ARCHETYPAL ANALYSIS: A NEW WAY TO SEGMENT MARKETS BASED ON EXTREME INDIVIDUALS

Shan Li
Commonwealth Bank of Australia

Paul Wang and Jordan Louviere
University of Technology, Sydney

Richard Carson
University of California, San Diego

Track: Marketing Research and Research Methodologies

Keywords: Segmentation, Choice Modelling, Mixture Models

We gratefully acknowledge Timothy Devinney (AGSM) and Patrice Auger (Melbourne Business School), Co-Principal Investigators (with J. Louviere) on grants that funded the data used in this paper (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").

Abstract

Segmenting consumers into groups has long been used to gain marketing insights (e.g., Frank, Massey & Wind 1972). Many ways to identify segments have been proposed (e.g., Kaufman & Rousseeuw 1990), although most share some common features, such as choosing representative local centered "objects" according to some criterion. The rationale for focusing on extreme individuals in segmenting markets was noted by Allenby and Glinter (1995), who argued such consumers can be important in new product introduction and switching contexts. We discuss and illustrate a new approach called "archetypal analysis" (AA) based on distance from "important extreme objects." AA has considerable potential for use in marketing, which we illustrate with two examples: 1) identifying segments from responses to attitudinal statements and 2) identifying segments from responses to discrete choice experiments.

Introduction to Archetypal Analysis

Archetypal analysis is a statistical technique proposed by Cutler and Breiman (1994). Cutler and Breiman note similarities and differences of AA compared with cluster centers and the concept of principal points (See Flurry 1990). Use of AA in the hard sciences is growing; for example AA has been applied to study air pollution in Los Angeles, the behaviour of flame cells, head dimensions and plasma in Tokama fusion reactors. Several properties of AA make it attractive for segmenting product markets or consumers. At the risk of oversimplification, taxonomic applications in marketing fall into two broad types: 1) identify and classify individuals into mutually exclusive and exhaustive groups, such as K-Means, Ward's method, etc; or 2) identify groups and let individuals be in more than one group or let group boundaries be fuzzy, such as probabilistic methods or mixture models. AA falls into the second type, and is a mixture model approach that identifies extreme individuals/cases and expresses all others as a probabilistic mixture of the identified extremes.

For a n individuals x m measures matrix $X_{i=1,n}$ (i indexes cases) AA seeks to find K 1 x m vectors, Z_k called archetypes that minimize the following residual sum of squares:

$RSS = \sum_i || X_i - \sum \alpha_{ik} Z_k||^2$,

subject to constraints: 1) the K archetypes lie inside and/or on the convex hull of X ($Z_k = \Sigma_i \beta_{ki} X_i$; $\beta_{ki} > 0$; $\Sigma_i \beta_{ki} = 1$), and 2) predictors of X_i are finite mixtures of archetypes ($\alpha_{ik} \geq 0$; $\Sigma_{k=1,p} \alpha_{ik} = 1$).

For *p* pre-specified archetypes, AA finds archetypes that define the smallest convex hull in a *k*-dimensional space defined by the *k* measures that best encompasses the data subject to constraints that enhance interpretation of results. Constraint one forces archetypes to be actual cases (or linear combinations of cases). Cases inside a convex hull defined by the archetypes incur no loss; cases outside the convex hull incur a loss according to criteria like the square of the projection distance to the hull. Archetypes are chosen to minimize this loss function. Briefly, for starting values of β_{ik}, AA uses a convex least-squares method (CLSM) to estimate α_{ik}, subject to the constraints. The α_{ik}'s, are used to solve for each β_{ik} in turn using the CLSM, holding other β_{ik} fixed. This process is repeated until RSS fails to improve. AA is subject to local minima; hence one uses different starting values to insure global solutions. Scree plots of RSS values for > 2 archetypes can be used to identify optimal numbers of archetypes by choosing p to be the number of archetypes beyond which RSS does not improve.

AA outputs include the location of the *p* archetypes in the *k*-dimensional space, and a vector of *p* coefficients that sum to one for each case and describes how similar each case is to each archetype; hence, they behave like the coefficients of a proper mixture distribution. For cases inside the convex hull, the coefficients are simply normalized distances to each archetype; hence, if one knows archetype locations and coefficient vectors one loses no information relative to a case's original location in *k*-dimensional space. For cases outside the convex hull, coefficients represent distance to the closest projection of the case onto the convex hull.

