ARCHETYPAL ANALYSIS

By

Adele Cutler
Mathematics and Statistics
Utah State University
Logan, UT 84322-3900

Leo Breiman
Department of Statistics
University of California
Berkeley, CA 94720

Technical Report No. 379
revised October 1993

Department of Statistics
University of California
Berkeley, CA 94720

ARCHETYPAL ANALYSIS

Adele Cutler

Mathematics and Statistics

Utah State University

Logan, UT 84322-3900

Leo Breiman

Department of Statistics

University of California

Berkeley, CA 94720

Abstract

Archetypal analysis represents each "individual" in a data set as a mixture of "individuals of pure type", or "archetypes". The archetypes themselves are restricted to be mixtures of the individuals in the data set. Archetypes are selected by minimizing the squared error in representing each individual as a mixture of archetypes. The usefulness of archetypal analysis is illustrated on a number of data sets. Computing the archetypes is a nonlinear least squares problem which is solved using an alternating minimizing algorithm.

KEY WORDS: Archetypes, principal components, convex hull, graphics, nonlinear optimization.

1. INTRODUCTION

For multivariate data $\{\boldsymbol{x}_i, i = 1,\ldots,n\}$ where each $\boldsymbol{x}_i$ is an m-vector $\boldsymbol{x}_i = (x_{1i},\ldots,x_{mi})^t$ an interesting problem is to find m-vectors $\boldsymbol{z}_1,\ldots,\boldsymbol{z}_p$ that characterize the "archetypal patterns" in the data. For instance, a data set analyzed by Flury and Riedwyl (1988) consists of 6 head dimensions for 200 Swiss soldiers. The purpose of the data was to help design face masks for the Swiss Army.

A natural question is whether there are a few "pure types" or "archetypes" of heads such that the 200 heads in the data base are mixtures of the archetypal heads.

One possible answer is provided by a variant of principal components. For given m-vectors $\boldsymbol{z}_1,\ldots,\boldsymbol{z}_p$, the linear combination $\sum_k \alpha_{ik}\boldsymbol{z}_k$ that best approximates $\boldsymbol{x}_i$ is defined as the minimizer of

$$\|\boldsymbol{x}_i - \sum_k \alpha_{ik}\boldsymbol{z}_k\|^2.$$

Then the "best patterns" $\boldsymbol{z}_1,\ldots,\boldsymbol{z}_p$ are the minimizers of

$$\sum_i \|\boldsymbol{x}_i - \sum_k \alpha_{ik}\boldsymbol{z}_k\|^2. \tag{1.1}$$

Without loss of generality, take $\boldsymbol{z}_1,\ldots,\boldsymbol{z}_p$ to be orthonormal. Then the minimizers of (1.1) maximize

$$\sum_{k=1}^{p} \boldsymbol{z}_k^t S \boldsymbol{z}_k \tag{1.2}$$

where $S = X^tX$. The maximizers of (1.2) are the eigenvectors of S corresponding to the p largest eigenvalues. Thus, if each $\boldsymbol{x}_i$ is centered at its mean, the solution is given by the principal components decomposition.

The "patterns" derived this way are usually not an answer to the problem posed above. For instance, the first four "patterns" found using the Swiss Army data do not correspond to any real or even fictitious heads. In some of the patterns, the distance between two points on the head is negative.

This is not surprising, given that the principal components approach nowhere requires either that the "patterns" resemble "pure types" in the data, or that each $\boldsymbol{x}_i$ be approximated by a mixture of the patterns (i.e. $\alpha_{ik} \geq 0$, $\sum_k \alpha_{ik} = 1$).

In archetypal analysis the patterns $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_p$ considered are mixtures of the data values $\{\boldsymbol{x}_i\}$. Furthermore, the only approximations to $\boldsymbol{x}_i$ allowed are mixtures of the $\{\boldsymbol{z}_k\}$.

More precisely, for fixed $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_p$ where

$$\boldsymbol{z}_k = \sum_j \beta_{kj} \boldsymbol{x}_j, k = 1, \ldots, p$$

and $\beta_{ki} \geq 0$, $\sum_i \beta_{ki} = 1$, define the $\{\alpha_{ik}\}$, $k = 1, \ldots, p$ as the minimizers of

$$\|\boldsymbol{x}_i - \sum_{k=1}^{p} \alpha_{ik} \boldsymbol{z}_k\|^2$$

under the constraints $\alpha_{ik} \geq 0$, $\sum_k \alpha_{ik} = 1$. Then define the archetypal patterns or archetypes as the mixtures $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_p$ that minimize

$$\sum_i \|\boldsymbol{x}_i - \sum_{k=1}^{p} \alpha_{ik} \boldsymbol{z}_k\|^2$$

and denote the minimum value by $RSS(p)$. For $p > 1$, the archetypes fall on the convex hull of the data (see Section 3). Thus the archetypes are extreme data-values

such that all of the data can be well-represented as convex mixtures of the archetypes. But the archetypes themselves are not wholly mythological, since each is constrained to be a mixture of points in the data.

In contrast to principal components analysis, archetype analysis does not nest, nor are the successive archetypes orthogonal to one another. As more archetypes are found, the existing ones can change to better capture the shape of the dataset. However, as we hope the examples will show, archetypes can give a uniquely informative way to understand multivariate data and curves.

The paper is organized as follows: Section 2 gives examples of archetype analysis as applied to data. Section 3 discusses the locations of the archetypes. Section 4 contains a description at the algorithm used to compute archetypes. Section 5 contains some results regarding convergence of the algorithm and Section 6 gives a brief summary.

Previous work that has the flavor of archetypal analysis is mainly based on principal components. A natural approach is to use the quantiles of the principal component scores to select "representative" individuals. For example Jones and Rice (1992) use principal components to summarize a large number of curves. The principal components themselves are informative, but additional information is obtained by selecting the curves corresponding to the median, minimum, and maximum values of the principal component score. However, such choices may be misleading, particularly if the

principal components themselves are difficult to interpret.

Flury and Tarpey (1992) suggest that if extreme curves are required, they might be chosen by considering those curves for which the Mahalanobis distance from the mean is large. However, the curves with large Mahalanobis distance may in fact be very similar to each other, and may not reflect the extremes present in the data.

The analysis of the Swiss Army data (Example 2.1) by Flury (1993) was based on "principal points", a concept similar to that of cluster centers. This method has also been used to get representative curves as an alternative to the Jones-Rice approach (see Flury 1990, 1993). One feature of principal points which is not shared by archetypes is that principal points is a concept for theoretical distributions.

Other related work is that of Woodbury and Clive (1974), who use maximum likelihood estimation based on grades of membership to derive "pure types". Similar ideas are also evident in latent class analysis (Lazarsfeld and Henry, 1968) and latent budget analysis (De Leeuw and van der Heijden 1991, van der Heijden et al. 1992).

2. EXAMPLES

The three following examples illustrate how archetypes can be used to understand data structure. The first example, involving head measurements of Swiss Army soldiers, is given because of its intuitive appeal. The second and third examples are more serious applications, involving air pollution and Tokamak fusion data.

2.1. *Swiss Army Head Dimension Data*

The Swiss Army data consists of six measurements on each head. Two are measures of the width of the face just above the eyes and just below the mouth. The 3rd is the distance from the top of the nose to the chin, the 4th the length of nose, and the 5th and 6th are the distances from the ear to the top of the nose and chin respectively. Figure 1 pictures the archetypal heads for $p = 2, 3, 4, 5$.

These pictures (Figure 1) are given as graphical illustration of the idea of archetypes. They are "extreme" or "pure" types as patterns such that each real individual can be well approximated by a mixture of the "pure types" or archetypes.

Figure 2 shows the values of $100 \times RSS(p)/RSS(1)$. In Section 3, we note that for $p = 1$, the single archetype is the mean of the $\{\boldsymbol{x}_i\}$. Thus $RSS(1)$ is simply the total sum-of-squares $\sum_i \|\boldsymbol{x}_i - \bar{\boldsymbol{x}}\|^2$ and the ratio $100 \times RSS(p)/RSS(1)$ measures the percent decrease in squared error when p archetypes are used to represent the data.

