2007-12-28

Extending the Information Partition Function: Modeling Interaction Effects in Highly Multivariate, Discrete Data

Paul C. Cannon
Brigham Young University - Provo

EXTENDING THE INFORMATION PARTITION FUNCTION: MODELING INTERACTION EFFECTS IN HIGHLY MULTIVARIATE, DISCRETE DATA

by

Paul C. Cannon

A thesis submitted to the faculty of

Brigham Young University

in partial fulfillment of the requirements for the degree of

Master of Science

Department of Statistics

Brigham Young University

April 2008

BRIGHAM YOUNG UNIVERSITY

GRADUATE COMMITTEE APPROVAL

of a thesis submitted by

Paul C. Cannon

This thesis has been read by each member of the following graduate committee and by majority vote has been found to be satisfactory.

Date	Dr. H. Dennis Tolley, Chair
Date	Dr. Del T. Scott
Date	Dr. Bruce J. Collings

BRIGHAM YOUNG UNIVERSITY

As chair of the candidate's graduate committee, I have read the thesis of Paul C. Cannon in its final form and have found that (1) its format, citations, and bibliographical style are consistent and acceptable and fulfill university and department style requirements; (2) its illustrative materials including figures, tables, and charts are in place; and (3) the final manuscript is satisfactory to the graduate committee and is ready for submission to the university library.

Date

Dr. H. Dennis Tolley
Chair, Graduate Committee

Accepted for the Department

Scott D. Grimshaw
Graduate Coordinator

Accepted for the College

Thomas W. Sederberg
Associate Dean, College of Physical and Mathematical Sciences

ABSTRACT

EXTENDING THE INFORMATION PARTITION FUNCTION:
MODELING INTERACTION EFFECTS IN HIGHLY MULTIVARIATE, DISCRETE DATA

Paul C. Cannon
Department of Statistics
Master of Science

Because of the huge amounts of data made available by the technology boom in the late twentieth century, new methods are required to turn data into usable information. Much of this data is categorical in nature, which makes estimation difficult in highly multivariate settings. In this thesis we review various multivariate statistical methods, discuss various statistical methods of natural language processing (NLP), and discuss a general class of models described by Erosheva (2002) called generalized mixed membership models. We then propose extensions of the information partition function (IPF) derived by Engler (2002), Oliphant (2003), and Tolley (2006) that will allow modeling of discrete, highly multivariate data in linear models. We report results of the modified IPF model on the World Health Organization's Survey on Global Aging (SAGE).

ACKNOWLEDGEMENTS

I would like to first thank Dr. H. Dennis Tolley. Under his tutelage I learned as much about philosophy, God, and Mammon as I did about statistics. Thanks also to James Oliphant who is the mastermind behind the IPF optimization algorithms. I also want to thank The Church of Jesus Christ of Latter-day Saints for providing a University which quickens the soul as well as the mind. Above all, I want to thank my supportive wife whose unfathomable generosity and sacrifice have allowed me to finish this work.

CONTENTS

CHAPTER

TABLES

Table

FIGURES

Figure

1. INTRODUCTION

> The explosive nature of exponential growth means it may only take a quarter of a millennium to go from sending messages on horseback to saturating the matter and energy in our solar system with sublimely intelligent processes. The ongoing expansion of our future superintelligence will then require moving out into the rest of the universe, where we may engineer new universes. (Ray Kurzweil, qtd. in "The Intelligent Universe")

1.1 The Information Age?

It is claimed that we live in the "information age," but this era would be more accurately called the "data age." We are confronted daily with a data deluge that has necessitated the creation of new jobs and special training to manage the overload. Instead of using all available data, researchers often settle for a small, convenient sample for research or decision-making. For example, millions of people use Google every day to search the internet, but few look past the first two or three pages of hits for information on the query. Most have neither time nor resources enough to spend countless hours searching through millions of documents or web pages to glean information, though there is certainly information buried in the millions of unchecked documents.

New methodologies are necessary to transform large amounts of data into information and to lift us out of the data age. When new methods become available to effectively use all available data, advances in science and technology will surge at a potentially unprecedented rate.

1.1.1 Deterrents of the Information Age

One major deterrent of the information age is the vast and largely untapped resource of information contained in textual documents. This huge data resource is

largely unavailable in the statistical decision-making process. This problem has been addressed by computer scientists in machine learning, artificial intelligence (AI), information retrieval, and other fields. New statistical methodologies should be developed in order to use the information contained in textual documents. Natural language processing (NLP) is a scientific area concerned with the modeling and use of natural language, either spoken or contained in text, in computer systems. A "natural" language is a language that is spoken or written by humans for common communication. Though NLP has received increased attention in many fields in the computer sciences, it has not been studied much in statistics.

NLP is actually a special case of categorical data analysis. Identifiability issues are very problematic when categorical data is highly multivariate. For example, if a researcher asked a set of n individuals 18 yes/no questions, the researcher would need to ask at least $2^{18} = 262,144$ individuals if he or she wanted to make any good inference on the population. The categorical nature of much of the data produced by technological advances is a second deterrent to the information age.

1.2 Thesis Outline

In this thesis, we review statistical methods of natural language processing (NLP), discuss various multivariate statistical methods, and discuss a general class of models described by Erosheva (2002) called generalized mixed membership models. We discuss the attributes shared by these models and discuss how they relate to the proposed methodology of this paper. We then propose a modified version of the information partition function (IPF) derived by Engler (2002), Oliphant (2003), and Tolley (2006). This adaptation of the IPF will model highly multivariate discrete data and is an effective way of reducing large amounts of data into a manageable format. The modified IPF will allow modeling of the interaction effects of highly multivariate discrete data and provide a way to more efficiently obtain updated probabilities of

interest at the individual level. This quick updating is crucial for internet companies and marketing applications.

The modified IPF can be used for several different applications. After describing the modifications of the IPF, we demonstrate its use on the World Health Organization's (WHO) Survey on Global Aging (SAGE). This dataset is used to show how the modified IPF can be used when the response variable of the model is continuous and when it is discrete.

For the continuous response case, the modified IPF is used in three ways. The first analysis uses the g_{ik}s from the IPF output as a summary variable of the observational factors in connection with main-effect variables from SAGE to build a linear model for a continuous health score variable. The second analysis estimates the g_{ik}s using only the health-related variables from SAGE, omitting socio-demographic variables, and defines three health sub-populations based on g_{ik} scores. The socio-demographic variables are then used to build a generalized linear model (GLM) to predict the probability of being in one of the health subgroups. This is similar to a propensity score in the statistical literature (Rosenbaum and Rubin 1983). Thirdly, the IPF is used with all of the variables and the g_{ik}s are used to define small health subgroups. These subgroups can be used by policy makers to determine how best to focus resources to move individuals in a particular health subgroup toward a better health subgroup.

The discrete response case demonstrates how the modified IPF is used when the response variable is categorical. The health score variable is discretized based on its quartiles. The IPF is then used to summarize the interaction effects as g_{ik}s, and the g_{ik}s are used with the main effects in a GLM to predict the probability of being in each health score quartile.

In Chapter 2 we discuss several multivariate statistical methods with an emphasis on latent variable models. Some developments from the 1960s in modeling

interaction effects in linear models with categorical predictors are also discussed. In Chapter 3 we provide an overview of natural language processing and current techniques. In Chapter 4 we discuss the general framework of and review several types of mixed-membership models, including the grade of membership (GoM) model and its relation to latent class models and latent Dirichlet allocation (LDA). We also review the hierarchical Bayesian structure of the mixed-membership model (HBMMM) established in Airoldi et al. (2006) and discuss its representation of LDA and GoM. We also review the Airoldi et al. (2006) formulation of a semi-parametric approach to the HBMMM based on the Dirichlet process prior.

In Chapter 5 we review the IPF as established by Engler (2002), Oliphant (2003) and Tolley (2006). In Chapter 6 we derive the modifications of the IPF and show how it relates to mixed membership models and modeling interaction effects in generalized linear models. In Chapters 7 and 8 we demonstrate the model with the analysis of the SAGE data for the continuous and discrete cases, respectively.

2. MULTIVARIATE METHODS

2.1 Introduction

In this chapter we review several methods of multivariate analysis that relate to the proposed IPF model. In Section 1 we discuss two approaches to multivariate analysis in general, then in subsequent sections elaborate on latent variable models, finite mixture models, and hidden Markov models. We also discuss some relevant work from the 1960s and '70s about partitioning interaction effects in linear models. The multivariate techniques relate to the theory and motivation of the IPF model, and the interaction effect modeling relates to how the IPF model and its modifications are used in data analysis.

2.2 Multivariate Analysis

Several techniques have been explored in multivariate statistics to reduce an overwhelmingly large data space into something manageable. There are several approaches to this problem that have been developed in various scientific areas. Principal component analysis, for example, maximizes the variance of linear combinations of the variables in a dataset based on the decomposition of a data matrix, $\mathbf{X}$. This is an example of a type of analysis that decomposes the data into fewer dimensions based on eigenvalues.

Another approach is to model the observed data as a manifestation of unobserved latent variables. That is, there is an underlying structure which cannot be observed directly, but which provides the probability model for what is observed. For example, in factor analysis the observed variables are modeled as linear combinations of latent variables and are used to account for the correlation structure among the

Table 2.1: This table shows the different types of latent variable models that exist for different types of manifest variables and the assumed type of latent variable.

		Latent Variable	
		Continuous	Categorical
Manifest Variable	Continuous	Factor Analysis	Latent Profile Analysis
	Categorical	Latent Trait Analysis	Latent Class Analysis

manifest variables (Rencher 2002). This method takes advantage of the correlation structure in the $\mathbf{X}$ matrix. These latent variable models relate to the IPF, and several of these models are discussed in the following sections.

2.3 Latent Variable Models

A latent variable model is characterized by the assumption that what is observed or measured on an individual is the manifestation of a set of latent variables that cannot be measured (Bartholomew and Knott 1999; Loehlin 1998; Rencher 2002). The type of models that fit into this category are determined by the type of observed variables, called manifest variables, and the type of the assumed latent structure. For example, a different model would be fit if the manifest variables were continuous and the latent variables were assumed to be discrete than if the manifest variables were discrete and the latent variables were assumed continuous. Table 2.1 shows which models are used under different circumstances.

2.4 Finite Mixture Models

The latent profile model is actually part of a broader class of models called mixture models. The form of the finite mixture model is formally defined as

$$p(y) = \sum_{i=1}^{k} \eta_i p(y|\theta), \tag{2.1}$$

where $p(y|\theta)$ is a probability distribution and the η_is are constrained to be greater than zero and to sum to unity (Fruhwirth-Schnatter 2006). The θs parametrize the distribution of the latent variables. This means that an individual, i, belongs to the kth population with different probability distributions with probability η_k. A finite mixture model is one that has a finite number of mixture components. Cluster analysis is one of the most common uses of the finite mixture model. Many clustering techniques are based on the decomposition of the overall variation in the data around the mean into within and between sums of squares. The usual criterion for clustering is to minimize the within sums of squares error for a specified number of groups. The finite mixture model can then serve as a model-based clustering method (Fruhwirth-Schnatter 2006; Loehlin 1998).

2.5 Hidden Markov Models

A hidden Markov model (HMM) is defined by a time- or space-dependent process that consists of an unobservable state variable that determines the probability model of the observed variables. It is a regular Markov process where the states are not observed or observable, and the observed outcome is conditionally dependent on this latent state. More formally, an HMM consists of an underlying Markov chain, X_t, which governs the observed stochastic process, Y_t, over time $t \geq 0$. For example, Y_t might be normally distributed with mean and variance determined by the state of X_t (Cappe et al. 2005). The conditional distribution of Y_t given X_t is similar to the finite mixture models discussed above, but it is also a function of time (Cappe et al. 2005; Elliot et al. 1995).

