Linear Clustering with Application to Single Nucleotide Polymorphism Genotyping

by

Guohua Yan

B.Sc., Liaocheng University, 1992
M.Sc., Beijing Normal University, 1995
M.Sc., The University of Windsor, 2003

A THESIS SUBMITTED IN PARTIAL FULFILMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

Doctor of Philosophy

in

The Faculty of Graduate Studies

(Statistics)

The University of British Columbia
(Vancouver)

June, 2008

Abstract

Single nucleotide polymorphisms (SNPs) have been increasingly popular for a wide range of genetic studies. A high-throughput genotyping technologies usually involves a statistical genotype calling algorithm. Most calling algorithms in the literature, using methods such as k-means and mixture-models, rely on elliptical structures of the genotyping data; they may fail when the minor allele homozygous cluster is small or absent, or when the data have extreme tails or linear patterns.

We propose an automatic genotype calling algorithm by further developing a linear grouping algorithm (Van Aelst et al., 2006). The proposed algorithm clusters unnormalized data points around lines as against around centroids. In addition, we associate a quality value, silhouette width, with each DNA sample and a whole plate as well. This algorithm shows promise for genotyping data generated from TaqMan technology (Applied Biosystems). A key feature of the proposed algorithm is that it applies to unnormalized fluorescent signals when the TaqMan SNP assay is used. The algorithm could also be potentially adapted to other fluorescence-based SNP genotyping technologies such as Invader Assay.

Motivated by the SNP genotyping problem, we propose a partial likelihood approach to linear clustering which explores potential linear clusters in a data set. Instead of fully modelling the data, we assume only the signed orthogonal distance from each data point to a hyperplane is normally distributed. Its relationships with several existing clustering methods are discussed. Some existing methods to determine the number of components in a data set are adapted to this linear clustering setting. Several simulated and real data sets are analyzed for comparison and illustration purpose. We also investigate some asymptotic properties of the partial likelihood approach.

A Bayesian version of this methodology is helpful if some clusters are sparse but there is strong prior information about their approximate locations or properties. We propose a Bayesian hierarchical approach which is particularly appropriate for identifying sparse linear clusters. We show that the sparse cluster in SNP genotyping datasets can be successfully identified after a careful specification of the prior distributions.

Table of Contents

List of Tables

List of Figures

Acknowledgements

Foremost I would like to thank my supervisors Dr. William J. Welch and Dr. Ruben H. Zamar, for their excellent guidance, continuous encouragement, great patience and generous support. I am indebted to them for many things I have learned which lay the foundation of my future career.

I am very grateful to other members of my supervisory committee, Dr. Paul Gustafson and Dr. Arnaud Doucet for their valuable comments and suggestions.

I am also very grateful to Dr. Harry Joe and Dr. Lang Wu for their great help and valuable advice both in statistics and in my job application process. Also I would like to thank them for teaching me playing table tennis although they must have been frustrated that I am a hopeless learner.

Many thanks to Dr. Matías Salibián-Barrera, Dr. Raphael Gottardo, Dr. Nancy Heckman and Dr. Jiahua Chen for their help and valuable suggestions. I also thank Dr. Jason Loeppky for his help on my proposal.

Special thanks to Ms. Loubna Akhabir and Ms. Treena McDonald for providing me the data and explaining to me the genomic background.

I would also like to thank Ms. Elaine Salameh, Ms. Christine Graham, Ms. Rhoda Morgan, Ms. Peggy Ng and Ms. Viena Tran for their help with administrative matters.

Furthermore, I thank my fellow students for their help throughout my program. I choose not to list their names here but their help and friendship will be remembered forever.

Finally, I would like to acknowledge the support of the University of British Columbia through a University Graduate Fellowship.

To my parents, my wife and my daughter

Statement of Co-Authorship

This thesis is completed under the supervision of Dr. William J. Welch and Dr. Ruben H. Zamar.

Chapter 2 is co-authored with Dr. William J. Welch, Dr. Ruben H. Zamar, Ms. Loubna Akhabir and Ms. Treena McDonald. I conducted the data analysis and prepared a draft of the manuscript.

Chapter 3 is co-authored with Dr. William J. Welch and Dr. Ruben H. Zamar. My main contribution is the derivation of the partial likelihood. I conducted the data analysis and prepared a draft of the manuscript.

Chapter 4 is co-authored with Dr. William J. Welch and Dr. Ruben H. Zamar. My main contributions are the formulation of the model and the specification of the prior distribution. I conducted the data analysis and prepared the manuscript.

Chapter 5 is co-authored with Dr. William J. Welch and Dr. Ruben H. Zamar. I derived the asymptotic properties and prepared the manuscript.

Chapter 1

Introduction

This thesis work investigates methods to detect linearly shaped clusters in a dataset. It is motivated by a clustering problem in SNP (single nucleotide polymorphism) genotyping and much effort of this thesis has been devoted to automatic genotype calling algorithms in TaqMan SNP genotyping technology. In this chapter, we introduce the SNP genotyping problem, review several clustering algorithms used in the rest of the thesis and give an outline of the thesis.

1.1 SNP genotyping

1.1.1 SNPs and their applications

A single nucleotide polymorphism (SNP, pronounced as "snip") is a single-base variation in a genome. The genetic code is specified by the four nucleotide "letters": A (adenine), C (cytosine), T (thymine) and G (guanine). There are two complimentary DNA strands. It is sufficient to consider only one. SNP variation occurs when a single nucleotide, such as an A, is replaced by one of the other three letters C, G or T at a particular base of the target strand. An example of a SNP is the alteration of the DNA segment AAGGTTA to ATGGTTA, where the second A in the first snippet is replaced with a T (www.ncbi.nlm.nih.gov). A SNP variation usually involves only two nucleotides; the two possible nucleotides are called two SNP alleles. Throughout we shall use "X" to generically represent the common wild-type allele and "Y" the variant allele. The DNA sequences of any two individuals are mostly identical and SNPs are found about every $250 - 350$ base pairs in the human genome (Niu et al., 2002).

Approximately 3 to 5 percent of a person's DNA sequence codes for the production of proteins, most SNPs are found outside of these "coding sequences". SNPs found within a coding sequence, cSNPs, are of particular interest to researchers because they are more likely to alter the biological function of a protein (www.ncbi.nlm.nih.gov). Moreover, the abundance of SNPs makes them useful markers for genetic association studies that work

to localize the genes involved in complex diseases or adverse drug reactions. The popularity of SNPs is also due to their usual biallelic property which makes them amenable to automated genotyping. (Ranade et al., 2001).

1.1.2 TaqMan SNP genotyping technology

Determination of the alleles at a single nucleotide polymorphism site is called genotyping. The nature of large-scale association studies requires rapid, reliable and cost-effective SNP genotyping. Various SNP genotyping technologies have been developed. The choice of a technology depends on whether a few SNPs are to be typed in many individuals or many different SNPs are to be examined in a few individuals (Callegaro et al., 2006).

The TaqMan SNP Genotyping Assay (Applied Biosystems) is a widely used fluorescence-based high-throughput genotyping technology suitable for the former case. In this method, the region flanking the SNP site of interest is amplified in the presence of two probes each specific for one or the other allele. Probes haves a fluor, called "reporter" at one end but do not fluoresce when free in solution because they have a "quencher" at the other end that absorbs fluorescence from the reporter. During the amplification in which many copies of the same sequence are produced, the probe specifically base-paired with the target is unwound, its reporter liberated from the quencher and the fluorescence is increased. The presence of two probes, each labelled with a different fluor, allows the detection of both alleles in a single tube (De La Vega et al., 2005; Ranade et al., 2001).

In the TaqMan SNP genotyping technology, DNA samples of many individuals are arranged in a 96- or 384-well plate and they are amplified simultaneously. For each individual, two quantitative fluorescent signals are measured at the end of amplification for the two alternative SNP alleles, indicating their presence or absence. The pair of signals for an individual forms a point in a scatterplot. Ideally there are four distinct clusters in a scatterplot. The NTC (no template control) cluster lies in the lower-left corner, close to the origin, containing negative control cases (a few cases that do not have DNA samples) and samples that fail to amplify. In the lower-right, upper-left, and upper-right corners are three clusters presumably containing samples of wild-type homozygotes, variant homozygotes, and heterozygotes, respectively.

Genotype calls for individual samples are made by a clustering algorithm in the propriety Sequence Detection Software (Applied Biosystems). However, considerable manual intervention of an expert operator is required to assess the data quality, to set fluorescent signal thresholds and to decide the

genotypes, especially when the variant allele Y is rare.

The accuracy of SNP genotype calls is critical to later studies. Even the slightest amount of genotyping errors can lead to serious consequences on haplotype analysis, linkage analysis, genetics distance estimation and background linkage disequilibrium estimation (Kang et al., 2004). For a brief review of clustering algorithms in fluorescence-based SNP genotyping, see Chapter 2. One aim of this thesis to develop a reliable automated SNP genotype calling algorithm in the TaqMan SNP genotyping technology that involves minimal manual intervention and should also work well even the variant allele homozygous cluster YY is sparse. We conclude this subsection by noting that any genotyping calling algorithm proposed in the TaqMan SNP genotyping technology should be applicable to other fluorescence-based genotyping method, such as Invader Assay (Mein et al., 2000).

In next section, we review several competing clustering algorithms used in the rest of the thesis.

1.2 Review of several clustering algorithms

1.2.1 MCLUST: normal mixture model-based clustering

Finite mixture model has long been proposed for clustering problems. See for example, Banfield and Raftery (1993), Fraley and Raftery (1998), Fraley and Raftery (2002), MacLachlan and Peel (2000). In this approach, data are assumed arising from a mixture of probability distributions; each component of the mixture is regarded as a cluster. This model-based method has some advantage over a heuristic approach. First of all, it is somehow data dependent via a variety of parameter restrictions coupled with a model selection mechanism, for example in the packages MCLUST (Fraley and AE, 2006) and EMMIX (McLachlan et al., 1999). Second, some early proposed heuristic clustering criteria, including the most widely used $k$means method, were later formulated in model frameworks. A statistical model helps people understand for what data sets a particular clustering algorithm is likely to work well. Third, within a model framework, many statistical procedures are readily applicable. For example, maximum likelihood is naturally used as an optimizing criterion; the estimated probabilities of a data point conforming to the components is an appealing measure of uncertainty of the resulting classification; a component in a mixture model, which has a clear meaning in the assumed model, can be interpreted as a cluster; the determination of clustering methods and the number of clusters is then a model selection problem.

MCLUST is a contributed R package for normal mixture modeling and model-based clustering. It provides a number of functions for normal mixture model-bases clustering, classification likelihood-based hierarchical clustering, normal mixture model-based density estimation and discriminant analysis etc. In this subsection, we review only the normal mixture model-based clustering which is relevant to the thesis.

Given observations $\mathbf{x}_1, \ldots, \mathbf{x}_n$, MCLUST assumes a normal mixture model

$$p(\mathbf{x}_i|\boldsymbol{\theta}) = \sum_{k=1}^{K} p_k \phi_k(\mathbf{x}_i|\boldsymbol{\mu}_k, \Sigma_k), \quad i = 1, \ldots, n,$$

where K is the number of components, p_k is the mixing proportion, $p_k > 0$, $\sum_{k=1}^{K} p_k = 1$, $\phi(\cdot|\boldsymbol{\mu}_k, \Sigma_k)$ is the multivariate normal density with mean vector $\boldsymbol{\mu}_k$ and covariance matrix Σ_k and $\boldsymbol{\theta}$ is the collection of all parameters.

Banfield and Raftery (1993) proposed a general framework for geometric cross-cluster constraint by parameterizing covariance matrices through eigenvalue decomposition

$$\Sigma_k = \lambda_k D_k A_k D_k^T,$$

where D_k is the orthogonal matrix of eigenvectors, A_k is a diagonal matrix whose elements are proportional to the eigenvalues, and λ_k is an associated constant of proportionality. The orientation of principal components of Σ_k is determined by D_k, while A_k determines the shape of the density contours; λ_k specifies the volume of the corresponding ellipsoid. Characteristics (orientation, shape and volume) of distributions are usually estimated from the data and can be allowed to vary between clusters or constrained to be the same for all clusters. This parameterization is very flexible; it includes but is not restricted to earlier proposals such as equal-volume spherical variance ($\Sigma_k = \lambda I$) which has a close relationship with the k-means method, constant variance ($\Sigma_k = \Sigma$) and the most general unconstrained variance.

The EM algorithm is used for maximum likelihood estimation; details are omitted. All models corresponding to the above parameterization possibilities are usually tried. Selection of the number K of components/clusters as well as models is through the Bayesian Information Criterion (BIC), which relates to the Bayes factor. The BIC has the form

$$\text{BIC} = 2l - \log(n)v,$$

where l is the log-likelihood of the model and v is the number of independent parameters in the model.

For noisy data, MCLUST adds a first order Poisson process to the normal mixture model,

$$p(\mathbf{x}_i|\boldsymbol{\theta}) = \frac{p_0}{V} + \sum_{k=1}^{K} p_k \phi_k(\mathbf{x}_i|\boldsymbol{\mu}_k, \Sigma_k),$$

where V is the hyper-volume of the data region, $p_0 > 0$ and $\sum_{k=0}^{K} p_k = 1$.

1.2.2 MIXREG: mixture of linear regressions

Mixtures of linear regressions may be regarded as a special case of mixture models. The response of interest is assumed univariate. Let $(y_1, \mathbf{x}_1)$, ..., $(y_n, \mathbf{x}_n)$ be the observations. The mixture of regression models is

$$p(y_i|\boldsymbol{\theta}, \mathbf{x}_i) = \sum_{k=1}^{K} p_k \phi(y_i|\mathbf{x}_i^T \boldsymbol{\beta}_k, \sigma_k^2), \quad i = 1, \ldots, n,$$

where K is the number of components, p_k is a mixing proportion, $p_k > 0$, $\sum_{k=1}^{K} p_k = 1$, $\boldsymbol{\beta}_k$ is the vector of regression coefficients, $\phi(\cdot|\mathbf{x}_i^T \boldsymbol{\beta}_k, \sigma_k^2)$ is the univariate normal density with mean $\mathbf{x}_i^T \boldsymbol{\beta}_k$ and variance σ_k^2. For ease of notation, we assume that an intercept is already included in $\boldsymbol{\beta}_k$ if necessary.

The EM algorithm can be used for maximum likelihood estimation. MIXREG is also a contributed R package (Turner, 2006); the computational details are in Turner (2000).

The package allows for the selection of equal variances ($\sigma_k^2 = \sigma^2$) or unequal variances. Model selection can be performed using BIC outside the package MIXREG.

1.2.3 LGA: linear grouping algorithm

Van Aelst et al. (2006) proposed a linear grouping algorithm to detect linear patterns in a dataset. It combines ideas of k-means, orthogonal regression and resampling and can uncover linear patterns when most traditional algorithms do not work well. LGA is the corresponding contributed R package (Harrington, 2007). Suppose that we are to uncover k linear groups in a data set with n points in d dimensions. The linear grouping algorithm works as follows:

1. *Initialization.* Starting values are generated by randomly selecting k mutually exclusive subsets of d points (d-subsets). For each of these d-subsets, a hyperplane through the d points is computed.

2. *Updating group assignment.* Each point is assigned to its closest hyperplane in the sense of orthogonal distance; each hyperplane is recomputed from the updated grouping using orthogonal regression.

3. *Iterative refinement.* The objective function is set to be the aggregated sum of the squared orthogonal distances of the data points from their closest hyperplane. Repeat step 2 until no significant improvement is attained. A moderate number of iterations (say, 10) is usually sufficient.

