Parametric classification with non-normal data
by Alan Ray Willse

A thesis submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Statistics
Montana State University

Abstract:
This thesis is concerned with parametric classification of non-standard data. Specifically, methods are developed for classifying two of the most common types of non-Gaussian distributed data: data with mixed categorical and continuous variables (often called mixed-mode data), and sparse count data. Both supervised and unsupervised methods are described. First, a promising, recently proposed method that uses finite mixtures of homogeneous conditional Gaussian distributions (Lawrence and Krzanowski, 1996) is shown to be non-identifiable. Identifiable finite mixtures of homogeneous conditional Gaussian distributions are obtained by imposing constraints on some of the model parameters. Then, in contrast, it is shown that supervised classification of mixed-mode data using the homogeneous conditional Gaussian model can sometimes be improved by relaxing parameter constraints in the model; specifically, certain features of the continuous variable covariance matrix — such as volume, shape or orientation — are allowed to differ between groups. In addition, the use of latent class and latent profile models in supervised mixed-mode classification is investigated. Finally, mixtures of over-dispersed Poisson latent variable models are developed for unsupervised classification of sparse count data. Simulation studies suggest that for non-Gaussian data these methods can significantly outperform methods based in Gaussian theory.

PARAMETRIC CLASSIFICATION WITH NON-NORMAL DATA

by

Alan Ray Willse

A thesis submitted in partial fulfillment
of the requirements for the degree

of

Doctor of Philosophy

in

Statistics

MONTANA STATE UNIVERSITY-BOZEMAN
Bozeman, Montana

November 1999

APPROVAL

of a thesis submitted by

Alan Ray Willse

This thesis has been read by each member of the thesis committee and has been found to be satisfactory regarding content, English usage, format, citations, bibliographic style, and consistency, and is ready for submission to the College of Graduate Studies.

Nov 1, 1999   Robert J Boik
Date   Robert J. Boik
Chairperson, Graduate Committee

Approved for the Major Department

11/1/99   John Lund
Date   John Lund
Head, Mathematical Sciences

Approved for the College of Graduate Studies

11-2-99   Bruce R. McLeod
Date   Bruce McLeod
Graduate Dean

STATEMENT OF PERMISSION TO USE

In presenting this thesis in partial fulfillment for a doctoral degree at Montana State University-Bozeman, I agree that the Library shall make it available to borrowers under rules of the Library. I further agree that copying of this thesis is allowable only for scholarly purposes, consistent with "fair use" as prescribed in the U. S. Copyright Law. Requests for extensive copying or reproduction of this thesis should be referred to University Microfilms International, 300 North Zeeb Road, Ann Arbor, Michigan 48106, to whom I have granted "the exclusive right to reproduce and distribute copies of the dissertation for sale in and from microform or electronic format, along with the non-exclusive right to reproduce and distribute my abstract in any format in whole or in part."

Signature Alan Willse

Date 10/29/99

ACKNOWLEDGEMENTS

I would like to thank my adviser, Robert Boik, for his support and patience during the preparation of this thesis; Marty Hamilton, for his encouragement and guidance on statistical applications at the Center for Biofilm Engineering; the rest of my graduate committee — Jim Robison-Cox, John Borkowski and Bill Quimby — for their advice throughout my graduate studies; the faculty, students and staff at the Center for Biofilm Engineering, for a stimulating work environment; and Julie and Michael, for everything else. This work was partially supported by the National Science Foundation under NSF Cooperative Agreement EEC-8907039 with Montana State University.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

This thesis is concerned with parametric classification of non-standard data. Specifically, methods are developed for classifying two of the most common types of non-Gaussian distributed data: data with mixed categorical and continuous variables (often called mixed-mode data), and sparse count data. Both supervised and unsupervised methods are described. First, a promising, recently proposed method that uses finite mixtures of homogeneous conditional Gaussian distributions (Lawrence and Krzanowski, 1996) is shown to be non-identifiable. Identifiable finite mixtures of homogeneous conditional Gaussian distributions are obtained by imposing constraints on some of the model parameters. Then, in contrast, it is shown that supervised classification of mixed-mode data using the homogeneous conditional Gaussian model can sometimes be improved by relaxing parameter constraints in the model; specifically, certain features of the continuous variable covariance matrix — such as volume, shape or orientation — are allowed to differ between groups. In addition, the use of latent class and latent profile models in supervised mixed-mode classification is investigated. Finally, mixtures of over-dispersed Poisson latent variable models are developed for unsupervised classification of sparse count data. Simulation studies suggest that for non-Gaussian data these methods can significantly outperform methods based in Gaussian theory.

CHAPTER 1

Introduction

Classification problems abound in the natural and social sciences. Volumes have been written about classification with continuous variables, especially variables that are normally distributed (see, for example, McLachlan (1992) or Ripley (1992)). Much less work has been done on non-continuous data — for example, data containing both categorical and continuous variables, or vectors of counts — even though such data are frequently encountered in practice. In this thesis, methods are developed to fill some of the gaps in classification with non-continuous data. Specifically, methods are developed for mixed categorical and continuous data, and for multidimensional count data. The methods use ideas from discriminant analysis, cluster analysis, and latent variable models. The distinction between these methods will be made in the remainder of this chapter. As will be seen, latent variable methods can play a significant role in classification efforts.

The basic problem in classification is to assign an entity (e.g., a person, document) to one or more of K groups (e.g., disease class, topic) based on some measures $\mathbf{X} = (X_1, \ldots, X_p)'$ taken on the entity. A distinction is made between *supervised* and *unsupervised* classification. In supervised classification (also known as discriminant analysis), observations made on entities with known group membership are available. These observations are used to develop a rule for classifying future observations, or observations without group labels. For example, suppose the variables $X_1, \ldots, X_p$ describe symptoms of some disease, and that the true disease status can be determined only after a laborious and costly medical procedure. To avoid unnecessary medical procedures, the disease status of most individuals must be predicted from data collected from those few individuals who underwent the medical procedure. Supervised classification methods are routinely used in medical settings to diagnose diseases and to prognose outcomes of risky medical procedures.

In unsupervised classification (also known as cluster analysis), no group labels are known. In some cases, there is prior understanding of the types of groups (for example, diseased or not diseased). In the absence of a gold standard (e.g., for emerging diseases) individuals may be clustered and classified into groups based on their observed symptom variables. The groups might be given the labels *diseased* and *not diseased.* In some cases the goal of unsupervised classification is to discover group structure in a dataset. A major problem is to decide how many groups are in the data and then to characterize the groups. For example, we might wish to cluster a large collection of documents into groups of related topics. In this thesis both supervised and unsupervised methods are considered.

This thesis focuses on classification of non-continuous data. Specifically, two types of data structures are considered:

1. Data containing mixtures of categorical and continuous variables. This type of data will be referred to as mixed-mode.

2. Sparse multivariate count data.

Datasets with mixed categorical and continuous variables are often encountered in practice. It is common to standardize these datasets by either 1) categorizing the continuous variables and applying categorical variable methods, or 2) treating the categorical variables as continuous and applying continuous variable methods. Clearly, information is lost with either approach. As an alternative, Krzanowski (1975, 1980, 1993) developed a parametric approach to analyzing mixed-mode data. In this model, known as the conditional Gaussian model, the continuous variables have a different multivariate normal distribution at each possible combination of categorical variable values. Research on the conditional Gaussian model has been driven by the growing interest in Bayesian Belief Networks, which frequently employ conditional Gaussian models when mixed variables are present. The conditional Gaussian model, and a special case known as the location model, will be described in more detail in Chapters 2 and 3, where the models are exploited in the development of both supervised and unsupervised methods for classifying mixed-mode data.

In Chapter 4 unsupervised methods are developed for classification with sparse multivariate count data. In sparse multivariate count data, multiple counts are observed for each entity, and many of the counts are very small or zero. Chapter 4 describes how such sparse count data are routinely collected in secondary ion mass spectrometry. In the analysis of textual data, a document is often represented by a vector $\mathbf{X} = (X_1, \ldots, X_T)'$, where T is the number of unique terms, or words, in some collection of documents, and X_i is the number of times (i.e. count) the i^{th} term occurs in the document. A given document will contain only a fraction of the unique term in the collection, so many counts will be zero. The positive counts tend to be very low. Thus, textual data analysis must contend with sparse multivariate count data. If the data weren't sparse (i.e., if the counts weren't so small), it might be possible to apply continuous variable methods to the count data following some transformation. For example, the Anscombe transform of a random variable X is given by

$$Y = t(X) = 2\sqrt{X + 3/8}.$$

If $X \sim$ Poisson(λ) and λ is large, then Y is approximately normally distributed with variance 1. When λ is small, the transformed variables are not approximately normal. In this case, we postulate that classification can be improved by modeling the count data with more appropriate multivariate count distributions. This is done in Chapter 4, where the multivariate count distributions are described by latent variable models. A latent variable is introduced to "explain" the correlations among observations within a cluster. Observations conforming to the latent variable model are clustered using a finite mixture model. In a finite mixture model, an observation's group membership is treated as an unobservable, or latent, variable. Thus the clustering algorithm contains two levels of latent variables. A brief discussion of latent variables, and their use in this thesis, is considered next.

In its most general definition, a latent variable model is any model with a variable that is unobservable (or latent). If $\mathbf{X}$ is a vector of observable variables, and $\mathbf{Z}$ is a vector of latent variables, then the density of the observable variables may be written as

$$f(\mathbf{x}) = \int_z h(\mathbf{z}) g(\mathbf{x}|\mathbf{z}) d\mathbf{z}. \tag{1.1}$$

If $\mathbf{Z} \sim$ Mult$(1; \mathbf{p})$, where $\mathbf{p} = (p_1, \ldots, p_k)'$, then (1.1) is a finite mixture model with mixing parameters $p_1, \ldots, p_k$. Thus, finite mixture models are special types of latent variable models. These types of latent variable models are used in Chapters 2 and 4.

More commonly, latent variable models are defined by the notion of conditional independence, so that, conditional on the value of the latent variable, the observable variables are taken to be independent. In this sense the latent variables are said to explain (the associations among) the observable variables. In this definition the dimension of the latent variables $\mathbf{Z}$ is taken to be smaller (usually much smaller) than the dimension of the observable variables $\mathbf{X}$. These types of models are used in Chapters 3 and 4.

In the three main chapters of this thesis (Chapters 2, 3 and 4) new methods are developed for classifying non-standard data types. These three chapters provide a cohesive argument that better classification can be achieved if the data structure is properly accounted for.

Chapter 2 considers the problem of unsupervised classification for mixed-mode data. After reviewing existing approaches to the problem, a promising approach based on finite mixtures of conditional Gaussian distributions is shown to be non-identifiable. Then identifiable finite mixtures of conditional Gaussian distributions are developed.

In Chapter 3, conditional Gaussian models are developed for supervised classification. Parsimonious models which relax the assumption of common within-cell dispersion matrices are considered.

Finally, unsupervised methods for sparse count data are developed in Chapter 4. The methods, based on finite mixtures of latent variable models, compare favorably with methods that transform the variables (using, for example, the Anscombe transform) and then apply normal variable methods.

CHAPTER 2

Identifiable Finite Mixtures of Location Models for Clustering Mixed-Mode Data

Finite mixture models have become popular tools for cluster analysis, especially when it is reasonable to make distributional assumptions about observations within each group. Titterington, Smith and Makov (1985) and McLachlan and Basford (1988) provide comprehensive reviews of finite mixture applications in cluster analysis.

Suppose that an observation $\mathbf{x}$ has arisen from exactly one of g distinct groups, denoted $G_1, \ldots, G_g$, where the density of an observation from G_i is $g_i(\mathbf{x}; \mathbf{\Psi}_i)$. The parameter vector $\mathbf{\Psi}_i$ is generally unknown. If α_i is the relative size of G_i $(0 < \alpha_i < 1;\ \sum_{i=1}^{g} \alpha_i = 1)$, then the density of a randomly selected observation is

$$f(\mathbf{x}) = \sum_{i=1}^{g} \alpha_i g_i(\mathbf{x}; \mathbf{\Psi}_i). \tag{2.1}$$

Model (2.1) is a finite mixture model with mixing parameters α_i $(i = 1, \ldots, g)$. The mixing parameters also are known as prior group probabilities. Finite mixture models are suitable for multiple group analysis — in our case cluster analysis — when group labels are unknown. The posterior probability that $\mathbf{x}_h$ belongs to G_i is

$$\tau_i(\mathbf{x}_h; \mathbf{\Psi}) = \Pr(G_i | \mathbf{x}_h, \mathbf{\Psi}_i, \alpha_i) = \frac{\alpha_i g_i(\mathbf{x}_h; \mathbf{\Psi}_i)}{\sum_{l=1}^{g} \alpha_l g_l(\mathbf{x}_h; \mathbf{\Psi}_l)}.$$

If misclassification costs are equal, then observation $\mathbf{x}_h$ is assigned to the group for which the posterior probability is greatest. That is, the classification rule is

$$\text{assign } \mathbf{x}_h \text{ to } G_i \text{ if } \max_{1 \le l \le g} \tau_l(\mathbf{x}_h; \mathbf{\Psi}) = \tau_i(\mathbf{x}_h; \mathbf{\Psi}). \tag{2.2}$$

In practice, the parameters α_i and Ψ_i $(i = 1, \ldots, g)$ usually are estimated from the sample $\mathbf{x}_1, \ldots, \mathbf{x}_n$ which is to be clustered, and the estimates are substituted in (2.2) for classification. Before the finite mixture model can be used for cluster analysis, a decision must be made about the form of the group conditional densities $g_i(\mathbf{x}; \Psi_i)$. For continuous data it is often reasonable to assume multivariate normal group conditional densities. Maximum likelihood estimates of the parameters can be obtained by treating the unobserved group labels as missing data and using the EM algorithm (McLachlan and Krishnan, 1997; Redner and Walker, 1984).

When observations are made on both categorical and continuous variables — in which case we say the data are mixed-mode, or mixed — the multivariate normal assumption is not realistic. Everitt (1988) constructed finite mixture models for this case by assuming that each categorical variable is obtained from an underlying continuous variable by thresholding. The underlying (unobserved) continuous variables and the observed continuous variables are assumed to be jointly multivariate normal within each group, with common covariance matrix. This model will be referred to as the *underlying variable mixture model.*

The categorical variables in the underlying variable mixture model are ordinal. That is, the levels of each categorical variable are determined by ordered threshold values of an underlying continuous variable. Because the categorical variables provide no information about the means and variances of the underlying continuous variables, Everitt (1988) takes the means to be 0 and the variances to be 1. The threshold values are allowed to vary across variables and groups. The category probabilities are determined by the threshold values. In practice the method is limited to one or two categorical variables (Everitt and Merette, 1990), because for q categorical variables estimation of the parameters requires q-dimensional numerical integration at each iteration of the EM algorithm. Fitting the model can be numerically intractable for large q.

Lawrence and Krzanowski (1996) proposed a finite mixture model for mixed-mode data that avoids the numerical integration required by the underlying variable mixture model. They assumed that the group-conditional densities conform to the location model for mixed variables. The location model has been successfully applied in discriminant analysis problems (Krzanowski, 1993). In the graphical models literature it is called the homogeneous Conditional Gaussian model (Whittaker, 1990). The finite mixture of location models will be called the *location mixture model.* In addition to greater numerical tractability, the location mixture model promises more flexibility than the underlying variable mixture model because it doesn't impose any orderings of the categories in each categorical variable, and it doesn't impose structure on the conditional means.

Unfortunately, the great flexibility of the location mixture model leads to multiple distinct sets of parameter values that yield identical mixture densities; that is, the model in its unrestricted form is not identifiable. This is demonstrated in the next section. Then identifiable location mixture models are obtained by imposing restrictions on the conditional means of the continuous variables. The restricted models are assessed in a simulation experiment.

Location Mixture Model

The Conditional Gaussian distribution decomposes the joint distribution of mixed-mode data as the product of the marginal distribution of the categorical variables and the conditional distribution of the continuous variables given the categorical variables. The latter distribution is assumed to be multivariate normal. The categorical variables can be uniquely transformed to a single discrete variable

$w \in \{w_1, \ldots, w_m\}$, where m is the number of distinct combinations (i.e., locations) of the categorical variables, and w_s is the label for the s^{th} location. If there are q categorical variables and the j^{th} variable has c_j categories ($j = 1, \ldots, q$) then $m = \prod_{j=1}^{q} c_j$. The associations among the original categorical variables are converted into relationships among the discrete probabilities $\Pr(w_s) = p_s$. Following Lawrence and Krzanowski (1996), a sample of mixed-mode data will be denoted by

$$\mathbf{x} = (\mathbf{x}'_{11} \ldots \mathbf{x}'_{1n_1} \mathbf{x}'_{21} \ldots \mathbf{x}'_{2n_2} \cdots \mathbf{x}'_{m1} \ldots \mathbf{x}'_{mn_m})'$$

where $\mathbf{x}_{sh}$ is a $p \times 1$ vector of continuous variables for the h^{th} observation at location w_s, and n_s is the number of observations at w_s. Within w_s, the Conditional Gaussian model states that $\mathbf{x}_{sh} \sim \mathrm{N}(\boldsymbol{\mu}_s, \boldsymbol{\Sigma}_s)$. The homogeneous Conditional Gaussian model, also called the *location model*, is obtained by restricting the covariance matrix to be the same for all locations and for all groups (if there is additional grouping structure).

In their finite mixture application, Lawrence and Krzanowski (1996) assumed that each vector $\mathbf{x}_{sh}$ ($h = 1, \ldots, n_s$; $s = 1, \ldots, m$) belongs to one of g distinct groups, $G_1, \cdots, G_g$, but that the group labels are unknown. They assumed that observations within each group conform to a location model, so that $\Pr(w = w_s | G_i) = p_{is}$ and, in G_i, $\mathbf{x}_{sh} \sim \mathrm{N}(\boldsymbol{\mu}_{is}, \boldsymbol{\Sigma})$. In G_i the joint probability that an observation is from w_s and has continuous variable vector $\mathbf{x}_{sh}$ is

$$g_i(\mathbf{x}_{sh}, w_s; \boldsymbol{\Psi}) = p_{is} h(\mathbf{x}_{sh}; \boldsymbol{\mu}_{is}, \boldsymbol{\Sigma}),$$

where $\boldsymbol{\Psi}$ contains all unknown parameters and $h(\mathbf{x}_{sh}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the pdf of a $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ random variable. The joint probability that a random observation with unknown group membership is from w_s and has continuous variable vector $\mathbf{x}_{sh}$ is

$$f(\mathbf{x}_{sh}, w_s; \boldsymbol{\Psi}) = \sum_{i=1}^{g} \alpha_i g_i(\mathbf{x}_{sh}, w_s; \boldsymbol{\Psi}) = \sum_{i=1}^{g} \alpha_i p_{is} h(\mathbf{x}_{sh}; \boldsymbol{\mu}_{is}, \boldsymbol{\Sigma}), \qquad s = 1, \ldots, m. \quad (2.3)$$

The α_i are group mixing parameters. The parameters $\{\alpha_i\}$ and $\{p_{is}\}$ satisfy the constraints

$$\sum_{i=1}^{g} \alpha_i = 1 \quad \text{and} \quad \sum_{s=1}^{m} p_{is} = 1 \ \forall i. \tag{2.4}$$

Lawrence and Krzanowski (1996) describe how the unknown parameters in (2.3) can be estimated using the EM algorithm. The conditional group means $\boldsymbol{\mu}_{is}$ $(i = 1, \ldots, g; s = 1, \ldots, m)$ are unrestricted in $\Re^p$. If, at each location, the means are the same for each group, then $g = 1$ is sufficient and the mixture model is degenerate. This paper is concerned with non-degenerate models. We therefore assume that any two groups have different means at some location (i.e., for each $i \neq i'$, $\boldsymbol{\mu}_{is} \neq \boldsymbol{\mu}_{i's}$ for some s). The $p \times p$ common covariance matrix $\boldsymbol{\Sigma}$ is assumed to be positive definite.

Model (2.3) is called the location mixture model. In this paper it will sometimes be called the *unrestricted* location mixture model to distinguish it from the restricted location mixture models which are introduced later in this chapter.

Identifiability

A parametric family of probability models is said to be identifiable if distinct parameter values determine distinct members of the family. That is, a family $\{p(\mathbf{x}; \Theta)\}$ is identifiable if for Θ and Θ' in the family's parameter space, $p(\mathbf{x}; \Theta) \equiv p(\mathbf{x}; \Theta') \Rightarrow \Theta = \Theta'$. In finite mixture models, different representations corresponding to a simple relabeling of group indexes are considered equivalent, so identifiability is required only up to a relabeling of group indexes. In the location mixture model the parameter sets $\boldsymbol{\Psi} = \{\alpha_i, p_{is}, \boldsymbol{\mu}_i, \Sigma\}$ and $\boldsymbol{\Psi}' = \{\alpha_i', p_{is}', \boldsymbol{\mu}_i', \Sigma'\}$ are considered to be *equivalent* if they can be made identical by permuting group labels. Otherwise they are distinct. For example, the parameter set $\boldsymbol{\Psi} = \{\alpha_i, p_{is}, \boldsymbol{\mu}, \Sigma\}$ for $g = 2$ groups and m locations is equivalent to the parameter set $\boldsymbol{\Psi}'$ obtained by $\alpha_1' = \alpha_2$, $\alpha_2' = \alpha_1$, $\Sigma' = \Sigma$, and, for all s, $p_{1s}' = p_{2s}$, $p_{2s}' = p_{1s}$, $\boldsymbol{\mu}_{1s}' = \boldsymbol{\mu}_{2s}$, and $\boldsymbol{\mu}_{2s}' = \boldsymbol{\mu}_{1s}$. Accordingly, the location mixture model (2.3) is identifiable if, for each $s = 1, \ldots, m$ and for all $\mathbf{x}_{sh} \in \Re^p$

$$\sum_{i=1}^{g} \alpha_i p_{is} h(\mathbf{x}_{sh}; \boldsymbol{\mu}_{is}, \Sigma) = \sum_{i=1}^{g} \alpha_i' p_{is}' h(\mathbf{x}_{sh}; \boldsymbol{\mu}_{is}', \Sigma') \Rightarrow \boldsymbol{\Psi} \text{ and } \boldsymbol{\Psi}'\text{are equivalent.} \qquad (2.5)$$

Yakowitz and Spragins (1968) provide some useful results for establishing the identifiability of finite mixture models.

To examine the identifiability of the unrestricted location mixture model in (2.3), it is convenient to define $f_{is} = \alpha_i p_{is}$ $(i = 1, \ldots, g; s = 1, \ldots m; \sum \sum f_{is} = 1)$. It follows from (2.4) that

$$\alpha_i = \sum_{s=1}^{m} f_{is}, \qquad p_{is} = \frac{f_{is}}{\alpha_i}. \qquad (2.6)$$

Consider the case of $m = 2$ locations and $g = 2$ groups. This model defines $mg = 4$ clusters of continuous observations with relative frequencies f_{is} and associated means $\boldsymbol{\mu}_{is}$.

If there is another set of parameters $\boldsymbol{\Psi}' = \{\alpha_i', p_{is}', \boldsymbol{\mu}_{is}', \Sigma'\}$, distinct from $\boldsymbol{\Psi} = \{\alpha_i, p_{is}, \boldsymbol{\mu}_{is}, \Sigma\}$, such that (2.5) is satisfied, then the location mixture model is not identifiable. Such a set of parameters can be obtained by permuting group labels at some locations but not at others, or by permuting group labels differently at different locations. Consider permuting (or swapping) group labels for cluster frequencies and conditional means at the second location, but not at the first location, so that cluster frequencies after permutation are (in prime notation) $f_{11}' = f_{11}$ and $f_{21}' = f_{21}$ at location 1, and $f_{12}' = f_{22}$ and $f_{22}' = f_{12}$ at location 2. Parameter values for both labelings – denoted by $\boldsymbol{\Psi}$ and $\boldsymbol{\Psi}'$ – are given in Table 1.

Clearly $\sum_{i=1}^{2} f_{is}' h(\mathbf{x}_{sh}; \boldsymbol{\mu}_{is}', \Sigma) = \sum_{i=1}^{2} f_{is} h(\mathbf{x}_{sh}; \boldsymbol{\mu}_{is}, \Sigma)$ for $s = 1, 2$ and $\forall \mathbf{x}_{sh} \in \Re^p$. Thus the distinct parameter sets $\boldsymbol{\Psi}$ and $\boldsymbol{\Psi}'$ — both in the parameter space for the location mixture model — satisfy (2.5). It follows that (2.3) is not identifiable for $m = 2$ and $g = 2$. (It may happen that the $\{f_{is}'\}$ are all the same, which implies

		$\boldsymbol{\Psi}$	$\boldsymbol{\Psi}'$
	α_1	$f_{11}+f_{12}$	$f_{11}+f_{22}$
	α_2	$f_{21}+f_{22}$	$f_{12}+f_{21}$
location 1	p_{11}	$\frac{f_{11}}{f_{11}+f_{12}}$	$\frac{f_{11}}{f_{11}+f_{22}}$
	p_{21}	$\frac{f_{21}}{f_{21}+f_{22}}$	$\frac{f_{21}}{f_{12}+f_{21}}$
	$\boldsymbol{\mu}_{11}$	$\boldsymbol{\theta}_{11}$	$\boldsymbol{\theta}_{11}$
	$\boldsymbol{\mu}_{21}$	$\boldsymbol{\theta}_{21}$	$\boldsymbol{\theta}_{21}$
location 2	p_{12}	$\frac{f_{12}}{f_{11}+f_{12}}$	$\frac{f_{22}}{f_{11}+f_{22}}$
	p_{22}	$\frac{f_{22}}{f_{22}+f_{21}}$	$\frac{f_{12}}{f_{12}+f_{21}}$
	$\boldsymbol{\mu}_{12}$	$\boldsymbol{\theta}_{12}$	$\boldsymbol{\theta}_{22}$
	$\boldsymbol{\mu}_{22}$	$\boldsymbol{\theta}_{22}$	$\boldsymbol{\theta}_{12}$

Table 1: Two distinct sets of parameters that give equivalent expressions for the unrestricted location mixture model (2.3) for the case of $m = 2$ locations and $g = 2$ groups. The parameter set $\boldsymbol{\Psi}'$ is obtained from $\boldsymbol{\Psi}$ by permuting group labels at the second location but not at the first location. Group/location cluster frequencies are represented by the parameters $f_{is} = \alpha_i p_{is}$.

that the $\{p'_{is}\}$ and $\{\alpha'_i\}$ are all the same. The parameter sets $\boldsymbol{\Psi}$ and $\boldsymbol{\Psi}'$ will still be distinct, because the $\boldsymbol{\mu}_{is}$'s are assumed in general to be different). The model can be made identifiable by imposing restrictions on $\{\boldsymbol{\mu}_{is}\}$, as will be shown in the next section.