We now consider the Fruhwirth-Schnatter (2006) development of the HMM. Consider the properties of the state variable X_t. This is a stationary Markov process that can be defined by a transition matrix ξ, where

$$\xi_{jk} = Pr(X_t = k|X_{t-1} = j).$$

If X_t is an ergodic, irreducible, aperiodic Markov chain, its stationary distribution is defined by a set of probabilities $\eta = \eta_1, ..., \eta_K$, where

$$Pr(X_0 = k|\xi) = \eta_k.$$

Let

$$Y_t|X_t = k \sim T(\theta_k),$$

where $T(\theta_k)$ is a parametric distribution family and has density $p(y_t|\theta_k)$. Thus, the marginal distribution of Y_t is

$$p(y_t|\vartheta) = \sum_{k=1}^{K} p(y_t|X_t = k, \vartheta)Pr(X_t = k|\vartheta).$$

The unconditional distribution of Y_t is then defined as a finite mixture of $T(\theta)$ distributions with ergodic probabilities η_k. The HMM is then defined as

$$p(y_t|\vartheta) = \sum_{k=1}^{k} p(y_t|\theta_k)\eta_k. \tag{2.2}$$

This looks very similar to the finite mixture model 2.1, but the η_ks are ergodic probabilities estimated as the components of an unobserved process. For HMM estimation techniques, see the chapters on estimation found in Elliot et al. (1995), Cappe et al. (2005), and Fruhwirth-Schnatter (2006).

A classic example of an HMM is the stock market, which is often described as a bull or a bear market. The stock market could be modeled as a two-state HMM where Wall Street investors would use different trading strategies depending on whether they were in a bear or a bull market. This type of model is often called a Markov-switching model where the conditional distribution of Y_{t+1} depends not only on X_{t+1}, but also previous values of Y (Cappe et al. 2005). While Markov-switching models can be set up in the framework of HMMs they are often treated as a separate class of models.

2.6 Modeling Interaction Effects

The models discussed in the prior sections of this chapter deal primarily with multivariate models that relate to the theory behind the IPF model. In this section we review research that relates to how the modified IPF will be used to model interaction effects.

Mandel (1969) describes a method of partitioning interaction in an analysis of variance (ANOVA) situation when the response term is quantitative. Consider an ANOVA with two factors. The typical ANOVA model is written as

$$y_{ij} = \mu + \alpha_i + \beta_j + \eta_{ij},$$

where μ is an overall mean, α_i is the effect of factor A, β_j is the effect of factor B, and η_{ij} represents the interaction between the two treatments. This reduces to the additive model if η_{ij} is assumed to be a random variable with mean zero and standard deviation σ. Otherwise, according to Mandel, the interaction term can be partitioned into a multiplicative component as

$$\eta_{ij} = \theta u_i v_j + \theta' u'_i v'_j + \theta'' u''_i v''_j + ... + \epsilon_{ij}$$

where ϵ_{ij} is a random variable with mean zero and standard deviation σ.

The θs in this interaction term are estimated using the least squares estimation of the residuals

$$r_{ij} = y_{ij} - \hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j,$$

which is a vector, $\mathbf{r}$. It turns out that under certain constraints of the u_i and v_j, the estimates of $\theta', \theta'', ...$ are the eigenvalues of $\mathbf{S} = \mathbf{r}\mathbf{r}'$, and u is the associated eigenvector (Mandel 1969). The vector, v, is shown to be the eigenvector of $\mathbf{r}'\mathbf{r}$. The interaction terms are partitioned according to eigenvalues and eigenvectors. Gollob (1968) develops a similar method of decomposing the interaction effects into multiplicative components using principal components. Most of the work on this type of interaction

modeling was done in the late '60s and early '70s and has not been developed much since. The extensions of the IPF discussed in this thesis develop a new framework to model interaction effects.

3. NATURAL LANGUAGE PROCESSING

3.1 Introduction

In the last chapter we discussed multivariate statistical methods in general. In this chapter we demonstrate how many of these methods have been specifically applied to the field of Natural Language processing. A "natural" language is defined in the literature as a written or spoken language used for common communication between humans. Natural Language Processing (NLP) consists of several steps for understanding texts or spoken words: (1) morphological analysis, (2) syntactic analysis, (3) semantic analysis, and (4) discourse integration. Statistical NLP incorporates several probabilistic methods and models; each of these processes are used to obtain better understanding of natural languages. Though there is not an integrated system that performs all of these functions simultaneously, there have been advances in statistical NLP that are helpful in each of the four aspects of NLP. These advances provide insight into using textual documents to enrich statistical inference. This chapter provides an overview of current statistical methods used in NLP for each of the functions of NLP. In Section 1 we discuss the steps of NLP, and in Section 2 we discuss current research in statistical NLP. For a comprehensive treatment of statistical NLP methods, see Manning (1999).

3.2 Functions of NLP

3.2.1 Morphological Analysis

Morphological analysis is designed to dissect individual words to help the researcher derive meaning from those words. The primary use of morphological analysis in NLP is word-sense disambiguation. This is the process of discovering the contex-

tual meaning of individual words that have multiple meanings. For example, the word "bank" could refer to the land adjacent to a river or a financial institution, among other things (Manning 1999). While there have been several approaches to word sense disambiguation, Hidden Markov Models (HMM) have become the most widely used method and seem to outperform most other methods.

HMMs use the previous N words to predict the meaning of the $(N+1)$th word. This is known as an N-Gram model, and such models are generally trained on a corpus of text. A corpus generally includes thousands of published documents that represent the type of text to be analyzed. Some common corpora include the Brown corpus and the Wall Street Journal corpus.

3.2.2 Syntactic Analysis

Syntactic analysis considers sequences of words and how they relate to each other within a sentence. This breaks a sentence into parts of speech and extracts whatever meaning the word order contains. There are several probabilistic parsing techniques, and there is some debate as to which is the optimal method. Optimality is usually determined based on the individual problem. Probabilistic Context Free Grammars (PCFG) are commonly used because they are simple and they lend themselves nicely to grammar tree structures (Manning 1999). PCFGs are also trained from a corpus. The idea is to find the most probable sequence of words. The phrase "The man sold the dog biscuits" has several possible meanings. A man could be (1) selling a dog named Biscuits, (2) selling biscuits to a dog, or (3) selling dog biscuits. PCFG's would use the probabilities estimated from a corpus of one word being followed by another (or possibly an N-gram model) to find the most likely parse. In our dog biscuit example this process would discover the most likely direct object (what is being sold).

3.2.3 Semantic Analysis

Semantic analysis is the key to using information contained in textual documents in statistical analysis because it deals primarily with the meaning of a text. Semantic analysis analyzes the structures generated by the syntactic processing and derives the meaning of the sentence based on its structure. This is considered the holy grail of NLP. It uses the syntactic and morphological analysis to determine the most likely meaning of a sentence. Many models assume that there is a hidden or latent semantic structure from which the words are generated. This assumption leads to a latent variable approach to modeling semantics, borrowing heavily from models developed in the social sciences and psychology, such as factor analysis and structural equations. One such model is latent semantic analysis (LSA).

LSA boils a text down into its most fundamental components. It first strips a text of all non-contextual words (eliminating words like "the," "and," "of," etc.) The meaningful words left over are then broken down into semantic root meanings, thus incorporating morphological structure. Frequencies are often assigned weights determined by frequency of occurrence in the document and frequency of occurrence in the entire corpus. The most common weight is the tf-idf weight, which stands for term frequency–inversed document frequency. The term frequency is calculated as

$$tf = \frac{n_i}{\sum_k n_k}, \tag{3.1}$$

where n_i is the number of occurrences of word i in a document and the denominator is the total number of words in the document. The inverse document frequency is calculated as

$$idf = \log \frac{|D|}{|d_i \supset t_i|}, \tag{3.2}$$

where $|D|$ is the total number of documents and $|d_i \supset t_i|$ is the number of documents where word t_i appears. The term weight tf-idf is calculated $tf - idf = tf * idf$ for each term in the document. Thus, word counts are weighted by the relative

frequency within the document and within the corpus. The weighted frequencies of morphologically similar words in a contingency table are then used to group similar documents using singular value decomposition (SVD) or factor analysis. Though SVD and factor analysis work remarkably well, both methods require the assumption of a Gaussian error structure, which is incorrect for count data. LSA can be thought of as a variable reduction or classification method based on the underlying meaning of a document, assigning each document a unique, latent "meaning" variable.

Hofmann (1999) proposes a probabilistic LSA (PLSA) based on the likelihood principle, which is more appropriate for discrete data. PLSA models each word as a sample from a mixture distribution where each mixture component represents a topic. Mixture distributions in general define a probability distribution of the form

$$p_X(x) = \sum_{k=1}^{K} a_k h(x|\lambda_k), \tag{3.3}$$

with the constraint that $a_k \geq 0$ for all $k = 1, ..., K$, and $\sum_k a_k = 1$, where K is the number of components in the mixture and $h(x|\lambda_k)$ is a probability distribution parameterized by λ_k. For PLSA, the a_ks represent the proportion of the document that belongs to a single topic. This means that a single document belongs to one or more topics.

PLSA falls short as a complete language model because it only offers a probabilistic model at the document level. PLSA is also prone to overfitting the training corpus because the parameter estimates of the mixture distributions are directly linked to the corpus. Latent Dirichlet allocation (LDA), established by Blei et al. (2003), is a generalization of PLSA. It is a model that extends topic sharing across corpora and treats the mixture components as a random variable. This ameliorates the problem of overfitting.

It should also be noted that both LSA and PLSA assume the number of latent variables, K, is fixed and must be decided on before the analysis is performed. Probabilistic approaches have been found to perform at least as well as regular LSA

and have the benefit of a unified mathematical framework. De Freitas and Barnard (2000) describe a Bayesian LSA that could potentially allow for K to be estimated by the data. Each of these approaches, LSA, PLSA, and LDA, is a "bag-of-words" method which does not exploit the phrase structure in documents.

3.2.4 Discourse Integration

Discourse integration describes how a sentence or paragraph is understood in the context of a document or how a collection of documents relates to another collection. Various types of hierarchical models are often used for discourse integration. LDA is essentially a hierarchical mixture model and can add another layer of the hierarchy to incorporate discourse integration.

3.3 Current Research in NLP

NLP is generally used to provide machines with the ability to use or understand natural language. Though various statistical methods are used for different components of NLP, there is not yet a unified model that will incorporate all of the components of NLP. Recent research has made progress in the direction of a unified language model. Wang et al. (2002) propose a model based on a latent maximum entropy principle, which combines the syntactic N-gram model with latent semantic analysis, allowing hidden features to be captured in the model. Erosheva (2002) shows that probabilistic LSA is a special case of a more general class of models called mixed-membership models. Mixed-membership models are extended to the hierarchical Bayesian framework in Airoldi et al. (2006). Both Erosheva (2002) and Airoldi (2006) show that LDA also fits into the mixed-membership framework. This general form of latent variable models can easily be extended to incorporate a hierarchical structure into documents and corpora. Instead of assigning a document a latent "meaning" variable, documents can have partial membership in topics, and topics

can be shared across corpora. This methodology incorporates discourse integration and semantic analysis. The GoM model and the Rasch model also fit into this general framework (Erosheva 2006). The Rasch model is a variation on the latent class model commonly used in psychology. In the next chapter we will discuss this general class of models.

NLP models can easily be broken down into main topics because documents generally have a stated purpose. By extending the IPF to model interaction effects, we might better understand the more subtle nuances of the information contained in documents. The rest of this thesis focuses on how to extend the IPF to model these subtleties in categorical data generally, but the tools can be specifically applied to NLP.