4. *Resampling.* Repeat steps $1 - 3$ a number of times (say, 100) and select the solution with the lowest aggregated sum of squared orthogonal distances. This solution is refined further as in step 3 until no improvement is achieved.

The idea of silhouette width (Rousseeuw, 1987) is adapted to the linear grouping scenario to measure the strength of the group assignment. Denote by $s_1(i)$ and $s_2(i)$ the orthogonal distances of point i from its closest and second closest hyperplanes, respectively. The silhouette width for point i is defined as

$$w(i) = 1 - \frac{s_1(i)}{s_2(i)}.$$

The GAP statistic (Tibshirani et al., 2001) is used to determine the number of linear groups in a data set.

1.3 Outline of the thesis

In Chapter 2, we propose a genotype calling algorithm by further adapting the linear grouping algorithm (Van Aelst et al., 2006). A key feature of this algorithm is that it applies to unnormalized fluorescent signals.

Whereas there is no explicit statistical model in Chapter 2, a partial likelihood approach to linear clustering is proposed in Chapter 3. It borrows ideas from normal mixture models and mixtures of regressions and provides an extension to the heuristic linear grouping algorithm. The SNP genotyping problem is revisited in this model framework.

The method in Chapter 3 is more flexible than that in Chapter 2, but still cannot handle sparse (homozygous variant) clusters well. In Chapter 4, we move to a Bayesian hierarchical approach to the SNP genotyping problem. With careful specification of the prior structures, the Bayesian approach is able to handle a sparse variant allele homozygous cluster.

Chapter 5 includes the technical details of the asymptotic results for the partial likelihood approach in Chapter 3.

Chapter 6 summarizes a few possible future research directions.

Bibliography

Banfield, J. and A. Raftery (1993). Model-Based Gaussian and Non-Gaussian Clustering. *Biometrics 49*(3), 803–821.

Callegaro, A., R. Spinelli, L. Beltrame, S. Bicciato, L. Caristina, S. Censuales, G. De Bellis, and C. Battaglia (2006). Algorithm for automatic genotype calling of single nucleotide polymorphisms using the full course of TaqMan real-time data. *Nucleic Acids Research 34*(7), e56.

De La Vega, F., K. Lazaruk, M. Rhodes, and M. Wenz (2005). Assessment of two flexible and compatible SNP genotyping platforms: TaqMan snp genotyping assays and the SNPlex genotyping system. *Mutation research. Fundamental and molecular mechanisms of mutagenesis 573*(1-2), 111–135.

Fraley, C. and R. AE (2006). mclust Version 3 for R: Normal Mixture Modeling and Model-based Clustering. Technical report, Technical Report 504, University of Washington, Department of Statistics.

Fraley, C. and A. Raftery (1998). How Many Clusters? Which Clustering Method? Answers Via Model-Based Cluster Analysis. *The Computer Journal 41*(8), 578.

Fraley, C. and A. Raftery (2002). Model-Based Clustering, Discriminant Analysis, and Density Estimation. *Journal of the American Statistical Association 97*(458), 611–632.

Harrington, J. (2007). *lga: Tools for linear grouping analysis (LGA)*. R package version 1.0-0.

Kang, H., Z. Qin, T. Niu, and J. Liu (2004). Incorporating Genotyping Uncertainty in Haplotype Inference for Single-Nucleotide Polymorphisms. *The American Journal of Human Genetics 74*(3), 495–510.

MacLachlan, G. and D. Peel (2000). *Finite mixture models*. J. Wiley.

McLachlan, G., D. Peel, K. Basford, and P. Adams (1999). The EMMIX software for the fitting of mixtures of normal and t-components. *Journal of Statistical Software 4*(2).

Mein, C., B. Barratt, M. Dunn, T. Siegmund, A. Smith, L. Esposito, S. Nutland, H. Stevens, A. Wilson, M. Phillips, et al. (2000). Evaluation of Single Nucleotide Polymorphism Typing with Invader on PCR Amplicons and Its Automation. *Genome Research 10*(3), 330–343.

Niu, T., Z. Qin, X. Xu, and J. Liu (2002). Bayesian Haplotype Inference for Multiple Linked Single-Nucleotide Polymorphisms. *Am. J. Hum. Genet 70*, 157–169.

Ranade, K., M. Chang, C. Ting, D. Pei, C. Hsiao, M. Olivier, R. Pesich, J. Hebert, Y. Chen, V. Dzau, et al. (2001). High-Throughput Genotyping with Single Nucleotide Polymorphisms. *Genome Research 11*(7), 1262–1268.

Rousseeuw, P. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics 20*, 53–65.

Tibshirani, R., G. Walther, and T. Hastie (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society. Series B (Statistical Methodology) 63*(2), 411–423.

Turner, T. (2000). Estimating the propagation rate of a viral infection of potato plants via mixtures of regressions. *Journal of the Royal Statistical Society Series C(Applied Statistics) 49*(3), 371–384.

Turner, T. (2006). *mixreg: Functions to fit mixtures of regressions.* R package version 0.0-2.

Van Aelst, S., X. Wang, R. Zamar, and R. Zhu (2006). Linear grouping using orthogonal regression. *Computational Statistics and Data Analysis 50*(5), 1287–1312.

Chapter 2

Automatic genotype calling of single nucleotide polymorphisms using a linear grouping algorithm

2.1 Background

2.1.1 TaqMan SNP genotyping

Single nucleotide polymorphisms (SNPs) make up about 90% of all human genetic variations. They occur approximately every 100 to 300 bases along the 3.2-billion-base human genome. They have been investigated as popular biological markers for a wide range of genetic studies. Such studies require reliable and cost-effective high-throughput genotyping technologies. Various SNP genotyping technologies have been developed and commercialized. Shi (2001) provides a review of some SNP genotyping technologies. De La Vega et al. (2005) give an elaborate assessment of two popular SNP genotyping technologies, TaqMan SNP Genotyping Assay and the SNPlex Genotyping System (Applied Biosystems). Kang et al. (2004) give a brief overview of three widely used high-throughput technologies, the TaqMan SNP Genotyping Assay, the OLA (Oligonucleotide Ligation Assay) and the MassARRAY system. Although they employ different mechanisms, the most popular high-throughput technologies share the same conceptual framework: for each DNA sample, two quantitative signal intensities are measured after amplification for two alternative SNP alleles, indicating their presence or absence; these pairs of signal intensities are used to call genotypes by an ex-

A version of this chapter will be submitted for publication. Authors: Guohua Yan, William J. Welch, Ruben H. Zamar, Loubna Akhabir and Treena McDonald.

pert manually or by a clustering algorithm. The two alleles are generically called X and Y hereafter.

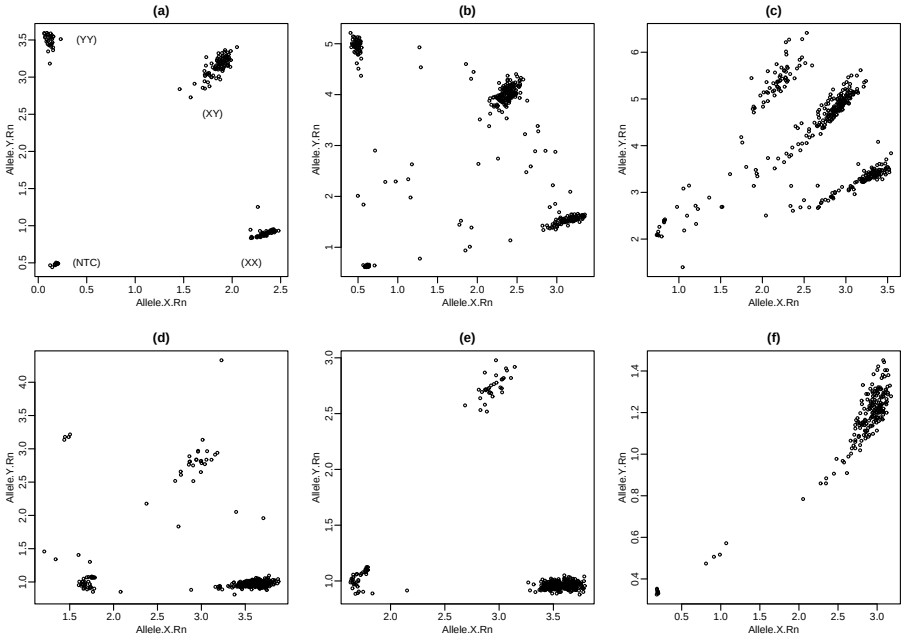

Figure 2.1: Scatterplots of ROX-normalized data for (a): a good plate; (b) & (c): messy plates; (d): a plate with only a few points in the YY cluster; (e): a plate with no points in the YY cluster; (f): a plate with only one cluster.

Figure 2.1 illustrates several scenarios that are typical in genotyping data from the TaqMan SNP Genotyping Assays (Applied Biosystems) at the James Hogg iCAPTURE Centre. In this technology, two alleles are labelled by fluorescent dyes VIC and FAM; in addition, a third dye ROX, which is assumed unchanged during the amplification, is used to normalize the data. The proprietary ABI PRISM® 7900HT Sequence Detection System Plate Utility Software (referred to as SDS system hereafter) actually uses VIC/ROX and FAM/ROX (referred to as ROX-normalized data hereafter) to make genotype calls. In Figure 2.1, ROX-normalized signals are shown.

Ideally, there are four distinct clusters in a scatterplot as in Figure 2.1 (a). The NTC (no template control) cluster lies in the lower-left corner, close to the origin, containing negative control cases (a few cases that do not have DNA samples) and samples that fail to amplify. In the lower-right, upper-left and upper-right corners are three clusters labelled XX, YY

and XY, presumably containing samples of wild-type homozygotes, variant homozygotes, and heterozygotes, respectively. Owing to various artifacts, however, segregation can be poor with points lying between clusters, which make the calls less trustworthy (Figure 2.1 (b) and (c)). Furthermore, in some plates there are only a few points or even no points in the variant allele homozygous cluster YY, which often causes classical clustering algorithms to fail or to make genotype calls incorrectly (Figure 2.1 (d) and (e)). To our knowledge, the proprietary clustering algorithm incorporated in the SDS system cannot make genotyping calls in this situation and one has to call manually. In extreme cases, there is only one cluster present (Figure 2.1 (f)), which usually indicates that something has gone wrong in the amplification procedure.

2.1.2 Review of some genotype calling algorithms

For a small project, it is possible to make genotype calls manually. In most cases, it is not hard for an expert to perform this job, and the "eyeballing" procedure usually gives reasonable results due to its sophisticated incorporation of prior information. For large-scale studies, however, manual scoring can become a daunting challenge. Furthermore, humans are likely to make errors due to fatigue or oversight when a job becomes routine and different readers may have different views (Kang et al., 2004; van den Oord et al., 2003). van den Oord et al. (2003) conducted a study on the error rates of manual scoring and three statistical procedures and report that these statistical procedures uniformly outperform the manual procedure in error rates.

Perhaps the most widely used clustering method in the SNP genotyping literature is the classical k-means, possibly with minor modification (Akula et al., 2002; Olivier et al., 2002; Ranade et al., 2001). Olivier et al. (2002) notice that k-means algorithms often split one genotype group of the data, especially when the variant allele homozygous cluster has only a few data points or when there is no sharp distinction between one of the genotype clusters and the NTC cluster. They develop a new algorithm, the CA algorithm, that initially examines all data points and determines the approximate location of the centroid for each cluster by dividing the space into sections for each expected cluster. They report that the heuristic algorithm works better than classical k-means.

Ranade et al. (2001) and Akula et al. (2002) assign a "quality score" to each sample. After the genotype assignments are made, they assume that each genotype cluster is bivariate normally distributed. For each sample,

they calculate the probability density value of the normal distribution for its genotype cluster as the quality score that this sample is called. Akula et al. (2002) also assume equal covariance matrices across the clusters, arguing that this assumption gives conservative but better results than otherwise. Lovmar et al. (2005) propose using "silhouette scores" to assess the quality of SNP genotype clusters, the idea of which originates from Rousseeuw (1987). They report that the measure is satisfactory and empirically the genotypes can be unequivocally assigned without manual inspection when the silhouette score is greater than 0.65.

van den Oord et al. (2003) propose a mixture model approach, in which each cluster is assumed to follow a normal distribution. The number of component clusters and the initial values of the parameters are set upon inspecting the scatterplot. This is a more delicate approach than the k-means approach since it considers the elliptical structures in the assignments of genotypes and the k-means algorithm can be regarded as a special case of the mixture models approach. Kang et al. (2004) go one step further in this direction. They assume a bivariate t-mixture model, in which a t distribution with small degrees of freedom is assumed for each cluster. They argue that clusters from SNP genotyping data usually have heavier tails than the normal distribution and report that the t-mixture model is less sensitive to outlying points and has advantages over the normal mixture model or the k-means approach. Fujisawa et al. (2004) also use a model-based approach in which only angle data are used.

2.1.3 ROX-normalized versus unnormalized data

The datasets illustrated in Figure 2.1 are after ROX-normalization. An important feature of our approach is that it uses data *without* ROX-normalization. We now discuss four reasons why we prefer to work with the unnormalized data.

First, the motivation for ROX-normalization is presumably to form spherical clusters for which classical clustering algorithms such as k-means can be used. Inspecting the ROX-normalized scatterplots, there are quite a few plates for which the normalization does not produce reasonable spherical structures as expected, e.g., Figure 2.1 (c).

Second, the incorporation of ROX intensities is aiming to correct systematic biases in the chemistry such as plate to plate variation. However, there are scenarios where some points may be pulled away from their home clusters merely by this correction and fail to be called a genotype; by the same token, points which fail to amplify may be assigned a genotype incorrectly

due to this correction. In the left panel of Figure 2.2, points a, b, c are away

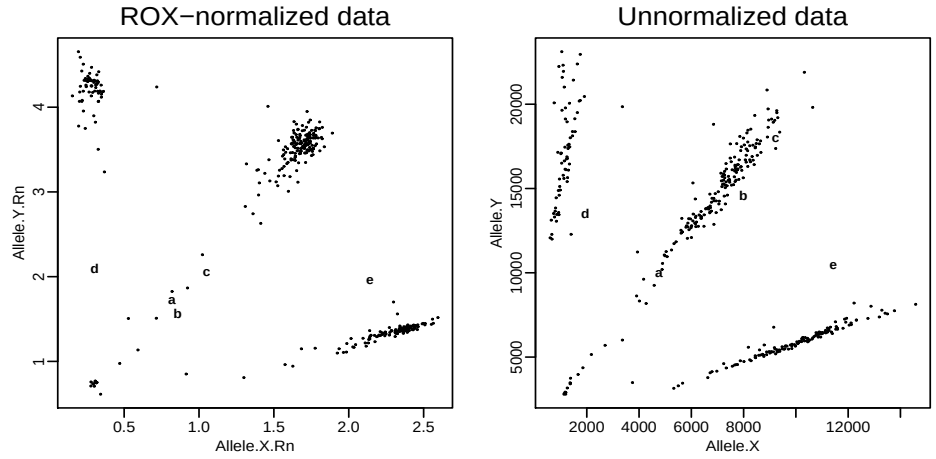

Figure 2.2: Scatterplots of unnormalized and ROX-normalized data for a plate. Both points and letters represent samples.

from the heterozygous cluster when the data are ROX-normalized, but are within a linear arranged cluster in the right panel when the unnormalized data are used. Similarly, point d is away from the variant allele homozygous cluster in the left panel, but is much closer to the variant allele homozygous cluster in the right panel. The proprietary software assigns very low quality values for these points. Intuitively, these points "should" be classified into one of the genotype clusters, or at least with higher "confidence" than their "quality" values indicate. Point e is outlying in both panels; in this case, normalization will not help any way.

