Archetypal Analysis: Three Case Studies

Anna Quach* Adele Cutler†

Abstract

Archetypal analysis was first introduced by Adele Cutler and Leo Breiman in 1994. In their paper they presented three examples: Swiss soldiers, air pollution, and Tokamak fusion. We extend the work with three additional case studies including nutrition data from the Cache County Memory and Aging Study, community attachment data provided by the Knight Foundation, and leaf shape data.

Key Words: archetypal analysis; unsupervised learning; cluster analysis; dimension reduction; dietary patterns; shape analysis

1. Introduction

Two popular approaches in unsupervised learning are cluster analysis and principal component analysis (PCA). In cluster analysis, we assume the data fall into more or less distinct subgroups called clusters, with either no data or sparse data between clusters. Cluster means or medians are then used to summarize the cluster centers and these cluster centers are then used for interpretation of the data. One difficulty of this approach is that if the data don't fall into distinct clusters, the various cluster analysis methods can all give quite different results and it's not clear that any of them give meaningful cluster centers. PCA does not assume clusters but attempts to summarize data by producing orthogonal directions that "explain" the variance of the data in such a way that the first PC explains the most variance, the second explains the most subject to being orthogonal to the first, etc. The coefficients for transforming the data into the PCA directions are used to interpret the principal components.

One advantage of cluster analysis over PCA is that the cluster centers are comparable to the data points themselves and can be interpreted accordingly. Interpretation in PCA is less convenient, because we are interpreting directions instead of data-like objects. On the other hand, cluster analysis can be unsatisfactory if clusters don't exist in the data. Archetypal analysis is one way to bridge the gap between the techniques, by producing data-like objects called archetypes that summarize the shape of the dataset but do not assume the existence of clusters.

Cutler and Breiman (1994) first introduced archetype analysis in 1994 and used it to analyze the shape of heads of Swiss soldiers, air pollution, and Tokamak fusion. Since then, archetypal analysis has been applied in areas such as spatio–temporal dynamics (Stone and Cutler, 1996), market research (Li et al., 2003), microarray analysis (Thøgersen et al., 2013), informetrics (Seiler and Wohlrabe, 2013), and behavior learning for video games (Sifa and Bauckhage, 2013).

In this paper, we briefly review archetypal analysis and provide three case studies. The case studies comprise shape analysis of leaves from the poplar, *Populus szechuanica var. tibetica*, food frequency questionnaire (FFQ) data from the Cache County Memory and Aging Study, and survey data from 26 communities in the United States provided by the

*Utah State University, Department of Mathematics and Statistics, 3900 Old Main Hill, Logan, UT 84322–3900, USA. E–mail: aquach4@hotmail.com

†Utah State University, Department of Mathematics and Statistics, 3900 Old Main Hill, Logan, UT 84322–3900, USA. Phone: 435 797 2761, Fax: 435 797 1822, E-mail: adele.cutler@usu.edu

Knight Foundation. To the best of our knowledge, archetypal analysis has yet been applied in these fields.

2. Archetypal Analysis

Archetypal analysis finds a set of archetypal or pure types that are convex linear combinations (or "mixtures") of data points. Each observation is itself approximated by a mixture of the archetypes so that the mean squared error between the observed data points and the archetype-approximated data points is minimized. The result is that the convex hull of the archetypes approximates the convex hull of the original dataset.

More precisely, archetypal analysis finds a set of k archetypes that are mixtures of data points:

$$\mathbf{z}_j = \sum_{i=1}^{n} \beta_{ji}\mathbf{x}_i \text{ for } j = 1, \ldots, k,$$

where

$$\beta_{ji} \geq 0, \sum_{i=1}^{n} \beta_{ji} = 1.$$

Each observation is itself approximated by a mixture of the archetypes:

$$\mathbf{x}_i \approx \sum_{j=1}^{k} \alpha_{ij}\mathbf{z}_j \text{ for } i = 1, \ldots, n,$$

where

$$\alpha_{ij} \geq 0, \sum_{j=1}^{k} \alpha_{ij} = 1.$$

The combined values of the α's and β's are chosen to minimize

$$RSS = \sum_{i=1}^{n} ||\mathbf{x}_i - \sum_{j=1}^{k} \alpha_{ij}\mathbf{z}_j||^2. \tag{1}$$

The algorithm alternates in finding the best archetypes, $\mathbf{z}_1, \ldots, \mathbf{z}_k$, and the best α_{ij}. The approach is an alternating least squares algorithm and is described in Cutler and Breiman (1994).

