doi:10.2498/cit.1001390

# Some Analyses of Interval Data

Lynne Billard

Department of Statistics, University of Georgia, Athens, USA

Contemporary computers bring us very large datasets, datasets which can be too large for those same computers to analyse properly. One approach is to aggregate these data (by some suitably scientific criteria) to provide more manageably-sized datasets. These aggregated data will perforce be symbolic data consisting of lists, intervals, histograms, etc. Now an observation is a $p$-dimensional hypercube or Cartesian product of $p$ distributions in $\mathbf{R}^p$, instead of the $p$-dimensional point in $\mathbf{R}^p$ of classical data. Other data can be naturally symbolic. We give a brief overview of interval-valued data and show briefly that it is important to use symbolic analysis methodology since, e.g., analyses based on classical surrogates ignore some of the information in the dataset.

*Keywords:* aggregate data, symbolic data analysis, principal component analysis, p-dimensional hypercube, interval-valued data, divisive clustering method

## 1. Introduction

Suppose $\mathbf{Y} = (Y_1, \cdots, Y_p)$ is a $p$-dimensional random variable. Classical data values of $\mathbf{Y}$ are points in $p$-dimensional space $\mathbf{R}^p$. In contrast, symbolic data are hypercubes in $\mathbf{R}^p$ or a Cartesian product of $p$ distributions. Symbolic data typically are in the form of lists, intervals, or modal values; the most common form of modal data is histogram-valued data, but it can include other formats such as models, or "histograms" weighted by possibilities, necessities, capacities, credibilities, and the like. A classical value is a special case. See Billard and Diday (2006) for detailed descriptions of symbolic data.

Symbolic data arise in a number of different ways. One way is as the result of aggregation of (usually) large or enormous datasets. The aggregation can occur simply to produce a dataset of more manageable size in order to conduct appropriate analyses; or it can occur as a result of some scientific question(s) of interest. Take the data of Table 1, extracted from a large dataset of (over 200) variables. Table 1(a) shows the values for individual cardiac patients, and gives details of the admitting hospital, age, and smoking history. In this particular study, interest centered on the survival rates for patients following the different pathways (where pathways were identified by hospital, units such as cardiology, or intensive care, or . . ., to which a patient was admitted). Thus, the statistical "observation" of study was not the patient but the pathway. Therefore, observed $Y$ values for a pathway are those obtained after aggregating over all patients who followed that specific pathway. Thus, in Table 1(b), the patients who followed the path-

| Patient | Hospital | Age | Smoker |
|---|---|---|---|
| Patient 1 | Hospital 1 | 74 | heavy |
| Patient 2 | Hospital 1 | 78 | light |
| Patient 3 | Hospital 2 | 69 | no |
| Patient 4 | Hospital 2 | 73 | heavy |
| Patient 5 | Hospital 2 | 80 | light |
| Patient 6 | Hospital 1 | 70 | heavy |
| Patient 7 | Hospital 1 | 82 | heavy |
| Patient 8 | Hospital 3 | 74 | heavy |
| ⋮ | ⋮ | ⋮ | ⋮ |

*Table 1.(a)* Classical Values.

| Hospital | Age | Smoker |
|---|---|---|
| Hospital 1 | [70, 82] | {light 1/4, heavy 3/4} |
| Hospital 2 | [69, 80] | {no, light, heavy} |
| Hospital 3 | [74, 74] | {heavy} |
| ⋮ | ⋮ | ⋮ |

*Table 1.(b)* Symbolic Values.

way ‘Hospital 1’ had ages over the interval [70, 82]. Here, $Y_1 =$ age becomes an interval-valued symbolic value. The variable $Y_2 =$ smoker becomes a list or multi-valued symbolic value, and can be modal multi-valued as for Hospital 1 or a list as for Hospital 2. Classical values are special cases of symbolic data. Thus, the point value $a \equiv [a, a]$, and the categorical value $c \equiv \{c\}$; see Hospital 3 in Table 1(b).

There are many possible aggregations. Clearly, those driven by relevant scientific questions are best. Medical insurance companies may not be particularly interested in your details for a specific visit to a physician, hospital, clinic, . . ., but rather are interested in the aggregate of your visits over a time period such as 5-years. Or, they may be interested in 20-year-old males, or 30-year-old females; or, in those with lung cancer collectively or by age × gender, or . . ., and so on. The automobile manufacturer is less interested in the type of car you purchased but rather in the automobile purchases of 40-year-olds (or 40-year-old males), or of purchases of cars by model and make, etc.

Some data are naturally symbolic. For example, Table 2 contains values for $Y_1 =$ pileus cap width, $Y_2 =$ stipe length, $Y_3 =$ stipe thickness, and $Y_4 =$ edibility of species of mushrooms (from Billard and Diday, 2006, Table 3). Thus, for the species *arorae* the cap width is $Y_1 = [3, 8]$. Any one mushroom from that species has a specific length, e.g., $Y_1 = 5.2$, i.e., a classical point value; but it cannot be said that all mushrooms of that species have the same $Y_1$(=5.2, say) value. Notice $Y_4$ however is a classical value for each species.