Third, the undisclosed, proprietary algorithm used by the SDS system for clustering, and hence calling genotypes, seems to have difficulty with samples with extreme (small or large) ROX values, as evidenced in Figure 2.3. The curve in Figure 2.3 shows that the empirical chance of calling a genotype by the proprietary algorithm decreases when the ROX value is too low or too high.

Unnormalized data, like in Figure 2.2, often show well-separated clusters, but they are along *lines*. The linear grouping algorithm (LGA) (Van Aelst et al., 2006) can identify such clusters. When we apply LGA to the unnormalized data, the problem of low-call rate for extreme ROX values is much alleviated. Figure 2.4 contrasts the silhouette width (a measure of calling quality) versus ROX value when the LGA algorithm is applied to unnormal-

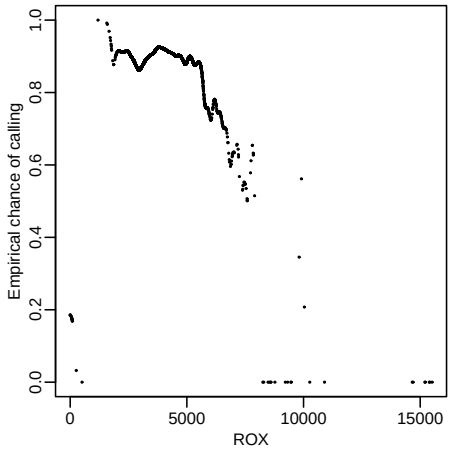

Figure 2.3: Probability of calling a genotype versus ROX

ized data and when the k-means method is used for ROX-normalized data. For k-means, the definition of silhouette width in Rousseeuw (1987) is used; for the LGA, the definition of silhouette width in Van Aelst et al. (2006) is used (see also equation (2.1)).

Last, the positions of positive control points in scatterplots of ROX-normalized data suggest the use of unnormalized data. In a TaqMan SNP genotyping assay, individual samples are usually arranged in a 96- or 384-well plate and are amplified simultaneously. For quality assessment purpose, some wells have DNA samples with known genotypes (in our example 12 in a 384-well plate), called positive controls, while some other wells have no DNA samples (typically 8 in a 384-well plate), called negative controls or no template controls (NTC). The positions of control points in one plate are shown in Figure 2.5. The positive control points are represented by "P". Note that quite a few control points are outside of their corresponding genotype clusters in the left panel where ROX-normalized data are used. In the right panel, however, these points are well within their corresponding linear clusters.

In conclusion, we shall use unnormalized data hereafter in our method of making genotype calls.

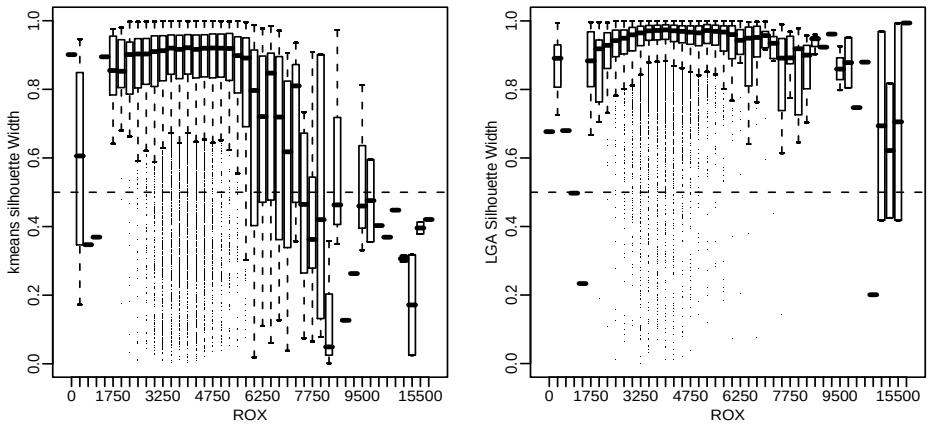

Figure 2.4: Silhouette width versus ROX value. Left panel: k-means results using ROX-normalized data; right panel: LGA result using unnormalized data.

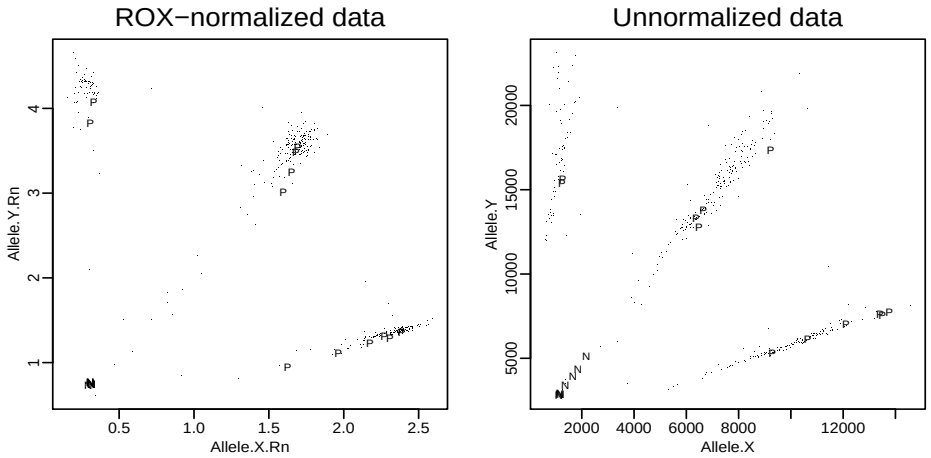

Figure 2.5: Control points in scatterplots of ROX-normalized and unnormalized data for a plate. Letter "P" represents a positive control point and "N" a negative control point.

2.2 Results

As an illustration, we apply the LGA approach to the unnormalized data in Figure 2.2. The silhouette threshold is set to be 0.75, which means a point is called if its distance from the nearest line is one quarter of its distance from the second nearest line. We set the signal threshold empirically. Let m_x and m_y be the median values for X signals ("Allele.X") and Y signals ("Allele.Y") respectively. Any points with X signal less than $0.5m_x$ and Y signal less than $0.5m_y$ are not called. The results are in Figure 2.6. More points are called compared with the SDS software.

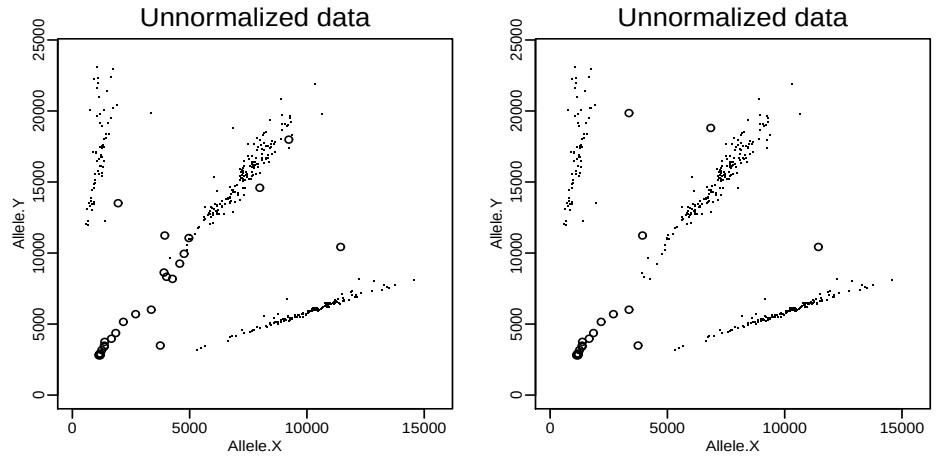

Figure 2.6: Left panel: the calling result of SDS software displayed in the scatterplot of unnormalized data. Right panel: the scoring results of LGA using unnormalized data. The symbol ∘ represents noncalls.

2.3 Conclusions

We have proposed an automatic genotype calling algorithm by taking advantage of a linear grouping algorithm (Van Aelst et al., 2006). The proposed method uses unnormalized signals and clusters points around lines as against centroids. In addition, we associate a quality value, silhouette width (Rousseeuw, 1987; Van Aelst et al., 2006), with each DNA sample and a whole plate as well. This algorithm shows promise for genotyping data from TaqMan technology (Applied Biosystems). The algorithm is reliable.

It could be potentially adapted to other fluorescent-based SNP genotyping technologies as well, such as the Invader Assay.

2.4 Methods

In the proposed calling algorithm, unnormalized data are used and each genotype cluster is represented by a straight line.

2.4.1 Data preprocessing

In a TaqMan SNP genotyping assay, the quality of a plate is monitored by making sure that all positive controls go to the correct genotype clusters and negative controls stay in the NTC clusters. Without well to well contamination, the negative controls should have low signals for both VIC and FAM dyes (horizontal and vertical coordinates in a scatterplot, respectively) since there are no samples in these wells to be amplified. It is advantageous that we empirically discard negative control points and a portion (5% or 10%, say) of points with very low signals (close to the origin) prior to applying LGA to locate the lines. Negative control points should not contribute to the lines which decide the orientations of the genotype clusters. (Some points with low signals are assigned to lines afterwards.)

2.4.2 Fitting lines using LGA

In SNP genotyping, there are at most three clusters, ignoring the NTC cluster. Two clusters are possible when there are very few or no points in the upper-left, YY cluster. We shall fit three lines and two lines to the data separately using LGA. LGA minimizes aggregated sum of squared orthogonal distances of points to their closest hyperplanes. We made some modification specific to the genotyping setting.

Grid based initialization. LGA usually depends on multiple starts to have large probability of attaining the global minimum. In the genotyping setting, it is computationally affordable to try a sequence of initializations such that the global minimum is very likely attained. We set the intercepts to 0 for all lines and choose the slopes from a set of values $\{s_1, \cdots, s_m\}$, where $\text{atan}(s_i)$ are an equally spaced grid from 0 to $\pi/2$, since the slopes should always be positive in the SNP genotyping setting.

Given a set of intercepts and slopes, points are assigned to their nearest lines; these lines are recalculated. These two steps are iterated until convergence. Denote by $s_1(i)$ and $s_2(i)$ the orthogonal distances of point i from its

closest and second closest lines, respectively. The silhouette width for point i is defined as

$$w(i) = 1 - \frac{s_1(i)}{s_2(i)}. \tag{2.1}$$

From the multiple starts, we choose the solution that has the largest average silhouette width. (This criterion respects small clusters and penalizes two lines which are too close.)

In addition, we shall restrict the slope for the XX cluster be smaller than that of the XY cluster and the slope of the XY cluster be smaller than that of the YY cluster. Another restriction is added such that there are few points in YY than XX and XY. For each LGA, we associate the lines with smallest, second-smallest and largest slopes to XX, XY, and YY, respectively. We simply discard solutions in which any slope is negative or there are more points in the cluster with the largest slope than one of the other clusters.

Number of lines. We also choose between the best two line solution and the best three line solution according to this average silhouette width. If the three line solution has large average silhouette width, each line represents a genotype cluster; if the two line solution is chosen, they correspond to XX and XY.

2.4.3 Genotype assignment

For genotype assignment, we could empirically set a signal threshold beforehand, below which a point is not called, labelled as "Undetermined", and assigned to the "NTC" cluster. The threshold can be set based on previous genotyping practice in a laboratory. In this case, a point beyond the threshold is assigned to its closest line and its genotype is called accordingly; a silhouette width value is calculated for each point to measure adequacy of the fitted line structures. A cutoff point for the silhouette widths is also predetermined such that outlying points with low silhouettes are also labelled as "Undetermined".

Our code adapts LGA to deal with thresholding of points as follows. Let

$$m_x = \text{median}(x_i),$$

and

$$m_y = \text{median}(y_i),$$

where x_i and y_i are the X signal and Y signal of sample i respectively. Points with $x_i \leq 0.5m_x$ and $y_i \leq 0.5m_y$ are not called.

As an alternative to an empirical signal threshold, we could penalize points closer to the origin by modifying the definition of silhouette width. For example,

$$w(i) = 1 - \frac{s_1(i) + c}{s_2(i) + c}, \tag{2.2}$$

where the tuning parameter c is also set empirically such that points too close to the origin are not called.

Another version we tried is as follows. Let

$$r_i = \sqrt{x_i^2 + y_i^2},$$

be the distance of point i from the origin. Let

$$m = \min\{r_i\},$$

and

$$M = \text{median}\{r_i\}.$$

Let

$$c_i = \left(\frac{\min(r_i, M) - \frac{m}{2}}{M - \frac{m}{2}} \right)^{\frac{1}{2}}.$$

The adjusted value for silhouette width s_i is

$$s_i^* = c_i s_i. \tag{2.3}$$

This downweights the silhouette score for points with weak x_i and y_i signals.

2.4.4 Quality assessment of DNA samples and plates

The silhouette width for a DNA sample serves as a quality value of the genotype call. In addition, for each plate, the average silhouette width can be computed. An expert may decide upon a threshold such that plates with lower average silhouette width are regarded as unreliable. Meanwhile, the average silhouette width of positive controls in a plate and the average silhouette width of negative controls are also indications of the quality of genotype calls for a specific plate.

2.5 Discussion

The existing methods assume spherical or elliptical clusters. The unnormalized signals tend to fall on lines, however, forcing normalization to try to

produce clusters with elliptical shapes, but the ROX normalization signal may be noisy. We take a quite different approach, identifying the linear clusters associated with the unnormalized signals. As seen in Figure 2.6, fewer uncalled samples results. We call only samples with at least one strong signal. The definition of "strong" will vary from plate to plate. We normalize internally by not calling points having weak signals relative to median signals, rather than externally with respect to ROX.

Bibliography

Akula, N., Y. Chen, K. Hennessy, T. Schulze, G. Singh, and F. McMahon (2002). Utility and accuracy of template-directed dye-terminator incorporation with fluorescence-polarization detection for genotyping single nucleotide polymorphisms. *Biotechniques 32*(5), 1072–1078.

De La Vega, F., K. Lazaruk, M. Rhodes, and M. Wenz (2005). Assessment of two flexible and compatible SNP genotyping platforms: TaqMan snp genotyping assays and the SNPlex genotyping system. *Mutation research. Fundamental and molecular mechanisms of mutagenesis 573*(1-2), 111–135.

Fujisawa, H., S. Eguchi, M. Ushijima, S. Miyata, Y. Miki, T. Muto, and M. Matsuura (2004). Genotyping of single nucleotide polymorphism using model-based clustering. *Bioinformatics 20*, 5.

Kang, H., Z. Qin, T. Niu, and J. Liu (2004). Incorporating Genotyping Uncertainty in Haplotype Inference for Single-Nucleotide Polymorphisms. *The American Journal of Human Genetics 74*(3), 495–510.

Lovmar, L., A. Ahlford, M. Jonsson, and A. Syvänen (2005). Silhouette scores for assessment of SNP genotype clusters. *feedback*.

Olivier, M., L. Chuang, M. Chang, Y. Chen, D. Pei, K. Ranade, A. de Witte, J. Allen, N. Tran, D. Curb, et al. (2002). High-throughput genotyping of single nucleotide polymorphisms using new biplex invader technology. *Nucleic Acids Research 30*(12), e53.

