INTERDISCIPLINARY CENTER
FOR ECONOMIC SCIENCE
GEORGE MASON UNIVERSITY

Exploring Network Behavior Using Cluster Analysis

Rong Rong and Daniel Houser

October 2014

Discussion Paper

Interdisciplinary Center for Economic Science
4400 University Drive, MSN 1B2, Fairfax, VA 22030
Tel: +1-703-993-4719   Fax: +1-703-993-4851
ICES Website: http://ices.gmu.edu
ICES RePEc Archive Online at: http://edirc.repec.org/data/icgmuus.html

Exploring Network Behavior Using Cluster Analysis

Rong Rong[a] and Daniel Houser[b]

[a] Department of Economics, Weber State University, rongrong@weber.edu

[b] ICES, Department of Economics, George mason University, dhouser@gmu.edu

February, 2014

Abstract: Innovation occurs in network environments. Identifying the important players in the innovative process, namely "the innovators", is key to understanding the process of innovation. Doing this requires flexible analysis tools tailored to work well with complex datasets generated within such environments. One such tool, cluster analysis, organizes a large data set into discrete groups based on patterns of similarity. It can be used to discover data patterns in networks without requiring strong ex ante assumptions about the properties of either the data generating process or the environment. This paper reviews key procedures and algorithms related to cluster analysis. Further, it demonstrates how to choose among these methods to identify the characteristics of players in a network experiment where innovation emerges endogenously.

JEL classification: C46, C81

Keywords: cluster analysis, k-means algorithm, innovation, networks, laboratory experiment

I. Introduction

Innovation often occurs in networked environments. A player in these networks may play the role of either “innovator” or “follower”. To identify the characteristics of players is a crucial first step towards understanding the process of innovation and economic growth. More generally, researchers in social science often need to classify individual behavior data into meaningful groups so that we can better describe the differences and similarities among individuals.

When natural features, such as gender, age or income, are obviously driving the change of the variable of interest (in our case the level of innovation) then one can form hypotheses regarding the nature of differences between groups and, subsequently, use statistical methods such as regression analysis to validate or reject these hypotheses. Unfortunately, such a priori interpretations of data are not always available. An advantage to cluster analysis is that it does not require strong ex ante assumptions about the data generating process. As a numerical method for classification, cluster analysis allocates large and complicated datasets into discrete groups.

As early as the 1920s, psychologists were interested in the composition of ability. Some claimed all ability could be explained using two factors (Spearman, 1904), others argued that there were more divisions, such as verbal, arithmetic, memory and spatial. Left unanswered were the number of low-level abilities and the way they relate to each other. This question inspired Robert Tryon to develop the first cluster analysis algorithm, then leading to the development of the first cluster analysis software BC TRY in the 1960s (Tryon, 1932; Tryon,1935; Tryon and Beiley, 1966).

Since then, numerous mathematical algorithms have been proposed to improve the performance of clustering (Everitt et al 2011). Due to its simplicity and wide applicability, cluster analysis has been commonly used for data analysis in fields ranging from astronomy (Rosenburg, 1910; Babu and Feigelson, 1996 for a review), biology (Kerr and Chirchill, 2001;Witten and Tibshirani, 2010), psychology (Johnson, 1967; Farmer et al, 1983; Borgen and Barnett, 1987; Hay et al,1996) and anthropology (Clarke, 1968; Sutton and Reinhard, 1995), marketing (see Punj and Stewart, 1983 for a review), to increasingly in economics (Fisher, 1969; Hirschberg et al, 1991; El-Gamal and Grether, 1995; Slater and Zwirlein, 1996; Houser, et al, 2004; Yamamori et al, 2008; Adomavicius et al, 2012).

Walter Fisher was the first economist to systematically study the problem of classification. In his 1969 book *Clustering and Aggregation in Economics*, he foretold the increasing complexity of quantification in social variables and stressed “the need for systematic and scientific simplification” of social science data through clustering[1]. The discussion regarding the methods of clustering disappeared in economics for a long time after Fisher’s book was published. In 1960s and 1970s, the fields that saw new developments and applications using clustering methods were largely psychology and anthropology.

El-Gamal and Grether (1995) revived economists’ interest in uncovering behavioral strategies from complex data. They developed a pseudo-Bayesian approach to classify behavioral strategies used by individuals in games. The method is loosely related to finite mixture density clustering. Houser et al (2004) developed a related method in which the nature and the number of decision rules are determined simultaneously.

Substantial time elapsed from Fisher’s original work to the time empirical economists began to apply cluster analysis to real-world datasets. Among the few studies that implement cluster analysis, a variety of topics are included. Hieschberg et al (1991) identify clusters for welfare measures across countries using multiple hierarchical agglomerative clustering methods. Slater and Zwirlein (1996) adopt a slightly different hierarchical method using Ward’s minimum variance as clustering criteria[2]. They allocated 303 S&P 400 companies into 8 distinct groups in which some were classified as “stable maintainers” and others “leveraged strategists”.

Recently, a few experimental economists started to use cluster analysis to identify behavioral patterns among subjects. DeRubeis et al (2007) investigates the difference on the transmission pattern of sexually transmitted disease. The authors clustered individuals based on their demographic and clinical characteristics and separated the social network analysis for each cluster. Yamamori et al (2008) found three types of dictators in a modified dictator game with communication using Ward’s minimum variance hierarchical clustering. Adomavicius et al (2012) found that bidders in their auction experiment could be categorized into three behavioral groups using k-means clustering.

[1] The methods reviewed in Fisher (1969) is somewhat different from the cluster analysis defined by its current literature. The author did relate these clustering and aggregation methods to the general literature of cluster analysis.

[2] The difference and relations between cluster method and cluster criteria will be detailed in Section 2.

Given the level of complexity of innovative behavior in networks and the absence of pre-specified hypothesis on players' characteristics in these networks, it is natural to extend the use of cluster analysis to this context. The goal of this paper is to (1) review cluster analysis methods that are straightforward and easily implementable and (2) provide a concrete example of implementing this technique in a network dataset where we identify the "innovators" without pre-specifying their characteristics. Two key questions must be answered before implementing any clustering procedure[3]: which method should be used for the clustering analysis; and which method should be used to discover the number of clusters in the data. As these two decisions are made separately, we review them in separate section of the paper.

In Section II, we begin with a discussion of various clustering criteria and how they are used to find clusters in one's data. Since finding exact solutions in cluster analysis can be extremely computationally burdensome, semi-optimal clustering algorithms, such as k-means and k-median algorithms, are discussed. Section III reviews procedures for cluster analysis and discusses different methods used for each procedure. In addition to the choice of clustering methods, one also needs to choose how to determine the "correct" number of clusters. Section IV reviews two common approaches for this, the Silhouette width and the Calinski-Harabatz index. Section V introduces an example relevant to the study of innovation in networks and provides a sample analysis using data from a laboratory experiment related to innovation. The final section concludes.

II. Measures Used in Clustering

With optimization cluster analysis one develops indices and criteria to know in a mathematically precise way how "close" or far apart objects are to each other. There are many schools of thought regarding clustering.