## 2. Symbolic or Classical Analysis?

In the absence of techniques to analyse symbolic data directly, it is tempting to use classical surrogates. To do so, however, loses information contained in the data. Consider the three different realizations of the random variable $Y$ = weight, viz., $Y_1 = 135$, $Y_2 = [132, 138]$, $Y_3 = [129, 141]$. Let these be three samples each of size $m = 1$. It is easily shown that the sample mean is $\bar{Y}_1 = \bar{Y}_2 = \bar{Y}_3 = 135$; the sample variance is $S_1^2 = 0$, $S_2^2 = 3$, $S_3^2 = 12$. [See (1) below for the definition of $S^2$ for interval-valued observations.] Thus, had we taken the midpoints as a classical surrogate for our analyses, all three samples would have produced the same answer. Yet, since $S_i^2 \neq S_{i'}^2$, $i \neq i'$, it is clear that the samples are differently valued. In this case, the differences revolve around the internal variation of symbolic observations. Classical observations have no internal variation. Therefore, the use of classical surrogates implies that these internal variations are not taken into account.

Bertrand and Goupil (2000) have shown that under the assumption that values are uniformly distributed within each interval, the sample variance (of each variable $Y$) of a set of observations $Y_u = [a_u, b_u], u = 1, \ldots, m$, is

$$S^2 = \frac{1}{3m} \sum_{u=1}^{m} (a_u^2 + a_u b_u + b_u^2) - \frac{1}{4m^2} \left[ \sum_{u=1}^{m} (a_u + b_u) \right]^2 \tag{1}$$

| $w_u$ | Species | Pileus Cap Width | Stipe Length | Stipe Thickness | Edibility |
|---|---|---|---|---|---|
| $w_1$ | *arorae* | [3.0, 8.0] | [4.0, 9.0] | [0.50, 2.50] | U |
| $w_2$ | *arvenis* | [6.0, 21.0] | [4.0, 14.0] | [1.00, 3.50] | Y |
| $w_3$ | *benesi* | [4.0, 8.0] | [5.0, 11.0] | [1.00, 2.00] | Y |
| $w_4$ | *bernardii* | [7.0, 6.0] | [4.0, 7.0] | [3.00, 4.50] | Y |
| $w_5$ | *bisporus* | [5.0, 12.0] | [2.0, 5.0] | [1.50, 2.50] | Y |
| $w_6$ | *bitorquis* | [5.0, 15.0] | [4.0, 10.0] | [2.00, 4.00] | Y |
| $w_7$ | *califorinus* | [4.0, 11.0] | [3.0, 7.0] | [0.40, 1.00] | T |
| . . . | . . . | . . . | . . . | . . . | . . . |

*Table 2.* Mushrooms.

and the sample mean is

$$\bar{Y} = \frac{1}{2m}\sum_{u=1}^{m}(a_u + b_u). \qquad (2)$$

This sample variance can be shown to satisfy (see Billard, 2007)

$$mS^2 = \frac{1}{3}\sum_{u=1}^{m}\Big[(a_u-\bar{Y}_u)^2+(a_u-\bar{Y}_u)(b_u-\bar{Y}_u) + (b_u-\bar{Y}_u)^2\Big] + \sum_{u=1}^{m}\Big[\frac{a_u+b_u}{2}-\bar{Y}\Big]^2 \qquad (3)$$

where the midpoint of each observation $[a_u, b_u]$ is

$$\bar{Y}_u = (a_u + b_u)/2. \qquad (4)$$

That is, the Total Sum of Squares (SS) is

$$mS^2 = \text{Total SS} = \text{Within SS} + \text{Between SS}. \qquad (5)$$

The Between SS term represents the variation between the midpoints of the observations. The Within SS is the sum of the internal variations of the observations. When all observations are classically valued, $a_u = b_u = \bar{Y}_u$, and so Within SS $= 0$, as a special case. When the interval midpoints are used as classical surrogates, this Within SS component of the total variation is ignored. Hence, the results will differ from the (correct) symbolic analyses results.

Exploiting the analogous relationship (5) for sums of products, we can show that for observations $\mathbf{Y}_u = (Y_{1u}, Y_{2u})$ with $Y_{ju} = [a_{ju}, b_{ju}]$,

$$\begin{aligned} Cov(Y_1, Y_2) = \frac{1}{6m}\sum_{u=1}^{m}\Big[&2(a_{1u}-\bar{Y}_1)(a_{2u}-\bar{Y}_2) \\ &+ (a_{1u}-\bar{Y}_1)(b_{2u}-\bar{Y}_2) \\ &+ (b_{1u}-\bar{Y}_1)(a_{2u}-\bar{Y}_2) \\ &+ 2(b_{1u}-\bar{Y}_1)(b_{2u}-\bar{Y}_2)\Big] \end{aligned} \qquad (6)$$

As for the variances, analyses using classical surrogates such as interval midpoints lose the within observations "covariances". Suitable adjustments for those interval observations where the internal distribution is non-uniform follow through.