Ranade, K., M. Chang, C. Ting, D. Pei, C. Hsiao, M. Olivier, R. Pesich, J. Hebert, Y. Chen, V. Dzau, et al. (2001). High-Throughput Genotyping with Single Nucleotide Polymorphisms. *Genome Research 11*(7), 1262–1268.

Rousseeuw, P. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics 20*, 53–65.

Shi, M. (2001). Enabling large-scale pharmacogenetic studies by high-throughput mutation detection and genotyping technologies. *Clin Chem 47*(2), 164–172.

Van Aelst, S., X. Wang, R. Zamar, and R. Zhu (2006). Linear grouping using orthogonal regression. *Computational Statistics and Data Analysis 50*(5), 1287–1312.

van den Oord, E., Y. Jiang, B. Riley, K. Kendler, and X. Chen (2003). FD-TDI SNP Scoring by Manual and Statistical Procedures: A Study of Error Rates and Types. *BioTechniques 34*(3), 610–624.

Chapter 3

A partial likelihood approach to linear clustering

3.1 Introduction

In this paper we are concerned with detecting linearly shaped clusters in a data set. This work is motivated by a clustering problem in a SNP (single nucleotide polymorphism) genotyping setting. Data of this type are bivariate, consisting of two signals, and conform mostly to linear clusters. More detailed discussion of these genotyping data appears in Section 3.7.

Van Aelst et al. (2006) proposed a method called the Linear Grouping Algorithm (LGA) for linear clustering problems. For earlier approaches to linear clustering, see the references therein. There is a need for algorithms specialized in linear clustering. First, the usual mixture of normal or t distributions (see for example, Banfield and Raftery, 1993; Fraley and Raftery, 1998, 2002; MacLachlan and Peel, 2000, among others) aims at elliptical structures in a data set; when a data set displays highly linear patterns, especially when these patterns are not well separated from each other, we need a different criterion for detecting linear patterns. Second, most earlier linear clustering approaches assume that a response variable is available supervising the search for linear clusters, see for example, Spath (1982), DeSarbo and Cron (1988) and Turner (2000). Van Aelst et al. (2006) have a simulated data set illustrating that the selection of a response variable is not trivial. In that data set, two linear clusters are obvious in the first two dimensions; the third dimension is merely noise. If the second variable is selected as response, the mixture of regressions method by Spath (1982) successfully recovers the linear structures; it fails when the third variable is used as the response. Chen et al. (2001) proposed a method to detect linear

A version of this chapter has been submitted for publication. Authors: Guohua Yan, William J. Welch and Ruben H. Zamar.

structures one by one without assuming a response variable in the computer vision context.

We propose in this paper a partial likelihood approach and compare it with existing algorithms such as LGA (Harrington, 2007), MCLUST (Fraley and Raftery, 2007) and MIXREG (Turner, 2006). Our approach borrows ideas from finite mixture model-based clustering and from mixture of regressions. Each linear cluster is characterized by a hyperplane; the orthogonal deviation of a data point from that hyperplane is assumed to be normally distributed but the position of the data on the hyperplane is not modelled. With this strategy, all variables are treated symmetrically, as compared with the mixture of regressions approach.

The rest of the paper is organized as follows. In Section 3.2, we describe a partial likelihood-based objective function, leading to a model-based linear clustering algorithm, MLC. In Section 3.3, an EM algorithm to maximize the partial likelihood function is presented. In Section 3.4, we discuss the asymptotic properties of MLC. In Section 3.5, we briefly discuss its relationships with existing algorithms such as MCLUST, LGA and MIXREG. In Section 3.6, we propose several methods for determining the number of linear clusters. In Section 3.7, several real and simulated datasets are used to compare MLC with LGA, MCLUST and MIXREG. Some closing remarks are given in Section 3.8.

3.2 Partial likelihood-based objective function for linear clustering

3.2.1 Partial likelihood for orthogonal regression

First, we formulate a partial likelihood representation for orthogonal regression (see for example, Fuller, 1987). To model a single linear cluster, assume that the vector-valued data $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are independently drawn from a random mechanism represented by a random vector $\mathbf{X}$ in a d-dimensional space. We do not model the distribution of $\mathbf{X}$ except for the deviation from an also unknown hyperplane.

Let $\{\mathbf{x} : \mathbf{a}'\mathbf{x} - b = 0\}$ be a hyperplane. For identifiability purpose, we assume that $\mathbf{a}'\mathbf{a} = 1$ and that the first nonzero element of $\mathbf{a}$ is positive. We assume that the signed orthogonal deviation $\mathbf{a}'\mathbf{X} - b$ from the hyperplane $\{\mathbf{x} : \mathbf{a}'\mathbf{x} - b = 0\}$, is normally distributed, i.e.,

$$\mathbf{a}'\mathbf{X} - b \sim N(0, \sigma^2). \tag{3.1}$$

Since $\mathbf{a}$ and b are unknown, $\mathbf{a}'\mathbf{X} - b$ is also unknown. Nevertheless, we can still estimate $\mathbf{a}$, b and σ^2 through the following partial likelihood function

$$\prod_{i=1}^{n} N(\mathbf{a}'\mathbf{x}_i - b; 0, \sigma^2), \tag{3.2}$$

where $N(\cdot\ ; 0, \sigma^2)$ is the density function of the normal distribution with mean 0 and variance σ^2.

Let $\bar{\mathbf{x}}$ and S be the sample mean and sample covariance matrix of $\mathbf{x}_1, \ldots, \mathbf{x}_n$ respectively. Then the maximum partial likelihood estimate $\hat{\sigma}^2$ of σ^2 is the smallest eigenvalue of S; the maximum partial likelihood estimate $\hat{\mathbf{a}}$ of $\mathbf{a}$ is the (standardized) eigenvector associated with $\hat{\sigma}^2$, which is not necessarily unique. Finally, the maximum partial likelihood estimate $\hat{b}$ of b is $\hat{b} = \hat{\mathbf{a}}'\bar{\mathbf{x}}$. See for example Fuller (1987).

Note that model (3.1) and the partial likelihood (3.2) treat the components of $\mathbf{x}$ symmetrically.

3.2.2 Partial likelihood for linear clustering

Now we assume that the data $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are independently drawn from a more complicated random mechanism, still represented by a random vector $\mathbf{X}$ in a d-dimensional space which we do not model fully. The data now lie around K hyperplanes $\{\mathbf{x} : \mathbf{a}_k'\mathbf{x} = b_k\}$, $k = 1, \ldots, K$. Let $\mathbf{Z} = (Z_1, \ldots, Z_K)'$ be a random vector indicating these hyperplanes, where $Z_k = 1$ with probability p_k for $k = 1, \ldots, K$. Let $\mathbf{p} = (p_1, \ldots, p_K)'$. We assume that, conditional on $Z_k = 1$,

$$\mathbf{a}_k'\mathbf{X} - b_k \sim N(0, \sigma_k^2), \quad k = 1, \ldots, K.$$

Let $\mathbf{z}_1, \ldots, \mathbf{z}_n$ be the corresponding unobservable indicators for the data $\mathbf{x}_1, \ldots, \mathbf{x}_n$. Let $\boldsymbol{\kappa}$ be the collection of component parameters, $\boldsymbol{\kappa} = (\mathbf{a}_1', b_1, \sigma_1^2, \ldots, \mathbf{a}_K', b_K, \sigma_K^2)'$, and $\boldsymbol{\theta} = (\boldsymbol{\kappa}', \mathbf{p}')'$.

When the indicators $\mathbf{z}_1, \ldots, \mathbf{z}_n$ are regarded as unknown parameters, the partial likelihood function for parameters $(\boldsymbol{\theta}', \mathbf{z}_1', \ldots, \mathbf{z}_n')'$ is

$$L(\boldsymbol{\theta}, \mathbf{z}_1, \ldots, \mathbf{z}_n | \mathbf{x}_1, \ldots, \mathbf{x}_n) = \prod_{i=1}^{n} \prod_{k=1}^{K} [p_k N(\mathbf{a}_k'\mathbf{x}_i - b_k; 0, \sigma_k^2)]^{z_{ik}}. \tag{3.3}$$

This is a so-called classification likelihood. See for example Scott and Symons (1971) and Banfield and Raftery (1993) in the context of mixture

models for elliptical clusters. In this approach, the parameters $\boldsymbol{\theta}$ and indicators $\mathbf{z}_1, \ldots, \mathbf{z}_n$ are chosen to maximize (3.3) and data points are classified according to these indicators.

In another view (see for example, Fraley and Raftery, 1998, 2002), the indicators $\mathbf{z}_1, \ldots, \mathbf{z}_n$ are regarded as realizations of random vectors $\mathbf{Z}_1, \ldots, \mathbf{Z}_n$, which in turn are an independent and identically distributed sample from $\mathbf{Z}$. After integrating out the indicators $\mathbf{z}_1, \ldots, \mathbf{z}_n$, the partial likelihood function for $\boldsymbol{\theta}$ is

$$L(\boldsymbol{\theta}|\mathbf{x}_1, \ldots, \mathbf{x}_n) = \prod_{i=1}^{n} \sum_{k=1}^{K} p_k N(\mathbf{a}_k' \mathbf{x}_i - b_k; 0, \sigma_k^2). \tag{3.4}$$

The latter is a mixture likelihood approach. Celeux and Govaert (1993) did simulation studies of classification likelihood approaches and mixture likelihood approaches in normal mixture models for elliptical clusters and reported that no likelihood method uniformly outperforms the others. It is noted that the classification approach cannot consistently estimate parameters due to the "all-or-nothing" classification bias (Bryant and Williamson, 1978; Ganesalingam, 1989; Marriott, 1975). In our experience with linear clustering, both approaches give practically very similar clustering results. We shall pursue in this paper the mixture likelihood (3.4), as many existing model selection criteria like BIC can be used and allows for a more feasible asymptotic theory (see Section 3.4).

As in the usual normal mixture model for elliptical clusters, the partial likelihood function (3.4) is unbounded: when a cluster consists of only points lying on a hyperplane, the contribution of each of these points to the partial likelihood tends to infinity as the variance tends to zero. The infinity occurs on the boundary of the parameter space.

Hathaway (1985) proposed a constrained formulation of the maximum likelihood estimation in the univariate normal mixture model. Specifically, the author added a constraint on the standard deviations,

$$\min_{1 \le i \ne j \le K} (\sigma_i / \sigma_j) \ge c > 0, \tag{3.5}$$

where c is a known constant determined *a priori*. We shall incorporate the constraint (3.5) into the partial likelihood function (3.4) (See Step 2 in the EM algorithm in Section 3.3). The solution may depend on the choice of the constant c. A usual strategy is to decrease c gradually and monitor the resulting solutions (Hathaway, 1986). [Chen and Kalbfleisch (1996); Ciuperca et al. (2003) proposed penalized approaches to this unboundedness problem which are also readily applicable to linear clustering although not

adopted in this paper; on the other hand, the Hathaway's constraint can also be regarded as a form of penalization on the likelihood.]

The partial likelihood function (3.4) naturally brings the clustering problem into a finite mixture model framework. Standard EM algorithms can be adapted to maximize (3.4); once the maximum partial likelihood estimate $\hat{\boldsymbol{\theta}}$ is obtained, data point $\mathbf{x}_i$ can be assigned to the component with the largest posterior probability. The probabilities are given by

$$\hat{w}_{ik} = \frac{\hat{p}_k N(\hat{\mathbf{a}}_k'\mathbf{x}_i - \hat{b}_k; 0, \hat{\sigma}_k^2)}{\sum_{k=1}^K \hat{p}_k N(\hat{\mathbf{a}}_k'\mathbf{x}_i - \hat{b}_k; 0, \hat{\sigma}_k^2)},\ i = 1, \ldots, n;\ k = 1, \ldots, K, \tag{3.6}$$

which also serve as a measure of uncertainty of classifying data point $\mathbf{x}_i$.

3.3 The EM algorithm

Now we describe an EM algorithm for maximizing the partial likelihood (3.4). It is straightforward and works well in our experience.

The completed log partial likelihood is

$$l(\boldsymbol{\theta}|\mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{z}_1, \ldots, \mathbf{z}_n) = \sum_{i=1}^n \sum_{k=1}^K z_{ik}\{\log(p_k) + \log(N(\mathbf{a}_k'\mathbf{x}_i - b_k; 0, \sigma_k^2))\}. \tag{3.7}$$

In the E-step, with $\boldsymbol{\theta}^{(t-1)}$ from the previous iteration, we have

$$\begin{aligned} w_{ik}^{(t)} &\equiv E(Z_{ik}|\boldsymbol{\theta}^{(t-1)}, \mathbf{x}_1, \ldots, \mathbf{x}_n) \\ &= \frac{p_k^{(t-1)} N(\mathbf{a'}_k^{(t-1)}\mathbf{x}_i - b_k^{(t-1)}; 0, \sigma_k^{2(t-1)})}{\sum_{k=1}^K p_k^{(t-1)} N(\mathbf{a'}_k^{(t-1)}\mathbf{x}_i - b_k^{(t-1)}; 0, \sigma_k^{2(t-1)})}. \end{aligned} \tag{3.8}$$

In the M-step, we have

$$p_k^{(t)} = \frac{\sum_{i=1}^n w_{ik}^{(t)}}{n}. \tag{3.9}$$

Let

$$\bar{\mathbf{x}}_k^{(t)} = \frac{\sum_{i=1}^n w_{ik}^{(t)} \mathbf{x}_i}{\sum_{i=1}^n w_{ik}^{(t)}},$$

and

$$\Sigma_k^{(t)} = \sum_{i=1}^n w_{ik}^{(t)} (\mathbf{x}_i - \bar{\mathbf{x}}_{(k)}^{(t)})(\mathbf{x}_i - \bar{\mathbf{x}}_{(k)}^{(t)})'.$$

Then

$\mathbf{a}_k^{(t)}$ is the eigenvector associated with the smallest eigenvalue of $\Sigma_k^{(t)}$, (3.10)

$$b_k^{(t)} = \mathbf{a}'^{(t)}_k \bar{\mathbf{x}}^{(t)}_{(k)}, \tag{3.11}$$

and

$$\sigma_k^{2(t)} = \frac{\sum_{i=1}^n w_{ik}^{(t)} (\mathbf{a}'^{(t)}_k \mathbf{x}_i - b_k^{(t)})^2}{\sum_{i=1}^n w_{ik}^{(t)}}. \tag{3.12}$$

If all σ_k^2 in equation (3.4) are assumed to be equal, then the common value is estimated by

$$\sigma^{2(t)} = \frac{\sum_{i=1}^n \sum_{k=1}^K w_{ik}^{(t)} (\mathbf{a}'^{(t)}_k \mathbf{x}_i - b_k^{(t)})^2}{n}. \tag{3.13}$$

The MLC algorithm is described as follows.

1. Initialize with $\boldsymbol{\theta}^{(0)}$. To initialize the EM algorithm, we adopt the strategy of Van Aelst et al. (2006). We randomly select K mutually exclusive subsets of $d+1$ observations from $\mathbf{x}_1, \ldots, \mathbf{x}_n$. For $k = 1, \ldots, K$, we compute the maximum partial likelihood estimates $\mathbf{a}_k^{(0)}$, $b_k^{(0)}$ and $(\sigma_k^2)^{(0)}$ from the kth subset of $d+1$ observations (see Subsection 3.2.1). The proportions $p_k^{(0)}$ are all set as $1/K$. The initial values are then $\boldsymbol{\theta}^{(0)} = ((\mathbf{a}'_1)^{(0)}, b_1^{(0)}, (\sigma_1^2)^{(0)}, \ldots, (\mathbf{a}'_K)^{(0)}, b_K^{(0)}, (\sigma_K^2)^{(0)}, p_1^{(0)}, \ldots, p_K^{(0)})'$.

