2010-09-10

Application of Convex Methods to Identification of Fuzzy Subpopulations

Ryan Lee Eliason
Brigham Young University - Provo

Application of Convex Methods to Identification of Fuzzy Subpopulations

Ryan L. Eliason

A project submitted to the faculty of
Brigham Young University
in partial fulfillment of the requirements for the degree of

Master of Science

H. Dennis Tolley, Chair
David A. Engler
David G. Whiting

Department of Statistics

Brigham Young University

December 2010

ABSTRACT

Application of Convex Methods to Identification of Fuzzy Subpopulations

Ryan L. Eliason

Department of Statistics

Master of Science

In large observational studies, data are often highly multivariate with many discrete and continuous variables measured on each observational unit. One often derives subpopulations to facilitate analysis. Traditional approaches suggest modeling such subpopulations with a compilation of interaction effects. However, when many interaction effects define each subpopulation, it becomes easier to model membership in a subpopulation rather than numerous interactions. In many cases, subjects are not complete members of a subpopulation but rather partial members of multiple subpopulations. Grade of Membership scores preserve the integrity of this partial membership. By generalizing an analytic chemistry concept related to chromatography-mass spectrometry, we obtain a method that can identify latent subpopulations and corresponding Grade of Membership scores for each observational unit.

Keywords: Grade of Membership scores, archetype, maximum entropy, fuzzy partitioning

ACKNOWLEDGMENTS

My success as a student in completely due to this department's gracious service to me. The professors here have devoted many hours to helping me and many other students around me. I would like to thank the department for everything that they have done for me.

Dr. Tolley has been a great role model and mentor for me. He is truly one of the most brilliant minds of the statistics field. It is an honor to get to research with him. He has taught me how to research and do high quality statistical work. He has been extremely patient and helpful. I would like to thank him for the great opportunity to work on this project. He came up with almost all of the ideas. This project mearly scratches the surface of the topic. With Dr. Tolley's great ideas, many statisticians could spend lifetimes on this work.

I would also like to thank Drew Johnson, a BYU Math student who generously helped me understand matrix algebra related to Simplicies.

I would like to thank my parents for their tremendous service. Thanks to them, I was raised with good morals and a strong desire to work hard. They have also given me numerous rides to school and cooked me great meals. They are the best parents any one could hope for.

CONTENTS

CHAPTER 1

INTRODUCTION

In highly multivariate observational studies, one often facilitates analysis through classical approaches, such as a compilation of interaction effects. The use of interactions comes from experimental design where the factors can have antagonistic or synergistic effects that add to the main effect factors. With a large number of covariates, it is often challenging to identify which interaction effects are necessary to characterize the population. In addition, the potentially high number of interaction effects often uses many degrees of freedom, which is a luxury in a highly multivariate setting. Koch et al. (1977) propose an alternative approach based on the idea of subpopulations. Based on the belief that subjects are not identically distributed from a single distribution, but rather are from a larger mixture distribution, subpopulations become valuable in partitioning the variance associated with a response variable. Koch et al. (1977) claim the interactions between observed factor variables may well be indicative of latent subpopulations and not simply the effect resembling an interaction between two or more design variables. In this case, these latent subpopulations are not apparent in the data but are only implicitly characterized by the patterns of response in the observational variables. Interaction effects collectively characterize each subpopulation. Thus, identifying latent subpopulations seems more appropriate than modeling numerous interactions.

Traditionally, if observational units come from one of the particular subpopulations, then a sample of such individuals is viewed as being from a mixture distribution comprised of each subpopulation distribution. In this case, each individual is from one explicit subpopulation. The probabilities of the mixture represent the probability that an individual is a member of each specific subpopulation. In many situations, subjects themselves can be

considered mixtures of multiple subpopulations. Such individuals are only vague or fuzzy members of each subpopulation. In this sense, each observational unit can often be viewed as a composition distribution, having a set of coefficients denoting the strength of their membership in each subpopulation.

As an example, consider an informed citizen voting in an election. They most likely do not agree with one party on every issue. Instead, they might agree with one party on 40% of the issues and the other party on 60% of the issues. The key here is that the two theoretical parties are fixed entities that most citizens do not perfectly agree with. Each citizen agrees partially with both parties. A political survey would then identify the voter as a partial member of each party. The survey might give a score to the citizen denoting their placement on a spectrum relating the parties. One might refer to this value as a Grade of Membership (GoM) score because it denotes the strength of membership that the subject has in a particular subpopulation.

This situation is very similar to that of chemical analysis using, say chromatographic separation coupled with mass spectrometry. Such analyses are common in analytic chemistry. The technique characterizes the compounds that compose a chemical composition. Consider that there exists a number of compounds contained in the chemical composition. When analyzing such a composition, samples are taken over time. Only a few compounds are represented at each particular time. This is done by using a technique referred to generally as chromatography. Chromatography separates or partially separates compounds over time. The sample observed at a particular time can be viewed as a stratified observation of the chemical composition. The relative expression rate of compounds governing the sample changes across time. In other words, the sample can be viewed as a composition of the pure compounds. Each sample can be accounted for fully by the set of pure components. These pure compounds are sometimes referred to as pure variables. Mass spectrometry tools can identify discriminant scores that denote the relative presence of a particular pure variable

for a particular sample. In the statistics field, these discriminant scores can be thought of as GoM scores.

If we want to define GoM scores for our highly multivariate observational study, we can obtain a new set of analysis tools based on the methods that have been used in conjunction with chromatography-mass spectrometry. To illustrate this, consider each observational unit as being a fuzzy composition of subpopulations. The difference between the chromatography-mass spectrometry approaches and the statistics analysis theory is that in chemistry, all compounds which are used in a composition have already been characterized. If a sample at a particular time is found to be 90% salt, then salt is certainly one of the compounds that governs the composition. In statistics, however, we do not enter the analysis knowing what the subpopulations are. Thus, we wish to determine the set of latent subpopulations that best characterize the sample, and ultimately the population. In addition, we want to know the strength of membership for each observational unit with regard to each latent subpopulation. By generalizing the methods used in mass spectrometry analysis, we can define subpopulations and corresponding GoM scores for each observational unit.

The information partition function (IPF), originally derived by Oliphant (2003) and Engler (2002), is an alternative method to define subpopulations and create GoM scores for a sample. The IPF can be derived based on the second law of thermodynamics (Cannon 2008). The fundamental assumption is that highly informative GoM scores can be obtained by maximizing the entropy that exists in a system. The IPF is an iterative solution to obtain GoM scores under this assumption. The IPF exhibits a great deal of variability due to the algorithm's initialization process. Our methodology stands alone, but it can also be used to initialize the IPF.

1.1 Outline

We first give a background on the current relevant work. We then introduce a novel approach to obtain GoM scores, founded on the chromatography-mass spectrometry process. We will

outline this methodology in great detail. Then, using our methodology, we analyze an Introdutory Statistics Course data set. Finally, we use cross-validation to show usefulness of the GoM scores.

CHAPTER 2

LITERATURE REVIEW

2.1 Overview

In the following sections, we present brief backgrounds on Grade of Membership methodologies, gas chromatography-mass spectrometry techniques, interaction modeling, and the information partition function.

2.2 Grade of Membership

When fitting a Grade of Membership model, we may view the vector of measurements on each individual as responses to a questionnaire. For this we need to set up a framework of variables and indexers to work with. Following Oliphant (2003), we define the following variables and indexers:

$$
\begin{aligned}
i &= \text{index on observational units,} \\
j &= \text{index on questions,} \\
l &= \text{index on possible responses to question } j, \\
k &= \text{index on pure variables,} \\
y_{ijl} &= \begin{cases} 1 \text{ if observational unit } i \text{ chose response } l \text{ on question } j \\ 0 \text{ otherwise,} \end{cases} \\
g_{ik} &= \text{Grade of Membership score for unit } i \text{ with regard to pure variable } k, \text{ and} \\
\lambda_{kjl} &= \text{probability of response } l \text{ on question } j \text{ for pure type } k.
\end{aligned}
$$

With this framework, we find the estimates of the $\mathbf{g}_{ik}s$ and $\boldsymbol{\lambda}_{kjl}s$ that maximize the joint likelihood, following Oliphant (2003), as

$$L = \prod_i \prod_j \prod_l \left(\sum_k g_{ik} \lambda_{kjl} \right)^{y_{ijl}}, \tag{2.1}$$

where each of the $\mathbf{g}_{ik}s$ and $\boldsymbol{\lambda}_{kjl}s$ are nonnegative and sum to unity over k and l respectively.

A key issue in determining the maximum likelihood estimate of $\mathbf{g}_{ik}$ and $\boldsymbol{\lambda}_{kjl}$ in 2.1 is initiation of the algorithm. Since there is often a constant, c, such that $\mathbf{g}_{ik}c$ and $c^{-1}\boldsymbol{\lambda}_{kjl}$ satisfy the constraints on $\mathbf{g}_{ik}$ and $\boldsymbol{\lambda}_{kjl}$, these parameters have an identifiability problem. However, initial starting values can often resolve such a problem. In this case, the initial starting values should be informed. The algorithm we develop provides informed initial values for likelihood estimation.

2.3 Mass Spectrometry Characterization

The methodology described below is based on a chemistry technique presented by Grande and Manne (1999). In determining the composition of a chemical sample, the relative amounts of each compound in the sample is of primary interest. Therefore the problem can be standardized as one in which the proportion of each compound present in a sample is of interest. Grande and Manne (1999) develop an algorithm to create a simplex surrounding the data with endpoints denoting pure compounds. This convex representation makes it possible to methodically consider chemical compounds as an n-way mixture distribution.

The determination of composition follows two steps: the first is to separate the sample into subsamples where each consists of one, or a few, of the compounds from the sample. The method of choice entails chromatography (Giddings 1965). In this case, the different subsets are separated vertically in time. Thus, the first compound is separated before the second. The remaining compounds are separated in time order as well. The order of the separation is based on a physical property called retention time. Unfortunately, the separation is rarely crisp. In other words, pollution from the second and third compounds contaminate the first

compound's separation. In this case, the earliest samples of the first compound are more pure. Toward the end of the time in which the first compound is separating, the second and third get mixed into the sample. Similarly, with other compounds separated in time, contamination from neighboring compounds enters the analysis.

The second step is to identify the molecules separated using mass spectrometry. This step consists of scans of the sample over time and results in a highly multivariate response profile from each scan. When contamination occurs, these scans will show response profiles for a pure sample initially and then a less pure sample over time until the first compound is removed and the second compound is brought into focus. Soon the second compound will be brought into complete focus, denoting a second pure compound. The process continues until a number of pure compounds have been identified. To visualize this process, consider Figure 2.1. Figure 2.1 is suggestive of distinct pure compounds that govern the chemical mixture.

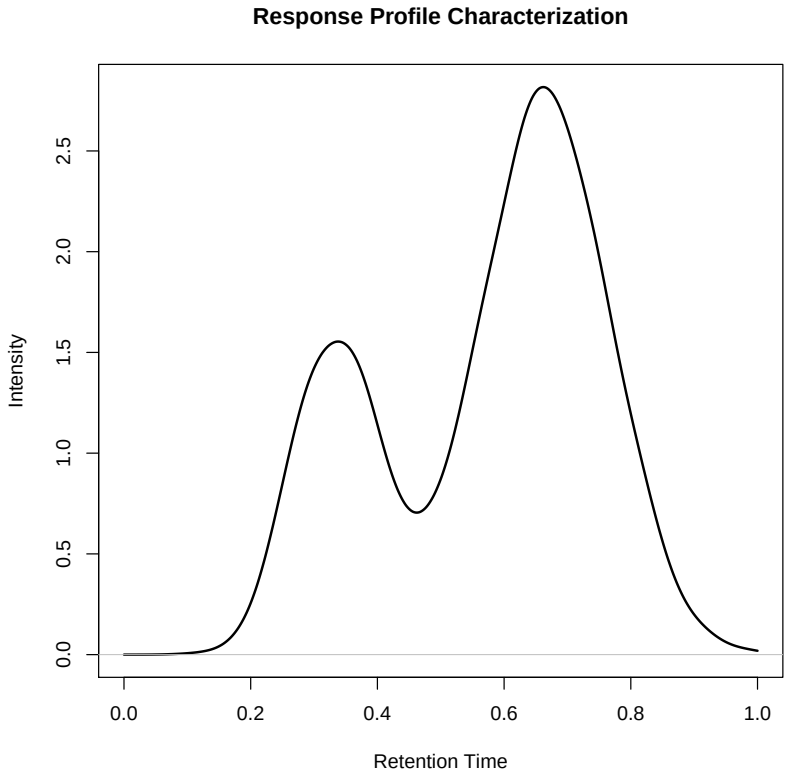

Figure 2.1: Response Profile Characterization. Ion ratio intensity changes across retention time.