4. GENERALIZED MIXED-MEMBERSHIP MODELS

4.1 Review

The general framework of mixed-membership models is established by Erosheva (2002); Erosheva, Fienberg, and Lafferty (2004); and Airoldi et al. (2006), and is a generalization of the finite mixed-membership models discussed in Section 2.3. Mixed-membership models perform fuzzy or soft clustering. In many applications it is unrealistic to assume that an observation belongs exclusively to a single cluster or subpopulation. A scientific publication, for instance, might simultaneously contain relevant information regarding chemistry, biology, and physics. This would mean that the publication has partial membership in each cluster (chemistry, biology, and physics). In the following section we will establish the general framework of the mixed-membership model based on Erosheva et al. (2004) and Erosheva (2002).

The general formulation of mixed-membership models is based on assumptions at the population level, the subject level, the latent variable level, and the sampling scheme. The following assumptions are taken from Erosheva (2002) and Airoldi et al. (2006).

4.1.1 Population-Level Assumptions

At the population level it is assumed that there is a latent structure responsible for the $\mathbf{J}$ manifest variables observed for each individual i, where $i = 1, 2, ..., I$. Each of the k subpopulations, where $k = 1, 2, ..., K$, is characterized by a probability distribution $f(x_j|\theta_{jk})$, where x_j are the manifest variables for an individual and θ_{kj} is a vector of relevant parameters. The observations are assumed to be exchangeable, or conditionally independent, given class membership (Erosheva 2006). This means that, given the class, individuals are independent.

4.1.2 Subject-Level Assumptions

It is assumed at the subject level that each individual has a membership vector, $\mathbf{g} = (g_1, g_2, ..., g_K)$, which has length equal to the number of latent variables. Each component of $\mathbf{g}$ represents the degree of membership in each of the latent groups. The probability distribution of x_j given the membership vector, $\mathbf{g}$, is

$$Pr(x_j|\mathbf{g}) = \sum_{k=1}^{K} g_k f(x_j|\theta_{jk}). \tag{4.1}$$

The x_j are assumed to be conditionally independent of one another given $\mathbf{g}$.

4.1.3 Latent Variable-Level Assumptions

At the latent variable level the mixed-membership scores $\mathbf{g}$ can be assumed to be fixed or random effects. For the fixed-effects mixed-membership model the conditional probability of observing x_j is

$$Pr(x_j|\mathbf{g}, \theta) = \sum_{k=1}^{K} g_k f(x_j|\theta_{jk}), \tag{4.2}$$

where the g_k are modeled as fixed effects.

In some cases it is reasonable to assume that the mixed-membership scores $\mathbf{g}$ are random realizations from a distribution D parameterized by α. If this is the case, the GMMM is a mixed-effects model with random effect $\mathbf{g}$. For the mixed-effects mixed-membership model,

$$Pr(x_j|\theta, \alpha) = \int \left(\sum_{k=1}^{K} g_k f(x_j|\theta_{jk}) \right) dD_\alpha(\mathbf{g}). \tag{4.3}$$

All types of mixed membership models including GoM, LDA, PLSA, and others differ only in what assumptions are made at each of these levels.

4.1.4 Sampling Scheme

In some instances it is possible that there are multiple independent replications of the J manifest variables for an individual. The sampling scheme denotes

$\{x_1^{(r)}, ..., x_J^{(r)}\}_{r=1}^R$ as R replications of **J** variables of a subject. For the random effects model the conditional probability of such a set of R replications is

$$Pr(\{x_1^{(r)}, ..., x_J^{(r)}\}_{r=1}^R | \alpha, \theta) = \int \left(\prod_{j=1}^{J} \prod_{r=1}^{R} \sum_{k=1}^{K} g_k f(x_j^{(r)} | \theta_{jk}) \right) dD_\alpha(\mathbf{g}). \tag{4.4}$$

It should also be noted that it is not necessary for **J** to be the same across subjects, nor for **R** to be the same across observed variables.

4.2 Hierarchical Bayesian Mixed-Membership Models

Erosheva (2002, 2003) and Airoldi et al. (2006) detail the hierarchical Bayesian mixed-membership model (HBMMM) representation of the GoM and LDA models. PLSA is also discussed in Erosheva (2002). The hierarchical Bayesian model requires specification of $p(\mathbf{x}|\mathbf{g}, \lambda)$, prior distributions for λ and $\mathbf{g}$, and hyper-prior distributions (depending on whether or not one is fitting the fixed effects or random effects model). The assumptions concerning the nature of membership scores, whether they are fixed or random, must be defined in this step. Because of the constraints on $\mathbf{g}$, the Dirichlet distribution is a natural prior choice; however, according to Erosheva (2002) if complex dependencies exist between latent groups the Dirichlet may be a poor choice.

Airoldi et al. (2006) discuss several strategies for model specification. One of the main challenges of mixed-membership models is determining the number of latent groups, K. If a researcher has strong prior belief and a strong understanding of the underlying structure of a population of interest K can be chosen before the analysis. However, in most unsupervised learning scenarios there is little knowledge of the number of latent factors. Airoldi et al. (2006) suggest using a Dirichlet process prior on the number of latent groups. This semi-parametric Bayesian approach allows for K to be estimated from the data.

4.3 The GoM Model

The grade of membership model was first introduced by Woodbury (1974). Subsequent developments were made by Manton et al. (1994) which established the GoM model in the framework of fuzzy set theory. Their model assumes that the population of interest can be modeled as a set of extreme profiles, often called "pure types." In their model, the g_{ik}s represent the degree of membership of each individual i to the kth pure type. The g_{ik} scores vary between 0 and 1 over the k groups and $\sum_{k=1}^{k} g_{ik} = 1$ for all individuals. If all elements of the set of g_{ik}s is either 0 or 1, then this reduces to a crisp cluster analysis. The likelihood of the GoM model is

$$L = \prod_i \prod_j \prod_l \left(\sum_k g_{ik} \lambda_{kjl} \right)^{y_{ijl}}, \tag{4.5}$$

where both g_{ik} and λ_{kjl} are constrained to be greater than zero and to sum to unity. This is noticeably similar to the fixed-effects version of the mixed-membership model in equation 2.2.

Potthoff, Manton, and Woodbury (2000) generalize the GoM model by assuming random membership scores generated from a Dirichlet distribution. Manton et al. (1994) describe the unconditional likelihood of the GoM model for random membership scores as

$$L = \int \prod_i \prod_j \prod_l \left(\sum_k g_{ik} \lambda_{kjl} \right)^{y_{ijl}} dD_\alpha(\mathbf{g}), \tag{4.6}$$

integrating over the random g_{ik} scores. This is identical to what Erosheva (2002) refers to as the "marginal" likelihood.

As a cautionary note, there is a major difference between the GoM model and other clustering methods. Though the g_{ik}s are bound between 0 and 1 and constrained to sum to unity, they are not to be interpreted as probabilities. The g_{ik}s do not represent the probability of membership in one of K groups and are not to be interpreted in the same way as discriminant function scores or other clustering

methods. The g_{ik}s are interpreted as the proportion of membership in each group. For example, an elderly individual might be neither totally incapacitated nor totally independent. That individual would have part membership in both groups if he or she had difficulty with certain activities.

Erosheva (2002) develops the GoM model as a generalization of a latent trait model. This is a subtle difference from the Woodbury (1974) and Manton et al. (1994) formulations. Latent trait models are a type of latent class models that are widely used in psychology and the social sciences. These models assume that there are hidden, unobservable variables, such as intelligence or personality, that cannot be measured directly. Erosheva (2006) proves the equivalence of a constrained latent class model and the GoM model.

The modified IPF relates to the generalized mixed-membership models of Erosheva (2002). Oliphant (2003) shows that the grade of membership (GoM) is a linear approximation of the IPF and demonstrates the advantages of the IPF over the GoM in analyzing discrete multivariate data. The generalized IPF is shown to be a category of models for which mixed-membership models are linear approximations and can be used for the same type of analysis as the generalized mixed-membership models.

4.4 PLSA and LDA

Erosheva (2002) shows how probabilistic latent semantic analysis (PLSA) and latent Dirichlet allocation (LDA) are related and how they fit in the mixed-membership framework. The joint likelihood of PLSA as derived by Hofmann (1999) is

$$L = \prod_i \prod_m \left(\sum_k g_{ik} \lambda_{km} \right)^{y_{im}}, \tag{4.7}$$

which is essentially equivalent to the GoM likelihood in Equation 4.5. The LDA likelihood derived by Blei et al. (2003) with a Dirichlet prior is

$$L = \int \prod_i \prod_m \left(\sum_k g_{ik} \lambda_{km} \right)^{y_{im}} dD_\alpha(\mathbf{g}), \tag{4.8}$$

which is similar to the marginal GoM likelihood and the mixed-effects mixed-membership model in Equation 4.3. Thus, according to Erosheva (2002), the LDA model is a mixed-effects representation of Hofmann's PLSA.

5. THE INFORMATION PARTITION FUNCTION

The information partition function (IPF) has many similarities with a few of the models discussed already in this thesis, but it has a very different motivation and theory. Most of its theory comes from statistical mechanics and information theory and is based on maximizing a quantity called entropy. In this chapter we discuss entropy and develop the motivation of the IPF.

5.1 Entropy

In 1948, Claude Shannon revolutionized communication theory in a publication for Bell Labs. *A Mathematical Theory of Communication* established a unified mathematical framework for information theory. Shannon (1948) established a quantity called "entropy" defined as

$$H = -\sum_i p_i log(p_i), \tag{5.1}$$

where $p_i \equiv P\{X = x_i\}$. It is called entropy because of its relation to the thermodynamic quantity of the same name in statistical mechanics. Shannon established entropy as a measure of the uncertainty of a random variable in an information system. Cover and Thomas (1991) provide a comprehensive introduction to information theory and its general use in statistical inference.

5.1.1 Newton's Second Law

In the first part of the twentieth century, physicists were struggling to apply Newtonian physics to the motion of gases in contained systems and other subatomic behavior. On such a micro level it was necessary to find a way to describe these systems without ever knowing the exact path of particles. Quantum physicists had accepted probability as a fact of subatomic particles and incorporated it into their physical theories as if it were a necessary truth.

In 1957 Edwin T. Jaynes used Shannon's information theory to establish statistical mechanics as inferential science as opposed to a physical theory. Instead of assuming that probability was part of the physical world, Jaynes established probability in mechanics as a measure of our limited state of knowledge. Physicists should therefore find the probability distribution of the path of particles that imposes the fewest assumptions about the system as a whole. Jaynes (1957) shows that maximizing entropy leads to the least biased estimator given our current state of knowledge. In his words it is "maximally noncommittal with regard to missing information." Entropy, according to Jaynes, is "a unique, unambiguous criterion for the amount of uncertainty represented by a discrete probability distribution." Thus, finding the probability distribution that maximizes entropy is the proper way to describe thermodynamic systems.

5.1.2 The Partition Function

The partition function is a formula physicists devised to analyze thermodynamic systems (Oliphant 2003). It is defined as

$$Z = \sum_{s} exp\left(\frac{-E_s(V,N)}{\mathbf{k}T}\right) \tag{5.2}$$

and describes a link between the micro and macro levels. $E_s(V,N)$ represents the energy of macro-state $\mathbf{s}$ with volume, V, and number of particles, N. T is the temperature of the system and $\mathbf{k}$ is the Boltzmann constant. Equation 5.2 is essential in deriving the information partition function.

5.2 The Information Partition Function

There are multiple ways to derive the information partition function (IPF). In this section we review the approach followed by Oliphant (2003) and the approach followed by Tolley (2006) to derive the IPF.