2. If constraint (3.5) is not satisfied, go back to step 1; otherwise go to step 3.

3. Update $\boldsymbol{\theta}^{(0)}$ for a predefined number of iterations or until the improvement in the partial likelihood (3.7) is less than a predefined threshold. At iteration t,

   - E-step. Update $w_{ik}^{(t)}$ by equation (3.8).
   - M-step. Update $p_k^{(t)}$, $a_k^{(t)}$, $b_k^{(t)}$ and $(\sigma_k^2)^{(t)}$ by equations (3.9), (3.10), (3.11) and (3.12), respectively. If $c = 1$, which corresponds to equal variances across clusters, use (3.13) in place of (3.12).

4. Repeat steps 1 – 3 for a predefined number of times. The final cluster labels $\hat{\mathbf{z}}_1, \ldots, \hat{\mathbf{z}}_n$ and parameter estimates $\hat{\boldsymbol{\theta}}$ are the solution that has the largest completed log partial likelihood (3.4) and satisfies constraint (3.5).

For the classification likelihood (3.3), we can use a so-called CEM (classification EM) algorithm. A C-step is inserted between the E-step and M-step of step 3. In the C-step, $\mathbf{x}_i$ is assigned to the cluster with the largest $w_{ik}^{(t)}$, i.e., the maximum $w_{ik}^{(t)}$ is replaced by the value 1 and all other $w_{ik}^{(t)}$ take the value 0. If the maximum value of $w_{ik}^{(t)}$ is not unique, choose one of the tying clusters at random.

3.4 Asymptotic properties

Let

$$f(\mathbf{x};\boldsymbol{\theta}) = \sum_{k=1}^{K} p_k N(\mathbf{a}_k'\mathbf{x} - b_k; 0, \sigma_k^2),$$

which is the contribution of one observation to the partial likelihood (3.4). The constrained parameter space is

$$\Theta_c = \left\{ \begin{array}{l} \boldsymbol{\theta} = (\mathbf{a}_1, b_1, \sigma_1^2, \ldots, \mathbf{a}_K, b_K, \sigma_K^2, p_1, \ldots, p_K) : \\ \mathbf{a}_k'\mathbf{a}_k = 1, -\infty < b_k < \infty, \sigma_k > 0, k = 1, \ldots, K. \\ \min_{i,j} \sigma_i/\sigma_j \geq c > 0, 0 < p_k < 1, \sum_{k=1}^{K} p_k = 1. \end{array} \right\}.$$

Assuming that the data arise from a probability distribution measure P, we have the following results.

Theorem 3.1 *Let P be an absolutely continuous probability measure with finite second moments. Let*

$$g(\boldsymbol{\theta}) \equiv \int \log f(\mathbf{x};\boldsymbol{\theta}) dP(\mathbf{x}).$$

Then the supremum of g over Θ_c is finite and attainable.

Since the function $f(\cdot, \boldsymbol{\theta})$ is not a density function, this theorem and the following ones cannot be proved by verifying some regularity conditions. The proofs find their roots in Wald (1949), Redner (1981), Hathaway (1983, 1985) and García-Escudero et al. (2007) and are available on request.

Theorem 3.2 *Let P be an absolutely continuous probability measure with finite second moments. Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be a random sample from P. Then a global maximizer of the partial likelihood function $L(\boldsymbol{\theta} | \mathbf{X}_1, \ldots, \mathbf{X}_n)$ in (3.4) over Θ_c exists almost surely if $n \geq Kd + 1$.*

Note that points in Θ_c are not identifiable for $f(\mathbf{x};\cdot\ )$. The function $f(\mathbf{x};\cdot\ )$ remains the same if we permute the labels 1, ..., K; p_{k_1} and p_{k_2} are not identifiable if $(\mathbf{a}_{k_1}, b_{k_1}, \sigma^2_{k_1}) = (\mathbf{a}_{k_2}, b_{k_2}, \sigma^2_{k_2})$. Thus the consistency result is in a quotient topology space. Let $\sim$ be an equivalent relation on Θ_c such that $\boldsymbol{\theta}_1 \sim \boldsymbol{\theta}_2$ if and only if $f(\mathbf{x};\boldsymbol{\theta}_1) = f(\mathbf{x};\boldsymbol{\theta}_2)$ almost surely in P. Denote by Θ_c^q the quotient topological space consisting of all equivalent classes of $\sim$. For a point $\boldsymbol{\theta}_0$ that maximizes $g(\boldsymbol{\theta}) = \int \log f(\mathbf{x};\boldsymbol{\theta}) dP(\mathbf{x})$, its equivalent class in Θ_c^q is denoted by $\boldsymbol{\theta}_0^q$.

Theorem 3.3 (Consistency). *Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be a sample from an absolutely continuous probability measure P with finite second moments. Let $\hat{\boldsymbol{\theta}}^{(n)}$ be a global maximizer of the partial likelihood function $L(\boldsymbol{\theta}|\mathbf{X}_1, \ldots, \mathbf{X}_n)$ in (3.4) over Θ_c. Then $\hat{\boldsymbol{\theta}}^{(n)} \to \boldsymbol{\theta}_0^q$ almost surely in the topological space Θ_c^q.*

Let

$$v_1(\boldsymbol{\theta}) = \mathrm{E}\left\{ \left(\frac{\partial \log f(\mathbf{X};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right) \left(\frac{\partial \log f(\mathbf{X};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right)' \right\},$$

and

$$v_2(\boldsymbol{\theta}) = \mathrm{E}\left\{ \frac{\partial^2 \log f(\mathbf{X};\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \right\},$$

where the expectations are taken with respect to P.

Theorem 3.4 (Asymptotic normality). *Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be a sample from an absolutely continuous probability measure P with finite sixth moments. Let $\hat{\boldsymbol{\theta}}^{(n)}$ be a subsequence of global maximizers of the partial likelihood function $L(\boldsymbol{\theta}|\mathbf{X}_1, \ldots, \mathbf{X}_n)$ in (3.4) over Θ_c, which tends to an interior point $\boldsymbol{\theta}_0$. Then*

$$\sqrt{n}(\hat{\boldsymbol{\theta}}^{(n)} - \boldsymbol{\theta}_0) \xrightarrow{L} N(0, v(\boldsymbol{\theta}_0)),$$

where $v(\boldsymbol{\theta}_0) = [v_2(\boldsymbol{\theta}_0)]^+ v_1(\boldsymbol{\theta}_0) [v_2(\boldsymbol{\theta}_0)]^+$ and A^+ is the Moore-Penrose inverse of A.

In equation (3.6), $\hat{w}_{ik}$ is a function of $\mathbf{x}_i$ and $\hat{\boldsymbol{\theta}}$. Denote $\hat{w}_{ik} = h_k(\mathbf{x}_i, \hat{\boldsymbol{\theta}})$, $k = 1, \ldots, K$. By the Delta method, we have

Corollary 3.1 *Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be a sample from an absolutely continuous probability measure P with finite sixth moments. Let $\hat{\boldsymbol{\theta}}^{(n)}$ be a subsequence of global maximizers of the partial likelihood function $L(\boldsymbol{\theta}|\mathbf{X}_1, \ldots, \mathbf{X}_n)$ in*

(3.4) over Θ_c, which tends to an interior point $\boldsymbol{\theta}_0$. Let $\mathbf{x}$ be a data point. Then for $k = 1, \ldots, K$,

$$\sqrt{n}(h_k(\mathbf{x}, \hat{\boldsymbol{\theta}}^{(n)}) - h_k(\mathbf{x}, \boldsymbol{\theta}_0)) \xrightarrow{L} N(0, [h_k^{(0)}(\mathbf{x}, \boldsymbol{\theta}_0)]' v(\boldsymbol{\theta}_0)[h_k^{(0)}(\mathbf{x}, \boldsymbol{\theta}_0)]),$$

where

$$h_k^{(0)}(\mathbf{x}, \boldsymbol{\theta}_0) = \frac{\partial h_k(\mathbf{x}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \Bigg| \boldsymbol{\theta} = \boldsymbol{\theta}_0 \quad .$$

Using this corollary, we can build approximate confidence intervals for w_{ik} by replacing $\boldsymbol{\theta}_0$ with $\hat{\boldsymbol{\theta}}$ and hence evaluate the clustering of a data point.

3.5 Relationships with other clustering methods

3.5.1 With LGA

In the LGA of Van Aelst et al. (2006), the objective function is the aggregated sum of squared orthogonal distances of the data points to their closest hyperplanes. Using our notation, LGA minimizes

$$d(\boldsymbol{\kappa}, \mathbf{z}_{1:n}) = \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik}(\mathbf{a}_k' \mathbf{x}_i - b_k)^2. \tag{3.14}$$

In the classification likelihood (3.3), if we assume that the variances σ_k^2 are equal across all components (or equivalently, $c = 1$ in (3.5)) and the mixing proportions p_k are equal, the log completed partial likelihood is

$$l(\boldsymbol{\kappa}, \mathbf{z}_{1:n}) = \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \log[N(\mathbf{a}_k' \mathbf{x}_i - b_k; 0, \sigma^2)]. \tag{3.15}$$

It is straightforward to check that the minimization of (3.14) and the maximization of (3.15) are equivalent. Hence the iterative procedure in LGA coincides with the CEM algorithm, which is discussed in Section 3.3, for maximizing (3.15).

In this sense, the proposed partial likelihood approach, or the classification partial likelihood (3.3), is an extension and a model for LGA. The proposed approach permits one dispersion parameter σ_k^2 for each cluster while LGA implicitly uses one dispersion parameter for all clusters. Furthermore, in our model framework, we are able to use the membership probability w_{ik} (see (3.6)) to assess the uncertainty of clustering a data point while an ad hoc measure, silhouette width, is used in LGA.

3.5.2 With normal mixture models

In the M-step in Section 3.3, we can see that the proposed approach is closely related to normal mixture models. The difference resides in the E-step: the proposed approach weighs distances to hyperplanes while the latter compares distances to cluster centres.

In the use of normal mixture models one is forced to estimate the means and also the covariance matrix of each component. Banfield and Raftery (1993) reparameterized the covariance matrices and lead to a variety of parsimonious covariance matrix structures. The unrestricted covariance matrix requires $K(d + d(d + 1)/2)$ parameters and the most parsimonious model, equal spherical covariance matrices, requires only $Kd + 1$ parameters. The proposed approach can be regarded as a parsimonious simplification of another type which is appropriate for highly linear data sets. It requires $d + 1$ parameters per cluster, i.e., d first order parameters for the location of a hyperplane and one second order parameter for the dispersion around the hyperplane.

3.5.3 With mixture of ordinary regressions

As mentioned in the introduction, a mixture of ordinary regressions approach requires a response variable for guidance. A response variable suitable for clustering purpose may not exist. Furthermore, due to the "regression to the mean" phenomenon, the regression hyperplanes are different when different variables are selected as the response. The proposed approach treats each variable symmetrically, and therefore is more suitable in a clustering setting.

3.6 Choosing the number of linear clusters

In some problems, the number of linear clusters is obvious from subject-matter knowledge; in other problems, it has to be estimated from the data. Too many components exploit the randomness of the data and make it hard to interpret the clustering; too few clusters lose information and may be misleading.

3.6.1 Bootstrapping the partial likelihood ratio

A natural approach is to apply the partial likelihood ratio test to a sequences of hypotheses $K = K_0$ against $K = K_0 + 1$, starting from $K_0 = 1$. However,

$-2\ln\lambda$, where λ is the likelihood ratio, does not have a usual asymptotic χ^2 distribution under the null hypothesis. The reason is that the regularity conditions are violated as the parameter space under the null hypothesis is on the edge of the parameter space under the alternative hypothesis when the former is embedded into the latter. Much theoretic research has been done on likelihood ratios for mixture problems (Chen and Chen, 2001a,c,b; Chen et al., 2001; Chen, 1998; Chen and Kalbfleisch, 2005; Everitt and Hand, 1981; Thode Jr et al., 1988, among others).

We adapt the bootstrap approach in McLachlan (1987) to our setting. The log likelihood ratio statistic $-2\log\lambda$ for the test of the null hypothesis of $K = K_1$ versus the alternative $K = K_2$ can be bootstrapped as follows. Let $\hat{\boldsymbol{\theta}}$ be the maximum partial likelihood estimate of all parameters from the original sample under the null hypothesis. For $i = 1, \ldots, n$, sample $\mathbf{z}_i$ from the multinomial distribution with probabilities ($\hat{p}_1$, ..., $\hat{p}_{K_1}$); if $z_{ik} = 1$, sample orthogonal distance e_i from $N(0, \hat{\sigma}_k^2)$. The ith data point in the bootstrap sample is $\mathbf{x}_i + (e_i - \hat{b}_k)\hat{\mathbf{a}}_k$. Suppose B copies of n samples are generated. The value of $-2\log\lambda_i$ is computed after fitting mixture models for $K = K_1$ and $K = K_2$, $i = 1, \ldots, B$. The significance level of the test is obtained by comparing $-2\log\lambda$ with the empirical distribution of $-2\log\lambda_1, \ldots, -2\log\lambda_B$.

3.6.2 Information criteria

The partial likelihood approach enables us to use an approximate Bayes factor to determine the number of linear clusters. Banfield and Raftery (1993) used a heuristically derived approximation to twice the log Bayes factor, called AWE (approximate weight of evidence) to determine the number of clusters. Fraley and Raftery (2002) used BIC (Bayesian information criterion) as a more reliable approximation to twice the log Bayes factor,

$$\mathrm{BIC}(K) = 2L(K) - v_K \log(n), \tag{3.16}$$

where $L(K)$ is the log partial likelihood of a mixture of K components and v_K is the number of independent parameters to be estimated in a K component mixture. As stated in Fraley and Raftery (2002), although the regularity conditions that underly the approximation are not satisfied in finite mixture models, there are some theoretical and practical support of its use for determining the number of clusters. Keribin (2000) showed that the estimator of the number of clusters based on BIC is consistent; Fraley and Raftery (1998, 2002) included a range of applications in which estimating the number of clusters based on BIC has given good results.

A large number of criteria have been proposed in the literature, including NEC (Biernacki et al., 1999; Celeux and Soromenho, 1996), ICL (Biernacki et al., 2000), and cross validation likelihood (Smyth, 2000). The performance of some of these criteria were compared by Biernacki et al. (1999).

As we are adopting a partial likelihood approach, in principle almost all methods proposed in a model framework are approximately applicable although we do not fully model the data. As pointed out by Everitt et al. (2001), "it is advisable not to depend on a single rule for selecting the number of groups, but to synthesize the results of several techniques". In our experience, though, all the methods mentioned above work well if there do exist obvious linear patterns in a data set.

3.7 Examples

3.7.1 Simulated data I

One data set of size 300 in two dimensions is simulated to illustrate the different behaviours of MCLUST, LGA and the proposed MLC. The x_{i1} are uniformly sampled from the interval $(0, 15)$; for the first 250 points, $x_{i2} = x_{i1} + 3\epsilon_i$ and for the last 50 points, $x_{i2} = 8 + 1.5x_{i1} + \epsilon_i$, where ϵ_i are independent and identically standard normal distributed. The data are shown in the top-left panel of Figure 3.1.