One method adopts a bottom-up approach where the closest two objects are grouped first and then a third objects that are closest to the two[4] are added, so on and so forth. This method gradually forms a tree-like cluster result which gives its name "Hierarchical clustering". The

[3] An exception arises when one uses finite mixed density approaches for cluster analysis. In this case both questions are answered at the same time.

[4] Depending on the sub-school of thought, the similarity of an object to a group of objects could be evaluated by the distance of the object from the mean, the centroid, or the farthest or the closest object of the group.

hierarchical cluster analysis has a natural implication in taxonomy where objects bear similarity at different levels and join groups that are not necessarily horizontally comparable. An example is the classification of plants where genus, family and variety are groups formed at different levels of similarity. However, when studying clusters in social science data, researchers are often interested in parallel group structures that contain the entire dataset. This specific goal is achieved with another clustering method, optimization clustering.

The goal of optimization clustering is to allocate optimally all objects into a few groups[5] so that the aggregate distance within a group is small and the distance between groups is large. As this method provides a way to place individuals into flexible decision rule categories, and is straightforward and easily applicable to almost all behavioral datasets, we believe that the method bears relevance to the current discussion.

We introduce optimization clustering by describing each step of the clustering procedure. It starts with distance measures which calculate how close and far apart an object (or a group) is from another object (or another group). Built on the distance measures, we then discuss a variety of (dis-)similarity indices developed to aggregate these distance measures for any particular group. Different similarity indices are then combined to become the goal of the maximization (or minimization) problem. We introduce these goals (also known as optimization criteria) one by one. Finally, we demonstrate how clustering algorithms, like k-means and k-median, provide quasi-optimal solutions for the computationally impossible clustering problems.

II.1 Distance Measures

The starting point of many clustering investigations is an $n \times p$ multivariate matrix X with n observations each of which are described with p distinct characteristics. For behavioral datasets, this can be interpreted as a matrix of n individuals with each individual having p descriptive variables, such as gender, age, choices, etc.

A variety of distance measures have been proposed to measure quantitatively the distance between objects from a set of categorical or continuous observations (see, e.g., Jajuga et al,

[5] The number of groups is a choice variable for the researchers. Methods to choose the number of groups are discussed in Section 3.

2003). Categorical data are usually measured in terms of similarity, while continuous data are commonly measured in dissimilarity (or distance). These two types of measures are mostly interchangeable as they carry the same amount of information regarding distance.

When individual measures are binary, one may use the Matching Coefficient or Jaccard Coefficient as a distance measure. For each pair of individuals, the following table counts the matches and mismatches in the p variables.

Table 1. Counts of matches and mismatches for two individual *i* and *j*

		Individual *j*		
		1	0	Total
Individual *i*	1	a	b	$a+b$
	0	c	d	$c+d$
	Total	$a+c$	$b+d$	$p=a+b+c+d$

The Matching Coefficient approach simply calculates the ratio of one-one and zero-zero matches over the total number of characteristics p.

$$s_{ij}=(a+d)/(a+b+c+d) \qquad (1)$$

Alternatively, the Jaccard Coefficient ignores the zero-zero matches when calculating the similarity. Therefore, the Jaccard Coefficient is:

$$s_{ij}=a/(a+b+c) \qquad (2)$$

This is particularly useful when the absence of a large number of attributes may not necessarily lead to a high degree of similarity. For example, in biology, lacking similar attributes when comparing certain plants with certain insects does not lead to a high degree of similarity between them. Therefore, the principle to choose between the above two coefficients depends on the characteristics of the variables. When co-absence is considered informative, one may use the Matching Coefficient, otherwise the Jaccard Coefficient should be used[6].

[6] Similar coefficients have been proposed by Rogers and Tanimoto (1960), Sneath and Sokal (1973) and Gower and Legendre (1986). Their proposed coefficients vary the weight on the mismatches.

When each variable has more than two categories, the similarity measure s_{ijk} is constructed for each variable: when two individual i and j are the same on the kth variable, s_{ijk} equals one, and is zero otherwise. The measure is then averaged over all p variables. The over-all similarity measure between individual i and j is calculated as:

$$\mathrm{s}_{\mathrm{ij}}=\frac{1}{\mathrm{p}}\sum_{\mathrm{k}=1}^{\mathrm{p}}\mathrm{s}_{\mathrm{ijk}} \qquad (3)$$

Alternatively, one can also divide multiple categories into two subsets, then convert the original data into binary datasets and finally apply the Matching Coefficient or Jaccard Coefficient approach as in equation 2 and 3. However, whether it is proper to divide categories into two subsets may depend on the specific dataset and the research question one wishes to address.

When each individual has their characteristics measured as a continuous variable, distance between two individuals i and j are typically quantified by a dissimilarity index $\mathrm{d}_{\mathrm{ij.}}$ A variety of dissimilarity measures are proposed, among which Euclidean distance is the most commonly used one:

$$d_{ij}=\left[\sum_{k=1}^{p}(x_{ik}-x_{jk})^2\right]^{1/2} \qquad (4)$$

where x_{ik} and x_{jk} are, respectively, the *k*th variable value of the *p*-dimensional observations for individual i and j. This distance measure has the appealing property that the d_{ij} can be interpreted as physical distances between two *p*-dimensional points $x_i=(x_{i1},x_{i2}...x_{ip})$ and $x_j=(x_{j1},x_{j2}...x_{jp})$ in Euclidean space. Alternatively, city block distance measures the dissimilarity of individuals on a a rectilinear configuration[7].

$$d_{ij}=\sum_{k=1}^{p}\left|x_{ik}-x_{jk}\right| \qquad (5)$$

Where x_{ik} and x_{jk} are defined in the same manner as it is in Euclidean distance. Both of the above two measures are special cases of the general Minkowski distance with r=2 and r=1 respectively:

[7] It is also known as the Manhattan distance or taxicab distance as it is measures the travelling distance between two points on the street when city blocks are organized chess-board style.

$$d_{ij} = \left(\sum_{k=1}^{p} \left| x_{ik} - x_{jk} \right|^{r} \right)^{1/r} \quad (r \geq 1) \qquad (6)$$

In some cases, the data may contain both categorical and continuous variables. It is possible to construct a single measure by combining distance measures either with or without certain weighting function.

Notice that even though the distance measures mentioned above for categorical data are measuring distance in similarity while those for continuous data is in dissimilarity, in most cases, these two measure are interchangeable using the following formula[8]:

$$d_{ij} = \sqrt{1 - s_{ij}} \qquad (7)$$

In the following discussion, we assume the distance is measured in, or has been converted to, dissimilarity.

II.2 Dissimilarity Index

Whichever distance measure one may choose, one can form the dissimilarity matrix D by stacking the distance between all pairs of objects. In behavioral datasets, therefore, each row or column of a dissimilarity matrix corresponds to an individual. Each entry reflects a quantitative measure of dissimilarity between a particular pair of objects.

An informative clustering should include groups such that the distance between objects in the same group is small, while the distance between groups is large. Based on this simple principle, a variety of so-called “dissimilarity indices” (formed by taking combinations of distance measures) have been suggested.