5.2.1 Statistical Mechanics Approach

Consider a dataset containing a J-dimensional contingency table where the rows are all possible combinations of responses to J questions. This is the same structure as the data for the GoM model. Let $l = (l_1, l_2, ..., l_J)^t$ denote a vector of possible outcomes such that l indexes the table of cells. Entropy is then defined as

$$H = -\sum_i \sum_l p_{il} log(p_{il}) \tag{5.3}$$

and is considered to be the information contained in the physical system (i.e. gas in a container). Entropy adds across i because of independence across individuals. The next step is to maximize entropy under the constraints

$$\sum_l p_{il} = 1 \text{ and} \tag{5.4}$$

$$\sum_i \sum_l p_{il} g_{ik} \sum_j w_{kjl_j} = E_k, \tag{5.5}$$

where w_{kjl_j} represents the distribution of a fixed amount of energy E of type k through all the J elements of l. Equation 5.5 is known as the energy constraint and is used in statistical mechanics in the equipartition theorem.

Using Lagrange multipliers λ_k for the energy constraint (5.5), $k = 1, ..., K$ and $\mu_i, i = 1, ..., n$ for constraint (5.4), the constrained Lagrange equation is expressed as

$$\begin{aligned} L_g = &-\sum_i \sum_l p_{il} log(p_{il}) + \\ &\sum_i \mu_i \left(\sum_l p_{il} - 1 \right) + \\ &\sum_k \lambda_k \left(\sum_i \sum_l p_{il} g_{ik} \sum_j w_{kjl_j} - E_k \right), \end{aligned} \tag{5.6}$$

which is maximized by finding the gradient of the L_g, ∇L_g, and setting it equal to zero. This generates a system of equations with $IL + I + K$ equations and the same number of unknowns. The solution to this system of equations with respect to p_{il} is

$$p_{il} = \prod_i exp(-\sum_k g_{ik} \lambda_k w_{kjl_j}). \tag{5.7}$$

Now, define $\lambda_{kjl_j} = \lambda_k w_{kjl_j}$ as the posterior distribution of energy. Using the multinomial identity, $p_{il} = \prod_j p_{ijl}$, the likelihood is

$$L = \prod_i \prod_j \prod_l exp(-\sum_k g_{ik}\lambda_{kjl_j})^{y_{ijl}} \tag{5.8}$$

with constraints

$$\begin{aligned} \sum_l exp(-\sum_k g_{ik}\lambda_{kjl_j}) &= 1 \\ g_{ik} &\geq 0 \\ \sum_k g_{ik} &= 1. \end{aligned} \tag{5.9}$$

This is the information partition function as derived by Oliphant (2003). This likelihood resembles a cross between the partition function (5.2) and the GoM likelihood (4.5). Oliphant then shows that the GoM model is a linear approximation of the IPF with an opposite slope using the MacLaurin expansion.

5.2.2 Intrinsic Data Model Derivation of the IPF

The second approach to the derivation of the IPF comes from Tolley (2006). Consider the setup as a questionnaire of J questions given to N individuals. Each question has a finite number, L_j, of possible answers. This questionnaire paradigm corresponds to the statistical mechanics derivation, where N is the number of particles in a contained system and L_j is the number of degrees of freedom of each of the particles. Note that "degrees of freedom" here indicates the complete description of particles in a microstate system, not the statistical quantity.

For the questionnaire example, j is the index of questions, and l is the answer to question j; and let i indicate the individual responding to the set of questions. X_{ijl} is the random variable of answers of the j^{th} question for individual i. $X_{ijl} = 1$ if the response to question j is l for individual i, $X_{ikl} = 0$ if the response is otherwise.

Individuals are described by their answer profiles; $l = (l_1, ..., l_J)$ and $X_{il} = 1$ if individual i has profile l. $\prod_j L_j$ is the number of possible response profiles of a single

individual. Let p_{il} be the probability that individual i has profile l and let p_{ijl} be the marginal probability of individual i answering question j with answer l; thus,

$$p_{ijl} = \sum_{l:l_j=l} p_{il}. \tag{5.10}$$

In the case that the individuals are randomly selected, the probability model becomes a multinomial distribution with $\prod_j L_j$ cells.

Suppes and Zanotti (1981) demonstrate that if a joint probability distribution exists for the X_{ijl} for all i, $j = 1, ..., J$, and $l = 1, ..., L$, then there exists a discrete, finite-state random variable denoted by Z such that, conditional on Z, the answers to the J questions are independent within an individual. In other words, for two individuals with the same value of Z, we would get the same answers, up to random noise, from both individuals responding to the questions. Z captures all of the information about the two individuals up to random noise. From the Suppes and Zanotti result, the Z variables are the set of intrinsic data that maximizes entropy because it contains all information about the data without extraneous assumptions. It is possible to choose a subset of these intrinsic variables in such a way that the ignored information is a marginalization that, if it satisfies the Cox axioms (Jaynes 2003), is a probability distribution (Tolley 2006). The following definitions are required to determine Z:

$$\begin{aligned}
K &= \text{ Number of the levels of Z,} \\
\pi_{ik} &= Pr(Z = k), k = 1, ..., K, \\
\omega_{ikjl} &= Pr(Z = k|X_{ijl} = 1), \\
\gamma_{ikjl} &= Pr(X_{ijl} = 1|Z = k), \\
\lambda_{ik} &= \text{ Lagrange multiplier for each } k, k = 1, ..., K, \text{and} \\
\mu_i &= \text{ Lagrange multiplier for the sum of the profile probabilities.}
\end{aligned}$$

With these definitions, Tolley (2006) formulated a family of probability models

that represent the uncertainty due to model choice. There is uncertainty associated with the choice of probability models, but the distribution that maximizes entropy defined in Equation 5.1 is the model that uses the fewest assumptions possible on the probability structure. Recall that entropy is a measure of uncertainty and, according to Jaynes (2003), models that allow for little uncertainty limit the ability of the data to speak for itself. Thus, the probability model that must be chosen is the one which maximizes entropy with the constraint that the π_{ik} are fixed for all k. Tolley (2006) provides the following lemma, which examines what it means to hold the π_{ik} constant with regard to p_{il}.

Lemma 1: *Under the conditions above, if the random variable Z exists and J is fixed, holding* π_{ik} *fixed for k=1,...,K is equivalent to holding*

$$\sum_{j=1}^{J}\sum_{l=1}^{L} p_{ijl}\omega_{ikjl} = C, \tag{5.11}$$

where C is some constant value.

Entropy is then maximized under the constraints in Equations 5.10 and 5.11 using Lagrange multipliers. The system of equations derived from differentiating the Lagrange representation with respect to each p_{il} is of the form

$$\frac{\partial}{\partial p_{il}}\left[\sum_{i=1}^{N}\sum_{l} p_{il}\ln(p_{il}) + \mu_i\left(\sum_{l} p_{il} - 1\right) + \left(\sum_{k=1}^{K}\lambda_{ik}\sum_{j=1}^{J}\sum_{l=1}^{L_j}\sum_{l:l_j=l} p_{il}\omega_{ikjl} - C\right)\right] = 0 \tag{5.12}$$

for all i and l.

Solving for p_{il} gives

$$p_{il} = \exp\left(-1 - \mu_i - \sum_{j=1}^{J}\sum_{k=1}^{K}\lambda_{ik}\omega_{ikjl_j}\right). \tag{5.13}$$

In this expression, μ_i is the Lagrange multiplier for the constraint in Equation 5.10, and λ_{ik} is the Lagrange multiplier for the constraints associated with Equation 5.11. The λ_{ik} are the realizations of the random variable Z for individual i. Equation (5.13) is used to parameterize the likelihood, which provides the basis for estimating

realizations of Z. This is equivalent to the information partition function derived by Oliphant (2003) up to a normalizing constant.

Because the IPF is derived from maximizing entropy it is a maximally non-committal model for discrete, multivariate data. This is a desirable quality, especially in an unsupervised learning language model. The IPF as developed by Tolley (2006), Oliphant (2003), and Engler (2002) has huge potential to solve many problems and overcome the weaknesses of the methods described in Chapters 2, 3, and 4. It has already been shown to outperform GoM type models by Oliphant (2003), but there is still room for improvement. In the next chapter we propose three crucial modifications of the IPF that may make it an even more powerful tool in discrete data analysis and statistical NLP. These changes will be a step forward in developing methodologies that will help manage the data deluge.

6. PROPOSED IPF MODIFICATIONS

6.1 Introduction

Though the IPF model has been shown to be extremely versatile and powerful, it still has room for improvement. In this chapter we propose two significant modifications to the IPF that make this tool more efficient and adaptable by (1) initializing the IPF by using a non-parametric clustering algorithm and (2) overcoming the dominance of main effects by removing a few of the most influential variables prior to fitting the IPF. The IPF will be used to reduce large amounts of data into a manageable format, allowing many of the the observational factor variables of highly multivariate categorical data to be used in predictive modeling. This will provide a way to get updated probabilities of interest at the individual level more quickly. In this chapter we develop the proposed modifications of the IPF.

6.2 Modifications

One of the deficiencies of the original IPF algorithm is that it appears to be overwhelmed by main effects. For example, in modeling the voting behavior of the 109th senatorial congress, the IPF easily divides Democrats and Republicans, as seen in Figure 6.1. This is relatively uninteresting for the most part, though it does highlight those who are truly moderate.

The goal of modifying the IPF is to expand its use in linear models with highly multivariate categorical data. This provides a way to use all of the variables in a way that might not be possible with other standard methods when the data are sparse. The IPF is modified in two major ways: (1) by the initialization of the algorithm and (2) through use of the IPF as a data reduction method to model interaction effects.

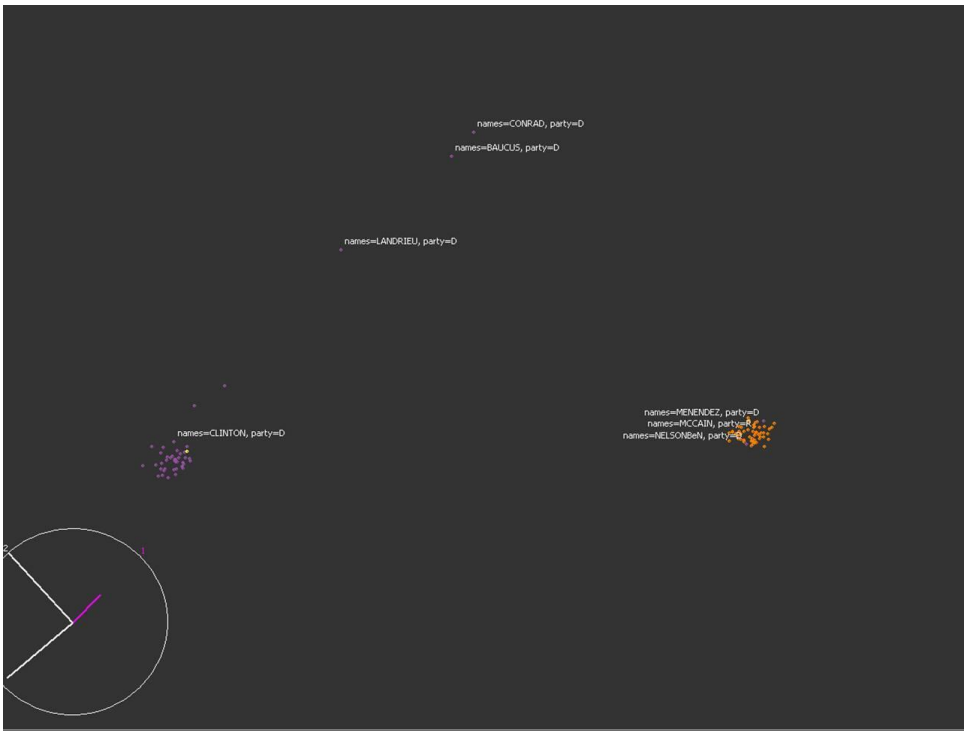


Figure 6.1: Regular IPF output for the 109th US Senate

These modifications also provide a way to update probabilities for individuals more quickly. The following subsections discuss these alterations and their implications.