Choosing the number of archetypes is subjective and we face similar issues to those faced in cluster analysis. The number of archetypes can be as large as the sample size because a convex hull for n data points can have up to n vertices. As with cluster analysis, increasing the number of archetypes will improve the fit, in other words, the residual sum of squares (RSS) in equation 1 will decrease. Using n archetypes will result in an RSS of 0. We discuss choosing the number of archetypes in the case studies.

In order to draw a representation of the convex hull, we let the position of k archetypes be the vertices of a regular k-sided convex polygon. A vector of k coefficients is used to approximate $\mathbf{x}_i$, that is, $\mathbf{x}_i \approx \alpha_{i1}\mathbf{z}_1 + \alpha_{i2}\mathbf{z}_2 + \ldots + \alpha_{ik}\mathbf{z}_k$. If we map $\mathbf{z}_i$ to the ith vertex $\boldsymbol{\mu}_i$ of a regular convex polygon then $\mathbf{x}_i$ can be approximated by $\alpha_{i1}\boldsymbol{\mu}_1 + \alpha_{i2}\boldsymbol{\mu}_2 + \ldots + \alpha_{ik}\boldsymbol{\mu}_k$. The α values sum up to one for each observation, and α_{ij} can be thought of as measuring the extent to which the ith data point resembles the jth archetype. Note that the order in which we map the archetypes to the vertices of the polygon can change the appearance of the representation significantly.

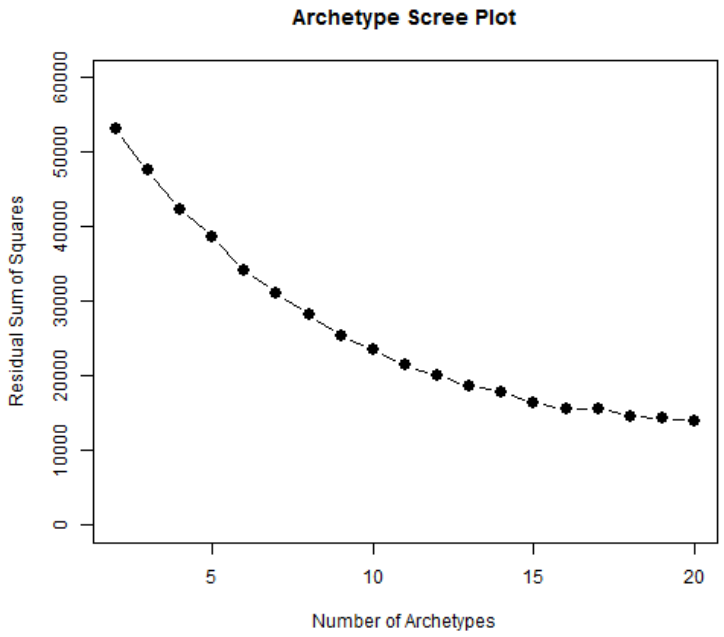

Figure 1: The suggested number of archetypes to keep, using the elbow criterion, is 4, 9, or 16.

3. Case Studies

We illustrate how archetypes can be used to understand data structure using three case studies. The first case study involves nutrition data from the Cache County Memory and Aging Study, the second case study comes from data provided by the Knight Foundation, and the last case study is on leaf shape data.

3.1 Cache County Memory and Aging Study Nutrition Data

Examining associations between nutrient intake and disease is common in the nutrition literature. However, many studies analyzing single nutrients and disease-related illnesses have provided mixed results, e.g., studies on Alzheimer's disease. Single nutrient analysis is a poor representation of the complexity of diets. Defining dietary patterns using food groups may provide better insights into the relationship between diet and the risk of disease.

Diet is assessed among elderly participants 65 years of age or older from the Cache County Study on Memory, Health and Aging by using a 142-item food frequency questionnaire (FFQ) administered in 1995. The 142-items are grouped into 40 food groups and are energy adjusted. We defined dietary patterns using archetypal analysis on the 40 energy adjusted food groups.