Separating across retention time, we can get an informative view of the process. This separation is illustrated further in Figure 2.2. A particular snapshot, say retention time equals 1.5, is the fingerprint of a particular pure compound. Other snapshots may suggest a convolution between two or more pure compounds. As retention time increases, the relative ion ratio composition changes dramatically. In fact, it actually moves from one pure compound to another. Snapshots taken between pure compounds are thought of as "polluted." This means the they are some linear combination of multiple pure compounds.

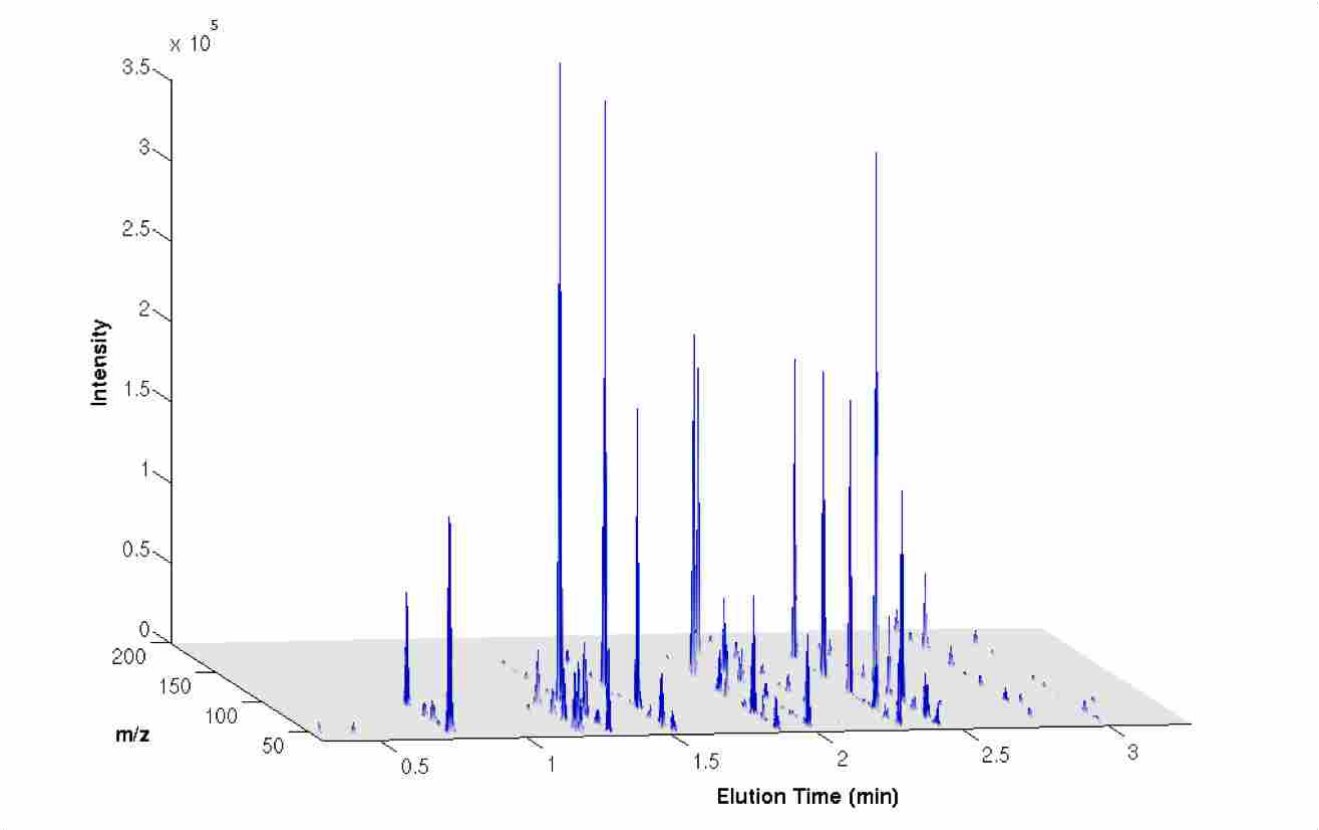

Figure 2.2: Full Data. By separating the ion ratios (m/z) across retention (elution) time, we can see unique compositional elements of the mixture.

The approach presented by Grande and Manne (1999) is novel in that it considers this ion ratio data in terms of a time series based on retention time within a simplex. By normalizing the data, Grande and Manne (1999) consider the relative intensities of each ion ratio at each retention time. This approach can be seen in Figure 2.3. Grande and Manne (1999) plot the results of a principal components analysis on the normalized ion ratio data. This plot illustrates that the profile of ion ratios has a particular pattern which is evident by introducing retention time. When the composition sharply changes from one time to another, the composition of ion ratios is sharply changing. This is evidence that the endpoints of the simplex, where the sharp changes occur, are pure compounds which compose the mixture. Samples taken at retention times between these pure compounds are referred to as polluted or contaminated. They are linear combinations of the pure compounds with the weights

denoting the relative similarity to each particular compound. In other words, they are a composition of the pure compounds mixed together.

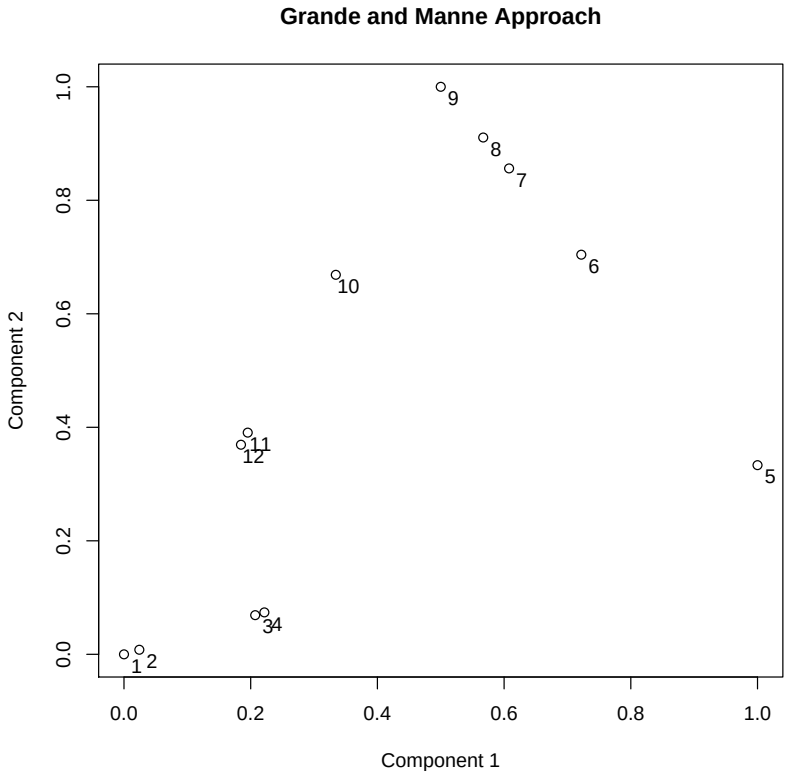

Figure 2.3: Grande and Manne Approach. Samples 1, 5, and 9 are the pure compounds that govern this particular chemical mixture.

2.4 Interaction Modeling

The most recent substantial advancements in the partitioning of interaction effects come from work dating over 30 years ago. Cannon (2008) notes, in regard to classical analysis of variance, that most major developments occurred no later than the 1970s. Interaction modeling stems from the analysis of variance partitioning work by Mandel (1969). His work points out that particular portions of the interaction signal are estimated best by the eigenvalues of $\mathbf{r}'\mathbf{r}$, where $\mathbf{r}$ is the vector of residuals from the analysis of variance model. Gollob (1968) suggests partitioning the interaction signal via principal components analysis.

The higher order categorical interaction theory is based on work by Koch et al. (1977). As mentioned above, Koch et al. (1977) claim that modeling membership in particular subpopulations may have more merit than classical approaches. The classical approach to dealing with subpopulations, as mentioned above, is to opt for a compilation of interaction effects. Under this paradigm, each subject's corresponding subpopulation membership may be a replacement for the predictor variables that would typically be used.

2.5 IPF Background

The IPF is the product of compositional data analysis theory from many sources. Early work by Tolley and Manton (1992) analyze properties of discrete Grade of Membership models and develop the theoretical basis for the IPF. Later work by Engler (2002) and Oliphant (2003) refine the IPF.

To define the IPF likelihood, Oliphant (2003) uses the same variables and indexers as before, but changes the interpretation of λ_{kjl}: the LaGrange multiplier for constraint on question j for pure type k. Note that λ_{kjl} is related to, but not equal to, the probability of response l on question j for pure type k. With this framework, Oliphant (2003) gives the joint likelihood as

$$L = \prod_i \prod_j \prod_l \exp\left(-\sum_k g_{ik}\lambda_{kjl}\right)^{y_{ijl}}, \tag{2.2}$$

where $\mathbf{g}_{ik}s$ meet the nonnegativity and unity constraints.

In addition, Oliphant (2003) maximizes the likelihood subject to the constraint that

$$\sum_l \exp\left(-\sum_k g_{ik}\lambda_{kjl}\right) = 1. \tag{2.3}$$

The key to solving the likelihood for IPF modeling is getting reasonable starting values. One application of the methodology presented here is to produce such initial starting values.

CHAPTER 3

METHODS

Similar to determining the purity of a chemical sample, compositional data analysis only focuses on relative amounts of measurements on a certain subpopulation. We can then standardize the interaction analysis as one in which the proportion of each subpopulation present in a sample is of interest. Generalizing the algorithm by Grande and Manne (1999) allows us to put the purest possible subpopulation subjects as endpoints of a simplex containing the data. This makes it possible to consider subjects as coming from an n-way mixture distribution. We refer to extreme points in the polyhedral representation of the data as "pure variables."

In chemistry, the pure variables are easy to visualize as the true components that comprised the mixture. This concept does not easily translate to a statistical mindset. In statistics, the pure variables are hypothetical subjects composing endpoints for a simplex that contains all or most of the data set subjects. Additionally, the hypothetical subjects characterize a pure type. A pure type is a stereotype, similar to a subpopulation, which characterizes subjects that are highly similar to a particular pure variable. Similarity is measured by the strength of the corresponding GoM score. Examining the likely response profiles of these subjects implicitly defines what a pure variable looks like. The algorithm developed by Grande and Manne (1999) only allows existing data subjects to be chosen as pure variables. In general, theory suggests that a subject not in the data set could best summarize a pure selection of a subpopulation.

3.1 Algorithm Overview

The first step of the algorithm is to standardize measurements (which are potentially taken from different spaces or scales). This will ensure that each measurement is given a compa-

rable influence on the algorithm. Next, we project normalized subjects into a low number of dimensions. In the lower dimensional space, the pure variables are determined methodically. Once the pure variables are obtained, we find each subject to be a convex combination of the pure variables. This new composition distribution can be taken as a consistent set of initial $\mathbf{g}_{ik}s$ to be used in the IPF.

3.2 Preliminaries

We use the following notation:

N is the number of subjects.

M is the number of measurements taken on each subject.

$\mathbf{D}$ is a matrix of the data with subjects as columns.

$\mathbf{S}_m$ is the standard deviation of the m^{th} measurement across the N subjects.

$\mathbf{D}_m$ is the m^{th} row vector of $\mathbf{D}$.

$\mathbf{X}_m$ is the m^{th} row vector of $\mathbf{X}$.

$\mathbf{X}_m$ is $\frac{\mathbf{D}_m}{\mathbf{S}_m}$.

