Archetypal Analysis: A New Way To Segment Markets Based On Extreme Individuals

Gui Li
Commonwealth Bank of Australia

Paul Wang
University of Technology, Sydney

Jordan Louviere
University of Technology, Sydney

Richard Carson
University of California, San Diego

The data used in this paper were contributed by: Timothy Devinney (AGSM) and Patrice Auger (Melbourne Business School), Co-Principal Investigators (with J. Louviere) on grants that funded the data collection (ARC-LRG A00102981, "Consumer Assessment of Social Product Features"; and City University of Hong Kong Exploratory, Small-Scale, or Developmental Studies Research Grant 9030655, "Measuring the Utility Value of Ethical Consumerism")

Archetypal Analysis: A New Way To Segment Markets Based On Extreme Individuals

Introduction

Segmenting consumers into groups has long been a useful way of gaining insight into marketing problems (eg, Frank, Massey, and Wind, 1972). A number of different statistical approaches to accomplish such partitionings have been proposed over the years (eg, Kaufman and Rousseeuw, 1990). A common feature of these approaches has been to chose representative "objects" which are locally "centered" according to some criteria. In this paper, a new technique, archetypal analysis, which is based upon distance from "important extreme objects" is presented. The approach appears to show considerable promise for marketing applications and we present examples of its use in several marketing contexts using both products and consumers as the objects.

The importance of looking at extreme consumers in segmenting markets recently was recognized by Allenby and Glinter (1995), who argued that extreme consumers may be the most important from the perspective of the new product introduction and switching behavior. Their methodological approach involved a hierarchical Bayesian random-effects model applied to conjoint data. The approach we propose has some conceptual similarities but a very different statistical foundation.

Archetypal Analysis

Archetypal analysis is a new statistical data analysis technique proposed by Cutler and Breiman (1994) that has found considerable application in the hard sciences. While briefly noting the similarity and differences of the approach to cluster centers, and particularly, the concept of principal points (Flurry, 1990), Cutler and Breiman developed archetypal analysis (hereafter, "AA") primarily as an alternative to principal components analysis. AA has been used to model a number of different physical phenomena such as air pollution in Los Angeles, behavior of flame cells, head dimensions, and plasma in Tokama fusion reactors.

As an approach to segmenting products markets or consumers, AA has a number of interesting properties. For a pre-specified number of archetypes (p), the approach finds archetypes that define the smallest convex hull in the k-dimensional space defined by the k variables being used which can best encompass the data subject to a set of constraints designed to enhance the interpretability of the results. The major constraint imposed is that the archetypes must be actual objects or simple linear combinations of objects which appear in the data set being analysed. Objects contained inside the convex hull defined by the archetypes incur no loss while objects outside of the convex hull incur a loss according to some criteria such as the square of the projection distance to the convex hull. The archetypes are chosen to minimize this loss function.

The information returned from the AA procedure include the location of the p archetypes in the k -dimensional space and a vector of p coefficients for each object which describe the relationship of the object to each p archetypes. This vector of coefficients for each object must sum to one, and hence, can be considered to be similar to the coefficients of a proper mixture distribution. For objects inside the convex hull, the coefficients are simply the normalized distances to each of the p archetypes. For such an object, knowledge of the location of the archetypes and the vector of coefficients entails no loss of information relative to the object's original location in the k

-dimensional space. For objects outside the convex hull, the coefficients represent distance to the closest projection of object onto the convex hull.

In certain respects AA can be seen as a form of fuzzy or probabilistic clustering. Objects located at an archetype receive a coefficient of one for that archetype and coefficients of zero for all other archetypes. Objects not located at an archetype can be seen as mixtures of these pure types. That is, their respective coefficients on the p archetypes will be less than one for all p archetypes, and the magnitude of each individual coefficient reflects the relative proximity to a specific archetype. The distribution of the coefficients of each archetype across a data set provides useful summary information, and it is possible to perform additional analyses using these mixture coefficients (not illustrated in this paper).