6.2.1 Initializing the IPF

The first way to modify the IPF is to improve the initialization of the g_{ik}s and the λs by using a non-parametric clustering algorithm. This will allow flexibility in choosing the number of pure types, k, and finding initial clusters that will speed up the maximization of the likelihood. As it stands, the IPF algorithm requires a variable of interest which acts as the state variable to initialize the g_{ik}s and the λs. This is not quite the same as a response variable in linear models, but it does serve as an initial grouping variable and is one of the variables in the data. In the questionnaire derivation of the IPF, the initial state variable would be one question that is thought to best separate the underlying groups. This variable also implicitly determines the number of latent variables.

The modified IPF begins this initialization with a non-parametric clustering technique, such as k-means or k-nearest neighbors. As discussed earlier this is not an ideal analysis for discrete data, but this only serves as an initial step. The number of pure types is chosen in this step. The state vector from the cluster analysis is used as the initial state variable. In this thesis we use the CLARA algorithm, which is good for large datasets.

6.2.2 Identifying Main Effects

It is not always immediately clear which variables will be removed from the IPF model as main effects. There are two main ways in which main effects can be determined. The first is to decide which variables to use based on prior knowledge. Usually this is based on previous studies or conventional wisdom. For example, in the World Health Survey on Global Aging (SAGE) likely main effects are gender, age,

and location (urban/rural). These are identified as important main effects due to the significance of these health factors in the literature.

In many other cases it is not clear which variables to use as the main effects. For these situations, using mutual information can determine which variables best account for the separation of groups. Mutual information is a quantity that measures the interdependence of two variables and can be used iteratively to determine how much of the variation between groups can be explained by adding another variable. It is formally defined as

$$\sum_{y \in Y} \sum_{x \in X} p(x,y) log \frac{p(x,y)}{p(x)p(y)}. \tag{6.1}$$

The top two or three variables that best describe the difference can be used as main effects. Using a greedy algorithm, the variable that contributes to most of the variation between groups is selected first, then subsequent variables are put together as a tuple. The next variable that best explains the separation of that tuple is selected and added to the tuple. This continues until all of the variation between the groups defined by tuples is accounted for. The first few variables that best describe variation between groups can be selected as main effects.

After the main effects are determined they are set aside. If there is a response variable it should also be removed from the data for prediction. The IPF algorithm is then used to reduce all remaining variables into g_{ik}s. Neither the prediction variable nor the main effects are used in calculating the g_{ik}s. The data are then set up as a contingency table with the main effects and the response variable with the g_{ik}s acting as covariates.

Once the IPF model is fit, the g_{ik}s are easily estimable for a new observation. A polytomous logistic regression model is fit using the X variables to predict the individual's g_{ik}s. A classification tree can also be used. Once the g_{ik}s are fit for the new individual, that individual's response can be predicted using regular categorical data analysis techniques.

This adaptation of the IPF combines a powerful, sound data-reduction technique with the flexibility of categorical data analysis. After the model is fit, secondary analyses can be performed on the g_{ik}s to determine regional mutual information. That is, individuals with similar g_{ik}s will tend to exhibit common behavior and could be defined by similar profiles. These profiles can be determined by using mutual information to determine which variables best define membership in a cluster of individuals with similar g_{ik} profiles.

In the next two chapters, we describe two illustrative datasets and demonstrate how the modified IPF algorithm can be used to analyze these types of data. In Chapter 7, the SAGE data analysis illustrates how the modified IPF is used as a predictive model. In Chapter 8, the WHS data analysis shows how policy-makers could use this methodology.

7. SAGE: CONTINUOUS RESPONSE

7.1 Introduction

In this chapter, the SAGE data is described and the modified IPF is used to build a predictive model to determine an individual's health score. Section 1 provides discussion about the SAGE data, mentions two questions of interest to the WHO, and describes how the modified IPF is used to answer those questions. Sections 2 and 3 discuss the results of the modified IPF is answering the WHO questions. These analyses are a unique and informative method of data analysis that should provide decision-makers with the information needed to assist in understanding various health subgroups in aging populations. Section 4 discusses the results from a simulation study that compares the predictive root mean squared error for two different models.

7.2 SAGE

The Study on Global Aging and Adult Health (SAGE) by the World Health Organization (WHO) is designed to determine the health status of individuals in aging populations. It is necessary for the WHO to provide relevant information to decision-makers in order to prepare for an aging population.

The data contain information from 23 survey variables for 1,437 individuals in a pilot study. These variables include age group, sex, urban/rural location, education level, marital status, 4-meter walk time, number of overnight stays in a health care facility, number of inpatient and outpatient visits, activities of daily living (ADLs), and responses to several self-reported health questions about arthritis, angina, stroke, and others. The response variable is health score and is a continuous score based on the responses to a series of self-reported health questions exogenous to the data

analyzed here. The score is between 16 and 67 where a lower score represents better health.

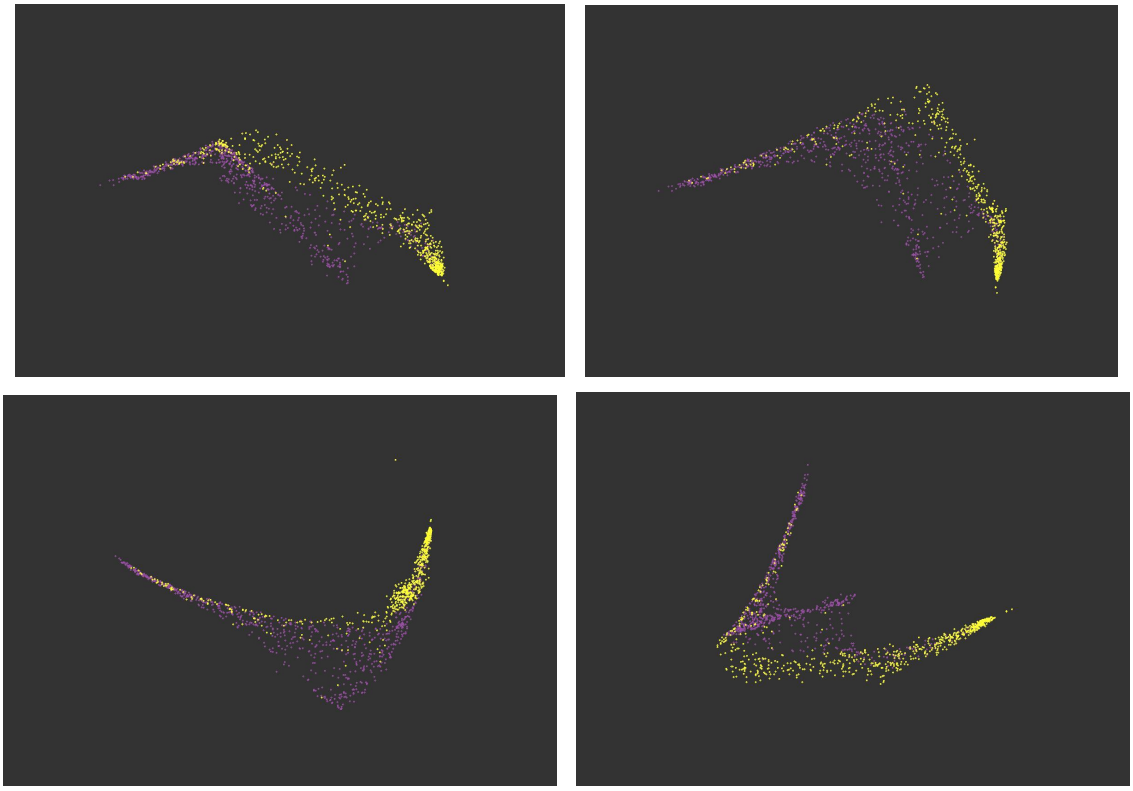
In discussing the SAGE data with WHO researchers, we discovered two major questions of interest: (1) how to build a linear model to predict health score based on the survey data and (2) how to define health subgroups to help policy-makers better understand how to allocate resources. The modified IPF can be used to answer each of these research questions. The first analysis uses the g_{ik}s from the IPF output as a summary variable of the observational factors in connection with main effect variables from SAGE to build a linear model for a continuous health score variable. Secondly, the IPF is used with all of the variables and the g_{ik}s are used to define small health groups. These subgroups can be used by policy makers to determine how best to focus resources to move individuals in a particular health subgroup toward a better health subgroup.

7.3 Health Score Linear Model

For the SAGE analysis, the IPF was first fit using all of the variables. A second IPF model was fit ignoring the three main effects age, sex, and location. The models were compared to see what changes, if any, occurred in the structure of the g_{ik}s. The IPF models for both cases were initialized using the cluster vector from a non-parametric, large-sample clustering algorithm (CLARA). We chose five clusters, which translate into five pure types in the IPF. The cluster vector was used to initialize the groups and determine the number of pure types in the model. Figure 7.1 shows four views of the data from the full model in g_{ik} space. It is colored by sex to demonstrate the dominance of the main effects. The full model shows dominant separation based on gender. This might not be too problematic in some situations, but a more sensitive model might reveal more subtleties in the data.

For the reduced model, we estimated the g_{ik}s for the SAGE data while ignoring

Figure 7.1: Four views of IPF clusters based on the all of the variables, colored by sex to show the dominance of main effects.



age, gender, and location. Health Score (SRH16) was also ignored in order to be used in the analysis of covariance (ANCOVA) model discussed later.

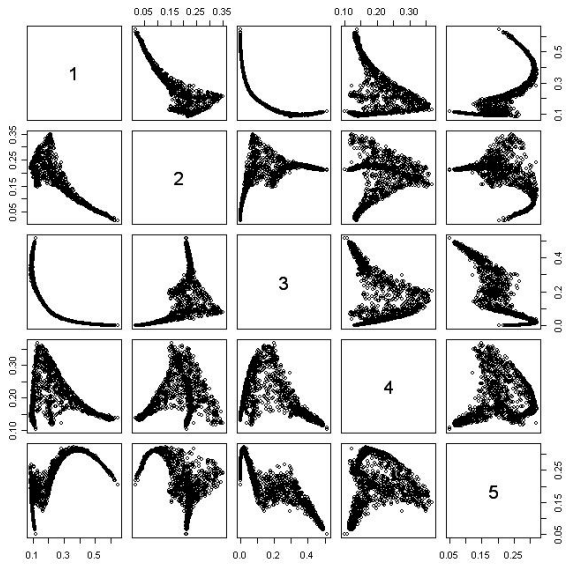
Removing the main effects did not significantly alter the overall structure of the g_{ik}s, but it did smear things out slightly. Figure 7.2 shows scatter plots for each of the g_{ik}s plotted against each other for both the full model and the interaction model. Notice that the structure is similar overall, but the interaction model has smoothed out the influence of the main effects. This hyperbolic shape shows the underlying distribution of the interactions and serves as an effectual "health spectrum".

One of the main problems of categorical data analysis in highly multivariate settings is the identifiability of parameters in sparse contingency tables. The modified IPF is an effective way to model interaction effects in cases where there is sparse categorical data. The g_{ik}s from the IPF output represent the information contained in the variables. This includes the information contained in the main effects of those variables as well as the interaction terms. The IPF algorithm is used to reduce the dimensionality of the interaction terms into continuous variables, namely the g_{ik}s, and use the g_{ik}s in connection with the main effect variables in an ANCOVA setting. The logarithm of health score was found to be approximately normally distributed, as shown by a normal probability plot in Figure 7.3. For this analysis, ANCOVA was performed using log(SRH16) as the response.