| $w_u$ | $Y_1$ | $Y_2$ | $Y_3$ |
|---|---|---|---|
| | **Pulse Rate** | **Systolic Pressure** | **Diastolic Pressure** |
| $w_1$ | [44, 68] | [90, 110] | [50, 70] |
| $w_2$ | [60, 72] | [90, 130] | [70, 90] |
| $w_3$ | [56, 90] | [140, 180] | [90, 100] |
| $w_4$ | [70, 112] | [110, 142] | [80, 108] |
| $w_5$ | [54, 72] | [90, 100] | [50, 70] |
| $w_6$ | [70, 100] | [134, 142] | [80, 110] |
| $w_7$ | [72, 100] | [130, 160] | [76, 90] |
| $w_8$ | [76, 98] | [110, 190] | [70, 110] |
| $w_9$ | [86, 96] | [138, 180] | [90, 110] |
| $w_{10}$ | [86, 100] | [110, 150] | [78, 100] |
| $w_{11}$ | [53, 55] | [160, 190] | [205, 219] |
| $w_{12}$ | [50, 55] | [180, 200] | [110, 125] |
| $w_{13}$ | [73, 81] | [125, 138] | [78, 99] |
| $w_{14}$ | [60, 75] | [175, 194] | [90, 100] |
| $w_{15}$ | [42, 52] | [105, 115] | [70, 82] |

*Table 3.* Blood Dataset.

To illustrate, take the blood data of Table 3 (taken from Billard and Diday, 2006, Table 3.5). Here, $Y_1$ = pulse rate, $Y_2$ = systolic pressure, and $Y_3$ = diastolic pressure. These data may have arisen by aggregating values for individuals making up the respective categories $\Omega = \{w_1, \cdots, w_{15}\}$. Or, since it is known that pulse rates and blood pressure values typically fluctuate considerably, the categories $\{w_u\}$ can be measurements over time on single individuals. The sample mean and variance for each variable obtained from (1) and (2), as well as the covariances from (6) and hence the correlation functions, are shown in Table 4(a).

It is often the case that logical dependency rules exist to maintain the integrity of the data and/or to perform the role of data cleaning. Here, since diastolic pressure is less than systolic pressure, $Y_3 < Y_2$, the observation $w_{11}$, by violating this axiom, should be omitted as a data cleaning exercise. The resulting sample means, variances, covariances and correlations are given in Table 4(b). In this case, the $Y_2$ and $Y_3$ values are such that all $Y_3/Y_2$ values violate the rule $v : Y_3 < Y_2$. More generally, aggregation of individual values, each of which has $Y_3 < Y_2$, could produce a symbolic datapoint of, e.g.,

| | $Y_1$ | $Y_2$ | $Y_3$ |
|---|---|---|---|
| | **Pulse Rate** | **Systolic Pressure** | **Diastolic Pressure** |
| (a) | $\bar{Y}_1 = 72.43$ | $\bar{Y}_2 = 140.27$ | $\bar{Y}_3 = 95.67$ |
| No Rules | $s_1^2 = 272.50$ | $s_2^2 = 922.40$ | $s_3^2 = 1204.13$ |
| | $s_1 = 16.51$ | $s_2 = 30.37$ | $s_3 = 34.70$ |
| | $Cov(Y_1, Y_2) = 53.21$ | $Cov(Y_2, Y_3) = 674.99$ | $Cov(Y_1, Y_3) = 63.60$ |
| | $\rho(Y_1, Y_2) = 0.106$ | $\rho(Y_2, Y_3) = 0.640$ | $\rho(Y_1, Y_3) = -0.111$ |
| (b) | $\bar{Y}_1 = 73.45$ | $\bar{Y}_2 = 137.79$ | $\bar{Y}_3 = 87.36$ |
| $Y_3 < Y_2$ | $s_1^2 = 265.94$ | $s_2^2 = 890.60$ | $s_3^2 = 253.25$ |
| | $s_1 = 16.31$ | $s_2 = 29.84$ | $s_3 = 15.91$ |
| | $Cov(Y_1, Y_2) = 105.65$ | $Cov(Y_2, Y_3) = 411.47$ | $Cov(Y_1, Y_3) = 95.80$ |
| | $\rho(Y_1, Y_2) = 0.217$ | $\rho(Y_2, Y_3) = 0.866$ | $\rho(Y_1, Y_3) = 0.369$ |

*Table 4.* Descriptive Statistics.

$(Y_2, Y_3) = ([160, 190], [170, 185])$, say. Now, only that portion bounded by the vertices (160, 170), (160, 185), (185, 185), (170, 170) forming a hexagon in $\mathbf{R}^2$ is a valid region. This is analogous to the baseball dataset analysed in detail in Billard and Diday (2006), q.v.

## 3. Principal Component Analysis

Chouakria (1998) and Billard et al. (2008) have proposed a principal component methodology based on the vertices of the data hypercubes. They show that the $\nu$th symbolic principal component is

$$Y^*_{u\nu} = [y^a_{u\nu}, y^b_{u\nu}], \quad \nu = 1, \ldots, s \leq p, \qquad (7)$$

where

$$y^a_{u\nu} = \min_{k \in L_u}\{y^u_{\nu k}\}, \quad y^b_{u\nu} = \max_{k \in L_u}\{y^u_{\nu k}\} \qquad (8)$$

where $y^u_{\nu k}$ is the $\nu$th principal component for the vertex $k$ of the hypercube of observation $w_u$ and $L_u$ is the set of vertices associated with $w_u$. The resulting $\nu = 1$ and $\nu = 2$ principal components PC$\nu$ for the blood data of Table 3 are shown in Table 5(a) and plotted in Figure 1.