2.2. *Air Pollution Data*

This data consists of measurements of data relevant to air pollution in the Los Angeles Basin in 1976. There are 330 complete cases consisting of daily measurements on the variables

- ozone (OZONE)
- 500 millibar height (500MH)

- wind speed (WDSP)
- humidity (HMDTY)
- surface temperature (STMP)
- inversion base height (INVHT)
- pressure gradient (PRGRT)
- inversion base temperature (INVTMP)
- visibility (VZBLTY)

These data were standardized to have mean zero and variance one, and archetypes were computed. Figure 3 is a graph of $100 \times RSS(p)/RSS(1)$. We focus on three archetypes.

Figure 4 displays the percentile value of each variable in an archetype as compared to the data. For example, the height of the first bar for OZONE in Archetype 1 is 92. This indicates that the OZONE value in archetype 1 is in the 92nd percentile of the 330 OZONE readings in the data.

Archetype 1 is high in OZONE, 500MH, HMDTY, STMP, INVTMP and low in INVHT and VZBLTY. This indicates a typical hot summer day. The nature of the other two archetypes is less clear. The PRGRT is predominantly measured in the north-south direction. A low percentile value indicates a large negative pressure

gradient, and a high value, a large positive gradient. The differences in PRGRT and WDSP in archetypes 2 and 3 indicates a connection with air mass motion in the basin. The temperatures are lower in archetype 3, so it seems to represent cooler days toward winter.

We can get more insight by looking at another graphical representation. With three archetypes z_1, z_2, z_3, the vector of variables x_i for the ith day is best approximated by the mixture $x_1 \simeq \alpha_{i1}z_1 + \alpha_{i2}z_2 + \alpha_{i3}z_3$. There is a simple way to get a two-dimensional data representation. Let μ_1, μ_2, μ_3 be the vertices of a two-dimensional equilateral triangle, and map $z_i \rightarrow \mu_i$, $i = 1, 2, 3$. Then we represent x_i by $\alpha_{i1}\mu_1 + \alpha_{i2}\mu_2 + \alpha_{i3}\mu_3$.

Figure 5 a) - d) gives such plots separately for each of the four seasons. Clearly, the summer days cluster close to the 1st archetype. Spring mixes mainly the 1st and 3rd; Fall, the 1st and 2nd; and Winter the 2nd and 3rd.

The archetype mixture coefficients can also be used to see how the individual variables vary as functions of archetypes. For instance, let the mixture coefficients of the ith day's data be α_{i1}, α_{i2}, α_{i3}. If O_i is the OZONE value for the ith day, we would generally expect O_i to be large if α_{i1} is close to one, and smaller otherwise (see Figure 4).

To make this more specific, O_i was regressed on terms of up to 3rd degree in α_{i1}, α_{i2}, α_{i3} (actually only on terms in α_{1i}, α_{i2} since $\alpha_{i1} + \alpha_{i2} + \alpha_{i3} = 1$). The resulting

prediction equation for OZONE as a function of α_1, α_2, α_3 is plotted as a surface in Figure 6(a). As noted there, the R^2 of the equation is .85.

The values are normalized before fitting so that zero represents the lowest value in the 330 data values of OZONE and one, the highest. The vertical pole in the Figure has height one.

The results show that OZONE is well determined by the mixture coefficients, nearly zero between archetypes 2 and 3 and rising toward the maximum near archetype 1. Other plots give interesting but different information. For instance Figure 6(b) of the INVTMP shows moderate temperature at archetype 2 increasing to a maximum at archetype 1, and with an R^2 of .93.

The plot of INVHT (Figure 6(c)) has $R^2 = .69$ with an interesting nonlinearity between archetypes 1 and 2, staying close to its minimum value until almost halfway to archetype 2. All of the variables have R^2 at around .8 or higher except for inversion height (.69), wind speed (.45) and visibility (.37). The plot of WDSP is given in Figure 6(d).

Although this data set has been extensively studied in the literature (Breiman and Friedman (1985), Hastie and Tibshirani (1990), among others), the archetypal analysis reveals new aspects. The data (except for two variables) can be surprisingly well-represented as a mixture of three archetypal days.

This analysis is a vest-pocket edition of the problem that initiated this study

of archetypes. The EPA has funded elaborate computer models to simulate the production of ozone in the lower atmosphere. Hundreds of chemical equations are embedded in the codes. The usual running time is (or was) as slow as real time, i.e. a 24 hour computer run is needed to simulate 24 real hours.

Given this, in a typical project, only a few days can be modeled. The problem becomes to select data representing a few "prototypical" days. This selection problem led to the idea of archetypes.

2.3. *Tokamak Fusion Data*

A Tokamak resembles a giant hollow donut filled with hot plasma. In each run, a strong external magnetic field is imposed. A current is induced in the plasma inside the donut, and causes the lines of magnetic flux to spiral. Physical theory has not been able to accurately model the complex plasma conditions. So understanding the statistical structure of the experimental results is an important undertaking. In particular, one outstanding problem has been to understand how the shapes of the temperature profiles relate to the covariates. Pioneering work on this issue has been done by Kurt Riedel and coworkers (see Riedel and Imre 1993, McCarthy, Riedel, et al. 1991, Kardaun, Riedel et al. 1990). Archetypal analysis gives another view.

We use a data set containing 40 temperature profiles from the Tokamak Fusion Test Reactor at the Princeton Plasma Physics Laboratory (see Hiroe et al. 1988). Each profile consists of 61 plasma temperature measurements (in KeV) at values of

the radius ranging from 1.8m to 3.2m. Figure 7 is a plot of log temperature vs radius for the 40 profiles.

In each of the 40 runs, there were 5 global covariates

- ESF: edge safety factor
- LPC: log plasma current (Amperes)
- TMF: toroidal magnetic field (Tesla)
- LVG: loop voltage (Volts)
- LPD: log particle density (particles per cubic meter).

The edge safety factor, the most important covariate, is related to the spiraling of the toroidal magnetic field lines generated by the Tokamak current.

Start by smoothing the curves using smoothing splines. Results are in Figure 8. To focus on the shapes rather than on scale differences, we ignored the regions of radius $R \leq 2.2$ and $R \geq 3.0$ where there was little shape difference, and used only 35 values for each curve. The curves were shifted up or down to have the same value at $R = 2.2$, and then divided by their average over the remaining R-range (see Figure 9).

Archetypes were extracted, treating each curve as a point in 35-dimensional space, and $100 \times RSS(p)/RSS(1)$ graphed in Figure 10. We focus on 3 archetypes (Figure

11). The two-dimensional representation is given in Figure 12, and shows that most of the curves are mixtures of archetypes 1 and 3, but with some significant pulls toward archetype 2.

The surface plots of the five covariates, scaled in the same way as the ozone surface plots are in Figure 13 a) - e). Because there are only 40 data points, the regression used only the linear and quadratic terms in the mixture coefficients giving 5 independent variables.

The R^2 were

- ESF .71
- LPC .44
- TMF .24
- LVG .19
- LPD .07

Given that we are using 40 cases and 5 variables, some of these R^2 are substantial.

The surface plots show that the archetypes and the covariates are associated as follows

Archetype	ESF	LPC	TMF	LVG
1	low	high	moderate	high
2	low	moderate	low	moderate
3	high	low	high	low

This archetypal analysis gives some new and interesting insights into the relationships between the temperature profiles and the covariates. Much more extensive statistical work needs to be done in this area.

3. LOCATION OF THE ARCHETYPES

The following proposition helps in understanding the nature of archetypes.

PROPOSITION 1. Let $\mathcal{C}$ be the convex hull of $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$. Let $\mathcal{S}$ be the set of data points on the boundary of $\mathcal{C}$ and let N be the cardinality of $\mathcal{S}$.