An important advantage of archetypal segmentation in contexts where the variables are product attributes or consumer characteristics is that archetypes are almost always easy to interpret because they represent extreme combinations of attributes or extreme combinations of consumer characteristics. In contrast, in more traditional clustering procedures, the centers of clusters all tend to be somewhat in the middle of the k-dimensional space by construction. Indeed, it is this property of standard cluster analysis techniques that often makes it difficult to understand exactly how clusters differ from one another if k is even of moderate size. Having said that, typically can be a large number of possible observations in any data set that might qualify as being extreme and there is nothing particularly interesting about simply choosing *any* extreme observation. Instead, AA defines the p most important objects that are supported by the data in so far as they encompass most of the other objects inside the convex hull defined by the objects, with the remaining objects lying not too far outside the convex hull. By dealing with an encompassing convex hull, one also avoids the imposition of artificial orthogonallity constraints which underlie many standard cluster approaches.

Interesting archetypal solutions always will involve p 3, and there is always a finite number of possible archetypes which is less than or equal to the number of observations. In practice, a fairly small number of archetypes is usually sufficient to incorporate most of the information in a data set. Some insight into the nature of archetypes can be gained by noting that for p=1, a single archetype implies a vector of the means of the k variables used to define the space. For p=2, the two archetypes define a separating hyperplane which for k=2 is simply a linear regression line.

The procedure used to estimate archetypes with a squared error loss function is based upon iterative applications of a convex least squares algorithm. The procedure is quick and easy to estimate but can converge to a local minimum. Starting the procedure from a number of randomly chosen starting point usually quickly identifies archetype locations which minimize the loss function. We consider various improvements to Cutler and Brieman's proposed procedure which should make it easier to use by marketing practitioners. We also consider the implications of standardizing the data before performing the analysis, the properties of statistical tests for determining the optimal number of archetypes and the possibility of using alternative loss functions.

Empirical Illustrations Of AA

We illustrate use of archetypal segmentation in two common marketing research applications: 1) identifying segments from responses to attitudinal statements, and 2) identifying segments from responses to discrete choice experiments.

The first application involves a sample of 600 consumers intercepted in malls in six countries. Respondents completed a battery of 20 sets of questions that involved different combinations of 16 issues listed in Table 1. A balanced incomplete block design was used to make the 20 sets to insure that each set contained four issues. Respondents identified, respectively, the most and least important issues in each set. Finn and Louviere (1992) showed how to use "Best-Worst Scaling" to measure each issue for each respondent on a difference scale; we used their approach to measure respondents' stated importance for each of the 16 issues. The sample means and associated standard errors are in Table 1, together with the results for each archetype. We developed and tested a procedure for conducting archetypal analysis in MATLAB, and extracted 2 to 12 archetypes. The residual sums of squares for these solutions indicated little further improvement after five archetypes. As previously discussed, archetypes are unique individuals who represent extremes in the dataset. These extreme five individuals are shown in Table 1. Their interpretation is as follows:

- A1 = considers animals and genetically modified materials unimportant; considers safety most important.
- A2 = considers sexual rights and recyclable packaging unimportant; considers safe working conditions, genetically modified materials, unions and human rights relatively important.
- A3 = considers use of animal byproducts unimportant; considers gender, racial and religious rights and use of genetically modified materials important.
- A4 = consider use of genetically modified materials unimportant; considers biodegradability and product disposability important.
- A5 = considers packaging and recycling to be unimportant; considers human, animal, gender, religious and racial rights important.

Minimum wages and good living conditions were consistently average in importance across all archetypes, and so do not distinguish the groups.