Archetypal analysis identifies extreme dietary patterns of elderly participants of Cache County and doesn't assume there are any clusters within the data. Similar to PCA, we can use a scree plot (Figure 1) to get a suggested number of archetypes to keep. The scree plot shows the residual sum of squares (RSS) calculated from equation 1 against the corresponding number of archetypes, starting at 2 archetypes. The suggested number of archetypes can be determined by using an elbow criterion or choosing the number of archetypes where the RSS does not improve noticeably if more archetypes are included. To

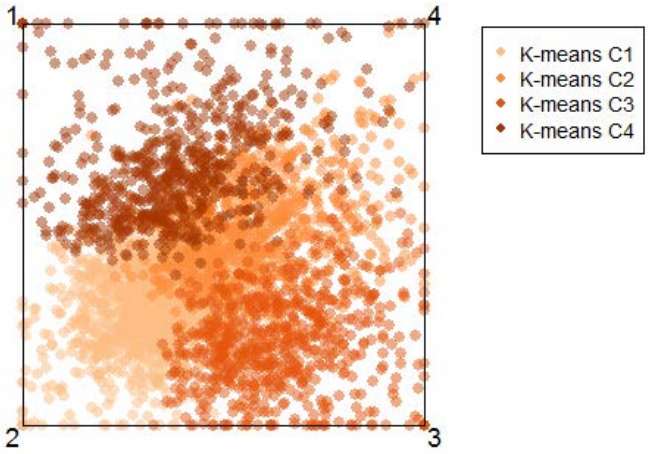

Figure 2: Each point represent the diet of an elderly participant. The diet is described as a mixture of the extreme diets, archetypes one to four.

use the elbow criterion, simply look for the elbow and keep the number archetypes before the elbow. Looking at (Figure 1) this suggests using 4, 9 or 16 archetypes. Keeping a larger number of archetypes will give us a smaller RSS and capture the shape of the data better, however, it may increase the difficulty of interpreting our results. For these data we selected 4 archetypes. Figure 2 is the representation of the four archetype solution. Each point represents an elderly participant in the convex hull (Figure 2). The points with labels 1, 2, 3, and 4 are the archetypes and the elderly participants closest to any of the four archetypes are those with extreme or unusual diets. The points are colored according to the clusters found using a k-means clustering with 4 clusters on the same data. The plot shows that there is considerable overlap between the clusters, at least according to this representation. Interpreting the cluster means is challenging because they are relatively close together. In contrast, the four archetypal diets were quite easily characterized. Figure 3 shows a dot chart of the percentile profile of the archetype, ranked according to the first archetype. We see that archetype 1 is high fruit and vegetable intake, but low intake of sweets, refined grains, and high calorie foods. Similar plots for the other archetypes (not shown) show that archetype 2 has a high intake of sweets, drinks (alcohol, coffee, tea), and low intake of fruits, vegetables, and grains. Archetype 3 has a high intake of refined grains, margarine, and instant foods, but low intake of whole grains, fruit, and desserts. Finally, archetype 4 has a high intake of whole grains, pizza and low fat dairy, but low intake of refined grains, Mexican food, organ meats, and alcohol.

3.2 Community Attachment Data

From 2008 to 2010, the Knight Foundation (Knight Foundation, 2014) conducted the Soul of the Community (SOTC) study (Knight Foundation, 2013). The Knight Foundation is a foundation that supports transformational ideas, helps communities succeed, invests in civic innovations, and promotes civic innovation and robust engagement. Together with