The non-identifiability of the unrestricted location mixture model is due to indeterminacy of group labels at each location. This group indeterminacy is illustrated in Figure 1 for $m = 2$ locations, $g = 2$ groups and $p = 2$ continuous variables. The triangles represent cluster means at location 1, and the squares represent means at location 2. Cluster frequencies are given beside the means. Locations of the clusters are known and labeled, but group labels within the locations are unknown. Group labels can be assigned in two nonredundant ways. The first labeling, in which clusters from the same group are connected by solid lines, can be described by the location mixture model with probability parameters $\alpha_1 = .6, p_{11} = 1/3$, and $p_{21} = 3/4$ (assuming that the clusters are conditionally MVN with common covariance matrix). The second labeling, represented by dashed lines, can be described by the location

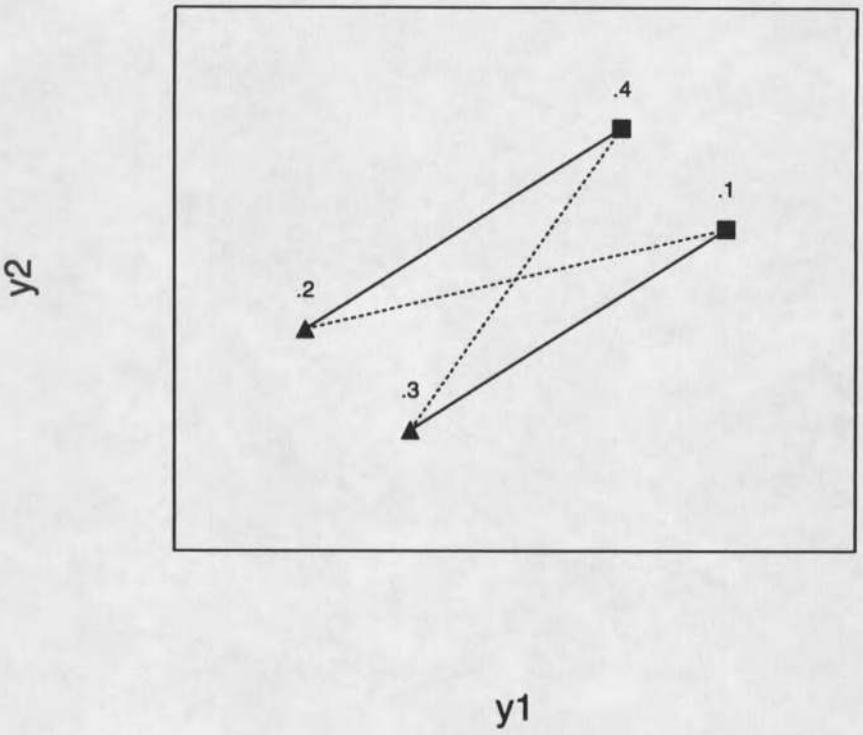

Figure 1: Four cluster means for a hypothetical 2-group, 2-location mixture model. Conditional means at the first location are represented by triangles. Conditional means at the second location are represented by squares. In the unrestricted location mixture model group labels can be assigned in two nontrivial ways. The two respective labelings are represented by connecting clusters by solid lines and by dashed lines.

mixture model with probability parameters $\alpha_1 = .3, p_{11} = 2/3$, and $p_{21} = 3/7$. These two labelings, which provide equivalent expressions for (2.3), offer different views of the group structure of the data. Not only are the mixing parameters and the location probabilities different, but the relationships between the conditional means and the groups and locations also are different. In the first labeling, the difference between the group conditional means is the same at both locations (that is, there is parallel structure). In the second labeling, the group ordering of conditional means depends on the location (that is, there is group by location interaction). It seems that the best we can do with the unrestricted location mixture model is to obtain a separate cluster analysis within each location, and then use expert knowledge to assign group labels within locations.

For the case $m = 2, g = 2$ there are two distinct parameter sets providing equivalent expressions for any mixture representation (2.3). For the general case of m locations and g groups there are $(g!)^{m-1}$ distinct parameter sets. Let Ψ be any parameter set in the parameter space of model (2.3) with g groups, m locations and p continuous variables. Consider permuting group labels within locations. At each location there are g clusters, which can be assigned group labels in $g!$ ways. To avoid obtaining parameter sets that result from the same permutations of group indexes at all locations, the group labels at the first location are not permuted. There are $(g!)^{m-1}$ different ways to label the groups at the remaining $m - 1$ locations. Thus, (2.5) holds for distinct sets of parameters and it follows that (2.3) is not identifiable.

Unlike many non-identifiable models which have infinitely many parameter representations, the unrestricted location mixture model only has finitely many representations. Given a maximum likelihood solution of parameter estimates, $(g!)^{m-1} - 1$ other distinct solutions having equal likelihood can be obtained.

Example

Lawrence and Krzanowski (1996) conducted a simulation study to evaluate the ability of the unrestricted location mixture model to recover group structure and to classify observations. For each replication 20 observations were generated from each of two 4-variate normal populations, one with mean $(0,0,1,1)$ and the other with mean $(0,0,6,6)$. The populations had common covariance matrix

$$\boldsymbol{\Sigma} = \begin{pmatrix} 2 & 1 & 1 & 1 \\ 1 & 2 & 1 & 1 \\ 1 & 1 & 2 & 1 \\ 1 & 1 & 1 & 3 \end{pmatrix}.$$

The first two variables were dichotomized by thresholding at 0, giving a sample with 2 binary variables (or $m = 4$ locations) and 2 continuous variables. An observation with binary variables y_1 and y_2 was assigned to location $s = 1 + y_1 + 2y_2$.

The unrestricted location mixture model was fit for each of 50 replications. Cases were classified to groups by matching the recovered groups with the original (known) groups. The authors always chose that matching which yielded fewest misclassifications. The average misclassification rate for the location mixture model in the Lawrence and Krzanowski (1996) simulation was 31.4%. A baseline rate, for comparison, can be obtained as follows. Suppose that group assignments are made *randomly* with probability 1/2 for each group and the groups are matched to minimize misclassification rate. Then the expected misclassification rate for N observations can be shown to be

$$\frac{1}{2}\left(1 - \left(\frac{1}{2}\right)^{N-1} \binom{N}{[N/2]}\right),$$

where [ ] is the greatest integer function. For $N = 40$, the expected misclassification rate under random assignment is 44%. Estimates of the continuous variable means (presumably the group means were averaged over the four locations so that $\boldsymbol{\mu}_i = \sum_{s=1}^{4} p_{is}\boldsymbol{\mu}_{is}$) were found in the Lawrence and Krzanowski (1996) simulation to be

(2.52, 2.53) and (4.47, 4.47) with standard errors about .1. The authors attributed this excessive shrinkage of the parameter estimates from (1,1) and (6,6) toward the overall mean to the large number of misclassified individuals (which, in turn, could be attributed to shrinkage of parameter estimates). Three alternative explanations for the excessive shrinkage of parameter estimates are given next.

First, shrinkage of mean parameter estimates in mixture models is possible when the assumed form of the underlying group densities is incorrect. In the simulation study, the conditional distributions of the continuous variables given location and group are *not* multivariate normal. The underlying variable mixture model of Everitt (1988) — which assumes that the binary variables are obtained by dichotomizing underlying normal variables — is the correct model for this data.

Second, shrinkage can result from careless application of the EM algorithm. It is well known that log-likelihood surfaces for mixture models are often flat with many local maxima, so the EM algorithm should be applied many times with different starting parameter values to increase the chance of obtaining global maxima. The most common approach to obtain different starting values is to select each posterior probability $\tau_i(\mathbf{x}_{sh}, w_s; \mathbf{\Psi})$ uniformly on (0,1), and then standardize to satisfy the constraint $\sum_{i=1}^{g} \tau_i(\mathbf{x}_{sh}, w_s; \mathbf{\Psi}) = 1 \ \forall s, h$. Initial estimates for the mean parameters are obtained using equation (14) in Lawrence and Krzanowski (1996). These initial estimates of conditional means will all tend to be close to the overall mean (that is, shrinkage will be apparent in the initial estimates). If the EM algorithm isn't allowed to converge, or if the algorithm isn't re-run for enough starting values, shrinkage of mean parameter estimates may result. In their simulation experiment, the authors applied the EM algorithm with 50 different random starts for each replication. Though they didn't state their EM convergence criteria, it is plausible that they obtained global maxima for most or all replicates.

A third explanation for shrinkage is simply that the location mixture model is not identifiable. In fact, we can obtain the shrinkage estimates found by Lawrence and Krzanowski (1996) by averaging the (true) conditional means over all $(2!)^{4-1} = 8$ different parameterizations that yield equivalent location mixture models. Although the conditional distributions of the continuous variables are not MVN and have no apparent closed form expressions, conditional means and variances can be found by numerical integration. In one group the conditional means of continuous variables at the four locations are (.16, .16), (1.00, 1.00), (1.00, 1.00), and (1.84, 1.84) for locations 1,2,3 and 4. In the other group the conditional means are (5.16, 5.16), (6.00, 6.00),(6.00, 6.00), and (6.84, 6.84). The group conditional location probabilities are 1/3, 1/6, 1/6, and 1/3 in both groups. Within each group, overall means are obtained as a weighted average of the location conditional means, where the weights are the location probabilities. The overall means are (1.00, 1.00) and (6.00, 6.00) for the two groups. The true within location/group covariance matrix varies slightly among locations (if the data truly conformed to the location model there would be no differences among locations). The weighted average of the true covariance matrix over all locations is

$$\begin{pmatrix} 1.5 & .5 \\ .5 & 2.5 \end{pmatrix}.$$

Table 2 lists the conditional mean parameters for all $(2!)^{4-1} = 8$ permutations of group labels within locations. Permutations 2-8 were obtained by fixing group labels at location 1, and permuting group labels at locations 2-4. In this simplistic example the group/location cluster frequencies are the same for all permutations, so it follows from (2.6) that $\{\alpha_i\}$ and $\{p_{is}\}$ are the same for all permutations. In each permutation the group with the lowest overall mean, computed by $\sum_{s=1}^{4} p_{is}\boldsymbol{\mu}_{is}$, is labeled "low", and the group with the highest overall mean is labeled "high". The

Permutation	Group	Loc 1	Loc 2	Loc 3	Loc 4	Average
1 (true)	G_1	.16	1.00	1.00	1.84	1.00 (low)
	G_2	5.15	6.00	6.00	6.85	6.00 (high)
2	G_1	.16	6.00	1.00	1.84	1.83 (low)
	G_2	5.15	1.00	6.00	6.85	5.17 (high)
3	G_1	.16	1.00	6.00	1.84	1.83 (low)
	G_2	5.15	6.00	1.00	6.85	5.17 (high)
4	G_1	.16	1.00	1.00	6.85	2.67 (low)
	G_2	5.15	6.00	6.00	1.84	4.33 (high)
5	G_1	.16	6.00	6.00	1.84	2.67 (low)
	G_2	5.15	1.00	1.00	6.85	4.33 (high)
6	G_1	.16	6.00	1.00	6.85	3.50
	G_2	5.15	1.00	6.00	1.84	3.50
7	G_1	.16	1.00	6.00	6.85	3.50
	G_2	5.15	6.00	1.00	1.84	3.50
8	G_1	.16	6.00	6.00	6.85	4.34 (high)
	G_2	5.15	1.00	1.00	1.84	2.66 (low)
Average	low					2.46
	high					4.54
Simulation	low	se=.1				2.52
estimates	high	se=.1				4.47

Table 2: Continuous variable mean parameters for the eight permutations in simulation study. For all permutations group probabilities are 1/2, and location probabilities are 1/6, 1/3, 1/3, and 1/6 for locations 1, 2, 3 and 4. Simulation estimates are from Lawrence and Krzanowski (1996).

average means for the "low" and "high" groups over all permutations are 2.46 and 4.54. Lawrence and Krzanowski (1996) estimated the group means to be 2.52 and 4.47, with standard error about .1. Thus, they estimated well (within 1 se) the group means averaged over all permutations, although they intended to estimate the group means for the first permutation only. Apparently, the excessive shrinkage in their parameter estimates can be attributed to the non-identifiability of the model, which the authors did not mention in their paper.

In the next section identifiable location mixture models are obtained by imposing restrictions on the conditional mean parameters $\boldsymbol{\mu}_{is}$. We might expect an identifiable model to attain lower misclassification rates in the simulation example than the unrestricted, non-identifiable model. The next section confirms the expected result.

Restricted Location Mixture Models

All restricted models considered in this paper are obtained by constraining the conditional mean parameters, $\boldsymbol{\mu}_{is}$, so all models can be completely specified by their conditional mean structure. The unrestricted model will be denoted by $[\boldsymbol{\mu}_{is}]$.

A simple identifiable model can be obtained by imposing the restriction $\boldsymbol{\mu}_{is} = \boldsymbol{\mu}_i \ \forall i, s$. That model is denoted by $[\boldsymbol{\mu}_i]$. The model may be too restrictive, however, because it ignores any differences in conditional means across locations (i.e., the continuous variables are taken to be independent of the categorical variables). The restriction is relaxed in the additive model $[\boldsymbol{\mu}_i + \boldsymbol{\theta}_s]$ where $\boldsymbol{\theta}_1$ is taken to be $\mathbf{0}$. The parameter $\boldsymbol{\mu}_i$ is interpreted as the conditional mean of the continuous variable vector at location 1 of G_i, and $\boldsymbol{\theta}_s$ is the difference in the conditional means between location 1 and location s. The difference, $\boldsymbol{\theta}_s$, is assumed to be the same for all groups. This invariance of $\boldsymbol{\theta}_s$ across groups induces a *parallel structure* in the conditional means, where the difference between conditional means for any two groups is the same at all locations.

Next consider the identifiability of $[\boldsymbol{\mu}_i + \boldsymbol{\theta}_s]$. The structure of $[\boldsymbol{\mu}_i + \boldsymbol{\theta}_s]$ is not preserved under the permutations of group labels within locations (the source of non-identifiability of the unrestricted location mixture model). To see this, let π_s be a permutation of group labels $(1, \ldots, g)$ at location s $(s \neq 1)$, where $\pi_s(i)$ is the permuted value of the original group label i at w_s. No labels are permuted at location 1, so $\pi_1(i) = i \; \forall i$. The structure of the model $[\boldsymbol{\mu}_i + \boldsymbol{\theta}_s]$ is preserved only if, for each s, there is a unique $\boldsymbol{\theta}_s^*$ that satisfies

$$\boldsymbol{\mu}_{\pi_s(i)} + \boldsymbol{\theta}_s^* = \boldsymbol{\mu}_i + \boldsymbol{\theta}_s$$

for all i. There is no unique solution, because $\boldsymbol{\mu}_i - \boldsymbol{\mu}_{\pi_s(i)}$ can never be the same for all i (except in the degenerate case where the conditional means are the same for all groups). Thus the structure of $[\boldsymbol{\mu}_i + \boldsymbol{\theta}_s]$ is not preserved by the permutations. Although this does not constitute a formal proof of the identifiability of $[\boldsymbol{\mu}_i + \boldsymbol{\theta}_s]$, it does demonstrate that the type of non-identifiability revealed in the previous section for the unrestricted model is not possible with this restricted model.

The model $[\boldsymbol{\mu}_i + \boldsymbol{\theta}_s]$ can be written in the form $[\boldsymbol{\mu}_i + \mathbf{B}\mathbf{u}_s]$ where $\mathbf{u}_s$ is an $m - 1$ dimensional location covariate containing all main effect and interaction terms of the categorical variables at location s, and $\mathbf{B}$ is a $p \times (m-1)$ matrix of regression coefficients. For example, if there are three binary variables y_1, y_2 and y_3, then the observation (y_1, y_2, y_3) is assigned to location $s = 1 + \sum_{j=1}^{q} y_j 2^{j-1}$. If $(y_1, y_2, y_3) =$

(1, 1, 0), then $s = 4$ and

$$\mathbf{u}_4 = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_1y_2 \\ y_1y_3 \\ y_2y_3 \\ y_1y_2y_3 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}.$$

The regression matrix $\mathbf{B} = [\mathbf{b}_1, \ldots, \mathbf{b}_{m-1}]$ contains the same information as the location parameters $(\boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_m)$. For example, using $\boldsymbol{\theta}_s = \mathbf{B}\mathbf{u}_s$, it follows that $\boldsymbol{\theta}_4 = \mathbf{b}_1 + \mathbf{b}_2 + \mathbf{b}_4$ in the three binary variable case.

A more restrictive model is obtained if the location covariate vector contains main effects and possibly some – but not all – interaction terms. In this model, the location covariate vector $\mathbf{u}$ has length $r < m - 1$ and $\mathbf{B}$ is $p \times r$ (the location covariate vector containing all main effects and all interaction terms is called the *saturated* location covariate vector). A special case that will be considered in the examples is the main effects only model, denoted $[\boldsymbol{\mu}_i + \mathbf{B}\mathbf{y}_s]$. Because the models $[\boldsymbol{\mu}_i]$ and $[\boldsymbol{\mu}_i + \mathbf{B}\mathbf{y}_s]$ are obtained from $[\boldsymbol{\mu}_i + \boldsymbol{\theta}_s]$ by imposing constraints on the regression matrix $\mathbf{B}$, their identifiability follows from the identifiability of $[\boldsymbol{\mu}_i + \boldsymbol{\theta}_s]$.

Categorical variables with more than two levels can be handled by coding the category levels with dummy binary variables. Suppose there are q categorical variables and the j^{th} variable has c_j levels. For $j = 1, \ldots, q$ and $l = 1, \ldots, c_j - 1$ define the binary variable

$$y_j^{(l)} = \begin{cases} 1 \text{ if the } j^{th} \text{ variable is at level } l \\ 0 \text{ if the } j^{th} \text{ variable is not at level } l. \end{cases}$$

If all the binary variables are 0, the categorical variable is at level c_j. At most one of the variables $y_j^{(1)}, \ldots, y_j^{(c_j-1)}$ can be 1, so there can be no interactions among them. The *saturated* location covariate vector contains $\sum_{j=1}^{q}(c_j - 1)$ main effects (of dummy binary variables),$\sum\sum_{j<k}(c_j - 1)(c_k - 1)$ first order interaction terms, $\cdots$, and

$\prod_{j=1}^{q}(c_j-1)$ $(q-1)^{th}$ order interactions. So the saturated location vector has

$$\sum_{j=1}^{q}(c_j-1)+\sum\sum_{j<k}(c_j-1)(c_k-1)+\cdots+\prod_{j=1}^{q}(c_j-1)=\prod_{j=1}^{q}c_j-1=m-1$$

binary elements. If there are two categorical variables, each with three levels, then $(y_1^{(1)},y_1^{(2)},y_2^{(1)},y_2^{(2)})$ is assigned to location $s=1+\sum_{j=1}^{q}\sum_{l=1}^{c_j-1}y_j^{(l)}l\prod_{l=1}^{j-1}c_l$. If $y_1^{(1)}=1$ and $y_2^{(2)}=1$ then $s=8$ and

$$\mathbf{u}_8=\begin{pmatrix} y_1^{(1)}\\ y_1^{(2)}\\ y_2^{(1)}\\ y_2^{(2)}\\ y_1^{(1)}y_2^{(1)}\\ y_1^{(1)}y_2^{(2)}\\ y_1^{(2)}y_2^{(1)}\\ y_1^{(2)}y_2^{(2)} \end{pmatrix}=\begin{pmatrix}1\\0\\0\\1\\0\\1\\0\\0\end{pmatrix}.$$

The location covariate vector containing only main effects terms is, at location 8,

$$\mathbf{y}_8=\begin{pmatrix}1\\0\\0\\1\end{pmatrix}.$$

Whatever the choice of the location covariate $\mathbf{u}_s$, the model $[\boldsymbol{\mu}_i+\mathbf{B}\mathbf{u}_s]$ has structural similarity to the underlying variable mixture model proposed by Everitt (1988). The underlying variable mixture model assumes that the q categorical variables are obtained by thresholding q (unobservable) underlying continuous variables contained in $\mathbf{v}$, say. The unobservable variable $\mathbf{v}$ and the observable continuous variable $\mathbf{x}$ are assumed to be jointly multivariate normal, with common covariance matrix. The conditional expectation of $\mathbf{x}$ given $\mathbf{v}$ in G_i has the form

$$E(\mathbf{x}|\mathbf{v},G_i)=\boldsymbol{\mu}_i+\mathbf{B}\mathbf{v}, \tag{2.7}$$

which is the same form as the conditional expectation of $\mathbf{x}$ given location covariate $\mathbf{u}$ in G_i for the model $[\boldsymbol{\mu}_i+\mathbf{B}\mathbf{u}_s]$. After $\mathbf{v}$ is categorized the conditional distribution of $\mathbf{x}$

is no longer normal in the underlying variable model, and the conditional expectation no longer has form (2.7). Nonetheless, the restricted location models $[\boldsymbol{\mu}_i + \mathbf{B}\mathbf{u}_s]$ may still provide good approximations to the underlying variable mixture models. If the threshold values are the same for all groups (as in the Lawrence and Krzanowski simulation), then the conditional means will have parallel structure, and the restricted location model should provide an excellent approximation to Everitt's model. If the threshold values differ between groups, then the conditional means will not have parallel structure. There is a practical limit, however, to the range of values that the threshold parameters can take if we require that some observations be made at each location. Within this practical range (say between -1.5 and 1.5) the conditional mean structure may not deviate substantially from parallel structure, and the restricted location mixture models may still provide good approximations. An example of this is given later. If the parallel structure models don't provide adequate approximations, less restrictive models may be tried.

An even less restrictive model than $[\boldsymbol{\mu}_i + \boldsymbol{\theta}_s]$ can be obtained by allowing the regression matrices $\mathbf{B}$ to vary across groups, which gives the additive plus multiplicative model $[\boldsymbol{\mu}_i + \mathbf{B}_i\mathbf{u}_s]$. If the location covariate vector contains all main effects and all interaction terms, then $[\boldsymbol{\mu}_i+\mathbf{B}_i\mathbf{u}_s]$ is equivalent to the unrestricted location model, and hence is not identifiable. If at least one interaction term is excluded it can be shown that the structure $[\boldsymbol{\mu}_i+\mathbf{B}_i\mathbf{u}_s]$ is not, in general, preserved by the permutations π_s. But the structure *is* preserved for certain parameter values, which can lead to equivocal results in practice. For example, if differences between group means are the same at all locations (ie, when $[\boldsymbol{\mu}_i + \mathbf{B}\mathbf{u}_s]$ holds), then the structure of $[\boldsymbol{\mu}_i + \mathbf{B}_i\mathbf{u}_s]$ is preserved under the permutations discussed in the previous section, so the model is not identifiable. This is illustrated in Table 3 for 2 groups, 2 binary variables and 1 continuous variable. In representation A the difference between group means is 5 at

		Location			
	Group	1 (0,0)	2 (1,0)	3 (0,1)	4 (1,1)
Label A	G_1	0	2	2	4
	G_2	5	7	7	9
Label B	G_1	0	2	7	9
	G_2	5	7	2	4

Table 3: Continuous means for a single continuous variable conforming to the model $[\boldsymbol{\mu}_i + \mathbf{B}_i y]$ under two different group labelings. Conditional means at label B were obtained from conditional means at label A by swapping group labels at locations 3 and 4. The two labelings yield equivalent mixture densities. Label A also conforms to the model $[\boldsymbol{\mu}_i + \mathbf{B}y]$.

all locations. Corresponding mean parameter values are $\mu_1 = 0$, $\mu_2 = 5$, $\mathbf{B}_1 = (2, 2)$ and $\mathbf{B}_2 = (2, 2)$. Representation B is obtained by swapping group labels at locations 3 and 4. It also has structure $[\mu_i + \mathbf{B}_i \mathbf{y}_s]$ with parameters $\mu_1 = 0$, $\mu_2 = 5$, $\mathbf{B}_1 = (2, 7)$ and $\mathbf{B}_2 = (2, -3)$. Because of these identifiability problems, the model $[\boldsymbol{\mu}_i + \mathbf{B}_i \mathbf{u}_s]$ will not be pursued further in this thesis.

Estimation

Let $\mathbf{x} = (\mathbf{x}'_{11} \ldots \mathbf{x}'_{1n_1} \cdots \mathbf{x}'_{m1} \cdots \mathbf{x}'_{mn_m})'$ be a sample of p-dimensional continuous variables at m locations where n_s is the number of observations at w_s and $N = \sum_{s=1}^{m} n_s$ is the total number of observations. If observations are not made at each location, then we require that the rank of $\{\mathbf{u}_s\}_{w_s \in \text{sample}}$ be r, where $\mathbf{u}_s$ is $r \times 1$. Let $\mathbf{z}_{sh} = (z_{1sh}, \ldots, z_{gsh})$ be an unobservable g-dimensional group indicator vector for the h^{th} observation at w_s, so that $z_{ish} = 1$ if $\mathbf{x}_{sh} \in G_i$ and $z_{ish} = 0$ if $\mathbf{x}_{sh} \notin G_i$. Maximum likelihood estimates of the parameters in the model $[\boldsymbol{\mu}_i + \mathbf{B}\mathbf{u}_s]$ can be computed by treating z_{ish} as missing and using the EM algorithm.

The complete data log-likelihood is

$$L_c = \sum_{i=1}^{g} \sum_{s=1}^{m} \sum_{h=1}^{n_s} z_{ish} \{\log \alpha_i + \log p_{is} + \log h(\mathbf{x}_{sh}; \boldsymbol{\mu}_{is}, \boldsymbol{\Sigma})\}$$

where $h(\mathbf{x}_{sh}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the pdf of a $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ random variable evaluated at $\mathbf{x}_{sh}$, and

$\boldsymbol{\mu}_{is} = \boldsymbol{\mu}_i + \mathbf{B}\mathbf{u}_s$. In the E-step, we compute $Q = E_z^{\psi}(L_c)$ where the expectation is taken with respect to the conditional distribution of the unobserved data $\{\mathbf{z}_{sh}\}$ given the observed data and current parameter estimates $\boldsymbol{\Psi}$. Because L_c is linear in the unobserved data, the expectation is easily obtained by replacing each z_{ish} with $\hat{z}_{ish} = \tau_i(\mathbf{x}_{sh}, w_s; \hat{\boldsymbol{\Psi}})$, where

$$\tau_i(\mathbf{x}_{sh}, w_s; \boldsymbol{\Psi}) = \frac{\alpha_i p_{is} \exp\{-\frac{1}{2}(\mathbf{x}_{sh} - \boldsymbol{\mu}_{is})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}_{sh} - \boldsymbol{\mu}_{is})\}}{\sum_{l=1}^{g} \alpha_l p_{ls} \exp\{-\frac{1}{2}(\mathbf{x}_{sh} - \boldsymbol{\mu}_{ls})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}_{sh} - \boldsymbol{\mu}_{ls})\}} \tag{2.8}$$

is the posterior probability that $\mathbf{x}_{sh}$ belongs to G_i.

In the M-step, Q is maximized subject to the constraints $\sum_{i=1}^{g} \alpha_i = 1$ and $\sum_{s=1}^{m} p_{is} = 1\ \forall i$. Using the method of Lagrange multipliers we maximize without constraint the expression

$$Q' = Q - \lambda \left(\sum_{i=1}^{g} \alpha_i - 1 \right) - \sum_{i=1}^{g} \gamma_i \left(\sum_{s=1}^{m} p_{is} - 1 \right)$$

where λ and $\{\gamma_i\}$ are Lagrange multipliers. This yields updated probability parameter estimates

$$\hat{\alpha}_i = \frac{1}{N} \sum_{s=1}^{m} \sum_{h=1}^{n_s} \hat{z}_{ish} \tag{2.9}$$

and

$$\hat{p}_{is} = \frac{1}{N\hat{\alpha}_i} \sum_{h=1}^{n_s} \hat{z}_{ish}. \tag{2.10}$$

Estimating equations for the parameters $\boldsymbol{\mu}_i$ and $\mathbf{B}$ are

$$N\hat{\alpha}_i \boldsymbol{\mu}_i = \sum_{s=1}^{m} \sum_{h=1}^{n_s} \hat{z}_{ish}(\mathbf{x}_{sh} - \mathbf{B}\mathbf{u}_s), \qquad i = 1, \ldots, m \tag{2.11}$$

and

$$\mathbf{B} \sum_{i=1}^{g} \sum_{s=1}^{m} \sum_{h=1}^{n_s} \hat{z}_{ish} \mathbf{u}_s \mathbf{u}_s' = \sum_{i=1}^{g} \sum_{s=1}^{m} \sum_{h=1}^{n_s} \hat{z}_{ish}(\mathbf{x}_{sh} - \boldsymbol{\mu}_i)\mathbf{u}_s'. \tag{2.12}$$

M-step estimates for $\boldsymbol{\mu}_i$ $(i = 1, \ldots, g)$ and $\mathbf{B}$ can be found by solving (2.11) and (2.12) simultaneously.

