Bernie Harris: A Life in Statistics
A session celebrating the life and work of
Bernard Harris (1926-2011):

Bernie Harris' Contributions to Cluster Analysis

Stanley L. Sclove*

Abstract

Aspects of Bernie Harris' activity in ASA and other organizations and his contributions to *cluster analysis* via the moment-preservation method and graph theory are outlined.

Key Words: Cluster analysis, graph theory, moment-preservation method; ASA Risk Analysis Section

1. Introduction

1.1 Risk Analysis Section

This paper outlines Bernie Harris' contributions to the theory of *cluster analysis*. However, in ASA, Bernie was better known not as a developer of cluster analysis but rather for activity in the field of reliability, in technical aspects of national security, and as a founder of the Risk Analysis Section in the early 1990s, along with Lee Abramson, Harry Martz, Lisa Weissfeld, and others. This section grew out of an ASA committee, advisory to the U.S. Nuclear Regulatory Commission. When the committee's work was done, Bernie and others had the idea of forming a related ASA section, rather than just disbanding. Bernie was instrumental in drawing up a charter and recruiting members. Among the Section's first officers were, in addition to Bernie and myself, Lee R. Abramson, U.S. Nuclear Regulatory Commission, Robert F. Bordley (General Motors Research Labs), Harry F. Martz (Los Alamos National Lab), and Lisa Weissfeld (Biostatistics, Graduate School of Public Health, University of Pittsburgh). Mark Vangel, a speaker in this session, was a section officer in the late 1990s. David Banks, organizer of this session, was a section officer for a number of years.

The variety of fields of section officers reflects the varied interests of its members, including medicine and biostatistics (safety and efficacy of treatments and pharmaceuticals), finance (risk management and insurance), and operations (product quality; worker safety; corporate and national security). Some of these interests overlap appreciably with those of other ASA sections.

*Stanley L. Sclove (Ph.D., Mathematical Statistics, Columbia University; B.A., Applied Honors Mathematics, Dartmouth College) is a Professor in the Department of Information & Decision Sciences at the University of Illinois at Chicago (UIC) and Coordinator of the Business Statistics Area of the PhD Program in Business Administration. In addition to UIC he has taught at Carnegie-Mellon, Northwestern, and Stanford. Sclove is Secretary/Treasurer of the Classification Society and the Representative of the Risk Analysis Section of the American Statistical Association to its Council of Sections. Address: Department of Information & Decision Sciences (MC 294), College of Business Administration, University of Illinois at Chicago, 601 S. Morgan St., Chicago, IL 60607-7124. The author thanks David Banks for organizing the session and Richard Johnson for chairing it.

1.2 Some of Harris' Classification/Cluster Analysis-Related Activities

Bernie's involvement in classification and clustering went back at least to 1976, when he was a member of the Organizing Committee of the Conference on Classification and Clustering chaired by John Van Ryzin. Subsequent involvement included giving a Short Course on Combinatorics at the 1984 Army Design of Experiments Conference (now the Army Conference on Applied Statistics), New Mexico State University, Las Cruces, and being on the Organizing Committee for the 1985 Army Design of Experiments Conference, Madison. Bernie regularly attended annual meetings of the Classification Society of North America (CSNA, now the Classification Society), beginning with the 1995 meeting in Denver.

The Classification Society is an interdisciplinary and international organization promoting the scientific study of classification and clustering (including systematic methods of creating classifications from data), and disseminating scientific and educational information related to its fields of interest. It is a member of IFCS, the International Federation of Classification Societies, consisting of about twenty national or regional classification societies.

Harris gave presentations of graph-theoretic methods of verifying clusters at CSNA 1995 Denver, CSNA-NT 1996 Amherst, and CSNA 2003 Tallahassee. Bernie hosted CSNA 2002 Madison. In 2002-2003, he chaired the CSNA Membership Committee.

2. Harris' Research Contributing to the Theory of Cluster Analysis

Cluster analysis is the grouping of similar individuals or objects. Bernie made a couple of types of contributions to this field, in terms of the *moment-preservation method* and the *graph-theoretic approach*. These contributions were in talks, esp. at meetings of the Classification Society, the International Federation of Classification Societies (IFCS) and the International Statistical Institute (ISI), and in short courses at such meetings, and in related papers.