(i) If $p = 1$, choosing $\boldsymbol{z}$ to be the sample mean minimizes RSS.

(ii) If $1 < p < N$, there is a set of archetypes $\{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_p\}$ on the boundary of $\mathcal{C}$ which minimize RSS.

(iii) If $p = N$, choosing $\{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_p\} = \mathcal{S}$ results in $RSS = 0$.

Proof. In each case, it is easily verified that the proposed archetypes are mixtures of the data. It remains to show that the archetypes minimize the RSS. For (i), the sample mean is the unconstrained minimizer of the RSS. For (ii), suppose without

loss of generality that z_1 is strictly interior to $\mathcal{C}$, let

$$z(t) = z_j + t(z_1 - z_j), \text{ for } t > 1 \text{ and } j \neq 1,$$

and choose t so that $z(t)$ is on the boundary of $\mathcal{C}$. For $z_1, \ldots, z_p$ fixed, RSS is minimized with respect to the α's by choosing $\sum_{k=1}^{p} \alpha_{ik} z_k$ to be the point in the convex hull of $z_1, \ldots, z_p$ that is closest to x_i. But the convex hull of $z(t), z_2, \ldots, z_p$ contains the convex hull of $z_1, \ldots, z_p$, so $z(t), z_2, \ldots, z_p$ provide a larger set over which to minimize (1) with respect to the α's. For (iii), the convex hull of $z_1, \ldots, z_p$ is $\mathcal{C}$, so $RSS = 0$.

The editor raised the question of where the archetypes of simple distributions are located. In general, the locations are quite data dependent, and sensitive to outliers. Since analytic results seem formidable, the following simulation was done:

1) Generate a sample of size 1000 from $N(\mu, \sum)$ where $\mu = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \sum = \begin{pmatrix} 1 & .8 \\ .8 & 1 \end{pmatrix}$. Discard any points outside the 95% density contour, i.e. with Mahalanobis Distance $\geq \chi^2_2(.95)$.

2) Fit 4 archetypes.

3) Repeat 100 times.

The plots of all archetypes are in Figure 14. They cluster around the ends of the

major and minor axes of the 95% density contour. This is what we expected, but its comforting to get confirmation.

4. THE ARCHETYPE ALGORITHM

Let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ be n m-dimensional data points. The problem is to find $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_p$ where

$$\boldsymbol{z}_k = \sum_{j=1}^{n} \beta_{kj} \boldsymbol{x}_j$$

and $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_p$ minimize

$$RSS = \min_{\{\alpha_{ik}\}} \sum_{i=1}^{n} \left\| \boldsymbol{x}_i - \sum_{k=1}^{p} \alpha_{ik} \boldsymbol{z}_k \right\|^2 \tag{4.1}$$

subject to the constraints

$$\begin{aligned}
\alpha_{ik} &\geq 0 \quad \text{for} \quad i = 1, \ldots, n; \; k = 1, \ldots, p \\
\sum_{k=1}^{p} \alpha_{ik} &= 1 \quad \text{for} \quad i = 1, \ldots, n \\
\beta_{kj} &\geq 0 \quad \text{for} \quad k = 1, \ldots, p; \; j = 1, \ldots, n \\
\sum_{j=1}^{n} \beta_{kj} &= 1 \quad \text{for} \quad k = 1, \ldots, p.
\end{aligned}$$

The $\boldsymbol{z}_k$'s are the *archetypes*; the algorithm used to compute them will be called the *archetype algorithm.*

The residual sum of squares in (4.1) can be written

$$RSS = \sum_{i=1}^{n} \left\| \boldsymbol{x}_i - \sum_{k=1}^{p} \alpha_{ik} \sum_{j=1}^{n} \beta_{kj} \boldsymbol{x}_j \right\|^2,$$

and the archetype problem is to find α's and β's to minimize this RSS subject to the constraints above. This problem could be solved using a general-purpose constrained nonlinear least squares algorithm, but this may prove impractical for all but the smallest problems. Instead, we propose an alternating constrained least squares algorithm.

4.1 Alternating Optimization

The algorithm alternates between finding the best α's for a given set of x-mixtures, and finding the best x-mixtures for a given set of α's. Each step requires the solution of several convex least squares (CLS) problems of the form

> Given $\mathbf{u}$ and $\mathbf{t}_1, \ldots, \mathbf{t}_q$, find $w_1, \ldots, w_q$ to minimize $\|\mathbf{u} - \sum_{k=1}^{q} w_k \mathbf{t}_k\|^2$ subject to $w_k \geq 0$ for $k = 1, \ldots, q$ and $\sum_{k=1}^{q} w_k = 1$.

At each step, the sum of squares (4.1) is reduced, and the algorithm stops when the reduction is sufficiently small.

First consider finding the best α's for a given set of x-mixtures $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_p$. Finding the best α's requires the solution of n CLS problems, namely, minimizing

$$\left\| \boldsymbol{x}_i - \sum_{k=1}^{p} \alpha_{ik} \boldsymbol{z}_k \right\|^2$$

for each i, subject to $\alpha_{ik} \geq 0$ for $k = 1, \ldots, p$ and $\sum_{k=1}^{p} \alpha_{ik} = 1$. Each of the n CLS problems has m observations and p variables.

Now consider finding the best x-mixtures for the current set of α's. If all but one of the x-mixtures are held constant, we show that the remaining one can be found by solving a CLS problem. More precisely, if z_l is the x-mixture of interest, let

$$\begin{aligned} \mathbf{v}_i &= \left(x_i - \sum_{k \neq l}^{p} \alpha_{ik} z_k \right) / \alpha_{il} \\ \bar{\mathbf{v}} &= \sum_{i=1}^{n} \alpha_{il}^2 \mathbf{v}_i \Big/ \sum_{i=1}^{n} \alpha_{il}^2 \end{aligned}$$

and write (4.1) as

$$\begin{aligned} f &= \sum_{i=1}^{n} \alpha_{il}^2 \left\| \mathbf{v}_i - z_l \right\|^2 \\ &= \sum_{i=1}^{n} \alpha_{il}^2 \left\| \mathbf{v}_i - \bar{\mathbf{v}} \right\|^2 + \sum_{i=1}^{n} \alpha_{il}^2 \left\| \bar{\mathbf{v}} - z_l \right\|^2 . \end{aligned}$$

Since the first term does not depend on z_l, minimizing f is equivalent to minimizing $\|\bar{\mathbf{v}} - z_l\|^2 = \|\bar{\mathbf{v}} - \sum_{j=1}^{n} \beta_{lj} x_j\|^2$, subject to the constraints $\beta_{lj} \geq 0$ for $j = 1, \ldots, n$ and $\sum_{j=1}^{n} \beta_{lj} = 1$, which is a CLS problem with n variables and m observations.

The entire collection of x-mixtures is found by cycling through the set, optimizing with respect to each x-mixture in turn, until the improvement in the objective function from an entire pass is smaller than some prescribed tolerance. The resulting x-mixtures are the archetypes.

4.2 Implementation

Many methods are available for solving CLS problems. The method used to develop and test the algorithm is a penalized version of the NNLS algorithm in Lawson

and Hanson (1974). In particular, we obtain $\tilde{\mathbf{u}}$ and $\tilde{\mathbf{t}}_1, \ldots, \tilde{\mathbf{t}}_p$ by adding an extra element M to $\mathbf{u}$ and $\mathbf{t}_1, \ldots, \mathbf{t}_p$, where M is large. Then

$$\left\| \tilde{\mathbf{u}} - \sum_{k=1}^{p} w_k \tilde{\mathbf{t}}_k \right\|^2 = \left\| \mathbf{u} - \sum_{k=1}^{p} w_k \mathbf{t}_k \right\|^2 + M^2 \left\| 1 - \sum_{k=1}^{p} w_k \right\|^2,$$

which is minimized under non-negativity restrictions. For large M, the second term dominates and forces the equality constraint to be approximately satisfied, while maintaining the non-negativity constraint. The speed of the archetype algorithm is determined by the efficiency of the CLS method. The penalized non-negative least squares method is appealing because it can be used when the number of variables is larger than the number of observations. However, it is still quite slow, and alternative convex least squares procedures are being developed (Cutler 1993).