The solution of estimating equations (2.11) and (2.12) is

$$\hat{\mathbf{B}} = (\mathbf{A} - \frac{1}{N}\mathbf{G})(\mathbf{E} - \frac{1}{N}\mathbf{F})^{-1} \tag{2.13}$$

and

$$\hat{\boldsymbol{\mu}}_i = \frac{1}{N\hat{\alpha}_i}(\mathbf{c}_i - \hat{\mathbf{B}}\mathbf{d}_i), \qquad i = 1, \ldots, g, \tag{2.14}$$

where

$$\begin{aligned}
\mathbf{c}_i &= \sum_{s=1}^{m}\sum_{h=1}^{n_s} \hat{z}_{ish}\mathbf{x}_{sh}, \qquad i = 1, \ldots, g \\
\mathbf{d}_i &= \sum_{s=1}^{m}\sum_{h=1}^{n_s} \hat{z}_{ish}\mathbf{u}_{s}, \qquad i = 1, \ldots, g \\
\mathbf{A} &= \sum_{i=1}^{g}\sum_{s=1}^{m}\sum_{h=1}^{n_s} \hat{z}_{ish}\mathbf{x}_{sh}\mathbf{u}_s' \\
\mathbf{E} &= \sum_{i=1}^{g}\sum_{s=1}^{m}\sum_{h=1}^{n_s} \hat{z}_{ish}\mathbf{u}_{s}\mathbf{u}_s' \\
\mathbf{F} &= \sum_{i=1}^{g} \frac{1}{\hat{\alpha}_i}\mathbf{d}_i\mathbf{d}_i' \\
\mathbf{G} &= \sum_{i=1}^{g} \frac{1}{\hat{\alpha}_i}\mathbf{c}_i\mathbf{d}_i'.
\end{aligned} \tag{2.15}$$

The covariance matrix is estimated as

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{N}\sum_{i=1}^{g}\sum_{s=1}^{m}\sum_{h=1}^{n_s} \hat{z}_{ish}(\mathbf{x}_{sh} - \hat{\boldsymbol{\mu}}_i - \hat{\mathbf{B}}\mathbf{u}_s)(\mathbf{x}_{sh} - \hat{\boldsymbol{\mu}}_i - \hat{\mathbf{B}}\mathbf{u}_s)'. \tag{2.16}$$

Parameter estimates for the models $[\boldsymbol{\mu}_i + \boldsymbol{\theta}_s]$ and $[\boldsymbol{\mu}_i + \mathbf{B}\mathbf{y}_s]$ can be obtained by appropriate choice of the location covariates $\mathbf{u}_s$. Parameter estimates for the model $[\boldsymbol{\mu}_i]$ can be obtained by setting $\hat{\mathbf{B}} = \mathbf{0}$ in (2.14).

The EM algorithm alternately updates (2.8) (E-step) and (2.9)-(2.16) (M-step). The procedure requires starting values for the iterations. Starting values can be obtained by randomly selecting posterior probabilities uniformly on (0,1), and then standardizing to satisfy $\sum_{i=1}^{g} \tau_i(\mathbf{x}_{sh}, w_s; \boldsymbol{\Psi}) = 1 \; \forall s, h$. Alternatively, the sample can be partitioned into g groups and initial parameter estimates computed using

(2.9)-(2.16) assuming group labels are known (ie, $z_{ish} \in \{0,1\}$). Ideally, this initial partition would be found by another cluster analysis method, perhaps using only observations on the continuous variables. Because of the possibility of multiple local maxima, the EM algorithm should be applied several times from different starting values.

The development and estimation of the restricted location mixture models assumes that the number of groups, g, is known. In practice g is often unknown and a statistical heuristic such as Bayes Information Criterion (BIC) can be employed to aid the choice of g. This heuristic suggests selecting the model for which

$$\text{BIC} = -2(\text{maximized log-likelihood}) + 2\log(\text{N})(\text{number of free parameters})$$

is a minimum. In applications, use of BIC should be balanced with expert judgement. The difficult problem of choosing the number of clusters is not pursued here. In the following examples, the number of groups is assumed known.

Examples

Two simulation experiments were run to assess the performance of the new methods. The experiments are described next.

Simulation 1

The simulation example of Lawrence and Krzanowski (1996) was revisited to compare the performance of the three nested models $[\boldsymbol{\mu}_i] \subset [\boldsymbol{\mu}_i + \mathbf{B}\mathbf{y}_s] \subset [\boldsymbol{\mu}_i + \boldsymbol{\theta}_s]$ in parameter recovery and classification. For each of 50 replications, the EM algorithm was applied 11 times: 10 times with randomly selected starting values and once with starting values determined by classification assignments from an initial k-means cluster analysis of the continuous variables. The solution with the largest log-

	$[\boldsymbol{\mu}_i]$	$[\boldsymbol{\mu}_i + \mathbf{By}]$	$[\boldsymbol{\mu}_i + \boldsymbol{\theta}_s]$	$[MVN_2]$
mean	1.82 (4.55%)	.64 (1.6%)	.79 (2.03%)	1.36 (3.4%)
median	1 (2.5%)	1 (2.5%)	1 (2.56%)	1 (2.5%)
minimum	0	0	0	0
maximum	9 (22.5%)	3 (7.5%)	3 (7.69%)	6 (15%)

Table 4: Misclassifications for Simulation Experiment 1 ($n_1 = n_2 = 20$). Results for $[\boldsymbol{\mu}_i + \boldsymbol{\theta}_s]$ are based on 49 replications (see text).

likelihood value was retained. One simulated dataset (of $n_1 + n_2 = 40$ observations) contained no observations from location 3. This did not affect the estimation of $[\boldsymbol{\mu}_i]$ or $[\boldsymbol{\mu}_i + \mathbf{By}_s]$, but it did affect the estimation of $[\boldsymbol{\mu}_i + \boldsymbol{\theta}_s]$. Infinite parameter estimates were obtained, because the data were silent about $\boldsymbol{\theta}_3$. This replicate is omitted in the summary statistics reported for $[\boldsymbol{\mu}_i + \boldsymbol{\theta}_s]$.

A mixture model with multivariate normal component densities and homogeneous variance was fit for the two continuous variables for comparison. This model is denoted $[\mathrm{MVN}_2]$. When only the two continuous variables are used, the true misclassification rate is 2.6%. If the two latent continuous variables were observable, and parameters known, then the true misclassification rate would be .62% (the first two variables, though marginally distributed the same in both groups, enhance group separation due to their correlations with the last two variables). The true misclassification rate under Everitt's (1988) model was estimated by Monte Carlo simulation to be 1.1%.

Misclassification rates for the simulations are compared in Table 4. All methods performed well. The models $[\boldsymbol{\mu}_i + \mathbf{By}_s]$ and $[\boldsymbol{\mu}_i + \boldsymbol{\theta}_s]$ performed slightly better than the others.

Tables 5 and 6 compare average estimates of the parameters $\{\boldsymbol{\mu}_{is}\}$ and $\{p_{is}\}$. True parameter values are also given (though we should not forget that the location model is not the correct model for these data – the continuous variables are not

Model	Group	location 1 $(0,0)$	location 2 $(1,0)$	location 3 $(0,1)$	location 4 $(1,1)$
$[\boldsymbol{\mu}_i]$	G_1	(5.96, 5.97)	(5.96, 5.97)	(5.96, 5.97)	(5.96, 5.97)
	G_2	(1.00, .98)	(1.00, .98)	(1.00, .98)	(1.00, .98)
		se≈.10			
$[\boldsymbol{\mu}_i + \mathbf{By}]$	G_1	(5.22, 5.21)	(5.99, 6.03)	(6.06, 6.00)	(6.83, 6.83)
	G_2	(.17, .15)	(.93, .97)	(1.01, .94)	(1.77, 1.76)
		se≈.10			
$[\boldsymbol{\mu}_i + \boldsymbol{\theta}_s]$	G_1	(5.26, 5.22)	(6.03, 6.16)	(6.04, 6.10)	(6.88, 6.85)
	G_2	(.17, .10)	(.94, 1.05)	(.95, .98)	(1.78, 1.73)
		se≈.10			
true values	G_1	(5.16, 5.16)	(6.00, 6.00)	(6.00, 6.00)	(6.84, 6.84)
	G_2	(.16, .16)	(1.00, 1.00)	(1.00, 1.00)	(1.84, 1.84)

Table 5: Average estimates (and their standard errors) and true values of conditional means for Simulation Experiment 1 ($n_1 = n_2 = 20$).

conditionally MVN). The models $[\boldsymbol{\mu}_i + \mathbf{By}_s]$ and $[\boldsymbol{\mu}_i + \boldsymbol{\theta}_s]$ recover the parameters well. They also recover the within group/ location covariance matrix better than the model $[\boldsymbol{\mu}_i]$ does. The true value of the covariance matrix is

$$\begin{pmatrix} 1.5 & .5 \\ .5 & 2.5 \end{pmatrix}.$$

The average estimate for model $[\boldsymbol{\mu}_i + \mathbf{By}_s]$ was

$$\begin{pmatrix} 1.46 & .47 \\ .47 & 2.21 \end{pmatrix}$$

with standard error about .05 for all entries. The average estimate for model $[\boldsymbol{\mu}_i + \boldsymbol{\theta}_s]$ was

$$\begin{pmatrix} 1.41 & .46 \\ .46 & 2.16 \end{pmatrix}$$

with standard error about .05 for all entries.

Simulation 2

A second simulation experiment was performed to assess the models on less well separated groups. Observations were generated from one of two 4-variate normal

Model	Group	location 1 $(0,0)$	location 2 $(1,0)$	location 3 $(0,1)$	location 4 $(1,1)$
$[\boldsymbol{\mu}_i]$	G_1	.31	.18	.17	.35
	G_2	.37	.19	.16	.28
		se≈.02			
$[\boldsymbol{\mu}_i + \mathbf{By}]$	G_1	.34	.18	.17	.31
	G_2	.33	.18	.16	.32
		se≈.10			
$[\boldsymbol{\mu}_i + \boldsymbol{\theta}_s]$	G_1	.34	.18	.17	.31
	G_2	.33	.18	.17	.32
		se≈.10			
true values	G_1	.33	.17	.17	.33
	G_2	.33	.17	.17	.33

Table 6: Average estimates (and their standard errors) and true values of location probabilities $\{p_{is}\}$ for Simulation Experiment 1 ($n_1 = n_2 = 20$).

populations, one with mean (1,0,5,5) and one with mean (0,1,2,2). The populations had common covariance matrix

$$\boldsymbol{\Sigma} = \begin{pmatrix} 2 & 1 & 1 & 1 \\ 1 & 2 & 1 & 1 \\ 1 & 1 & 2 & 1 \\ 1 & 1 & 1 & 3 \end{pmatrix}.$$

As in the first experiment, the first two binary variables were dichotomized by thresholding at 0.

This is equivalent (using Everitt's convention) to sampling from multivariate normal populations with means (0,0,5,5) and (0,0,2,2) and common covariance matrix

$$\begin{pmatrix} 1 & \frac{1}{2} & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{2} & 1 & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 2 & 1 \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 1 & 3 \end{pmatrix},$$

and thresholding at $-1/\sqrt{2}$ and 0 for the two underlying variables in the first group, and at 0 and $-1/\sqrt{2}$ in the second group. This interpretation emphasizes that the threshold values are different for the two groups.

For each of 50 replications, samples of size $n_1 = n_2 = 100$ were drawn from the two populations. Misclassification rates are compared in Table 7. The models

	$[\boldsymbol{\mu}_i]$	$[\boldsymbol{\mu}_i + \mathbf{By}]$	$[\boldsymbol{\mu}_i + \boldsymbol{\theta}_s]$	$[MVN_2]$
mean	42.9 (21.45%)	18.38 (9.15%)	19.84 (9.92%)	28.62 (14.31%)
median	37.5 (18.75%)	18 (9.00%)	17.5 (8.75%)	27.5 (13.75%)
minimum	16 (8.0%)	7 (3.5%)	7 (3.5%)	13 (6.5%)
maximum	82 (41%)	36 (18%)	64 (32.0%)	57 (28.5%)

Table 7: Misclassifications for Simulation Experiment 2 ($n_1 = n_2 = 100$).

$[\boldsymbol{\mu}_i + \mathbf{By}_s]$ and $[\boldsymbol{\mu}_i + \boldsymbol{\theta}_s]$ performed best. Their respective mean misclassification rates of 9.15% and 9.92% are lower than the true (or optimal) misclassification rate for [MVN$_2$], which is 12.26%. The realized mean misclassification rate for [MVN$_2$] was 14.31%. The true misclassification rate under Everitt's (1988) model (assuming parameters are known) was estimated by Monte Carlo simulation to be 7.2%.

Discussion

The unrestricted location mixture model proposed by Lawrence and Krzanowski (1996) is not identifiable. The identifiable models proposed in this paper can be useful if the additive assumption (ie, $\boldsymbol{\mu}_{is} = \boldsymbol{\mu}_i + \boldsymbol{\theta}_s$) is reasonable. This assumption is often approximately true when the categorical variables are derived from underlying continuous variables. Computation in the restricted models is more tractable than computation in Everitt's (1988) underlying variable model, which in practice is limited to one or two categorical variables. Estimation of the parameters in the restricted models does not require numerical integration, so there is no computational limit to the number of categorical variables that the model can handle (though there is the practical limit of sample size).

The restricted location mixture models can be profitably extended in two directions. First, the categorical variables can be more parsimoniously modeled, perhaps with loglinear or latent class models. This is particularly important when the sample is small or boundary value solutions for p_{is} are obtained. Second, the homogeneous variance assumption can be relaxed by allowing the group/ location dispersion matrix to vary across groups, locations, or both. Parsimonious representations can be obtained by imposing structure on the dispersion matrices. Celeux and Govaert (1995) describe a parsimonious parameterization of multivariate normal mixture models with unequal group dispersion matrices based on eigenvalue decomposition of the group dispersion matrices. This approach can be extended to location mixture models.

CHAPTER 3

Conditional Gaussian Discriminant Analysis with Constraints on the Covariance Matrices

Krzanowski (1975, 1980, 1993) developed parametric methods for discriminant analysis with mixed categorical and continuous variables. He assumed that within each group, observations conform to a conditional Gaussian distribution. In the conditional Gaussian model, the continuous variables have a different multivariate normal distribution at each possible combination of categorical variable values. This model has received much attention recently in the graphical models literature (Whittaker, 1990).

Suppose we wish to discriminate between K groups, $G_1, \ldots, G_K$, based on the vector $\mathbf{w}' = (\mathbf{y}', \mathbf{x}')$, where $\mathbf{y}' = (y_1, \ldots, y_q)$ is a vector of q categorical variables, and $\mathbf{x}' = (x_1, \ldots, x_p)$ is a vector of p continuous variables. The categorical variables can be uniquely transformed to an m-state discrete variable $w \in \{w_1, \ldots, w_m\}$, where m is the number of distinct combinations (i.e., locations) of the categorical variable values, and w_s is the label for the s^{th} location. If the j^{th} variable has c_j categories $(j = 1, \ldots, q)$, then $m = \prod_{j=1}^{q} c_j$. Let $p_{is} = Pr(w = w_s | G_i)$. In G_i, the joint probability of observing location w_s and continuous vector $\mathbf{x}$ is

$$g_i(w_s, \mathbf{x}) = p_{is} h(\mathbf{x}; \boldsymbol{\mu}_{is}, \boldsymbol{\Sigma}_{is}),$$

where $h(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the pdf of a $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ random variable. An observation $(w_s, \mathbf{x})$ is assigned to a group according to Bayes Rule (Anderson, 1984). If misclassification costs are equal and prior group probabilities are given by $\alpha_1, \ldots, \alpha_K$, then Bayes Rule is:

$$\text{assign } (w_s, \mathbf{x}) \text{ to } G_i \text{ if } \max_{1 \leq t \leq K} \alpha_t g_t(w_s, \mathbf{x}) = \alpha_i g_i(w_s, \mathbf{x}). \tag{3.1}$$

Taking the log of $\alpha_i g_i(w_s, \mathbf{x})$, the classification region for group G_i can be written as

$$\mathbf{R}_i = \{w \in \{w_1, \cdots, w_m\}, \mathbf{x} \in \Re^p : q_{is}(\mathbf{x}) \geq q_{ts}(\mathbf{x}) \quad \forall t = 1, \ldots, K\}$$

where the classification functions $q_{is}(\mathbf{x})$ are given by

$$q_{is}(\mathbf{x}) = \mathbf{x}'\mathbf{A}_{is}\mathbf{x} + \mathbf{b}'_{is}\mathbf{x} + c_{is}$$

with

$$\mathbf{A}_{is} = -\frac{1}{2}\mathbf{\Sigma}_{is}^{-1}, \qquad \mathbf{b}_{is} = \mathbf{\Sigma}_{is}^{-1}\boldsymbol{\mu}_{is}$$

$$c_{is} = \log \alpha_i + \log p_{is} - \frac{1}{2}\log|\mathbf{\Sigma}_{is}| - \frac{1}{2}\boldsymbol{\mu}'_{is}\mathbf{\Sigma}_{is}^{-1}\boldsymbol{\mu}_{is}.$$

The classification rule depends on the parameters $p_{is}, \boldsymbol{\mu}_{is}$ and $\mathbf{\Sigma}_{is}$, which usually are unknown. In practical applications, parameter estimates obtained from a training sample of classified observations are substituted in (3.1). Because these estimates are subject to sampling error, the classification rule (i.e., plug-in Bayes Rule) is no longer optimal. The performance of the classification rule depends on the precision of the estimates (Flury, Schmid and Narayanan, 1994). More efficient parameter estimates can be obtained by imposing constraints on the parameter space. For example, in normal theory (Gaussian) discriminant analysis, the covariance matrices often are assumed to be the same for all groups. In the conditional Gaussian setting, Krzanowski (1975, 1980, 1993) took the covariance matrices to be the same across all groups and locations (i.e., $\mathbf{\Sigma}_{is} = \mathbf{\Sigma}\ \forall i, s$), so that

$$g_i(w_s, \mathbf{x}) = p_{is}h(\mathbf{x}; \boldsymbol{\mu}_{is}, \mathbf{\Sigma}). \tag{3.2}$$

Model (3.2) is called the *homogeneous* conditional Gaussian model in the graphical models literature, and the location model in the statistics literature.

For a training sample, let the p dimensional vector $\mathbf{x}_{ish}$ denote the h^{th} continuous observation at location w_s of group G_i, and let n_{is} denote the number of observations made at location w_s of G_i. The total number of observations from G_i is given by n_i, and the total number of observations over all groups is N. Maximum likelihood estimates of the parameters in (3.2) are given by

$$\hat{p}_{is} = \frac{n_{is}}{n_i}, \qquad \hat{\boldsymbol{\mu}}_{is} = \bar{\mathbf{x}}_{is} = \frac{1}{n_{is}} \sum_{h=1}^{n_{is}} \mathbf{x}_{ish} \tag{3.3}$$

and

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{N} \sum_{i=1}^{g} \sum_{s=1}^{m} \sum_{h=1}^{n_{is}} (\mathbf{x}_{ish} - \hat{\boldsymbol{\mu}}_{is})(\mathbf{x}_{ish} - \hat{\boldsymbol{\mu}}_{is})'.$$

An unbiased estimate of the covariance matrix,

$$\tilde{\boldsymbol{\Sigma}} = \frac{N}{N - mK} \hat{\boldsymbol{\Sigma}},$$

often is used in place of $\hat{\boldsymbol{\Sigma}}$.

Sometimes additional constraints on the parameter space are necessary. When the sample size is small compared to the number of locations, there will likely be locations for which no data are present in the training sets. Also, there will be some locations with very few individuals present in the training sample; the parameters for these locations will be poorly estimated. To obtain reasonable parameter estimates at all locations in this case, Krzanowski (1975, 1980) proposed that the categorical data be modeled with a reduced-order loglinear model. In his applications he used either first-order (main-effects only) or second-order (main effects and first-order interaction) models. If the categorical data consists of q binary variables, then the second-order loglinear model for probability of location w_s in group G_i is

$$\log p_{is} = \boldsymbol{\theta}_i' \mathbf{u}_{p,s}$$

where $\mathbf{u}_{p,s}$ is a *location covariate vector* for p_{is} containing an intercept term and the values of all main effects and first order interactions of the binary variables at the s^{th}

location. The subscript p is a reminder that the location covariate vector is for p_{is}. For example, if there are three binary variables y_1, y_2 and y_3, then the observation $(y_1, y_2, y_3) = (1, 1, 0)$ is assigned to location $s = 4$ using the location assignment rule $s = 1 + \sum_{j=1}^{3} y_j 2^{j-1}$, and

$$\mathbf{u}_{p,4} = \begin{pmatrix} 1 \\ y_1 \\ y_2 \\ y_3 \\ y_1 y_2 \\ y_1 y_3 \\ y_2 y_3 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 0 \\ 1 \\ 0 \\ 0 \end{pmatrix}.$$

The parameters $\{\boldsymbol{\theta}_i\}$ can be estimated using Newton-Raphson methods. Categorical variables with more than two levels can be similarly handled by coding the category levels with dummy binary variables (Krzanowski, 1980).

Likewise, the continuous mean vector can be modeled as a linear function of the location covariate vector $\mathbf{u}_{\mu,s}$:

$$\boldsymbol{\mu}_{is} = \mathbf{B}_i \mathbf{u}_{\mu,s}.$$

The parameters $\{\mathbf{B}_i\}$ can be estimated independently of $\boldsymbol{\theta}_i$ using multivariate regression results of Anderson (1984, chapter 8). Details are given in Krzanowski (1975). The location covariate vector used in the model for p_{is} need not be the same as that used in the model for $\boldsymbol{\mu}_{is}$. For example, $\mathbf{u}_{p,s}$ could code for main effects only whereas $\mathbf{u}_{\mu,s}$ could code for main effects as well as first order interactions.

Because of the homogeneous variance assumption $\boldsymbol{\Sigma}_{is} = \boldsymbol{\Sigma}$ in (3.2), a separate linear discriminant analysis is conducted at each location. Thus, discriminant analysis based on model (3.2) shall be referred to as L-LDA (for linear location discriminant analysis). L-LDA has been shown to outperform competing methods when there is interaction between the groups and the categorical variables (Krzanowski, 1993).

In applications, the homogeneous variance assumption, though parsimonious, may not be realistic. Krzanowski (1993) suggested that models with heterogeneous variances be developed to cater to various types of dispersion heterogeneity. Later, Krzanowski (1994) considered the consequences of allowing the dispersion matrices to differ between groups, but not between locations within a group (i.e.,$\Sigma_{is} = \Sigma_i$). In Krzanowski's (1994) model, a separate *quadratic* discriminant analysis is performed at each location. In an example with a relatively small sample and heterogeneous variances, Krzanowski (1994) found that the quadratic location discriminant analysis (Q-LDA) performed only slightly better than L-LDA, reflecting the tradeoff between fitting a more appropriate model but estimating many more parameters.

This is a familiar problem in Gaussian discriminant analysis. Linear discriminant analysis (LDA) outperforms quadratic discriminant analysis (QDA) when group covariance matrices are identical (i.e., when the model assumptions for LDA are correct). But even when group covariance matrices are not identical, LDA may still outperform QDA, especially when sample sizes are modest. This suggests that, for small samples, the bias introduced by imposing theoretically wrong constraints may be offset by the gain in precision from reducing the number of parameters (Flury, Schmid and Narayanan, 1994).

Several authors have proposed intermediate methods that avoid both the overparameterization of QDA and the oversimplification of LDA. Such methods attempt to capture the heterogeneity of the covariance matrices using as few parameters as possible. Friedman (1989) designed an intermediate classifier between LDA, QDA, and the nearest neighbor classifier by introducing regularization parameters. Flury, Schmid and Narayanan (1994) considered common principal components and proportional covariance models. More recently, Bensmail and Celeux (1996) developed intermediate models by parameterizing the covariance matrix for G_i in terms of its eigenvalue decomposition $\mathbf{\Sigma}_i = \rho_i \mathbf{\Gamma}_i \mathbf{\Lambda}_i \mathbf{\Gamma}_i'$, where $\rho_i = |\Sigma_i|^{1/p}$, $\mathbf{\Gamma}_i$ is the orthogonal matrix of eigenvalues of $\mathbf{\Sigma}_i$, and $\mathbf{\Lambda}_i$ is the diagonal matrix such that $|\mathbf{\Lambda}_i| = 1$, with the normalized eigenvalues of $\mathbf{\Sigma}_i$ on the diagonal in decreasing order. The parameter ρ_i determines the volume of the probability contours of G_i, $\mathbf{\Gamma}_i$ determines its orientation and $\mathbf{\Lambda}_i$ determines its shape. Intermediate, or regularized, models are obtained by allowing some but not all of these quantities to vary between groups. Common principal components ($\mathbf{\Sigma}_i = \rho_i \mathbf{\Gamma} \mathbf{\Lambda}_i \mathbf{\Gamma}'$) and proportional covariance models ($\mathbf{\Sigma}_i = \rho_i \mathbf{\Gamma} \mathbf{\Lambda} \mathbf{\Gamma}'$) are special cases of this approach. Bensmail and Celeux (1996) found that such intermediate models often outperform both LDA and QDA. Flury, Schmid and Narayanan (1994) found that proportional covariance discrimination performed well in a variety of situations. Even when the assumptions for LDA were correct, proportional covariance discrimination didn't do much worse that LDA.

In this paper we extend the singular value decomposition approach to regularization to the conditional Gaussian model for discriminant analysis with mixed-mode data. Our goals are 1) to discover the extent to which regularized models can outperform L-LDA and Q-LDA and 2) to explore parsimonious models that allow dispersion matrices to differ between locations. We will express the within-cell dispersion matrices as $\mathbf{\Sigma}_{is} = \rho_{is} \mathbf{\Gamma}_{is} \mathbf{\Lambda}_{is} \mathbf{\Gamma}'_{is}$, and we will hold some of the geometric quantities invariant across locations and/or groups. For example, in the model $\mathbf{\Sigma}_{is} = \rho_i \mathbf{\Gamma} \mathbf{\Lambda}_s \mathbf{\Gamma}'$, the volume parameter ρ_i varies between groups but not between locations, the shape parameter $\mathbf{\Lambda}_s$ varies between locations but not groups, and the orientation $\mathbf{\Gamma}$ is invariant to both location and group. This model will be denoted by $[\rho_i \mathbf{\Gamma} \mathbf{\Lambda}_s \mathbf{\Gamma}']$. We will also consider the diagonal family of covariance matrices, where $\mathbf{\Sigma}_{is} = \rho_{is} \mathbf{\Lambda}_{is}$, with $\mathbf{\Gamma}_{is} = \mathbf{I}$, and the spherical family, where $\mathbf{\Sigma}_{is} = \rho_{is} \mathbf{I}$.

These three families of models are described more fully in the next section. In addition, two parsimonious models that allow covariance matrices to differ between locations are derived. In the first model, loglinear constraints are placed on the geometric parameters ρ_{is} and $\mathbf{\Lambda}_{is}$. In the second model, a discrete latent variable (which defines latent classes) is introduced to simplify the conditional structure of the model. Maximum likelihood estimates of the parameters for these models are derived. The regularized models are compared with L-LDA and Q-LDA. Finally, other possible approaches to regularized discriminant analysis are discussed.

Models

By allowing each of the geometric quantities to vary by group, location, neither, or both, we can obtain 64 models from the general SVD family $\mathbf{\Sigma}_{is} = \rho_{is}\mathbf{\Gamma}_{is}\mathbf{\Lambda}_{is}\mathbf{\Gamma}'_{is}$, 16 models from the diagonal family $\mathbf{\Sigma}_{is} = \rho_{is}\mathbf{\Lambda}_{is}$, and 4 models from the spherical family $\mathbf{\Sigma}_{is} = \rho_{is}\mathbf{I}$. A total of 84 models are possible. For a given data set, we might select the model that minimizes the sample-based estimate of future misclassification risk. Thus, to avoid excessive computation, it may be desirable to reduce the number of models under consideration. To obtain parsimonious models, it is reasonable to omit from consideration those models involving the greatest number of parameters. The orientation $\mathbf{\Gamma}_{is}$ of a probability contour is described by $p(p-1)/2$ functionally independent parameters, the shape $\mathbf{\Lambda}_{is}$ is described by $p-1$ functionally independent parameters, and the size is described by a single parameter ρ_{is}. The most parsimonious models are obtained by holding $\mathbf{\Gamma}_{is}$, and possibly $\mathbf{\Lambda}_{is}$, invariant across locations and groups.

One strategy is to consider only those models that satisfy the following conditions.

1. At least one geometric feature is invariant to both location and group.
2. Only the size parameter is allowed to vary across both locations and groups.