With d_{lv}^{qk} defined as the dissimilarity between the *l*th object in the *q*th group and the *v*th object in the *k*th group, the following equations gives a simple example of an index that measures heterogeneity within group m:

[8] Gower (1966) showed that if a similarity matrix S, with element s_{ij}, is nonnegative definite, then the matrix D, with elements d_{ij} defined by equation 5 is Euclidean.

$$h_1(m) = \sum_{l=1}^{n_m} \sum_{v=1, v \neq l}^{n_m} (d_{lv}^{mm})^2 \qquad (8)$$

Intuitively, this index is the sum of squared dissimilarities between two objects that belong to the same group m.

Another commonly used similar index measures the sum of squared dissimilarities between an object in a cluster group m and the mean of objects in group m. It is also known as the trace of within-group dispersion matrix[9]. This index comprises the foundation for the k-means clustering algorithm which we will discuss later.

$$h_2(m) = \frac{1}{2n_m} \sum_{l=1}^{n_m} \sum_{v=1}^{n_m} (d_{lv}^{mm})^2 \qquad (9)$$

The final index we note here uses the smallest sum of distances to quantify dissimilarity of a group:

$$h_3(m) = \min_{v=1,\dots n_m} \left[\sum_{l=1}^{n_m} d_{lv}^{\mathrm{mm}} \right] \qquad (10)$$

where a reference object *v* is connected with all other objects in the group *m* to form a star, which then determines the sum of distance of the group. Since the smallest sum of distance is achieved when the reference object *v* is at the center of the group, the index is often referred to as the “star index”. $h_3(m)$ index is used in the k-median algorithm.

All three indices mentioned above measure the dissimilarity within the group m and ignore the information about the distance between group m and other groups. Separation indices are designed to capture this information. One commonly used separation index takes form $h_1(m)$ but now instead of summing over within group distance, the distance $d_{ml,kv}$ captures the dissimilarity between the object l from group m and the object v from a different group k.

$$h_4(m) = \sum_{l=1}^{n_m} \sum_{k \neq m} \sum_{v=1}^{n_k} (d_{lv}^{mk})^2 \qquad (11)$$

[9] The dispersion matrix is derived from multivariate matrix X directly without constructing the dissimilarity matrix D. These two methods are mathematically equivalent, hence we omit the discussion of the other method.

As separation indices are mostly capturing the same information as in dissimilarity indices[10] and that the current computer algorithms tend to use the latter, we will refer readers who are interested in other separation indices to Everitt et al (2010).

II.3 Clustering Criteria

Having chosen an index to represent a group's dissimilarity, clustering criteria can be defined by aggregating these group measures over all groups. The aggregation can be defined as the sum of dissimilarity over all groups as in $c_1(n,g)$, or as the maximum or minimum dissimilarity among groups as in $c_2(n,g)$ or $c_3(n,g)$ below:

$$c_1(n,g) = \sum_{m=1}^{g} h(m) \qquad (12)$$

$$c_2(n,g) = \max_{m=1,\dots g} [\mathrm{h(m)}] \qquad (13)$$

$$c_3(n,g) = \min_{m=1,\dots g} [\mathrm{h(m)}] \qquad (14)$$

One of the most commonly used clustering criteria combines $c_1(n,g)$ with dissimilarity index $h_2(m)$ to represent the total sum of within group dissimilarity. The criterion can also be shown equivalent to the within-group sum-of-squares criteria derived directly from the $n \times p$ multivariate matrix X.

$$c_1^*(n,g) = \sum_{m=1}^{g} h_2(m) = \sum_{m=1}^{g}\sum_{l=1}^{n_m} (d_l^{m\bar{m}})^2 = \sum_{m=1}^{g}\sum_{l=1}^{n_m} (x_l^m - \bar{x}^m)'(x_l^m - \bar{x}^m) \qquad (15)$$

Intuitively, when the above $c_1^*(n,g)$ clustering criterion is minimized, agents put into the same cluster share descriptive variables most similar to each other as compared to when they are allocated based on any other alternative clustering outcome.

[10] Roughly speaking, the sum of squared distance of the sample comprises two parts: the within group sum of squares and the between group sum of squares. Since the total sum of squared distance is constant, minimizing within group sum of squares, the dissimilarity index mentioned earlier, is equivalent to maximizing the between group sum of squares, the separation index.

There are a few features of the above clustering criterion of which any user should be aware. First, the method is scale dependent. For data that contains variables measured on different scales, one may reach different solutions from the same raw data standardized in different manners. Second, this clustering criterion imposes a “spherical” structure on the clusters and is unlikely to find clusters of other shapes, for example, agents that are separated into a few layers. Other clustering criteria exist to circumvent these two features[11]. However, any clustering approach has its advantages and disadvantages, and one must evaluate approaches within the context of particular applications.

III. Clustering Procedure—K-means and K-median Clustering

Ideally, one would consider all combinations of objects and choose the one that yields the lowest dissimilarity index within each group[12]. However, when the number of objects is large, it becomes infeasible to do this. Indeed, Liu (1968) provides the exact number of possible partitions one must consider in order to cluster n objects into g groups:

$$N(n,g)=\frac{1}{g!}\sum_{m=1}^{g}(-1)^{g-m}\binom{g}{m}m^{n} \qquad (16)$$

That is, in order to partition 100 network agents into 5 groups, the number of possible combinations to examine is about 6.6×10^{67}. The task becomes impossible even with modern computational power when the population under analysis comprises hundreds, if not thousands, of agents. This excessive computational burden has led scholars to develop numerical search algorithms to approximate clustering solutions. Here we review the two most commonly used numerical algorithms, k-means and k-median, both of which involve iterative updating processes for partitions and group centroids.

[11] Attempts to create clustering criteria less restrictive regarding the cluster’s shape include Scott and Symons(1971), Symons(1981), Murtagh and Raftery(1984), Banfield and Raftery(1993) and Celeux and Govaert(1995)

[12] Indices that measure the separation between groups are also used in many other methods. We refer interested readers to Everrit et al (2011)

- **K-means Algorithm:**

As stated in its name, the k-means algorithms emphasize the mean of the clusters. Generally speaking, all k-means algorithms involve iterative updates of clusters by simultaneously relocating objects into the cluster whose ***mean*** is closest and then recalculating cluster means. Particularly, all k-means algorithms contain the following four steps:

(1) g initial seeds are defined for each cluster by a p-dimensional vector, $\tilde{x}^m = (\tilde{x}_1^m, \tilde{x}_2^m, ..., \tilde{x}_p^m)$ where $\tilde{x}_k^m$ stands for the kth characteristic of the initial seed of cluster m. The squared Euclidean distance between the ith object and the initial seed of cluster m is simply calculated as:

$$d_{i\tilde{x}^m}^2 = \sum_{k=1}^{p} (x_{ik} - \tilde{x}_k^m)^2 \qquad (17)$$

By comparing the result of equation (X) for an object with each initial seed (there are g of them), we allocate the object to the cluster where the result is minimized.

(2) After all objects have been allocated to one cluster or another, the mean of the cluster is obtained by taking average over all objects that falls into each cluster. This is done for each dimension of the p characteristics:

$$\bar{x}^m = (\bar{x}_1^m, \bar{x}_2^m, ..., \bar{x}_p^m) \qquad (18)$$

The above mean of clusters $\bar{x}^m$ can then replace the initial seeds $\tilde{x}^m$ and be used to calculate the squared distance between each object and each cluster centroid as in equation (X). Objects are again moved to the cluster which yields the lowest squared distance measure.

