CLUSTER AND DISCRIMINANT ANALYSIS ON TIME-SERIES AS A RESEARCH TOOL

James K. Galbraith and Lu Jiaqing

Galbraith@mail.utexas.edu, Jqlu@uts.cc.utexas.edu

LBJ School of Public Affairs

The University of Texas at Austin

Austin, Texas 78713

UTIP Working Paper Number 6

ABSTRACT

This paper presents a procedure for studying industrial performance and related issues such as changes in the wage structure. This procedure combines cluster analysis and discriminant analysis as a package, and applies this package to time series data. This enables us to organize industrial data into groups with similar wage or performance histories and then to extract summary time-series showing the main pattern of variation in performance between groups.

JEL Classification Codes: C1,C4,C63,E32,J31,L13,L16,O38; Key Words: Cluster, Discriminant, Time-Series.

First Draft: February 1997
This Paper: January 30, 1999

1. Introduction

This paper presents a procedure for studying industrial performance and related issues such as change in the wage structure. The procedure combines cluster and discriminant analysis, and applies them to time series data to explore, first, the group pattern, and then the forces that promote the formation of that group pattern. This procedure can be applied to many fields for which time series data are available on a single key measure of behavior for a large number of related entities -- for example wages by industry or occupation or expenditure by account in a study of government budgets.

The use of dated information as a tool for classification is well-established in disciplines such as geology, paleontology, archeology, and even in biology and developmental psychology. For example, Chiodi (1989) uses time-series height and arm span data to classify children, and Hirsch and DuBois (1991) classify children based on the similarities in behavior through time. So far as we know, however, the present sequence of cluster and discriminant analysis on multi-variate time series data had not been done until Galbraith and Calmon's work on industrial wage rates (1990, 1994, 1996). Further work in this area includes Galbraith (1998) and Ferguson and Galbraith (1998) on American industrial performance and wage structures, Galbraith and Kim (1998) on Korean industrial policy, and Calistri (in progress) on industrial structures and wage change in the OECD. Very recently, Kakizawa, Shumway and Taniguchi (1998) have published a full development and empirical application of closely related techniques to a problem in seismology, namely that of distinguishing earthquakes from nuclear explosions.

A more formal presentation of our procedure is now needed. In Section 2, we will explain the theoretical linkage between wages or earnings and industrial performance underlying the use of the former as attribute variables in a cluster-and-discriminant analysis. In addition, we will propose a measurement of industrial performance, total payroll per production hour, which has practical and theoretical advantages in certain cases. Section 3 will present the method of cluster-discriminant analysis, and section 4 will offer an example to illustrate step-by-step the application of the procedure.

2. Wages, Industrial Performance and the P-measure

The first step for cluster and discriminant analysis is to choose characteristic or attribute variables for the objects to be clustered. For analyses of industrial performance, Galbraith and Calmon propose the year-to-year change of average wages by standard industrial classification (SIC) category as a performance measure. The notion of industry-specific labor rents is helpful in motivating this choice. If capital markets clear, but labor markets don't, we should expect that rates of return on investment equalize across industries but that rates of pay will not. Hence, there will be industry-specific pay differentials. There is a persuasive body of information to this effect, summarized in Katz and Summers (1989) and strongly seconded in an important paper by Blanchflower, Oswald and Sanfey (1996). The burden of this analysis is that scarce factors, such as human skill, eventually capture the monopoly rents that an industry's market position may earn.

The Katz-Summers argument is essentially static, based on the degree of monopoly power enjoyed by an industry at a particular moment of time. But if degrees of monopoly change (and who would deny it?) then surely the industry-specific labor rents will also change. And if that is so, the patterns of change through time can serve as markers of similarity and difference in economic performance among and between industries. When a pattern of wage changes is essentially identical in two separate industrial subclassifications over a long period of time, it becomes unlikely that this is accidental. Instead, similar effects result from structural characteristics that produce like reactions to common causes. That being so, patterns of similar effects can be used to classify industries according to structural similarity, even if one has no direct measure of what the structural similarities may be.

A drawback of the change in average wage rates by industrial group as a performance measure is that there may occur intra-industry distributional shifts such as from production workers to salaried employees, and these may confound the use of wage change as a proxy for industrial performance. When data is available, we therefore suggest *total payroll per production worker hour*, a measure that Galbraith (1998) calls the "P-measure," as a better measurement of industrial performance.

The P-measure is closely related to industrial productivity, and its change is closely related to industrial productivity growth. It also reflects the changes in market power and position, such as improvements in technology or reductions in cost due to outsourcing

(across industrial lines or national boundaries). In contrast to the change in hourly production worker wages, the P-measure is unaffected by shifts in the allocation of earnings within an industry, say from production workers to non-production workers or vice versa, which are not necessarily related to industrial performance. The P-measure is also a better measure of industrial performance over time than say, value of total shipments per hour would be, since the latter may be affected by pass-through of variations in materials prices, while the P-measure is not.