In order to help clarify the visualization and interpretation of these principal components, it is further proposed to retain for use in (8) only those vertices $x^u_k$ whose contribution

$$Ctr(\mathbf{x}^i_k, PC\nu) = \frac{(y^i_{\nu k})^2}{[d(\mathbf{x}^i_k, \mathbf{G})]^2}$$

exceeds a specified $\alpha$. In (9), $x^u_k$ is the vertex $k$ in $L_u$, $d(\cdot, G)$ is the Euclidean distance from that vertex and the centroid $G$ of all observations. Eqn (8) directly corresponds to $\alpha = 0$. When $\alpha = 0.2$, the principal components $PC\nu$, $\nu = 1, 2$, are as shown in Table 5(b); the number of vertices $n_\nu$ which satisfies this condition, $\nu = 1, 2$, is also shown. The resulting $PC\nu$, $\nu = 1, 2$, are plotted in Figure 2.

The clusters become more apparent in Figure 2. Thus, we see that observations $\{w_1, w_2, w_5, w_{15}\} = C_1$ constitute one cluster; $\{w_4, w_6, w_7, w_8, w_9, w_{10}\} = C_2$ form a second cluster, with maybe $w_{13}$ part of this second cluster if not its own cluster; $\{w_3, w_{12}, w_{14}\}$ are a single cluster $C_3$ (or perhaps two clusters); and finally $\{w_{11}\} = C_4$ is a cluster on its own.

Figure 3 is the plot of the $PC1$ and $PC2$ obtained when the interval midpoints are used as classical surrogates. Unlike the distinctness of $w_3$ and $w_{14}$ in the symbolic analysis, these observations would seem to be part of the central cluster $C_2$. The degree of these differences depends on the varied lengths of the intervals. More importantly, however, the relative sizes of the principal component hypercubes reflect those of the data hypercubes. For example, comparing those for $w_3$ and $w_{14}$ in Figure 1 (or in Figure 4 below), we see how the data hypercube for $w_4$ fits "inside" that of $w_{14}$ as do also the respective principal component hypercubes. This phenomenon cannot be observed in any classical analysis.