For this data set, MCLUST chooses a model of three clusters with different volumes, different shapes and different orientations by the BIC criterion. LGA favours one cluster using the gap statistic. MLC selects two clusters by BIC, ICL, NEC and the bootstrapping likelihood ratio test (see Table 3.1). We set the threshold $c = 0.05$. The result of MIXREG, which is omitted, is similar to that of MLC as we generate the data in favour of it. Figure 3.1 displays the clustering results from MCLUST, LGA and MLC when the numbers of clusters are all set as 2.

This simulated data helps us understand how these algorithms work. MCLUST tries to find elliptical structures, but the two linear clusters have data distributions along the lines not well modelled by normal distributions. For LGA, when two clusters are roughly parallel and touching, it imposes a structure to the data set and tends to partition the data set into bands with equal width.

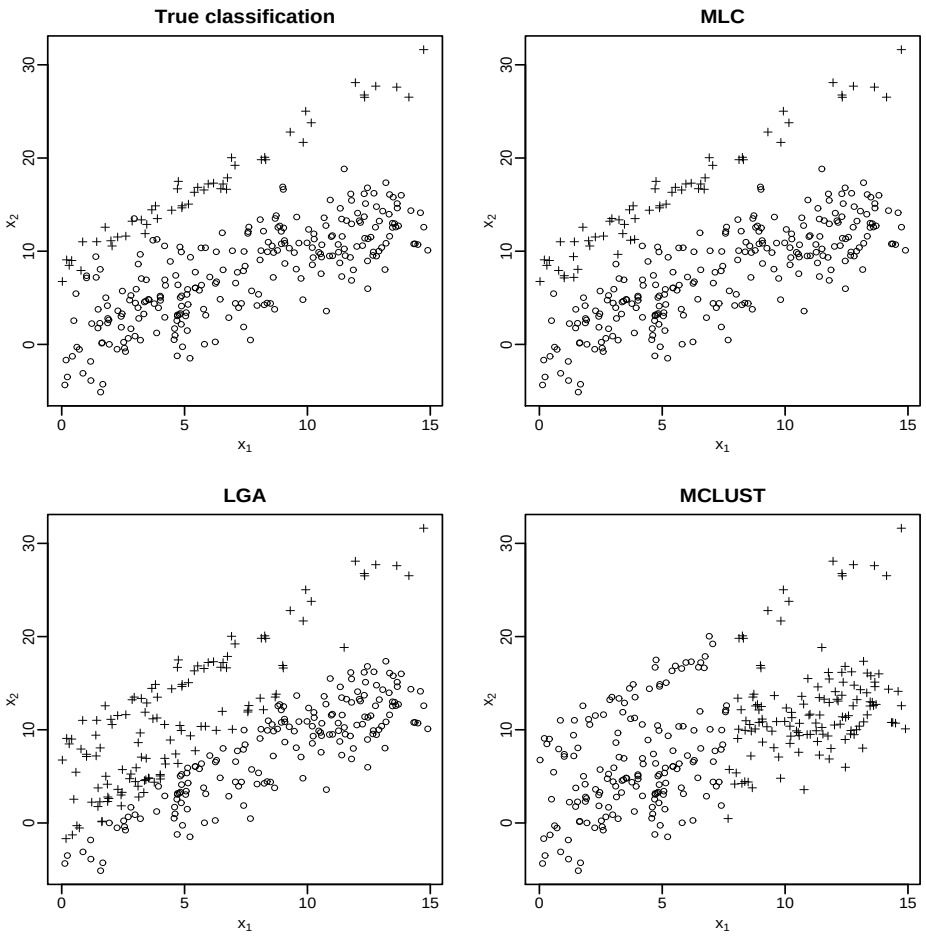

Figure 3.1: Comparison of clustering results of simulated data I. Upper-left panel displays the true classification; upper-right, lower-left and lower-right are that of MLC, LGA and MCLUST.

Table 3.1: Criteria for choosing the numbers of clusters, K, when MLC is applied to simulated data I. "Boot p-value" is the p-value using the bootstrapping partial likelihood ratio test for $H_0 : K = K_0$ vs $H_1 : K = K_0 + 1$ where 99 bootstrapping samples are used

	Criterion			
K	BIC	ICL	NEC	Boot p-value
1	-1541.61	–	–	0
2	**-1461.63**	**-1477.28**	**0.14**	**1**
3	-1468.25	-1712.08	0.69	–

3.7.2 Simulated data II

Another dataset is simulated to demonstrate the difference between MIXREG and the other three algorithms, MLC, MCLUST and LGA. This dataset has 200 data points in four dimensional space. For the first 100 data points, the first two components, x_{i1} and x_{i2}, are linearly related while the third and the fourth components, x_{i3} and x_{i4}, are randomly scattered:

$$x_{i1} = t_i + \epsilon_{i1}, \;\; x_{i2} = 0.5t_i + \epsilon_{i2}, \;\; x_{i3}, x_{i4} \sim N(0, 5^2), \text{ for } i = 1, \ldots, 100. \tag{3.17}$$

Here $\epsilon_{ij} \sim N(0, 1)$ and $t_i \sim N(0, 5^2)$. For the remaining 100 data points, the first two components are randomly scattered while the last two components are linearly related:

$$x_{i1}, x_{i2} \sim N(0, 5^2), \;\; x_{i3} = t_i + \epsilon_{i3}, \;\; x_{i4} = 0.5t_i + \epsilon_{i4}, \text{ for } i = 101, \ldots, 200. \tag{3.18}$$

Again, $\epsilon_{ij} \sim N(0, 1)$ and $t_i \sim N(0, 5^2)$. The pairwise scatterplots of the simulated data are in Figure 3.2.

From the way the dataset is simulated, it is not possible to find a natural response variable to discriminate the two linear clusters. As a result, MIXREG does not work well for this dataset and the two clusters are not well separated. The misclassification matrices of the four algorithms are in Table 3.2. MLC, MCLUST and LGA have the same good performance here. Note that (3.17) and (3.18) generate (elongated) multivariate normal clusters in x_1 and x_2 or in x_3 and x_4, respectively. The assumptions of MCLUST are therefore satisfied, but MLC nonetheless has the same performance.

3.7.3 Australia rock crab data

Now we consider the rock crab data set of Campbell and Mahon (1974) on the genus *Leptograpsus*. We focus on the 100 blue crabs, 50 of which are

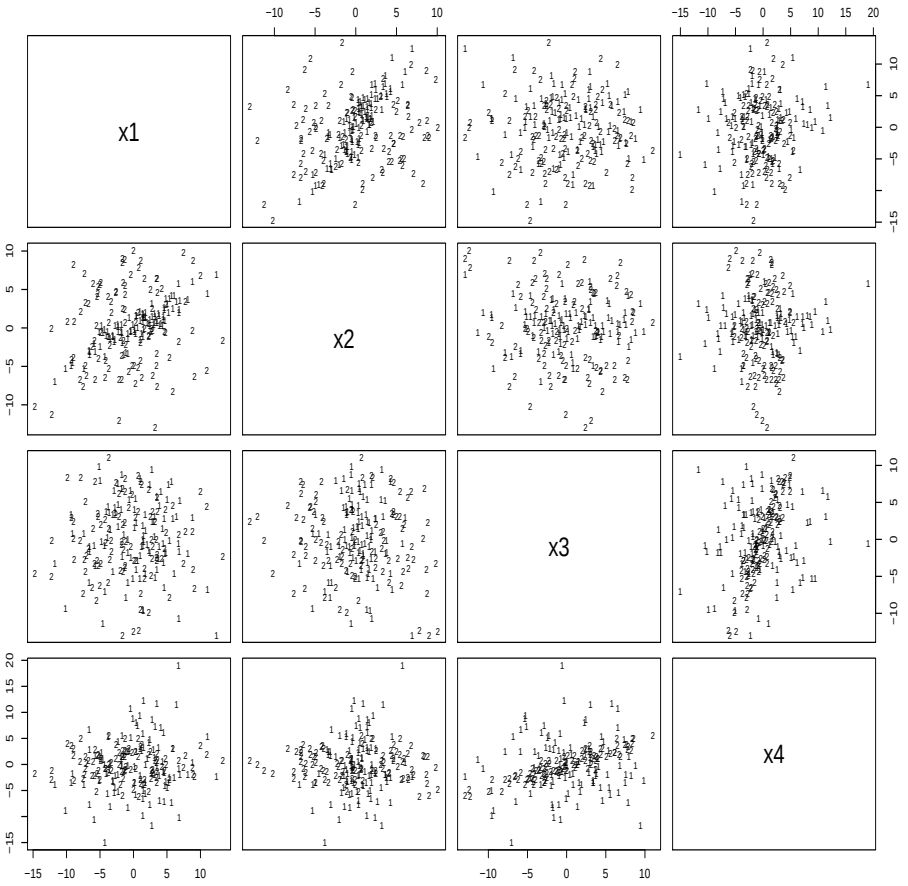

Figure 3.2: Pairwise scatterplots of simulated data II.

Table 3.2: Misclassification matrices of simulated data II from MLC, MCLUST, LGA and MIXREG.

	MLC, MCLUST, LGA		MIXREG					
			Resp= x_1 or x_2		Resp= x_3		Resp= x_4	
True class	1	2	1	2	1	2	1	2
1	88	12	26	74	5	95	63	37
2	9	91	25	75	10	90	75	25

Table 3.3: Criteria for choosing the number of clusters, K, when MLC is applied to the blue crab data. "Boot p-value" is the p-value using the bootstrapping partial likelihood ratio test for $H_0 : K = K_0$ vs $H_1 : K = K_0 + 1$ where 99 bootstrapping samples are used.

	Criterion			
K	BIC	ICL	NEC	Boot p-value
1	476.13	–	–	0
2	**518.88**	**500.77**	**0.24**	**0.63**
3	506.35	446.83	0.51	–

males and 50 are females. Each crab has five measurements, FL, the width of frontal lip, RW, the rear width, CL, the length along the midline, CW, the maximum width of the carapace, and BD, the body depth in mm. As in Peel and McLachlan (2000), we are interested in the classification of male and female crabs. Peel and McLachlan (2000) had 18 crabs misclassified applying a mixture of two t distributions using all the five measurements. Using a mixture of two normal distributions resulted in one more crab being misclassified.

Inspecting all pairwise scatter plots of the data set, RW and CL display two fairly well separated lines as in Figure 3.3. We apply MLC, MCLUST, LGA and MIXREG to these two measurements.

For this data set, MCLUST chooses two clusters with different volumes, different shapes and different orientations by the BIC criterion. LGA chooses two clusters by the gap statistic. MLC also chooses two clusters by the various criteria in Table 3.3. MCLUST has 13 cases misclassified. LGA and MLC both have 7 cases misclassified. Inspecting the posterior probabilities of classification, MLC assigns roughly equal probabilities to the two clusters for two of the misclassified cases. In other words, there is extreme uncertainty about the assignment of these two cases, and their misclassifications are not surprising. There are no such diagnostic probabilities for LGA. MIXREG has eight cases misclassified when RW is regarded as the response; it has seven cases misclassified if CL is taken as the response.

If we take logarithms of RW and CL, only five crabs are misclassified by LGA and MLC. MIXREG has five cases misclassified, if CL is taken as the response; it has 15 cases if RW is taken as the response. The results are summarized in Table 3.4. The clustering results are displayed in Figure 3.4.

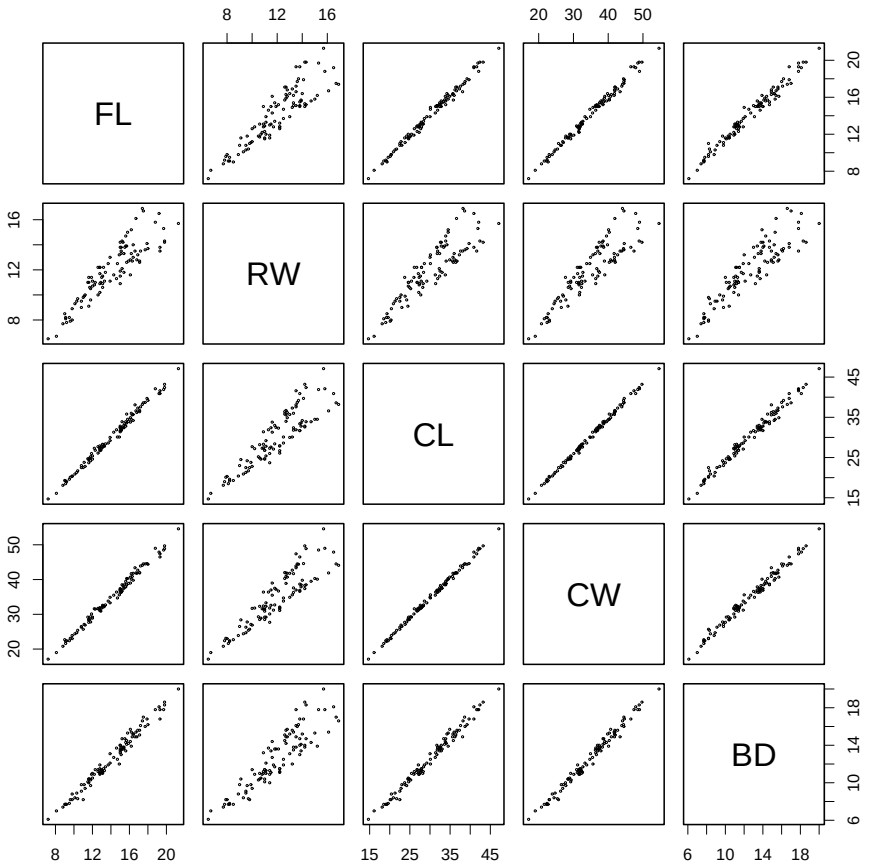

Figure 3.3: Pairwise scatterplots of the blue crab data.

Table 3.4: Misclassification matrices of the blue crab data (using log scales) from MLC, MCLUST, LGA and MIXREG.

	MLC		MCLUST		LGA		MIXREG (Response: CL)		MIXREG (Response: RW)	
True class	1	2	1	2	1	2	1	2	1	2
1	50	0	50	0	50	0	50	0	38	12
2	5	45	14	36	5	45	5	45	3	47

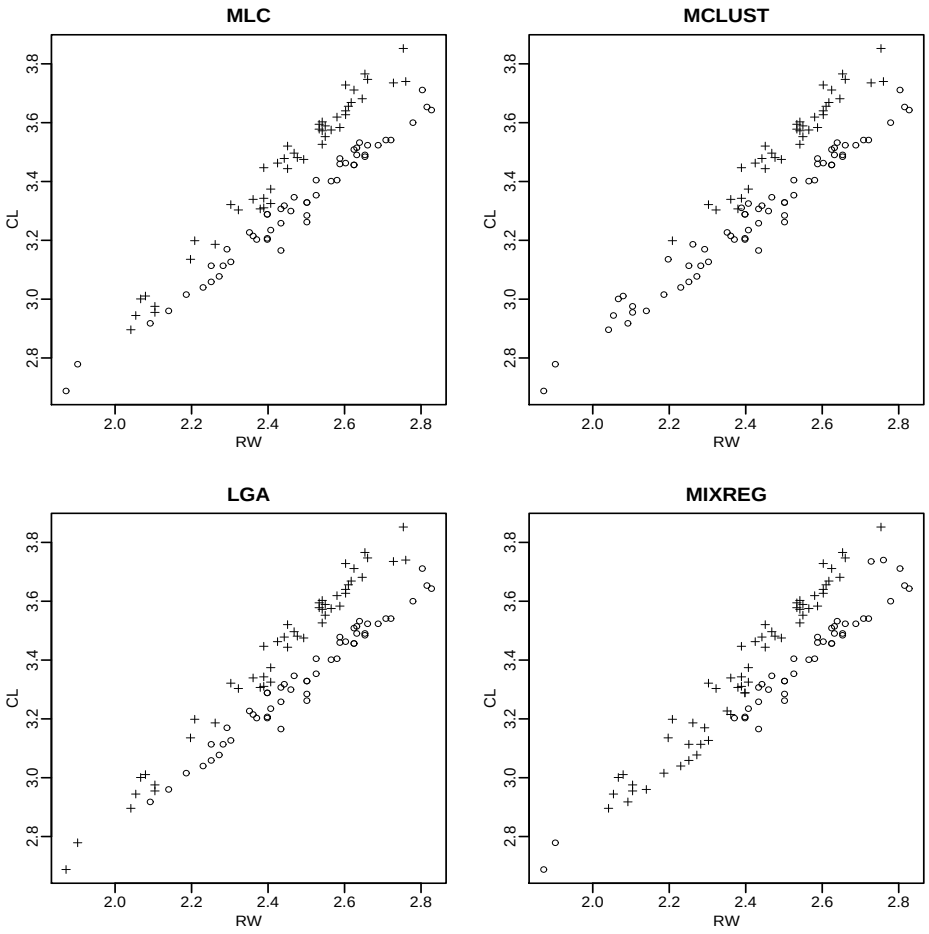

Figure 3.4: Clustering results of the blue crab data (using log scales) from MLC, MCLUST, LGA and MIXREG. CL is used as the response in MIXREG.