The use of percentage rates of change of our performance variable, rather than the level, has an economic justification and also technical advantages. From an economic standpoint, we are interested in the change in performance through time; this is a matter of rates of change rather than of initial levels. As a technical matter, since cluster analysis is sensitive to the units and scale of variables, a change in scale of one of the measures can change the implicit weight of the characteristic being measured, and hence the group structure. But if we use annual percentage change, we can be free of units and scale problems because each measurement is of the same form as any other, and scale-altering forces such as inflation do not affect our analysis.

3. Cluster-Discriminant Analysis

Cluster analysis is a technique used to classify objects into homogeneous groups or clusters based on their similarities in some attribute or characteristic variables. For example, the technique may be used to group flowers by visual characteristics, or students according to their pattern of scores on a series of tests. For details about cluster analysis, please refer to Lorr (1983), Anderberg (1973), Aldenderfer and Blashfield (1984), and Everitt (1974). Informative use of cluster analysis has recently been made in this journal by Hirschberg and Slottje (1994).

Suppose there are N industrial sectors or objects, such as those based on SIC codes at the 3- or 2-digit level. The attribute variables are percentage rates of change of our performance measure -- the P-measure in this case -- for each year. Each element ? L(i, t) in the matrix **? L** of order N x (T-1) can therefore represent the annual percentage change in performance of industry i at year t. The complete row of values across variables (years) is called the industry's profile, and cluster analysis classifies industries into groups by the similarity between profiles. Geometric similarity can be measured by Euclidean distance, which is defined as

$$d_{ij} = \sqrt{S_t}(? L_{it} \& ? L_{jt})^2$$

where d_{ij} is an element of **D**, the N x N matrix of distances between objects.

Deciding the structural model for the expected clusters, as well as the clustering method or algorithm which can generate this cluster structure, is the most important step. There are two kinds of cluster structure, the chained cluster and the compact cluster. According to Lorr (1983, p. 18), a chained or serpentine cluster is a category of objects in which every member is more like *one* other member than it is like any object not in the category. The compact or ellipsoidal cluster is a category of which all members are more like *every* other member than they are like objects in any other subgroup; such clusters exhibit "high mutual similarity." For the comparison of industrial groupings, we have a strong preference for high mutual similarity, and therefore for compact structure.

There are many available clustering algorithms and still more are under development. All of these methods fall into two major categories, single-level cluster methods and multilevel hierarchical methods, with the choice typically depending on the problem. But according to Lorr (1983, p. 20), hierarchical methods are often preferred. One of the reasons is that hierarchical methods tend to reflect a developmental or evolutionary pattern or sequence. For this reason, most biologists favor this kind of method, and since our data are historical in character, so do we.

We choose Ward's Method, a hierarchical method, also known as the Minimum-Variance Method (Ward, 1963). This method begins by treating each object as a separate group, so that no information is missing. At each step afterwards, group or cluster numbers are reduced from N to N-1, N-2, ... 2, 1 in such a way that a specified objective

function is minimized at each step. The objective function Ward chose is the increase in the total error sum of squares -- or the geometric distance from each data point to the center of its cluster -- due to the merger of two objects, clusters, or objects-and-clusters to form a new, more encompassing cluster. Details are in Anderberg (1973, pp 147-8). In our case, the error sum of squares for cluster g is:

$$E_g = R_g - \frac{1}{m_g} \sum_t T_{tg}^2 \qquad (2)$$

where

$$T_{tg} = \sum_i \Delta L_{itg} \qquad (3)$$

is the sum of changes of the P-measure at year t for industries in the g-th cluster; and

$$R_g = \sum_t \sum_i \Delta L_{itg}^2 \qquad (4)$$

is the sum of squared changes of the P-measure in all years for all industries in the g-th cluster. Here, m_g is the number of industries in cluster g, and ΔL_{itg} is the change of the P-measure at time t for the i-th of m_g industries in the g-th cluster.

The increase in the total error sum of squares due to the merger of clusters g and h to form the new cluster k is

$$\Delta E_{gh} = E_k - (E_g + E_h) \qquad (5)$$

By Ward's method, in each step, cluster g and h will be merged to form a new cluster k if they satisfy

$$\text{Min}(? E_{gh}) \tag{6}$$

for all possible values of g and h, contingent on the clustering achieved at the previous step. The next question is when to stop, and this is essentially a matter of deciding when "too much" information is being lost by forcing dissimilar objects to associate; that is, when the minimum increase in the error-sum-of-squares has become too large. The semi-partial R^2 criterion can be used to choose this point (cf. Lorr 1983, p. 99). There is an element of judgment about applying this criterion, as we shall see below.