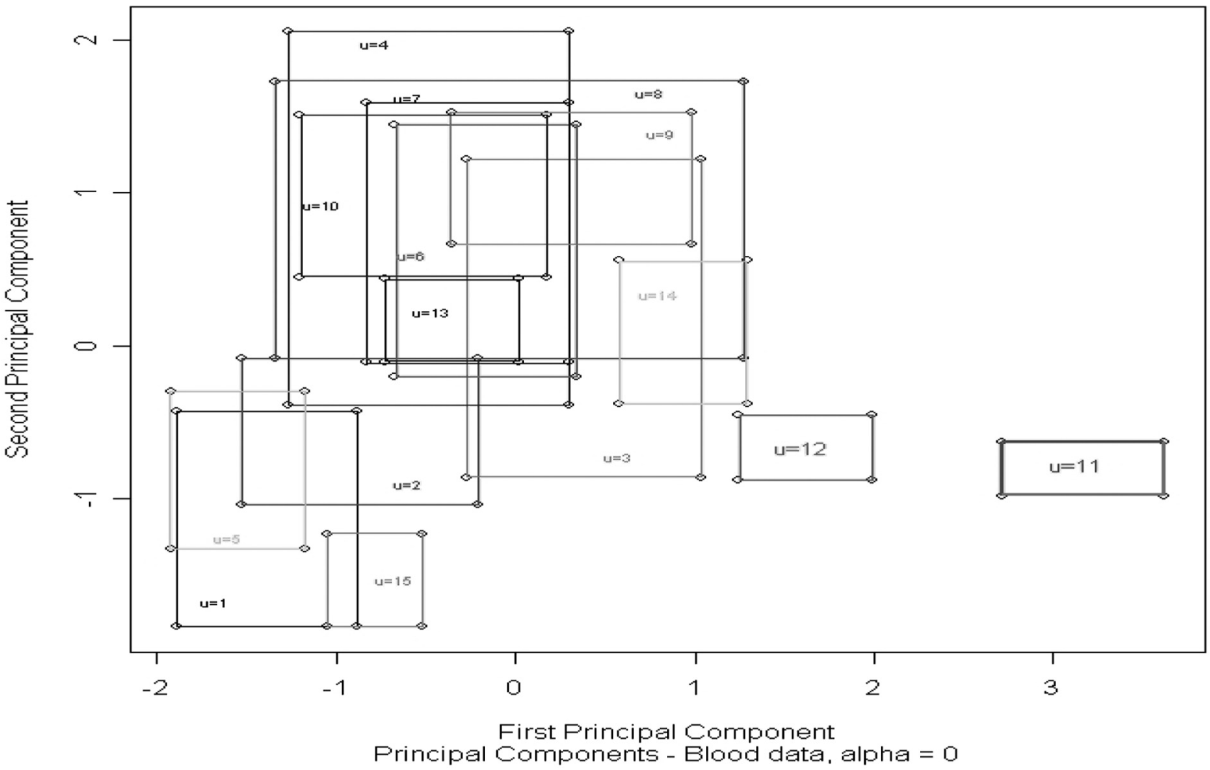

*Figure 1.*

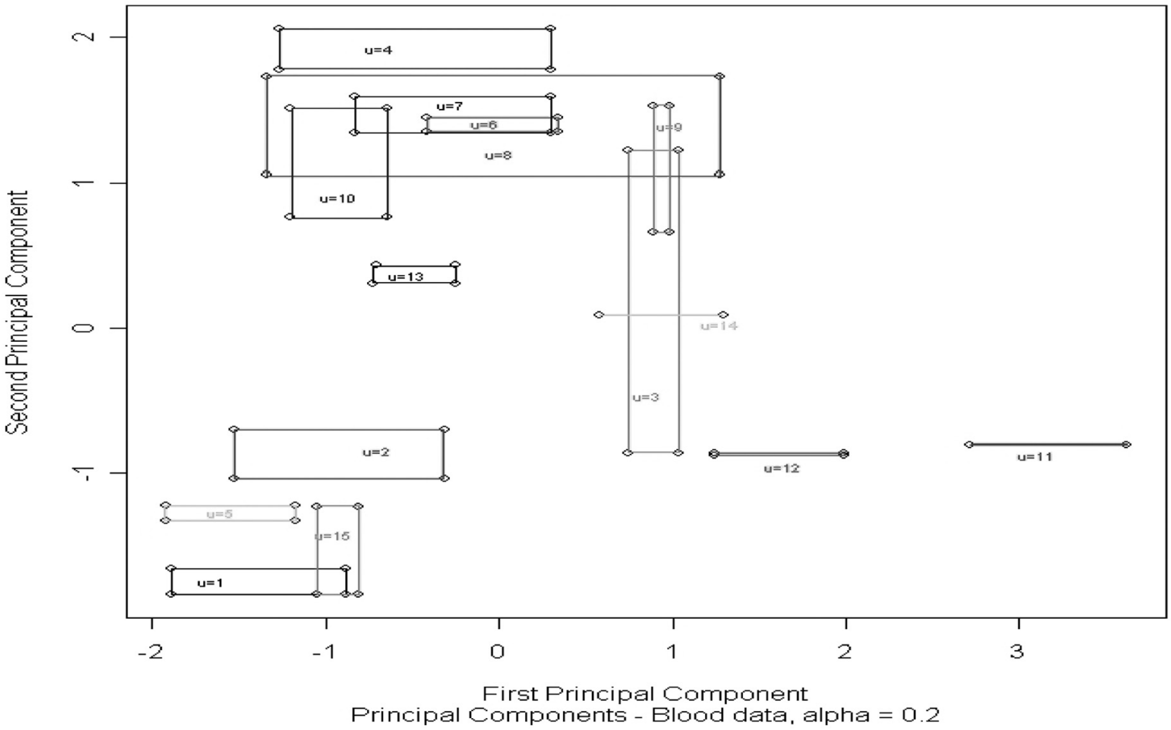

*Figure 2.*

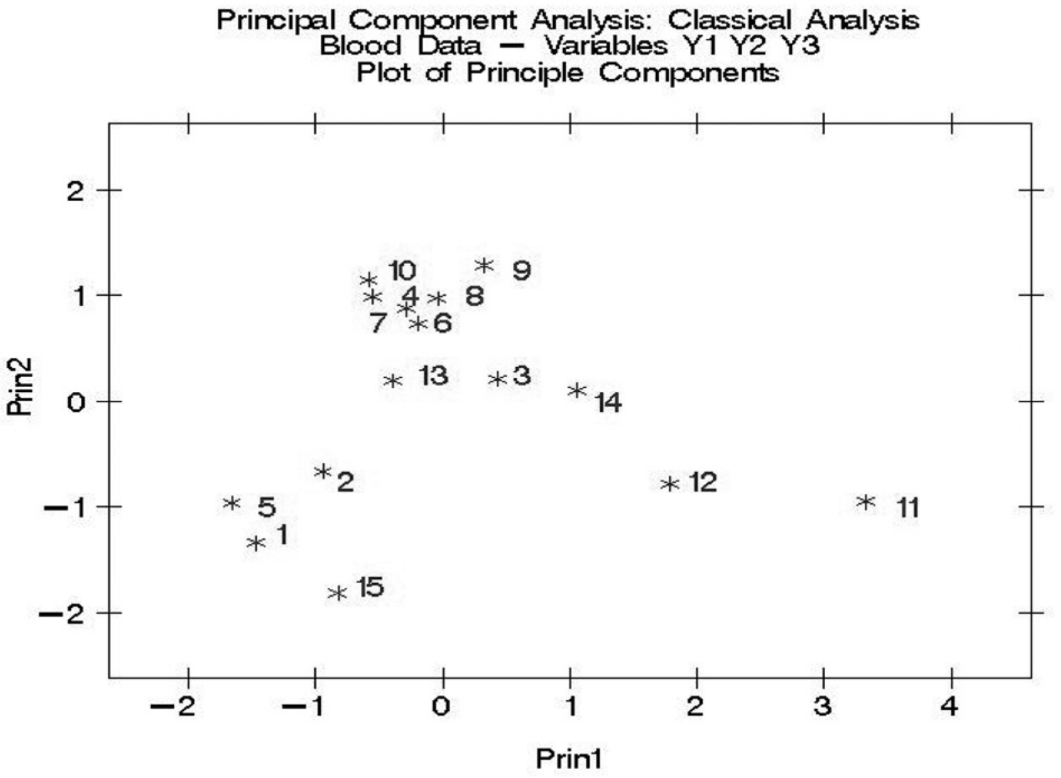

*Figure 3.*

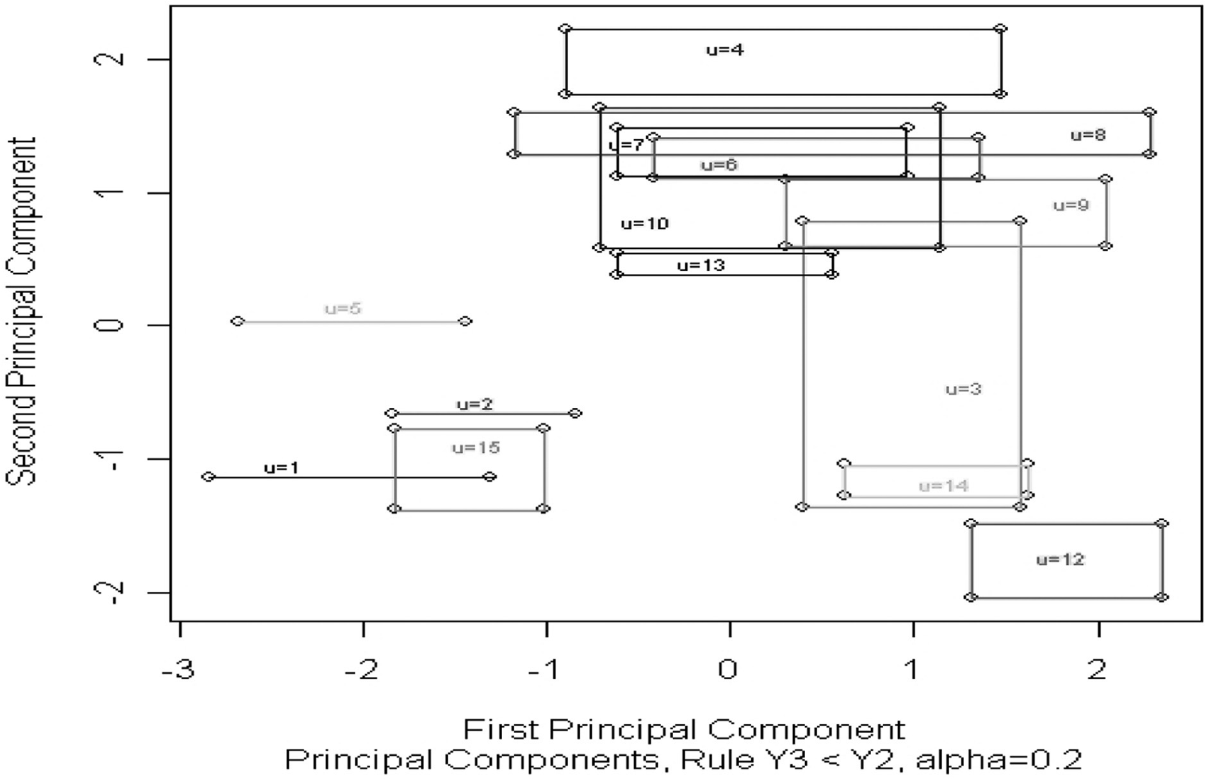

*Figure 4.*

| | **(a)** $\alpha = 0$ | | **(b)** $\alpha = 0.2$ | | |
|---|---|---|---|---|---|
| $u$ | **PC1** | **PC2** | **PC1** | **PC2** | $n_v$ $^+$ |
| 1 | [-1.891, -0.881] | [-1.843, -0.428] | [-1.891, -0.881] | [-1.843, -1.663] | 8, 4 |
| 2 | [-1.529, -0.214] | [-1.040, -0.085] | [-1.529, -0.317] | [-1.040, -0.703] | 7, 4 |
| 3 | [-0.267, 1.040] | [-0.862, 1.216] | [0.748, 1.040] | [-0.862, 1.216] | 3, 6 |
| 4 | [-1.266, 0.302] | [-0.386, 2.059] | [-1.266, 0.302] | [1.777, 2.059] | 4, 4 |
| 5 | [-1.925, -1.171] | [-1.328, -0.301] | [-1.925, -1.171] | [-1.328, -1.228] | 8, 4 |