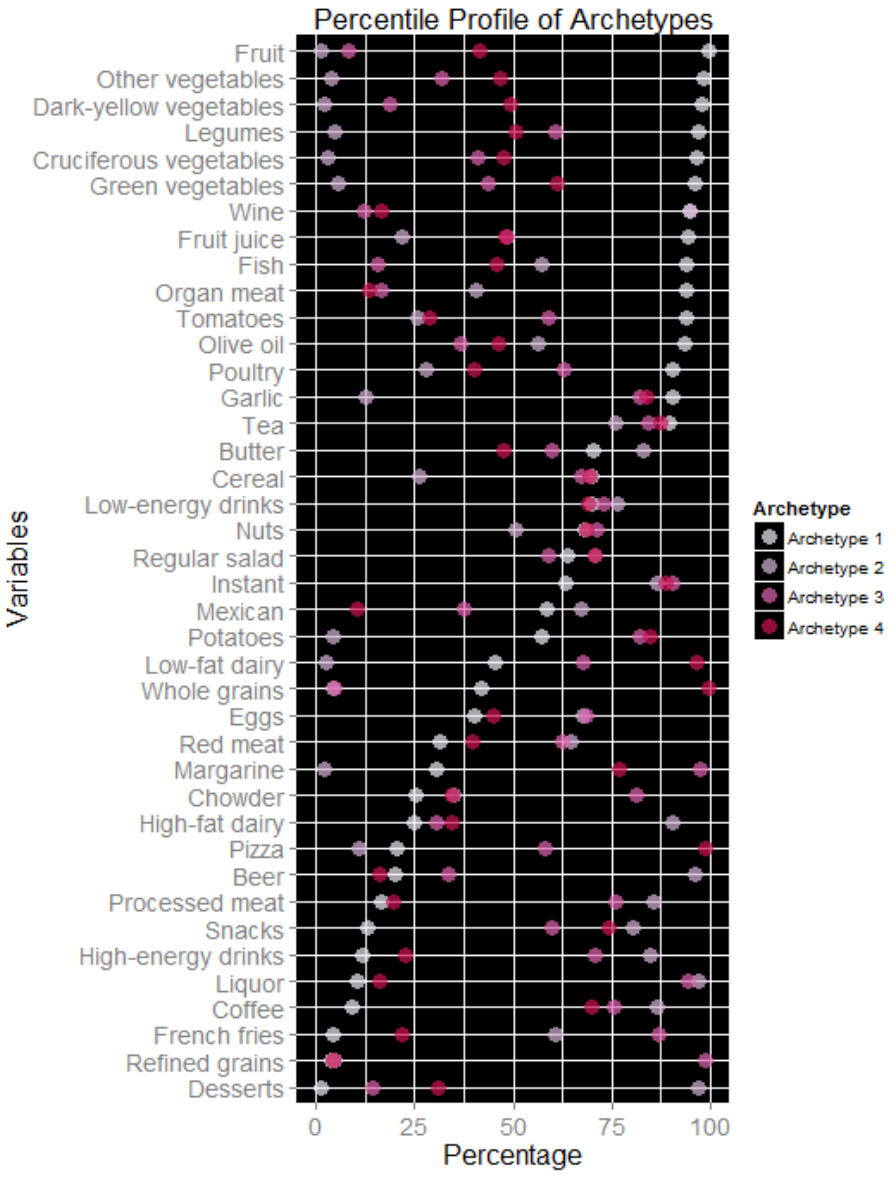

Figure 3: Dotchart of the percentile of the the first archetype. The energy adjusted food groups are ranked by the percentile of the first archetype.

Gallup, they collected data from more than 47,800 participants from 26 communities across the United States. But, unfortunately, the 26 communities were not randomly selected. Rather, these are all the communities where the Knight Foundation has been active in the past (Knight Foundation, 2015). The participants were given a survey with more than 200 questions each year. The questions asked for demographic information, the participants' feelings towards their community, location, and much more.

The analysis was included as part of the bi-annual 2013 Data Expo competition of the Sections on Statistical Graphics and Statistical Computing of the American Statistical Association (ASA). Additional details of the 2013 Data Expo can be found in Hofmann (2013) and in Hofmann and Wickham (2016). Quach et al. (2013) contains the preliminary attempt to assess attachment to a community.

Characterizing communities with similar attachment status in terms of the respondents'

answers to the survey questions is of interest. Attachment is measured at 3 different levels for each participant: not attached, neutral and attached. We used archetypal analysis to identify extreme or unusual communities within the data in the year 2008 and investigated what made these particular communities similar or dissimilar to other communities with respect to their attachment status.

Prior to using archetypal analysis, we used random forests to predict attachment status using the responses to the survey questions for all subjects. The 6 most important survey questions are paraphrased below, where unless otherwise noted, the responses are between 1 (very bad) and 5 (very good).

- **q3c:** The community has a good reputation to outsiders or visitors who do not live here: 1 (strongly disagree), ..., 5 (strongly agree)
- **q5:** If you had the choice of where to live would you rather:

    1. stay in your neighborhood
    2. move to another neighborhood
    3. move outside of your community
    4. move to another city and state

- **q6:** How would you compare how the community is as a place to live today compared to five years ago?
- **q8d:** How good is your community for families with young children?
- **q7i:** How good is your community as a place to meet people and make friends?
- **q7m:** How good is your community in terms of how much people care about each other?