$\mathbf{X}$ is the standardized matrix of the data with rows $\mathbf{X}_m$.

$\mathbf{p}$ is a normalization column vector of length M.

K is the rank of the simplex we will construct where $K < M$. This implicitly assumes that a simplicial subspace exists such that the barycentric coordinates are the coefficients of the convex representation of the composition. This barycentric representation is designed to explain most of the variability of the observed M dimensional data vectors, $\mathbf{X}_n$.

3.3 Supervised Variable Selection

We have the option to use supervised variable selection before starting algorithm. If we exercise this option, then only the selected variables will be used in the algorithm. In addition to this option, we can also incorporate supervision through our normalization process which we describe in the next section. In order to select a supervised option, we require a set of

training data where the true state of each composition sample is known. This state can be considered as a response variable.

One supervised variable selection approach is to use the classic decision tree algorithm developed by Breiman et al. (1984). This is a good approach in situations where measurements have nonlinear effects and interactions exist between measurements. The general approach of growing a Breiman tree is to recursively split the sample based on a single covariate that separates the average response of the resulting subsamples. The problem with fitting a tree, however, is that it often overfits the particular sample (Breiman 2001).

To remove the bias, we have selected to use Random Forests (Breiman 2001). Random Forests provide a theoretically unbiased variable selection methodology because criteria are aggregated over a bootstrap generated ensemble of trees. This ensemble will not "overfit" a particular sample. Our approach is to select the variables which are used in $\boldsymbol{\xi}\%$ or more of the bootstrap generated trees. Note here that $\boldsymbol{\xi}$ is a tuning parameter which could be set to, say 50%.

3.4 Normalization

To normalize a data set, we divide each entry by the dot product of that column vector and a normalizing vector $\mathbf{p}$. The normalized data set, $\mathbf{Y}$, is given column-wise by

$$\mathbf{Y}_n = \frac{\mathbf{X}_n}{\mathbf{p} \cdot \mathbf{X}_n}, \tag{3.1}$$

where $\mathbf{Y}_n$ is the n^{th} column of $\mathbf{Y}$, $\mathbf{X}_n$ is the n^{th} column of $\mathbf{X}$, and $\mathbf{p}$ is the normalization vector.

Cluster Based Singular Vectors

The choice of a proper normalization vector is vital because a slight change in $\mathbf{p}$ results in more weight on a certain measurement. A common selection for $\mathbf{p}$ is the first column of the

$\mathbf{U}$ matrix, found from a singular value decomposition of $\mathbf{X}$ (Grande and Manne 1999). The singular value decomposition is formed as $\mathbf{X} = \mathbf{UDV}'$, with $\mathbf{U}$ and $\mathbf{V}$ orthonormal and $\mathbf{D}$ as a diagonal matrix with singular values $\mathbf{d}_k \geq 0$.

We have found that a more stable approach is to use the singular value decomposition of a centroid dataset based on $\mathbf{X}$. An algorithm to determine a clustered normalization vector is given below.

(1) Using a supervised or unsupervised clustering algorithm, create clustered subject groups (approximately 1 cluster for each 10 subjects).

(2) Find the singular value decomposition on the matrix of centroids (cluster means) instead of the original $\mathbf{X}$.

We denote the cluster based $\mathbf{U}$, from step 2, as $\mathbf{C}$ and the normalization vector as $\mathbf{C}_1$, the first column of $\mathbf{C}$.

3.5 Barycentric Coordinates

Since we are interested in determining GoM scores, which are closely related to discriminant scores, empty dimensions in $\mathbf{Y}$ may corrupt the integrity of discrimination. In other words, our GoM scores are less informative if based on empty dimensions. To correct for this possibility, we give the option of a data reduction, resulting in $\boldsymbol{\alpha}$ values which denote Barycentric coordinates relative to a simplex surrounding the data. Determining the $\boldsymbol{\alpha}$ values can be viewed as an intermediate step to obtaining the GoM scores. Define r as an arbitrary passed parameter denoting the number of dimensions that the data can fill completely. r should be set greater than K. To defend this, assume that the data can adequately fill only r dimensions. Then, it would be a poor assumption to conclude that more than r subpopulations exist because the data has no more than r significant pieces of information that could possibly discriminate subpopulations.

The $\boldsymbol{\alpha}$ values are given by

$$\boldsymbol{\alpha} = \mathbf{T}'\mathbf{Y}, \tag{3.2}$$

where $\mathbf{T}$ denotes the second through r^{th} columns of $\mathbf{C}$.

If the process is to be completely automated, then r can be set as K plus 3. As will be discussed in the next section, the number of pure variables, K, can be latently estimated through the Broken-Stick method. Setting r as K plus 3 allows for the perhaps valuable information from the singular vectors below the Broken-Stick cut line to still have an influence on the GoM scores.

3.6 Pure Variables

Defining the number of Pure Variables

A simplex is defined by endpoints. Following the spirit of the chemistry concepts discussed previously, we consider these endpoints as being pure variables or archetypes. The number of pure variables, K, can be estimated using the Broken-Stick method (Jackson 1993). Randomly generated data would yield correlation matrices whose squared singular values would follow a Broken-Stick distribution. Thus, we can compare our observed squared singular values, based on the variable correlation matrix, to the Broken-Stick distribution in order to estimate the number of values that are significantly higher than those which random data would yield. The Broken-Stick expected values are given by

$$\mathbf{E}_i = \sum_{k=i}^{q} \frac{1}{k}, \tag{3.3}$$

where q is the number of measurements taken on each subject. These expected values, given in a vector as $\mathbf{E}$, can be compared to the observed squared singular values in order to determine a reasonable number of pure types. We estimate the optimal number of pure

variables to be the number of positive entries in the row vector $\mathbf{d}^2 - \mathbf{E}$, where $\mathbf{d}$ is a vector of the singular values from the singular value decomposition on $\mathbf{X}$. To illustrate this point, consider Figure 3.1. In Figure 3.1, we observe a crossing point between observed squared singular values, from the Ionosphere data which we discuss below, and the values that would have been observed if data were randomly generated in an uncorrelated fashion. It is worth noting that the sum of the observed squared singular values equals the sum of the expected values. Thus, there is a guaranteed solution for any nontrivial data.

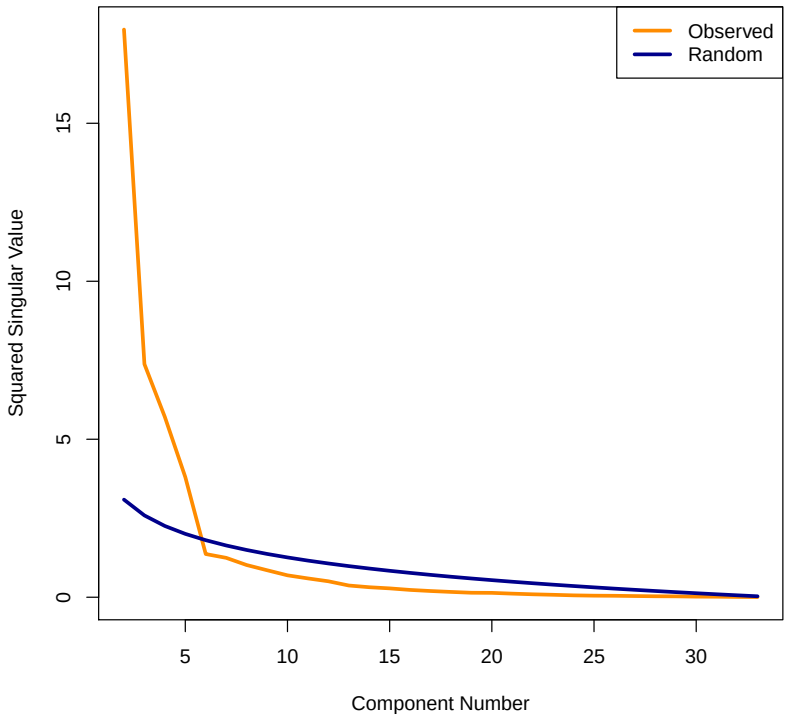

Figure 3.1: Broken-Stick Method. The observed squared singular values are higher than the expected values until component six. This implies that five is a reasonable value of K.

Weights

In $\boldsymbol{\alpha}$ space, distance from one subject to another subject can be viewed as a discriminant score denoting how similar the two subjects are. However, typical Euclidean distance is

inappropriate because the $\boldsymbol{\alpha}$ space is created from a number of singular vectors which do not have equal singular values. To correct for this, any distance metric needs to incorporate weights given by the singular values. To form a set of weights, we create the row vector $\mathbf{w}$ given by

$$\mathbf{w}_i = \frac{d_i}{\sum_{i=1}^{r} d_i}, \tag{3.4}$$

where d_i is the i^{th} diagonal entry of the singular value matrix previously obtained from the data.

Pure Variable 1

An optimal point for the first pure archetype is one far away from the center of the $\boldsymbol{\alpha}$'s, the origin. To determine the first pure variable, find the subject index k_1 that corresponds to the largest $\mathbf{w}\boldsymbol{\alpha}^2$ entry, where $\boldsymbol{\alpha}^2$ denotes squaring each entry in $\boldsymbol{\alpha}$. The first pure variable is given by $\boldsymbol{\lambda}_1 = \boldsymbol{\alpha}_{k_1}$.

Pure Variable 2

The second pure variable is the point farthest away from the first pure variable. To determine which point this is, we create a transformed $\boldsymbol{\alpha}$ space: $\boldsymbol{\alpha}^{(2)}$. This is following notational conventions developed by Grande and Manne (1999). Note that (2) denotes $\boldsymbol{\alpha}$ space 2 and not $\boldsymbol{\alpha}^2$. To create $\boldsymbol{\alpha}^{(2)}$, center $\boldsymbol{\alpha}$ about the first pure variable and find the distances of the other $\boldsymbol{\alpha}$ points from it. $\boldsymbol{\alpha}^{(2)}$ is given by

$$\boldsymbol{\alpha}^{(2)} = \boldsymbol{\alpha} - \boldsymbol{\alpha}_{k_1}\mathbf{j}, \tag{3.5}$$

where $\mathbf{j}$ is an N length row vector with 1 in each entry.

The second pure variable, $\boldsymbol{\lambda}_{k_2}$, is the column of $\boldsymbol{\alpha}$ corresponding to the largest entry of the vector $\mathbf{w}(\boldsymbol{\alpha}^{(2)})^2$. Note that $\boldsymbol{\lambda}_{k_2}$ was determined in a space different from $\boldsymbol{\lambda}_{k_1}$. The index, k_2, was determined in $\boldsymbol{\alpha}^{(2)}$ space, but $\boldsymbol{\lambda}_{k_2}$ is a column vector of $\boldsymbol{\alpha}$, not $\boldsymbol{\alpha}^{(2)}$. The pure variables are of ultimate use in the original space, $\boldsymbol{\alpha}$, but we use them temporarily in the uninterpreted space as a means to obtain future pure variables.

Pure Variable 3

The third pure variable is the column of $\boldsymbol{\alpha}$ which is orthogonally farthest away from the first two pure variables. To obtain this variable, we Gram-Schmidt orthogonalize $\boldsymbol{\alpha}^{(2)}$ to the first 2 pure variables. The projection matrix, $\mathbf{M}$, for the orthogonalization is defined as

$$\mathbf{M} = \mathbf{I} - \frac{\boldsymbol{\alpha}_{k_2}^{(2)} \boldsymbol{\alpha}_{k_2}^{(2)'}}{\boldsymbol{\alpha}_{k_2}^{(2)'} \boldsymbol{\alpha}_{k_2}^{(2)}}, \tag{3.6}$$

where $\mathbf{I}$ is an identity matrix.

$\boldsymbol{\alpha}^{(3)}$ is given by

$$\boldsymbol{\alpha}^{(3)} = \mathbf{M}\boldsymbol{\alpha}^{(2)}. \tag{3.7}$$

$\boldsymbol{\lambda}_{k_3}$ is the column of $\boldsymbol{\alpha}$ corresponding to the largest entry of the vector $\mathbf{w}(\boldsymbol{\alpha}^{(3)})^2$.

