasets have been presented with more random variables than were actually used in the accompanying example. Those 'other' variables can be used by the reader as exercises. Indeed, most datasets provided (in any one example or section) can be analyzed using techniques developed in other sections; again these are left as exercises for the reader.

The reader who wants to learn how to analyze symbolic data can go directly to Chapter 3. However, Section 2.1 would serve as a introduction to the types of data and the notation used in subsequent chapters. Also, Section 2.3 with its comparison between classically adapted analyses of symbolic data provides some cautionary examples that can assist in understanding the importance of using symbolic analytic methods on symbolic data.

Symbolic data appear in numerous settings, all avenues of the sciences and social sciences, from medical, industry and government experiments and data collection pursuits. An extensive array of examples has been included herein drawing upon datasets from as wide a range of applications as possible. Having said that, some datasets are used a number of times in order to provide a coherency as well as a comparative dialog of different methodologies eliciting different information from the same data. In this way, a richer knowledge base of what is contained within a dataset can be exposed, providing for a broader set of interpretations and scientific conclusions.

The datasets used are publicly available (at http://www.stat.uga.edu/faculty/LYNNE/Lynne.html) source references as to where they can be found are provided (usually when they first appear in the text). All can be obtained from the authors. As is evident to the reader/user, some examples use the entire dataset presented in the text's illustrations, while others use a portion of the data, be that some but not all of the observations or some but not all of the random variables. Therefore, the user/reader can take other portions, or indeed the complete dataset, and thence analyze the data independently as exercises. Some are provided at the end of each chapter. Since several of the methodologies presented can be applied to such data, these alternative analyses can also be performed as yet another extensive range of potential exercises. The reader might be encouraged to work through the details of the methodology independently even if only to persuade him/herself that the technique has been understood. The SODAS software package (available free at http://www.ceremade.dauphine.fr/%7Etouati/sodas-pagegarde.htm) provides the computational power for most of the methodology presented in this text.

In some sense this volume owes its origins to the review paper by Billard and Diday (2003) and the edited volume by Bock and Diday (2000). Those publications form initial attempts to bring together the essential essence of what symbolic data are. The forthcoming edited volume by Diday and Noirhomme (2006) complements the present volume in so far as it provides algorithmic and computational details of some of the software routines in the SODAS software package.

References

Billard, L. and Diday, E. (2003). From the Statistics of Data to the Statistics of Knowledge: Symbolic Data Analysis. *Journal of the American Statistical Association* 98, 470–487.

Bock, H.-H. and Diday, E. (eds.) (2000). *Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information from Complex Data*. Springer-Verlag, Berlin.

Diday, E. and Noirhomme, M. (eds.) (2006). *Symbolic Data and the SODAS Software*. John Wiley & Sons, Ltd, Chichester.

Schweizer, B. (1984). Distributions are the Numbers of the Future. In: *Proceedings of The Mathematics of Fuzzy Systems Meeting* (eds. A. di Nola and A. Ventes). University of Naples, Naples, Italy, 137–149.