3. If orientation varies by location or group, then shape must be invariant to location and group.

The first two conditions apply to all three families; the third condition applies only to the general family $\mathbf{\Sigma}_{is} = \rho_{is}\mathbf{\Gamma}_{is}\mathbf{\Lambda}_{is}\mathbf{\Gamma}'_{is}$. This strategy reduces the number of models under consideration to 30. Table 8 lists all 30 models, and gives the number of functionally independent covariance parameters for K groups, m locations and p continuous variables. The first model, $[\rho\mathbf{\Gamma}\mathbf{\Lambda}\mathbf{\Gamma}']$, is the traditional (homogeneous covariance) location model, which leads to L-LDA. In the next five models (M2–M6), the dispersion matrices are invariant to location. These models represent compromises between L-LDA and Q-LDA.

The next five models (M7–M11) are identical to models M2–M6, except their geometric features differ between locations but not groups. These models result in separate linear discriminant analysis at each location. In models M12–M20 the dispersion matrices differ between locations and groups. Models in which the orientation $\mathbf{\Gamma}_{is}$ differs between locations and groups generally involve a large number of parameters. Proportional covariance models (where only ρ_{is} varies) are generally the most parsimonious. In the diagonal models, the orientations $\mathbf{\Gamma}_{is}$ are identity matrices, which don't require estimation. In the spherical models, the orientations are not identified and can be assumed to be identity matrices without loss of generality. Hence, the diagonal and spherical models contain fewer parameters than the SVD models.

In this chapter we give special attention to the following geometric shapes that have shown promise in Gaussian discriminant analysis.

- (homogeneous covariance) $[\rho\mathbf{\Gamma}\mathbf{\Lambda}\mathbf{\Gamma}']$ and $[\rho\mathbf{\Lambda}]$.
- (proportional covariance) $[\rho_i\mathbf{\Gamma}\mathbf{\Lambda}\mathbf{\Gamma}']$ and $[\rho_i\mathbf{\Lambda}]$. Flury, Schmid, and Narayanan (1994) recommended that proportional discrimination be tried whenever the

Model	Number of parameters	$m = 4$ $K = 2$ $p = 5$	Comments
M1.) $[\rho\mathbf{\Gamma}\mathbf{\Lambda}\mathbf{\Gamma}']$	$p(p+1)/2$	**15**	L-LDA
M2.) $[\rho_i\mathbf{\Gamma}\mathbf{\Lambda}\mathbf{\Gamma}']$	$K-1+p(p+1)/2$	**16**	proportional cov.
M3.) $[\rho_i\mathbf{\Gamma}_i\mathbf{\Lambda}\mathbf{\Gamma}'_i]$	$K+Kp(p-1)/2+p-1$	26	
M4.) $[\rho_i\mathbf{\Gamma}\mathbf{\Lambda}_i\mathbf{\Gamma}']$	$Kp+p(p-1)/2$	**20**	CPC
M5.) $[\rho\mathbf{\Gamma}_i\mathbf{\Lambda}\mathbf{\Gamma}'_i]$	$p+Kp(p-1)/2$	25	
M6.) $[\rho\mathbf{\Gamma}\mathbf{\Lambda}_i\mathbf{\Gamma}']$	$1+p(p-1)/2+K(p-1)$	19	
M7.) $[\rho_s\mathbf{\Gamma}\mathbf{\Lambda}\mathbf{\Gamma}']$	$m-1+p(p+1)/2$	18	proportional cov.
M8.) $[\rho_s\mathbf{\Gamma}_s\mathbf{\Lambda}\mathbf{\Gamma}'_s]$	$m+mp(p-1)/2+p-1$	48	
M9.) $[\rho_s\mathbf{\Gamma}\mathbf{\Lambda}_s\mathbf{\Gamma}']$	$mp+p(p-1)/2$	30	CPC
M10.) $[\rho\mathbf{\Gamma}_s\mathbf{\Lambda}\mathbf{\Gamma}'_s]$	$p+mp(p-1)/2$	45	
M11.) $[\rho\mathbf{\Gamma}\mathbf{\Lambda}_s\mathbf{\Gamma}']$	$1+m(p-1)+p(p-1)/2$	27	
M12.) $[\rho_i\mathbf{\Gamma}_s\mathbf{\Lambda}\mathbf{\Gamma}'_s]$	$K+mp(p-1)/2+p-1$	46	
M13.) $[\rho_i\mathbf{\Gamma}\mathbf{\Lambda}_s\mathbf{\Gamma}']$	$K+p(p-1)/2+m(p-1)$	28	
M14.) $[\rho_s\mathbf{\Gamma}_i\mathbf{\Lambda}\mathbf{\Gamma}'_i]$	$m+p-1+Kp(p-1)/2$	28	
M15.) $[\rho_s\mathbf{\Gamma}\mathbf{\Lambda}_i\mathbf{D}']$	$m+K(p-1)+p(p-1)/2$	22	
M16.) $[\rho_{is}\mathbf{\Gamma}\mathbf{\Lambda}\mathbf{\Gamma}']$	$mK-1+p(p+1)/2$	22	proportional cov.
M17.) $[\rho_{is}\mathbf{\Gamma}_i\mathbf{\Lambda}\mathbf{\Gamma}'_i]$	$mK+p-1+Kp(p-1)/2$	32	
M18.) $[\rho_{is}\mathbf{\Gamma}_s\mathbf{\Lambda}\mathbf{\Gamma}'_s]$	$mK+p-1+mp(p-1)/2$	52	
M19.) $[\rho_{is}\mathbf{\Gamma}\mathbf{\Lambda}_i\mathbf{\Gamma}']$	$mK+K(p-1)+p(p-1)/2$	26	
M20.) $[\rho_{is}\mathbf{\Gamma}\mathbf{\Lambda}_s\mathbf{\Gamma}']$	$mK+m(p-1)+p(p-1)/2$	34	
M21.) $[\rho_i\mathbf{\Lambda}]$	$K+p-1$	6	proportional cov.
M22.) $[\rho_s\mathbf{\Lambda}]$	$m+p-1$	8	proportional cov.
M23.) $[\rho_{is}\mathbf{\Lambda}]$	$mK+p-1$	12	proportional cov.
M24.) $[\rho\mathbf{\Lambda}_i]$	$1+K(p-1)$	9	
M25.) $[\rho\mathbf{\Lambda}_s]$	$1+m(p-1)$	17	
M26.) $[\rho\mathbf{\Lambda}]$	p	5	
M27.) $[\rho\mathbf{I}]$	1	1	
M28.) $[\rho_i\mathbf{I}]$	K	2	proportional cov.
M29.) $[\rho_s\mathbf{I}]$	m	4	proportional cov.
M30.) $[\rho_{is}\mathbf{I}]$	mK	8	proportional cov.
other models:			
M31.) $[\mathbf{\Sigma}_{is}]$	$mKp(p+1)/2$	120	
M32.) $[\mathbf{\Sigma}_s]$	$mp(p+1)/2$	60	
M33.) $[\mathbf{\Sigma}_i]$	$Kp(p+1)/2$	**30**	Q-LDA

Table 8: Some constrained covariance models, and the number of functionally independent covariance parameters for m locations, K groups, and p continuous variables. The third column gives the number of covariance parameters for $m = 4, K = 2$ and $p = 5$.

assumption of equality of covariance matrices seems questionable. In the worst case – if the covariance matrices are equal – proportional discrimination may perform slightly worse than linear discrimination. When covariance matrices are unequal, proportional discrimination often outperforms both linear discrimination and more theoretically correct quadratic discrimination methods with moderate sample sizes.

- (common principal components) $[\rho_i \mathbf{\Gamma} \mathbf{\Lambda}_i \mathbf{\Gamma}']$.
- (general heterogeneous covariance) $[\mathbf{\Sigma}_i]$. This results in the quadratic location discriminant analysis (Q-LDA) considered by Krzanowski (1994).

These models can be extended to allow the geometric features to vary between locations.

- (proportional covariance) $[\rho_{is} \mathbf{\Gamma} \mathbf{\Lambda} \mathbf{\Gamma}']$ and $[\rho_{is} \mathbf{\Lambda}]$. These models allow for a group by location interaction in the volume of the covariance matrix.
- (common principal components-like) $[\rho_i \mathbf{\Gamma} \mathbf{\Lambda}_s \mathbf{\Gamma}']$. Cluster volumes differ between groups, and cluster shapes differ between locations.

Other models from the three families are possible, but they will generally include more parameters than those models listed in Table 8. If sample sizes are large, then we should try to find the model that best fits the data (because this will generally lead to better discrimination). But if samples are small or moderate, a premium should be placed on parsimonious models.

More Parsimonious Covariance Models for Location

When the number of locations is large (this is common in practice — 5 binary variables define 32 locations), models which allow geometric features to vary between locations can contain an exhorbitant number of parameters. In this section we consider models which allow a more parsimonious representation. Two approaches will be considered.

1. (Reduced Models) Loglinear restrictions can be imposed on the geometric parameters.
2. (Latent Class Models) The locations can be clustered using latent class models, with the geometric features homogeneous within latent classes.

Reduced Models

The first approach is an extension of the reduced location model, where location probabilities and conditional means are modeled as functions of location covariates. We also can place loglinear restrictions on the geometric parameters. These models result in smoothed estimates for the geometric parameters, in much the same way that loglinear models produce smooth estimates of the probability parameters. We will consider loglinear restrictions for the volume parameter, ρ_{is}, and the shape parameter, $\mathbf{\Lambda}_{is}$.

If $\mathbf{u}_{\rho,s}$ is a known $r \times 1$ location covariate vector containing an intercept term, main effects and possibly some interaction terms of the categorical variables at location s, then a reduced model for ρ_{is} is

$$\rho_{is} = \exp(\mathbf{a}_i'\mathbf{u}_{\rho,s})$$

or

$$\log \rho_{is} = \mathbf{a}_i'\mathbf{u}_{\rho,s}, \tag{3.4}$$

where $\mathbf{a}_i$ $(i = 1, \ldots, K)$ are unknown regression coefficients. The exponential parameterization ensures that the volume parameter ρ_{is} is positive.

A restricted model for the diagonal matrix $\mathbf{\Lambda}_{is}$ is

$$\mathbf{\Lambda}_{is} = \text{diag}\{\exp(\mathbf{b}'_{ij}\mathbf{u}_{\lambda,s})\}_{j=1}^{p}, \tag{3.5}$$

where $\mathbf{b}_{ij}$ $(i = 1, \ldots, K; j = 1, \ldots, p)$ are unknown regression coefficients. The exponential parameterization ensures that the diagonal elements of $\mathbf{\Lambda}_{is}$ are positive (which is required because the covariance matrix $\mathbf{\Sigma}_{is}$ is positive definite). To satisfy the constraint $|\mathbf{\Lambda}_{is}| = 1$, the parameters $\mathbf{b}_{ij}$ $(i = 1, \ldots, K; j = 1, \ldots, p)$ must satisfy the constraint

$$\sum_{j=1}^{p} \mathbf{b}'_{ij}\mathbf{u}_{\lambda,s} = 0 \quad (i = 1, \ldots, K; s = 1, \ldots, m).$$

This constraint is not required in the common principal components model $[\rho_{is}\mathbf{\Gamma}\mathbf{\Lambda}_{is}\mathbf{\Gamma}']$ if we define, for computational convenience, $\mathbf{A}_{is} = \rho_{is}\mathbf{\Lambda}_{is}$, and impose the loglinear restriction on $\mathbf{A}_{is}$:

$$\mathbf{A}_{is} = \text{diag}\{\exp(\mathbf{b}'_{ij}\mathbf{u}_{\rho,s})\}_{j=1}^{p}.$$

The performance of the loglinear restricted CPC model will be studied in Section 4.

These reduced models can lead to significant parameter savings. Consider, for example, the proportional covariance model $[\rho_{is}\mathbf{\Gamma}\mathbf{\Lambda}\mathbf{\Gamma}']$. For $K = 3$ groups, $q = 5$ binary variables (hence $m = 32$ locations), and $p = 5$ continuous variables, the model requires estimation of 110 covariance parameters. The first-order reduced model (3.5) requires estimation of only 32 covariance parameters. The second-order reduced model (3.5) requires estimation of 62 covariance parameters.

Latent Class Models

When there are many locations, it may be prudent to reduce their number. Latent class analysis (LCA) is one way to do this. An extensive review of LCA, including an exhaustive bibliography and a review of software can be found at John Ubersax' home page (http://members.xoom.com/XOOM/jubersax).

Latent class analysis is a statistical method for analyzing multivariate categorical data. It has been widely used in psychiatry, sociology, and medical diagnosis applications. LCA was motivated by the desire to find subtypes of related cases (i.e., latent classes) in a set of observations. For example, if the observations are vectors of binary variables, where all variables are considered to be indicators of some disease, a latent class analysis might partition the observations into two classes: diseased and undiseased. Thus, LCA can be considered a method for clustering categorical data. The clustering is performed using finite mixture models (Titterington, Smith and Makov, 1985). In the location model, we can use LCA to reduce many locations to a few new locations, defined by the latent classes.

In LCA, latent classes are defined by the criterion of conditional independence. The observed variables are taken to be independent within latent classes. Conditional independence models have fared well in comparative studies of discriminant analysis methods for categorical data. In naive Bayes discriminant analysis, variables are taken to be statistically independent within each group. Chang (1980) found naive Bayes discriminant analysis to perform as well as or better than several other methods for classifying multivariate binary observations. An advantage of this approach is its simplicity, and the relatively small number of parameters that need to be estimated. Naive Bayes is still the preferred method for supervised classification of textual data (Dumais, et. al., 1998).

In some applications, the assumptions of naive Bayes may be so unrealistic that the model doesn't perform well. In a comparative study of methods for classifying head injury patients based on categorical observations, Titterington, et. al. (1981) found that classification by naive Bayes could be improved if groups were partitioned into subclasses (latent classes) so that the variables are statistically independent within each subclass. This LCA approach performed better than all other methods for this particular dataset. Latent class discriminant analysis has not been pursued much in the literature. Dillon and Mulani (1989) developed latent class discriminant analysis models for market research applications, where the primary goal is to understand the relationships between the subtypes of consumers (latent classes) and the products they bought (groups). Their model allowed for both categorical and continuous variables and all variables were taken to be mutually independent within a latent class. In the following, we will exploit latent class models in several different ways. In some special cases the resulting discriminant analysis method will be equivalent to that of Dillon and Mulani (1989), but in general they will be different.

Before describing the discrimination methods, we first describe multiple group latent class analysis. For simplicity of notation, we take all categorical variables to be binary. The generalization to polytomous variables is straightforward. For binary variables $\mathbf{y}_i = (y_{i1}, \cdots, y_{iq})'$ from group G_i, a latent class model with T latent classes, $C_1, \ldots, C_T$, is given by

$$f_i(\mathbf{y}) = \sum_{t=1}^{T} \eta_{it} \prod_{j=1}^{q} \pi_{jt}^{y_{ij}} (1 - \pi_{jt})^{1-y_{ij}} \tag{3.6}$$

where $\eta_{it} = Pr(C_t|G_i)$, $\sum_{t=1}^{T} \eta_{it} = 1$, and $\pi_{jt} = Pr(Y_j = 1|C_t)$. The number of latent classes T is generally much smaller than q. In this model, the conditional response probabilities, defined by π_{jt}, are the same for all groups, but the class sizes, defined by η_{it}, are different between groups. Other options are possible for this multiple group latent class model (see Clogg, 1993, for details). In our model, the T latent classes are common to all groups. Note that model (3.6) is a finite mixture model, where the variables in each mixture component are independent.

There are several ways that LCA can be incorporated into the mixed-mode discrimination problem. We will consider four of them here.

1. Substantive LCA. In some LCA applications, the latent classes have real, phys-

ical meaning. For example, in medical diagnostics, often there is no gold standard for patient diagnosis. The patient must be diagnosed based on imperfect indicators of disease. LCA is a natural model for diagnosis when the disease indicators are independent within a diagnostic group. The latent classes are literally interpreted as diseased or undiseased groups. In addition to classification, the latent class model provides estimates of diagnostic accuracy (e.g., sensitivity, specificity, positive predictive value). If, in the mixed-mode discrimination problem, the categorical variables are indeed imperfect indicators of diagnostic status (and the groups $G_1, \ldots, G_K$ are not diagnostic groups but correspond to some other classification), then LCA is a natural model for the categorical data. In this case, the number of latent classes might be known in advance. The latent classes might be considered new locations. The continuous variables – not necessarily indicators of disease status – may or may not be independent of the categorical variables conditional on disease status (i.e., latent class membership). We shall keep that option open.

Within G_i, the model is

$$g_i(\mathbf{y}, \mathbf{x}) = \sum_{t=1}^{T} \eta_{it} \prod_{j=1}^{q} \pi_{jt}^{y_{ij}} (1 - \pi_{jt})^{1-y_{ij}} h(\mathbf{x}|\boldsymbol{\mu}_{ist}, \boldsymbol{\Sigma}_{ist}). \tag{3.7}$$

The conditional distribution of the continuous variables is allowed, in general, to depend on group, location, and latent class. In most applications it is probably unnecessary to condition the continuous variables on both latent class and location, since latent classes and locations contain much of the same information. If we take $\boldsymbol{\mu}_{ist} = \boldsymbol{\mu}_{it}$, and $\boldsymbol{\Sigma}_{ist} = \boldsymbol{\Sigma}_{it}$, then we are assuming that $\mathbf{x}$ and $\mathbf{y}$ are conditionally independent, given latent class membership. If we take $\boldsymbol{\mu}_{ist} = \boldsymbol{\mu}_{is}$, and $\boldsymbol{\Sigma}_{ist} = \boldsymbol{\Sigma}_{is}$, then the conditional density h can be pulled outside of the summation and estimated independently of the categorical data. In this case, the latent class model can be considered an alternative to the loglinear model for the categorical data. This is considered next.

2. LCA as an alternative to loglinear model. In the reduced location models loglinear restrictions are placed on the location probabilities

$$\log p_{is} = \boldsymbol{\theta}_i' \mathbf{u}_{p,s},$$

primarily to reduce the number of parameters that need to be estimated. In some cases – for example, if the categorical variables are all indicators of disease status – it might be more reasonable to model the location probabilities with latent class models. When the latent classes are common to all groups, this LCA approach can result in significant reduction of parameters. Consider, for example, the case of $q = 5$ binary variables, all indicators of a particular disease, so that there are $T = 2$ latent classes. For $K = 2$ groups, the first order loglinear model requires estimation of 10 parameters, the second order loglinear model requires estimation of 20 parameters, and the latent class model requires estimation of 11 parameters. For $K = 4$ groups, the first order loglinear model requires estimation of 20 parameters, the second order loglinear model requires estimation of 40 parameters, and the latent class model requires estimation of just 13 parameters.

3. LCA as a tool for dimension reduction. LCA is sometimes used as a method for data reduction – that is, for reducing a large number of categorical variables to a more manageable number. Using this approach, we perform mixed-mode discriminant analysis in two steps. In the first step, multiple group LCA is performed on the categorical variables, and the observation $\mathbf{y}_h$ from G_i is allocated to the latent class for which it has greatest posterior probability,

$$Pr(C_t|\mathbf{y}_h, G_i) = \frac{\eta_{it} \prod_{j=1}^{q} \pi_{jt}^{y_{jh}} (1-\pi_{jt})^{1-y_{jh}}}{\sum_{l=1}^{T} \eta_{il} \prod_{j=1}^{q} \pi_{jl}^{y_{jh}} (1-\pi_{jl})^{1-y_{jh}}}.$$

Note that locations may be allocated to latent classes differently across groups because latent class probabilities are different between groups. In the second step, we apply the discriminant analysis procedures previously described, using latent class assignments as new "locations". In this approach, LCA is used as a pre-processing step prior to mixed-mode discriminant analysis.

4. LCA and LPA for nonparametric density estimation. Mixture models have been used as nonparametric density estimators. In a discriminant analysis application with continuous data, Hastie, Tibshirani, and Buja (1997) modeled group-conditional densities with multivariate normal finite mixture models. The rationale of the method is that in many cases, a single prototype (i.e., mean value) is not sufficient to describe the data within a group.

   We can take a related approach for mixed-mode data. We model the group-conditional density of the mixed-mode data with a finite mixture model. We take variables within each mixture component to be statistically independent. This results in a latent class model for the categorical data and a latent profile model for the continuous data. In general, our model is a latent structure model. The conditional independence assumption may seem severe, but we assume that any statistical dependence between the variables can still be captured at the cost of a potential large number of components. The number of components will be chosen by cross-validation. Within group G_i the density is

$$f_i(\mathbf{y},\mathbf{x}) = \sum_{t=1}^{T} \eta_{it} \left[\prod_{j=1}^{q} \pi_{ij}^{y_j} (1-\pi_{ij})^{1-y_j} \right] \left[\prod_{l=1}^{p} (2\pi\sigma_{lit}^2)^{-1/2} \exp\left(-\frac{1}{2\sigma_{lit}^2} (x_l - \mu_{lit})^2 \right) \right].$$

   When $\sigma_{jit}^2 = \sigma_{jt}^2$, this model reduces to the model of Dillon and Mulani (1989).

Estimation

This section is divided into three parts. In the first part, maximum likelihood estimates for covariance parameters in the full conditional Gaussian model are derived. In the second part, parameter estimation is described for reduced conditional Gaussian models. In the third part, estimation details are given for latent class location models.

Full Models

In the full model, maximum likelihood estimates for the location probabilities and conditional mean parameters are given by (3.3). Maximum likelihood estimates for the covariance parameters are obtained – independently of the probability and conditional mean parameter estimates – by minimizing the objective function

$$F = \sum_{i=1}^{K}\sum_{s=1}^{m} n_{is} \log|\mathbf{\Sigma}_{is}| + \sum_{i=1}^{K}\sum_{s=1}^{m} \mathrm{tr}(\mathbf{\Sigma}_{is}^{-1}\mathbf{W}_{is}).$$

In terms of the geometric features $\rho_{is}, \mathbf{\Gamma}_{is}$, and $\mathbf{\Lambda}_{is}$, the objective function is

$$F = p\sum_{i=1}^{K}\sum_{s=1}^{m} n_{is} \log \rho_{is} + \sum_{i=1}^{K}\sum_{s=1}^{m} \frac{1}{\rho_{is}} \mathrm{tr}(\mathbf{\Gamma}_{is}\mathbf{\Lambda}_{is}^{-1}\mathbf{\Gamma}_{is}'\mathbf{W}_{is}),$$

where

$$\mathbf{W}_{is} = \sum_{h=1}^{n_{is}} (\mathbf{x}_{ish} - \hat{\boldsymbol{\mu}}_{is})(\mathbf{x}_{ish} - \hat{\boldsymbol{\mu}}_{is})'$$

is the within location/group scatter matrix. Similarly, define the location, group, and overall scatter matrices as

$$\mathbf{W}_{\cdot s} = \sum_{i=1}^{K} \mathbf{W}_{is},$$

$$\mathbf{W}_{i\cdot} = \sum_{s=1}^{m} \mathbf{W}_{is},$$

and

$$\mathbf{W} = \sum_{i=1}^{K}\sum_{s=1}^{m} \mathbf{W}_{is}.$$

We first outline a general iterative procedure for estimating the geometric parameters ρ_{is}, $\mathbf{\Gamma}_{is}$ and $\mathbf{\Lambda}_{is}$ in the SVD family. The procedure reduces to closed form solutions in special cases. In other cases the procedure can be simplified. Appendix B describes estimation details for some common models.

- For fixed $\mathbf{\Gamma}_{is}$ and $\mathbf{\Lambda}_{is}$,

$$\hat{\rho}_{is} = \frac{1}{pn_{is}}\mathrm{tr}(\mathbf{\Gamma}_{is}\mathbf{\Lambda}_{is}^{-1}\mathbf{\Gamma}'_{is}\mathbf{W}_{is}).$$

If $\rho_{is} = \rho_i$, then

$$\hat{\rho}_i = \frac{1}{pn_{i\cdot}}\sum_{s=1}^{m}\mathrm{tr}(\mathbf{\Gamma}_{is}\mathbf{\Lambda}_{is}^{-1}\mathbf{\Gamma}'_{is}\mathbf{W}_{is}).$$

If $\rho_{is} = \rho_s$, then

$$\hat{\rho}_s = \frac{1}{pn_{\cdot s}}\sum_{i=1}^{K}\mathrm{tr}(\mathbf{\Gamma}_{is}\mathbf{\Lambda}_{is}^{-1}\mathbf{\Gamma}'_{is}\mathbf{W}_{is}).$$

If $\rho_{is} = \rho$, then

$$\hat{\rho} = \frac{1}{pN}\sum_{i=1}^{K}\sum_{s=1}^{m}\mathrm{tr}(\mathbf{\Gamma}_{is}\mathbf{\Lambda}_{is}^{-1}\mathbf{\Gamma}'_{is}\mathbf{W}_{is}).$$

- For fixed ρ_{is} and $\mathbf{\Gamma}_{is}$, the mle of $\mathbf{\Lambda}_{is}$ minimizes the function

$$f(\mathbf{\Lambda}_{is}) = \frac{1}{\rho_{is}}\mathrm{tr}(\mathbf{\Gamma}_{is}\mathbf{\Lambda}_{is}^{-1}\mathbf{\Gamma}'_{is}\mathbf{W}_{is}) = \frac{1}{\rho_{is}}\mathrm{tr}(\mathbf{\Lambda}_{is}^{-1}\mathbf{\Gamma}'_{is}\mathbf{W}_{is}\mathbf{\Gamma}_{is}).$$

By Corollary 1 of Appendix A,

$$\hat{\mathbf{\Lambda}}_{is} = \frac{\mathrm{diag}(\mathbf{\Gamma}'_{is}\mathbf{W}_{is}\mathbf{\Gamma}_{is})}{|\mathrm{diag}(\mathbf{\Gamma}'_{is}\mathbf{W}_{is}\mathbf{\Gamma}_{is})|}.$$

If $\mathbf{\Lambda}_{is} = \mathbf{\Lambda}_i$, then

$$f(\mathbf{\Lambda}_i) = \mathrm{tr}(\mathbf{\Lambda}_i^{-1}\sum_{s=1}^{m}\frac{1}{\rho_{is}}\mathbf{\Gamma}'_{is}\mathbf{W}_{is}\mathbf{\Gamma}_{is}),$$

and by Corollary 1 of Appendix A,

$$\hat{\mathbf{\Lambda}}_i = \frac{\sum_{s=1}^{m}\frac{1}{\rho_{is}}\mathbf{\Gamma}'_{is}\mathbf{W}_{is}\mathbf{\Gamma}_{is}}{|\sum_{s=1}^{m}\frac{1}{\rho_{is}}\mathbf{\Gamma}'_{is}\mathbf{W}_{is}\mathbf{\Gamma}_{is}|^{1/p}}.$$

Similarly, if $\mathbf{\Lambda}_{is} = \mathbf{\Lambda}_s$, then

$$\hat{\mathbf{\Lambda}}_s = \frac{\sum_{i=1}^K \frac{1}{\rho_{is}} \mathbf{\Gamma}'_{is} \mathbf{W}_{is} \mathbf{\Gamma}_{is}}{|\sum_{i=1}^K \frac{1}{\rho_{is}} \mathbf{\Gamma}'_{is} \mathbf{W}_{is} \mathbf{\Gamma}_{is}|^{1/p}}.$$

If $\mathbf{\Lambda}_{is} = \mathbf{\Lambda}$, then

$$f(\mathbf{\Lambda}) = \text{tr}(\mathbf{\Lambda}^{-1} \sum_{i=1}^K \sum_{s=1}^m \frac{1}{\rho_{is}} \mathbf{\Gamma}'_{is} \mathbf{W}_{is} \mathbf{\Gamma}_{is}),$$

and by Corollary 1 of Appendix A,

$$\hat{\mathbf{\Lambda}} = \frac{\sum_{i=1}^K \sum_{s=1}^m \frac{1}{\rho_{is}} \mathbf{\Gamma}'_{is} \mathbf{W}_{is} \mathbf{\Gamma}_{is}}{|\sum_{i=1}^K \sum_{s=1}^m \frac{1}{\rho_{is}} \mathbf{\Gamma}'_{is} \mathbf{W}_{is} \mathbf{\Gamma}_{is}|^{1/p}}.$$

- For fixed ρ_{is} and $\mathbf{\Lambda}_{is}$, the mle of $\mathbf{\Gamma}_{is}$ minimizes the function

$$f(\mathbf{\Gamma}_{is}) = \text{tr}(\mathbf{\Gamma}_{is} \mathbf{\Lambda}_{is}^{-1} \mathbf{\Gamma}'_{is} \mathbf{W}_{is}) = \text{tr}(\mathbf{L}'_{is} \mathbf{\Gamma}_{is} \mathbf{\Lambda}_{is}^{-1} \mathbf{\Gamma}'_{is} \mathbf{L}_{is} \mathbf{\Omega}_{is}),$$

where $\mathbf{W}_{is} = \mathbf{L}_{is} \mathbf{\Omega}_{is} \mathbf{L}'_{is}$ is the eigenvalue decomposition of $\mathbf{W}_{is}$. It follows from Theorem 3 of Appendix A that $\hat{\mathbf{\Gamma}}_{is} = \mathbf{L}_{is}$, which doesn't depend on $\mathbf{\Lambda}_{is}$ or ρ_{is}. If $\mathbf{\Gamma}_{is} = \mathbf{\Gamma}_i$, then

$$f(\mathbf{\Gamma}_i) = \sum_{s=1}^m \text{tr}(\mathbf{\Gamma}_i \mathbf{\Lambda}_{is}^{-1} \mathbf{\Gamma}'_i \mathbf{W}_{is} / \rho_{is}),$$

and $\hat{\mathbf{\Gamma}}_i$ can be obtained using Theorem 4, which is described in Appendix A. Similarly, if $\mathbf{\Gamma}_{is} = \mathbf{\Gamma}_s$, then

$$f(\mathbf{\Gamma}_s) = \sum_{i=1}^K \text{tr}(\mathbf{\Gamma}_s \mathbf{\Lambda}_{is}^{-1} \mathbf{\Gamma}'_s \mathbf{W}_{is} / \rho_{is}),$$

and $\hat{\mathbf{\Gamma}}_s$ can be obtained using Theorem 4. If $\mathbf{\Gamma}_{is} = \mathbf{\Gamma}$, then

$$f(\mathbf{\Gamma}) = \sum_{i=1}^K \sum_{s=1}^m \text{tr}(\mathbf{\Gamma} \mathbf{\Lambda}_{is}^{-1} \mathbf{\Gamma}' \mathbf{W}_{is} / \rho_{is}),$$

and $\hat{\mathbf{\Gamma}}$ can be obtained using Theorem 4.