| 6 | [-0.673, 0.341] | [-0.199, 1.441] | [-0.416, 0.341] | [1.346] | 3, 4 |
| 7 | [-0.834, 0.296] | [-0.106, 1.588] | [-0.834, 0.296] | [1.336, 1.588] | 4, 4 |
| 8 | [-1.344, 1.270] | [-0.079, 1.728] | [-1.344, 1.270] | [1.054, 1.728] | 4, 4 |
| 9 | [-0.359, 0.980] | [0.657, 1.525] | [0.895, 0.980] | [0.657, 1.525] | 2, 8 |
| 10 | [-1.203, 0.170] | [0.446, 1.507] | [-1.203, -0.647] | [0.762, 1.507] | 3, 6 |
| 11 | [2.717, 3.624] | [-0.984, -0.629] | [2.717, 3.624] | -0.807 | 8, 0 |
| 12 | [1.246, 1.994] | [-0.881, -0.449] | [1.246, 1.994] | [-0.881, -0.865] | 8, 2 |
| 13 | [-0.733, 0.016] | [-0.104, 0.433] | [-0.733, -0.249] | [0.308, 0.433] | 6, 3 |
| 14 | [0.577, 1.291] | [-0.379, 0.554] | [0.577, 1.291] | 0.088 | 8, 0 |
| 15 | [-1.052, -0.524] | [-1.840, -1.233] | [-1.052, -0.814] | [-1.840, -1.233] | 4, 8 |