The second illustration involves a choice experiment administered to convenience samples of City U of Hong Kong undergraduate business students, AGSM MBA students and Amnesty International members (Australia). The choice experiment involved 14 attributes, 12 varied over two levels and two varied over four levels (price and brand); additional brand levels were created by using available two-level columns to systematically vary presence/absence of brands within level four. Experimental subjects indicated whether or not (yes, not) they would consider buying 32 athletic shoes described by combinations of the attribute levels. The 32 1, 0 responses were used as input to the AA procedure. Analysis of residual sums of squares indicated that there was little improvement after 4 archetypes. The mixture weights for the four archetypes were extracted and used to weight MNL models for each of the four archetypal segments. These results are shown in Table 2, which also shows that the use of the archetypal information significantly reduced the overall MNL model likelihood, indicating that the AA solution provides useful statistical information about preference heterogeneity (2 x the sum of the separate AA log-likelihoods = 503.4, and is distributed as χ^2 for 72 df). The results reveal that compared to the overall sample, the archetypes can be described as follows based on the 95% confidence interval for the overall sample estimates:

- A1 = 28.3% of sample - significantly less likely to consider any shoes, more interested in shock absorbency, ventilation fabric/material, comfortable fit, dangerous work practices, proper accommodation for workers and brand 1; more negative to brand 5. This segment cares about shoe performance, but has a social conscience.
- A2 = 19.8% of sample - more likely to consider any shoes, less interested in use of child labor in manufacturing and proper accommodation for workers but more price sensitive. This segment emphasizes price regardless of labor practices.
- A3 = 36.2% of sample - more interested in ankle support, less in comfortable fit but very price sensitive. This segment is concerned about support and price.
- A4 = 16.6% of sample - least likely to consider any shoes, least interested in shock absorbency, ankle support, brand 1 and brand 11; most interested in weight, fabric/material, comfortable fit, child labor, workers paid minimum wages and brand 11.

References

Allenby, G.M. and J.L. Ginter (1995), "Using Extremes to Design Products and Segment Markets," *Journal of Marketing Research*, 32, 392-403.

Cutler, A. and I. Breiman (1994), "Archetypal Analysis," *Technometrics,* 36, 338-347.

Finn, A. and J.J. Louviere (1992) "Determining the Appropriate Response to Evidence of Public Concern: The Case of Food Safety," Journal of Public Policy and Marketing, 11, 1, 12-25, 1992.

Frank, R.E. and W.F. Massey, and Y. Wind (1972), *Market Segmentation* (Englewood Cliffs, NJ: Prentice Hall).

Flurry, B. (1990), "Principal Points," *Biometrika*, 77, 33-41.

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

Table 1: Archetypal Analysis Of Ethical Issues

	Sample, N=603		Arch 1	Arch 2	Arch 3	Arch 4	Arch 5
Ethical Issues	Mean	StdError	Scores	Scores	Scores	Scores	Scores
animal rights	-0.226	0.104	-5	-1	-1	-1	3
animal byproducts	-1.270	0.083	-4	-2	-3	-1	2
biodegradability	-0.433	0.090	2	0	-1	4	-2
recyclable materials	-1.227	0.074	-2	0	-2	1	-3
safety information	-0.478	0.095	4	-2	-1	0	0
human rights	3.015	0.081	1	3	2	2	4
recyclable package	-1.698	0.078	-2	-4	-1	1	-4
product disposability	-0.416	0.084	2	-1	-2	4	-3
minimum wages	0.355	0.068	-1	2	1	-2	0
unions allowed	-0.896	0.093	2	3	1	-2	-2
good living conditions	1.020	0.077	0	1	-1	-2	1
sexual rights	-0.521	0.105	1	-5	2	-1	1
safe working conditions	1.509	0.067	3	3	0	1	1
no child labor	1.852	0.093	2	3	0	-2	1
gm used	-1.119	0.084	-3	0	3	-4	-3
gender, racial, religious rights	0.532	0.104	0	0	3	2	4