Reduced Models

In this section we consider the estimation of models with restrictions on one or more of the parameters p_{is}, $\boldsymbol{\mu}_{is}$ or $\boldsymbol{\Sigma}_{is}$. The estimation details presented in this section are valid for any choice of location covariates $\mathbf{u}_{p,s}$, $\mathbf{u}_{\mu,s}$, $\mathbf{u}_{\lambda,s}$ and $\mathbf{u}_{\rho,s}$.

The log-likelihood function is

$$L = \sum_{i=1}^{K} \sum_{s=1}^{m} \sum_{h=1}^{n_{is}} \left[\boldsymbol{\theta}_i' \mathbf{u}_{p,s} - \frac{1}{2} \log |\boldsymbol{\Sigma}_{is}| - \frac{1}{2} (\mathbf{x}_{ish} - \mathbf{B}_i \mathbf{u}_{\mu,s})' \boldsymbol{\Sigma}_{is}^{-1} (\mathbf{x}_{ish} - \mathbf{B}_i \mathbf{u}_{\mu,s}) \right]$$

where $\boldsymbol{\Sigma}_{is}$ also may be modeled as a function of location covariates. The log-likelihood is maximized subject to the constraint on the probability parameters, $\sum_{s=1}^{m} p_{is} = \sum_{s=1}^{m} \exp(\boldsymbol{\theta}_i' \mathbf{u}_s) = 1 \ \forall i$.

The location probability regression coefficients $\boldsymbol{\theta}_i$ can be estimated independently of the other parameters using Newton-Raphson methods (McCullagh and Nelder, 1989, Chapter 6).

The conditional mean and covariance parameters can be estimated using the following iterative procedure, which successively estimates $\mathbf{B}_i$ conditional on $\boldsymbol{\Sigma}_{is}$, and then estimates $\boldsymbol{\Sigma}_{is}$ conditional on $\mathbf{B}_i$. The procedure yields closed-form solutions in special cases.

(Updating $\mathbf{B}_i$ given $\boldsymbol{\Sigma}_{is}$) Conditional on $\boldsymbol{\Sigma}_{is}$, the maximum likelihood estimate of $\mathbf{B}_i$ is the solution of the equation

$$\sum_{s=1}^{m} \boldsymbol{\Sigma}_{is}^{-1} \mathbf{x}_{is\cdot} \mathbf{u}_{\mu,s}' = \sum_{s=1}^{m} n_{is} \boldsymbol{\Sigma}_{is}^{-1} \mathbf{B}_i \mathbf{u}_{\mu,s} \mathbf{u}_{\mu,s}'.$$

Applying the vec operator to both sides of the equation, it follows that

$$\mathrm{vec}\hat{\mathbf{B}}_i = \left[\sum_{s=1}^{m} \left(n_{is} \mathbf{u}_{\mu,s} \mathbf{u}_{\mu,s}' \otimes \boldsymbol{\Sigma}_{is}^{-1} \right) \right]^{-1} \mathrm{vec} \left(\sum_{s=1}^{m} \boldsymbol{\Sigma}_{is}^{-1} \mathbf{x}_{is\cdot} \mathbf{u}_{\mu,s}' \right)$$

where

$$\mathbf{x}_{is\cdot} = \sum_{h=1}^{n_{is}} \mathbf{x}_{ish}.$$

This expression can be simplified for special cases.

- If the covariance matrix does not differ between locations (i.e., if $\mathbf{\Sigma}_{is} = \mathbf{\Sigma}_i$ or $\mathbf{\Sigma}_{is} = \mathbf{\Sigma}$), or if $\mathbf{u}_{\mu,s}$ is saturated, then the *unconditional* maximum likelihood estimate of $\mathbf{B}_i$ is

$$\hat{\mathbf{B}}_i = \left(\sum_{s=1}^{m} \mathbf{x}_{is\cdot} \mathbf{u}'_{\mu,s} \right) \left(\sum_{s=1}^{m} n_{is} \mathbf{u}_{\mu,s} \mathbf{u}'_{\mu,s} \right)^{-1} .$$

- In the proportional covariance model $[\lambda_{is}\mathbf{DAD}'] = [\lambda_{is}\mathbf{C}]$,

$$\hat{\mathbf{B}}_i = \left(\sum_{s=1}^{m} \frac{1}{\lambda_{is}} \mathbf{x}_{is\cdot} \mathbf{u}'_{\mu,s} \right) \left(\sum_{s=1}^{m} \frac{1}{\lambda_{is}} n_{is} \mathbf{u}_{\mu,s} \mathbf{u}'_{\mu,s} \right)^{-1} .$$

(Updating $\mathbf{\Sigma}_{is}$ given $\mathbf{B}_i$) The general approach outlined earlier applies to the conditional estimation of $\mathbf{\Sigma}_{is}$, where the scatter matrices $\mathbf{W}_{is}$ are computed using current estimates of $\mathbf{B}_i$. Modifications are needed for geometric features with loglinear restrictions. This will be illustrated for two common models – the proportional covariance model and the common principal components model.

- In the proportional covariance model $[\rho_{is}\mathbf{C}]$, where $\log \rho_{is} = \mathbf{a}'_i \mathbf{u}_{\Sigma,s}$, the objective function is

$$F = p \sum_{i=1}^{K} \sum_{s=1}^{m} n_{is} (\mathbf{a}'_i \mathbf{u}_{\Sigma,s}) + \sum_{i=1}^{K} \sum_{s=1}^{m} \exp(-\mathbf{a}'_i \mathbf{u}_{\Sigma,s}) \mathrm{tr}(\mathbf{C}^{-1} \mathbf{W}_{is}).$$

Maximum likelihood estimates of $\mathbf{a}_i$ (for fixed $\mathbf{C}$) can be obtained using the Newton-Raphson procedure

$$\hat{\mathbf{a}}_i^{\text{new}} = \hat{\mathbf{a}}_i^{\text{old}} - \mathbf{H}_i^{-1} \mathbf{g}_i,$$

with

$$\mathbf{g}_i = \sum_{s=1}^{m} \left[p n_{is} - \exp(-\mathbf{a}'_i \mathbf{u}_{\rho,s}) \mathrm{tr}(\mathbf{C}^{-1} \mathbf{W}_{is}) \right] \mathbf{u}_{\rho,s},$$

and

$$\mathbf{H}_i = \sum_{s=1}^{m} \exp(-\mathbf{a}'_i \mathbf{u}_{\rho,s}) \mathrm{tr}(\mathbf{C}^{-1} \mathbf{W}_{is}) \mathbf{u}_{\rho,s} \mathbf{u}'_{\rho,s}.$$

- In the common principal components model $[\mathbf{\Gamma A}_{is}\mathbf{\Gamma}']$, where $\mathbf{A}_{is}$ is the diagonal matrix with j^{th} diagonal entry $\exp(\mathbf{b}_{ij}'\mathbf{u}_{\rho,s})$, the objective function to minimize is

$$F = \sum_{i=1}^{K}\sum_{s=1}^{m} n_{is}\sum_{j=1}^{p}\mathbf{b}_{ij}'\mathbf{u}_{\rho,s} + \sum_{i=1}^{K}\sum_{s=1}^{m}\sum_{j=1}^{p}\exp(-\mathbf{b}_{ij}'\mathbf{u}_{\rho,s})c_{isj},$$

where c_{isj} is the j^{th} diagonal entry of the matrix $\mathbf{\Gamma}'\mathbf{W}_{is}\mathbf{\Gamma}$. Estimates of $\mathbf{b}_{ij}$ (for fixed $\mathbf{\Gamma}$) can be obtained using the Newton-Raphson procedure

$$\hat{\mathbf{b}_{ij}}^{\text{new}} = \hat{\mathbf{b}}_{ij}^{\text{old}} - \mathbf{H}_{ij}^{-1}\mathbf{g}_{ij},$$

where

$$\mathbf{g}_{ij} = \sum_{s=1}^{m}\left[n_{is} - \exp(-\mathbf{b}_{ij}'\mathbf{u}_{\rho,s})c_{isj}\right]\mathbf{u}_{\rho,s}$$

and

$$\mathbf{H}_{ij} = -\sum_{s=1}^{m}\exp(-\mathbf{b}_{ij}'\mathbf{u}_{\Sigma,s})c_{isj}\mathbf{u}_{\Sigma,s}\mathbf{u}_{\Sigma,s}'.$$

Latent Class Location Models

The two basic approaches to latent class location models shall be referred to as the two-step approach and the simultaneous approach. In the two-step approach, latent classes (the new locations) are assigned to observations following latent class analysis of the categorical data, and then locations models applied. The simultaneous approach is given by (3.23).

In the two-step latent class reduction approach, parameters in the multiple group latent class model (3.6) must be estimated so that response patterns (locations) can be assigned to latent classes (the "new" locations). Maximum likelihood estimates can be obtained using the EM algorithm.

The complete data log-likelihood is

$$L_c = \sum_{i=1}^{K}\sum_{h=1}^{n_i}\sum_{t=1}^{T} z_{ith}\{\log\eta_{it} + \sum_{j=1}^{q}[y_{jih}\log\pi_{jt} + (1-y_{jih})\log(1-\pi_{jt})]\}$$

where $z_{iht} = 1$ if $\mathbf{y}_{ih} \in C_t$. In the E-step, we compute $Q = \mathrm{E}_z^{\Psi}(L_c)$, where the expectation is taken with respect to the conditional distribution of the unobserved data $\{z_{ith}\}$ given the observed data and current parameter estimates $\mathbf{\Psi}$. Because L_c is linear in the unobserved data, the expectation is easily obtained by replacing each z_{iht} with $\hat{z}_{iht} = \tau_t(\mathbf{y}_{ih}; \hat{\mathbf{\Psi}})$, where

$$\tau_t(\mathbf{y}_{ih}; \mathbf{\Psi}) = \frac{\eta_{it} \prod_{j=1}^{q} \pi_{jt}^{y_{jih}} (1 - \pi_{jt})^{1-y_{jih}}}{\sum_{l=1}^{T} \eta_{il} \prod_{j=1}^{q} \pi_{jl}^{y_{jih}} (1 - \pi_{jl})^{1-y_{jih}}}$$

is the posterior probability that $\mathbf{y}_{ih}$ belongs to C_t.

In the M-step, Q is maximized subject to the constraints $\sum_{t=1}^{T} \eta_{it} = 1 \; \forall i$. Using the method of Lagrange multipliers, we maximize without constraint the expression

$$Q' = Q - \sum_{i=1}^{K} \gamma_i \left(\sum_{t=1}^{T} \eta_{it} - 1 \right)$$

where the γ_i are Lagrange multipliers. Q' is maximized by

$$\hat{\eta}_{it} = \frac{1}{n_i} \sum_{h=1}^{n_i} \hat{z}_{ith} \qquad (i = 1, \ldots, K; t = 1, \ldots, T)$$

and

$$\hat{\pi}_{jt} = \frac{\sum_{i=1}^{K} \sum_{h=1}^{n_i} \hat{z}_{ith} y_{jih}}{\sum_{i=1}^{K} \sum_{h=1}^{n_i} \hat{z}_{ith}} \qquad (j = 1, \ldots, q; t = 1, \ldots, T).$$

The EM algorithm alternately performs the E-step and M-step until parameter estimates have converged. The procedure requires starting values. Starting values can be obtained by randomly initializing posterior probabilities $\hat{z}_{ith}$ on $(0, 1)$ and then standardizing the uniform variates to satisfy $\sum_{t=1}^{T} \hat{z}_{ith} = 1 \; \forall i$. Because the algorithm may converge to local maxima, it should be rerun several times using different starting values to increase the chance that global maxima are obtained.

Parameters in the simultaneous model with augmented location covariate vectors also can be estimated using the EM algorithm. In this case the complete data log-likelihood is

$$L_c = \sum_{i=1}^{K} \sum_{s=1}^{m} \sum_{h=1}^{n_{is}} \sum_{t=1}^{T} z_{isht} \{ \log p_{ist} + \log h(\mathbf{x}_{ish}; \boldsymbol{\mu}_{ist}, \Sigma_{ist}) \}$$

where $z_{isht} = 1$ if $\mathbf{x}_{ish} \in C_t$.

As before, in the E-step we replace z_{isht} with $\hat{z}_{isht} = Pr^{\hat{\Psi}}(C_t|\mathbf{x}_{ish}, w_s)$. In the M-step, Q is maximized subject to the constraints $\sum_{s=1}^{m}\sum_{t=1}^{T} p_{ist} = 1 \ \forall i$.

The probability coefficients $\boldsymbol{\theta}_i (i = 1, \ldots, K)$ can be updated using Newton-Raphson methods. Let $\hat{\eta}_{ist} = \sum_{h=1}^{n_{is}} \hat{z}_{isht}$. Then

$$\hat{\boldsymbol{\theta}}_i^{\text{new}} = \hat{\boldsymbol{\theta}}_i^{\text{old}} - \mathbf{H}_i^{-1}\mathbf{g}_i$$

where

$$\mathbf{g}_i = \sum_{s=1}^{m}\sum_{t=1}^{T}[\hat{n}_{ts} - N\exp(\boldsymbol{\theta}_i'\mathbf{u}_{p,st})]\mathbf{u}_{p,st}$$

and

$$\mathbf{H}_i = -\sum_{s=1}^{m}\sum_{t=1}^{T} N\exp(\boldsymbol{\theta}_i'\mathbf{u}_{p,st})\mathbf{u}_{p,st}\mathbf{u}_{p,st}'.$$

The conditional mean coefficients $\mathbf{B}_i$ can be estimated as in the previous section, with slight modification. Let $\bar{\mathbf{x}}_{ist\cdot} = \sum_{h=1}^{n_{is}} \hat{z}_{isht}\mathbf{x}_{ish}$ and $\hat{n}_{ist} = \sum_{h=1}^{n_{is}} \hat{z}_{isht}$. Then

$$\text{vec } \hat{\mathbf{B}}_i = \left(\sum_{s=1}^{m}\sum_{t=1}^{T}\left(\hat{n}_{ist}\mathbf{u}_{\mu,st}\mathbf{u}_{\mu,st}' \otimes \boldsymbol{\Sigma}_{ist}^{-1}\right)\right)^{-1} \text{vec}\left[\sum_{s=1}^{m}\sum_{t=1}^{T}\boldsymbol{\Sigma}_{ist}^{-1}\hat{\mathbf{x}}_{ist\cdot}\mathbf{u}_{st}'\right]$$

The coefficients for the covariance matrix can be updated as in the previous section, with slight modification.

- In the proportional covariance model $[\rho_{it}\mathbf{C}]$, where $\log\rho_{it} = \mathbf{a}_i'\mathbf{u}_{\Sigma,t}$, the M-step objective function is

  $$F = p\sum_{i=1}^{K}\sum_{t=1}^{T}\hat{n}_{it}\log\rho_{it} + \sum_{i=1}^{K}\sum_{s=1}^{m}\sum_{t=1}^{T}\exp(-\mathbf{a}_i'\mathbf{u}_{\Sigma,st})\text{tr}(\mathbf{C}^{-1}\mathbf{W}_{it})$$

  where

  $$\mathbf{W}_{ist} = \sum_{s=1}^{m}\sum_{h=1}^{n_{is}}\hat{n}_{ist}(\mathbf{x}_{ish} - \boldsymbol{\mu}_{ist})(\mathbf{x}_{ish} - \boldsymbol{\mu}_{ist})'.$$

  The Newton-Raphson procedure is

  $$\hat{\mathbf{a}}_i^{\text{new}} = \hat{\mathbf{a}}_i^{\text{old}} - \mathbf{H}_i^{-1}\mathbf{g}_i$$

where

$$\mathbf{g}_i = \sum_{t=1}^{T}[p\hat{n}_{it} - \exp(-\mathbf{a}_i'\mathbf{u}_{\Sigma,t})\mathrm{tr}(\mathbf{C}^{-1}\mathbf{W}_{it})]\mathbf{u}_{\Sigma,st}$$

and

$$\mathbf{H}_i = \sum_{t=1}^{T}\exp(-\mathbf{a}_i'\mathbf{u}_{\Sigma,t})\mathrm{tr}(\mathbf{C}^{-1}\mathbf{W}_{it})]\mathbf{u}_{\Sigma,t}\mathbf{u}_{\Sigma,t}'$$

- In the common principal components model $[\mathbf{\Gamma R}_{it}\mathbf{\Gamma}']$, where $\mathbf{R}_{it}$ is the diagonal matrix with j^{th} diagonal entry $\exp(\mathbf{b}_{ij}'\mathbf{u}_{\Sigma,t})$, the M-step objective function is

$$F = \sum_{i=1}^{K}\sum_{t=1}^{T}\hat{n}_{it}\sum_{j=1}^{p}\mathbf{b}_{ij}'\mathbf{u}_{\Sigma,t} + \sum_{i=1}^{K}\sum_{t=1}^{T}\sum_{j=1}^{p}\exp(-\mathbf{b}_{ij}'\mathbf{u}_{\Sigma,t})c_{itj}$$

where c_{itj} is the j^{th} diagonal entry of the matrix $\mathbf{DW}_{it}\mathbf{D}'$. The Newton-Raphson procedure is

$$\hat{\mathbf{b}}_{ij}^{\mathrm{new}} = \hat{\mathbf{b}}_{ij}^{\mathrm{old}} - \mathbf{H}_{ij}^{-1}\mathbf{g}_{ij}$$

where

$$\mathbf{g}_{ij} = \sum_{t=1}^{T}[\hat{n}_{it} - \exp(-\mathbf{b}_{ij}'\mathbf{u}_{\Sigma,t})c_{itj}]\mathbf{u}_{\Sigma,t}$$

and

$$\mathbf{H}_{ij} = -\sum_{t=1}^{T}\exp(-\mathbf{b}_{ij}'\mathbf{u}_{\Sigma,t})c_{itj}\mathbf{u}_{\Sigma,t}\mathbf{u}_{\Sigma,t}'.$$

Examples

In this section we illustrate the potential value of some of the models. To do so, we compare the methods in three small simulation studies. Because of the large number of possible models, only a small subset of models were evaluated. Included are the full and reduced versions of $[\rho\mathbf{\Gamma\Lambda\Gamma}']$, $[\mathbf{\Sigma}_i]$, $[\rho_{is}\mathbf{\Gamma\Lambda\Gamma}']$, $[\rho_i\mathbf{\Gamma\Lambda}_i\mathbf{\Gamma}']$, and the simultaneous and two-step versions of the latent class location model.

Simulation 3

In this simulation, observations of five binary variables and two continuous variables were generated from each of two groups conforming to a latent class location model. We take the variables to be independent of group membership conditional on latent class membership; the latent classes, however, are distributed differently between the groups. We take $Pr(C_1|G_1) = .3$, $Pr(C_1|G_2) = .7$, $Pr(C_2|G_1) = .7$, and $Pr(C_2|G_2) = .3$. The response probabilities of the binary variables, conditional on latent class, are, in latent class 1, $\pi_{11} = .1,\pi_{21} = .2$, $\pi_{31} = .3$, $\pi_{41} = .4$, and $\pi_{51} = .6$. In latent class 2 the conditional response probabilities are $\pi_{12} = .9,\pi_{22} = .8$, $\pi_{32} = .7$, $\pi_{42} = .6$, and $\pi_{52} = .5$. The conditional means of the continuous variables are $(15, 20)$ in class 1 and $(20, 15)$ in class 2. The conditional covariances are

$$\mathbf{\Sigma}_1 = \begin{pmatrix} 4 & 2 \\ 2 & 4 \end{pmatrix}$$

and

$$\mathbf{\Sigma}_2 = \begin{pmatrix} 12 & 6 \\ 6 & 12 \end{pmatrix}$$

For each of 100 replications, n_i observations from group i were used to construct a classifier (n_i was varied from 50 to 1000). The classifiers were assessed by applying them on 100 independently generated observations (test set). Average error rates over the 100 replications are given in table 9. As should be expected, the latent class location model performed best, because the data were generated from that model. At $n_i = 50$, the LCLM had an average error rate of 29%. The 2-step LCLM had an average error rate of 32%. The next best method had an error rate of 36%. This example demonstrates that the latent class location model can be useful for certain data sets, especially if interpretation of latent classes is important. Similar results were obtained in simulations with better separated groups, as demonstrated

Model	$n_i = 50$	$n_i = 250$	$n_i = 1000$
full $[\rho\mathbf{\Gamma\Lambda\Gamma}']$	.42	.36	.32
reduced $[\rho\mathbf{\Gamma\Lambda\Gamma}']$	.36	.32	.32
full $[\mathbf{\Sigma}_i]$	.43	.36	.32
reduced $[\mathbf{\Sigma}_i]$	.36	.32	.32
full $[\rho_{is}\mathbf{\Gamma\Lambda\Gamma}']$	.42	.35	.31
reduced $[\rho_{is}\mathbf{\Gamma\Lambda\Gamma}']$	.36	.32	.31
simultaneous LCLM	.29	.29	.29
2-step LCLM	.32	.31	.29

Table 9: Average percent misclassifications for Simulation Experiment 3. Data were generated from the latent class location model (LCLM).

Model	$n_i = 50$	$n_i = 250$	$n_i = 1000$
full $[\rho\mathbf{\Gamma\Lambda\Gamma}']$	.23	.15	.13
reduced $[\rho\mathbf{\Gamma\Lambda\Gamma}']$	.14	.12	.12
full $[\mathbf{\Sigma}_i]$	.23	.15	.12
reduced $[\mathbf{\Sigma}_i]$	.15	.13	.12
full $[\rho_{is}\mathbf{\Gamma\Lambda\Gamma}']$	.23	.13	.12
reduced $[\rho_{is}\mathbf{\Gamma\Lambda\Gamma}']$	.15	.32	.31
simultaneous LCLM	.11	.11	.11
2-step LCLM	.13	.12	.12

Table 10: Average percent misclassifications for Simulation Experiment 4. Data were generated from the latent class location model (LCLM).

in Simulation 4.

Simulation 4

Conditions for Simulation 4 were identical to those for Simulation 3, except the conditional latent class distributions are given by $Pr(C_1|G_1) = .1$, $Pr(C_1|G_2) = .9$, $Pr(C_2|G_1) = .9$ and $Pr(C_2|G_2) = .1$, which leads to better separated groups. Average misclassification rates are given in Table 10. Again the latent class models perform best. There is not much difference between the different covariance models, probably because there were only two continuous variables in these examples. Using reduced models can improve classification performance for small datasets.

Simulation 5

In the next simulation, a different type of data structure was examined. One hundred observations of two binary and five continuous variables were generated from each of two groups, as follows. First, observations were generated from two multivariate normal populations with means $(0, 1, 4, 4, 4, 4, 4)$ and $(1, 0, 6, 6, 6, 6, 6)$, and covariance matrices

$$\boldsymbol{\Sigma}_1 = \begin{pmatrix} 2 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 2 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 3 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 4 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 5 \end{pmatrix},$$

and

$$\boldsymbol{\Sigma}_2 = \begin{pmatrix} 2 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 3 & 3 & 3 & 3 & 3 \\ 1 & 1 & 3 & 6 & 3 & 3 & 3 \\ 1 & 1 & 3 & 3 & 9 & 3 & 3 \\ 1 & 1 & 3 & 3 & 3 & 12 & 3 \\ 1 & 1 & 3 & 3 & 3 & 3 & 15 \end{pmatrix}.$$

As expected, the proportional covariance model (both full and reduced forms) performed best. Its average error rate over 100 replications was 19%. The average error rate for the location model was 28%. The quadratic location model had an error rate of 24%, and the CPC model had an average error rate of 23%. The full and reduced models had nearly identical performance, probably because there were only two binary variables.

Discussion

In previous research, the location model has been shown to be a powerful approach to classification with mixed-mode data. Sometimes, however, improvements can be made if the homogeneous covariance assumption is relaxed. This chapter described approaches to relaxing the homogeneous covariance assumption while estimating as few parameters as possible. In addition, latent class location models were considered as alternatives to the usual latent class formulation.

CHAPTER 4

Mixture Model Clustering of Correlated High-Dimensional Count Data

Routine collection of high-dimensional data has been facilitated by improvements in analytical instrumentation. For example, automated mass spectrometers now allow analytical chemists to rapidly collect mass spectra with hundreds of variables. Such instrumentation has spurred the development of methods for analyzing high-dimensional data.

Often, the analysis is concerned with partitioning the data into natural groupings. A parametric approach to this pattern recognition problem requires that distributional assumptions be made about observations within each group. Suppose that an observation $\mathbf{x}$ has arisen from exactly one of g distinct groups, denoted $G_1, \ldots, G_g$, where the density of an observation from G_i is $g_i(\mathbf{x}; \mathbf{\Psi}_i)$. The parameter vector $\mathbf{\Psi}_i$ is generally unknown. If nothing is known a priori about group structure, then inference about $\mathbf{\Psi}_i$ must be made indirectly by reference to the density of a randomly selected observation, which is given by the mixture model

$$f(\mathbf{x}) = \sum_{i=1}^{g} \eta_i g_i(\mathbf{x}; \mathbf{\Psi}_i),$$

where the η_i are mixing parameters which give the relative size of G_i $(0 < \eta_i < 1; \sum_{i=1}^{g} \eta_i = 1)$.

Let $\mathbf{x}_h$ denote the h^{th} observation. The posterior probability that $\mathbf{x}_h$ belongs to G_i is

$$\tau_i(\mathbf{x}_h; \mathbf{\Psi}) = \Pr(G_i | \mathbf{x}_h, \mathbf{\Psi}) = \frac{\eta_i g_i(\mathbf{x}_h; \mathbf{\Psi}_i)}{\sum_{l=1}^{g} \eta_l g_l(\mathbf{x}_h; \mathbf{\Psi}_l)},$$

where $\boldsymbol{\Psi} = (\eta_1, \ldots, \eta_g, \ldots, \boldsymbol{\Psi}_1, \ldots, \boldsymbol{\Psi}_g)$ contains all unknown parameters.