Pure Variable l

To obtain the l^{th} pure variable, we Gram-Schmidt orthogonalize $\boldsymbol{\alpha}^{(l-1)}$ to the set of pure variables that have currently been defined. The projection matrix, $\mathbf{M}$, is defined as

$$\mathbf{M} = \mathbf{I} - \frac{\boldsymbol{\alpha}_{k_{l-1}}^{(l-1)} \boldsymbol{\alpha}_{k_{l-1}}^{(l-1)'}}{\boldsymbol{\alpha}_{k_{l-1}}^{(l-1)'} \boldsymbol{\alpha}_{k_{l-1}}^{(l-1)}}. \tag{3.8}$$

$\boldsymbol{\alpha}^{(l)}$ is given by

$$\boldsymbol{\alpha}^{(l)} = \mathbf{M}\boldsymbol{\alpha}^{(l-1)}. \tag{3.9}$$

$\boldsymbol{\lambda}_{k_l}$ is the column of $\boldsymbol{\alpha}$ corresponding to the largest entry of the vector $\mathbf{w}(\boldsymbol{\alpha}^{(l)})^2$.

3.7 GoM Scores

The GoM scores are the $\boldsymbol{\alpha}$ values solved as a linear combination of the pure variables.

We can use linear algebra to solve for the partitioned GoM matrix, $\mathbf{G}$. Define $\boldsymbol{\Lambda}$ as a matrix of the first $K-1$ pure variables, as columns, centered by the K^{th} pure variable. Note that these pure variables are in the original $\boldsymbol{\alpha}$ space. $\boldsymbol{\Lambda}$ is given by

$$\boldsymbol{\Lambda} = \begin{bmatrix} (\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_K) & \cdots & (\boldsymbol{\lambda}_{K-1} - \boldsymbol{\lambda}_K) \end{bmatrix}. \tag{3.10}$$

The GoM scores are given by the partitioned formulation

$$\mathbf{G} = \begin{bmatrix} \boldsymbol{\Lambda}^{-1}(\boldsymbol{\alpha} - \boldsymbol{\lambda}_K \mathbf{j}) \\ \mathbf{j} - \mathbf{q}\boldsymbol{\Lambda}^{-1}(\boldsymbol{\alpha} - \boldsymbol{\lambda}_K \mathbf{j}) \end{bmatrix} \tag{3.11}$$

where $\boldsymbol{\Lambda}^{-1}$ denotes the Moore-Penrose generalized inverse of $\boldsymbol{\Lambda}$ and $\mathbf{q}$ is a row vector of 1s with length r. By the method of partitioning defined above, the GoM scores for each subject sum to 1. We prove this as Theorem 1.

Theorem 1: Let $\mathbf{H}$ be a matrix with arbitrary entries, r rows, and n columns. Define the partitioned matrix, $\mathbf{A}$, as

$$\mathbf{A} = \begin{bmatrix} \mathbf{H} \\ \mathbf{j} - \mathbf{qH} \end{bmatrix}$$

where $\mathbf{j}$ is a row vector of length n with 1 in each entry and $\mathbf{q}$ is a row vector of length r with 1 in each entry.

Then, every column in $\mathbf{A}$ sums to 1.

Theorem 1 Proof: Since $\mathbf{A}$ has $r + 1$ rows by construction, we can left multiply it by $\mathbf{t}$, an $r + 1$ length row vector with 1 in each entry, to obtain the column sums. Since

$$\mathbf{tA} = \mathbf{t} \begin{bmatrix} \mathbf{H} \\ \mathbf{j} - \mathbf{qH} \end{bmatrix} = \mathbf{qH} + 1(\mathbf{j} - \mathbf{qH}) = \mathbf{j},$$

each column of $\mathbf{A}$ sums to 1, regardless of the $\mathbf{H}$ elements.

Note that if a particular point is contained in the simplex, it's GoM scores will be non-negative. The Grande and Manne (1999) algorithm does not guarantee that all points will be contained in the resulting pure variable simplex. A simplex that does not contain all data set points is referred to as "affine." This means some GoM scores would be negative. Theoretically, a GoM score can not be negative because that would imply that the subject was more than completely orthogonal to a particular subpopulation. The next section describes our approach to reconcile this issue.

3.8 Extending Pure Variables

A convex combination is a representation of the data in which no points lie outside the simplex. Such a simplex can contain points on a corner, edge, or face, but no points are outside. Using the algorithm developed by Dr. David Wright of the BYU Mathematics Department, see Appendix B, we extend the pure variables such that the resulting simplex contains all

points as a convex combination. Since Wright's algorithm guarantees a containing simplex, no GoM scores will be negative.

Assume subject $\mathbf{e}$ has at least one negative GoM score. As previously described, we would refer to GoM scores for such a subject as being an affine combination of the uncentered pure variable set,

$$\mathbf{Q} = \begin{bmatrix} \boldsymbol{\lambda}_1 & \cdots & \boldsymbol{\lambda}_K \end{bmatrix}. \tag{3.12}$$

Equivalently,

$$\mathbf{e} = \sum_{i=1}^{m} \mathbf{s}_i \mathbf{v}_i + \sum_{j=1}^{n} \mathbf{t}_j \mathbf{w}_j, \tag{3.13}$$

where $\mathbf{v}_i$ are the columns of $\mathbf{Q}$ corresponding to nonnegative GoM scores, $\mathbf{w}_j$ are the columns of $\mathbf{Q}$ corresponding to negative GoM scores, $\mathbf{s}_i \geq 0$, and $\mathbf{t}_j < 0$. In other words, $\mathbf{s}_i$ are the nonnegative GoM scores and $\mathbf{t}_j$ are the negative GoM scores.

Define a column vector as

$$\mathbf{v}_i' = \left(\sum_{j=1}^{m} \mathbf{s}_j \right) \mathbf{v}_i + \sum_{j=1}^{n} \mathbf{t}_j \mathbf{w}_j. \tag{3.14}$$

Dr. Wright shows, as given in Appendix B, that

$$\mathbf{Q}' = \begin{bmatrix} \mathbf{v}_1{}' & \cdots & \mathbf{v}_m{}' & \mathbf{w}_1 \cdots \mathbf{w}_n \end{bmatrix} \tag{3.15}$$

represents a convex simplex for subject $\mathbf{e}$. Our application is to apply this algorithm in sequence on each affine subject. Thus, the resulting GoM scores will be nonnegative and sum to unity.

3.9 New Observations

When a new observation is observed, GoM scores can typically be obtained using 3.1 – 3.11. A complication arises, however, when the subject lies outside of simplex because their GoM scores will not meet the nonnegativity constraints. As mentioned before, this is a major issue theoretically because subjects can not have negative subpopulation membership. We assume that the simplex is fixed and we wish to parametrize the new subject's GoM scores with respect to it. To correct the issue, we project these subjects onto the previously determined simplex. We can accomplish this task using least squares regression. Consider a particular combination of pure variables. This combination defines an edge, face, etc. in $\boldsymbol{\alpha}$ space, depending on the number of pure variables in the combination. We can obtain the $\boldsymbol{\alpha}$ coordinates of the orthogonal projection for a newly observed subject, $\boldsymbol{\alpha}_{new}$, through the algorithm outlined below.

Consider a possible subset of pure variables, $\mathbf{L} \subset \mathbf{Q}$, with p nonrepeated elements of $\mathbf{Q}$. Then, the corresponding candidate projection is given by

$$\boldsymbol{\beta} = (\mathbf{V}'\mathbf{V})^{-1}\mathbf{V}'(\boldsymbol{\alpha}_{new} - \mathbf{L}_1), \tag{3.16}$$

where

$$\mathbf{V} = \begin{bmatrix} (\mathbf{L}_2 - \mathbf{L}_1) & \cdots & (\mathbf{L}_p - \mathbf{L}_1) \end{bmatrix}. \tag{3.17}$$

This algorithm will produce a number of candidate projections for each subject. The selected projection is the one with the minimum euclidean distance between the new subject's original (affine) GoM scores and the candidate projection. Note that simplex endpoints, the pure variables themselves, should also be considered as possible projections. Once the optimal projection is selected, GoM scores can be obtained by 3.11.

Back transforming GoM scores into $\boldsymbol{\alpha}$ coordinates

If one had a set of GoM scores for which he or she wanted to know the corresponding $\boldsymbol{\alpha}$ coordinates, then he or she could apply some matrix algebra to obtain the corresponding $\boldsymbol{\alpha}$ coordinates. Recall that solving for the GoM scores is done in partitions. The first partition, $\mathbf{G}_1$, is the only one necessary for this process. The desired $\boldsymbol{\alpha}$ coordinates are given by

$$\boldsymbol{\alpha} = \boldsymbol{\Lambda}\mathbf{G}_1 + \boldsymbol{\lambda}_K\mathbf{j}. \tag{3.18}$$

To see logic in this formula, consider the right side of 3.18 in parts. The first term multiplies the centered pure variables by the corresponding GoM scores (not including the last GoM score). This gives us the location of the point in an $\boldsymbol{\alpha}$ space centered around the last pure variable. To correct for the centering, we add $\boldsymbol{\lambda}_K$ to the space via term 2. While this transformation is not used in our algorithm, it may be useful in future research.

CHAPTER 4

ANALYSIS

4.1 Introductory Statistics Course Analysis

Introduction

In order to predict Statistics 221 student success, as measured by final grade in the course, data were gathered on 488 students in the Winter 2009 semester. These students were given a survey at the beginning of the semester which included 58 questions regarding "math anxiety" and 44 questions regarding "learning style." In addition, typical "on file" data such as gender, age, ACT math score, and high school GPA were available for each student.

The primary interest was to predict success in the course based on latently characterized subgroups. Such information would be useful in either discriminating admittance to the course, encouraging further prerequisites, or catering review sessions to different subgroups. Subjects self selected themselves, in order to receive extra credit in the course, and are certainly not a fully randomized representation of the courses student body. However, if preliminary results can be found with these data, then a stronger study could be implemented in future semesters. The data were obtained from BYU Statistics Department CSRs. A full detail of the data is given in Appendix A.

Features of the Data

The collection of covariates comes from three sources: the "math anxiety" survey, the "learning style" survey, and the "on file" material. A separate introduction to each source is necessary.

The 58 question "math anxiety" survey measured both students experiences with past math courses and feeling towards the math discipline in general. Students gave an ordinal rating ranging from *strongly disagree* to *strongly agree* for the majority of questions. Although a small number of additional questions were given which profiled each student's math/stat coursework history, most questions were straightforward opinion questions such as:

(a) Math and statistics are the same.

(b) Statistics conclusions are rarely presented in everyday life.

(c) Statistics involves massive computations.

The 44 question "learning style" survey measured student learning techniques. All binary responses, questions included:

(a) I understand something better after I

try it out.

think it through.

(b) Once I understand

all the parts, I understand the whole thing.

the whole thing, I see how the parts fit.

(c) When I get directions to a new place, I prefer

a map.

written instructions.

A number of "on file" variables were available. Due to distributional skewness, Age and high school GPA were dichotomized into 2 quantile factor levels each. Verification of

these findings is given via variable histograms in Figure 4.1 and Figure 4.2. Both histograms illustrate the heavy skewness in each respective distribution.