$^+$ $n_v$ = # of vertices retained

*Table 5.* Principal Components.

| | $\alpha = 0.2$ | | |
|---|---|---|---|
| $u$ | **PC1** | **PC2** | $n_v$ |
| 1 | [-2.838, -1.302] | [-1.149, -1.149] | 8, 1 |
| 2 | [-1.833, -0.836] | [-0.662, -0.662] | 6, 1 |
| 3 | [0.402, 1.579] | [-1.370, 0.777] | 5, 6 |
| 4 | [-0.896, 1.469] | [1.738, 2.229] | 4, 4 |
| 5 | [-2.677, -1.439] | 0.022 | 8, 0 |
| 6 | [-0.414, 1.351] | [1.113, 1.414] | 6, 4 |
| 7 | [-0.612, 0.964] | [1.115, 1.479] | 3, 4 |
| 8 | [-1.174, 2.282] | [1.285, 1.595] | 4, 2 |
| 9 | [0.297, 2.049] | [0.591, 1.099] | 7, 3 |
| 10 | [-0.713, 1.138] | [0.582, 1.635] | 3, 7 |
| 12 | [1.309, 2.352] | [-2.039, -1.498] | 8, 8 |
| 13 | [-0.622, 0.553] | [0.378, 0.541] | 5, 3 |
| 14 | [0.622, 1.619] | [-1.286, -1.046] | 8, 4 |
| 15 | [-1.821, -1.011] | [-1.386, -0.785] | 8, 7 |

*Table 6.* Principal Components, $Y_3 < Y_2$.

Finally, under the rule $v : Y_3 < Y_2$, the symbolic principal components $PCv$, $v = 1, 2$ for $\alpha = 0.2$ are shown in Table 6 and plotted in Figure 4. The effect of the invalid datapoint $w_{11}$ is immediately apparent by comparing Figure 2 and Figure 4.

## 4. Clusters

As another example of the distinctness of a symbolic analysis over a classical analysis, the divisive clustering method of Chavent (1997, 1998, 2000) is applied to the symbolic data in Table 3, under the rule $Y_3 < Y_2$. The resulting hierarchy is shown in Figure 5. When the divisive method is applied to the midpoints as classical surrogates, the hierarchy is as displayed in Figure 6. Not only do the hierarchies produce different clusters, the cutting criteria (after the first cut) differ, too. The difference is explained by the fact that in the symbolic analysis, all the information in the data is used, while in the classical analysis, some of the information is omitted.

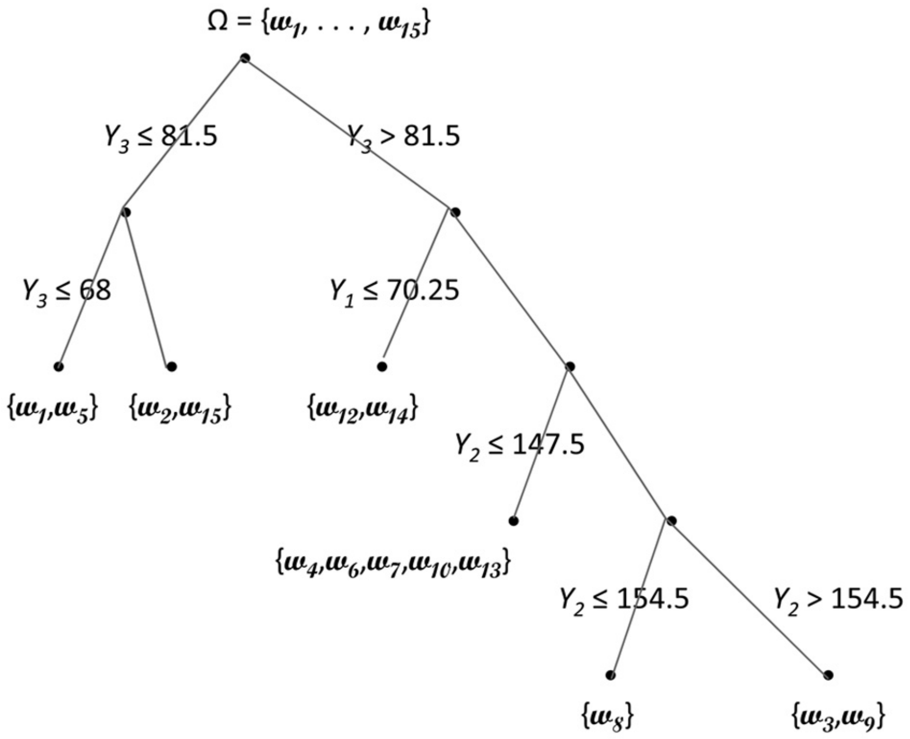

*Figure 5.* Divisive clustering on symbolic data, $Y3 < Y2$.