If misclasification costs are equal, the optimal classification rule assigns $\mathbf{x}_h$ to the group for which the posterior probability is greatest. That is, the classification rule is

$$\text{assign } \mathbf{x}_h \text{ to } G_i \text{ if } \max_{1 \leq l \leq g} \tau_l(\mathbf{x}_h; \boldsymbol{\Psi}) = \tau_i(\mathbf{x}_h; \boldsymbol{\Psi}). \tag{4.1}$$

In practice, the parameters η_i and $\boldsymbol{\Psi}_i$ $(i = 1, \ldots, g)$ can be estimated from the sample $\mathbf{x}_1, \ldots, \mathbf{x}_n$ which is to be clustered, and the estimates substituted in (4.1) for classification.

For continuous data, often it is reasonable to assume multivariate normal group conditional densities. When the data are high-dimensional, restrictions on the covariance matrices often are necessary to obtain efficient parameter estimates. A common approach is to take the variables to be independent within groups (i.e., restrict the covariance matrices to be diagonal).

In many applications the data are counts. If the counts are not too small, then it may be possible to transform the variables so that they are approximately normally distributed, and then use Gaussian mixture models or other continuous variable methods. For example, the Anscombe transform of a random variable X is given by

$$Y = t(X) = 2\sqrt{X + \frac{3}{8}}.$$

If $X \sim \text{Poisson}(\lambda)$, and λ is large, then Y is approximately normally distributed with variance 1 (Starck, Murtagh, and Bijaoui, 1998, p. 49). When λ is small the transformed variables are not approximately normal, and the group conditional densities should be based on count distributions (e.g., Poisson, negative binomial). Multivariate count distributions are complicated. Fortunately, in many applications in the physical sciences – for example, secondary ion mass spectrometry (SIMS) –theoretical and empirical evidence suggests that the variables (i.e., counts) are independent and

Poisson distributed under "ideal conditions." Thus, the independence model, also called the latent profile model in the latent variable models literature, is a good starting point for analyzing sparse count data.

In real data sets, however, the independence assumption often is not plausible, because data are collected subject to measurement error. For example, if the sensitivity of the count detector varies from run to run, then the multivariate counts will tend to vary together, depending on the sensitivity of the count detector. The sensitivity of the count detector can be thought of as an unobservable, or latent, variable, which induces a correlation between the variables. This variation in the data generation or data recording mechanism is known as *instrumental interference.*

The idea of a latent variable inducing correlations between observable variables is applicable to other fields as well. For example, in medical applications, a latent variable representing severity of illness may induce correlations in variables describing symptoms of patients from the same diagnostic class. In the absence of a gold standard for patient diagnosis, latent class models have been used to classify patients into diagnostic groups. Within a latent class (or diagnostic group), variables describing symptoms are assumed to be independent. But symptom variables for patients in the same diagnostic group often are correlated (thus violating the latent class model assumptions), and frequently this correlation can be "explained" by a unidimensional latent variable (which might be interpreted as "severity of illness"). To classify patients into one of two groups – diseased or not diseased – for some particular disease, Ubersax (1993, 1999) modified the two-class latent class model by introducing a continuous latent variable z, nominally interpreted as "severity of disease," so that the symptom indicators are assumed independent conditional on both the latent class and the continuous latent variable. The continuous latent variable is not distributed the same in both groups — the mean of the latent variable in the diseased group will be larger because patients in the diseased group are assumed to be more severely ill than those in the undiseased group. The modified latent class model, known as the located latent class model, is a special case of the underlying variable mixture model proposed by Everitt (Everitt, 1988; Everitt and Merette, 1990; Ubersax, 1999). Ubersax (1993) showed that patients can be more accurately classified by using the located latent class model than by using the conventional two-class latent class model.

In this chapter we take a related approach for clustering low count data. We modify the Poisson latent profile model to adjust for violations of the independence assumption by introducing a latent variable to "explain" the within-class correlations. We motivate the model with a SIMS application, which is described in the next section. For the models described in this chapter, unlike the located latent class models, it is assumed that the latent variable has the same distribution in all groups. Then the resulting latent variable mixture models are described, and estimation details given. The chapter concludes with simulated examples, and suggestions for future research.

An Example from Secondary Ion Mass Spectrometry

Secondary ion mass spectrometry (SIMS) has found widespread use in the physical and biological sciences. It is a particularly powerful technique for chemical analysis of surfaces and for identifying chemical constituents in unknown samples. Many recent advances in SIMS have been driven by demands of the semiconductor industry. In the 1996 Olympic Games, SIMS was used to detect anabolic steroids and other illegal substances in the urine of olympic athletes. SIMS is so widely used because it provides detailed molecular information about surfaces and unknown samples. In principle, it provides unambiguous identification of unknown chemical species. The goal is to provide a “fingerprint” profile of unknown species. For species identification, the profile might be matched against a library of “fingerprint” profiles for known species. Other techniques, by comparison, don’t provide as much detail at the molecular level, and usually don’t lead to unambiguous identification. Next we describe, briefly, the fundamentals of SIMS, giving just enough detail to explain how sparse count data might be obtained. Benninghoven (1994) and Benninghoven, Hagenhoff and Niehuis (1993) provide excellent reviews of SIMS capabilities and applications.

SIMS is a destructive technique, though the destruction is usually slight — typically only the top monolayer (i.e., the top layer of molecules) is destroyed. A sample is processed as follows. The sample is mounted under an ion gun, which bombards the sample with pulses of ions, called primary ions. This bombardment creates a “collision cascade,” in which the molecules in the sample are decomposed into neutral particles, positively charged ions, and negatively charged ions. The vast majority of the particles are neutral: typically only $10^{-6} - 10^{-2}$ of them are emitted as ions. The emitted ions are called secondary ions. The neutral particles sputter away, but the secondary ions are counted, using an ion detector designed to attract either the positively charged ions or the negatively charged ions. Further, the secondary ions are separated according to their atomic mass. The result of the analysis is a vector of secondary ion counts, one count for each atomic mass unit (or amu). The vector of counts — often graphically displayed as a spectrum of peaks — provides a “fingerprint” profile of the unknown sample. The dimension of the secondary ion count vector depends on the complexity of the sample. For example, for a sample of pure water (H_2O), positive ion counts will be observed only at 1 amu (H^+), 16 amu (O^+), 17 amu (OH^+) and 18 amu (H_2O^+). For more complex samples, a larger vector is required to capture the “fingerprint” of the sample.

In surface science applications, the analyst usually is interested only in the uppermost monolayer of a surface. This limits the number of surface molecules that can be analyzed if the surface area of the sample is small (for example, in SIMS imaging applications – where a separate spectrum is collected for each pixel – the surface area is frequently less than one square micron if high lateral resolution is required). The small number of molecules available for analysis, combined with the relative scarcity of secondary ions emitted per molecule, may lead to "fingerprint" profiles with very low counts (Schweiters, et. al., 1991). When the counts are small, it often is not possible to transform the variables to approximate normality (using, for example, the Anscombe transformation). Discrete distributions should be used instead.

Adriaens and Adams (1991) found that in repeat applications of SIMS to the same sample, ion counts are approximately Poisson distributed and independent. Thus, the Poisson latent profile model is a reasonable starting point for clustering a set of SIMS spectra. Let $\mathbf{x}_h = (x_{1h}, \ldots, x_{ph})'$ denote the collection of ion counts (i.e., the mass spectrum) for the h^{th} observation. In G_i, we take $x_{jh} \sim \text{Poisson}(\lambda_{ij})$. For g groups, the Poisson latent profile model is given by

$$f(\mathbf{x}_h) = \sum_{i=1}^{g} \eta_i \prod_{j=1}^{p} \frac{\exp(-\lambda_{ij})\lambda_{ij}^{x_{jh}}}{x_{jh}!}. \tag{4.2}$$

The Poisson parameters $\boldsymbol{\lambda}_i = (\lambda_{i1}, \ldots, \lambda_{ip})'$ represent the pure spectrum for the i^{th} chemical class. For a sample of n independent observations, $\mathbf{x}_1, \ldots, \mathbf{x}_n$, the unknown model parameters $\{\eta_i\}$ and $\{\lambda_{ij}\}$ can be estimated by the method of maximum likelihood using the EM algorithm (McLachlan and Krishnan, 1997; Willse and Tyler, 1998).

This model works well when there is no interference. In a SIMS imaging application, however, Willse and Tyler (1998) found that topographic differences over the area analyzed may induce a correlation among counts from the same pixel.

Typically, samples closer to the primary ion gun (i.e., samples at a "high topographic level") register larger counts than samples farther away from the primary beam gun (i.e., samples at a "low topographic level"), because the primary ions strike the closer samples with more velocity, emitting more secondary ions.

If topographic variation is not accounted for, chemical effects may be confounded with topographic effects, and the true chemical classes may be poorly separated. We can account for the topographic variation by introducing a latent variable z, which serves as a proxy for the unobserved "topographic level." We assume that the Poisson parameter λ_{ij} depends on the latent variable z through the log link (Moustaki and Knott, 1997)

$$\log \lambda_{ij} = \alpha_{ij} + \beta z. \tag{4.3}$$

Because the latent variable is measured on an arbitrary scale with arbitrary location, we take $z|G_i \sim \mathrm{N}(0,1)$. Initially, we assume that z has the same log-additive effect on all p variables – that is, that z is a random baseline. Thus, the coefficient on z does not depend on variable or group. Model (4.3) shall be referred to as the random baseline model. In the next section we will consider more general models, which allow the slope of the conditional response to vary between variables and/or groups.

Under model (4.3), we can remove, or annihilate, the effect of unobserved topography by standardizing the data prior to clustering. Let $M_h = \sum_{j=1}^{p} X_{jh}$ be the total count for observation h. Then, conditioning on M_h yields

$$\mathbf{X}_h | G_i, z, M_h = m_h \sim \text{Multinomial}(m_h, \boldsymbol{\pi}_i)$$

where $\boldsymbol{\pi}_i = (\pi_{i1}, \cdots, \pi_{ip})'$ is the vector of multinomial probabilities, which represent the standardized pure spectrum for the i^{th} chemical class, with

$$\pi_{ij} = \frac{\lambda_{ij}}{\sum_{l=1}^{p} \lambda_{il}} = \frac{\exp(\alpha_{ij} + \beta z)}{\sum_{l=1}^{p} \exp(\alpha_{il} + \beta z)} = \frac{\exp(\alpha_{ij})}{\sum_{l=1}^{p} \exp(\alpha_{il})}.$$

Note that the conditional distribution does not depend on z, so cluster analysis corrected for random baseline can be accomplished by fitting a mixture of multinomial distributions, which do not depend on the latent variable z. For example, if we take $\Pr(G_i|M_h = m_h) = \Pr(G_i) = \eta_i$, then the mixture model would be

$$f(\mathbf{x}_h|M_h = m_h) = \sum_{i=1}^{g} \eta_i \frac{m_h!}{x_{1h}! \ldots x_{ph}!} \prod_{j=1}^{p} \pi_{ij}^{x_{ih}}. \tag{4.4}$$

This approach was successfully applied by Willse and Tyler (1998) to a SIMS image with known chemical components (with modification for spatial correlation).

The multinomial approach successfully removes the topographic effect in the random baseline model (4.3), but information is lost in the approach. In many applications with no interference, different chemical components can be distinguished based solely on their total SIMS ion counts. In that case the assumption $Pr(G_i|M_h = m_h) = Pr(G_i)$ is unrealistic. Information about total counts is lost, and the components may be more difficult to distinguish. If z is distributed the same in all groups, can better group separation be attained by using model (4.3) directly ? This question is empirically investigated later in this chapter.

Sometimes the form of interference is more complicated than the random baseline model. Although there may be just one source of interference, different variables may respond differently to that interference. In the next section we extend model (4.3) to handle this more general case.

Mixtures of Poisson Latent Variable Models

Suppose that the correlations between the Poisson variables in G_i are induced by a single latent variable z, so that, conditional on z, the variables are independent. The latent variable z is taken to be standard normal in all groups. Following the development of generalized linear models (McCullagh and Nelder, 1989) and the general theory of latent variable models (Bartholomew, 1987), for the response on the j^{th} variable in the i^{th} group, the latent variable is related to the Poisson mean parameter λ_{ij} through the (canonical) log link:

$$\log \lambda_{ij}(z) = \alpha_{ij} + \beta_{ij} z. \tag{4.5}$$

Then, in G_i,

$$g_i(\mathbf{x}) = \int_{-\infty}^{\infty} \phi(z) \prod_{j=1}^{p} \frac{\exp(-\lambda_{ij}(z))\lambda_{ij}(z)^{x_j}}{x_j!} dz \tag{4.6}$$

where $\phi(z)$ is the standard normal density. The density in G_i is a latent variable model where the conditional responses of the manifest variables $\{x_j\}$ given the latent variable z conform to the Poisson distribution, with mean parameter λ_{ij} given by (4.5). Moustaki and Knott (1997) review a more general class of latent variable models for conditional responses conforming to the exponential family. They call their class of models generalized latent trait models. These models are generalizations of factor models for continuous observed data and of latent trait models for categorical observed data.

Note that z is measured with arbitrary direction: in the SIMS example z could increase with increasing topographic level or with decreasing topographic level – the substantive conclusion would be the same. If we replace z with $-z$, then $\log \lambda_{ij}(z) = \alpha_{ij} - \beta_{ij} z$. Thus, changing the sign of all the β_{ij} yields a substantively equivalent model.

The overall density for a randomly selected observation with unknown group membership is

$$f(\mathbf{x}) = \sum_{i=1}^{g} \eta_i \int_{-\infty}^{\infty} \phi(z) \prod_{j=1}^{p} \frac{\exp(-\lambda_{ij}(z))\lambda_{ij}(z)^{x_j}}{x_j!} dz, \tag{4.7}$$

which is a mixture of Poisson latent variable models. Finite mixtures of factor analysis models have been developed for continuous (MVN) observable data (Yung, 1997).

The located latent class model (Ubersax, 1993) is a special type of mixture model for binary and ordinal observed variables. Finite mixtures of latent trait models have not previously been developed.

Other models can be constructed by holding the slope parameter β_{ij} invariant across variables and/or groups. We will consider the following models.

- $[\alpha_{ij}+\beta_{ij}z]$ This is the unrestricted latent trait model, where the slope parameter is free to vary across variables and groups.
- $[\alpha_{ij} + \beta_j z]$ In this model the slope parameter of the conditional response is invariant to group.
- $[\alpha_{ij} + \beta_i z]$ In this model the slope parameter of the conditional response is the same for all variables within a group, but the parameter varies between groups. This is a type of random baseline model, where all variables within a group are affected equally by the interference.
- $[\alpha_{ij} + \beta z]$ Random baseline model. The conditional responses on all variables in all groups are affected equally by the interference.
- $[\alpha_{ij}]$ Latent profile model. The variables are independent within groups.

It is useful to recognize the hierarchical relationships between these classes of mixture models. When models are nested for a fixed number of groups g, they can be compared using likelihood ratio tests. Table 11 gives the degrees of freedom in the likelihood ratio tests (where applicable) for comparison of these five models. If $\hat{L}_1$ and $\hat{L}_2$ are maximized log-likelihoods for models M_1 and M_2, with $M_1 \subset M_2$, then the likelihood-ratio test statistic, given by

$$T = -2(\hat{L}_1 - \hat{L}_2)$$

is asymptotically distributed as $\chi^2(\mathrm{df})$ under the null hypothesis that M_1 is true.

Model	$[\alpha_{ij}]$	$[\alpha_{ij}+\beta z]$	$[\alpha_{ij}+\beta_i z]$	$[\alpha_{ij}+\beta_j z]$	$[\alpha_{ij}+\beta_{ij} z]$
$[\alpha_{ij}]$	–	1	g	p	pg
$[\alpha_{ij}+\beta z]$		–	$g-1$	$p-1$	$pg-1$
$[\alpha_{ij}+\beta_i z]$			–	NA	$p(g-1)$
$[\alpha_{ij}+\beta_j z]$				–	$g(p-1)$
$[\alpha_{ij}+\beta_{ij} z]$					–

Table 11: Degrees of freedom in likelihood ratio test comparisons of some nested models, with p response variables and g groups.

Because we take $z|G_i \sim \mathrm{N}(0,1)\ \forall i$, the slope parameter in the random baseline model can be interpreted as a variance component parameter. That is,

$$\beta z|G_i \sim \mathrm{N}(0,\beta^2),$$

so the model $[\alpha_{ij}+\beta z]$ could be written as $[\alpha_{ij}+z]$, where $z|G_i \sim \mathrm{N}(0,\beta^2)$. The model $[\alpha_{ij}+\beta_i z]$ has a similar random effects interpretation. The latent profile model $[\alpha_{ij}]$ also has a random effects interpretation, where the variance of the latent variable is taken to be 0.

In the previous section it was shown that in the random baseline model $[\alpha_{ij}+\beta z]$ the effect of the latent variable z can be annihilated by conditioning on sums of the count vectors. This type of annihilation also is possible for the model $[\alpha_{ij}+\beta_i z]$, but not for the models $[\alpha_{ij}+\beta_j z]$ and $[\alpha_{ij}+\beta_{ij} z]$.

Moments

The first and second moments of the observed variables can be obtained as follows. Within group G_i,

$$\begin{aligned} E(X_j|G_i) &= E[E(X_j|G_i,z)] \\ &= E[\lambda_{ij}(z)|G_i] \\ &= E[\exp(\alpha_{ij}+\beta_{ij}z)|G_i] \\ &= \exp(\alpha_{ij}+\beta_{ij}^2/2), \end{aligned}$$

$$\begin{aligned}
E(X_j^2|G_i) &= E[E(X_j^2|G_i, z)] \\
&= E[\lambda_{ij}(z) + \lambda_{ij}(z)^2|G_i] \\
&= E[\exp(\alpha_{ij} + \beta_{ij}z) + \exp(2\alpha_{ij} + 2\beta_{ij}z)|G_i]E(X_j|G_i) \\
&= \exp(\alpha_{ij} + \beta_{ij}^2/2) + \exp(2\alpha_{ij} + 2\beta_{ij}^2) \\
&= E(X_j|G_i) + E(X_j|G_i)^2 \exp(\beta_{ij}^2), \\
E(X_jX_k|G_i) &= E[E(X_jX_k|G_i, z)] \\
&= E[\lambda_{ij}(z)\lambda_{ik}(z)|G_i] \\
&= E[\exp(\alpha_{ij} + \beta_{ij}z)\exp(\alpha_{ik} + \beta_{ik}z)|G_i] \\
&= \exp(\alpha_{ij} + \alpha_{ik} + (\beta_{ij} + \beta_{ik})^2/2) \\
&= E(X_j|G_i)E(X_k|G_i)\exp(\beta_{ij}\beta_{ik}).
\end{aligned}$$

Then

$$\begin{aligned}
Var(X_j|G_i) &= E(X_j^2|G_i) - E(X_j|G_i)^2 \\
&= \exp(\alpha_{ij} + \beta_{ij}^2/2) + \exp(2\alpha_{ij} + \beta_{ij}^2)[\exp(\beta_{ij}^2) - 1] \\
&= E(X_j|G_i) + E(X_j|G_i)^2[\exp(\beta_{ij}^2) - 1]
\end{aligned}$$

and

$$\begin{aligned}
Cov(X_j, X_k|G_i) &= E(X_jX_k|G_i) - E(X_j|G_i)E(X_k|G_i) \\
&= E(X_j|G_i)E(X_k|G_i)[\exp(\beta_{ij}\beta_{ik}) - 1].
\end{aligned}$$

The Poisson latent variable model is a type of overdispersed Poisson model. The dependencies between the variables are controlled by the β_{ij}'s. The variables X_j and X_k are independent in G_i if and only if $\beta_{ij} = 0$ or $\beta_{ik} = 0$. Note that X_j and X_k will be negatively correlated in G_i if β_{ij} and β_{ik} have different signs. In the random baseline models, β_{ij} and β_{ik} are forced to have the same sign by the restriction $\beta_{ij} = \beta_{ik} = \beta_i$ (or β). In the unrestricted model $[\alpha_{ij} + \beta_{ij}z]$ we might wish to force the β_{ij}'s to have the same sign without forcing them to be equal. As previously discussed, changing the signs of all the β_{ij}'s yields a substantively equivalent model, so without loss of generality we consider how to force the β_{ij}'s to all be positive. A straightforward approach is to reparameterize β_{ij} as $\beta_{ij} = \exp(a_{ij})$, thus forcing it to be positive. This model will not be pursued here, but a method for estimating the parameters might be derived by paralleling the approach to estimating $[\alpha_{ij} + \beta_{ij} z]$.

The unconditional moments are

$$\begin{aligned}
E(X_j) &= \sum_{i=1}^{g} \eta_i E(X_j|G_i) \\
&= \sum_{i=1}^{g} \eta_i \exp(\alpha_{ij} + \beta_{ij}^2/2), \\
Var(X_j) &= \sum_{i=1}^{g} \eta_i [Var(X_j|G_i) + (E(X_j|G_i) - E(X_j))^2] \\
Cov(X_j, X_k) &= \sum_{i=1}^{g} \eta_i [Cov(X_j, X_k|G_i) + (E(X_j|G_i) - E(X_j))(E(X_k|G_i) - E(X_k))].
\end{aligned}$$

Alternative Parameterization of Random Baseline Models

If the random baseline model $[\alpha_{ij} + \beta z]$ is reparameterized by taking $\alpha_{ij}^* = \exp(\alpha_{ij})$ and $z^* = \exp(\beta z)$, then the Poisson parameter is a linear function of the transformed latent variable z^*: $\lambda_{ij}(z^*) = \alpha_{ij}^* z^*$, where $z^*|G_i \sim \text{lognormal}(0, \beta^2)$. A similar interpretation can be given to the model $[\alpha_{ij} + \beta_i z]$, with $z^* = \exp(\beta_i z)$, but not to the models $[\alpha_{ij} + \beta_j z]$ and $[\alpha_{ij} + \beta z]$, because in these models the distribution of z^* depends on the variable indicator.

This parameterization doesn't necessarily simplify the random baseline models, because the integration over the latent variable in (4.6) is still analytically intractable. But, because the choice of latent distribution is largely arbitrary, it motivates the search for a positive valued latent distribution which simplifies the integration.

If we choose a gamma prior distribution for the latent variable instead of a lognormal distribution, then the integral over the latent variable can be analytically evaluated. Suppose $\lambda_{ij}(z) = \alpha_{ij}z$, where, within G_i, z has a gamma distribution with mean 1 and variance $1/\theta$:

$$h(z|G_i) = \frac{\theta^\theta}{\Gamma_\theta} z^{\theta-1}\exp(-\theta z), \qquad z > 0.$$

Then the density of the observed variables in G_i is

$$\begin{aligned} g_i(\mathbf{x}) &= \int_0^\infty h(z|G_i)\prod_{j=1}^p g_i(x_j|z)dz \\ &= \int_0^\infty \frac{\theta^\theta}{\Gamma_\theta} z^{\theta-1}\exp(-\theta z)\prod_{j=1}^p \frac{(\alpha_{ij}z)^{x_j}\exp(-\alpha_{ij}z)}{x_j!}dz \\ &= \frac{\theta^\theta}{\Gamma_\theta}\prod_{j=1}^p \frac{\alpha_{ij}^{x_j}}{x_j!}\int_0^\infty z^{\theta+m-1}\exp[-(\theta+\sum_j \alpha_{i\cdot})z]dz \\ &= \frac{\theta^\theta}{\Gamma_\theta}\frac{\Gamma_{\theta+m}}{(\theta+\alpha_{i\cdot})^{\theta+m}}\prod_{j=1}^p \frac{\alpha_{ij}^{x_j}}{x_j!}, \end{aligned} \tag{4.8}$$

where $m = \sum_{j=1}^p x_j$ and $\alpha_{i\cdot} = \sum_{j=1}^p \alpha_{ij}$. The integration in the last step can be performed by recognizing that the integrand is the kernel of a gamma distribution. The density (4.8) is a *negative multinomial* distribution, a multivariate generalization of the negative binomial distribution, with moments given by

$$E(X_j|G_i) = E[E(X_j|G_i,z)] = E(\alpha_{ij}z|G_i) = \alpha_{ij},$$

$$\begin{aligned} Var(X_j|G_i) &= E[Var(X_j|G_i,z)] + Var[E(X_j|G_i,z)] \\ &= E(\alpha_{ij}z|G_i) + Var(\alpha_{ij}z|G_i) \\ &= \alpha_{ij} + \alpha_{ij}^2\frac{1}{\theta}, \end{aligned}$$

$$E(X_jX_k|G_i) = E[E(X_jX_k|G_i,z)] = E(\alpha_{ij}\alpha_{ik}z^2|G_i) = \alpha_{ij}\alpha_{ik}\left(\frac{1}{\theta}+1\right),$$

and

$$Cov(X_j,X_k|G_i) = E(X_jX_k|G_i) - E(X_j|G_i)E(X_k|G_i)$$

$$= \alpha_{ij}\alpha_{ik}\left(\frac{1}{\theta}+1\right) - \alpha_{ij}\alpha_{ik}$$
$$= \alpha_{ij}\alpha_{ik}\frac{1}{\theta}.$$

The parameter $1/\theta$ plays the same role as $\exp(\beta^2)-1$ in the model $[\alpha_{ij}+\beta z]$. Mixtures of negative multinomial distributions will not be considered in this paper, though they might be considered as an alternative to the random baseline models $[\alpha_{ij}+\beta z]$ and $[\alpha_{ij}+\beta_j z]$. We expect the two approaches to yield similar results.

Comparison with Normal Models

It is useful to review normal theory methods that have been developed for clustering data which have been distorted by interference. In fact, if the counts are large, the data (perhaps following transformation) might be adequately modeled with normal models.

Suppose that the observed variables $X_1, \ldots, X_p$ are continuous, and normally distributed within a group. Then, in G_i, the one-dimensional (i.e., one latent variable) random baseline model can be written as

$$\mathbf{x} = \boldsymbol{\alpha}_i + \mathbf{1}_p\beta z + \mathbf{e}$$

$$z|G_i \sim N(0,1)$$

$$\mathbf{e}|G_i \sim N(\mathbf{0}, \boldsymbol{\Psi}_i), \qquad \boldsymbol{\Psi}_i \text{ diagonal}, \qquad \mathrm{Cov}(z, \mathbf{e}|G_i) = \mathbf{0}.$$

It follows that

$$\mathbf{X}|G_i \sim N(\boldsymbol{\alpha}_i, \beta^2\mathbf{J}_p + \boldsymbol{\Psi}_i),$$

where $\mathbf{J}_p = \mathbf{1}_p\mathbf{1}_p'$. In this model, a common approach to annihilating the random baseline effect is to center each observation, that is, to replace each observed $\mathbf{x}_h$ with $\mathbf{x}_h - \mathbf{1}_p\bar{x}_h$, where $\bar{x}_h = \frac{1}{p}\sum_{j=1}^{p} x_{jh}$ (Ge and Simpson, 1998). Centering is analogous to conditioning on total counts in the Poisson model. The mean-centered variables are normally distributed but not independent, so a finite mixture of independence models may not be adequate for clustering.

An alternative approach is to fit a finite mixture of factor models conforming to the (assumed) interference pattern. This approach allows for a more general class of models than the mean-centering approach, which is somewhat successful in annihilating the effect of the latent variable only when the slope parameter β is the same for all p variables. The factor analysis mixture model allows for a variety of assumptions about the slope parameter, and is easily generalized to multiple latent variables. This approach has many similarities with the Poisson latent variable mixture model. The general factor analysis mixture model is given by

$$f(\mathbf{x}) = \sum_{i=1}^{g} \eta_i g_i(\mathbf{x}; \boldsymbol{\Psi}_i) \tag{4.9}$$

where, within G_i,

$$\mathbf{x} = \boldsymbol{\alpha}_i + \mathbf{B}_i\mathbf{z} + \mathbf{e},$$

$$\begin{array}{rcll}
\mathbf{x} & \text{is} & \text{a } p \times 1 \text{ vector of observed variables} & \\
\mathbf{z} & \text{is} & \text{a } r \times 1 \text{ vector of latent variables} & \\
\boldsymbol{\alpha}_i & \text{is} & \text{a } p \times 1 \text{ vector of unknown intercept parameters} & \\
\mathbf{B}_i & \text{is} & \text{a } p \times r \text{ matrix of unknown factor loadings} & \\
\mathbf{z}|G_i & \sim & \mathrm{N}(\boldsymbol{\nu}_i, \boldsymbol{\Phi}_i) & \\
\mathbf{e}|G_i & \sim & \mathrm{N}(\mathbf{0}, \boldsymbol{\Psi}_i), & \boldsymbol{\Psi}_i \text{ diagonal} \\
\mathrm{Cov}(\mathbf{z}, \mathbf{e}|G_i) & = & \mathbf{0}. &
\end{array}$$

The within-group density of the observed variables is

$$\mathbf{X}|G_i \sim \mathrm{N}(\boldsymbol{\alpha}_i + \mathbf{B}_i\boldsymbol{\nu}_i, \mathbf{B}_i\boldsymbol{\Phi}_i\mathbf{B}_i' + \boldsymbol{\Psi}_i).$$

The mixture model (4.9) is not identifiable without imposing additional restrictions. Yung (1997) considered three types of restrictions which lead to three classes of identifiable models.