Initially, the x-mixtures may be chosen at random without replacement from the data. Some caution should be exercised in choosing initial x-mixtures that are not too close together, since this can cause slow convergence or convergence to a local optimum.

At any stage, if one or more of the x-mixtures is inside the convex hull of the others, $\sum_{i=1}^{n} \alpha_{il}^2$ can be zero. When this occurs, the archetype is redundant and may be replaced by that data point $\boldsymbol{x}_i$ for which $\|\boldsymbol{x}_i - \sum_{k=1}^{p} \alpha_{ik} \boldsymbol{z}_k\|^2$ is the largest.

5. CONVERGENCE

As with many alternating optimization algorithms, the archetype algorithm can

be shown to result in a fixed point of an appropriate transformation, but there is no guarantee that this will be a global minimizer of RSS.

First consider the inner loop used to compute the β's. Let

$$\begin{aligned}\boldsymbol{\beta} &= (\beta_{11},\ldots,\beta_{1n},\ldots,\beta_{p1},\ldots,\beta_{pn})^t \\ &= \left(\mathbf{b}_1{}^t,\ldots,\mathbf{b}_n{}^t\right)^t,\end{aligned}$$

for $\mathbf{b}_k = (\beta_{k1},\ldots,\beta_{kn})^t$. The inner loop produces iterates $\boldsymbol{\beta}_1,\boldsymbol{\beta}_2,\ldots$ by minimizing (1) with respect to the current $\mathbf{b}_k$ while holding the others fixed. This gives $f(\boldsymbol{\beta}_1) \leq f(\boldsymbol{\beta}_2) \leq \ldots$, and since the set $\mathcal{B}$ of feasible $\boldsymbol{\beta}$'s is compact, the iterations for the inner loop have a limit point $\boldsymbol{\beta}^\star$.

Treating RSS in (4.1) as a function of $\boldsymbol{\beta}$ gives

$$f(\boldsymbol{\beta}) = \sum_{i=1}^{n} \boldsymbol{x}_i{}^t\boldsymbol{x}_i - 2\mathbf{c}^t\boldsymbol{\beta} + \boldsymbol{\beta}^t H\boldsymbol{\beta}$$

where $\mathbf{c} = (c_{11},\ldots,c_{1n},\ldots,c_{p1},\ldots,c_{pn})^t$ for $c_{kj} = \alpha_{ik}\boldsymbol{x}_i{}^t\boldsymbol{x}_j$ and $H = X^tX \otimes A^tA$, where A has i, k element α_{ik}, $X = (\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n)$, and $\otimes$ denotes the direct product. Since X^tX and A^tA are both positive semi-definite, H is also positive semi-definite, so f is a convex function of $\boldsymbol{\beta}$.

PROPOSITION 2. The limit point $\boldsymbol{\beta}^\star$ minimizes f over $\mathcal{B}$.

Proof. If $g : \mathcal{A} \to \mathrm{R}$ is a convex continuously differentiable function and $\mathcal{A}$ is a convex subset of R^q, then g has a global minimum at $\mathbf{y}^\star \in \mathcal{A}$ iff $\nabla g(\mathbf{y}^\star)^t(\mathbf{y}-\mathbf{y}^\star) \geq 0\ \forall\, \mathbf{y} \in \mathcal{A}$. Let $\mathbf{b}_i^\star = (\beta_{i1}^\star,\ldots,\beta_{in}^\star)$ for $i = 1,\ldots,p$ and $f_k(\mathbf{b}) = f(\mathbf{b}_1^\star,\ldots,\mathbf{b}_{k-1}^\star,\mathbf{b},\mathbf{b}_{k+1}^\star,\ldots,\mathbf{b}_p^\star)$.

Then $f_k : \mathcal{B}_1 \rightarrow \mathrm{R}$ is a convex continuously differentiable function and $\mathcal{B}_1 = \{\mathbf{b} \in \mathrm{R}^n : b_i \geq 0 \text{ and } \sum_{i=1}^n b_i = 1\}$ is compact. Since $\mathbf{b}_k^\star$ minimizes f_k over $\mathcal{B}_1$, $\nabla f_k(\mathbf{b}_k^\star)^t(\mathbf{b}-\mathbf{b}_k^\star) \geq 0 \; \forall \; \mathbf{b} \in \mathcal{B}_1$. But $\nabla f(\boldsymbol{\beta}^\star)^t = \left(\nabla f_1(\mathbf{b}_1^\star)^t, \ldots, \nabla f_p(\mathbf{b}_p^\star)^t\right)$, so

$$\nabla f(\boldsymbol{\beta}^\star)^t(\boldsymbol{\beta} - \boldsymbol{\beta}^\star) = \sum_{k=1}^{p} \nabla f_k(\mathbf{b}_k^\star)^t(\mathbf{b}_k - \mathbf{b}_k^\star) \geq 0$$

$\forall \; \boldsymbol{\beta} = (\mathbf{b}_1^t, \ldots, \mathbf{b}_p^t)^t \in \mathcal{B}_1 \times \ldots \times \mathcal{B}_1 = \mathcal{B}$, and since f is convex, this implies that $\boldsymbol{\beta}^\star$ minimizes f over $\mathcal{B}$.

PROPOSITION 3. If A has rank p, then the x-mixtures $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_p$ which minimize (4.1) for fixed α's are unique.

Proof. Let $\boldsymbol{z}^t = (\boldsymbol{z}_1^t, \ldots, \boldsymbol{z}_p^t)$ and treat f as a function of $\boldsymbol{z}$. Then

$$f(\boldsymbol{z}) = \sum_{i=1}^{n} \boldsymbol{x}_i{}^t \boldsymbol{x}_i - 2\mathbf{d}^t\boldsymbol{z} + \boldsymbol{z}^t G \boldsymbol{z}$$

where $\mathbf{d}^t = (\mathbf{d}_1^t, \ldots, \mathbf{d}_p^t)$ for $\mathbf{d}_k = \sum_{i=1}^n \alpha_{ik}\boldsymbol{x}_i$, $G = I_m \otimes A^tA$, and I_m denotes the m by m identity matrix. Since A has rank p, G is positive definite so $f(\boldsymbol{z})$ is strictly convex. Now f is minimized over a convex set, so the constrained problem has a unique minimum.

Propositions 2 and 3 establish the convergence of the inner loop to x-mixtures that minimize (4.1) for fixed α's. Although the β's might not be unique, the corresponding x-mixtures are unique. That the α's minimize (1) for fixed x-mixtures is immediate. These results do not imply that the alternating optimization algorithm invariably converges to the global minimum of (4.1). Numerical experiments suggest

Table 1. Local Minima and Timings

Data Set	n	m	p	Trials until global min	Percentage local min	CPU sec per trial
Masks	200	6	2	1	0	1.5
			3	2	27	2.5
			4	1	51	4.5
			5	5	73	5.8
Pollution	330	9	2	1	0	2.2
			3	1	0	4.0
			4	1	30	8.5
			5	1	45	13.4
Tokamak	40	35	2	1	0	1.4
			3	2	37	1.5
			4	3	73	2.2
			5	3	76	4.1

that convergence to local minima or other stationary points becomes more of a problem as the number of archetypes required increases. For example, in the Swiss Army mask data, in 1000 random starts for computing two archetypes, all converged to the same solution.

But local minima problems occurred in computing 3 or more archetypes. To see the extent of the problem, 500 random starts were used in computing 2, 3, 4, and 5 archetypes for the mask data. This was repeated for the air pollution and Tokamak data discussed in Section 2.

The results are given in Table 1. The fifth column gives the number of trials until the global minimum was first found, the sixth column gives the percentage of times local minima were found in the 500 trials, and the last column gives the average CPU seconds per trial on a SPARC 10 processor.