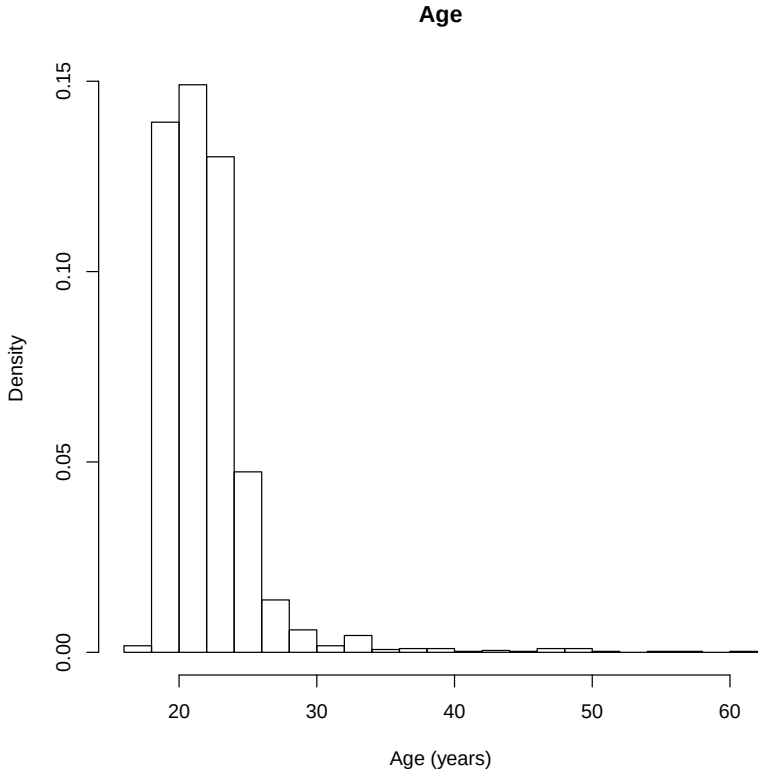

Figure 4.1: Reasoning for dichotomizing Age. The distribution of Age is skewed right. To prevent high aged persons from overly influencing the data, we dichotomize age.

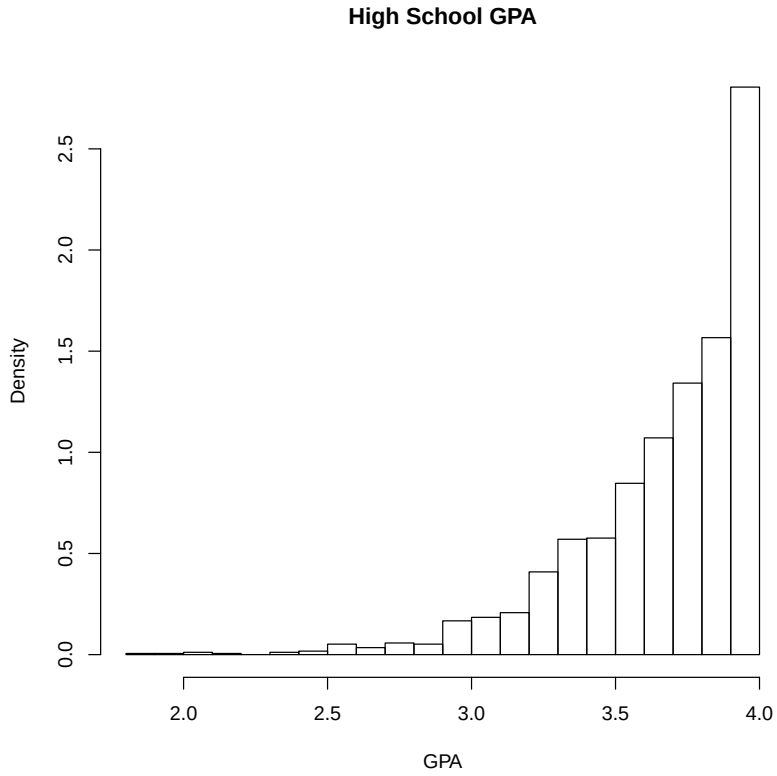

Figure 4.2: Reasoning for dichotomizing High School GPA. The distribution of GPA is skewed left. To prevent abnormally poor students from overly influencing the data, we dichotomize GPA.

High school GPA and math ACT had 77 and 80 missing entries respectively. These subjects are likely from a different population compared to the fully observed subjects. To account for this, additional indicator variables denoting missing values for each of these covariates were added. This methodology works better for factor variables than for continuous covariates. To account for this, math ACT was quantile dichotomized much like the process discussed above for high school GPA and age. No other covariates had missing values. Of all respondents, 63% were female. As is typical for STAT 221 students, the average number of years at BYU was 2.6.

Exploratory Analysis

Using the outlined methodology, we now analyze latent fuzzy subpopulations of the data. We selected to use the supervised options of the algorithm when analyzing the data. The supervised variable, which can be thought of as a response variable, was final grade in the course. Note that we have not utilized the IPF yet. We give the Broken-Stick plot to illustrate a weakness of the method. Note that although most analyzers would select around $K = 9$ for this data, the method selects 23. We performed this analysis using 9 pure types (denoted in dark red), although the choice is somewhat arbitrary.

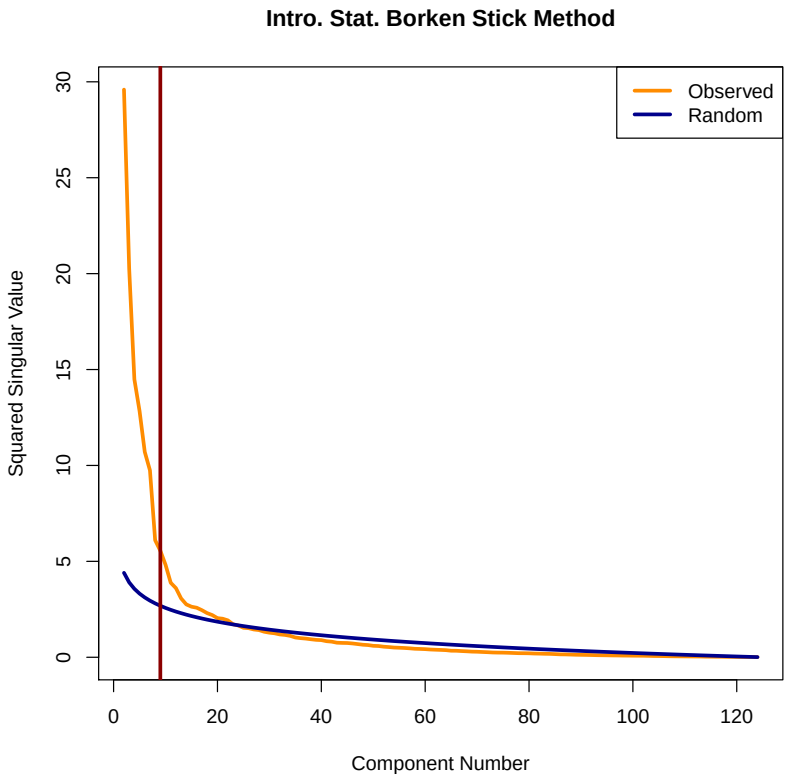

Figure 4.3: Broken-Stick Method Weakness.

A particular subject is defined as a heavy member of a particular pure type, say w, if their GoM w score lies 2 standard deviations above the mean GoM w score. The choice of 2 standard deviations is arbitrary and should be considered a tuning parameter for this

analysis. Note that using another reasonable approach would be to use GoM $w >$ a constant as an indicator of heavy membership. The problem with this approach in many cases is that the average GoM score changes drastically for different values of w.

By searching for similarities between the heavy members of each respective pure type, we generate a stereotype chart, given as Table 4.1. To create this chart we found the mean values of each measurements across heavy member groups of each pure type. Then when comparing the mean values of each measurement across pure types, we selected the variables where there was the largest discrepancy amongst pure types. The values are in percentage units. Note the comparison to the overall sample averages.

Table 4.1: Introductory Statistics Student Stereotypes.

Pure Type	Grade	AC3	Q52(i)	Q52(j)	L11	L12	L23	L25	L26	L37	L43	L44
1	74		100	100		43		21				
2	92			25			25	75			25	50
3	80				73		73	87	87	80	33	40
4	73	29	65									
5	81			89	58						26	
6	86	33	78	78	56		100	67				
7	70						82		64	9		36
8	77		31							92		
9	79	47				37						
Overall	80	22	61	67	41	20	62	53	47	50	17	27

We detail the Table 4.1 selected questions in Table 4.2.

Table 4.2: Key Questions Separating Stereotypes: Selected Question Detail.

Measurement	Question
AC3	Did not take Math ACT?
Q52(i)	Concerned about math formulas?
Q52(j)	Afraid that he or she won't understand material?
L11	Prefers text over pictures?
L12	Big picture easier than details?
L23	Written instructions over map?
L25	Think it through before trying it?
L26	Prefer creative writing over clarity?
L37	Considers themself reserved?
L43	Has trouble picturing details of a place?
L44	Tries to make ties between concepts?

To illustrate the implicit interaction substance that the GoM variables posses, consider Figure 4.4.

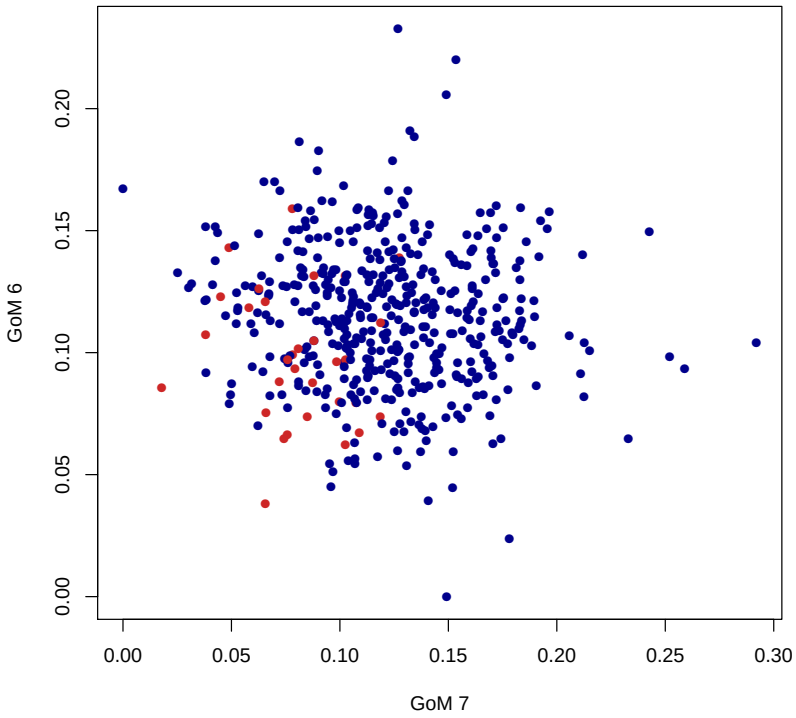

Figure 4.4: Implicit Interaction Structure. The students colored red are those with high Math ACT scores, high high school GPAs, and L37.

Coloring a couple of the GoM variables by a 3 way interaction reveals that the GoM variables themselves include interaction substance. The students colored red are those with high Math ACT scores, high high school GPAs, and L37. These students averaged a final grade of 92.0%, as compared to the class average of 79.9%. One might conclude that this region of the GoM space is filled with reserved students with strong academic backgrounds. The GoM scores may be more valuable than the interaction classification by itself because they preserve the integrity of discrimination intensity. This may suggest that our initial hypothesis was correct: the GoM space is more valuable than the collection interaction effects. We say this because the interaction variable is a binary piece of information, but the GoM space is continuous. For illustration, it may be the case that the farther down a subject's GoM 6 and GoM 7 scores are, the more similar they are to the stereotype described. In other

words, the grade of their membership may give more information about their similarity to the particular subpopulation. This concept has the potential to be extremely valuable in prediction improvement.

Prediction Models

To show the value of the GoM scores in grade prediction, consider Table 4.3 and Table 4.4. Table 4.3 gives a model summary for a typical regression approach. To decide which covariates to include in the typical regression model, we utilized stepwise AIC selection. In Table 4.3, effects are given in the order they were selected. The multiple R^2 for the typical regression model was .220 on 10 degrees of freedom. Our GoM regression approach is to use a few main effects along with the GoM variables as covariates. We remove one of the GoM variables to prevent collinearity due to the GoM sum to unity constraint. The main effect was AC2, which indicated if the student was in the top third of Math ACT scores. The multiple R^2 for the GoM regression model was .180 on 10 degrees of freedom. Table 4.4 gives a model summary our GoM regression approach.

Table 4.3: Typical Regression Model. Effects are given in the order they were selected.