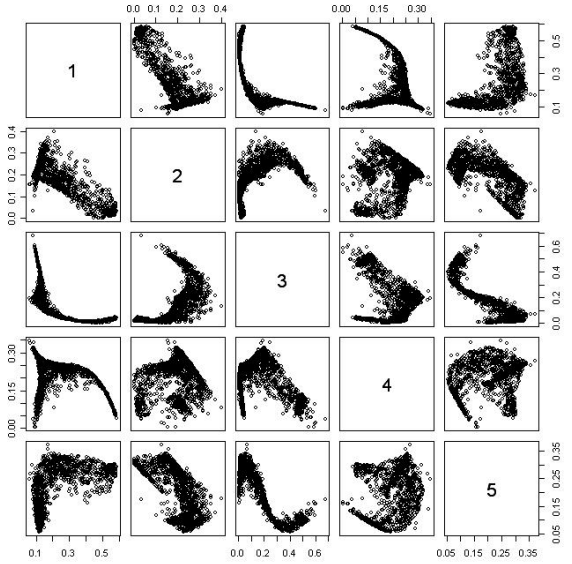
7.3.1 Results

Table 7.1 shows the Type I and Type III sums of squares for the ANCOVA based on the log of SRH16. The first g_{ik} was removed to avoid multicollinearity because of the constraint that the g_{ik}s sum to unity.

We also performed a stepwise variable selection using the AIC as a selection criterion, which showed that the first and the third g_{ik}s were the most important variables in determining health score, even more so than the main effects. This

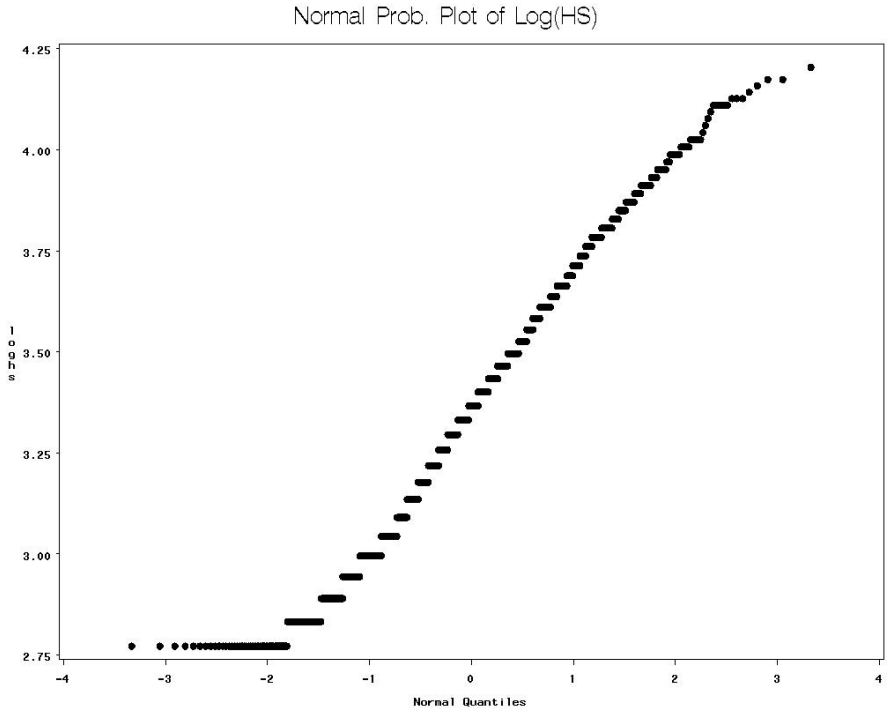


(a)



(b)

Figure 7.2: Scatterplots of each g_{ik} plotted against each other for the full model (a) and the interaction model (b). Notice the similar overall structure. This shows that removing the main effects smooths out the individual g_{ik}s without losing the overall structure.



(a)

Figure 7.3: Normal Probability Plot of the Log(SRH16). The normality assumption seemed adequately reasonable for ANCOVA.

Table 7.1: Type I and Type III sums of squares for the ANCOVA model using log(SRH16) as the response. X2, ..., X5 represent the g_{ik} effects.

Source	DF	Type I SS	Mean Square	F Value	Pr > F
age	4	19.571	4.893	99.53	<.0001
q1025-sex	1	11.722	11.722	238.46	<.0001
q0104-ur	1	0.582	0.582	11.84	0.0006
X2	1	21.920	21.920	445.90	<.0001
X3	1	26.253	26.253	534.04	<.0001
X4	1	1.445	1.445	29.40	<.0001
X5	1	0.228	0.228	4.63	0.0316

Source	DF	Type III SS	Mean Square	F Value	Pr > F
age	4	0.844	0.211	4.29	0.0019
q1025-sex	1	0.368	0.368	7.49	0.0063
q0104-ur	1	1.080	1.080	21.96	<.0001
X2	1	2.507	2.507	51.01	<.0001
X3	1	13.980	13.980	284.39	<.0001
X4	1	1.153	1.153	23.46	<.0001
X5	1	0.228	0.228	4.63	<.0316

Table 7.2: Stepwise Selection Summary

Step	Effect Entered	Effect Removed	Number Effects In	Number Parms In	AIC
0	Intercept	1	1	6629.3025	
1	X1		2	2	5777.6297
2	X3		3	3	5617.6481
3	q0104.ur		4	4	5603.2401
4	X5		5	5	5591.1239
5	age		6	9	5586.9166
6	q1025.sex		7	10	5583.6809*

Table 7.3: Type I sums of squares for the model $y = g_{i1} + g_{i3} + g_{i5} + sex + age + ur$. The main effects are put last in order to determine the relevance of the main effects after adjusting for the most important g_{ik}s.

Source	DF	Type I SS	Mean Square	F Value	Pr > F
X1	1	62.90566460	62.90566460	1278.74	<.0001
X3	1	16.19840160	16.19840160	329.28	<.0001
X5	1	0.39118133	0.39118133	7.95	0.0049
age	4	0.66802250	0.16700563	3.39	0.0090
q1025_sex	1	0.32970468	0.32970468	6.70	0.0097
q0104_ur	1	1.12917478	1.12917478	22.95	<.0001

indicates that the information contained in the g_{ik}s, or the summary of the interaction terms, is as important as the main effects. Table 7.2 shows the results.

Because g_{i1} was omitted in the original model shown in Table 7.1 and found to be a significant predictor, we reran the linear model to examine the Type I sums of squares for the model $y = g_{i1} + g_{i3} + g_{i5} + sex + age + ur$. The main effects were placed last in the model after g_{i1}, g_{i3}, and g_{i5} in order to see the strength of the main effects after adjusting for the g_{ik}s. Table 7.3 shows the Type I sums of squares. It is clear that the effects of g_{i1} and g_{i3} are much more practically significant than the rest of the variables.

7.4 Predictive Results

Using the g_{ik}s as summary variables of many discrete variables in linear models seems to work well, but it is possible that using one or two raw variables might do as well. To test the difference in predictive power between the model that uses g_{ik}s as covariates and the model with the best predicting raw variables, we performed a simulation study. To determine which raw variables best predicted health score, a stepwise variable selection method was used. Sex, "activ", and lungs were the three variables selected for the model. The variable "activ" is the response to the question "Overall in the last 30 days, how much difficulty did you have with work or

household activities?" . Possible responses include none, mild, moderate, severe, and extreme/cannot do. "Lungs" is the answer to the yes/no question "Have you ever been diagnosed with chronic lung disease?"

After deciding which variables to use, the predictive power of the models

$$y = gi1 + gi3 + gi5 \tag{7.1}$$

and

$$y = sex + activ + lungs \tag{7.2}$$

were compared in a simulation study. Only the best three of the g_{ik}s were used to make the models comparable in the number of variables.

We then performed a simulation study for each of these models. Each simulation run (1) randomly removed 100 of the 1,437 observations, (2) fit the model using the remaining 1,337 observations, (3) used the model to predict the health score of the 100 omitted observations, and (4) calculated the predictive root mean squared error (RMSE). This process was repeated for 1,000 samples. Figure 7.4 shows the distributions of the predictive RMSE. The blue line represents the g_{ik} model and the red represents the best subset model. The g_{ik} model outperforms the best subset model even without the use of the main effects. This shows that there is a lot of information contained in the g_{ik}s.

7.4.1 Adding Main Effects

It was also of interest to see if the predictive model was improved by adding the main effects to the model that only used g_{ik}s. Another simulation was performed adding the main effects to the g_{ik} model. The results are shown in Figure 7.5. The predictive RMSE is practically unchanged when we add the main effects to the base model containing only the g_{ik}s. This is further evidence of the strength of the infor-

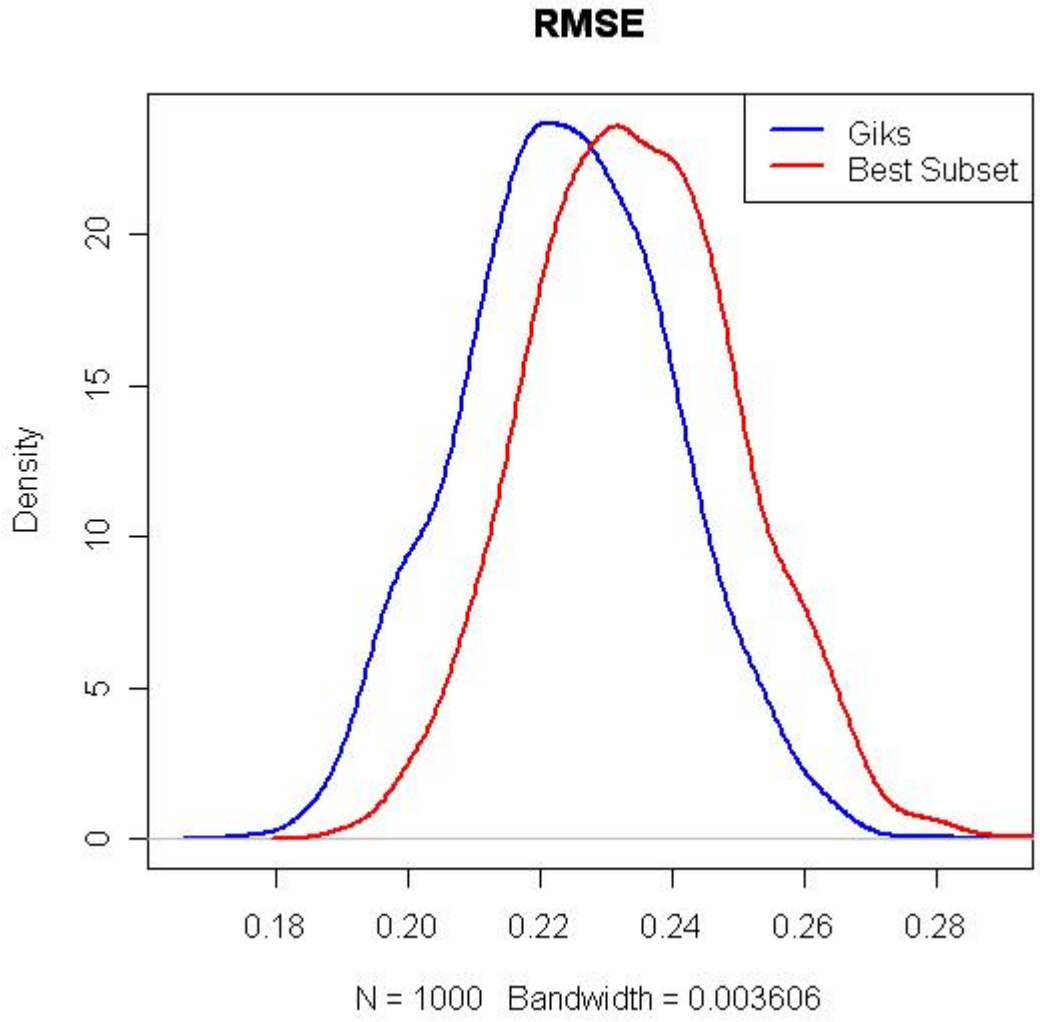


Figure 7.4: Predictive RMSE. Comparing simulation results for two models. The blue represents the model $y = gi1 + gi3 + gi5$ and the red represents $y = sex + activ + lungs$. Notice that the model with the g_{ik}s outperforms the best subset model even without using the main effects. This shows that there is a lot of predictive information contained in the g_{ik}s.