All examples given in this paper have been validated as global minima through the use of repeated random starts.

6. CONCLUDING REMARKS

Archetype analysis gives a simple and useful way of looking at multivariate data. Archetypes are relatively easy to interpret and the mixture coefficients can provide interesting information about the structure of the data. Past this, the examples speak for themselves.

Sometimes the variables used in the analysis should be standardized before com-

puting archetypes – sometimes not. The Swiss head dimension and air pollution data were standardized, but not the Tokamak data. When to use standardization depends on one's sense about the data.

Since the archetypes are located on the boundary of the convex hull of the data, the procedure can be sensitive to outliers. Robust versions could be developed using convex hull peeling or the outlyingness idea of Donoho and Gasko (1992).

In the more serious examples, we worked with 3 archetypes getting various graphical displays such as the mixture triangles and surface plots. Suppose that more than 3 archetypes are needed for a reasonable approximation to the data. We think that analogous graphical displays can be made using an appropriate mapping of the archetypes to the plane. Some experimentation has been done along these lines with encouraging results, but we leave the issue to future work.

The FORTRAN code for archetype analysis and an interface to S is available from the first author or electronically from adele@sunfs.math.usu.edu. The authors would like to thank Mark Hansen for suggesting the displays in Figures 5 and 12, and Mike Windham for suggestions and discussions relating to this work.

Kurt Riedel was of invaluable assistance in guiding us through the Tokamak maze, and transmitted the data we used. Thanks are also due to Ken Fowler (UCB Physics Department) and Mort Levine (Lawrence Berkeley Lab (ret)) for trying to educate us about Tokamaks.

REFERENCES

Breiman, L. and Friedman, J.H. (1985), "Estimating Optimal Transformations in Multiple Regression and Correlation," *JASA*, V.80, No. 391, 580–619.

Cutler, A. (1993), "A Branch and Bound Algorithm for Convex Least Squares," *Communications in Statistics: Simulation and Computation*, V.22, No. 2, 305–321.

De Leeuw, J. and van der Heijden, P.G.M. (1991), "Reduced Rank Models for Contingency Tables," *Biometrika*, V.78, No. 1, 229–232.

Donoho, D.L. and Gasko, M. (1992), "Breakdown Properties of Location Esti mates Based on Halfspace Depth and Projected Outlyingness," *The Annals of Statistics*, V.20, No. 4, 1803–1827.

Flury, B. (1990), "Principal Points," *Biometrika*, 77, 33–41.

Flury, B. (1993), "Estimation of Principal Points," *Applied Statistics*, V.42, No. 1, 139–151.

Flury, B. and Riedwyl, H. (1988), "Multivariate Statistics, A Practical Approach" London: Chapman and Hall.

Flury, B. and Tarpey, T. (1992), "Representing a Large Collection of Curves: a Case for Principal Points," *Unpublished Manuscript.*

Hastie, T. and Tibshirani, R. (1990), *Genealized Additive Models*, Chapman and Hall.

Hiroe, A. et al. (1988), "Scale Length Study in T.F.T.R.," Princeton Plasma Physics Laboratory Report #2576.

Jones, M.C. and Rice, J.A. (1992), "Displaying the Important Features of Large Collections of Similar Curves," *The American Statistician*, 46, 140–145.

Kardaun, O.J.W.F. Riedel, K.S. et al (1990), "A Statistical Approach to Plasma Profile Analysis," Max-Plank-Institut für Plasmaphysic, IPP 5/35.

Lawson, C.L. and Hanson, R.J. (1974), *Solving Least Squares Problems*, New Jersey: Prentice-Hall.

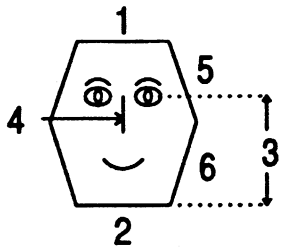
Lazarsfeld, P.F. and Henry, W. (1968), *Latent Structure Analysis*, Boston: Houghton Mifflin.

McCarthy, P.J., Riedel, K.S. et al (1991), "Scalings and Plasma Profile Parameterization of Asdex High Density Ohmic Discharges," *Nuclear Fusion*, V.31, No. 9, 1595–1633.

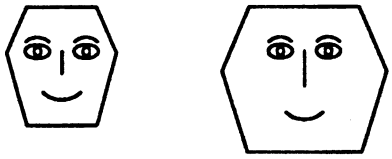
Riedel, K.S. and Imre, K. (1993), "Smoothing Spline Growth Curves with Covariates," *Communications in Statistics: Theory and Methods*, V.22, No. 7, 1795–1818.

van der Heijden, P.G.M., Mooijaart, A. and De Leeuw, J. (1992), "Constrai ned Latent Budget Analysis," *Sociological Methodology*, 22, 279–320.

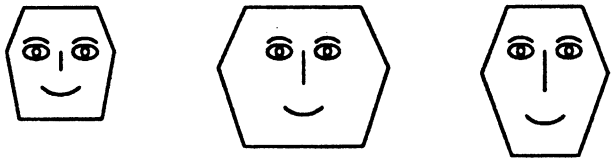
Woodbury, M.A. and Clive, J. (1974), "Clinical Pure types as a Fuzzy Part ition," *Journal of Cybernetics*, V.4, No. 3, 111–121.



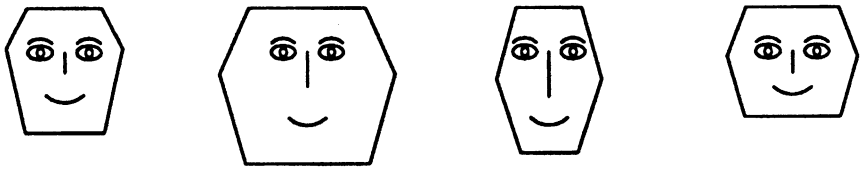
Two Archetypes



Three Archetypes



Four Archetypes



Five Archetypes

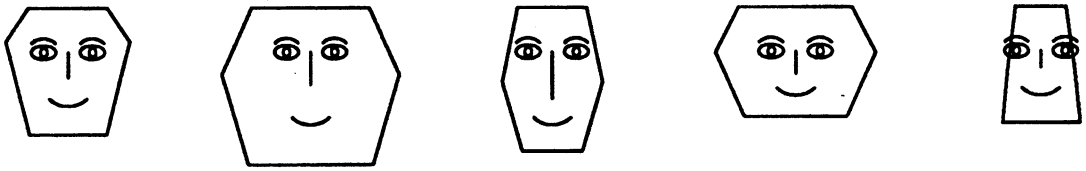


Figure 1. Archetypes for Head Dimension Data.

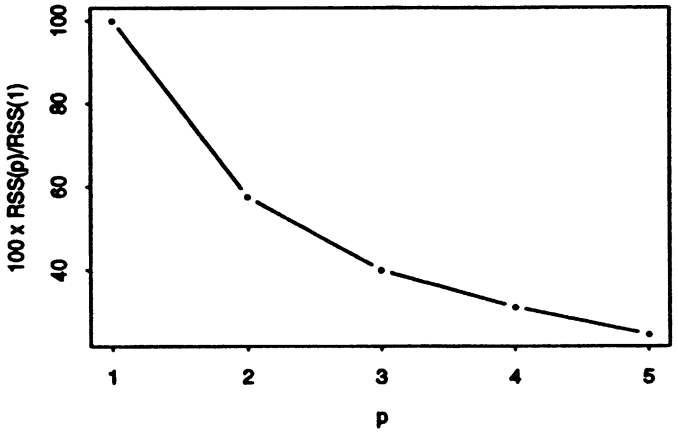


Figure 2. Percent RSS for Swiss Army Archetypes.