(3) The step (2) is repeated. For each repetition, the old cluster mean is replaced by the one calculated from the latest membership. The process repeats until no objects change membership.

Although all k-means algorithms attempt to minimize within-group sum of squared deviations from (group) mean, they may differ from each other in details. Depending on the specific dataset used, these differences may have substantial impact on the clustering results[13]. Here we trace a few important differences of these most popular algorithms.

[13] We have found substantial differences in K-means clustering results produced by the standard packages in Stata, R and Matlab. We traced it to differences in the specific numerical algorithms used by each package.

First, the methods of initialization affect the final clustering results. The simplest suggestion, currently used in SPSS, chooses g random data points as initial cluster seeds (MacQueen, 1967). A slightly different method randomly partition all data points into g mutually exclusive groups and use the group mean as initial seeds (Steinley 2003). These two methods both rely on the random process, therefore may yield a different clustering result each time the algorithm is performed.

Various deterministic methods also exist. Astrahan (1970) suggest a two parameter method as follows: before initialization, two distance d1 and d2 are specified. Then for each data point, a density index is calculated as the number of objects that are less or equal to d1 distance away from the object. The object that yields the highest density is selected as the first seed. Objects that are within the distance of d2 to the first seed are removed from the consideration. A second seed is selected if it has the highest density among the remaining objects. The objects that are within distance d2 to the second seeds are removed. The process continues until all g seeds are determined. A similar process was suggested by Ball and Hall (1965) and implemented in the PROC FASTCLUS procedure in SAS. Although other types of random or deterministic processes exist (see Milligan, 1980 and Bradley and Fayyad 1998 for examples), Steinley (2003) suggest that the most robust method that outperform most of the arbitrary initialization rules is to use multiple random restarts (in order of thousands) and pick the one result that gives the smallest clustering criteria value. *Kmeans* package in R allow the user to specify the number of restart.

Second, to further minimize the squared distance as in equation (X), some algorithm suggests to introduce an additional stage of single-object reallocation process after the group reallocation has been settled (Spath, 1980; Hartigan and Wong, 1979). Specifically, after performing the standard iterative process (1)-(3) mentioned above, if there is an object in cluster m such that

$$\frac{n_m}{n_m - 1}(d_i^{m\bar{m}})^2 > \frac{n_{m'}}{n_{m'} - 1}(d_i^{m\bar{m}'})^2 \qquad (19)$$

The object *i* should be moved from cluster *m* to cluster *m'* and the squared distance (as in equation (X)) is reduced. The objects will be checked and moved if necessary one after another until no further improvement can be achieved by this process[14].

- **K-median Algorithm:**

In more recent years, the k-median algorithm has received increasing attention (Kaufman and Rousseeuw, 1990; spath, 1985; Hansen and Jaumard, 1997; Kohn et al, 2010). This algorithm relocates an object to a group whose ***median*** is the closest to it according to certain distance measure. Numerically, the specific clustering procedure proceeds like k-means except that the clustering criteria in equation (6) is replaced by

$$c_2^*(n,g) = \sum_{m=1}^{g} \sum_{l=1}^{n_m} \left| (x_l^m - \breve{x}^m) \right| \qquad (20)$$

Where $\breve{x}^m$ refers to the median vector of the *mth* cluster. The original idea of using median instead of mean is to reduce the influence of outliers. However, Garcia-Escudero and Gordaliza (1999) pointed out that k-median method can also be as affected by outliers as k-means since the "joint" selection of two medians are unlikely to be as robust in terms of centralization as when only one random variable is involved.

Variations of k-median algorithm also exist in terms of how initial seeds are selected and how objects are swapped between clusters. PAM (Partitioning Around Medoids), developed by Kaufman and Rousseeuw (1990) and implemented in the *pam* package of R language, is one of the most popular one. The algorithm sets the objective function as the sum of distance between each object and its nearest medoid. The initial seeds in PAM are chosen by a greedy built phase[15] where the seed is added one after another and only the one that brings the largest improvement on the objective function will be selected.

Once the built phase completes, a multi-iteration swapping stage begins. For each iteration, a medoid object i and a non-medoid object j will be selected that brings the largest improvement on the objective function if i and j are switched. The iterations continue until no improvement is possible. Since in both built phase and swapping phase, there are many pairs of objects to go

[14] The *kmeans* package in Matlab and R adopt this two-phase iterative algorithm.

[15] In programming, greedy algorithms refer to the ones that are based on heuristics who find locally optimal choice.

through to find the largest improvement, the original PAM algorithm is very time consuming with large dataset and increasing number of clusters[16].

IV. Methods for Choosing the Number of Clusters

Independent of the choice of clustering criteria and algorithms introduced above, one also needs to choose the method to determine the number of clusters. The past literature has recommended many methods that are algorithmic, graphical or formulaic. All of these methods are based on some logical heuristics. To judge which method is better at recovering the number of clusters, Milligan and Cooper (1985) conducted a Monte Carlo analysis to compare 30 of the most popular ones and concluded that the top performer is the one suggested by Calinski and Harabasz (1974) (which we denote by C-H)[17]. Another popular method readily available in many commercial packages is Silhouette Width. The output of this method includes a visualization giving direct clue on the performance of clustering under different numbers of clusters. We review Silhouette Width in this paper as well.

IV.1 C-H Index

C-H (1974) suggested that the optimal number of clusters, g*, should maximize the following value C(g):

$$C(g) = \frac{trace(B)}{g-1} \bigg/ \frac{trace(W)}{n-g} \qquad (21)$$

where

$$B = \sum_{m=1}^{g} n_m (\bar{x}^m - \bar{x})(\bar{x}^m - \bar{x})' \qquad (22)$$

representing the between-group dispersion matrix, and

[16] The same authors also developed a similar but less deterministic method CLARA (Clustering LARge Applications), implemented in R language. This method could reduce the computing time significantly when a dataset is large. Meanwhile, STATA implements its *cluster kmedians* command in a similar way as in the basic k median algorithm as described at the beginning of this subsection.

[17] Another successful technique developed by Duda and Hart (1973) works with hierarchical cluster methods. The network data do not fit these types of cluster analysis.

$$W = \sum_{m=1}^{g} \sum_{l=1}^{n_m} (x_l^m - \bar{x}^m)(x_l^m - \bar{x}^m)' \qquad (23)$$

representing the within-group dispersion matrix, both of which derive from the original multivariate matrix X.

IV.2 Silhouette Width

The Silhouette Width index is first mentioned in Rousseeuw(1987). His paper argues that due to the absence of visualization for the quality of cluster, it is hard to tell whether an object is well-classified or misclassified. He then proposed the index and the plot of Silhouette Width to visualize the quality of cluster. Interestingly, the Silhouette Width Index has become increasingly popular as a way to choose the number of clusters and has been adopted by most commercial packages along with the Calinki-Harabatz Index we introduced above.