1. In the first class of models, the "factor-to-variable transformation mechanism," defined by $\boldsymbol{\alpha}_i$ and $\mathbf{B}_i$, is taken to be the same for all groups. All group differences are assumed to be generated by differences in the latent distribution. These assumptions are commonly made in regular multiple-group factor analysis. In the SIMS application it is reasonable to assume that the latent variable is distributed the same for all groups, so this model is not pursued in this paper.

2. In the second class of models, the latent distributions and the factor loadings $\mathbf{B}_i$ are taken to be the same for all groups. Only $\boldsymbol{\alpha}_i$, and possibly $\boldsymbol{\Psi}_i$, vary between groups. The Poisson analog of this model is the model $[\alpha_{ij} + \beta_j z]$. If, in the one latent dimension case we make the restriction $\mathbf{B} = \mathbf{1}_p\beta$, we obtain a model similar to $[\alpha_{ij} + \beta z]$. Annihilation by mean-centering (in the normal model) or conditioning on total (in the Poisson model) is possible for $[\alpha_{ij} + \beta z]$ but not for $[\alpha_{ij} + \beta_j z]$.

3. In the third class of models, the latent distribution is the same in all groups, but the factor-to-variable transformation mechanism is unrestricted. The groups are characterized by their factor-to-variable transformation mechanisms. The Poisson analog is $[\alpha_{ij} + \beta_{ij} z]$.

Yung described an EM algorithm for estimating the parameters in these three classes of models. He did not consider the random baseline restrictions ($[\alpha_{ij} + \beta_i z]$ and $[\alpha_{ij} + \beta z]$).

Estimation

In this section, algorithms are derived for maximum likelihood estimation of parameters for the Poisson latent profile model, the multinomial mixture model, and the Poisson latent variable mixture model.

Poisson Model $[\alpha_{ij}] = [\lambda_{ij}]$

Maximum likelihood estimates of the mixture model parameters can be obtained by treating the unobserved group labels as missing data and applying the EM algorithm (McLachlan and Krishnan, 1997). Let $\mathbf{y}_h = (y_{1h}, \ldots, y_{gh})'$ be the g-dimensional group indicator vector for the h^{th} observation, so that $y_{ih} = 1$ if the h^{th} individual belongs to G_i. The vector $\mathbf{y}_h$ is not observed.

The complete-data log-likelihood for the sample $\mathbf{x}_1, \ldots, \mathbf{x}_n$ from the Poisson independence model (4.2) is

$$L_c = \sum_{h=1}^{n} \sum_{i=1}^{g} y_{ih} \left\{ \log \eta_i + \sum_{j=1}^{p} [\lambda_{ij} + x_{ih} \log \lambda_{ij} - \log x_{jh}!] \right\}.$$

The EM algorithm is a two-step procedure. In the E-step (Expectation step) we compute $Q = E_z^{\Psi}(L_c)$, where the expectation is taken with respect to the conditional distribution of the unobserved data $\mathbf{y}_h$ $(h = 1, \cdots, n)$ given the observed data and current parameter estimate $\hat{\boldsymbol{\Psi}}$. Because L_c is linear in the unobserved data, the expectation is easily obtained by replacing each y_{ih} with $\hat{h}_{i|h} = \tau_i(\mathbf{x}_h; \hat{\boldsymbol{\Psi}})$, where

$$\tau_i(\mathbf{x}_h; \boldsymbol{\Psi}) = \frac{\eta_i \prod_{j=1}^{p} \exp(-\lambda_{ij}) \lambda_{ij}^{x_{jh}}}{\sum_{l=1}^{g} \eta_l \prod_{j=1}^{p} \exp(-\lambda_{lj}) \lambda_{lj}^{x_{jh}}} \tag{4.10}$$

is the posterior probability that individual h belongs to G_i.

In the M-step (Maximization step), Q is maximized subject to the constraint $\sum_{i=1}^{g} \eta_i = 1$. The result is

$$\hat{\eta}_i = \frac{1}{n} \sum_{h=1}^{n} \hat{h}_{i|h} \quad \text{and} \quad \hat{\lambda}_{ij} = \frac{1}{n\hat{\eta}_i} \sum_{h=1}^{n} \hat{h}_{i|h} x_{ih}. \tag{4.11}$$

The EM algorithm alternately updates (4.10) and (4.11). The procedure requires starting values for the iterations. Starting values can be obtained by randomly initializing posterior probabilities $\tau_i(\mathbf{x}_h; \Psi)$ uniformly on (0,1), and then standardizing the uniform variates to satisfy $\sum_{i=1}^{g} \tau_i(\mathbf{x}_h; \Psi) = 1$ for all h. It is well known that log-likelihood surfaces for mixture models are often flat with many local maxima, so the EM algorithm should be applied several times with different starting parameter values to increase the chance of obtaining global maxima. Classifications are made on the basis of $\tau_i(\mathbf{x}_h; \hat{\Psi})$.

Multinomial Model

In the multinomial mixture model (4.4), the complete-data log-likelihood for the sample $\mathbf{x}_1, \ldots, \mathbf{x}_n$ is

$$L_c = \sum_{h=1}^{n} \sum_{i=1}^{g} y_{ih} \left\{ \log \eta_i + \sum_{j=1}^{p} x_{ih} \log \pi_{ij} + \log \left(\frac{m_h!}{x_{1h}! \ldots x_{ph}!} \right) \right\}.$$

The E-step is accomplished by replacing each y_{ih} with $\hat{h}_{i|h} = \tau_i(\mathbf{x}_h; \hat{\Psi})$, where

$$\tau_i(\mathbf{x}_h; \Psi) = \frac{\eta_i \prod_{j=1}^{p} \pi_{ij}^{x_{jh}}}{\sum_{l=1}^{g} \eta_l \prod_{j=1}^{p} \pi_{lj}^{x_{ih}}}. \tag{4.12}$$

In the M-step, we obtain

$$\hat{\eta}_i = \frac{1}{n} \sum_{h=1}^{n} \hat{h}_{i|h} \quad \text{and} \quad \hat{\pi}_{ij} = \frac{\sum_{h=1}^{n} \hat{h}_{i|h} x_{jh}}{\sum_{h=1}^{n} \hat{h}_{i|h} m_h}.$$

Unrestricted Latent Variable Model $[\alpha_{ij} + \beta_{ij} z]$

Estimation of the model $[\alpha_{ij} + \beta_{ij} z]$, described by (4.5) and (4.6), is problematic because integration over the latent variable z is required. Using Gauss-Hermite quadrature, we approximate the (standard normal) distribution of the latent variable by a finite discrete distribution

$$\begin{pmatrix} z_1, & \dots, z_K \\ w_1, & \dots, w_K \end{pmatrix}$$

where z_k is the k^{th} mass point and w_k is the corresponding probability. A particularly elegant description of Gauss-Hermite quadrature applied to latent variable models can be found in Sammel et. al. (1997). Using Gauss-Hermite quadrature, the mixture density (4.7) is approximated by the density

$$f(\mathbf{x}) = \sum_{i=1}^{g} \sum_{k=1}^{K} \eta_i w_k g(\mathbf{x}_h | G_i, z_k, \mathbf{\Psi}), \tag{4.13}$$

where

$$g(\mathbf{x}_h | G_i, z_k, \mathbf{\Psi}) = \prod_{j=1}^{p} \frac{\exp(-\lambda_{ij}(z_k))\lambda_{ij}(z_k)^{x_j}}{x_j!}. \tag{4.14}$$

Model (4.13) is a finite mixture model with gK components and mixing proportions $\eta_i w_k$.

Let $\mathbf{v}_h = (v_{1h}, \dots, v_{kh})'$ be a latent variable indicator vector for the h^{th} observation, so that $v_{kh} = 1$ if $z = z_k$ (i.e., if subject h belongs to the k^{th} latent level). Similarly, let $\mathbf{y}_h = (y_{1h}, \dots, y_{gh})'$ be the group indicator vector, so that $y_{ih} = 1$ if subject h is from G_i. The unknown parameters in (4.13) can be estimated by treating the $\mathbf{v}_h$ and $\mathbf{y}_h$ as missing data and using the EM algorithm.

The contribution of the h^{th} observation to the complete data log likelihood is

$$\log f(\mathbf{x}_h, \mathbf{v}_h, \mathbf{y}_h) = \log f(\mathbf{y}_h) + \log f(\mathbf{x}_h, \mathbf{v}_h | \mathbf{y}_h)$$

$$= \sum_{i=1}^{g} y_{ih} \log \eta_i + \sum_{i=1}^{g} \sum_{k=1}^{K} \sum_{j=1}^{p} v_{kh} y_{ih} \left[x_{jh}(\alpha_{ij} + \beta_{ij} z_k) - \exp(\alpha_{ij} + \beta_{ij} z_k) \right].$$

The complete data log-likelihood is

$$L_c = \sum_{h=1}^{n} \sum_{i=1}^{g} y_{ih} \log \eta_i + \sum_{h=1}^{n} \sum_{i=1}^{g} \sum_{k=1}^{K} \sum_{j=1}^{p} v_{kh} y_{ih} \left[x_{jh}(\alpha_{ij} + \beta_{ij} z_k) - \exp(\alpha_{ij} + \beta_{ij} z_k) \right]. \tag{4.15}$$

In the E-step, we compute $Q = E^{\Psi}_{v,y}(L_c)$, where the expectation is taken with respect to the distribution of the unobserved data $\{\mathbf{v}_h, \mathbf{y}_h\}$ conditional on the observed data and current parameter estimates $\hat{\boldsymbol{\Psi}}$. In the second term in (4.15), we replace $v_{kh}y_{ih}$ with

$$\hat{h}_{ki|h} = \Pr(z_k, G_i|\mathbf{x}_h) = \frac{\hat{\eta}_i w_k g(\mathbf{x}_h|G_i, z_k, \hat{\boldsymbol{\Psi}})}{\sum_{i=1}^{g}\sum_{k=1}^{K} \hat{\eta}_i w_k g(\mathbf{x}_h|G_i, z_k, \hat{\boldsymbol{\Psi}})}, \tag{4.16}$$

where $g(\mathbf{x}_h|G_i, z_k, \boldsymbol{\Psi})$ is defined in (4.14). In the first term in (4.15) we replace y_{ih} with $\hat{h}_{i|h} = \Pr(G_i|\mathbf{x}_h) = \tau_i(\mathbf{x}_h; \hat{\boldsymbol{\Psi}}) = \sum_{k=1}^{K} h_{ki|h}$. The result is

$$Q = \sum_{h=1}^{n}\sum_{i=1}^{g} \hat{h}_{i|h} \log \eta_i + \sum_{h=1}^{n}\sum_{i=1}^{g}\sum_{k=1}^{K}\sum_{j=1}^{p} \hat{h}_{ki|h}[x_{jh}(\alpha_{ij} + \beta_{ij}z_k) - \exp(\alpha_{ij} + \beta_{ij}z_k)].$$

Let

$$\hat{N}_i = \sum_{h=1}^{n} \hat{h}_{i|h}, \qquad \hat{N}_{ki} = \sum_{h=1}^{n} \hat{h}_{ki|h}, \qquad \bar{x}_{kij} = \sum_{h=1}^{n} x_{jh}\hat{h}_{ki|h}.$$

$\hat{N}_i$ is the expected number of individuals in G_i, $\hat{N}_{ki}$ is the expected number of individuals at latent level z_k of G_i, and $\bar{x}_{kij}$ is average response of the j^{th} variable at latent level z_k of group G_i. Then

$$Q = \sum_{i=1}^{g} \hat{N}_i \log \eta_i + \sum_{i=1}^{g}\sum_{k=1}^{K}\sum_{j=1}^{p}[\bar{x}_{kij}(\alpha_{ij} + \beta_{ij}z_k) - \hat{N}_{ki}\exp(\alpha_{ij} + \beta_{ij}z_k)].$$

In the M-step, Q is maximized subject to the constraint $\sum_{i=1}^{g} \eta_i = 1$. This yields updated mixing parameter estimates

$$\hat{\eta}_i = \frac{\hat{N}_i}{n}.$$

The parameters α_{ij} and β_{ij} can be updated independently of all other parameters using the following Newton-Raphson procedure. Define $\boldsymbol{\theta}_{ij} = (\alpha_{ij}, \beta_{ij})'$, and $\mathbf{u}_k = (1, z_k)'$. Then, for $i = 1, \ldots, g$ and $j = 1, \ldots, p$,

$$\hat{\boldsymbol{\theta}}_{ij}^{\text{new}} = \hat{\boldsymbol{\theta}}_{ij}^{\text{old}} - \mathbf{H}_{ij}^{-1}\mathbf{g}_{ij}$$

where

$$\mathbf{g}_{ij} = \frac{\partial Q}{\partial \hat{\boldsymbol{\theta}}_{ij}} = \sum_{k=1}^{K}[\bar{x}_{kij} - \hat{N}_{ki}\exp(\hat{\boldsymbol{\theta}}_{ij}'\mathbf{u}_k)]\mathbf{u}_k$$

and

$$\mathbf{H}_{ij} = \frac{\partial^2 Q}{\partial \hat{\boldsymbol{\theta}}_{ij}\partial \hat{\boldsymbol{\theta}}_{ij}'} = -\sum_{k=1}^{K}\hat{N}_{ki}\exp(\hat{\boldsymbol{\theta}}_{ij}'\mathbf{u}_k)\mathbf{u}_k\mathbf{u}_k'.$$

Random Baseline Model $[\alpha_{ij} + \beta z]$

In the unrestricted model $[\alpha_{ij} + \beta_{ij} z]$, the parameters α_{ij} and β_{ij} can be estimated within the M-step in g separate two-dimensional Newton-Raphson optimizations. The EM algorithm is thus computationally feasible for high-dimensional data. In the models with restrictions on β_{ij}, the E-step is performed just like in the unrestricted model, but in the M-step the Newton-Raphson optimizations are a little more complicated because the Hessian matrix $\mathbf{H}$ is no longer rank two. The algorithms are still computationally feasible, however, because $\mathbf{H}$ will be shown to be "mostly diagonal."

In the random baseline model $[\alpha_{ij} + \beta z]$, the $gp+1$ parameter vector $\boldsymbol{\theta}$ defined by

$$\boldsymbol{\theta} = (\beta, \alpha_{11}, \ldots, \alpha_{1p}, \ldots, \alpha_{g1}, \ldots, \alpha_{gp})'$$

is updated in the Newton-Raphson steps

$$\hat{\boldsymbol{\theta}}^{\text{new}} = \hat{\boldsymbol{\theta}}^{\text{old}} - \mathbf{H}^{-1}\mathbf{g},$$

where

$$\mathbf{g} = \frac{\partial Q}{\partial \hat{\boldsymbol{\theta}}} \text{ and } \mathbf{H} = \frac{\partial^2 Q}{\partial \hat{\boldsymbol{\theta}}\partial \hat{\boldsymbol{\theta}}'}.$$

The gradient vector $\mathbf{g}$ contains the derivatives

$$\frac{\partial Q}{\partial \alpha_{ij}} = \sum_{k=1}^{K}[\bar{x}_{kij} - \hat{N}_{ki}\exp(\alpha_{ij} + \beta z_k)] \qquad (i = 1, \ldots, g; j = 1, \ldots, p)$$

and

$$\frac{\partial Q}{\partial \beta} = \sum_{i=1}^{g}\sum_{k=1}^{K}\sum_{j=1}^{p} z_k[\bar{x}_{kij} - \hat{N}_{ki}\exp(\alpha_{ij} + \beta z_k)].$$

It is convenient to partition the Hessian matrix as

$$\mathbf{H} = \begin{pmatrix} a_{11} & \mathbf{a}'_{21} \\ \mathbf{a}_{21} & \mathbf{A}_{22} \end{pmatrix},$$

where a_{11} is the scalar

$$\frac{\partial^2 Q}{\partial \beta^2} = -\sum_{i=1}^{g}\sum_{k=1}^{K}\sum_{j=1}^{p} \hat{N}_{ki} z_k^2 \exp(\alpha_{ij} + \beta z_k),$$

$\mathbf{a}_{21}$ is the *gp*-vector of cross derivatives

$$\frac{\partial^2 Q}{\partial \alpha_{ij}\partial \beta} = -\sum_{k=1}^{K} \hat{N}_{ki} z_k \exp(\alpha_{ij} + \beta z_k),$$

and $\mathbf{A}_{22}$ is the $gp \times gp$ diagonal matrix with diagonal entries

$$\frac{\partial^2 Q}{\partial \alpha_{ij}^2} = -\sum_{k=1}^{K} \hat{N}_{ki} \exp(\alpha_{ij} + \beta z_k).$$

Note that, for $i \neq i'$ or $j \neq j'$,

$$\frac{\partial^2 Q}{\partial \alpha_{ij}\partial \alpha_{i'j'}} = 0,$$

so the off-diagonal elements of $\mathbf{A}_{22}$ are all 0. Thus $\mathbf{H}$ is almost diagonal, and can be inexpensively inverted using the following well-known result for partitioned matrices (see Morrison, 1990, p.67). Let

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix}.$$

Then

$$\mathbf{A}^{-1} =$$

$$\begin{pmatrix} (\mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21})^{-1} & -(\mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21})^{-1}\mathbf{A}_{12}\mathbf{A}_{22}^{-1} \\ -\mathbf{A}_{22}^{-1}\mathbf{A}_{21}(\mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21})^{-1} & \mathbf{A}_{22}^{-1} + \mathbf{A}_{22}^{-1}\mathbf{A}_{21}(\mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21})^{-1}\mathbf{A}_{12}\mathbf{A}_{22}^{-1} \end{pmatrix}$$

provided that the required inverses exist. For the random baseline model $[\alpha_{ij} + \beta z]$ we only have to compute the inverse of the *diagonal* matrix $\mathbf{A}_{22}$. We also need the inverse of $(\mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21})$, which in this case is the scaler $a_{11} - \mathbf{a}'_{21}\mathbf{A}_{22}^{-1}\mathbf{a}_{21}$.

Within-group Random Baseline Model $[\alpha_{ij} + \beta_i z]$

A similar Newton-Raphson procedure can be applied to the within-group random baseline model $[\alpha_{ij} + \beta_i z]$, where the $gp + g$ parameter vector $\boldsymbol{\theta}$ is

$$\boldsymbol{\theta} = (\beta_1, \ldots, \beta_g, \alpha_{11}, \ldots, \alpha_{g1}, \ldots, \alpha_{gp})'.$$

In the Newton-Raphson step

$$\hat{\boldsymbol{\theta}}^{\text{new}} = \hat{\boldsymbol{\theta}}^{\text{old}} - \mathbf{H}^{-1}\mathbf{g},$$

the gradient vector $\mathbf{g}$ contains the derivatives

$$\frac{\partial Q}{\partial \alpha_{ij}} = \sum_{k=1}^{K} [\bar{x}_{kij} - \hat{N}_{ki} \exp(\alpha_{ij} + \beta_i z_k)], \qquad (i = 1, \ldots, g; j = 1, \ldots, p)$$

and

$$\frac{\partial Q}{\partial \beta_i} = \sum_{k=1}^{K} \sum_{j=1}^{p} z_k [\bar{x}_{kij} - \hat{N}_{ki} \exp(\alpha_{ij} + \beta_i z_k)], \qquad (i = 1, \ldots, g).$$

As before, we partition the Hessian matrix as

$$\mathbf{H} = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix},$$

where $\mathbf{A}_{11}$ is the $g \times g$ diagonal matrix with diagonal entries

$$\frac{\partial^2 Q}{\partial \beta_i^2} = -\sum_{k=1}^{K} \sum_{j=1}^{p} \hat{N}_{ki} z_k^2 \exp(\alpha_{ij} + \beta_i z_k),$$

$\mathbf{A}_{12}$ is the $g \times gp$ matrix of cross derivatives

$$\frac{\partial^2 Q}{\partial \beta_i \partial \alpha_{ij}} = -\sum_{k=1}^{K} \hat{N}_{ki} z_k \exp(\alpha_{ij} + \beta_i z_k),$$

where $\frac{\partial^2 Q}{\partial \beta_{i'} \partial \alpha_{ij}} = 0$ for $i' \neq i$,

$$\mathbf{A}_{21} = \mathbf{A}_{12}',$$

and $\mathbf{A}_{22}$ is the $gp \times gp$ diagonal matrix with diagonal entries

$$\frac{\partial^2 Q}{\partial \alpha_{ij}^2} = -\sum_{k=1}^{K} \hat{N}_{ki} \exp(\alpha_{ij} + \beta_i z_k).$$

The matrix inversion formula for partitioned matrices given in the previous section requires inverses for the matrices $\mathbf{A}_{22}$ and $\mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}$. $\mathbf{A}_{22}$ is diagonal, and inexpensively invertible. $\mathbf{A}_{11}$ is diagonal, and $\mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}$ can be shown to be diagonal, as follows. We can write $\mathbf{A}_{12}$ as

$$\mathbf{A}_{12} = \begin{pmatrix} \mathbf{a}_1' & \mathbf{0}' & \cdots & \mathbf{0}' \\ \mathbf{0}' & \mathbf{a}_2' & & \mathbf{0}' \\ \vdots & & \ddots & \vdots \\ \mathbf{0}' & \mathbf{0}' & \cdots & \mathbf{a}_g' \end{pmatrix},$$

where $\mathbf{a}_i = (a_{i1}, \cdots, a_{ip})'$ is a p vector containing the derivatives

$$\frac{\partial^2 Q}{\partial \beta_i \partial \alpha_{ij}}, \qquad (j = 1, \ldots, p).$$

Let $\mathbf{d} = (\mathbf{d}_1', \cdots, \mathbf{d}_g')'$ be the gp vector of diagonal entries of $\mathbf{A}_{22}^{-1}$, with $\mathbf{d}_i = (d_{i1}, \cdots, d_{ip})'$ corresponding to the partial derivatives for the i^{th} group. Then $\mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}$ is a $g \times g$ diagonal matrix with diagonal entries given by

$$\sum_{j=1}^{p} d_{ij} a_{ij}^2, \qquad (i = 1, \ldots, g).$$

Thus $\mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}$ is diagonal and inexpensively invertible.

Group Invariant Model $[\alpha_{ij} + \beta_j z]$

In the group invariant model $[\alpha_{ij} + \beta_j z]$, the $gp + p$ parameter vector is

$$\boldsymbol{\theta} = (\beta_1, \cdots, \beta_p, \alpha_{11}, \ldots, \alpha_{1p}, \ldots, \alpha_{g1}, \ldots, \alpha_{gp})'.$$

The gradient vector $\mathbf{g}$ in the Newton-Raphson algorithm contains the derivatives

$$\frac{\partial Q}{\partial \alpha_{ij}} = \sum_{k=1}^{K} [\bar{x}_{kij} - \hat{N}_{ki} \exp(\alpha_{ij} + \beta_j z_k)], \qquad (i = 1, \ldots, g; j = 1, \ldots, p)$$

and

$$\frac{\partial Q}{\partial \beta_j} = \sum_{k=1}^{K} \sum_{i=1}^{g} z_k [\bar{x}_{kij} - \hat{N}_{ki} \exp(\alpha_{ij} + \beta_j z_k)].$$

The Hessian matrix is

$$\mathbf{H} = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix},$$

where $\mathbf{A}_{11}$ is the $p \times p$ diagonal matrix with diagonal entries

$$\frac{\partial^2 Q}{\partial \beta_j^2} = -\sum_{k=1}^{K} \sum_{i=1}^{g} \hat{N}_{ki} z_k^2 \exp(\alpha_{ij} + \beta_j z_k),$$

$\mathbf{A}_{12}$ is the $p \times gp$ diagonal matrix with diagonal entries

$$\frac{\partial^2 Q}{\partial \beta_j \partial \alpha_{ij}} = -\sum_{k=1}^{K} \hat{N}_{ki} z_k \exp(\alpha_{ij} + \beta_j z_k),$$

$$\mathbf{A}_{21} = \mathbf{A}_{12}',$$

and $\mathbf{A}_{22}$ is the $gp \times gp$ diagonal matrix with diagonal entries

$$\frac{\partial^2 Q}{\partial \alpha_{ij}^2} = -\sum_{k=1}^{K} \hat{N}_{ki} \exp(\alpha_{ij} + \beta_j z_k).$$

As in the model $[\alpha_{ij} + \beta_i z]$, the matrix $\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}$ is diagonal. Hence $\mathbf{H}$ is inexpensively inverted using the inversion formula for partitioned matrices.

Factor Scores

In the SIMS application, we are primarily interested in partitioning the observations into distinct chemical classes. The factor scores are of secondary interest. There may be occasions, however, when factor scores also are of interest. For example, in SIMS image segmentation, the interference or random baseline may be caused by variations in topography. We may be interested in creating separate topographic and chemical maps. The topographic map would consist of predicted factor scores at each pixel. We will use, as factor score for the h^{th} observation,

$$\hat{z}_h \doteq \hat{\mathrm{E}}(z_h | \mathbf{x}_h) = \sum_{k=1}^{K} z_k \hat{h}_{k|h}, \quad \text{where} \quad \hat{h}_{k|h} = \sum_{i=1}^{g} \hat{h}_{ki|h}.$$

Examples

Several simulations were run to compare the performance of the Poisson latent profile model, the multinomial model, and the Poisson latent variable mixture models on parameter recovery and classification. The following discussion will consider the performance of the methods on data generated from the random baseline model $[\alpha_{ij} + \beta z]$, and the unrestricted latent variable model $[\alpha_{ij} + \beta_{ij} z]$. We focus on differences between the methods in classification performance.

Simulation 6

In Simulation 6, 2000 observations were generated from a mixture of two Poisson random baseline models $[\alpha_{ij} + \beta z]$ with $p = 10$ variables for various choices of β. In all simulations the mixing parameters were $\eta_1 = \eta_2 = 1/2$, and the intercept parameters $\boldsymbol{\alpha}_i = (\alpha_{i1}, \cdots, \alpha_{ip})$ were

$$\boldsymbol{\alpha}_1 = (-3, -2, -1, 0, 1, -3, -2, -1, 0, 1), \qquad \boldsymbol{\alpha}_2 = (0, 0, 0, 0, 0, 0, 0, 0, 0, 0).$$

For each value of β, 100 replications of the experiment were performed. In each replication, misclassification rates were computed for the Poisson latent profile model, the multinomial mixture model, the Poisson random baseline mixture model $[\alpha_{ij}+\beta z]$, and the unrestricted Poisson latent variable mixture model $[\alpha_{ij} + \beta_{ij} z]$. In addition, three continuous variable clustering methods were applied to the data:

- M_1: K-means cluster analysis applied to the raw count data.
- M_2: K-means cluster analysis applied to Anscombe transformed data.

- M_3: The raw count data are Anscombe transformed, then mean-centered, then K-means cluster analysis applied.

Cases were classified to groups by matching the recovered groups (determined by posterior probabilities) with known groups. Because of the arbitrary labeling of the recovered groups, two matchings are possible (i.e., identify recovered group 1 with true group 1, or identify recovered group 2 with true group 1). We always chose that matching which yielded fewest misclassifications. This introduced an optimistic bias in reported error rates, but the bias decreases rapidly with increasing sample size.

Average misclassification rates are reported in Table 12.