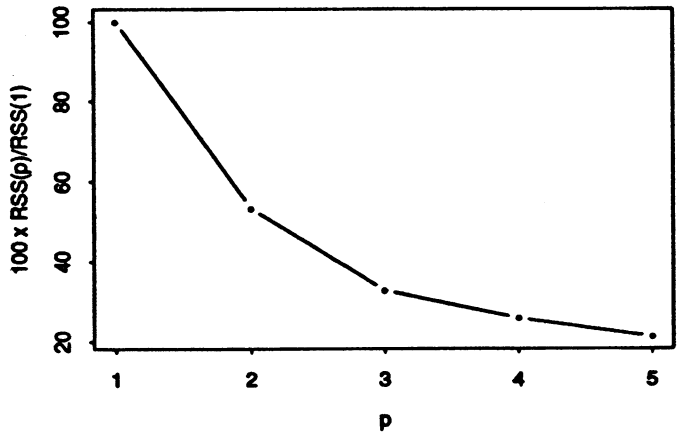


Figure 3. Percent RSS for Air Pollution Archetypes.

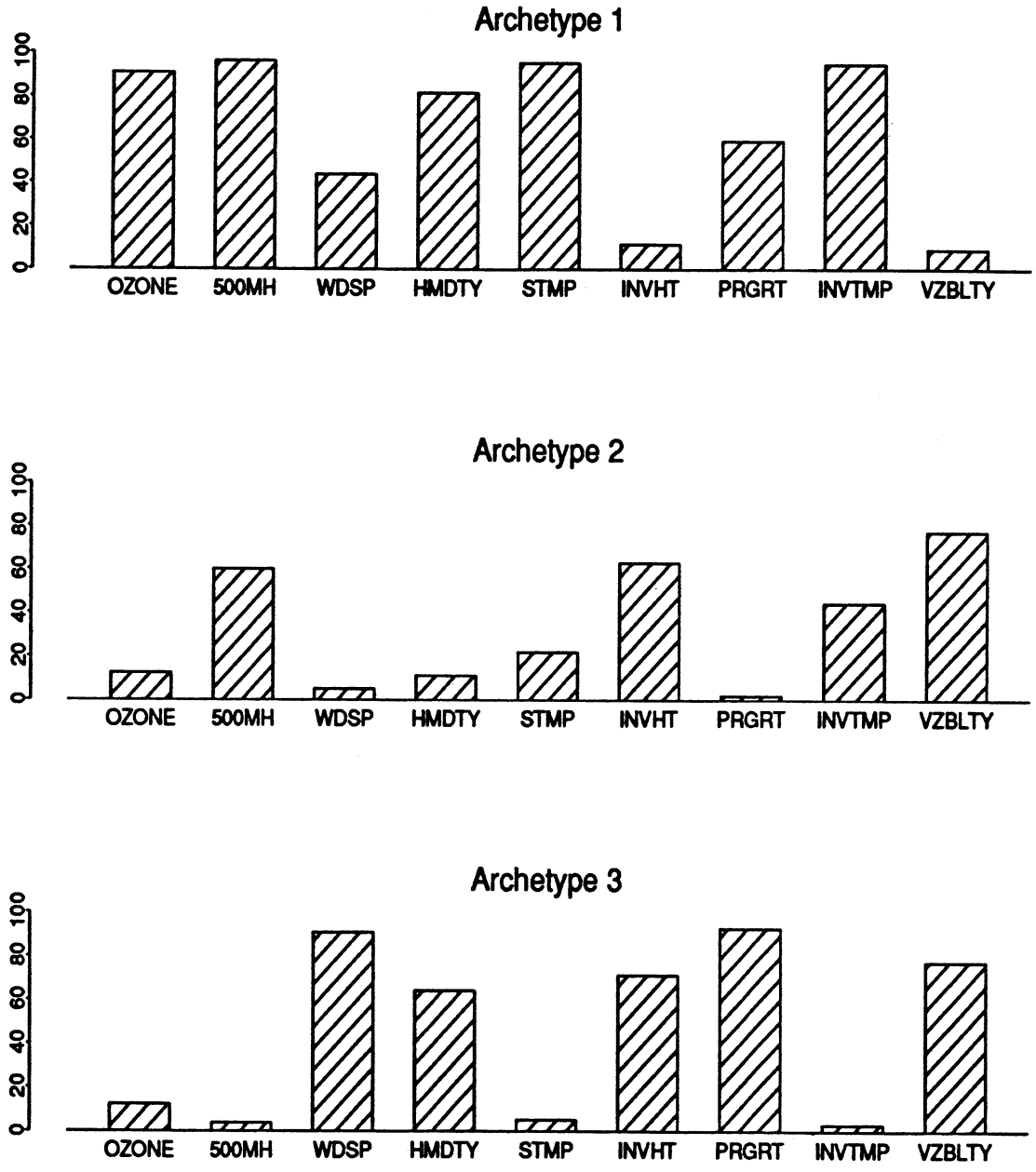


Figure 4. Percentile Profiles of Air Pollution Archetypes.

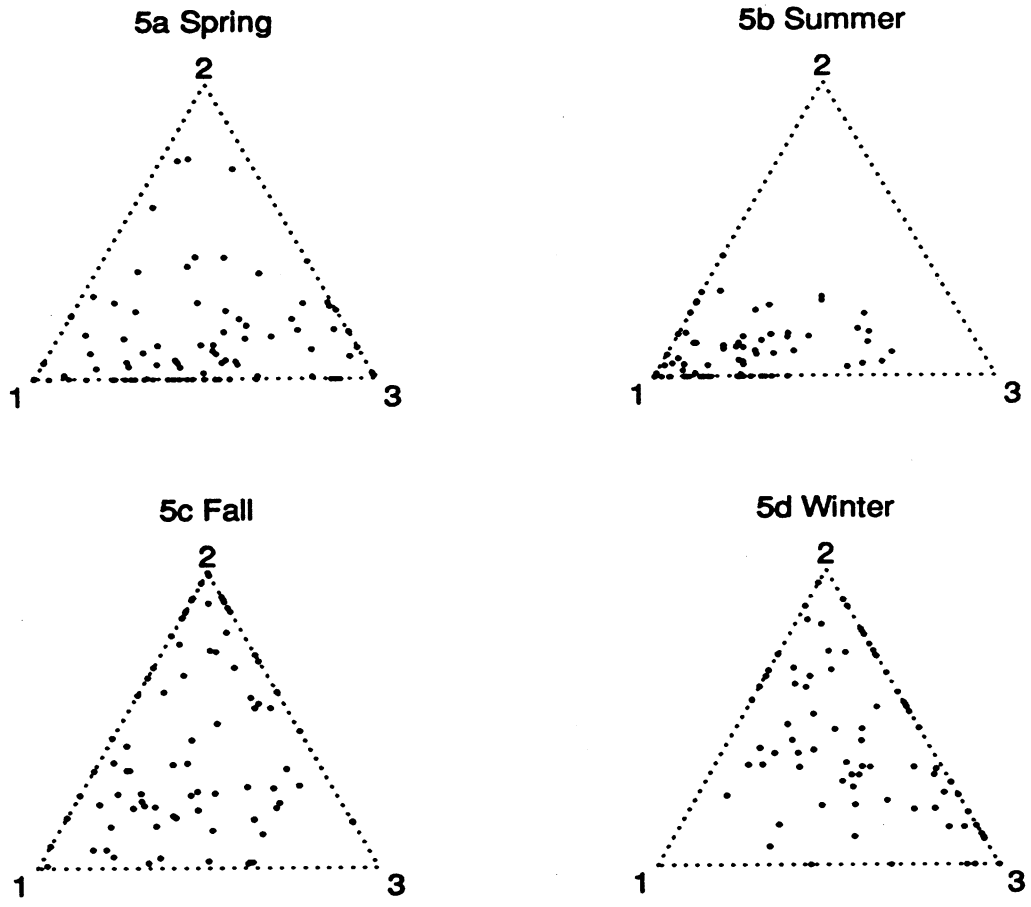


Figure 5. Mixture Plots for Air Pollution Archetypes.

