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
Date

Robert J. Boik
Chairperson, Graduate Committee

Approved for the Major Department

11/1/99
Date

John Lund
Head, Mathematical Sciences

Approved for the College of Graduate Studies

11-2-99
Date

Bruce R. 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 Wilbe

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 $g_i(\mathbf{x}; \boldsymbol{\Psi}_i)$. The parameter vector $\boldsymbol{\Psi}_i$ 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

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

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

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

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

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

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

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

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 $\{p(\mathbf{x}; \boldsymbol{\Theta})\}$ is identifiable if for $\boldsymbol{\Theta}$ and $\boldsymbol{\Theta}'$ in the family's parameter space, $p(\mathbf{x}; \boldsymbol{\Theta}) \equiv p(\mathbf{x}; \boldsymbol{\Theta}') \Rightarrow \boldsymbol{\Theta} = \boldsymbol{\Theta}'$. 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 $\boldsymbol{\Psi} = \{\alpha_i, p_{is}, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}\}$ and $\boldsymbol{\Psi}' = \{\alpha_i', p_{is}', \boldsymbol{\mu}_i', \boldsymbol{\Sigma}'\}$ are considered to be *equivalent* if they can be made identical by permuting group labels. Otherwise they are distinct. For example, the parameter set $\boldsymbol{\Psi} = \{\alpha_i, p_{is}, \boldsymbol{\mu}, \boldsymbol{\Sigma}\}$ for $g = 2$ groups and m locations is equivalent to the parameter set $\boldsymbol{\Psi}'$ obtained by $\alpha_1' = \alpha_2$, $\alpha_2' = \alpha_1$, $\boldsymbol{\Sigma}' = \boldsymbol{\Sigma}$, 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 $\mathbf{x}_{sh} \in \Re^p$

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

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

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

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 $\boldsymbol{\Psi}' = \{\alpha'_i, p'_{is}, \boldsymbol{\mu}'_{is}, \boldsymbol{\Sigma}'\}$, distinct from $\boldsymbol{\Psi} = \{\alpha_i, p_{is}, \boldsymbol{\mu}_{is}, \boldsymbol{\Sigma}\}$, 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 $\sum_{i=1}^{2} f'_{is} h(\mathbf{x}_{sh}; \boldsymbol{\mu}'_{is}, \boldsymbol{\Sigma}) = \sum_{i=1}^{2} f_{is} h(\mathbf{x}_{sh}; \boldsymbol{\mu}_{is}, \boldsymbol{\Sigma})$ 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

		Ψ	Ψ'
	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 Ψ' is obtained from Ψ 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 Ψ and Ψ' 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

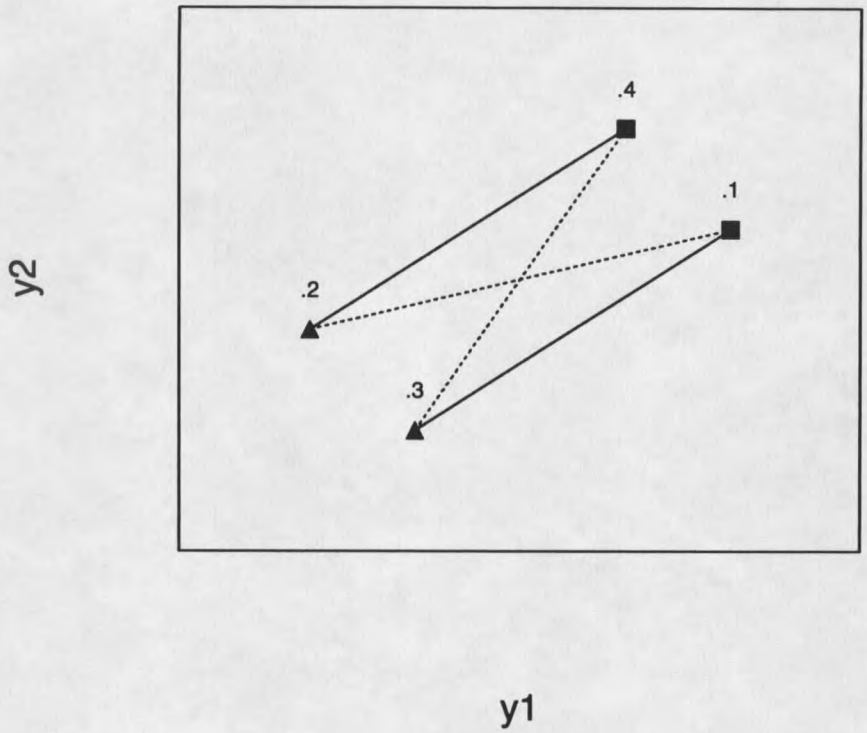

Figure 1: Four cluster means for a hypothetical 2-group, 2-location mixture model. Conditional means at the first location are represented by triangles. Conditional means at the second location are represented by squares. In the unrestricted location mixture model group labels can be assigned in two 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

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

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

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

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 $[\boldsymbol{\mu}_i + \mathbf{B}\mathbf{y}_\mathbf{s}]$. Because the models $[\boldsymbol{\mu}_i]$ and $[\boldsymbol{\mu}_i + \mathbf{B}\mathbf{y}_\mathbf{s}]$ 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 [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$\approx$.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$\approx$.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$\approx$.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$\approx$.02			
$[\boldsymbol{\mu}_i + \mathbf{By}]$	G_1	.34	.18	.17	.31
	G_2	.33	.18	.16	.32
		se$\approx$.10			
$[\boldsymbol{\mu}_i + \boldsymbol{\theta}_s]$	G_1	.34	.18	.17	.31
	G_2	.33	.18	.17	.32
		se$\approx$.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:

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

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

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

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

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

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

The classification rule depends on the parameters p_{is}, $\boldsymbol{\mu}_{is}$ and $\boldsymbol{\Sigma}_{is}$, 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., $\boldsymbol{\Sigma}_{is} = \boldsymbol{\Sigma}$ $\forall i, s$), so that

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

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.,$\mathbf{\Sigma}_{is} = \mathbf{\Sigma}_i$). 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 $\rho_i = |\mathbf{\Sigma}_i|^{1/p}$, $\mathbf{\Gamma}_i$ 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