For a given clustering result, the Silhouette width indices, denoted by s(i), are calculated for each object i=1,2,…,n, which are then combined into a Silhouette plot. Individual silhouette width s(i) is defined as:

$$s(i) = \frac{\min\limits_{C \neq M(i)} \frac{1}{n_C} \sum\limits_{\substack{k \notin M(i) \\ k \in C}} d(i,k) - \frac{1}{n_{M(i)}} \sum\limits_{\substack{j \in M(i) \\ j \neq i}} d(i,j)}{\max[\frac{1}{n_{M(i)}} \sum\limits_{\substack{j \in M(i) \\ j \neq i}} d(i,j), \min\limits_{C \neq M(i)} \frac{1}{n_C} \sum\limits_{\substack{k \notin M(i) \\ k \in C}} d(i,k)]} \qquad (24)$$

where M(i) refers to the cluster that contains object i, $n_{M(i)}$ refers to the number of objects in cluster M(i) and C refers to any cluster other than M(i).

The first term in the numerator refers to the minimum average distance of an object to all members of another cluster. It calculates the average distance from i to all members of an arbitrary cluster C. After the average distance is calculated for all arbitrary clusters, the closest cluster (in terms of distance to object i) is used.

The second term in the numerator refers to the within cluster average distance for object i. The term simply calculates the distance between object i and each other object in the same cluster and then takes an average. The denominator is the maximum of the two terms that appear in the numerator.

From the above formula, it is easy to see that s(i) would increase as object i is closer to other objects in the same group and farther away from objects in other groups. However, more characteristics of the index are revealed by evaluating s(i) under three different conditions.

First, note that if $\frac{1}{n_m}\sum_{\substack{j\in m(i)\\ j\neq i}} d(i,j) < \min_{c\neq m(i)} \frac{1}{n_c}\sum_{\substack{k\notin m(i)\\ k\in m(c)}} d(i,k)$, then s(i) can be simplified as

$1-\frac{\frac{1}{n_m}\sum_{\substack{j\in m(i)\\ j\neq i}} d(i,j)}{\min_{c\neq m(i)} \frac{1}{n_c}\sum_{\substack{k\notin m(i)\\ k\in m(c)}} d(i,k)}$ .That is s(i) is always positive and approaches 1 as the measure of within dissimilarity (the numerator) is much smaller than the measure of the smallest between dissimilarity (the denomenator).

Similarly, consider the opposite case where $\frac{1}{n_m}\sum_{\substack{j\in m(i)\\ j\neq i}} d(i,j) > \min_{c\neq m(i)} \frac{1}{n_c}\sum_{\substack{k\notin m(i)\\ k\in m(c)}} d(i,k)$. Under this condition, s(i) can be simplified as $\frac{\min_{c\neq m(i)} \frac{1}{n_c}\sum_{\substack{k\notin m(i)\\ k\in m(c)}} d(i,k)}{\frac{1}{n_m}\sum_{\substack{j\in m(i)\\ j\neq i}} d(i,j)}$-1, which is always a negative number and approaches -1 if within dissimilarity is large and the between dissimilarity if small. That is to say that the silhouette width index defined as in Rousseeuw(1987) is an index between -1 and 1 with a higher positive number indicating a better clustering quality.

In practice, one should choose the number of clusters that maximizes the average Silhouette Width across all objects.

V. Analyzing Network Data Using Cluster Analysis—An Example

V.1 The Dataset

To demonstrate the usefulness of cluster analysis in studying innovation in networks, we borrow the data from an experimental study that looked at individual behavior in a networked innovation game (Rong and Houser 2012). The study contains the repeated choice data from 160 subjects. Each subject is involved in a decision making game where they can earn money by either choosing to pay a high cost to provide a public good (representing costly but beneficial innovation) or choosing to pay a low cost to link to others who provide the public good (representing the follower or free-rider). Therefore, for each subject and each period, the dataset contains the contribution decision (1 if contributing to public goods, 0 if not contributing to public goods) and the linking decision (1 if linked to others, 0 if not linked to others) for each subject.

There are several treatments designed to mimic different market institutions which arguably could affect the level of innovation. The authors found significant difference between each institution. However, it is interesting to understand how each institution works to generate the difference in innovation. This is a task in which cluster analysis can play an important role. We use the dataset from that study to demonstrate how to use cluster analysis in this context and what level of new knowledge can be obtained from this exercise.

Our analysis proceeds in two steps. First, we estimate for each individual the parameters that characterize the way they make decisions given the information they have during their decision time. Then, we use cluster analysis to group similar individuals according to how their decision depends on the information they have. We call this dependence “decision rules”. In particular, we run a linear regression for each individual with the repeated decisions on contributing to public goods (or not) as a binary dependent variable. We regress this contribution decision on a constant, a dummy for whether investing is rational and a “history index” characterizing the subjects contribution behavior in the previous two rounds (see also Kurzban and Houser, 2005). After this analysis, individuals are characterized by the three estimated coefficients from their regression results. We have 142 subjects in our sample[18].

[18] We drop 18 subjects in this process, as there is zero variation in their dependent variables therefore we cannot estimate the coefficient for those subject using regression analysis.

In the second step, we implement the k-means algorithm to cluster these estimates into groups of behavioral rules.

The purpose of this analysis is to draw inferences about the behavioral rules of individuals in various treatments. We found that the difference in treatment design leads to different behavior rule usage. Note that our maintained assumption is that behavioral rules in all treatments are formed using elements from a menu of information that are finite and identical (in this case, decisions could be either "rationality dependent", "history dependent" or "constant level determined"), but that different treatments lead to rules that differ at the level of usage on each of this information. Without ex ante knowledge of what kind of weights people may put on each piece of information, we use cluster analysis to detect them. Cluster analysis allows us to explore behaviors among individuals without the need to pre-define the nature or number of possible rules (see also Houser et al, 2004).

V.2. Behavioral Rule Parameters

The independent variables we include in our regressions are meant to capture a person's: (i) base rate willingness to contribute to public goods (captured by the regression's constant); (ii) consistency with individual rationality (captured by the a dummy variable that takes value one if it is optimal to contribute); and (iii) propensity to form a "habit" of choice in the sense that they do what they did before (captured by the variable indicating the lagged decisions for the past 2 rounds). Equations 25 specifies our regression equations for contribution decision:

$$contribution_{i,t} = \beta_1 * rational^{p}_{i,t} + \beta_2 * \sum_{s=1}^{2} contribution_{i,t-s} + \beta_3 + \varepsilon_{i,t} \qquad (25)$$

where

$$rational^{p}_{i,t} = \begin{cases} 1, \text{ if it is optimal from subject i to contribute to public goods at round t} \\ \quad \text{according to individual rationality criteria} \\ 0, \text{ otherwise} \end{cases}$$

