

Clustering in the Service of the Public's Health

Leslie Lenert

*VA Medical Center 111N1, Section on Health Services Research
3350 La Jolla Village Drive
San Diego, CA 92161 U.S.A.
llenert@ucsd.edu*

Alfred Lin

*Stanford University, Department of Statistics
Sequoia Hall
Stanford, CA 94305-4065 U.S.A.
alfred@stat.stanford.edu*

Richard Olshen

*Stanford University, Department of Health Research and Policy
HRP Redwood Building
Stanford, CA 94305-5405 U.S.A.
olshen@stat.stanford.edu*

Catherine Sugar

*University of Southern California, Department of Information and
Operations Management
3670 Trousdale Parkway, Bridge Hall Room 400 G
Los Angeles, CA 90089-1421 U.S.A.
sugar@usc.edu*

1. Abstract

Our purpose here is to provide an overview of some recent developments in cluster analysis and their applications to health services research. To illustrate, we show how k -means clustering can be used to partition a population of patients with depression into clinically relevant health states and to analyze how they change over time. Among the issues discussed are preprocessing of data, identifying the "right" number of clusters, implementation via the Lloyd algorithm, reverse engineering of health state descriptions from clusters, and longitudinal modeling of patients' transitions.

2. Introduction and the Depression Panel of the Medical Outcomes Study

Because resources for medical care are increasingly expensive and in some cases scarce, governments and societies must confront the problem of allocating them equitably. As an aid to setting policy it can be worthwhile to measure the *quality of life* for patients in different *health states*. By the latter we mean a partition of the space of "attributes," or "dimensions of health," that adequately describe a population of medical interest. When utilities are measured for the different states, policy can be based upon gains achieved as patients move from one to another (Lenert et al., 1999).

The population of this study is a sub-group of the Medical Outcomes Study (MOS), a large trial of the effectiveness of "managed care" medical practice conducted in the United States in the late 1980s. Participants were from a national screen of 22,239 patients of primary care medical

practices. After a careful two-step process, 2,194 were found to have symptoms of depression; and of them, 1,772 agreed to undergo a diagnostic interview. Thereby, 772 with clinical depressive disorders were identified.

To create a final panel of 974 patients, researchers employed weighted probability sampling, over-sampling the elderly and patients with cardiac disorders and depression and giving preference to patients who had already demonstrated willingness to complete the MOS health status questionnaire. Of the 974, we have data on the initial values of our "attributes," mental and physical scores, for 602; of these, there are follow-up scores for two subsequent years for 218. As of this writing it is not possible to say to what extent our inferences are biased because of the way in which patients actually became part of the sample of data we analyzed. See (Wells et al., 1996) and (Lenert et al., 1999) for details regarding the data.