AA also is a type of fuzzy or probabilistic clustering because cases located at an archetype have a coefficient = 1.0 for that archetype and coefficients = 0.0 for all others. Cases not located at an archetype are mixtures of these pure types, and their coefficients on each archetype = < 1.0 for all *p* archetypes. The magnitudes of the *p* coefficients for each case reflect relative proximity to each archetype. Distributions of the *p* coefficients provide useful information that can be analysed (not illustrated in this paper) by investigating relationships with individual differences measures (e.g., age, income, etc). When X measures are product attributes or consumer characteristics archetypes should be easy to interpret because they represent extreme combinations of attributes or consumer characteristics. In contrast, many cluster methods tend to find cluster centers that are near the middle of *k*-dimensional spaces by construction, which can make it hard to see how clusters differ from one even for moderate *k*. On the other hand, many cases in a data set might qualify as being "extreme" which begs the question of how to identify interesting and important extreme cases. AA does this by defining the *p* most important objects in the data that encompass most cases inside the convex hull defined by the cases, with remaining cases being outside the hull but relatively close to it. AA's use of an encompassing convex hull also avoids imposing artificial orthogonality constraints that underlie many cluster approaches.

Thus, interesting AA solutions involve $p \geq 2$, and a finite number of archetypes will exist that is less than/equal to the number of cases. In practice, a small number of archetypes usually is sufficient to capture most of the information.

Empirical Illustrations Using AA For Segmentation Purposes

We use AA to identify segments from 1) responses to attitudinal statements, and 2) responses to discrete choice experiments. In the first application a sample of 603 consumers were interviewed in shopping malls in six countries; they answered 20 sets of questions involving different combinations of 16 issues shown in Table 1. We used a balanced incomplete block design to make 20 sets of four issues based on the Finn and Louviere (1992) Best-Worst Scaling method. That is, respondents identify both the most and least important issues in each set, which can be used to measure how important each issue is to each respondent. Issue sample means and standard errors are in Table 1, along with results for each archetype. We programmed MATLAB to implement AA, and used it to extract 2 to 12 archetypes, and insured global minima by using 100 random starting values for each solution. Global RSS values for each solution suggested little improvement after five archetypes. The five individuals who represent the extremes or "pure types" are shown in Table 1.

Table 1: Archetypal Analysis Of Ethical Issues

	Sample N=603		Arch1	Arch2	Arch3	Achr4	Arch5
Ethical Issues Studied	**Mean**	**StdError**	**Scores**	**Scores**	**Scores**	**Scores**	**Scores**
animal rights	-0.226	0.104	-5	-1	-1	-1	3
animal by-products	-1.270	0.083	-4	-2	-3	-1	2
biodegradability	-0.433	0.09	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 packaging	-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 labour	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

The interpretation of the archetypes 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 by-products 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 archetypes.

The second example involves a choice experiment administered to samples of City U of Hong Kong undergraduate business students, AGSM MBA students and Amnesty International members (Australia). We varied 14 attributes (12 at two levels + two at four levels), creating extra brand levels from unused two-level columns. Respondents stated if they would/not consider buying shoes described by combinations of attribute levels. These 32 responses were used to calculate achetypes, and four archetypes were identified by inspection of the RSS values. Mixture weights for the archetypes were used to weight MNL models for each segment, and these results are in Table 2.

Table 2 shows that the mixture coefficients significantly reduce the overall MNL likelihoods; hence AA provides valuable 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 suggests that the archetypes can be described as follows based on the 95% confidence interval for the overall sample estimates:
• A1 - 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 - more likely to consider any shoes, less interested in use of child labour in manufacturing and proper accommodation for workers but more price sensitive. This segment emphasizes price regardless of labour practices.
• A3 - more interested in ankle support, less in comfortable fit but very price sensitive. This segment is concerned about support and price.
• A4 - 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 labour, workers paid minimum wages and brand 11.