3.7.4 Taqman single nucleotide polymorphism genotyping data

A single nucleotide polymorphism (SNP) is a single-base variation in a genome. The genetic code is specified by the four nucleotide "letters": A (adenine), C (cytosine), T (thymine) and G (guanine). A SNP involves usually only two possibilities of these four letters, referred to as allele X and allele Y. An individual has two copies of genetic information passed from both parents. So the genotypes at this specific genetic position has three possible types: homozygote for allele X, homozygote for allele Y and heterozygote for both alleles. SNP genotyping, determination of genotypes of individuals, plays an increasing role in genetics studies.

Taqman assay is a fluorescence-based high-throughput genotyping technology. Two fluorescent signals, x_1 and x_2, called Allele.X and Allele.Y below, are used to detect the presence or absence of the two alleles, and their intensities determine the genotypes. (There is a third fluorescent signal presumably for normalizing the signal intensities. However, its effectiveness is suspicious in some situations. We choose to use signals without this normalization.)

Blood samples of many individuals are assayed in wells of a plate simultaneously. For quality controls purpose, there are usually some blood samples with known genotypes, called positive controls, and there are some wells without genetic material, called negative controls.

In the scatterplot of a SNP data set, there are typically four clusters: three clusters for the three possible genotypes and one for negative controls. In addition, there may be some failed samples. A clustering algorithm is usually employed to classify the samples. Algorithms commonly used include k-means algorithms (Olivier et al., 2002; Ranade et al., 2001) and model-based algorithms (Fujisawa et al., 2004; Kang et al., 2004). Finding an algorithm that works for all plates/SNPs is a challenge.

We fit a mixture of five components to the SNP genotyping data, for the three genotype clusters, the negative controls, and failed samples, respectively. The three genotype components are modelled linearly with the mixture likelihood (3.4). The fourth component is modelled with a bivariate normal distribution, and the fifth is a uniform component for possible outlying points.

In the model fitting, we use the known labels for the negative controls. All the negative controls are assigned to the fourth component *a priori*, as are all points with both signals less than the corresponding maximum intensities of negative controls. Denote the set of these points by N. We

Table 3.5: Largest membership probabilities of the 7 points labelled in Figure 3.5 by MCLUST and MLC.

	Points						
	1	2	3	4	5	6	7
MLC	0.99	1.00	1.00	0.99	0.98	0.86	0.99
MCLUST	0.96	0.93	1.00	1.00	1.00	0.97	0.50

do not use the known labels of the positive controls; these samples are used only for evaluating the clustering results.

Hence, the modified mixture likelihood for the problem is

$$\begin{aligned} & L(\boldsymbol{\kappa}, p_1, \ldots, p_5 | \mathbf{x}_1, \ldots, \mathbf{x}_n) \\ = & \prod_{i \notin N} \left\{ \sum_{k=1}^{3} p_k N(\mathbf{a}_k' \mathbf{x}_i - b_k; 0, \sigma_k^2) + p_4 N(\mathbf{x}_i; \boldsymbol{\mu}, \Sigma) + p_5 \mathrm{U}(\mathbf{x}_i; R) \right\} \\ \times & \prod_{i \in N} N(\mathbf{x}_i; \boldsymbol{\mu}, \Sigma), \end{aligned} \tag{3.19}$$

where U is the uniform distribution on the rectangular region R given by $(\min(x_{i1}),\ \max(x_{i1})) \times (\min(x_{i2}),\ \max(x_{i2}))$. This mix of linear, elliptical and uniform components demonstrates the flexibility of our modelling approach.

Figure 3.5 displays the clustering results of one plate using four different algorithms: MLC, MCLUST, LGA and MIXREG. Note that the majority of the points lie on three lines, hence the need for linear clustering. The results of modified MLC and MCLUST are displayed in the upper-left and the upper-right panels of Figure 3.5. All the points assigned to one of the three genotype clusters by MCLUST are assigned exactly the same way by MLC. Points 1 to 7, however, are assigned by MCLUST as background noise while MLC classifies all these points to one of the genotype clusters. Indeed, we know that these points should belong to the genotype clusters, and furthermore, points 1, 3, 6 and 7 are positive controls, all classified correctly by MLC. Table 3.5 displays the largest membership probabilities of these seven points by MCLUST and MLC, respectively. For the first six of the seven points, the two methods are very confident of their different assignments. We speculate here that, MCLUST does not place these points in the genotype clusters because they appear to be outliers with respect to the normality assumptions. In contrast, MLC classifies a point only using the proximity to a line and does not model the position along a line.

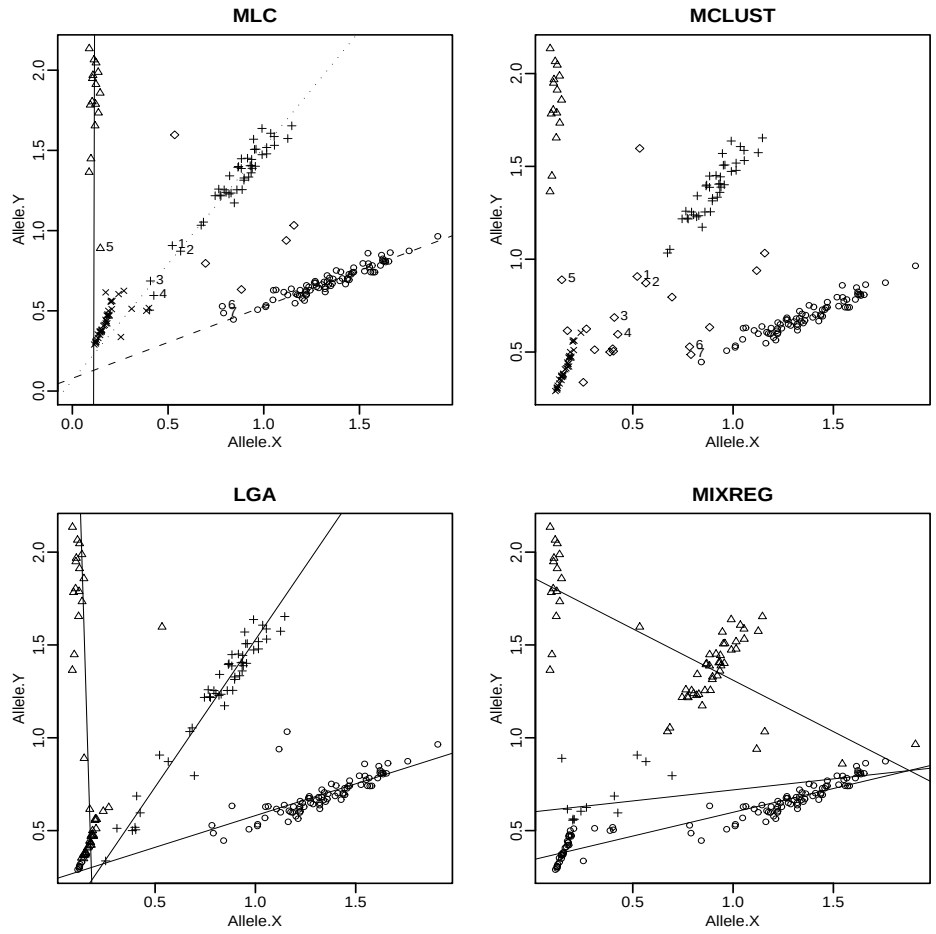

Figure 3.5: Clustering results from four methods applied to a Taqman data set plate. ○, + and △ denote the labels assigned for the three linear clusters (genotypes) by each method, × denotes points assigned as negative controls and ◇ denotes points assigned to the background cluster.

The lower-left panel of Figure 3.5 displays the results from LGA. Since LGA is formulated to deal with only linear clusters, we set the number of clusters to three. (If four or five clusters are chosen, the genotype clusters are mixed up.) The restriction to linear clusters means that LGA has difficulty separating the three genotype clusters from the negative controls and the failed samples. In the lower-right panel, we choose the signal for Allele Y as the response to implement MIXREG and set the number of regression lines to 3. MIXREG fails completely here. It finds only one of the three linear clusters. We obtained similar results if the signal for Allele X is chosen as the response.

3.8 Discussion

We have proposed a flexible model-based approach to linear clustering. It is flexible in the sense that only the deviation from a hyperplane is modelled parametrically; the position on the hyperplane is not modelled. The advantage of this approach is illustrated in the genotyping example, where the distribution along the line is complex and difficult to model. Furthermore, as was also illustrated in this example, we can incorporate elliptical clusters as necessary and a background cluster, borrowing from standard model-based clustering.

Robustness to outliers is desirable, as the assumption of normal deviations around hyperplanes is sensitive to large deviations in the orthogonal direction. In addition to the inclusion of a uniform background cluster (Banfield and Raftery, 1993), one option would be to use a heavier tailed distribution, for example, Student's t distribution with small degrees of freedom or with degrees of freedom depending on the data. This would adapt Peel and McLachlan (2000)'s EM algorithm for t mixture models from the elliptical context to the linear. The adaptation is straightforward but computationally more expensive. Further ideas include estimating the component covariance matrices in the M-step in a robust way, for example, trimming off some points.

With $\mathbf{a}'\mathbf{x} = b$, we are specifying a hyperplane in $d-1$ dimensions. With little effort, this could be generalized to a mixture of partial likelihoods, each of which specifies a hyperplane of dimension $q < d$,

$$l(\boldsymbol{\kappa}, \mathbf{p}|\mathbf{x}_{1:n}) = \prod_{i=1}^{n} \sum_{k=1}^{K} p_k N(A_k'\mathbf{x}_i - \mathbf{b}_k; \mathbf{0}, \Sigma_k), \tag{3.20}$$

where A is of dimension $d \times (d-q)$, b is a vector of dimension $d-q$,

and Σ_k is a $(d-q) \times (d-q)$ covariance matrix for the deviation from the hyperplane. In the extreme case of a 0-dimension hyperplane, which is a point, we have the usual mixture of multivariate normal distributions. A mixture of components with various dimensions could be considered.

A Bayesian version of this methodology would be helpful if some clusters are sparse but there is strong prior information about their approximate locations or properties (e.g., the parameters defining lines).

Bibliography

Banfield, J. and A. Raftery (1993). Model-Based Gaussian and Non-Gaussian Clustering. *Biometrics 49*(3), 803–821.

Biernacki, C., G. Celeux, and G. Govaert (1999). An improvement of the NEC criterion for assessing the number of clusters in a mixture model. *Pattern Recognition Letters 20*(3), 267–272.

Biernacki, C., G. Celeux, and G. Govaert (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence 22*(7), 719–725.

Bryant, P. and J. Williamson (1978). Asymptotic behaviour of classification maximum likelihood estimates. *Biometrika 65*(2), 273–281.

Campbell, N. and R. Mahon (1974). A multivariate study of variation in two species of rock crab of genus Leptograpsus. *Australian Journal of Zoology 22*(3), 417–425.

Celeux, G. and G. Govaert (1993). Comparison of the mixture and the classification maximum likelihood in cluster analysis. *Journal of statistical computation and simulation(Print) 47*(3-4), 127–146.

Celeux, G. and G. Soromenho (1996). An entropy criterion for assessing the number of clusters in a mixture model. *Journal of Classification 13*(2), 195–212.

Chen, H. and J. Chen (2001a). Large sample distribution of the likelihood ratio test for normal mixtures. *Statistics and Probability Letters 52*(2), 125–133.

Chen, H. and J. Chen (2001c). The Likelihood Ratio Test for Homogeneity in Finite Mixture Models. *The Canadian Journal of Statistics/La Revue Canadienne de Statistique 29*(2), 201–215.

Chen, H. and J. Chen (2001b). The likelihood ratio test for homogeneity in the finite mixture models. *Canadian Journal of Statistics 29*(2), 201–215.

Chen, H., J. Chen, and J. Kalbfleisch (2001). A modified likelihood ratio test for homogeneity in finite mixture models. *Journal of the Royal Statistical Society: Series B (Methodological) 63*(1), 19–29.

Chen, H., P. Meer, and D. Tyler (2001). Robust regression for data with multiple structures. *2001 IEEE Conference on Computer Vision and Pattern Recognition 1*, 1069–1075.

Chen, J. (1998). Penalized likelihood-ratio test for finite mixture models with multinomial observations. *The Canadian Journal of Statistics 26*(4), 583–599.

Chen, J. and J. Kalbfleisch (1996). Penalized Minimum-Distance Estimates in Finite Mixture Models. *The Canadian Journal of Statistics/La Revue Canadienne de Statistique 24*(2), 167–175.

Chen, J. and J. Kalbfleisch (2005). Modified likelihood ratio test in finite mixture models with a structural parameter. *Journal of Statistical Planning and Inference 129*(1-2), 93–107.

Ciuperca, G., A. Ridolfi, and J. Idier (2003). Penalized Maximum Likelihood Estimator for Normal Mixtures. *Scandinavian Journal of Statistics 30*(1), 45–59.

DeSarbo, W. and W. Cron (1988). A maximum likelihood methodology for clusterwise linear regression. *Journal of Classification 5*(2), 249–282.

Everitt, B. and D. Hand (1981). Finite mixture distributions. *Monographs on Applied Probability and Statistics, London: Chapman and Hall*.

Everitt, B., S. Landau, and M. Leese (2001). *Cluster Analysis*. Hodder Arnold.

Fraley, C. and A. Raftery (1998). How Many Clusters? Which Clustering Method? Answers Via Model-Based Cluster Analysis. *The Computer Journal 41*(8), 578.

Fraley, C. and A. Raftery (2002). Model-Based Clustering, Discriminant Analysis, and Density Estimation. *Journal of the American Statistical Association 97*(458), 611–632.

Fraley, C. and A. Raftery (2007). *mclust: Model-Based Clustering / Normal Mixture Modeling*. R package version 3.1-1.

Fujisawa, H., S. Eguchi, M. Ushijima, S. Miyata, Y. Miki, T. Muto, and M. Matsuura (2004). Genotyping of single nucleotide polymorphism using model-based clustering. *Bioinformatics 20*, 5.