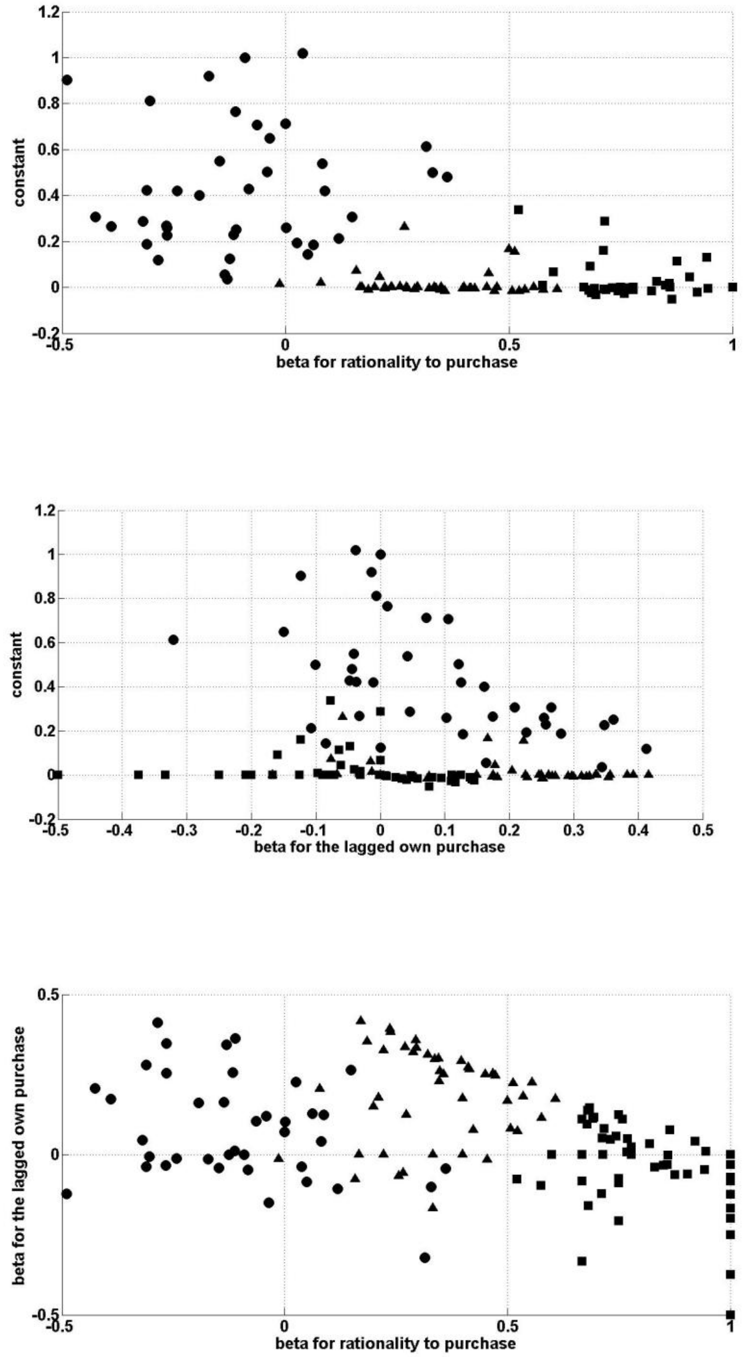
$$contribution_{i,t-s} = \begin{cases} 1, \text{ if subject i contributed to the public goods in round t-s} \\ 0, \text{ otherwise} \end{cases}$$

The above regressions are repeated for each individual. Each individual's estimates can be represented by a point in 3-space (See Appendix A).

VI.2. K-means Clustering

We implement our k-means cluster analysis, as well as cluster number selection, using R. Based on the C-H index, we find three clusters in contribution decisions (See Appendix B).

Figure 1. Projections of Estimates from Contribution Decision



(a)

(b)

(c)

The three panels of Figure 1 are the three 2-space projections of the estimates $\{\beta_1, \beta_2, \beta_3\}$ from regression on contribution decisions (Equation 25) into corresponding 2-space. Each point represents an individual's estimates from his/her contribution decisions regression. Points with the same marker belong to the same cluster.

It is clear from visual inspection that our clusters are well-separated. To provide statistical evidence on the strength of this separation, we analyze the separation along each independent variable's axis. Mann-Whitney tests find significant differences between all pairs of clusters in each axis ($p<0.001$), with the exception of the constants in the triangle and round clusters.

Not only are the clusters clearly separated, the location of the clusters also carries meaningful interpretation in our sample. Table 2 provides the mean estimate for each independent variable and for each cluster, and also reports whether that mean is significantly different from zero.

Table 2. The Mean of Estimates from Regression on Contribution Decision

	Square Cluster	Triangle Cluster	Round Cluster
Rational to contribute	0.8190 (0.0000)	0.3411 (0.0000)	-0.0978 (0.0054)
Lagged choice	-0.0408 (0.1480)	0.1745 (0.0000)	0.0782 (0.0120)
Base rate(constant)	0.0175 (0.2589)	0.0137 (0.7066)	0.4279 (0.0000)
Number of subjects	57	46	39

Note: p-value from Wilcoxon signed-rank test in parentheses

Based on the results from Table 2, we summarize the characteristics of the three clusters that define the three behavioral rules used by our subjects. Note that the decision rules below are not pre-specified. It is generated as a result of clustering.

(1) We define the cluster indicated with round markers as the "Rational" type. People that belong to this cluster are guided by the rationality of the current opportunity to contribute. They focus less on their past choices, and their base rate of investing is near zero.

(2) We define the cluster indicated by triangle markers as the “Habit” type. Subjects in this cluster are guided by rationality, but relatively less than the Rational type. Instead, their current decisions follow closely their past decisions.

(3) We define the cluster indicated by square markers as the “Dogmatic” type. We find that these subjects have the highest base rate of investing among all three types.

The clear separation of the three types of individuals in this experiment shows that innovation is not generated for the same reason for all people. Some people develop new ideas because it is optimal for them to do so. Some people innovate for the reason that they have done that before. The rest of the innovators choose to do so without concern for individual payoffs or their personal history. They are the dogmatic innovator.

Which types of innovators drive innovation in society and how can we promote their existence? These questions can be addressed by investigating how institutional characteristics in our various treatments affect the types of behavioral rules subjects use.

The level of innovation is lowest at the two treatments where subjects can make unconstrained choice[19]. This low level of innovation coincides with a concentration of Dogmatic type subjects (41.38% and 90% respectively) in both treatments. That is to say, having a concentration of players using the Dogmatic rule is not conducive to innovation. The unconstrained choice treatments may be unhelpful in generating innovation.

On the contrary, for the other two treatments that feature constrained choice sets, the data include relatively high levels of innovation. In those two “successful” treatments, the large majority of subjects (92%) choose to behave rationally or follow a habit (88.89%). We found zero dogmatic innovators in these two treatments.

In the last treatment where a medium level of innovation is observed, it is also the case that no subject belongs to the Dogmatic type.

The knowledge gained from cluster analysis provides a clear picture on which treatment design generates the most innovation and the reasons why that has happened: the behavioral rules shift away from the dogmatic innovator. This finding is not available in the absence of clustering

[19] The detail of the treatment design is of less importance to this study. We suggest the interested readers to find detailed description of the experiment in Rong and Houser (2012).

results and it would seem very difficult to come up with it as an ex-ante hypothesis. For these reasons, this example well-demonstrates the value of cluster analysis in the study of large and complex datasets.