Discriminant analysis is a multivariate technique which is used to examine the differences between two or more groups of objects with respect to several variables. The basic elements of a discriminant analysis are objects, group membership of objects and a set of attribute or characteristic variables. The goal of the analysis is to find discriminant function(s) which can differentiate groups, that is, can make group-means on the function(s) differ widely. For those who are interested in more technical details on discriminant analysis, please refer to Tatsuoka (1988) and Klecka (1980).

Since we get cluster membership from cluster analysis which classifies industries by annual changes of the P-measure, an intuitive way to construct a discriminant function is to linearly combine these annual changes, that is:

$$F = a_1? L_1 + a_2? L_2 + a_3? L_3 +...+ a_{t-1}? L_{t-1} \quad (7)$$

where ? L_t is change of the P-measure in year t, i = 1, 2, ..., T-1.

To get the coefficients a_1, a_2, ... $a_{T-1,}$ or simply the (t-1) vector **a,** consider the (T-1) dimensional matrices **B** and **W**, where the diagonal element of **B** is the sum of squared differences between groups for each year t=1 to T-1 and the diagonal element of **W** is the within-group sum of squared differences; off-diagonals are cross-products. The problem is to find **a** so that F differentiates group-means in such a way that minimizes within-group differences (**W**) and simultaneously maximizes between-group differences (**B**). This can be implemented by solving a maximization problem:

$$Max[(a'Ba)/(a'Wa)] \quad (8)$$

Applying differential calculus to (8), we get:

$$\partial/\partial a\,[a'\mathbf{B}a/a'\mathbf{W}a] = 0 \quad (9), \text{ or}$$

$$[\mathbf{W}^{-1}\mathbf{B} - ?\mathbf{I}]a = 0 \quad (10)$$

Here ? is the vector of eigenvalues associated with matrix $\mathbf{W}^{-1}\mathbf{B}$. Since **W** and **B** can be calculated from our data set, we can solve Equation (10) to get eigenvalues and associated eigenvectors **a**. In the literature of discriminant analysis, **a** are often called canonical roots of the discriminant functions. Suppose there are G groups from cluster analysis, the number of eigenvalues and eigenvectors is determined by the rank of $\mathbf{W}^{-1}\mathbf{B}$ (Klecka, 1980):

$$\text{Min}[(G-1), (T-1)] \tag{11}$$

Since in most cases, if not all, we run discriminant analysis with more years than the number of clusters, we can use (G-1) as the number of eigenvalues and eigenvectors safely. The discriminant function associated with a bigger eigenvalue should more powerfully explain the differences among groups than those with smaller eigenvalues; the eigenvalue ? is therefore called the discriminant criterion (Tatsuoka 1988 p. 213).

How many eigenvectors one should actually use depends on how many are needed to account for the between-group variations. If, for example, the first three functions associated with the three biggest eigenvalues can account for a high fraction of all discriminatory power, we are confident that three functions are sufficient. In practice, the acceptable proportion, as well as the optimum number of functions, depends on the problems at hand.

If we derive G-1 discriminant functions, they are:

$$F_1 = a_{11} ? L_1 + a_{12} ? L_2 + ... + a_{1(t-1)} ? L_{t-1}$$

$$F_2 = a_{21} ? L_1 + a_{22} ? L_2 + ... + a_{2(t-1)} ? L_{t-1}$$

.....

$$F_{G-1} = a_{(G-1)1} ? L_1 + a_{(G-1)2} ? L_2 + ... + a_{(G-1)(t-1)} ? L_{t-1} \quad (12)$$

in which the **a**s are known. Obviously, the **a** are actually a set of weights on annual changes of the P-measures. The weights, which are components of an eigenvector, form a (T-1) dimension vector in space. *But if we assume that the weight associated with a specified annual change is that-year-specific, we can reasonably assume that the sequence of weights for each year also form a time series. By doing so, a (T-1) dimension vector in space is converted into a one-dimensional time series with (T-1) values at (T-1) different time points.* This is the beauty of the present procedure and the one feature that we claim to have pioneered.

Based on the theoretical and empirical background of the problem at hand, we can next try to use historical economic data to match and identify these eigenvectoral time-series. For example, we might match the weights of F_1 with a GNP time series, and F_2 with the interest rate, and so on.. If we are successful in making such a match for some subset of our G-1 eigenvectors, we can infer that in those cases we have identified the economic forces underlying the differentiation of group behavior, and, because we have the eigenvalues, we also know the relative contribution of each force.