	Estimate	Std. Error	T score	P-value
(Intercept)	69.00	3.39	20.34	0.00
AC2	7.42	1.75	4.24	0.00
Q48	2.30	0.59	3.89	0.00
Q2	2.13	0.68	3.13	0.00
L25	4.41	1.42	3.10	0.00
Q47	-1.49	0.54	-2.77	0.01
L31	-3.91	1.63	-2.41	0.02
L22	-4.61	1.57	-2.93	0.00
L17	-3.72	1.41	-2.65	0.01
L10	3.29	1.46	2.26	0.02

Table 4.4: GoM Regression Model. Note that one main effect, AC3, was selected.

	Estimate	Std. Error	T score	P-value
(Intercept)	75.28	17.96	4.19	0.00
AC3	6.81	1.93	3.52	0.00
GoM1	-50.73	28.28	-1.79	0.07
GoM2	108.49	25.47	4.26	0.00
GoM3	-7.60	36.15	-0.21	0.83
GoM4	-20.57	39.39	-0.52	0.60
GoM5	17.61	19.58	0.90	0.37
GoM6	-7.10	29.84	-0.24	0.81
GoM7	-47.94	30.29	-1.58	0.11
GoM8	12.05	22.79	0.53	0.60

The reason that the GoM approach did not work as well as the typical approach is that the GoM space is highly nonlinear in nature. As a more appropriate model choice, consider the Breiman tree models proposed in Figure 4.5 and Figure 4.6. We constrained the models in this section to have 10 ending nodes so that they would be comparable to the regression type models. A typical Breiman tree is given as Figure 4.5. This model yielded a .278 multiple R^2 on 10 degrees of freedom. A GoM Breiman tree model is given as Figure 4.6. Similar to the GoM regression model, this model included AC3 and the GoM variables in the decision algorithm. This model yielded a multiple R^2 of .288, higher than the typical approach.

Since the GoM space is a data reduction of the original space, it would initially seem inconceivable that it would have more predictive value than the original data space. However, there is one important consideration: interactions are clearly defined and easier to manage in the GoM space by construction. We note that the GoM variables often act as compilation effects and are, for that reason, more powerful in determining splits than the original measurements.

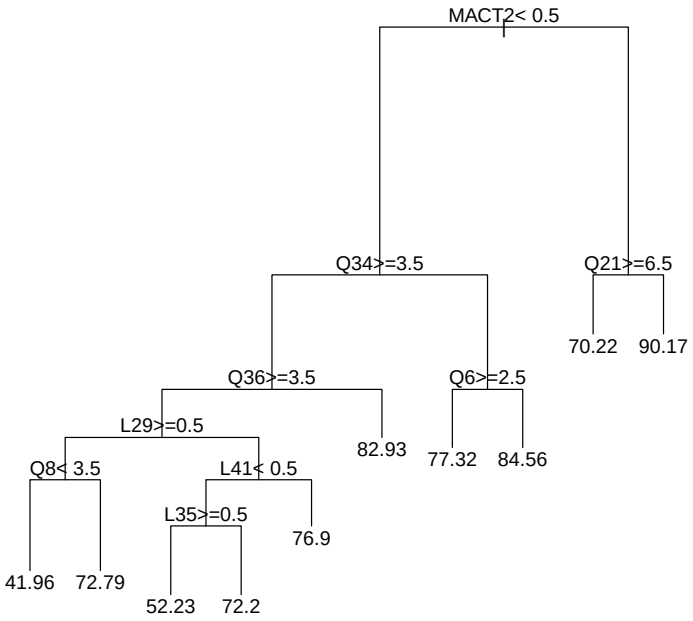

Figure 4.5: Typical Breiman Tree.

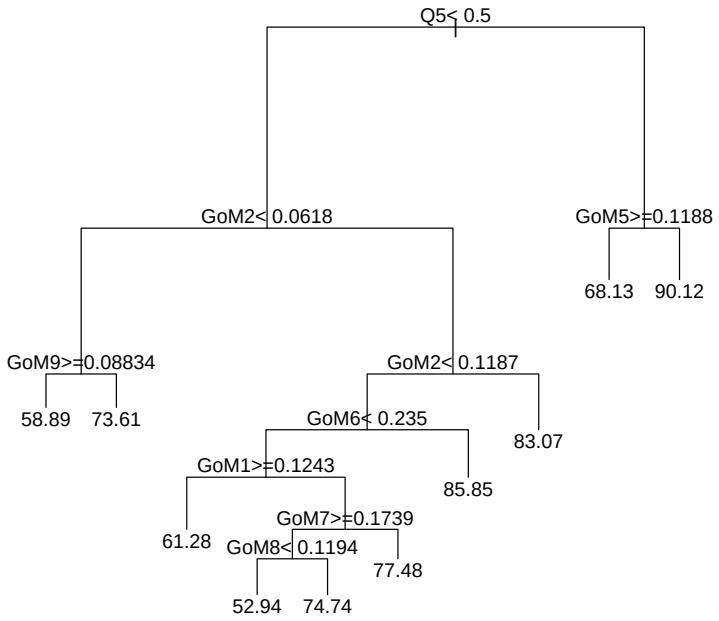

Figure 4.6: GoM Breiman Tree.

Extension to the IPF

Continuing the previous analysis, we introduce IPF GoM scores into the comparison. Using the GoM scores of our algorithm, we initialize the IPF. A scatter plot matrix of the IPF GoM structure is given as Figure 4.7. A particular snapshot of the GoM space, to illustrate the multidimensional correlation, is given as Figure 4.8.

Keep in mind that a GoM space contains information about the relationship between measurements with regard to latent fuzzy subpopulations. The nonlinear structure of this particular relationship is strong evidence that crisp subpopulation membership is not sufficient to capture the true underlying signal of the measurements. For this reason, an Analysis of Variance model on strict subpopulation membership is in reality a poor approximation to a model that appropriately utilizes the GoM space. In other words, the nonlinear rela-

tionship is evidence that a GoM space carries more information than strict subpopulation membership could provide.

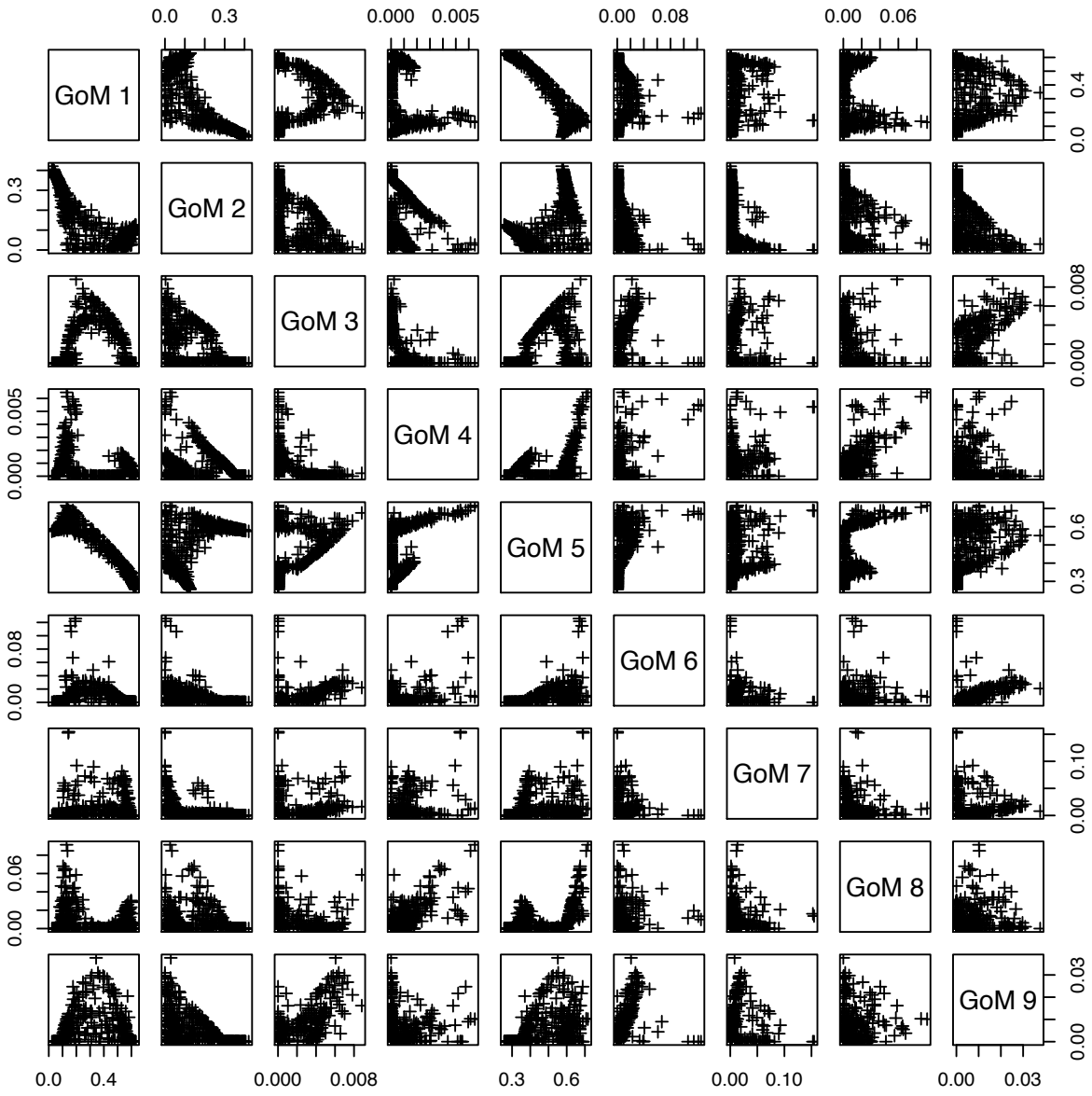

Figure 4.7: IPF GoM Scatter plot Matrix.

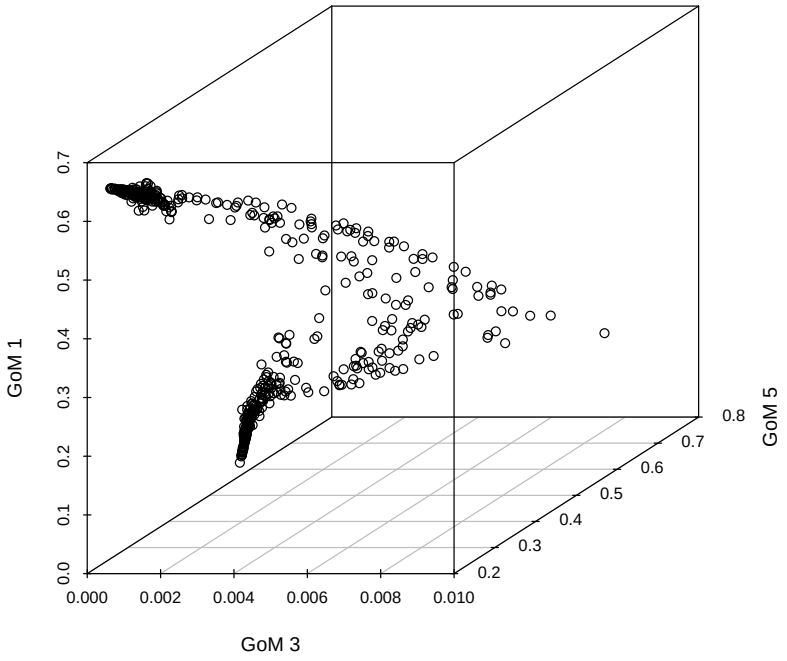

Figure 4.8: IPF GoM Snapshot.

Using the resulting IPF GoM scores and the same model methodology used before, we create two models: An regression model and a tree model. A summary of the IPF regression model is given in Table 4.5. The corresponding multiple R^2 was .178 on 10 degrees of freedom, which is about the same as the previous figures. The IPF tree model is given in Figure 4.9. This model's multiple R^2 was .271, which is also about the same as the previous figures. One substantial benefit of the IPF is the structure can be fit to a stochastic model.

Table 4.5: IPF GoM Regression Model Summary.