Table 2: MNL Model Using Archetype Mixtures Parameters As Weights

	Overall Sample MNL Model				Arch1		Arch2		Arch3		Arch4	
Effects	**Coeff**	**StdErr**	**T**	**P(T)**	**Coeff**	**P(T)**	**Coeff**	**P(T)**	**Coeff**	**P(T)**	**Coeff**	**P(T)**
Intercept	-0.578	0.098	-5.920	0.000	-1.111	0.000	0.598	0.002	-0.702	0.000	-1.154	0.000
ShockAbsorb	0.214	0.034	6.300	0.000	0.350	0.000	0.186	0.006	0.255	0.000	0.112	0.243
Weight	-0.193	0.034	-5.720	0.000	-0.195	0.010	-0.206	0.002	-0.210	0.001	-0.262	0.004
AnkleSupport	-0.155	0.033	-4.720	0.000	-0.112	0.126	-0.131	0.043	-0.292	0.000	-0.049	0.576
SoleDurability	0.146	0.034	4.340	0.000	0.113	0.134	0.174	0.009	0.203	0.001	0.135	0.150
Ventilation	0.156	0.034	4.550	0.000	0.233	0.003	0.138	0.043	0.131	0.031	0.171	0.082
FabricMater	-0.050	0.034	-1.500	0.140	0.086	0.267	-0.054	0.425	-0.107	0.079	-0.179	0.060
Reflection	0.006	0.034	0.180	0.860	-0.001	0.988	0.024	0.721	0.010	0.867	-0.029	0.762
ComfyFit	0.272	0.034	7.900	0.000	0.426	0.000	0.251	0.000	0.136	0.026	0.556	0.000
ChildLabor	-0.211	0.037	-5.780	0.000	-0.240	0.004	-0.129	0.080	-0.215	0.001	-0.402	0.000
MinWage	0.106	0.037	2.870	0.004	0.053	0.525	0.109	0.144	0.134	0.041	0.202	0.043
WorkDanger	-0.186	0.037	-5.070	0.000	-0.285	0.001	-0.118	0.114	-0.204	0.002	-0.237	0.017
WorkerAcc	0.110	0.037	3.020	0.003	0.246	0.003	0.013	0.865	0.090	0.167	0.171	0.088
Price	-0.010	0.001	-9.830	0.000	-0.012	0.000	-0.011	0.000	-0.013	0.000	-0.005	0.052
Brand1	0.189	0.072	2.630	0.009	0.552	0.002	0.133	0.309	0.117	0.407	-0.049	0.803
Brand2	0.153	0.072	2.130	0.034	0.021	0.908	0.116	0.374	0.249	0.070	0.281	0.137
Brand3	-0.075	0.193	-0.390	0.698	0.050	0.909	-0.162	0.662	0.053	0.878	-0.331	0.546
Brand4	-0.424	0.259	-1.640	0.102	-0.150	0.809	-0.420	0.353	-0.637	0.209	-0.092	0.893
Brand5	-0.067	0.265	-0.250	0.801	-1.103	0.241	0.221	0.603	-0.054	0.931	-0.213	0.760
Brand6	0.028	0.227	0.120	0.903	0.381	0.461	0.106	0.800	-0.082	0.848	0.057	0.927
Brand7	-0.201	0.215	-0.940	0.349	-0.316	0.557	-0.259	0.506	-0.146	0.709	-0.049	0.934
Brand8	-0.531	0.216	-2.460	0.014	-0.566	0.276	-0.601	0.125	-0.547	0.166	-0.520	0.385
Brand9	0.111	0.192	0.580	0.563	0.127	0.774	0.197	0.601	0.175	0.612	-0.038	0.943
Brand10	0.371	0.188	1.980	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							

Discussion and Conclusions

We described and discussed archetypal analysis, which was introduced to marketing by Carson and Louviere (1998 AMA ART Forum). We extended their paper by demonstrating that AA appears to be useful for discrete choice data. Indeed, a key advantage of AA relative to continuous and discrete mixture model approaches like latent class is that it is "model free" in so far as it does not impose a priori utility or choice model forms on data, but instead gives information on unique individuals who can be analysed to see if it is appropriate to impose particular models on the data, and if so, which model(s) appear to be most appropriate.

References

Allenby, GM & Ginter, JL 1995, “Using Extremes to Design Products and Segment Markets,” *Journal of Marketing Research*, vol.32, (November), pp. 392-403.

Cutler, A & Breiman, L 1994, “Archetypal Analysis,” *Technometrics*, vol. 36, no. 4, pp. 338-347.

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

Frank, RE, Massey, WF & Wind, Y 1972, *Market Segmentation,* Prentice Hall, Englewood Cliffs, NJ.

Flurry, B 1990, “Principal Points,” *Biometrika*, vol. 77, pp. 33-41.

Kaufman, L & Rousseeuw, P 1990, *Finding Groups in Data: An Introduction to Cluster Analysis*, John Wiley & Sons, New York.