Cases missing in any of the six variables were removed, giving a sample size of 13,256 cases for 2008. Averages were computed for each question for each community and archetypal analysis was applied to these averages. The number of archetypes was selected using a scree plot (not shown). In this case, the scree plot strongly suggested three archetypes for 2008.

A visualization of the convex hull of the archetypal analysis is shown in Figure 4 and an interpretation of the extreme communities is shown in Figure 5. Communities are colored by the dominating attachment status between the three attachment status levels of the community. One tie occurred between those that are neutral and attached in Long Beach, California, and was handled by categorizing that community as neutral. Communities closer to archetype 1 (Gary, Indiana; Detroit, Michigan; Miami, Florida) are dominated by those who are unattached to their community. Those communities closer to archetype 2 (Grand Forks, North Dakota; Columbus, Georgia; Aberdeen, South Dakota) display a mixture of attachment status and the few communities closest to archetype 3 (Bradenton, Florida; Biloxi, Mississippi) are attached to their community. To describe the archetypes themselves, we reference the dot chart in Figure 5. The points in the dot chart represent the percentile for each archetype for each variable as compared to the average response. For example, for archetype 1 variable q3c, the percentile is 11%, indicating that the q3c value in archetype 1 is in the 11th percentile among the average responses for the 26 communities. This would suggest that communities closer to archetype 1 would tend to respond “strongly

disagree" to q3c, that is, when asked if they think their community has a good reputation to outsiders or visitors who do not live in their community. Summarizing using the percentile profile dotchart, all archetypes are describing communities that want to move out of their city and state altogether (q5) and think that their community is much worse today than five years ago (q6). The differences lie in the variables, q3c, q8d, q7i, and q7m. For those particular variables, communities close to archetype 1 can be described as having negative feelings about their community and those close to archetype 2 and 3 are the opposite of this, that is, having more positive feelings.

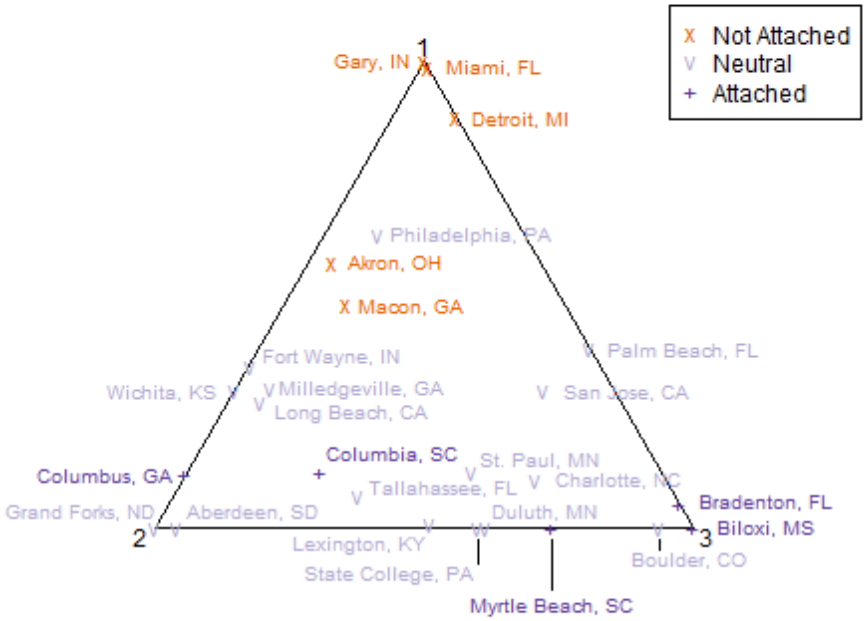

Figure 4: Graphical representation of the three archetype solution for 2008. The three points labeled 1, 2, and 3 are the archetypes. Communities are colored by the dominating group according to the three attachment status levels.