3. Moment-Preservation Method

Harris' contributions relating to the moment-preservation method were made at various meetings *ca.* 2002-2003, including those of ISI and IFCS.

Given a dataset, or a theoretical distribution, a problem is to fit several, say K "clusters" *e.g.,* K = 2 or 3 (or more). An approach is to find mass points (cluster *centers*) $c_1, c_2, \ldots, c_K$ and probabilities $p_1, p_2, \ldots, p_K$ such that the corresponding discrete distribution $\Pr\{X = c_k\} = p_k,$ where X denotes the corresponding random variable, **matches the moments** of the given distribution. Recall that if two distributions have a number of moments close together, then the two distributions will be close in appropriate ways. The moment-preservation method may be viewed as solving for a discrete (K-valued) approximation to the given distribution. The unknowns are $c_1, c_2, \ldots, c_K$ and $p_1, p_2, \ldots, p_K$. There are $2K$ unknowns, with one constraint (the probabilities add to 1), hence $2K - 1$ free unknowns. This can be considered as an extreme case of the finite mixture model where the component distributions are one-point distributions. The moment-preservation method consists of computing values of the unknowns by matching $2K - 1$ moments $m'_1, m'_2, \ldots, m'_{2K-1}$. The values m'_k could, remember, be sample moments or specified values of true moments.

The method can be viewed as one of discrete approximation to a distribution (data or theoretical). It reminds me of the elicitation process in Decision Risk Analysis or in

CPM/PERT: Obtain high, medium, and low values of a risk variable or activity time, with their (subjective) probabilities. Here, however, we're asking the data, not a person.

3.1 Details of the Moment-Preservation Method for K = 2

There are two mass points c_1, c_2 and two probabilities $p_1\, p_2,$ but $p_2 = 1 - p_1$ so there are only three free unknowns. Denote the (raw) moments by m'_k. (Given data $x_i,\ i = 1, 2, \ldots, n,\ m'_k = \sum_{i=1}^{n} x_i^k$.) Equations for the unknowns are $p_1 + p_2 = 1,\ p_1 c_1 + p_2 c_2 = m'_1,\ p_1 c_1^2 + p_2 c_2^2 = m'_2,\ p_1 c_1^3 + p_2 c_2^3 = m'_3$. The situation looks like this.

$$\underset{c_1}{|}\underset{\quad p_1 \,\propto\, m'_1 - c_1}{------}\underset{m'_1}{|}\underset{\quad p_2 \,\propto\, c_2 - m'_1}{----------}\underset{c_2}{|}\text{---------} > x$$

Figure. K = 2 mass points c_1, c_2 with probabilities p_1, p_2. The quantitites c_1, c_2, p_1, p_2 are taken as functions of m'_1, m'_2, m'_3.

The solution comes out as follows (Harris 2003). Let

$$a = [(m'_1 m'_3 - m'^2_2)/(m'_2 - m'^2_1)]/2,\ b = [(m'_1 m'_2 - m'_3)/(m'_2 - m'^2_1)]/2.$$

(Note $m'_2 - m'^2_1$ is the variance, m_2.) Then

$$c_1 = [-b - (b^2 - 4a)^{1/2}]/2,\ c_2 = [-b + (b^2 - 4a)^{1/2}]/2,$$

$$p_1 = (c_2 - m'_1)/(c_2 - c_1),\ p_2 = 1 - p_1 = (m'_1 - c_1)/(c_2 - c_1).$$

An alternative way of writing the solution (Tabatabai and Mitchell 1984) is

$$c_1 = m'_1 - s\sqrt{p_2/p_1},\ c_2 = m'_1 + s\sqrt{p_1/p_2},$$

where $s = \sqrt{m_2} = \sqrt{m'_2 - m'^2_1},\ p_2 = [1 + A\sqrt{1/(A^2+4)}]/2, p_1 = 1 - p_2,$ the quantity $A = m_3/m_2^{3/2}$ being the skewness. I have not checked in detail, but use of cental moments $m_1 = 0, m_2 = m'_2 - m'^2_1, m_3 = m'_3 - 3m'_2 m_1 + 2m'^3_1$ from the outset may simplify things. In terms of the variance $m_2,$ the equations in m'_1 and m'_2 give $p_1 p_2 (c_1 - c_2)^2 = m_2 = m'_2 - m'^2_1$. The problem might be even further simplified by working in terms of deviations $d_1 = c_1 - m'_1$ and $d_2 = c_2 - m'_1$.