	Estimate	Std. Error	T score	P-value
(Intercept)	268.51	358.37	0.75	0.45
AC3	6.86	1.83	3.75	0.00
GoM1	-185.22	346.87	-0.53	0.59
GoM2	-70.10	288.33	-0.24	0.81
GoM3	2271.34	1540.00	1.47	0.14
GoM4	4089.09	2677.31	1.53	0.13
GoM5	-250.99	399.48	-0.63	0.53
GoM6	-4.88	219.45	-0.02	0.98
GoM7	-143.82	222.63	-0.65	0.52
GoM8	0.22	226.45	0.00	1.00

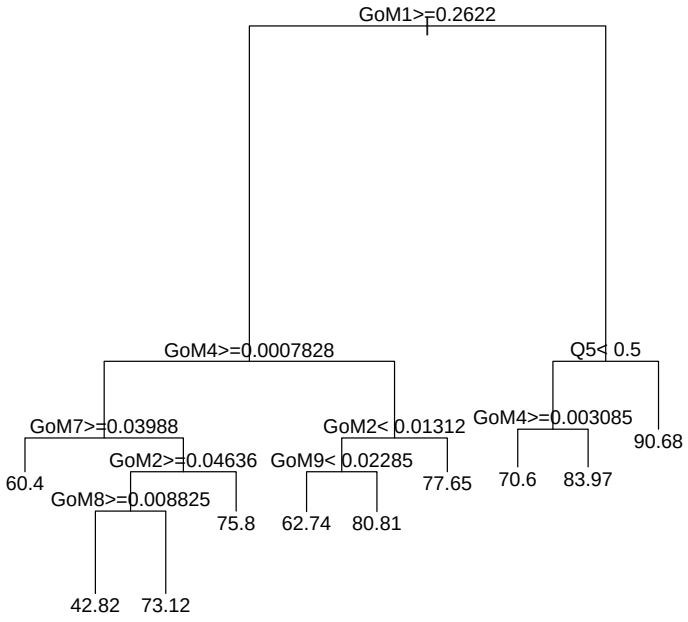

Figure 4.9: IPF GoM Tree.

4.2 Comparison to Other Methods

To validate our methodology, we will compare our results to that of Zarndt (1995). Zarndt (1995) compared a wide range of modeling algorithms using 10-fold cross-validation across

a high number of data sets. After introducing a few of his data sets, we will compare our results to his.

Brief Data Introductions

IONOSPHERE The Johns Hopkins University Ionosphere database was obtained from the UCI Machine Learning Repository. At each of 351 distinct Ionosphere coordinates, 16 antennas (2 readings each) were used to measure overall Ionosphere substance. The location was separately rated as good or bad with good denoting a reasonable amount of substance present. The goal was to build a model which could correctly classify locations as good or bad based on the 32 readings. The data were obtained at

`http://archive.ics.uci.edu/ml/machine-learning-databases/ionosphere/`

CLEVELAND HEART HEALTH Cleveland heart health data was obtained from the Cleveland Clinic Foundation, via the UCI Machine Learning Repository. Twelve discriminatory attributes were taken on each of 297 subjects in order to predict the multilevel categorical response variable. The response was strength of heart disease, measured ordinally from 0 to 4 with 0 denoting no evidence of disease and 4 representing heavy symptoms of disease. The data were obtained at

`http://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/`

MUSK Data were obtained on 476 molecules in order to classify the molecule as musky or not musky. The Musk scent is created from the chemical compound Muscon and is valuable in making various perfumes. In order to predict whether a particular molecule was musky, 166 distance measurements were taken. The data were obtained from UCI Machine Learning at

`http://archive.ics.uci.edu/ml/machine-learning-databases/musk/`

Comparison Results

To validate our methodology, we develop a Breiman decision tree using the GoM scores along with a small number of main effects. As a goal for further research, we point note that a stochastic model fit to the GoM space directly is the superior way to make predictions. However, that is beyond the breath of this work.

Since cross validation results fluctuate due to the random subsampling process, we must adjust our results to be comparable with Zarndt (1995). With our random subsamples, we apply one of the algorithms that Zarndt analyzed. The difference between our result and the Zarndt result can be considered as an additive bias between studies. We will refer to our bias correct result as "Corrected GoM." The results are given in Table 4.6. Keep in mind that these results are found using the default settings in our algorithm (no tuning). We would probably obtain marginally better results by changing where supervision is done or how many clusters to use. "Corrected GoM" is competitive in each case and is the best algorithm for the Ionosphere data. The number of main effects used is given in parenthesis by the data set name.

Table 4.6: Zarndt Model Comparison.

Data	Corrected GoM	Zarndt Average	Zarndt Best	Proportion Beat
Ionosphere (3)	92.3	87.5	92.0	16/16
Cleveland (0)	50.7	52.7	58.1	5/16
Musk (3)	80.3	79.3	83.4	4/9

4.3 Comparison to Breiman model for 10 Datasets

To give a feeling for the proportion of times that our GoM approach beats a typical Breiman et al. (1984) tree model, consider Table 4.7. The GoM approach performs better than the Breiman model in 5 of 10 cases. The first two data sets use the cross validation sum of squared errors metric (smaller is better). The remaining sets of data use the cross validation

correct classification rate (bigger is better). The number of main effects used in the GoM approach is given in parenthesis next to the data name. This number was set arbitrarily based on an initial Breiman model.

Table 4.7: Ten Dataset Breiman Comparison.

Data	GoM Approach	Typical Breiman
Intro. Stat (1)	15396	17187
SAGE (1)	7888	7756
Iono (0)	0.89	0.86
Heart (0)	0.56	0.56
Musk (3)	0.71	0.76
Glass (0)	0.43	0.45
Wine (3)	0.94	0.92
Sonar (3)	0.71	0.70
Robot (3)	0.96	0.98
Mushroom (0)	0.99	0.99

CHAPTER 5

CONCLUSIONS AND FURTHER RESEARCH

We have introduced a new powerful way to create GoM scores for large multivariate observational studies. We have shown that these GoM scores summarize interaction effects and have predictive value. In some cases, our GoM method outperforms a traditional Breiman tree model. In the case of the Ionosphere data, the GoM approach performed better than all other 16 methods compared by Zarndt (1995).

The next tool to develop is a stochastic prediction model fit to the GoM space. This model would likely use a transformed version of the GoM space, say the GoM odds ratios. Further work could be done in dealing with outliers. Outliers are often selected as pure variables with the current algorithm since they are very far away from the bulk of the data. In some cases, it might be better to project outliers onto an affine simplex around the bulk of the data before determining the pure variables. This might increase stability and heavy membership in corresponding pure types.

BIBLIOGRAPHY

Bajorski, P. (2004), "Simplex Projection Methods for Selection of Endmembers in Hyperspectral Imagery," *Center for Quality and Applied Statistics, Kate Gleason College of Engineering, Rochester Institute of Technology*, 4.

Breiman, L. (2001), "Random Forests," *Machine Learning*, 45, 5–32.

Breiman, L., Friedman, J., Stone, C., and Olshen, R. (1984), *Classification and Regression Trees*, Chapman and Hall/CRC.

Cannon, P. (2008), "Extending the Information Partition Function: Modeling Interaction Effects in Highly Multivariate Discrete Data," *M.S. Thesis, Department of Statistics, Brigham Young University, Provo, Utah.*

Engler, D. (2002), "An Approach to Probabilistic Record Linkage: Building a Model Through Probability as Extended Logic and Maximum Entropy," *M.S. Thesis, Department of Statistics, Brigham Young University, Provo, Utah.*

Giddings, J. (1965), *Dynamics of Chromatography*, Marcel Dekker.

Gollob, H. (1968), "A Statistical Model Which Combines Features of Factor Analytic and Analysis of Variance Techniques," *Psychometrika*, 33, No. 1, 73-115.

Grande, B.-V., and Manne, R. (1999), "Use of Convexity for Finding Pure Variables in Two-way Data from Mixtures," *Chemometrics and Intelligent Laboratory Systems*, 50, 19–33.

Jackson, D. (1993), "Stopping Rules in Principal Components Analysis: A Comparison of Heuristical and Statistical Approaches," *Ecology*, 74, 2204–2214.

Koch, G., Landis, J., Freeman, J., Freeman, D., and Lehnen, R. (1977), "A General Methodology for the Analysis of Experiments with Repeated Measurement of Categorical Data," *Biometrics*, 33, 133–58.

Mandel, J. (1969), "The Partitioning of Interaction in Analysis of Variance," *Journal of Research of the National Bureau of Standards*, 73b, No. 4, 309-328.

Oliphant, J. (2003), "The Information Partition Function," *M.S. Thesis, Department of Statistics, Brigham Young University, Provo, Utah.*

Tolley, H., and Manton, K. (1992), "Large Sample Properties of Estimates of Discrete Grade of Membership Models." *Annals of Statistical Mathematics*, 44, 85–95.

Tolley, H., Oliphant, J., and Eliason, R. (2010), "Modeling Aggregate Interaction Effects in Many Variable Observational Studies," *Statistics in Biopharmaceutical Research*, to appear.

Zarndt, F. (1995), "A Comprehensive Case Study: An Examination of Machine Learning and Confectionist Algorithms," *M.S. Thesis, Department of Computer Science, Brigham Young University, Provo, Utah.*

APPENDICES

APPENDIX A

INTRODUCTORY STATISTICS COURSE DATA DETAIL

The 221 Dataset was based on information from the winter 2009 semester. Two pre-semester surveys were combined with students academic history to predict 221 final grades. The continuous response variable used was the stat 221 final grade percentage given for each student.

A.1 Questionnaire

The questionnaire was designed to identify student attitudes about mathematics. There were 58 questions.

In questions 1-10, students were given a prompt to rate the statement as Strongly Disagree, Disagree, Undecided, Agree, or Strongly Agree.

In questions 11-45, students were given a prompt to rate the statment on a sclae of 1 to 7 with 7 representing strongly agree and 1 representing strongly disagree. In questions 46-48, students were given a prompt to rate the statment on a sclae of 1 to 7 with 7 representing high and 1 representing low. Questions 49-56 are detailed individually. The survey is included below.

1. It wouldn't bother me at all to take more math courses.

2. I have usually been at ease during math tests.

3. I have usually been at ease in math courses

4. I usually don't worry about my ability to solve math problems.

5. I almost never get uptight while taking math tests.

6. I get really uptight during math tests.
7. I get a sinking feeling when I think of trying hard math problems.
8. My mind goes blank and I am unable to think clearly when working mathematics.
9. Mathematics makes me feel uncomfortable and nervous.
10. Mathematics makes me feel uneasy and confused.
11. I will like statistics.
12. I will feel insecure when I have to do statistics problems.
13. Statistics is nothing like math.
14. I will have trouble understanding statistics because of how I think.
15. Statistics formulas are easy to understand.
16. Statistics is math.
17. I have no idea of what's going on in mathematics.
18. Statistics is worthless.
19. Statistics and math are completely different.
20. Statistics is a complicated subject.
21. Statistics should be a required part of my professional training.
22. Statistical skills will make me more employable.
23. I will have no idea of what's going on in statistics.
24. Statistics is not useful to the typical professional.
25. I will get frustrated going over statistics tests in class.

26. Math and statistics are very similar.

27. Statistical thinking is not applicable in my life outside my job.

28. I use statistics in my everyday life.

29. I will be under stress during statistics class.

30. I will enjoy taking statistics courses.

31. Statistics conclusions are rarely presented in everyday life.

32. Statistics is a subject quickly learned by most people.

33. Learning statistics requires a great deal of discipline.

34. I will have no application for statistics in my profession.

35. I will make a lot of math errors in statistics.

36. I am scared by statistics.

37. Math and statistics are the same.

38. Statistics involves massive computations.

39. I can learn statistics.

40. I will understand statistics equations.

41. Statistics is irrelevant in my life.

42. Statistics is highly technical.

43. I will find it difficult to understand statistics concepts.

44. Most people have to learn a new way of thinking to do statistics.

45. Math and statistics are only slightly related.

46. How good at mathematics are you?

47. How much computer experience have you had?

48. How confident are you that you can master introductory statistics material?

49. What is the most recent mathematics course you have completed?

    (a) High school algebra

    (b) High school calculus

    (c) College algebra

    (d) College calculus

    (e) Other

50. Please choose one and only one response from the following list. If you have taken more than one Statistics class before this semester please choose the response from items B through E that corresponds to the most recent class you have taken.

    (a) I have not taken a Statistics class before.