VI. Summary

Cluster analysis is an intuitive method to analyze complicated data sets. Without making strong assumptions regarding the data generating process, the method divides observations into discrete groups based on patterns of similarity. We reviewed key features of cluster analysis in this paper. First, we reviewed several distance measures appropriate for different types of measures (binary, categorical or continuous). We then illustrated how distance measures can be combined into dis-)similarity matrices and how these matrices are further used in forming clustering criteria. We also discussed the detail of two popular algorithms: k-means and k-median. Finally, we reviewed two indices, Calinski-Harabatz Index and Average Silhouette Width, used to discover the number of clusters in the data. We offered an example of this approach using experimental network data, and argued that individual decisions made in a network environment are often generated by complex behavioral rules that can be difficult to specify a priori. Such environments may particularly benefit from clustering methods.

Reference

Adomavicius, G., S. P. Curley, A. Gupta and P. Sanyal (2012): Effect of Information Feedback on Bidder Behavior in Continuous Combinatorial Auctions, *Management Science*,58:811-830

Anne E. Farmer, a, Peter McGuffinb, Edward L. Spitznagelc (1983): Heterogeneity in Schizophrenia: A Cluster-analytic Approach, *Psychiatry Research*, 8(1): 1–12

Astrahan, M. M. (1970): Speech Analysis by Clustering, or the Hyperphome Method, *Stanford Artificial Intelligence Project Memorandum AIM-124*. Stanford, CA: Stanford University.

Babu, G.J., E. D. Feigelson(1997):*Statistical Challenges in Modern Astronomy II*, Springer

Ball, G. H. and D. Hall, D. J. (1965): ISODATA: a Novel Method For Data Analysis and Pattern Classification Menlo Park, CA: Stanford Research Institute.

Banfield, J. D. and A. E. Raftery (1993): Model-based Gaussian and Non-Gaussian Clustering, *Biometrics,* 49: 803–821.

Borgen, F. H. and D.C. Barnett (1987): Applying Cluster Analysis in Counseling Research, *Journal of Counseling Psychology*, 34(4):456-468

Bradley, P. S. and U. M. Fayyad (1998): Refining Initial Points for k-means Clustering, *Machine Learning: Proceedings of the fifteenth International Conference* edited by J.Shavlik (pp. 91–99). San Francisco: Morgan Kaufmann.

Bushel, P. R., R.D. Wolfinger and G. Gibson (2007): Simultaneous Clustering of Gene Expression Data with Clinical Chemistry and Pathological Evaluations Reveals Phenotypic Prototypes, *BMC Systems Biology*, 23: 1–15.

Calinski, R. B. and J. Harabasz (1974): A Dendrite Method for Cluster Analysis, *Communications in Statistics*, 3:1–27.

Celeux, G. and G. Govaert (1995): Gaussian Parsimonious Clustering Models, *Pattern Recognition*, 28(5):781–793.

Clarke, D. L.(1968): *Analytical Archaeology*. Methuen

DeRubeis E, J.L. Wylie, D.W. Cameron, R.C. Nair and A.M. Jolly (2007): Combining Social Network Analysis and Cluster Analysis to Identify Sexual Network Types, *International Journal of STD & AIDS*, 18(11):754-9.

Duda, R. O. and P. E. Hart (1973)*: Pattern Classification and Scene Analysis*, John Wiley & Sons, Inc., New York.

El-Gamal, M.A. and D. M. Grether (1995): Are People Bayesian? Uncovering BehavioralStrategies*, Journal of the American Statistical Association***,** 90: 1137–1145.

Everitt, B.S., S. Landau, M. Leese, D. Stahl (2011): *Cluster Analysis*, John Wiley & Sons

Fisher, W.D. (1969): *Clustering and Aggregation in Economics*, The Johns Hopkins University Press

Garcia-Escudero, L. A. and A. Gordaliza (1999): Robustness of Properties of K-means and Trimmed K-means*, Journal of the American Statistical Association*, 94: 956–969.

Gower, J. C. (1966): Some Distance Properties of Latent Root and Vector Methods Used in Multivariate Analysis, *Biometrika*, 53: 325–338.

Gower, J. C. (1971): A General Coefficient of Similarity and Some of its Properties, *Biometrics*, 27: 857–872.

Gower, J. C. and P. Legendre (1986): Metric and Euclidean Properties of Dissimilarity Coefficients, *Journal of Classification*, 5: 5–48

Hansen, P. and B. Jaumard (1997): Cluster Analysis and Mathematical Programming, *Mathematical Programming*, 79: 191–215.

Hartigan, J. A. and M. A.Wong (1979): Algorithm AS 136: A k-means Clustering Algorithm, *Applied Statistics* 28: 100–108.

Hay, P. J., C.G. Fairburn and H.A. Doll (1996): The Classification of Bulimic Eating Disorders: A Community Based Study, *Psychological Medicine*, 26(4):801-812.

Hirschberg, J. G., E. Maasoumi and D. J. Slottje (1991): Cluster Analysis for Measuring Welfare and Quality of Life Across Countries", *Journal of Econometrics*, 50: 131-150.

Houser, D., M. Keane and K. McCabe (2004): Behavior in a Dynamic Decision Problem: an Analysis of Experimental Evidence using a Bayesian Type Classification Algorithm, *Econometrica*, 72(3): 781-822

Ichino, M. and H. Yaguchi(1994): Generalized Minkowski Metrics for Mixed Feature Type Data Analysis, *IEEE Transactions on Systems, Man and Cybernetics*, 24: 698–708.

Jajuga, K.,M. Walesiak and A. Bak(2003), On the General Distance Measure, *Exploratory Data Analysis in Empirical Research* (M. Schwaiger and O. Opitz, eds.) Springer-Verlag, Heidelberg.

Johnson, S. (1967): Hierarchical Clustering Schemes, *Psychometrika,* 32(3): 241-254

Kaufman, L. and P. Rousseeuw (1990): *Finding Groups in Data: An Introduction to Cluster Analysis*, New York: Wiley.

Kerr, M.K. and G.A. Churchill (2011): Bootstrapping Cluster Analysis: Assessing the Reliability of Conclusions from Microarray Experiments*,* 98(16):8961-5

Kohn, H. F., D. Steinley and M. J. Brusco (2010): The p-median Model as a Tool for Clustering Psychological Data, *Psychological Methods*, 15: 87–95.

Legendre, P. and A. Chodorowski (1977): A Generalisation of Jaccard's Association Coefficient for Q-analysis of Multi-state Ecological Data Matrices, *Ekologia Polska*, 25: 297–308.

Liu, G. L. (1968): *Introduction to Combinatorial Mathematics*, McGraw Hill, New York.

MacQueen, J. (1967): Some Methods of Classification and Analysis of Multivariate Observations *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability* (edited by Cam, Le, L. M. NeymanJ. (1) 281–297 Berkeley, CA: University of California Press.

Milligan, G. W. (1980): An Examination of the Effect of Six Types of Error Perturbation on Fifteen Clustering Algorithms, *Psychometrika,* 45:325–342.

Milligan, G. W. and M. C. Cooper (1985): An Examination of Procedures for Determining the Number of Clusters in a Data set*, Psychometrika,* 50: 159–179.