An intuitive way to show how discriminant functions i and j differentiate clusters is to plot their scores on functions i and j on a Root i -- Root j coordinate. The scores for each group can be calculated by simply substituting each year's group-mean of the P-measure change of each group into Equation (7). To show how discriminant functions differentiate each individual industry, follow a similar procedure using the annual values of change in the P-measure for each industry. The resulting scores, which are the vector inner-products of the weighting function **a** and the vector of rates of change in our performance measures, are scalars which can be plotted against a variety of variables to reveal cross-section relationships.

We will not present empirical examples in this paper, both to conserve space and avoid duplication with other work in the UTIP series. Presently the most complete applications in print are Galbraith (1998), Ferguson and Galbraith (1998), and Galbraith and Kim (1998). Additional applications will be made available on the UTIP site (http://utip.gov.utexas.edu) as they become available. In our repeated experience, this technique is useful for many social science problems where the essential problem is to identify the principal patterns of movement in time-series data sets involving blocks of observations on similar entities, such as firms or industries, where the appropriate group structure must be derived from the data itself.

ACKNOWLEDGMENTS

This research was sponsored in part by the Twentieth Century Fund (now The Century foundation) and by the Center for the Study of Western Hemispheric Trade. We thank the Ford Foundation and the Jerome Levy Economics Institute for supporting the publication of this paper in the present electronic series. We thank our fellow participants in the project on wage structures at the University of Texas at Austin, now the University of Texas Inequality Project: Maureen Berner, Amy Calistri, Pedro Conceiçao, Vidal Garza Cantu and Junmo Kim, for advice and encouragement. We especially acknowledge the important contribution of Paulo Du Pin Calmon to the original development of the procedure described in this essay.

REFERENCES

Aldenderfer, M. and Blashfield, R., 1984. *Cluster Analysis*. London: Sage.

Anderberg, M. R., 1973. *Cluster Analysis for Applications*, New York: Sage.

Blanchflower, D., Oswald, A, and Sanfey, P., 1996. "Wages, Profits and Rent-sharing," *Quarterly Journal of Economics*, February, **1**, 227-251.

Calistri, A., in progress. "Industrial Structures in the OECD: New Evidence from the STAN Database.".

Chiodi, M., 1989. "The Clustering of Longitudinal Multivariate Data when Time Series Data Are Short," in R. Coppi and S. Bolasco eds., *Multiway Data Analysis.* Amsterdam: North-Holland.

Ferguson, T. and Galbraith, J.K., 1998. "The American Wage Structure, 1920-1947. Bard College: Jerome Levy Economics Institute Working Paper.

Galbraith, J. K., 1998. *Created Unequal: The Crisis in American Pay*. New York, Free Press.

Galbraith, J.K., and Calmon, P., 1996. Wage Change and Trade Performance in US Manufacturing Industries, *Cambridge Journal of Economics*, 20, 433-450.

Galbraith, J.K., and Calmon, P., 1994. "Industries, Trade and Wages," in M. Bernstein and D. Adler (eds.), *Understanding American Economic Decline*, Cambridge: Cambridge University Press.

Galbraith, J.K., and Calmon, P., 1990. "Relative Wages and International Competitiveness," Austin: Lyndon B. Johnson School of Public Affairs Working Paper #56, January.

Galbraith, J.K. and Kim, J., 1998. "The Legacy of the HCI: An Empirical Analysis of Korean Industrial Policy," Journal of Economic Development (Seoul), Vol. 23, No. 1, June, 1-20.

Hirsch, B. and DuBois, D., 1991. "Self-Esteem in Early Adolescence: The Identification and Prediction of Contrasting Longitudinal Trajectories," *Journal of Youth and Adolescence*, 20, 1, 53-72.

Kakizawa, Y., Shumway, R.H. and Taniguchi, M., 1998. "Discrimination and Clustering for Multivariate Time Series," Journal of the American Statistical Association, Vol 93, No. 441, pp. 328-340.

Katz, L and Summers, L., 1989. "Industry Rents: Theory and Evidence," *Brookings Papers on Economic Activity: Microeconomics*, 209-290.

Klecka, W., 1980. *Discriminant Analysis*, Beverly Hills: Sage.

Lorr, M., 1983. *Cluster Analysis for Social Scientists*, San Francisco: Jossey-Bass.

Hirschberg, J and Slottje D., 1994. "An Empirical Bayes Approach to Analyzing Earnings Functions for Various Occupations and Industries," *Journal of Econometrics*, Vol 61, March: 65-79.

Tatsuoka, M .M., 1988. *Multivariate Analysis*, New York: Macmillan.

Ward, J. H. Jr. , 1963. "Hierarchical Grouping to Optimize an Objective Function," *Journal of the American Statistical Association, **58**, 236-244.*