

Graphs, Risks, and Clusters: The Work of Bernie Harris

A memorial session celebrating the life and work of Bernard Harris (1926-2011)

Bernie Harris'

Contributions to Cluster Analysis

Stan Sclove

Stanley L. Sclove
Department of Information & Decision Sciences
University of Illinois at Chicago

www.uic.edu/~slsclove

slsclove@uic.edu

Some of Bernie's classification-related activities

- 1976 Member, Organizing Committee (John Van Ryzin, chair), Conference on Classification and Clustering, Madison
- 1984 Short Course on Combinatorics, Army Design of Experiments Conference, New Mexico State University, Las Cruces
- 1985 Organizing Committee, Army Design of Experiments Conference, Madison
- 1995 and thereafter. Regular attendance at annual meetings beginning with CSNA 1995 Denver
- 2002 Host, CSNA 2002 Madison
- 2002-2003 Chair, CSNA Membership Committee
- 2003 Short Course, Graph-Theoretic Methods of Verifying Clusters, CSNA 2003 Tallahassee

Research contributing to the theory of cluster analysis

- Bernie made a couple of types of contributions to cluster analysis.
 - Moment-preservation method of cluster analysis
 - Graph-theoretic approach to cluster analysis
 - (Mel's talk of course relates to the latter.)
- These were in
 - talks, esp. at Classification Society and at ISI,
 - short courses, and
 - papers

Moment-Preservation Method of Cluster Analysis

Method: Find mass points (cluster centers) m_1, m_2, \dots, m_K and probabilities p_1, p_2, \dots, p_K such that the corresponding discrete distribution $\Pr\{X = m_k\} = p_k$ where X denotes the corresponding r.v., matches the given distribution (e.g., data distribution).

moments matching $2K - 1$ moments

$K = 2$ two mass points m_1, m_2 and one prob

$$p_1 \quad (p_2 = 1 - p_1)$$

$$p_1 m_1 + p_2 m_2 = \text{1st moment}$$

$$p_1 m_1^2 + p_2 m_2^2 = \text{2nd moment}$$

$$p_1 m_1^3 + p_2 m_2^3 = \text{3rd moment}$$

discrete approx. to a distribution (data or theoretical)

Reminds me of elicitation process in Decision Risk Analysis

$K = 3$ three mass points m_1, m_2, m_3 and two probs

$$p_1, p_2 \quad (p_3 = 1 - p_1 - p_2)$$

$$p_1 m_1 + p_2 m_2 + p_3 m_3 = \text{1st moment}$$

$$p_1 m_1^2 + p_2 m_2^2 + p_3 m_3^2 = \text{2nd moment}$$

$$p_1 m_1^3 + p_2 m_2^3 + p_3 m_3^3 = \text{3rd moment}$$

etc. 4th and 5th moments

Partitioning a distribution

I want to compare and contrast the moment-preservation method with *partitioning a distribution*.

$$K = 3$$

Normal distribution

27% rule

.27, .46, .27

$z = +.613, 0, -.613$ means = 1.22, 0, -1.22

$K = 9$: Stanines

moment preservation(moment matching) different

Moment preserving clustering

$$K = 3$$

Say the moments are the same as the standard Normal.

What do you get, and how does it compare with partitioning?

That is, say 1st moment = 0, 2nd = 1, 3rd = 0, 4th = 3

Solution is centers $-\sqrt{3}, 0, \sqrt{3}$, probs $1/6, 4/6, 1/6$

Compared with $-1.22, 0, 1.22$, probs $.27, .46, .27$ for partitioning

Graph-theoretic verification of clustering

CSNA 1995 Denver

CSNA-NT 1996 Amherst A Comparison of Various Combinatorial Tests for Verification of Clustering

Test statistics considered include:

complete subgraphs of order m , e.g., $m = 2$ or 3

components

I won't say much in particular about the graph-theoretic results, preferring to leave that to Classification Society members such as Mel Janowitz and Buck McMorris who have been more directly involved with those developments.

But in the next couple of slides I'll mention some surprising results that convey part of the flavor of the area.

Surprising probabilistic results in graphs

Graph = Vertices, Edges, *i.e.*, $G = (V, E)$

$$n = \#(V)$$

Consider a binary graph: any two points are connected (“like”) or not (“dislike”)

null model: link = “like”, “dislike” with prob. $1/2$, independently

N = size of largest complete subgraph (clique, cluster):
Distribution of N is quite spiked.

$n = 1000$ $\Pr\{N = 15\} > .8$, that is, 15 is the mode and has very high prob.

formula for mode in case of $p = 1/2$:

$$\text{mode} = \log_2 n - 2 \log_2 \log_2 n + 2 \log_2 \frac{e}{2} + 1$$

$n = 10^{10}$: with $p = .25$, $\Pr\{N = 30\} > .9997$ (Matula)

- Erdős and Spencer (1974). *Probabilistic Methods in Combinatorics*. Academic Press, Inc., New York.
- Harary (1969). *Graph Theory*. Addison-Wesley, Reading, PA.
- Some references to Bernie's cluster analysis related work, going back in time:
- Harris (2005). Review of Day and McMorris, *Axiomatic Consensus theory in Group Choice and Biomathematics*, SIAM Frontiers in Mathematics, Philadelphia, PA, **22**, 143-144.
- Harris (2003). The moment preservation method of cluster analysis. *Exploratory Data Analysis in Empirical Research: Proc. 25th Ann. Conf. of GfKI*, University of Munich, March 14-16, 2001, 98-103. Schwaiger and Opitz (eds.), Springer-Verlag Inc (Berlin; New York).

- Godehardt and Harris (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, Godehardt and Horsch (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, Godehardt, and Horsch (1995). Random multigraphs, classification and clustering. *Multivariate Data Analysis. 7th ed.* Prentice Hall (Pearson Education), Upper Saddle River, NJ.

Bibliography, cont'd

- Harris and Park (1994). A generalization of the Eulerian numbers with a probabilistic application. *Statistics & Probability Letters*, **20**, 37-47
- Harris (1968). Statistical inference in the classical occupancy problem: unbiased estimation of the number of classes. *JASA*, **63**, 837-847. Keywords: Multinomial distribution; Stirling numbers of the second kind
- Harris (1962). Determining bounds on expected values of certain functions. *Ann. Math. Statist.*, **33**, 1454-1456
- Harris (1960). Probability distributions related to random mappings. *Ann. Math. Statist.*, **31**, 1045-1062
- Matula (1977). Graph-theoretic techniques for cluster analysis algorithms. *Classification and Clustering*, J. Van Ryzin, ed., Academic Press, New York, 95-129.

(end of talk)

Thank you !