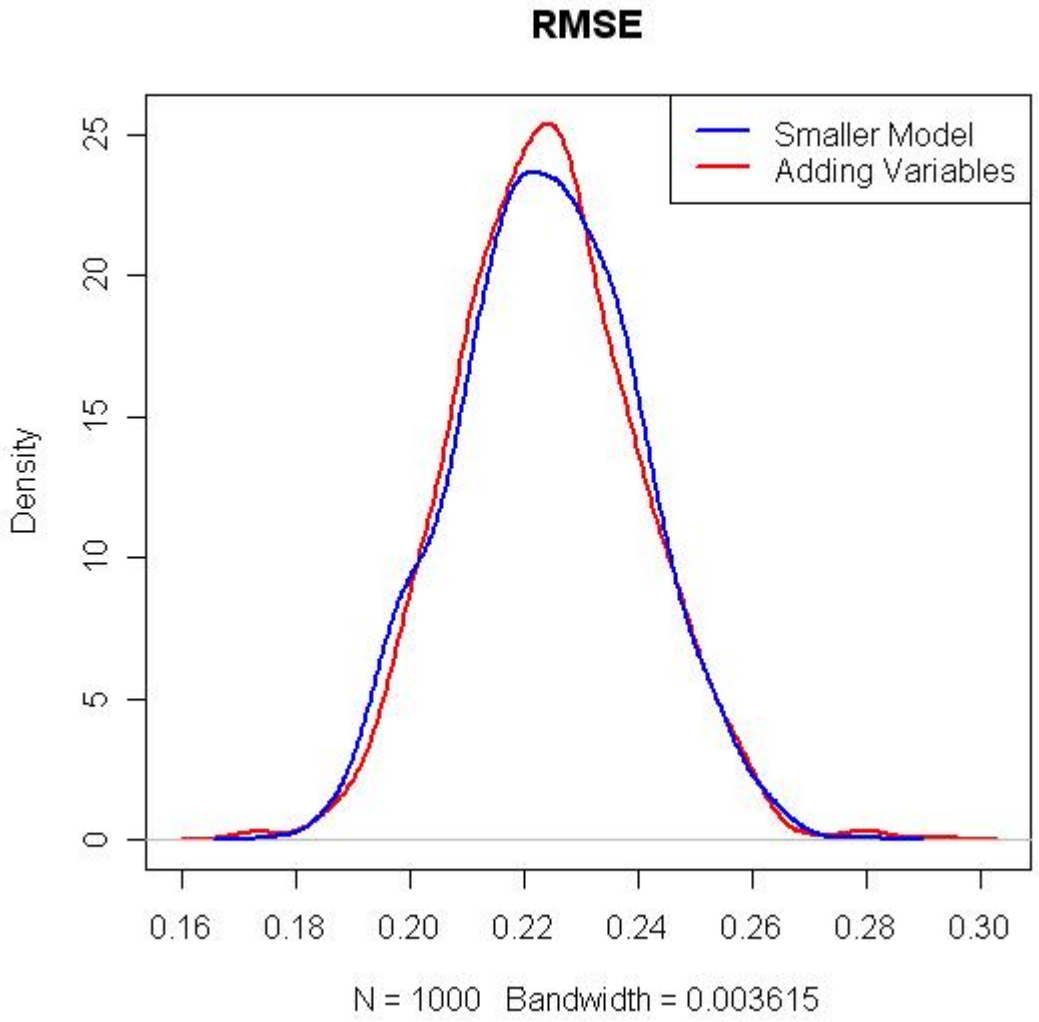


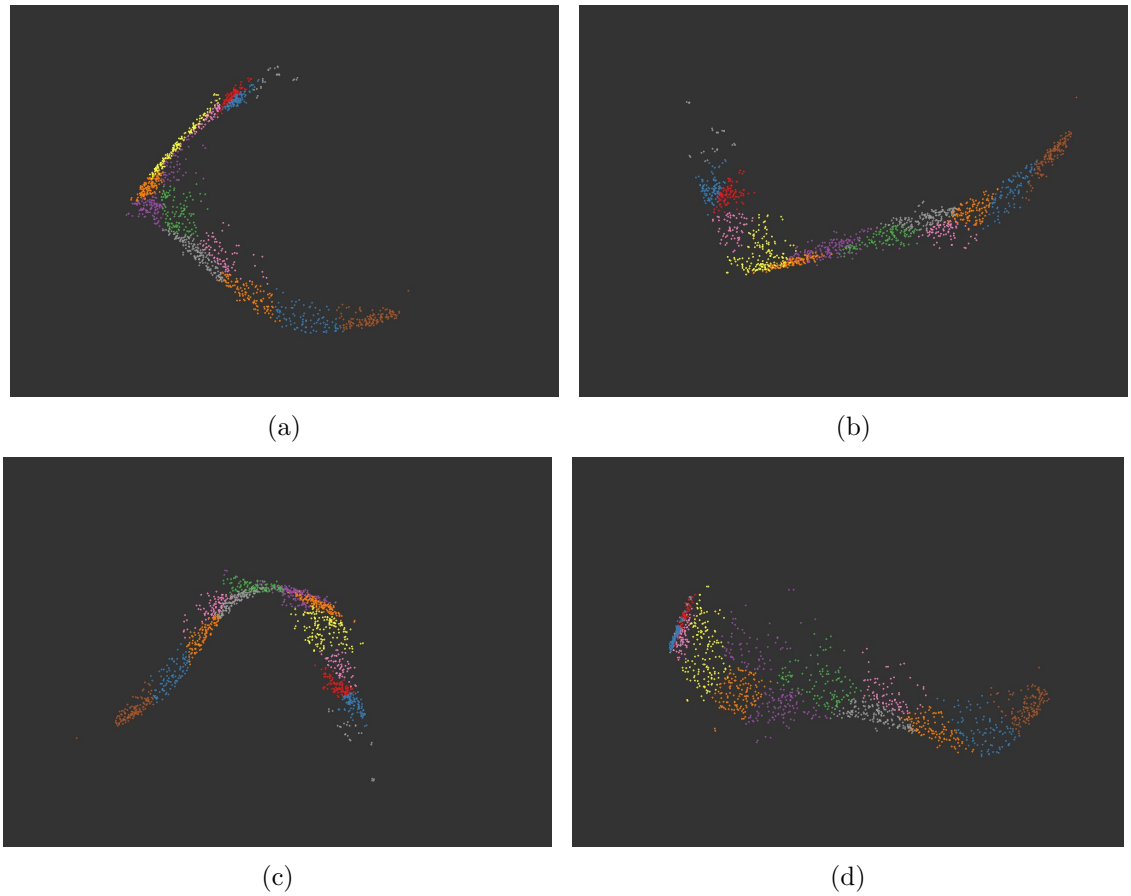
Figure 7.5: Predictive RMSE comparing simulation results of the g_{ik}-only model and the g_{ik} model with main effects. The blue represents the g_{ik}-only model, and the red represents the same model with the main effects. There is practically no difference in predictive RMSE. This is further evidence that the g_{ik}s contain a lot of information

mation contained in the g_{ik}s. This is not surprising in light of Table 7.3, which shows that the main effects are not nearly as practically significant.

7.5 Health Sub-Groups

The modified IPF can also be used in exploratory data analysis. After a model was fit, a secondary cluster analysis was performed on the g_{ik}s to define health sub-groups. After running several cluster analyses, 15 clusters seemed to break up the data into reasonable subgroups. The purpose of this secondary cluster analysis is to determine what variables define subgroups within g_{ik} space. Figure 7.6 shows the interaction model colored by health sub-groups.

After fitting the IPF to the interaction variables, it was of interest to determine the meaning of the health subgroups based on the g_{ik}s. The variables that best distinguish membership in these groups were determined using mutual information which iteratively finds which variables best describe group membership. Table 7.4 shows the variables and the levels of variables that characterize each of the clusters. The variables listed for each of the clusters are those that account for at least 90% of the variation between the groups. The three bolded variables for each cluster are the variables that best describe that group.



(a) (b)

(c) (d)

Figure 7.6: IPF clusters based on the interaction variables and colored by health sub-group.

Table 7.4: Cluster Meanings: most of the variation between groups is explained by seven or fewer variables. This table shows the variables and the levels of the variables that best describe each of the clusters. The bold-faced variables are the three most important for each cluster. The listed variables account for at least 90% of the variation for that group.

	HS	ADLs	4m	times	activ	health	education	outp	mar.stat
Clust 1	27.20	**39.28**	**5.96**	**4.63**	None-Mild	Moderate			
Clust 2	28.74	**41.77**	**6.04**	2.62	**Moderate-Mild**				
Clust 3	32.80	**50.84**	**5.83**	**4.74**	Mild-Moderate				
Clust 4	22.79	**34.64**			None-Mild	**Good-Moderate**	$\leq$ **2nd Compl**		
Clust 5	20.76	**30.73**	4.78			**Good**	$\leq$ **2nd Compl**	Half and Half	
Clust 6	39.27	**61.61**	**6.99**	**4.04**					
Clust 7	20.56	**29.51**	4.33	2.35		**Good**	$\leq$ **2nd Compl**		
Clust 8	30.22	**44.86**	**5.13**	**2.59**	Mild-Moderate				Mar-Wid
Clust 9	35.69	**54.80**	**5.82**	**4.99**					Wid-Mar
Clust 10	27.53	**40.74**	**5.03**	**2.50**			$\leq$Primary Compl		
Clust 11	45.30	**74.08**	**7.88**		**Severe**				
Clust 12	34.11	**52.27**	5.24	**3.12**	**Moderate**				
Clust 13	18.15	**26.23**	3.69			**Good**	Uniform	**Yes**	Mar
Clust 14	24.67	**38.28**	**5.05**	3.15			$\leq$ **2nd Compl**		
Clust 15	24.79	**34.41**	**5.21**			Good-Moderate	No Formal	**No**	

8. SAGE: DISCRETE RESPONSE

8.1 Introduction

This chapter demonstrates how the modified IPF can be used when the response is categorical. Section 2 outlines the parts of the GLM. The last section shows a simulation study to test the predictive power of two models. The first model uses only two g_{ik}s as predictors, and the second uses the two best variables, determined by a forward variable selection process.

8.2 Generalized Linear Model

For this analysis, the health scores were broken up into quartiles and a GLM was fit using the main effects and the g_{ik}s. These g_{ik}s are the same as those used in Chapter 7. Table 8.1 shows the parameter estimates and standard errors for the model. The model shows the probability of having a lower health index.