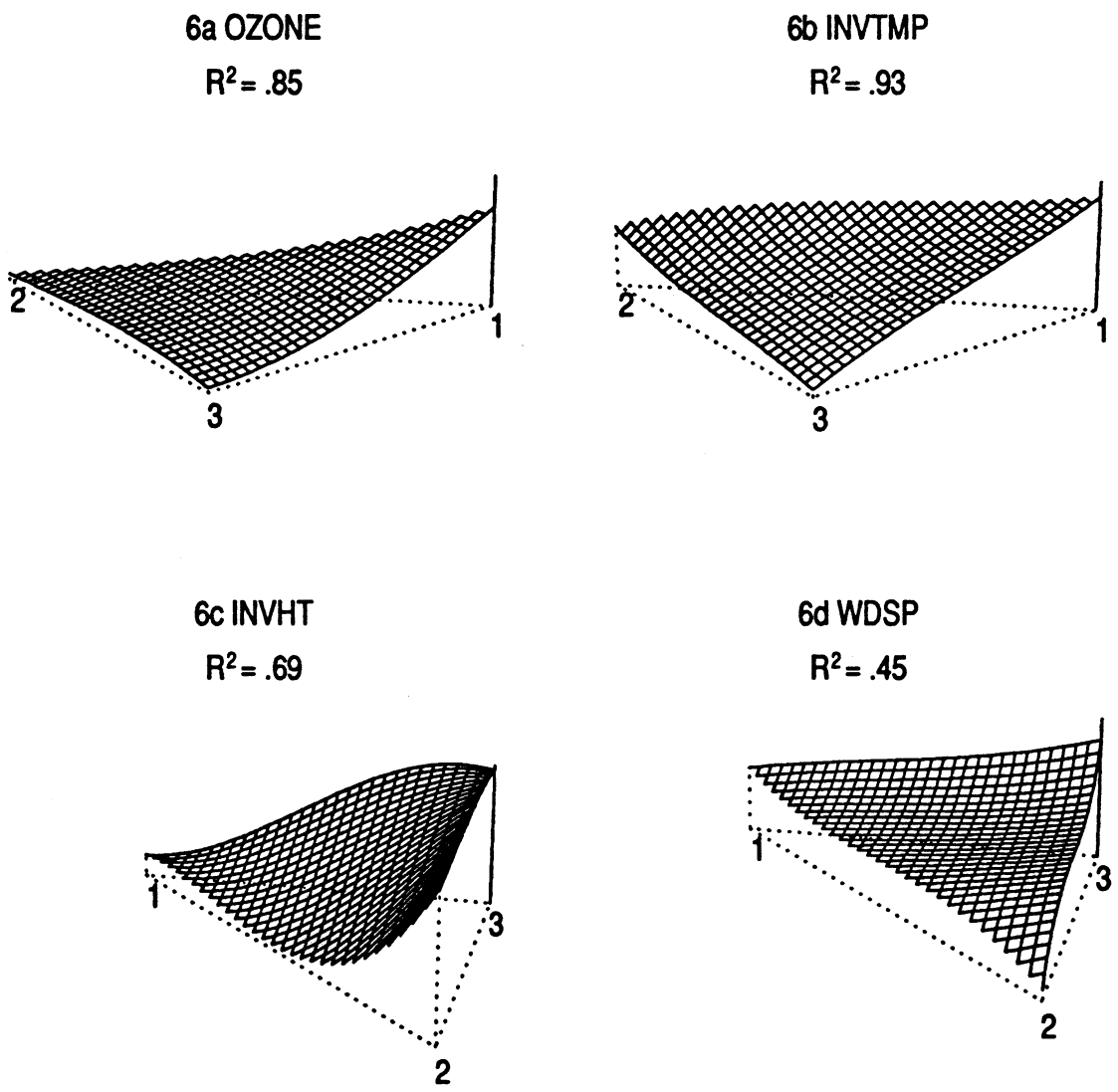


Figure 6. Perspective Plots of Regression Surfaces for Air Pollution Archetypes.

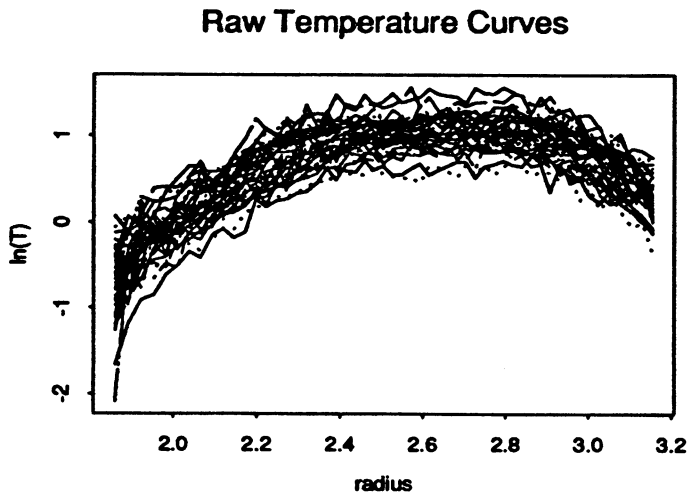


Figure 7. Raw log(Temperature) against radius.

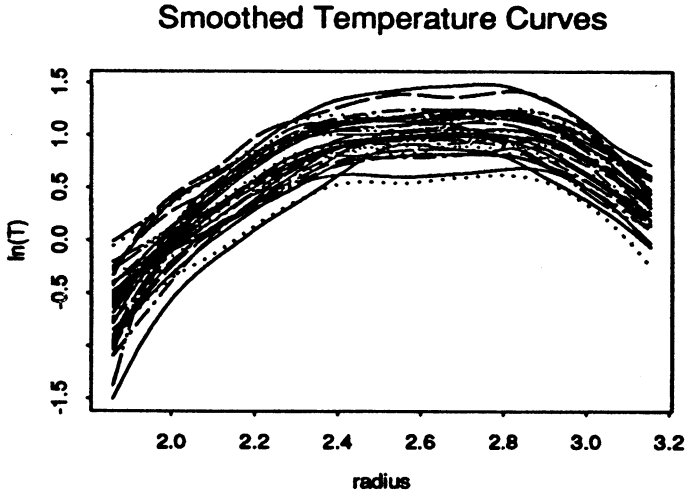


Figure 8. Smoothed log(Temperature) against radius.

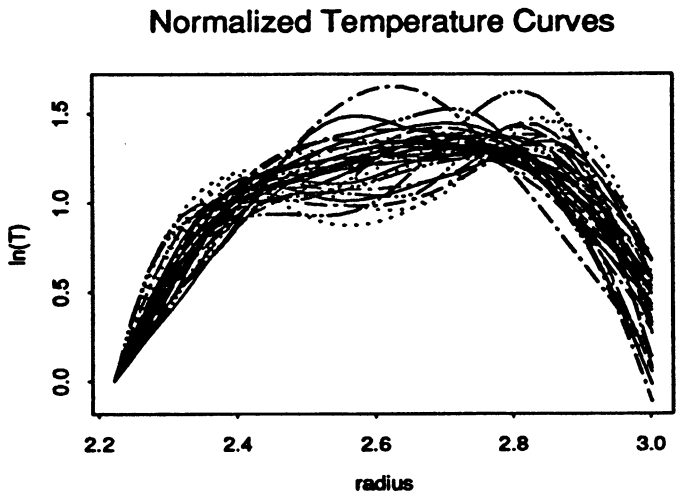


Figure 9. Normalized, smoothed, log(Temperature) against radius.