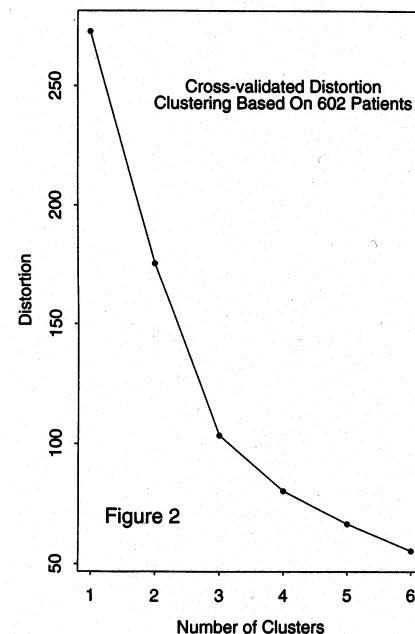
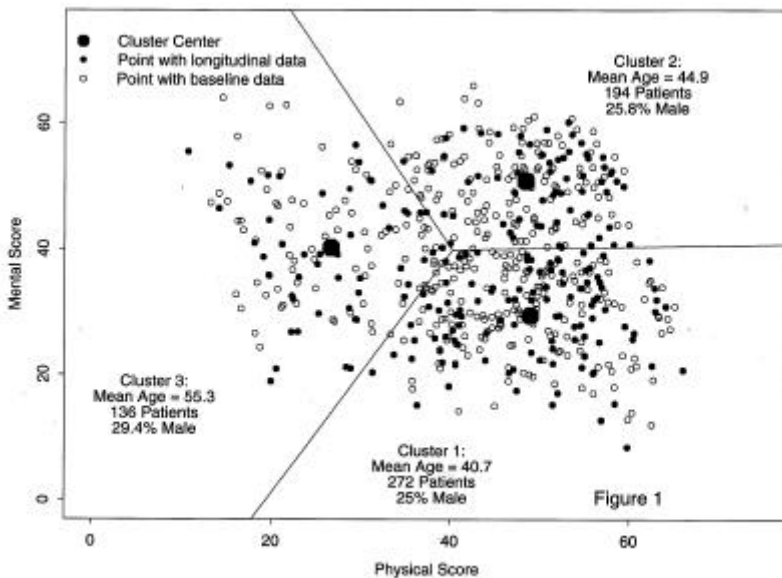
Typically in the past the dimensions of health have been selected by subject matter experts, with each dimension divided into several equally spaced levels. This entails a grid, for which cells represent health states and centers respective prototypical patients. This approach is patently inefficient. See (Sugar et al., 1998) and (Sugar, 1998). Some compression of necessarily subjectively chosen dimensions of health can be achieved by principal components, factor analysis, or another system of scoring. For us, data gathered by questionnaire have been reduced to simply "mental score" and "physical score." Our goal is to reduce them further, and therefore in the spirit that our overall population might consist of several distinct sub-populations, we cluster. The cluster centers then replace the grid centers as prototypical patients. Since algorithms for so doing are based upon notions of distance between vectors of data, scaling can be an issue. It is not in our application.

3. How to Cluster and How Many Clusters

We prefer k-means clustering in part because it makes no distributional assumptions save second moments being finite, so it is applicable to many problems. It has proven to be enormously useful in *lossy coding* of images and speech (see Gersho and Gray, 1992), and the extremal problem it solves is intuitively appealing. Suppose there are to be k clusters with respective centers c_1, \dots, c_k . $\|\cdot\|^2$ denotes "squared Euclidean length." Our data are $x_i: i=1,2, \dots$, and $ave\{.\}$ denotes the average of the numbers in $\{.\}$. Then the c_1, \dots, c_k we choose minimize $ave_i \{min_j \|x_i - c_j\|^2\}$. That is, the cluster centers minimize *distortion*, mean-squared distance from data to their respective centers. While there is no closed form solution to the minimization, S. Lloyd's celebrated 1957 alternating descent algorithm (Gersho and Gray) will always converge. Except in special circumstances there can be multiple minima, so it is imperative to try sufficiently many initializations to ensure to the extent possible that a global minimum has been found. Our implementation is with S-PLUS (Venables and Ripley, 1998).

The parameter k remains to be chosen. As was mentioned, in our application we hope that k is, at least approximately, the number of components of a mixture model that pertains to our data. In that sense we depart from applications to engineering, more specifically *vector quantization*, where the constraints are on bit rates of resulting codes. Thus, approximately $\log_2 k$ bits are required to identify each of the k clusters. We plot four-fold cross-validated distortion versus number of clusters in Figure 2. Were there k genuinely distinct components to the mixture, the curve - which is necessarily decreasing but possibly for noise in the cross-validation - should exhibit a "kink" when the abscissa is k . The kink appears to be at $k=3$. Two different analytical criteria also give "3" as "the right number of clusters." They have agreed in all examples we have studied. One (Sugar, 1998; Sugar et al., 1998) involves choosing k to maximize an ANOVA F-statistic for testing the null hypothesis the "distortion curve" is linear versus the alternative it is a linear spline with single knot at

k. The other (Lin et al., 1999) involves maximizing a suitably spaced second difference of the distortion curve. Information theory's rate-distortion theory suggests that in two dimensions a plot of the reciprocal of distortion against number of clusters should be informative (Sugar, 1998). It is. The smallest cluster in terms of numbers of members is interesting. Its subjects have lower physical but intermediate mental scores. They are considerably older than are members of the other two groups, and have a slightly higher percentage male.