8.3 Predictive Results

As with the continuous response case, we performed a simulation study to test the predictive performance of GLM using the g_{ik}s and the GLM using the best subset model. The same procedure was followed as before, but a forward selection method was used to determine which raw variables best predict health score quartile. Each simulation predicted which category the omitted observations would be in. The prediction was classified as either "right," "close," or "wrong." Because of the ordinal nature of the categories, predicting a 3 when the individual was in Category 4 is better than predicting a 2. A prediction that was within one place of the true category was classified as "close." Figure 8.1 shows the prediction distributions for correct, close,

Table 8.1: Parameter estimates for the GLM using the health score quartiles. The model shows the probabilities of having a lower health index.

Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits	Chi-Square	Pr > ChiSq
Intercept1		1	-6.45	0.5471	-7.5265 -5.3818	139.17	<0.0001
Intercept2		1	-4.635	0.5431	-5.6998 -3.5707	72.83	<0.0001
Intercept3		1	-2.777	0.5280	-3.8127 -1.7428	27.67	<0.0001
age	18-49	1	0.6499	0.3872	-0.1091 1.4088	2.82	0.0933
age	50-59	1	0.5712	0.2258	0.1287 1.0137	6.40	0.0114
age	60-69	1	0.5729	0.2250	0.1318 1.0139	6.48	0.0109
age	70-79	1	0.1504	0.2356	-0.3114 0.6121	0.41	0.5234
age	80+	0	0.0000	0.0000	0.0000, 0.0000	.	.
q1025-sex	Female	1	-0.2311	0.1116	-0.4498 -0.0124	4.29	0.0383
q1025-sex	Male	0	0.0000	0.0000	0.0000, 0.0000	.	.
q0104-ur	Rural	1	0.4239	0.1050	0.2181 0.6296	16.30	<0.0001
q0104-ur	Urban	0	0.0000	0.0000	0.0000, 0.0000	.	.
X2		1	5.8502	1.0003	3.8897 7.8107	34.20	<0.0001
X3		1	9.8560	0.7285	8.4281 11.2839	183.03	<0.0001
X4		1	3.5841	0.9676	1.6876 5.4805	13.72	0.0002
X5		1	2.7801	1.4458	-0.0537 5.6140	3.70	0.0545

and wrong predictions. It also shows the distribution of the fitted model's Akaike Information Criterion (AIC). The blue line is the distribution for the results of the model: srh = sex + gi1 + gi3 + gi5. The red is the distribution for the model srh = sex + activ + angi + lungs. They both seem to get about the same number of predictions correct, but the model with the g_{ik}s gets more predictions close and consequently fewer wrong.

The AIC is the likelihood penalized for the number of parameters in the model. It is a measure of model goodness of fit, and a model with smaller AIC is preferable. Figure 8.1a shows that the AIC is much lower in all simulation cases for the model which uses the g_{ik}s.

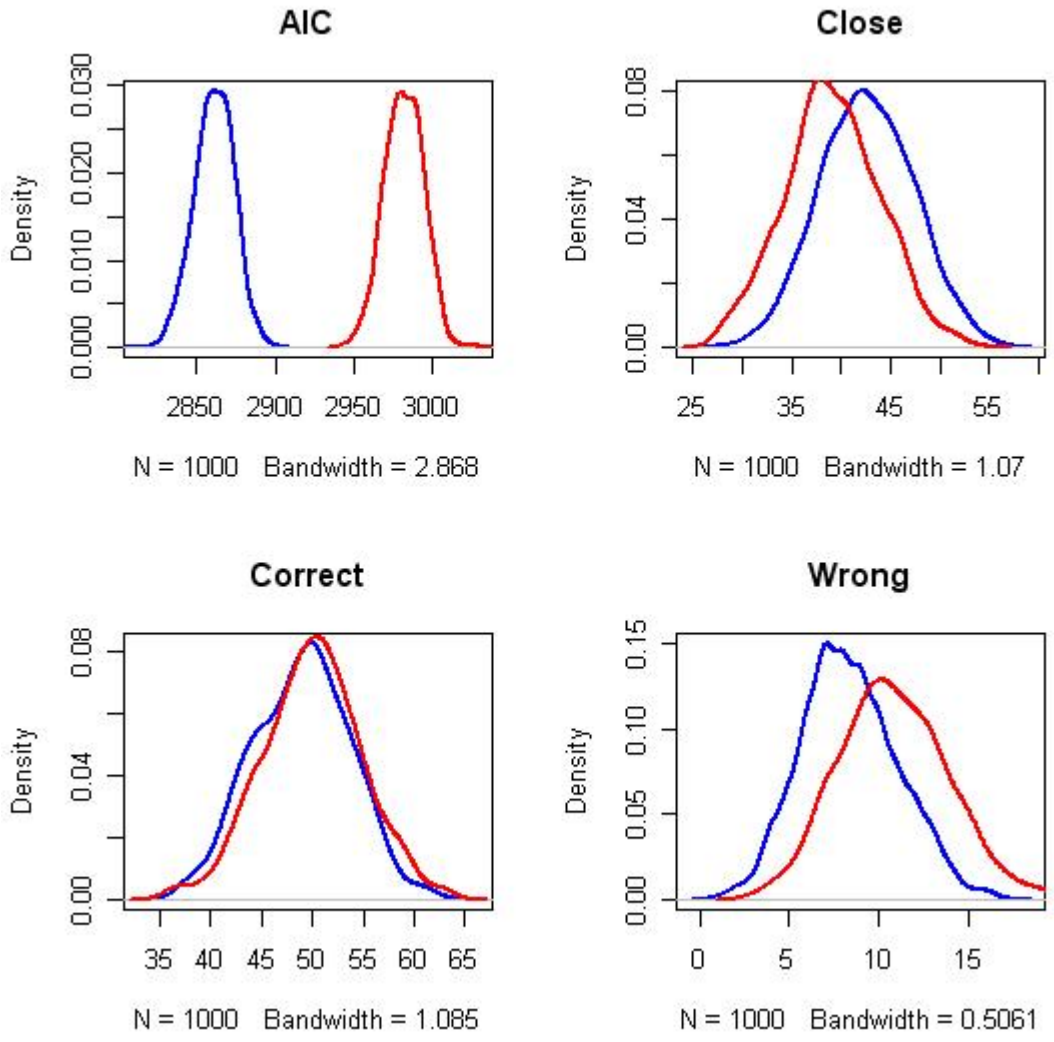


Figure 8.1: Simulation results for the discrete response case. Results from 1,000 samples of the SAGE data. The blue is the distribution for the results of the model srh = sex + gi1 + gi3 + gi5. The red is the distribution for the model srh = sex + activ + angi + lungs. They both seem to get the same number correct, but the model with the g_{ik}s gets more predictions close and, consequently, fewer wrong.

9. CONCLUSION AND DISCUSSION

In this thesis we have introduced a new, powerful method of using highly multivariate discrete data in linear models. We extended the Information Partition Function to be used as a way to represent all the information contained in the interaction terms of the World Health Organization's SAGE data and have demonstrated how powerful the g_{ik}s are in predicting health status.

There are several applications of the extended IPF that were not treated in this thesis. Further research in these applications might prove beneficial in moving us into the information age. Applications in natural language processing might help make the millions of online textual documents usable in statistical analysis. Other web-based applications such as collaborative filtering might also help to customize the internet for individual users.

BIBLIOGRAPHY

Airoldi, E. M., Fienberg, S. E., Joutard, C., and Love, T. M. (2006), "Discovering Latent Patterns with Hierarchical Bayesian Mixed-Membership Models," *CMU-ML-06-101, School of Computer Science, Carnegie Mellon University*.

Bartholomew, D. J. and Knott, M. (1999), *Latent Variable Models and Factor Analysis*, Oxford University Press Inc., 2nd ed.

Bishop, Y. M., Fienberg, S. E., and Holland, P. W. (1995), *Discrete Multivariate Analysis, Theory and Practice*, The MIT Press.

Blei, D., Ng, A., and Jordan, M. (2003), "Latent Dirichlet Allocation," *Advances in Neural Information Processing Systems*.

Cappe, O., Moulines, E., and Ryden, T. (2005), *Inference in Hidden Markov Models*, Springer Series in Statistics, Springer Science & Business Media, Inc.

Churchill, G. (1992), "Hidden Markov chains and the analysis of genome structure," *Computers & Chemistry*, 16, 107–115.

Cover, T. and Thomas, J. (1991), *Elements of Information Theory*, Wiley.

Elliot, R., Aggoun, L., and Moore, J. (1995), *Hidden Markov Models: Estimation and Control*, Springer Applications of Mathematics.

Engler, D. A. (2002), "An approach to probabilistic record linkage: Building a model through probability as extended logic and maximum entropy," Master's thesis, Brigham Young University, Provo Utah, Statistics Department.

Erosheva, E. (2002), "Grade of Membership and Latent Structure Models with Application to Disability Survey Data," Ph.D. thesis, Carnegie Mellon University, Pittsburgh.

— (2003), "Bayesian estimation of the Grade of Membership model," *Bayesian Statistics*, 7, 501–510.

— (2006), "Latent class representation of the Grade of Membership model; Technical Report No. 492," Tech. rep., Department of Statistics, University of Washington.

Erosheva, E. and Fienberg, S. (2004), "Bayesian Mixed Membership Models for Soft Classification; Working Paper no. 40." Tech. rep., Center for Statistics and the Social Sciences University of Washington.

Erosheva, E., Fienberg, S., and Lafferty, J. (2004), "Mixed–membership models of scientific publications," *Proceedings of the National Academy of Sciences*, 101, 5220–5227.

Fruhwirth-Schnatter, S. (2006), *Finite Mixture and Markov Switching Models*, Springer Series in Statistics, Springer Science and Business Media, LLC.

Gardner, J. (2007), *The Intelligent Universe: AI, ET, and the Emerging Mind of the Cosmos*, The Career Press Inc.

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

Griffiths, T. L. and Steyvers, M. (2004), "Finding Scientific Topics," *Proceedings of the National Academy of Sciences*, Vol. 101.

Hastie, T., Tibshirani, R., and Friedman, J. (2001), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer.

Hofmann, T. (1999), "Probabilistic Latent Semantic Analysis," *Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence*, http://www.cs.brown.edu/$\tilde{\text{t}}$h/papers/Hofmann–UAI99.pdf, (December 26, 2007).

Jaynes, E. (1957), "Information Theory and Statistical Mechanics," *Physical Review*, 106.

— (2003), *Probability Theory: The Logic of Science*, Cambridge Press.

Loehlin, J. C. (1998), *Latent Variable Models*, Lawrence Erlbaum Associates, Inc., Publishers, 3rd ed.

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

Manning, C. and Schütze, H. (1999), *Foundations of Statistical Natural Language Processing*, Massachusetts Institute of Technology, 2nd ed.

Manton, K., Woodbury, M., and Tolley, H. (1994), *Statistical Applications Using Fuzzy Sets*, Wiley–Interscience.

Oliphant, J. (2003), "The Information Partition Function," Master's thesis, Brigham Young University, Provo Utah, Statistics Department.

Potthoff, R., Manton, K., and Woodbury, M. (2000), "Dirichlet generalizations of latent–class models," *Journal of Classification*, 17, 315–353.

Rencher, A. C. (2002), *Methods of Multivariate Analysis*, Wiley Series in Probability and Statistics, John Wiley & Sons, Inc.

Rosenbaum, P. R. and Rubin, D. B. (1983), "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, 70 No. 1., 41–45.

Shannon, C. E. (1948), "A Mathematical Theory of Communication," *Bell System Technical Journal*, 27, 379–423.

Suppes, P. and Zanotti, M. (1981), "When are Probabilistic Explanations Possible?" *Synthese*, 48.

Tolley, H. (2006), "Derivation of the IPF Model," Personal correspondence regarding the justification of the intrinsic data model.

Woodbury, M. and Clive, J. (1974), "Clinical Pure Types as Fuzzy Partition," *Journal of Cybernetics*, 4, 111–121.