3.2 Moment-Preservation Method for K=3

There are mass points c_1, c_2, c_3 and probabilities $p_1, p_2\, p_3$. The equations are $p_1 + p_2 + p_3 = 1$ and $p_1 c_1^j + p_2 c_2^j + p_3 c_3^j = m'_j,\ j = 1, 2, 3, 4, 5$. I have not seen the details of a solution and do not pursue the matter here. Of course, given moments, a solution can be obtained by numerical methods.

3.3 Partitioning a Distribution

Contrast the moment-preservation method of Harris and others with *partitioning a distribution*, where "partitioning" means finding minimum mean squared error intervals. (The representative of an interval is then the conditional expectation, given that interval.) *E.g.,* for $K = 3$ for the Normal distribution, there is the "27% rule" (see, *e.g.,* Cox 1957): the proportions in the three intervals come out to be .27, .46, and .27. The intervals are approximately $(-\infty$, -0.613), (-0.613, +0.613), (+0.613, $\infty)$. The means of these intervals are approximately -1.22, 0, +1.22. The result used here is $\mathcal{E}[Z|a < Z < b] = \frac{\phi(b) - \phi(a)}{\Phi(b) - \Phi(a)}$,

Table 1: K=3: Comparing moment preservation and partitioning for the standard Normal distribution

moment-preservation	
centers $c_1 = -\sqrt{3},\ c_2 = 0,\ c_3 = \sqrt{3}$	probs $p_1 = 1/6, p_2 = 4/6, p_3 = 1/6$
partitioning solution	
centers $c_1 \approx -1.22,\ c_2 = 0,\ c_3 \approx 1.22$	probs $p_1 \approx .27, p_2 \approx .46, p_3 \approx .27$

where ϕ and Φ are the standard Normal p.d.f. and c.d.f., resp. Let's look at the results for K = 2 and 3. First consider $K = 3$. Say the given moments are the same as the standard Normal, the first one being 0, the second 1, the third 0, and the fourth 3. What results from the moment-preservation method, and how does it compare with partitioning?
Now consider the comparison for $K = 2$.

Table 2: K=2: Comparing moment preservation and partitioning for the standard Normal distribution

moment preservation	
centers $c_1 = -1, \quad c_2 = +1$	probs $p_1 = p_2 = 1/2$
partitioning	
centers $c_1 \approx -0.798, c_2 \approx +0.798$	probs $p_1 = p_2 = 1/2$

3.4 Stanines

It is worth mentioning another method of scoring, though it's not moment-preservation or partitioning. These are the *stanines*, for K = 9. They are based on intervals a half standard deviation wide. The values are 1, 2, 3, 4, 5, 6, 7, 8, 9. The boundaries are $-\infty$, -1.75, -1.25, -0.75, -0.25, +0.25, +0.75, +1.25, +1.75, ∞. The proportions in the nine categories are approximately 4, 7, 12, 17, 20, 17, 12, 7, and 4% . See Johari & Sclove (1976) for further discussion of stanines, and of partitioning for various families of distributions.

4. Harris' Contributions to Graph-Theoretic Clustering

Harris considered the graph-theoretic approach to clustering, giving presentations on the subject at CSNA 1995 Denver and CSNA-NT 1996 Amherst (a joint meeting of CSNA and the Numerical Taxonomy Society), with a presentation "A Comparison of Various Combinatorial Tests for Verification of Clustering." A *graph* is a collection of points (vertices) and edges (pairs of points, or lines). Write $G = (V, E)$. The vertices and edges can represent individuals and a binary relation defined on the set of individuals. A connected subgraph can be considered a cluster. Test statistics for clustering include the number of complete subgraphs of given order, say 2 or 3, and the number of components (a connected component of an undirected graph being a subgraph in which any two vertices are connected to each other by paths, and which is connected to no additional vertices.) I won't say much in detail about the graph-theoretic results except to mention some surprising probabilistic

results that convey the flavor of the area. Let $n = \#(V)$. Consider a *binary* graph: any two points are connected ("like") or not ("dislike"). A null model has link = "like", "dislike" with probabilities 1/2, independently. Let the random variable N be the size of largest complete subgraph (clique, cluster). The distribution of N is quite spiked. For example, for n = 1000 vertices, $\Pr\{N = 15\} > .8,$ that is, 15 is the mode and has very high probability. The expression for the mode in the case of $p = 1/2$ is $\log_2 n - 2\log_2 \log_2 n + 2\log_2 \frac{e}{2} + 1$. For $n = 10^{10},$ with $p = .25,$ $\Pr\{N = 30\} > .9997$ (see Matula 1977).