Table 2: MNL Model Estimations Using Archetype Mixtures Parameters As Weights

Attribute Effects	Overall MNL Model For Sample				Archetype 1		Archetype 2		Archetype 3		Archetype 4	
	Coeff	StdErr	T	P(T)	Coeff	P(T)	Coeff	P(T)	Coeff	P(T)	Coeff	P(T)
Intercept	-0.578	0.098	-5.92	0.000	-1.111	0.000	0.598	0.002	-0.702	0.000	-1.154	0.000
ShockAbsorb	0.214	0.034	6.30	0.000	0.350	0.000	0.186	0.006	0.255	0.000	0.112	0.243
Weight	-0.193	0.034	-5.72	0.000	-0.195	0.010	-0.206	0.002	-0.210	0.001	-0.262	0.004
Ankle Supprt	-0.155	0.033	-4.72	0.000	-0.112	0.126	-0.131	0.043	-0.292	0.000	-0.049	0.576
SoleDurability	0.146	0.034	4.34	0.000	0.113	0.134	0.174	0.009	0.203	0.001	0.135	0.150
Ventilation	0.156	0.034	4.55	0.000	0.233	0.003	0.138	0.043	0.131	0.031	0.171	0.082
FabricMater	-0.050	0.034	-1.50	0.140	0.086	0.267	-0.054	0.425	-0.107	0.079	-0.179	0.060
Reflection	0.006	0.034	0.18	0.860	-0.001	0.988	0.024	0.721	0.010	0.867	-0.029	0.762
ComfyFit	0.272	0.034	7.90	0.000	0.426	0.000	0.251	0.000	0.136	0.026	0.556	0.000
ChildLabor	-0.211	0.037	-5.78	0.000	-0.240	0.004	-0.129	0.080	-0.215	0.001	-0.402	0.000
MinWage	0.106	0.037	2.87	0.004	0.053	0.525	0.109	0.144	0.134	0.041	0.202	0.043
WorkDanger	-0.186	0.037	-5.07	0.000	-0.285	0.001	-0.118	0.114	-0.204	0.002	-0.237	0.017
WorkerAcc	0.110	0.037	3.02	0.003	0.246	0.003	0.013	0.865	0.090	0.167	0.171	0.088
Price	-0.010	0.001	-9.83	0.000	-0.012	0.000	-0.011	0.000	-0.013	0.000	-0.005	0.052
Brand1	0.189	0.072	2.63	0.009	0.552	0.002	0.133	0.309	0.117	0.407	-0.049	0.803
Brand2	0.153	0.072	2.13	0.034	0.021	0.908	0.116	0.374	0.249	0.070	0.281	0.137
Brand3	-0.075	0.193	-0.39	0.698	0.050	0.909	-0.162	0.662	0.053	0.878	-0.331	0.546
Brand4	-0.424	0.259	-1.64	0.102	-0.150	0.809	-0.420	0.353	-0.637	0.209	-0.092	0.893
Brand5	-0.067	0.265	-0.25	0.801	-1.103	0.241	0.221	0.603	-0.054	0.931	-0.213	0.760
Brand6	0.028	0.227	0.12	0.903	0.381	0.461	0.106	0.800	-0.082	0.848	0.057	0.927
Brand7	-0.201	0.215	-0.94	0.349	-0.316	0.557	-0.259	0.506	-0.146	0.709	-0.049	0.934
Brand8	-0.531	0.216	-2.46	0.014	-0.566	0.276	-0.601	0.125	-0.547	0.166	-0.520	0.385
Brand9	0.111	0.192	0.58	0.563	0.127	0.774	0.197	0.601	0.175	0.612	-0.038	0.943
Brand10	0.371	0.188	1.98	0.048	0.492	0.245	0.125	0.736	0.209	0.548	0.870	0.103
Brand11	0.446	---	---	---	0.512	---	0.544	---	0.663	---	0.084	---
-2LL	572.4				208.6		137.1		240.9.		148.8	