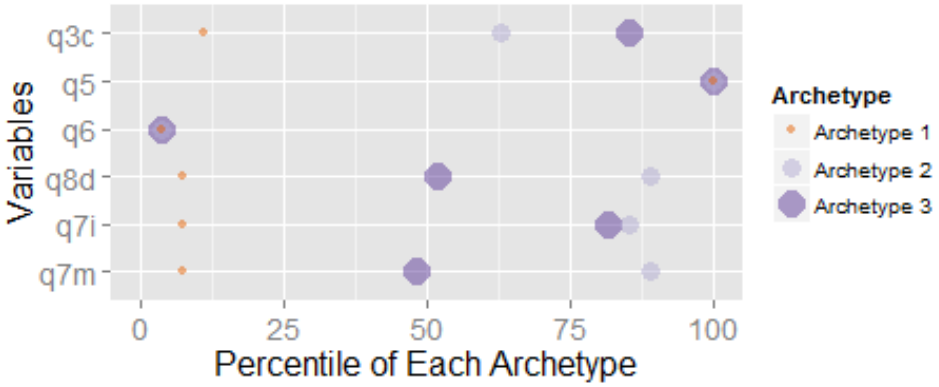

Figure 5: Graphical representation to aid with the interpretation of the three archetype solution for 2008. The points represent the percentile for each archetype for each variable.

3.3 Leaf Shape Data

In this case study, we consider leaves from the poplars of the Tacamahaca section. The poplar, *Populus szechuanica var. tibetica*, is grown naturally throughout the Tibet Plateau in the mountains at altitudes between 110 and 4600 meters (Hamzeh and Dayanandan, 2004). The data for this case study come from 106 poplar leaves, after preprocessing as described in Fu et al. (2013). Archetypal analysis was performed on the 360 radii of the leaves.

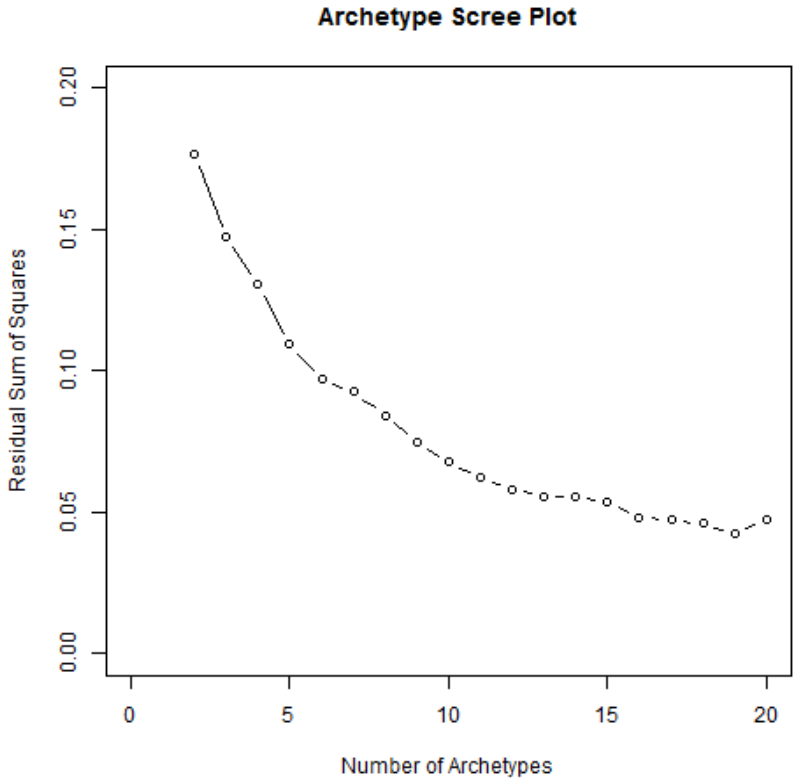

Figure 6: Scree plot for leaf data, giving the suggested number of archetypes to keep by using the elbow criterion. The residual sum of squares is determined by calculating equation 1 for each number of archetypes.

The scree plot in Figure 6 does not give clear guidance about how many archetypes to keep, so we consider 2, 3, 4 and 5 archetypes. The archetypal leaves for each of these solutions are displayed in Figure 7. A two archetypal solution identifies a leaf that's skinny and long and a leaf of opposite shape, i.e, fat and short. In this particular data set, the previous archetypal leaves happen to carry over to the next archetype solution. For example, in the three archetype solution, we see the same archetypal leaves as in the two archetype solution, with an additional archetype that is a leaf whose shape falls in between the previous two archetypes. The four archetype solution identifies an additional archetype similar to the short and fat leaf, but the base of the leaf differs. Last, the additional archetype solution in the five archetype solution is an outlier whose shape is quite long and skinny. If we were to based our decision of how archetypes to keep on this figure, we would keep either a three or four archetype solution since keeping a five archetype solution would make our interpretation more complex and a two archetype solution seems to be too simple.