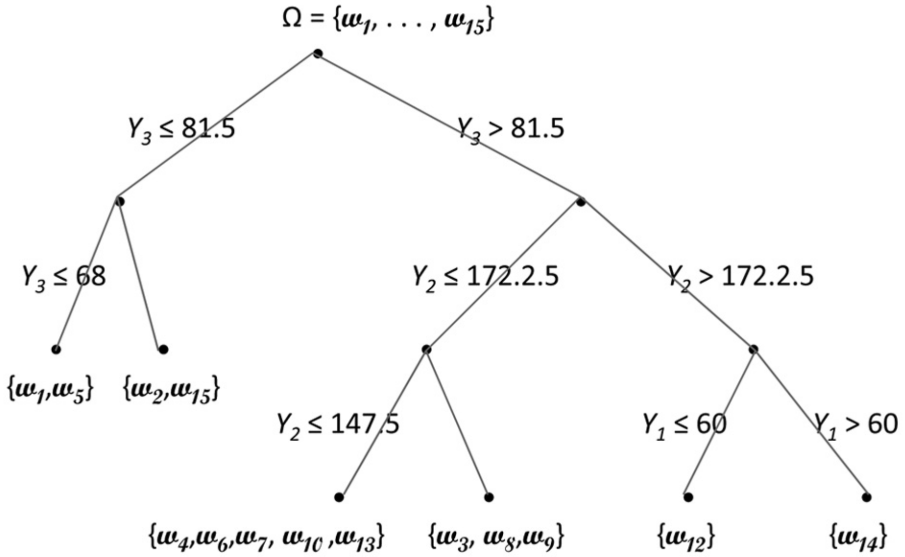

*Figure 6.* Divisive clustering on classical midpoints, $Y3 < Y2$.

## 5. Conclusion

With the advent of the modern computer, large datasets are ubiquitous. Aggregation across categories in these large datasets will inevitably produce symbolic data. Therefore, it is important that methodology be developed to analyse symbolic data. Some software exists and can be downloaded free from the web (`http://www.ceremade.dauphine.fr/%Etouati/sodas-pagegarde.htm`). Descriptions for their use can be found in Diday and Noirhomme-Fraiture (2008).

## References

[1] P. BERTRAND, F. GOUPIL, Descriptive statistics for symbolic data. In: *Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information from Complex Data* (eds. H.-H. Bock and E. Diday). Springer-Verlag, Berlin, (2000), 103–124.

[2] L. BILLARD, Dependencies and variation components of interval-valued data. In: *Selected Contributions in Data Analysis and Classification* (eds. P. Brito, P. Bertrand, G. Cucumel and F. de Carvalho), Springer, (2007), 3–12.

[3] L. BILLARD, A. CHOUAKRIA-DOUZAL, E. DIDAY, Symbolic principal components for interval-valued observations, (2007).

[4] L. BILLARD, E. DIDAY, *Symbolic Data Analysis: Conceptual Statistics and Data Mining*, (2006). John Wiley.

[5] L. BILLARD, E. DIDAY, Descriptive statistics for interval-valued observations in the presence of rules. *Computational Statistics*. 21, (2006), 187–210.

[6] M. CHAVENT, *Analyse de Donneés Symboliques Une Méthode Divisive de Classification*, Thése de Doctorat, Université Paris Dauphine, (1997).

[7] M. CHAVENT, A monothetic clustering algorithm. *Pattern Recognition Letters* 19, (1998), 989–996.

[8] M. CHAVENT, Criterion-based divisive clustering for symbolic data. In: *Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information from Complex Data* (eds. H.-H. Bock and E. Diday). Springer-Verlag, Berlin, (2000), 299–311.

[9] A. CHOUAKRIA, *Extension des Méthodes d'analyse Factorielle a des Données de Type Intervale*, Thése de Doctorat, Université Paris, Dauphine, (1998).

[10] E. DIDAY, M. NOIRHOMME-FRAITURE, (EDS.) *Symbolic Data Analysis and the SODAS Software*, (2008). Wiley.

*Received:* June, 2008
*Accepted:* September, 2008

*Contact address:*
Lynne Billard
Department of Statistics
University of Georgia
Athens, GA 30602, USA
e-mail: `lynne@stat.uga.edu`

LYNNE BILLARD is a University Professor at the University of Georgia and an Adjunct Professor at the Australian National University. She is a member and former president of American Statistical Association, as well as of International Biometric Society, and a member of Eastern North American Region (ENAR), International Statistics Institute and Institute of Mathematical Statistics. She has received awards for her work in statistics from American Statistical Association, University of Georgia and other statistical associations. Her research interests include stochastic processes with emphasis on model building, epidemic processes including AIDS research, sequential and time series analysis, statistical inference, symbolic and complex data analysis. She is a reviewer/referee for several journals, including Mathematical Reviews and Journal of American Statistical Association and an Associate Editor for Journal of Symbolic Data Analysis. She has published over 150 papers in international scientific journals and conference proceedings as well as several books.