    (b) I took a Statistics class in high school.

    (c) I took 221 here at BYU or some other place and am repeating the class.

    (d) I took a Statistics class other than 221 here at BYU.

    (e) I took a Statistics class at another college/university before I came to BYU.

51. Are you anxious or nervous regarding this class?

    (a) Yes

    (b) No

    If you answered YES to question 51 please answer question 52 and identify the reasons which most closely reflect the basis for your concern. Answer Agree or Disagree.

52. (a) I really don't know what to expect.

(b) Previous experience with Statistics was not positive.

(c) I am always a bit nervous at the beginning of a new course.

(d) I am worried about the grade I will get.

(e) I am concerned about the workload in the class.

(f) I do not do very well in math.

(g) It has been a long time since I have taken any math class.

(h) I do not do very well with story problems in math.

(i) I am concerned about complex formulas that might be in the class.

(j) I am afraid I won't understand the material.

(k) I have heard it is a very hard class.

(l) I am anxious because I am excited to learn the material in the class.

(m) I have heard a lot of horror stories about the class.

(n) I am concerned that I might not get my questions answered because of the class size.

The following six questions (53-58) cover concepts taught in Statistics 221. Please answer them as best you can. If you don't know the correct answer, select the last option.

53. What is the best reason for obtaining a random sample when you want to estimate a population parameter?

(a) It is the cheapest method.

(b) It results in the most representative estimate.

(c) It results in the estimate with the smallest standard deviation.

(d) It is the easiest method.

(e) I am unable to answer this question with reasonable certainty.

54. Which research method is most likely to support a causal relationship between two variables?

(a) A sample survey based on a simple random sample.

(b) An observational study based on a carefully selected large random sample.

(c) A comparative experiment implementing principles of randomization and replication.

(d) A correlational study that measures two variables on a large random sample of people of the same age, gender, and socioeconomic status.

(e) A matched pairs experiment using twins who were hand picked by the researcher to participate.

(f) I am unable to answer this question with reasonable certainty.

55. What is the purpose of a confidence interval for a population mean?

(a) To provide confidence in our sample mean.

(b) To give a range of plausible values for the population mean.

(c) To show how close our sample mean is to the population mean.

(d) To determine if the population mean takes on a particular value.

(e) I am unable to answer this question with reasonable certainty.

56. A social scientist did a study on gender and attitudes toward gun control, found a p-value of 0.042, and concluded there was a relationship. Which of the following represents a practical interpretation of the study?

(a) 4.2% of the respondents were in favor of gun control.

(b) One's attitude towards gun control is associated with one's gender.

(c) The difference between males and females who favored gun control was 4.2%.

(d) One's attitude towards gun control is not associated with one's gender.

(e) I am unable to answer this question with reasonable certainty.

57. When a result is 'statistically significant' this means that the result...

(a) has a very small probability of occurring by chance.

(b) is important enough that most people would believe it.

(c) is important enough to make a meaningful contribution to its subject area.

(d) has a very large probability of occurring by chance.

(e) I am unable to answer this question with reasonable certainty.

58. A newspaper report claims a margin of error of 4% when reporting the percentage of people in favor of a tax refund. Which of the following is the best explanation and interpretation of what margin of error is?

(a) A number that comes from the sampling distribution and tells us how close we are to the truth.

(b) A number that tells how much error can be expected because of chance variation.

(c) It is the population standard deviation divided by the square root of the sample size.

(d) A number that tells us how far our sample size is from what it ought to be.

(e) I am unable to answer this question with reasonable certainty.

A.2 Learning Style Survey

The dichotomous 44-question learning style survey gauged students learning techniques. The survey was given as follows:

1. I understand something better after I

(a) try it out.

(b) think it through.

2. I would rather be considered

   (a) realistic.

   (b) innovative.

3. When I think about what I did yesterday, I am most likely to get

   (a) a picture.

   (b) words.

4. I tend to

   (a) understand details of a subject but may be fuzzy about its overall structure.

   (b) understand the overall structure but may be fuzzy about details.

5. When I am learning something new, it helps me to

   (a) talk about it.

   (b) think about it.

6. If I were a teacher, I would rather teach a course

   (a) that deals with facts and real life situations.

   (b) that deals with ideas and theories.

7. I prefer to get new information in

   (a) pictures, diagrams, graphs, or maps.

   (b) written directions or verbal information.

8. Once I understand

   (a) all the parts, I understand the whole thing.

   (b) the whole thing, I see how the parts fit.

9. In a study group working on difficult material, I am more likely to

   (a) jump in and contribute ideas.

   (b) sit back and listen.

10. I find it easier

    (a) to learn facts.

    (b) to learn concepts.

11. In a book with lots of pictures and charts, I am likely to

    (a) look over the pictures and charts carefully.

    (b) focus on the written text.

12. When I solve math problems

    (a) I usually work my way to the solutions one step at a time.

    (b) I often see the solutions but then have to struggle to figure out the steps to get to them.

13. In classes I have taken

    (a) I have usually gotten to know many of the students.

    (b) I have rarely gotten to know many of the students.

14. In reading nonfiction, I prefer

    (a) something that teaches me new facts or tells me how to do something.

    (b) something that gives me new ideas to think about.

15. I like teachers

    (a) who put a lot of diagrams on the board.

(b) who spend a lot of time explaining.

16. When I'm analyzing a story or novel

(a) I think of the incidents and try to put them together to figure out the themes.

(b) I just know what the themes are when I finish reading and then I have to go back and find the incidents that demonstrate them.

17. When I start a homework problem, I am more likely to

(a) start working on the solution immediately.

(b) try to fully understand the problem first.

18. I prefer the idea of

(a) certainty.

(b) theory.

19. I remember best

(a) what I see.

(b) what I hear.

20. It is more important to me that an instructor

(a) lay out the material in clear sequential steps.

(b) give me an overall picture and relate the material to other subjects.

21. I prefer to study

(a) in a study group.

(b) alone.

22. I am more likely to be considered

(a) careful about the details of my work.

(b) creative about how to do my work.

23. When I get directions to a new place, I prefer

(a) a map.

(b) written instructions.

24. I learn

(a) at a fairly regular pace. If I study hard, I'll "get it."

(b) in fits and starts. I'll be totally confused and then suddenly it all "clicks."

25. I would rather first

(a) try things out.

(b) think about how I'm going to do it.

26. When I am reading for enjoyment, I like writers to

(a) clearly say what they mean.

(b) say things in creative, interesting ways.

27. When I see a diagram or sketch in class, I am most likely to remember

(a) the picture.

(b) what the instructor said about it.

28. When considering a body of information, I am more likely to

(a) focus on details and miss the big picture.

(b) try to understand the big picture before getting into the details.

29. I more easily remember

(a) something I have done.

(b) something I have thought a lot about.

30. When I have to perform a task, I prefer to

(a) master one way of doing it.

(b) come up with new ways of doing it.

31. When someone is showing me data, I prefer

(a) charts or graphs.

(b) text summarizing the results.

32. When writing a paper, I am more likely to

(a) work on (think about or write) the beginning of the paper and progress forward.

(b) work on (think about or write) different parts of the paper and then order them.

33. When I have to work on a group project, I first want to

(a) have "group brainstorming" where everyone contributes ideas.

(b) brainstorm individually and then come together as a group to compare ideas.

34. I consider it higher praise to call someone

(a) sensible.

(b) imaginative.

35. When I meet people at a party, I am more likely to remember

(a) what they looked like.

(b) what they said about themselves.

36. When I am learning a new subject, I prefer to

(a) stay focused on that subject, learning as much about it as I can.

(b) try to make connections between that subject and related subjects.

37. I am more likely to be considered

(a) outgoing.

(b) reserved.

38. I prefer courses that emphasize

(a) concrete material (facts, data).

(b) abstract material (concepts, theories).

39. For entertainment, I would rather

(a) watch television.

(b) read a book.

40. Some teachers start their lectures with an outline of what they will cover. Such outlines are

(a) somewhat helpful to me.

(b) very helpful to me.

41. The idea of doing homework in groups, with one grade for the entire group,

(a) appeals to me.

(b) does not appeal to me.

42. When I am doing long calculations,

(a) I tend to repeat all my steps and check my work carefully.

(b) I find checking my work tiresome and have to force myself to do it.

43. I tend to picture places I have been

    (a) easily and fairly accurately.

    (b) with difficulty and without much detail.

44. When solving problems in a group, I would be more likely to

    (a) think of the steps in the solution process.

    (b) think of possible consequences or applications of the solution in a wide range of areas.

A.3 Academic history Detail

The department has access to certain variables peratining to student academic history that were pooled into the data. These variables include:

1. High School GPA
2. Math ACT
3. Class Standing
4. Gender
5. Age

APPENDIX B

WRIGHT EXTENSION

Dr. David Wright, of the Brigham Young University Mathematics Department, developed the following theorem and proof.

Let $p_1, p_2 \ldots p_m$ be points in $\mathbb{R}^k$. We say that a point p in $\mathbb{R}^k$ is an *affine* combination of the points $p_1, p_2 \ldots p_m$ if p can be written $p = \sum_{i=1}^{m} \lambda_i p_i$ where $\sum_{i=1}^{m} \lambda_i = 1$. In case the $\lambda_i \geq 0$ we say that p is a *convex* combination of the points. The *convex* hull of a set of points in $\mathbb{R}^n$ is the set of all convex combinations of the points. The *affine* hull of a set of points in $\mathbb{R}^n$ is the set of all affine combinations of the points.

THEOREM: Let $S = \{v_1, v_2, \ldots, v_m, w_1, w_2, \ldots, w_n\}$ be a subset of $\mathbb{R}^k$. Suppose p is an affine combination of the points $v_1, v_2, \ldots, v_m, w_1, w_2, \ldots, w_n$ so that $p = \sum_{i=1}^{m} s_i v_i + \sum_{j=1}^{n} t_j w_j$ where $s_i \geq 0$, $t_i \leq 0$, $\sum_{i=1}^{m} s_i + \sum_{j=1}^{n} t_j = 1$, and $s = \sum_{i=1}^{m} s_i > 1$. Set $v_i' = sv_i + \sum_{j=1}^{n} t_j w_j$ and set $S' = \{v_1', v_2', \ldots, v_m', w_1, w_2, \ldots, w_n\}$. Then p lies the the convex hull of S'. Furthermore the convex hull of S lies in the convex hull of S'.

PROOF. The point p can be written as the convex combination $\sum_{i=1}^{m} \frac{s_i}{s} v_i'$. This sum is equal to $\sum_{i=1}^{m} \frac{s_i}{s} \left(sv_i + \sum_{j=1}^{n} t_j w_j \right) = \sum_{j=1}^{m} s_i v_i + \sum_{j=1}^{m} \frac{s_i}{s} \left(\sum_{j=1}^{n} t_j w_j \right) = \sum_{j=1}^{m} s_i v_i + \sum_{j=1}^{n} t_j w_j$. Now suppose q is a convex combination of S and can be written $q = \sum_{i=1}^{m} a_i v_i + \sum_{j=1}^{n} b_j w_j$ where $a_i \geq 0, b_j \geq 0$ and $\sum_{i=1}^{m} a_i + \sum_{j=1}^{n} b_j = 1$. Then $q = \frac{1}{s} \sum_{i=1}^{m} a_i (v_i' - \sum_{j=1}^{n} t_j w_j) + \sum_{j=1}^{n} b_j w_j$. When this expression for q is expanded, we get a linear combinations of the v_i' and w_j. Since all of the coefficients are non-negative ($-t_j \geq 0$), we only need to show that the sum of the coefficients is 1. Let $t = \sum_{j=1}^{n} t_j$, $a = \sum_{i=1}^{m} a_i$, and $b = \sum_{j=1}^{n} b_j$. The sum of the coefficients in the expression for q is $\frac{1}{s} \sum_{i=1}^{m} a_i - \frac{1}{s} \sum_{i=1}^{m} a_i \sum_{j=1}^{n} t_j + \sum_{j=1}^{n} b_j = \frac{a}{s} - \frac{at}{s} + b = \frac{a(1-t)}{s} + b = \frac{as}{s} + b = a + b = 1.$