Murtagh, F. and A. E. Raftery (1984): Fitting Straight Lines to Point Patterns, *Pattern Recognition*, 17:479–483.

Punj, G. and D. W. Stewart (1983): Cluster Analysis in Marketing Research: Review and Suggestions for Application, *Journal of Marketing Research* , 20(2):134-148

Rong, R. and D. Houser (2012): Growing Stars: A Laboratory Analysis of Network Formation, Working paper

Rosenburg, H. (1910): On the Relation Between Brightness and Spectral Type in the Pleiades [title translated in English], *Astronomische Nachrichten*, 186:71

Rousseeuw, P. J. (1987): Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis*, Journal of Computational and Applied Mathematics*, 20: 53–65.

Scott, A. J. and M. J.Symons (1971): Clustering Methods Based on Likelihood Ratio Criteria, *Biometrics*, 27: 387–398.

Slater, S. F. and T.J. Zwirlein(1996): The Structure of Financial Strategy: Patterns in Financial Decision Making, *Managerial and Decision Economics*, 17(3):253-266

Sneath, P. H. A. and R. R.Sokal(1973): Numerical Taxonomy, W. H. Freeman

Späth, H. (1980): *Cluster Analysis Algorithms for Data Reduction and Classification of Objects*, New York: Wiley

Späth, H. (1985): *Cluster Dissection and Analysis: Theory, FORTRAN Programs, Examples*, New York: Wiley.

Spearman, C. (1904): General intelligence, objectively determined and measured, *American Journal of Psychology*, 15: 201–292.

Steinley, D. (2003): K-means Clustering: What You Don't Know May Hurt You, *Psychological Methods*, 8: 294–304.

Sutton, M.Q. and K. J. Reinhard (1995): Cluster Analysis of the Coprolites from Antelope House: Implications for Anasazi Diet and Cuisine, *Journal of Archaeological Science*,22(6):741–750

Symons, M. J.(1981): Clustering Criteria and Multivariate Normal Mixtures, *Biometrics*, 37: 35–43

Tryon, R.C. (1932): Multiple Factors Vs Two Factors as Determiners of Ability, *Psychological Review*, 39: 324-51

Tryon, R. C. (1935): A Theory of Psychological Components—An Alternative to "Mathematical Factors," *Psychological Review*, 42: 425–454.

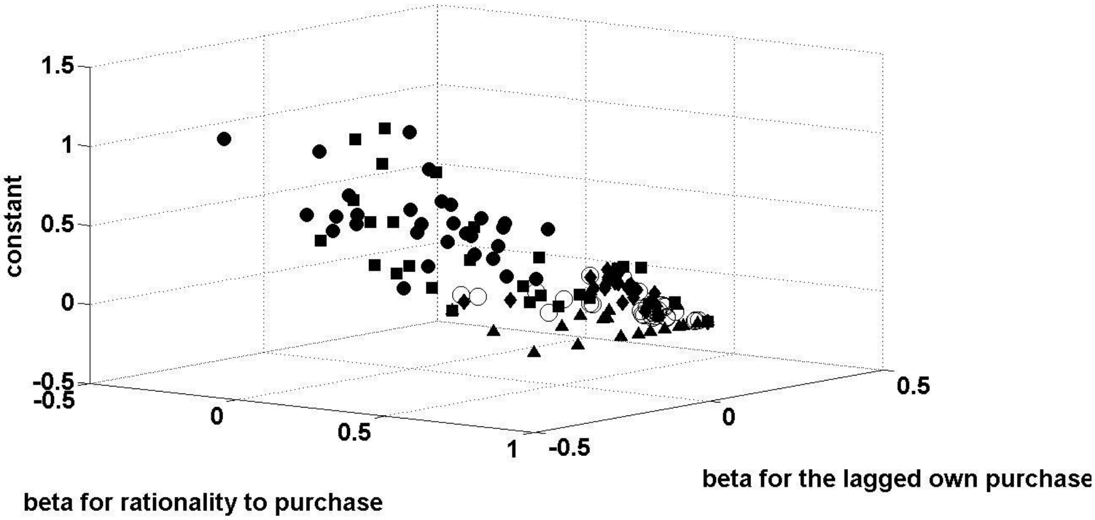
Tryon, R. C., and D.E. Bailey (1966): The BCTRY Computer System of Cluster and Factor Analysis, *Multivariate Behavioral Research*, 1:95-111

Witten, D.M. and R. Tibshirani(2010): Supervised Multidimensional Scaling for Visualization, Classification, and Bipartite Ranking, *Journal Computational Statistics & Data Analysis archive*, 55(1):789-801

Wright, C., T. Burns, P. James, et al. (2003): Assertive Outreach Teams in London: Models of Operation, *British Journal of Psychiatry*, 183: 132–138.

Yamamori, T, K. Kato, T. Kawagoe and A. Matsui (2008): Voice Matters in a Dictator Game, *Experimental Economics*, 11: 336–343

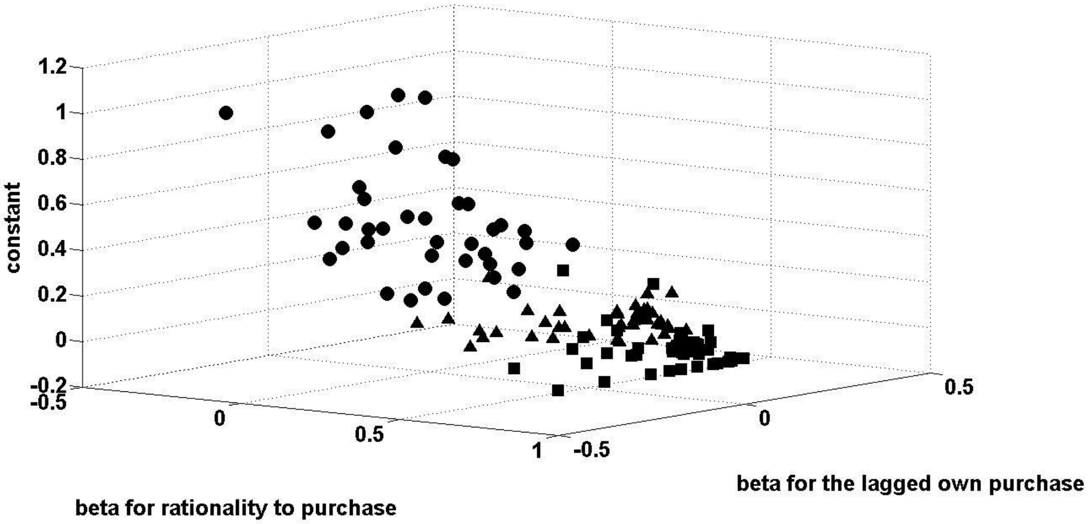
Appendix A: **3-space Plot for Individuals' Estimates by Treatment**



Note: different markers represent different treatments

■ -- Seq_B ; ▲ -- Seq_L; ● -- Sim_B; ♦ -- Sim_L; ○ -- Sim_L_NoRFR

Appendix B. 3-space Plot for Individuals' Estimates by Cluster


constant
beta for rationality to purchase
beta for the lagged own purchase