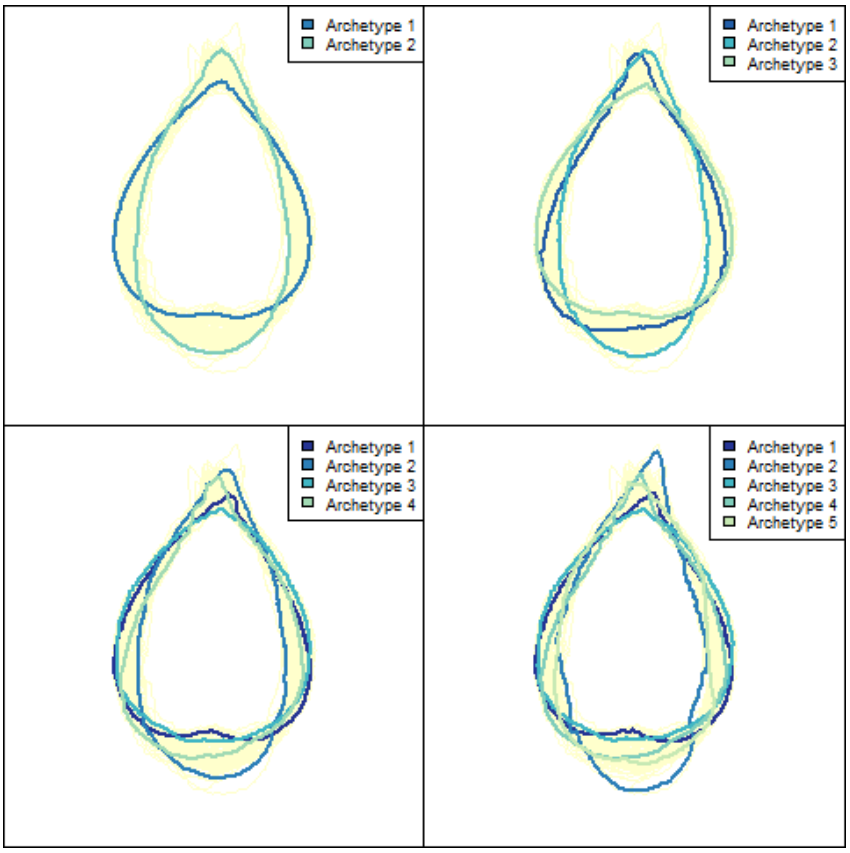

Figure 7: Each archetypal leaf is drawn for a two, three, four, and five archetype solution.

Figure 8 gives the visualization of the convex hull for four archetypes. At each archetype we visualize what the shape of the archetypal leaf would look in comparison to the 106 leaves. The points within the convex hull represent the leaves as mixtures of the four archetypal leaves. The points closest to a given archetype are the leaves that resemble the archetype most closely.

4. Discussion & Concluding Remarks

Archetypal analysis combines aspects of both cluster analysis and PCA and has been shown to provide new insights to the structure of multivariate data. The convex hulls in archetypal analysis can be used to identify clusters if they exist, or to examine covariates, as demonstrated in the case studies. Archetypal analysis is sensitive to outliers, but can be used to detect outliers. The outlier would be identified as an archetype and it would be isolated from the rest of the points. Outliers are determined by examining the polygonal representation of the convex hull.

5. Software

An open source implementation of archetypal analysis is available in R. However, we illustrate the use of archetypal analysis using Adele Cutler's implementation, which is available from her on request.