Observations:

1. The Poisson latent profile model performed well for small β but broke down with increasing β.

2. The three models, $[\alpha_{ij} + \beta z]$, $[\alpha_{ij} + \beta_{ij} z]$ and the multinomial, performed well for all values of β, though $[\alpha_{ij} + \beta z]$, which is the correct model for these data, performed slightly better than $[\alpha_{ij} + \beta_{ij} z]$ and the multinomial model. These three methods also performed significantly better than the continuous variable methods, probably because the observed counts were so low.

3. Of the three continuous variable methods, the Anscombe-transform, mean-centered approach performed best.

4. For all methods, misclassification rates increased with β. For the multinomial, $[\alpha_{ij} + \beta z]$ and $[\alpha_{ij} + \beta_{ij} z]$, the increase was very slight.

β	Poisson $[\alpha_{ij}]$	Multinomial	$[\alpha_{ij}+\beta z]$	$[\alpha_{ij}+\beta_{ij} z]$	M_1	M_2	M_3
0	10.2	10.7	10.3	10.2	26.6	15.4	15.7
.1	10.5	11.0	10.6	10.6	27.3	15.5	15.8
.3	11.3	11.4	10.7	10.7	35.1	17.2	16.3
.5	13.9	11.9	11.0	11.2	38.4	28.8	19.2
.6	23.2	11.9	11.0	11.2	40.6	38.3	19.5
.65	37.0	12.3	11.3	11.4	43.0	44.6	20.0
.7	43.3	12.3	11.3	11.4	43.2	46.9	21.7
1.0	47.0	13.5	12.6	12.7	47.3	48.9	30.4
1.1	48.6	13.7	12.7	13.0	47.8	49.1	32.5

Table 12: Average percent misclassifications for Simulation Experiment 6. Data were generated from random baseline model $[\alpha_{ij}+\beta z]$ for various β. The methods M_1 ,M_2 and M_3 correspond to K-means, Anscombe transformed K-means, and Anscombe transformed-mean centered K means.

Simulation 7

Conditions for Simulation 7 were identical to those for the first experiment, except the intercept parameters $\boldsymbol{\alpha}_2 = (\alpha_{21}, \ldots, \alpha_{1p})$ in G_2 are now

$$\boldsymbol{\alpha}_2 = (-3, -3, -3, -3, -3, -3, -3, -3, -3, -3).$$

The groups are more separated than in Simulation 5, but data generated from G_2 are extremely sparse. For $\beta = 0$ the mean response is $\exp(-3) = .0498$. For $\beta = 1$, the mean response is $\exp(-2.5) = .0821$. The expected number of observations with all zero counts is 61 % ($\beta = 0$) and 55 % ($\beta = 1$). Average misclassification rates are given in Table 13.

The Poisson latent variable mixture models $[\alpha_{ij} + \beta z]$ and $[\alpha_{ij} + \beta_{ij} z]$ significantly outperformed all other methods. Note that the multinomial mixture model completely breaks down for all β. To see why, consider the expression for the posterior probability (4.12) in the multinomial model. If $\mathbf{x}_h = \mathbf{0}$, then $\hat{h}_{i|h} = \frac{\eta_i}{\sum \eta_l}$, so the h^{th} observation is classified to the most prevalent group, regardless of the group's response profile. This is a result of the assumption $Pr(G_i|M_h = m_h) = Pr(G_i)$ in (4.4). The latent variable mixture model, by contrast, considers the sum of mean

β	Poisson $[\alpha_{ij}]$	Multinomial	$[\alpha_{ij} + \beta z]$	$[\alpha_{ij} + \beta_{ij} z]$	M_1	M_2	M_3
0	4.4	48.0	4.4	4.4	19.6	8.7	13.5
.3	5.9	48.0	5.3	5.3	22.4	10.4	14.4
.5	9.0	48.0	7.1	7.1	28.6	16.6	16.5
.7	15.7	47.3	8.5	8.6	36.0	21.2	19.3
1.0	29.4	46.5	12.2	12.6	43.5	31.6	28.3
1.1	32.4	45.5	13.4	13.7	44.9	33.7	30.2

Table 13: Average percent misclassifications for Simulation Experiment 7. Data were generated from random baseline model $[\alpha_{ij} + \beta z]$ for various β. The methods M_1, M_2 and M_3 correspond to K-means, Anscombe transformed K-means, and Anscombe transformed-mean centered K means.

Poisson $[\alpha_{ij}]$	Multinomial	$[\alpha_{ij} + \beta z]$	$[\alpha_{ij} + \beta_{ij} z]$	M_1	M_2	M_3
28.5	17.8	16.6	10.3	47.5	27.2	21.3

Table 14: Average percent misclassifications for Simulation Experiment 8. The methods M_1, M_2 and M_3 correspond to K-means, Anscombe transformed K-means, and Anscombe transformed-mean centered K means.

responses, $\sum \lambda_{ij}(z)$, in assigning $\mathbf{x}_h = \mathbf{0}$ to a group. This can be seen in equations (4.14) and (4.16).

Simulation 8

In Simulation 8, observations were generated from the unrestricted Poisson latent variable model $[\alpha_{ij} + \beta_{ij} z]$ with $\eta_1 = \eta_2 = 1/2$. The intercept and slope parameters $\boldsymbol{\alpha}_i = (\alpha_{i1}, \ldots, \alpha_{ip})$ and $\boldsymbol{\beta}_i = (\beta_{i1}, \ldots, \beta_{ip})$ are

$$\boldsymbol{\alpha}_1 = (-3, -2, -1, 0, 1), \qquad \boldsymbol{\alpha}_2 = (0, 0, 0, 0, 0),$$

and

$$\boldsymbol{\beta}_1 = (0, 0, 0, 0, 0), \qquad \boldsymbol{\beta}_2 = (0, 0, 0, 1, 1).$$

In each of 1000 replications of the experiment, 2000 observations were generated from the mixture distribution, and the methods applied. Misclassification rates are reported in Table 14. As expected, the model $[\alpha_{ij} + \beta_{ij} z]$ performed best.

Discussion

The results of the three simulation experiments reported in this chapter suggest that cluster analysis of low count data can be improved if 1) more realistic count distribution models are used instead of Gaussian or other continuous variable methods, and 2) the correlation structure of the variables is properly accounted for. More work needs to be done to determine the conditions under which these new methods might be useful. As with all statistical methods, the real value of the methods described in this paper depends on their applicability to real problems.

In the mixture models discussed in this chapter, the conditional distribution of the observed variables, $g_{ij}(x_j|z)$, are taken to be Poisson. The models are readily extended to allow conditional distributions to be any member of the exponential family. In the extended model,

$$g_{ij}(x_j|z) = \exp\left\{\frac{x_j\theta_{ij}(z) - b_{ij}(\theta_{ij}(z))}{\phi_{ij}} + c_{ij}(x_j, \phi_{ij})\right\}, \quad (i = 1, \ldots, g; j = 1, \ldots, p),$$

where the canonical parameter θ_{ij} is linear in the latent variable z

$$\theta_{ij}(z) = \alpha_{ij} + \beta_{ij} z.$$

In the Poisson response model (described by (4.5) and (4.6)),

$$g_{ij}(x_j|z) = \frac{\exp(-\lambda_{ij}(z))\lambda_{ij}(z)^{x_j}}{x_j!} = \exp\left\{[x_j \log\lambda_{ij}(z) - \lambda_{ij}(z)] - \log x_j!\right\}.$$

Thus,

$$\theta_{ij}(z) = \log\lambda_{ij}(z),$$

$$b_{ij}(\theta_{ij}(z)) = \exp(\theta_{ij}(z)) = \exp(\alpha_{ij} + \beta_{ij} z),$$

and

$$\phi_{ij} = 1.$$

Normal, gamma, and multinomial distributions can similarly be shown to be members of the exponential family. For the general case, the conditional expectation of the complete data log-likelihood $Q = E^{\Psi}_{v,y}(L_c)$ is

$$Q = \sum_{i=1}^{g} \hat{N}_i \log \eta_i + \sum_{i=1}^{g} \sum_{k=1}^{K} \sum_{j=1}^{p} [\bar{x}_{kij} \theta_{ij}(z_k) - \hat{N}_{ki} b_{ij}(\theta_{ij}(z_k))].$$

Unknown papameters can be estimated using Newton-Raphson methods.

This chapter was motivated by a problem in SIMS image segmentation, where it was noticed that without baseline correction chemical classes may be confounded with topographic classes. In an application, Willse and Tyler (1998) adapted Poisson and multinomial models to handle the spatial correlations among the pixels by introducing a locally dependent Markov random field as the probability distribution for class assignments. A similar approach could by applied to latent variable mixture models, with the goal of separating chemical effects (defined by the factor-to-variable transformation mechanism) from topographic effects (defined by the latent variable).

CHAPTER 5

Conclusion

The research presented in this thesis was motivated by the observation that non-standard data types are common in practice, but that there is a shortage of methods for analyzing these types of data. In practice, normal-based methods are often the method of choice. Indeed, many practitioners may be unfamiliar with other approaches. Often these methods are sufficient. But if the data have special, non-normal structure, we often can improve classification by more carefully modeling the data. That, in fact, is the main conclusion of this thesis: we can often improve classification by carefully modeling the data. Classification procedures for mixed-mode and multivariate count data were developed in Chapters 2, 3 and 4, and were shown to give better results than standard methods under special circumstances.

More work needs to be done to determine the conditions under which these methods will be expected to significantly outperform traditional methods. It would be useful to develop guidelines for the practitioner. This future research will likely be driven by the demands of applications, such as text analysis, which should spur the development of more specialized algorithms for analysis of mixed mode and multivariate count data.

APPENDICES

APPENDIX A

Some Useful Theorems for Covariance Model Estimation

Proofs of Theorems 1-3 and Corollaries 1-2 are given in Celeux and Govaert (1995). Theorem 4 is an adaptation of the FG algorithm of Flury and Gautschi (1986). The adaptation is given in Bensmail and Celeux (1996).

Theorem 1 *The minimum of* $tr(\mathbf{QM}^{-1})$ *with respect to the* $p \times p$ *symmetric matrix* $\mathbf{M}$ *where* $\mathbf{Q}$ *is positive definite and* $|\mathbf{M}| = 1$ *is* $p|\mathbf{Q}|^{1/p}$. *The minimizer is*

$$\mathbf{M} = \frac{\mathbf{Q}}{|\mathbf{Q}|^{1/p}}.$$

Corollary 1 *The minimum of* $tr(\mathbf{QM}^{-1})$ *with respect to the* $p \times p$ *diagonal matrix* $\mathbf{M}$ *where* $\mathbf{Q}$ *is positive definite and* $|\mathbf{M}| = 1$ *is* $p|diag(\mathbf{Q})|^{1/p}$. *The minimizer is*

$$\mathbf{M} = \frac{diag(\mathbf{Q})}{|diag(\mathbf{Q})|^{1/p}}.$$

Theorem 2 *The minimizer of* $tr(\mathbf{QM}^{-1}) + \alpha \log |\mathbf{M}|$ *with respect to the* $p \times p$ *symmetric matrix* $\mathbf{M}$ *where* $\mathbf{Q}$ *is positive definite and* α *is a positive real number is* $\mathbf{M} = \frac{1}{\alpha}\mathbf{Q}$.

Corollary 2 *The minimizer of* $tr(\mathbf{QM}^{-1}) + \alpha \log |\mathbf{M}|$ *with respect to the* $p \times p$ *diagonal matrix* $\mathbf{M}$ *where* $\mathbf{Q}$ *is positive definite and* α *is a positive real number is* $\mathbf{M} = \frac{1}{\alpha} diag(\mathbf{Q})$.

Theorem 3 *The minimum of tr(*$\mathbf{QAQ}^{-1}\mathbf{B}$*) with respect to the orthogonal matrix* $\mathbf{Q}$*, where* $\mathbf{A}$ *and* $\mathbf{B}$ *are diagonal matrices with diagonal terms* α_j *and* β_j *such that* $\alpha_1 \geq \alpha_2 \geq \cdots \geq \alpha_p$ *and* $\beta_1 \leq \beta_2 \leq \cdots \leq \beta_p$ *is tr(*$\mathbf{AB}$*)* $= \sum_j \alpha_j \beta_j$*, and the minimizer is the identity matrix.*

Theorem 4 *The* $p \times p$ *orthogonal matrix* $\mathbf{D}$ *minimizing*

$$f(\mathbf{D}) = \sum_{i=1}^{K} tr(\mathbf{D}\mathbf{A}_i^{-1}\mathbf{D}'\mathbf{W}_i)$$

where $\mathbf{A}_1, \cdots, \mathbf{A}_K$ *are fixed diagonal matrices and* $\mathbf{W}_i$ *are symmetric matrices can be obtained iteratively as follows.*

Step 1. Start with an initial solution $\mathbf{D} = (\mathbf{d}_1, \cdots, \mathbf{d}_p)$.

Step 2. For any l *and* $m \in \{1, \ldots, p\}, l \neq m$, *the pair* $(\mathbf{d}_l, \mathbf{d}_m)$ *is replaced with* $(\boldsymbol{\delta}_l, \boldsymbol{\delta}_m)$, *where* $\boldsymbol{\delta}_l$ *and* $\boldsymbol{\delta}_m$ *are orthonormal vectors, linear combinations of* $\mathbf{d}_l$ *and* $\mathbf{d}_m$, *such that*

$$\boldsymbol{\delta}_l = (\mathbf{d}_l, \mathbf{d}_m)\mathbf{q}_1 \qquad \textit{and} \qquad \boldsymbol{\delta}_m = (\mathbf{d}_l, \mathbf{d}_m)\mathbf{q}_2$$

where $\mathbf{q}_1$ *and* $\mathbf{q}_2$ *are orthonormal vectors in* $\Re^2$. *The vector* $\mathbf{q}_1$ *is the eigenvector associated with the smallest eigenvalue of the matrix* $\sum_{i=1}^{K}[(1/a_i^l) - (1/a_i^m)]\mathbf{Z}_i$, *where* $\mathbf{Z}_i = (\mathbf{d}_l, \mathbf{d}_m)'\mathbf{W}_i(\mathbf{d}_l, \mathbf{d}_m)$, *and* a_i^l *and* a_i^m *are the* l^{th} *and* m^{th} *diagonal entries of* $\mathbf{A}_i^{-1}$.

Step 3. Repeat Step 2 until the change in the estimate $\mathbf{D}$ *between successive iterations is sufficiently small.*

APPENDIX B

Estimation of Some Common Covariance Models

In the homogeneous covariance model $[\rho\mathbf{\Gamma}\mathbf{\Lambda}\mathbf{\Gamma}']$ the objective function reduces to

$$F = N\log|\mathbf{\Sigma}| + \mathrm{tr}(\mathbf{\Sigma}^{-1}\mathbf{W}),$$

which is minimized by $\hat{\mathbf{\Sigma}} = \frac{\mathbf{W}}{N}$.

In the homogeneous covariance diagonal model $[\rho\mathbf{\Lambda}]$,

$$F = pN\log\rho + \frac{1}{\rho}\mathrm{tr}(\mathbf{\Lambda}^{-1}\mathbf{W}),$$

which is minimized by

$$\hat{\mathbf{\Lambda}} = \frac{\mathrm{diag}(\mathbf{W})}{|\mathrm{diag}(\mathbf{W})|^{1/p}}$$

and

$$\hat{\rho} = \frac{1}{N}|\mathrm{diag}(\mathbf{W})|^{1/p}.$$

In the proportional covariance model $\mathbf{\Sigma}_{is} = \rho_{is}\mathbf{\Gamma}\mathbf{\Lambda}\mathbf{\Gamma}'$, it is convenient to write $\mathbf{C} = \mathbf{\Gamma}\mathbf{\Lambda}\mathbf{\Gamma}'$. Then the objective function to minimize is

$$F = p\sum_{i=1}^{K}\sum_{s=1}^{m} n_{is}\log\rho_{is} + \sum_{i=1}^{K}\sum_{s=1}^{m}\frac{1}{\rho_{is}}\mathrm{tr}(\mathbf{C}^{-1}\mathbf{W}_{is}).$$

The parameters can be estimated iteratively.

- For fixed $\mathbf{C}$,

$$\hat{\rho}_{is} = \frac{1}{pn_{is}}\mathrm{tr}(\mathbf{C}^{-1}\mathbf{W}_{is}).$$

- For fixed ρ_{is}, minimize with respect to $\mathbf{C}$ the function

$$f(\mathbf{C}) = \mathrm{tr}(\mathbf{C}^{-1}\sum_{i=1}^{K}\sum_{s=1}^{m}(\mathbf{W}_{is}/\rho_{is})).$$

By Corollary 1,

$$\hat{\mathbf{C}} = \frac{\sum_{i=1}^{K}\sum_{s=1}^{m}\frac{1}{\rho_{is}}\mathbf{W}_{is}}{|\sum_{i=1}^{K}\sum_{s=1}^{m}\frac{1}{\rho_{is}}\mathbf{W}_{is}|^{1/p}}.$$

Starting values for ρ_{is} can be obtained by

$$\hat{\rho}_{is} = \left| \frac{\mathbf{W}_{is}}{n_{is}} \right|^{1/p}.$$

Estimation of the proportional covariance model $[\rho_i \mathbf{\Gamma\Lambda\Gamma}'] = [\rho_i \mathbf{C}]$ is similar:

- For fixed ρ_{is}, by Theorem 1

$$\hat{\mathbf{C}} = \frac{\sum_{i=1}^{K} \frac{1}{\rho_i} \mathbf{W}_{i\cdot}}{|\sum_{i=1}^{K} \frac{1}{\rho_i} \mathbf{W}_{i\cdot}|^{1/p}}.$$

- For fixed $\mathbf{C}$,

$$\hat{\rho}_i = \frac{1}{pn_{i\cdot}} \mathrm{tr}(\mathbf{C}^{-1}\mathbf{W}_{i\cdot}).$$

Starting values for ρ_i can be obtained by

$$\hat{\rho}_i = \left| \frac{\mathbf{W}_{i\cdot}}{n_i} \right|^{1/p}.$$

Similarly, the models $[\rho_i \mathbf{\Lambda}]$ and $[\rho_{is} \mathbf{\Lambda}]$ can be obtained as above by setting $\mathbf{\Gamma} = \mathbf{I}$.

In the common principal components model $[\rho_i \mathbf{\Gamma\Lambda}_i \mathbf{\Gamma}']$, it is convenient to write $\mathbf{\Sigma}_{is} = \mathbf{\Gamma A}_i \mathbf{\Gamma}'$ where $\mathbf{A}_i = \rho_i \mathbf{\Lambda}_i$. Then the objective function to minimize is

$$F = \sum_{i=1}^{K} n_{i\cdot} \log |\mathbf{A}_i| + \sum_{i=1}^{K} \mathrm{tr}(\mathbf{\Gamma A}_i^{-1} \mathbf{\Gamma}' \mathbf{W}_{i\cdot}).$$

- For fixed $\mathbf{\Gamma}$, by Corollary 2,

$$\hat{\mathbf{A}}_i = \frac{1}{n_{i\cdot}} \mathrm{diag}(\mathbf{\Gamma W}_{i\cdot} \mathbf{\Gamma}').$$

- For fixed $\mathbf{A}_i$, $\hat{\mathbf{\Gamma}}$ can be obtained using Theorem 4.

In the model $[\rho_{is} \mathbf{\Gamma\Lambda}_s \mathbf{\Gamma}']$, the objective function is

$$F = p \sum_{i=1}^{K} n_{i\cdot} \log \rho_i + \sum_{i=1}^{K} \sum_{s=1}^{m} \frac{1}{\rho_i} \mathrm{tr}(\mathbf{\Gamma\Lambda}_s^{-1} \mathbf{\Gamma}').$$

Parameter estimates can be obtained iteratively as follows.

- For fixed ρ_i and $\mathbf{\Gamma}$, minimize with respect to $\mathbf{\Lambda}_s$ the function

$$f(\mathbf{\Lambda}_s) = \text{tr}(\mathbf{\Lambda}_s^{-1}\mathbf{\Gamma}\sum_{i=1}^{K}(\mathbf{W}_{is}/\rho_i)\mathbf{\Gamma}).$$

  By Corollary 1,

$$\hat{\mathbf{\Lambda}}_s = \frac{\sum_{i=1}^{K}\mathbf{\Gamma}'(\mathbf{W}_{is}/\rho_i)\mathbf{\Gamma}}{|\sum_{i=1}^{K}\mathbf{\Gamma}'(\mathbf{W}_{is}/\rho_i)\mathbf{\Gamma}|^{1/p}}.$$

- For fixed $\mathbf{\Gamma}$ and $\mathbf{\Lambda}_s$,

$$\hat{\rho}_i = \frac{1}{pn_i}\sum_{s=1}^{m}\text{tr}(\mathbf{\Gamma}\mathbf{\Lambda}_s^{-1}\mathbf{\Gamma}'\mathbf{W}_{is}).$$

- For fixed ρ_i and $\mathbf{\Lambda}_s$, minimize with respect to $\mathbf{\Gamma}$ the function

$$f(\mathbf{\Gamma}) = \sum_{i=1}^{K}\sum_{s=1}^{m}\frac{1}{\rho_i}\text{tr}(\mathbf{\Gamma}\mathbf{\Lambda}_{is}^{-1}\mathbf{\Gamma}'\mathbf{W}_{is}),$$

  which can be done using Theorem 4.

REFERENCES CITED

Adriaens, A. and Adams, F. (1991) Comparison between the precision characteristics of the magnetic and electrostatic peak switching systems in secondary ion mass spectrometry. *International Journal of Mass Spectrometry and Ion Processes*, **108**, 41-52.

Anderson, T.W. (1984) *An Introduction to Multivariate Statistical Analysis*, Second Edition, Wiley, New York.

Banfield, J.D. and Raftery, A.E. (1993) Model-based Gaussian and non-Gaussian clustering. *Biometrics*, **43**, 803-821.

Bartholomew, D.J. (1987) *Latent Variable Models and Factor Analysis*, Charles Griffin, London.

Benninghoven, A. (1994) Surface analysis by secondary ion mass spectrometry (SIMS). *Surface Science*, **299**, 246-260.

Benninghoven, A., Hagenhoff, B., and Niehuis, E. (1993) Surface MS: Probing realworld samples. *Analytical Chemistry*, **65**, 630-639.

Bensmail, H. and Celeux, G. (1996) Regularized Gaussian discriminant analysis through eigenvalue decomposition. *Journal of the American Statistical Association*, **91**, 1743-1748.

Celeux, G. and Govaert, G. (1991) Clustering criteria for discrete data and latent class models. *Journal of Classification*, **8**, 157-176.

Celeux, G. and Govaert, G. (1995) Gaussian parsimonious clustering models. *Pattern Recognition*, **28**, 781-793.

Clogg, C. C. (1993) Latent class models. In Arminger, G., Clogg, C. C., and Sobel, M. E. (eds), *Handbook of Statistical Modeling for the Social and Behavioral Sciences*, Plenum Press, New York.

Clogg, C. C. and Goodman, L. A. (1984) Latent structure analysis of a set of multidimensional contingence tables. *Journal of the American Statistical Association*, **79**, 762-771.

Dillon, W.R. and Mulani, N. (1989) LADI: A latent discriminant model for analyzing marketing research data. *Journal of Marketing Research*, **25**, 15-29.

Everitt, B.S. (1988) A finite mixture model for the clustering of mixed-mode data. *Statistics and Probability Letters*, **6**, 305-309.

Everitt, B.S. and Merette, C. (1990) The clustering of mixed-mode data: a comparison of possible approaches. *Journal of Applied Statistics*, **17**, 283-297.

Flury, B.W. (1984) Common principal components in k groups. *Journal of the American Statistical Association*, **79**, 892-897.

Flury, B.W. and Gautschi, W. (1986) An algorithm for simultaneous orthogonal transformation of several positive definite symmetric matrices to nearly diagonal form. *SIAM Journal of Scientific Statistical Computing*,**7**, 169-184.

Flury, B.W., Schmid, M.J. and Narayanan, A. (1993) Error rates in quadratic discrimination with constraints on the covariance matrices. *Journal of Classification*, **11**, 101-120.

Ge, N. and Simpson, D.G. (1998) Correlation and high-dimensional consistency in pattern recognition. *Journal of the American Statistical Association*, **93**, 995-1006.

Hastie, T., Tibshirani, R. and Buja, A. (1997) Flexible discriminant and mixture models. Technical Report Number 189, Division of Biostatistics, Stanford University.

Kargacin, M.E. and Kowalski, B.R. (1986) Ion intensity and image resolution in secondary ion mass spectrometry. *Analytical Chemistry*, **58**, 2300-2306.

Krzanowski, W.J. (1975) Discrimination and classification using both binary and continuous variables. *Journal of the American Statistical Association*, **70**, 782-790.

Krzanowski, W.J. (1980) Mixtures of continuous and categorical variables in discriminant analysis. *Biometrics*, **36**, 493-499.

Krzanowski, W.J. (1993) The location model for mixtures of categorical and continuous variables. *Journal of Classification*, **10**, 25-49.

Krzanowski, W.J. (1994) Quadratic location discriminant functions for mixed categorical and continuous data. *Statistics and Probability Letters*, **19**, 91-95.

Lawrence, C.J. and Krzanowski, W.J. (1996) Mixture separation for mixed-mode data. *Statistics and Computing*, **6**, 85-92.

McCullagh, P. and Nelder, J.A. (1989) *Generalized Linear Models*, Second Edition, Chapman and Hall, London.

McLachlan, G.J. and Basford, K.E. (1988) *Mixture Models: Inference and Applications to Clustering*, Marcel Dekker, New York.

McLachlan, G.J. and Krishnan, T. (1997) *The EM Algorithm and Extensions*, Wiley, New York.

Milligan, G.W. and Cooper, M.C. (1988) A study of standardization of variables in cluster analysis. *Journal of Classification*, **5**, 181-205.

Morrison, D.G. (1990) *Multivariate Statistical Methods*, Third Edition, McGraw-Hill, New York.

Moustaki, I. and Knott, M. (1997) Generalized latent trait models. Technical report LSERR36, London School of Economics.

Redner, R.A. and Walker, H.F. (1984) Mixture densities, maximum likelihood and the EM algorithm. *SIAM Review*, **26**, 195-239.

Sammel, M.D., Ryan, L.M., and Legler, J.M. (1997) Latent Variable Models for Mixed Discrete and Continuous Outcomes. *Journal of the Royal Statistical Society, Series B*, **59**, 667-678.

Schweiters, J., Cramer, H.G., Heller, T., Jurgens, U., Niehuis, E., Zehnpfenning, J. and Benninghoven, A. (1991) High mass resolution surface imaging with a time-of-flight secondary ion mass spectroscopy scanning microprobe. *Journal of Vacuum Science and Technology A*, **9**, 2864-2871.

Starck, J.L., Murtagh, F., and Bijaoui, A. (1998) *Image Processing and Data Analysis: The Multiscale Approach*, Cambridge University Press, Cambridge.

Titterington, D.M., Murray, G.D., Murray, L.S., Spiegelhalter, D.J., Skene, A.M., Habbema, J.D.F., and Gelpke, G.J. (1981) Comparison of discrimination techniques applied to a complex data set of head injury patients (with discussion). *Journal of the Royal Statistical Society, Series A*, **144**, 145-175.

Titterington, D.M., Smith, A.F.M., and Makov, U.E. (1985) *Statistical Analysis of Finite Mixture Distributions*, Wiley, Chichester.

Ubersax, J.S. (1993) Statistical modeling of expert ratings on medical treatment appropriateness. *Journal of the American Statistical Association*, **88**, 421-427.

Ubersax, J.S. (1999) Probit latent class analysis: conditional independence and conditional dependence models. *Applied Psychological Measurement* (In press. Preprint available at http://members.xoom.com/jsubersax).

Whittaker, J. (1990) *Graphical Models in Applied Multivariate Statistics*, Wiley, Chichester.

Willse, A. and Tyler, B. (1998) Multivariate methods for TOF-SIMS imaging, in Gillen, G., Lareau, D., Bennett, J. and Stevie, F. (eds) *Secondary Ion Mass Spectrometry: SIMS XI*, Wiley, New York.

Yakowitz, S.J. and Spragins, J.D. (1968) On the identifiability of finite mixtures. *Annals of Mathematical Statistics*, **40**, 1728-1735.

Yung, Y. (1997) Finite mixtures in confirmatory factor-analysis models. *Psychometrika*, **62**, 297-330.