More generally, whether or not two vertices are linked depends on the distance between them. The definition of distance is tricky in the multivariate case. Statistical (Mahalanobis) distance should be used, but the appropriate one is based on the within-groups covariance matrix, and in cluster analysis, by graph-theoretic methods or otherwise, the groups are only just being formed. So one needs to iterate between tentative assignments to clusters and updating the within-groups covariance matrix and hence updating the inter-point distances.

There is still much challenging research that can be done, related to and building upon the interesting work that Bernie Harris did.

REFERENCES

References on the Moment-Preservation Method and Partitioning

Cox, D. R. (1957). Note on grouping. *Journal of the American Statistical Association,* **52,** 543-547.

Delp, E. J., and Mitchell, O. R. (1979). Image compression using block truncation coding. *IEEE Trans. on Communications,* **COM-27,** 1335-1342.

Johari, S., and Sclove, S. L. (1976). Partitioning a distribution. *Communications in Statistics,* **A5,** 133-147.

Lin, J. C., and Tsai, W.-H. (1994). Feature-preserving clustering of 2-D data for two-class problems using analytical formulas: an automatic and fast approach. *IEEE Trans. on PAMI,* **16 (5),** 554-560.

Pei, S.-C, and Cheng, C.-M. A fast two-class classifier for 2D data using complex-moment-preserving principle. *Pattern Recognition,* **29(3),** 519-531.

Tabatabai, A. J., and Mitchell, O. R. (1984). Edge location to subpixel values in digital imagery. *IEEE Trans. on PAMI,* **6 (2),** 188-201.

References on Graph-Theory and Clustering

Erdös, P., and Spencer, J. (1974). *Probabilistic Methods in Combinatorics.* Academic Press, Inc., New York.

Harary, F. (1969). *Graph Theory*. Addison-Wesley, Reading, PA.

Matula, D. (1977). Graph-theoretic techniques for cluster analysis algorithms. *Classification and Clustering,* J. Van Ryzin, ed., Academic Press, New York, 95-129.

References to Harris' clustering-related work, going back in time

Harris, B. (2005). Review of W. H. E. Day and F. R. McMorris, *Axiomatic Consensus Theory in Group Choice and Biomathematics,* SIAM Frontiers in Mathematics, Philadelphia, PA, **22**, 143-144.

Harris, B. (2003). The moment preservation method of cluster analysis. *Exploratory Data Analysis in Empirical Research: Proc. 25th Ann. Conf. of GfKl,* University of Munich, March 14-16, 2001, 98-103. Schwaiger and Opitz (eds.), Springer-Verlag, Inc., Berlin and New York.

Godehardt, E., and Harris, B. (1997). Asymptotic properties of random interval graphs and their use in cluster analysis. *Probability Methods in Discrete Mathematics,* 19-30. Kolchin, Kozlov, Pavlov and Prokhorov (eds.), VSP International Science Publishers (Zeist, The Netherlands).

Harris, B., Godehardt, E. and Horsch, A. (1995). Random multigraphs, classification and clustering. *New Trends in Probability and Statistics, Vol. 3: Multivariate Statistics and Matrices in Statistics (Proceedings of the 5th Tartu Conference),* 279-286. Tiit, Kollo and Niemi (eds.), VSP International Science Publishers (Zeist, The Netherlands).

Harris, B., and Park, C. J. (1994). A generalization of the Eulerian numbers with a probabilistic application. *Statistics & Probability Letters,* **20,** 37-47

Harris, B. (1968). Statistical inference in the classical occupancy problem: unbiased estimation of the number of classes. *JASA,* **63**, 837-847.

Harris, B. (1962). Determining bounds on expected values of certain functions. *Ann. Math. Statist.,* **33**, 1454-1456

Harris, B. (1960). Probability distributions related to random mappings. *Ann. Math. Statist.,* **31,** 1045-1062.