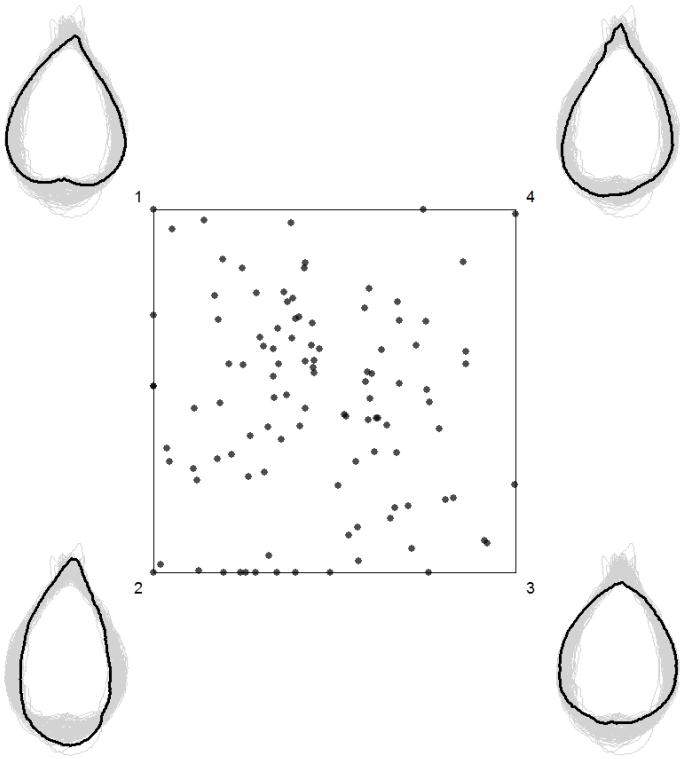

Figure 8: The convex hull of the four archetype solution includes a visualization of each archetypal leaf.

6. Acknowledgements

We acknowledge Guifang Fu (Department of Mathematics and Statistics, Utah State University) and Heidi Wengreen (Department of Nutrition and Food Sciences, Utah State University) for their support and extremely helpful discussions about the leaf data and nutrition data, respectively. Any errors in interpretation are ours alone. This research was supported by grants by NIH grant R01-AG-11380, R15-AG-037392, the General Mills Bell Institute of Health, the Agriculture Experiment Station at Utah State University, and NSF grant DMS-1413366.

References

Cutler, A., Breiman, L., 1994. Archetypal Analysis. Technometrics 36 (4), 338–347.

Fu, G., Bo, W., Pang, X., Wang, Z., Chen, L., Song, Y., Zhang, Z., Li, J., Wu, R., 2013. Mapping shape quantitative trait loci using a radius-centroid-contour model. Heredity 110 (6), 511–519.

Hamzeh, M., Dayanandan, S., 2004. Phylogeny of populus (salicaceae) based on nucleotide sequences of chloroplast trnt-trnf region and nuclear rdna. American journal of botany 91 (9), 1398–1408.

Hofmann, H., 2013. Soul of the Community. http://streaming.stat.iastate.edu/dataexpo/2013/ (last accessed November 12, 2013).

Hofmann, H., Wickham, H., 2016. Computational StatisticsForthcoming.

Knight Foundation, 2013. Soul of the Community. http://www.soulofthecommunity.org/ (last accessed November 12, 2013).

Knight Foundation, 2014. http://www.knightfoundation.org (last accessed May 23, 2014).

Knight Foundation, 2015. http://www.knightfoundation.org/about/ (last accessed March 3, 2015).

Li, S., Wang, P., Louviere, J., Carson, R., 2003. Archetypal analysis: A new way to segment markets based on extreme individuals. In: A Celebration of Ehrenberg and Bass: Marketing Knowledge, Discoveries and Contribution. Proceedings of the ANZMAC 2003 Conference. pp. 1674–1679.

Quach, A., Symanzik, J., Forsgren Velasquez, N., 2013. Soul of the Community: A First Attempt to Assess Attachment to a Community. In: 2013 JSM Proceedings. American Statistical Association, Alexandria, VA.

Seiler, C., Wohlrabe, K., 2013. Archetypal scientists. Journal of Informetrics 7 (2), 345–356.

Sifa, R., Bauckhage, C., 2013. Archetypical motion: Supervised game behavior learning with archetypal analysis. In: Computational Intelligence in Games (CIG), 2013 IEEE Conference on. IEEE, pp. 1–8.

Stone, E., Cutler, A., 1996. Introduction to archetypal analysis of spatio-temporal dynamics. Physica D: Nonlinear Phenomena 96 (1), 110–131.

Thøgersen, J. C., Mørup, M., Damkiær, S., Molin, S., Jelsbak, L., 2013. Archetypal analysis of diverse pseudomonas aeruginosa transcriptomes reveals adaptation in cystic fibrosis airways. BMC bioinformatics 14 (1), 279.