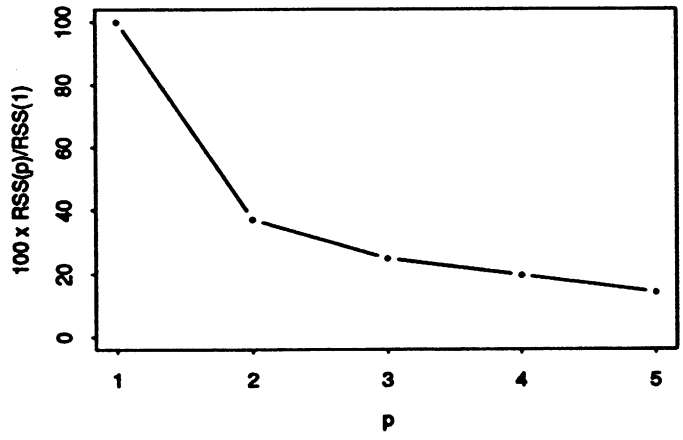


Figure 10. Percent RSS for Tokamak Archetypes.

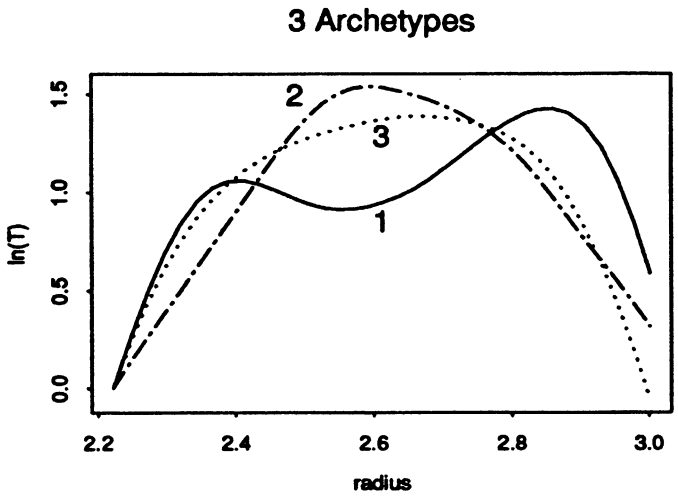


Figure 11. Three Archetypal Curves, Tokamak data.

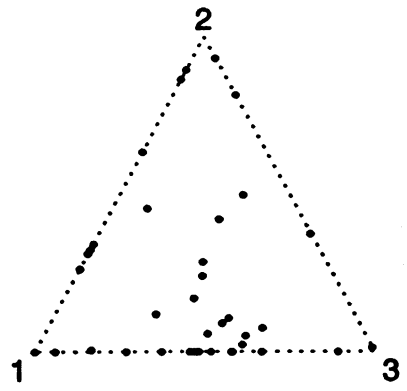


Figure 12. Mixture Plot for Three Tokamak Archetypes.

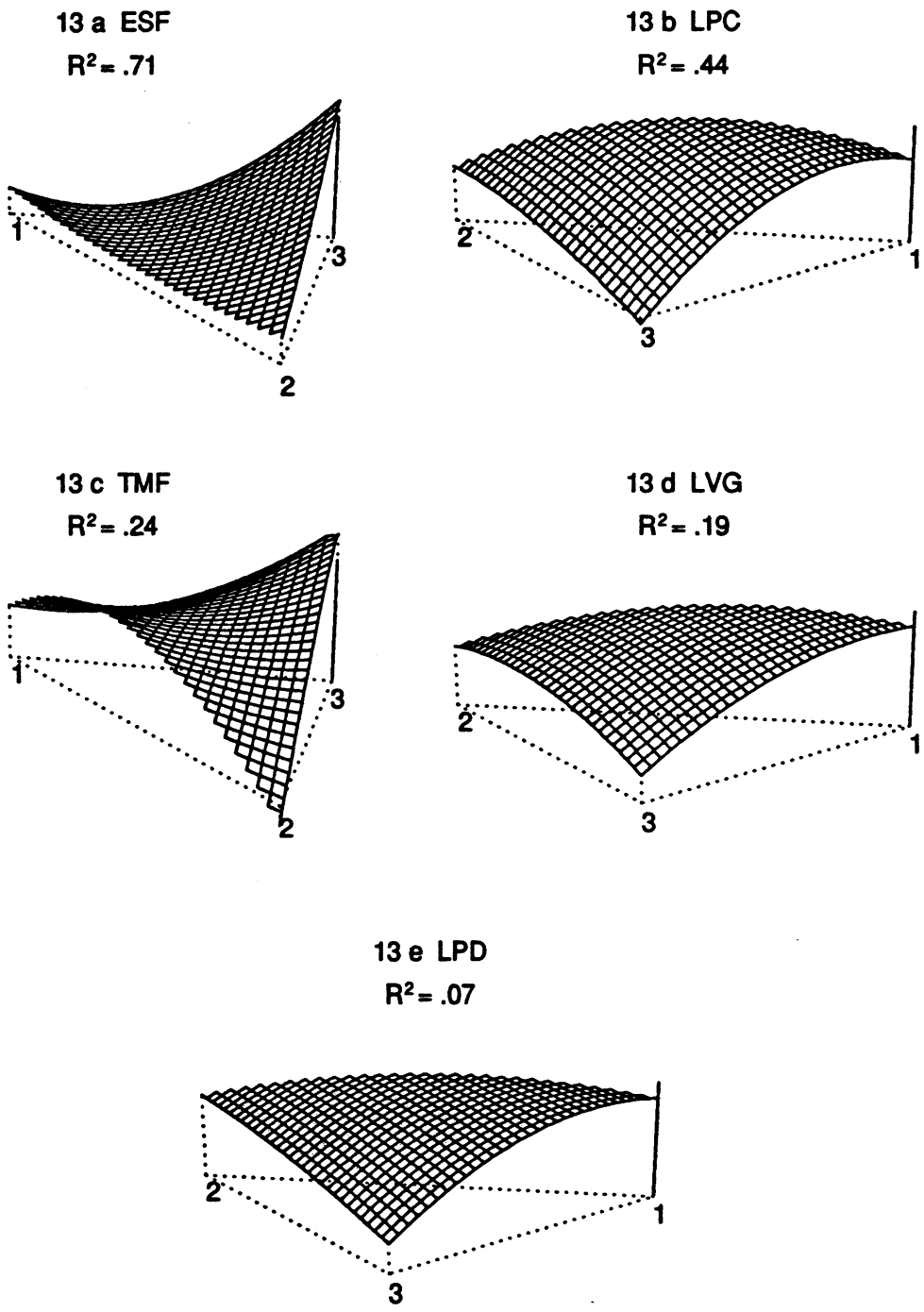


Figure 13. Perspective Plots of Regression Surfaces for Tokamak Archetypes.

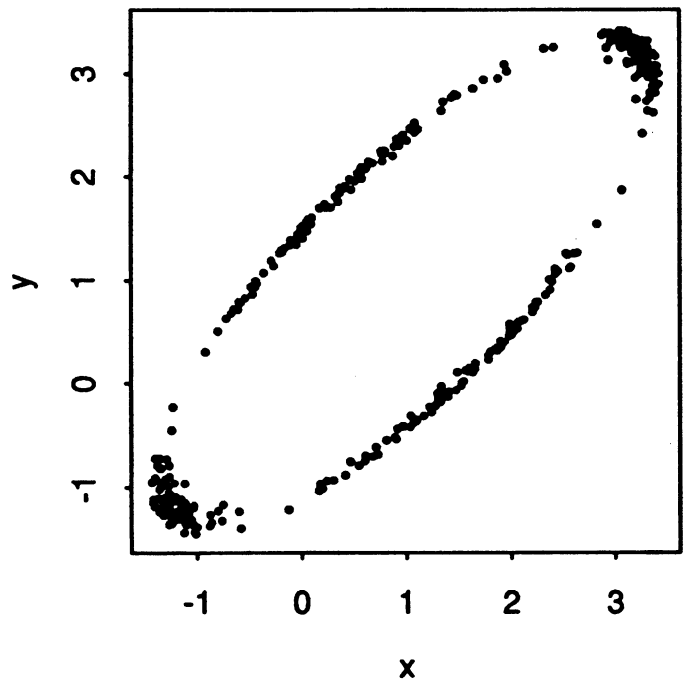


Figure 14. Simulated Bivariate Normal Data, Four Archetypes.