4. Longitudinal Data and Transitions Among States

As was stated, we have data gathered approximately annually for three successive years for 218 subjects. We used cluster membership for them to describe changes in health state, trying to learn, for example, whether to within noise their longitudinal paths are consistent with a Markovian hypothesis. We computed transition matrices for states defined by clusters. If we denote the respective cluster memberships for a patient at time i , $i = 0, 1, 2$ by $X(i)$, then it is clearly necessary and sufficient for the Markov hypothesis to hold that $X(0)$ and $X(2)$ be conditionally independent given $X(1)$. Now $X(1)$ takes three distinct values, and therefore there are three tests to perform. For each we employed the ordinary chi-square test statistic appropriate to a test for independence. For each of the three statistics there were at least four cells with expected value less than 5 under the model of (conditional) independence, so we did 1,000 permutations each and found the permutation distribution of each of the three statistics. (see Lazzeroni and Lange, 1997 for details.) Despite our concerns, the asymptotic chi-square and permutation attained levels of significance were nearly identical. Chi-square was 4.15 given $X(1)=1$, 10.65 given $X(1)=2$, and 20.18 given $X(1)=3$. Respective attained levels of significance are .41, .03, and $<.001$. Thus the extent to which the process is Markovian depends upon what $X(1)$ is. From a quick look at the data especially, it is clear that if the Markov hypothesis fails it is because there are some individuals who do not move, who are in the terminology of labor economics "stayers." Indeed, well over 40% of sample paths are of (i,i,i) form. It follows that our data may be not Markovian, but rather of "mover-stayer" form, that is, a mixture of Markov chains, one with each of the three states absorbing and the other unrestricted. (See Frydman, 1984.) This possibility seems plausible from context and is an area for our further research.

REFERENCES

Frydman, H. (1984). Maximum likelihood estimation in the mover-stayer model. *Journal of the American Statistical Association* 79, 632-638.

Gersho, A. and Gray, R.M. (1992). *Vector Quantization and Signal Compression*. Kluwer Academic Publishers. Boston.

Lazzeroni, L.C. and Lange K. (1997). Markov chains for Monte Carlo tests of genetic equilibrium in multidimensional contingency tables. *Annals of Statistics* 25, 138-168.

Lenert, L.A., Sherbourne, C.D., Lawrence, W.F., and Wells, K.F. (1999). A health index for depressive illnesses based on the SF-12. Submitted for publication.

Lin, A., Lenert, L.A., Hlatky, M.A., McDonald K.M., Olshen, R.A., and Hornberger, J. (1999). Clustering and the design of preference-assessment surveys in health care. *Health Services Research*, to appear.

Sugar, C.A. (1998). *Techniques for Clustering and Classification with Applications to Medical Problems*. Ph.D. Dissertation, Department of Statistics, Stanford University. Stanford, CA.

Sugar, C.A., Sturm, R., Lee, T.T., Sherbourne, C.D., Olshen, R.A., Wells, K.B., and Lenert, L.A. (1998). Empirically defined health states for depression from the SF-12. *Health Services Research* 33, 911-928.

Venables, W.N. and Ripley, B.D. (1998). *Modern Applied Statistics with S-PLUS* (Corrected Second Printing). Springer Verlag. New York.

Wells, K., Sturm, R., Sherbourne, C., and Meredith, L. (1996). *Caring for Depression*. Harvard University Press. Boston.

FRENCH RÉSUMÉ

Nous nous proposons de présenter un panorama de récents développements en cluster analysis et des applications qui en découlent, à la recherche sur l'allocation des services de santé. Afin d'illustrer notre propos, nous montrons comment la technique du k -means clustering peut être employée pour identifier, dans une population de patients souffrant de dépression, des groupes correspondant à des états de santé qui ont une signification clinique. Il est par ailleurs possible d'analyser comment l'état de santé de ces patients évolue dans le temps. Le pré-traitement des données, l'identification du nombre adéquat de clusters, l'implémentation à l'aide de l'algorithme de Lloyd et al modélisation longitudinale de l'évolution des patients figurent parmi les sujets traités.