Fuller, W. (1987). Measurement error models. *Wiley Series in Probability and Mathematical Statistics, New York: Wiley, 1987*.

Ganesalingam, S. (1989). Classification and Mixture Approaches to Clustering via Maximum Likelihood. *Applied Statistics 38*(3), 455–466.

García-Escudero, L., A. Gordaliza, R. San Martín, S. van Aelst, and R. Zamar (2007). Robust linear clustering. Preprint.

Harrington, J. (2007). *lga: Tools for linear grouping analysis (LGA)*. R package version 1.0-0.

Hathaway, R. (1983). *Constrained maximum-likelihood estimation for a mixture of m univariate normal distributions*. Ph. D. thesis, Rice University.

Hathaway, R. (1985). A Constrained Formulation of Maximum-Likelihood Estimation for Normal Mixture Distributions. *The Annals of Statistics 13*(2), 795–800.

Hathaway, R. (1986). A constrained EM algorithm for univariate normal mixtures. *Journal of Statistical Computation and Simulation 23*(3), 211–230.

Kang, H., Z. Qin, T. Niu, and J. Liu (2004). Incorporating Genotyping Uncertainty in Haplotype Inference for Single-Nucleotide Polymorphisms. *The American Journal of Human Genetics 74*(3), 495–510.

Keribin, C. (2000). *Sankhya: The Indian Journal of Statistics 62*, 49–66.

MacLachlan, G. and D. Peel (2000). *Finite mixture models*. J. Wiley.

Marriott, F. (1975). Separating mixtures of normal distributions. *Biometrics 31*(3), 767–769.

McLachlan, G. (1987). On Bootstrapping the Likelihood Ratio Test Stastistic for the Number of Components in a Normal Mixture. *Applied Statistics 36*(3), 318–324.

Olivier, M., L. Chuang, M. Chang, Y. Chen, D. Pei, K. Ranade, A. de Witte, J. Allen, N. Tran, D. Curb, et al. (2002). High-throughput genotyping of single nucleotide polymorphisms using new biplex invader technology. *Nucleic Acids Research 30*(12), e53.

Peel, D. and G. McLachlan (2000). Robust mixture modelling using the t distribution. *Statistics and Computing 10*(4), 339–348.

Ranade, K., M. Chang, C. Ting, D. Pei, C. Hsiao, M. Olivier, R. Pesich, J. Hebert, Y. Chen, V. Dzau, et al. (2001). High-Throughput Genotyping with Single Nucleotide Polymorphisms. *Genome Research 11*(7), 1262–1268.

Redner, R. (1981). Note on the Consistency of the Maximum Likelihood Estimate for Nonidentifiable Distributions. *The Annals of Statistics 9*(1), 225–228.

Scott, A. and M. Symons (1971). Clustering Methods Based on Likelihood Ratio Criteria. *Biometrics 27*(2), 387–397.

Smyth, P. (2000). Model selection for probabilistic clustering using cross-validated likelihood. *Statistics and Computing 10*(1), 63–72.

Spath, H. (1982). Fast algorithm for clusterwise linear regression. *COMPUT. 29*(2), 175–181.

Thode Jr, H., S. Finch, and N. Mendell (1988). Simulated Percentage Points for the Null Distribution of the Likelihood Ratio Test for a Mixture of Two Normals. *Biometrics 44*(4), 1195–1201.

Turner, T. (2000). Estimating the propagation rate of a viral infection of potato plants via mixtures of regressions. *Journal of the Royal Statistical Society Series C(Applied Statistics) 49*(3), 371–384.

Turner, T. (2006). *mixreg: Functions to fit mixtures of regressions.* R package version 0.0-2.

Van Aelst, S., X. Wang, R. Zamar, and R. Zhu (2006). Linear grouping using orthogonal regression. *Computational Statistics and Data Analysis 50*(5), 1287–1312.

Wald, A. (1949). Note on the Consistency of the Maximum Likelihood Estimate. *The Annals of Mathematical Statistics 20*(4), 595–601.

Chapter 4

Bayesian linear clustering with application to single nucleotide polymorphism genotyping

4.1 Introduction

This paper was motivated by a clustering problem in single nucleotide polymorphism (SNP) genotyping. A single nucleotide polymorphism (SNP, pronounced as "snip") is a single-base variation in a genome. The genetic code of life is specified by the four nucleotide "letters": A (adenine), C (cytosine), G (guanine) and T (thymine). There are two complementary DNA strands. It is sufficient to consider only one. SNP variation occurs when a single nucleotide, such as an A, is replaced by one of the other three letters C, G or T. One SNP usually involves only two letters, referred to generically throughout this paper as allele X and allele Y. An individual has two copies of genetic information passed from both parents. So the genotypes at a specific genetic position have three possibilities: homozygote for allele X, homozygote for allele Y and heterozygote for both alleles. SNP genotyping, determination of genotypes of individuals, plays an increasing role in genetics studies.

For a small project, it is possible to make genotype calls manually. In most cases, it is not hard for an expert to perform this job, and the "eyeballing" procedure usually gives reasonable results due to its sophisticated incorporation of prior information. For large-scale studies, however, manual

A version of this chapter will be submitted for publication. Authors: Guohua Yan, William J. Welch and Ruben H. Zamar.

scoring can become a daunting challenge. Typical SNP genotyping applications involve thousands of patients and hundreds of SNPs. Hence, reliable automated genotyping methods are highly needed.

TaqMan SNP Genotyping Assay (Applied Biosystems) is a fluorescence-based high-throughput genotyping technology. Blood samples of many individuals are arranged in a 96- or 384-well plate and are assayed simultaneously. Two fluorescent signals, x_1 and x_2, also called Allele.X and Allele.Y below, are used to detect the presence or absence of the two alleles. There is a third fluorescent signal presumably for normalizing the signal intensities. However, its effectiveness is suspicious in some situations. See Chapter 2.

For quality controls purpose, some blood samples with known genotypes are also included in a plate; these samples are called positive controls. As well, some wells do not have genetic material; these are called negative controls.

In the scatterplot of a SNP data set, there are typically four clusters, as in Figure 4.1. In the lower-right, upper-left, and upper-right corners are three clusters, presumably containing samples of wild-type homozygotes, variant homozygotes, and heterozygotes, respectively. In the lower-left corner, the cluster may contain negative controls and/or some failed samples. A clustering algorithm is usually employed to call the SNP genotypes. Algorithms commonly used in the literature include k-means (Olivier et al., 2002; Ranade et al., 2001) and model-based algorithms (Fujisawa et al., 2004; Kang et al., 2004). The proprietary software Sequence Detection System is included in a thermal cycler. Usually human intervention of an expert operator is usually needed to review the genotype calling results. Finding an algorithm that works well for all plates/SNPs is a challenge, especially when the variant allele homozygous genotype cluster is sparse, in which standard clustering algorithms often fail.

Several clustering algorithms k-means, MCLUST (Fraley and AE, 2006), LGA (Harrington, 2007), MIXREG (Turner, 2006) and MLC (Yan et al., 2008) are applied to the genotyping data in Figure 4.1. Figure 4.2 displays their corresponding clustering results. We can see that all of these algorithms fail to identify the five points in the upper-left corner as variant allele homozygotes while an expert operator would do with confidence. These algorithms fail because the variant allele homozygous genotype cluster is very sparse. This motivates us to adopt a Bayesian approach to incorporate available prior information. In addition, we found that it is convenient to model the genotype clusters as linear structures.

Most clustering methods and algorithms cluster data points around "centers". However, some data sets, as in the SNP genotype setting, form groups

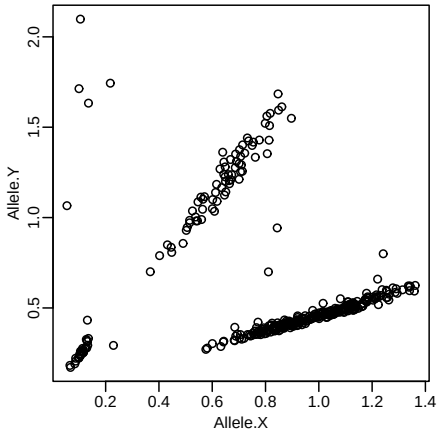

Figure 4.1: Scatterplot of one plate in which the variant allele homozygous cluster has only five points (upper-left).

through linear relationships and standard clustering techniques are usually not able to find these linear patterns. By linear clustering, we mean detecting these linear clusters in a data set.

Van Aelst et al. (2006) proposed a method called Linear Grouping Algorithm (LGA) for linear clustering problems without assuming a response variable. Yan et al. (2008) proposed a partial likelihood approach which models only the signed orthogonal distance from each data point to a hypothesized hyperplane and treats all variables symmetrically.

In this paper we introduce a hierarchical modeling approach, which is particularly appropriate for identifying sparse linear clusters. We show that the sparse cluster in our SNP genotyping dataset can be successfully identified after a careful specification of the prior distributions. The rest of this paper is organized as follows. In Section 2, we describe a hierarchical model framework for linear clustering. In Section 3, we discuss in details sampling issues. Our approach to linear clustering is illustrated with a relatively simple dataset (crab dataset) in Section 4. In Section 5, we revisit the SNP genotyping motivating dataset and show that a careful specification of the prior distributions is critical for the success of the clustering algorithm. A brief discussion follows in Section 6.

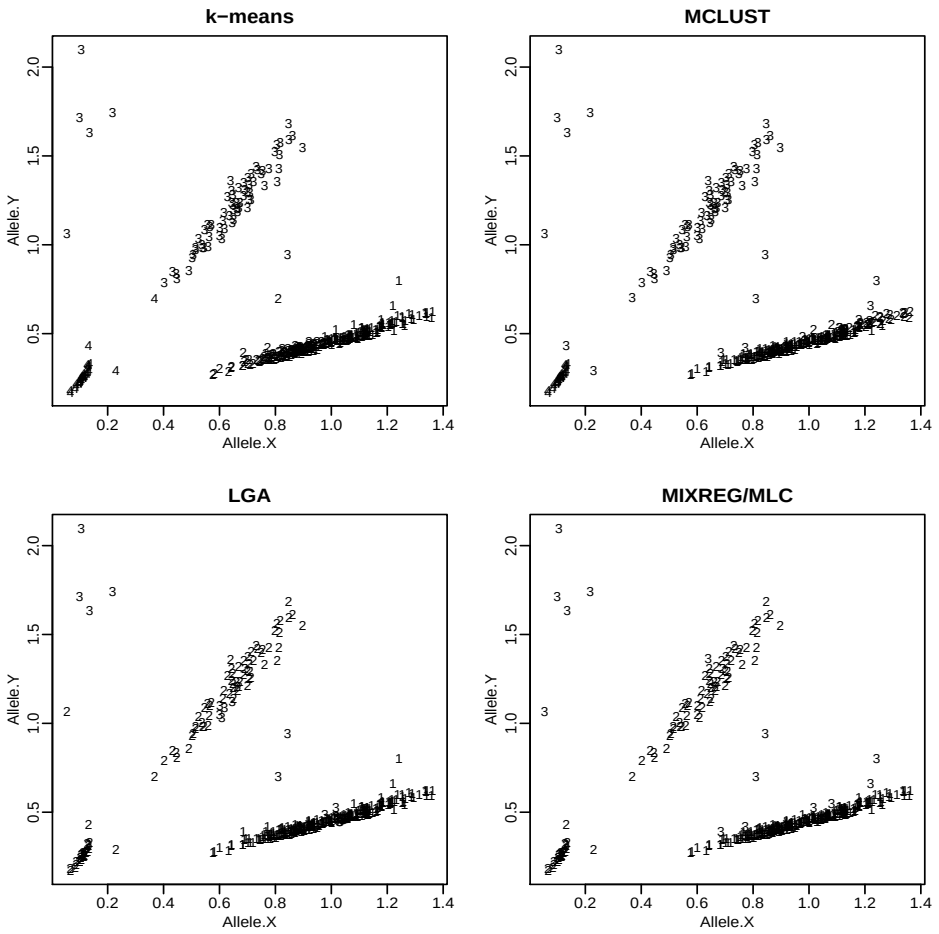

Figure 4.2: Clustering results of several clustering algorithms. For k-means and MCLUST, the number of clusters are set to 4; for the remaining algorithms, the number of lines are set to 3.

4.2 Model specification

4.2.1 Model for one cluster: orthogonal regression

We first introduce a Bayesian approach for orthogonal regression (see e.g., Fuller, 1987). To model a single cluster, assume that the vector-valued data $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are drawn by a random mechanism represented by a random vector $\mathbf{X}$ in a d-dimensional space. We do not model the distribution of $\mathbf{X}$ except for its deviation from an unknown hyperplane.

Let $\{\mathbf{x} : \mathbf{a}'\mathbf{x} - b = 0\}$ be a hyperplane. For identifiability purpose, we assume that $\mathbf{a}'\mathbf{a} = 1$ and that the first nonzero element of $\mathbf{a}$ is positive. Our prior information about this hyperplane is summarized by a prior distribution $\pi(\mathbf{a}, b)$. Given $\mathbf{a}$ and b, the signed orthogonal deviation $\mathbf{a}'\mathbf{X} - b$ from the hyperplane $\{\mathbf{x} : \mathbf{a}'\mathbf{x} - b = 0\}$ has a density function $p(\cdot \mid \sigma)$. Let $\pi(\sigma)$ be the prior distribution for σ. Denote $\boldsymbol{\theta} = (\mathbf{a}', b, \sigma)'$. The posterior distribution of $\boldsymbol{\theta}$ is

$$\pi(\boldsymbol{\theta}|\mathbf{x}_1, \ldots, \mathbf{x}_n) \propto \left\{ \prod_{i=1}^{n} p(\mathbf{a}'\mathbf{x}_i - b|\sigma) \right\} \pi(\mathbf{a}, b)\pi(\sigma). \tag{4.1}$$

Let $\bar{\mathbf{x}}$ and S be the sample mean and sample covariance matrix of $\mathbf{x}_1, \ldots, \mathbf{x}_n$ respectively. If we take $p(\cdot|\sigma)$ to be the normal density function $N(\cdot \mid 0, \sigma^2)$ and set $\pi(\mathbf{a}, b) \propto 1$ and $\pi(\sigma) \propto 1$, the maximum *a posteriori* estimator of $\boldsymbol{\theta}$ is $(\hat{\mathbf{a}}', \hat{b}, \hat{\sigma})'$, where $\hat{\sigma}^2$ is the smallest eigenvalue of S, $\hat{\mathbf{a}}$ is the (standardized) eigenvector associated with $\hat{\sigma}^2$ and $\hat{b} = \hat{\mathbf{a}}'\bar{\mathbf{x}}$. This is a connection with orthogonal regression, see for example, Fuller (1987); for the distributions of the eigenvalue $\hat{\sigma}^2$ and the eigenvector $\hat{\mathbf{a}}$ when $\mathbf{X}$ is normal, see Anderson (2003).

4.2.2 Model for linear clustering

Now we assume that the data $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are drawn by a more complicated random mechanism, still represented by random vector $\mathbf{X}$ in a d-dimensional space which we do not model fully. The data now lie around K hyperplanes $\{\mathbf{x} : \mathbf{a}_k'\mathbf{x} = b_k\}$, $k = 1, \ldots, K$, for which our prior knowledge is summarized in $\pi(\mathbf{a}_1, b_1, \ldots, \mathbf{a}_K, b_K)$. Let $\mathbf{Z} = (Z_1, \ldots, Z_K)'$ be a random vector indicating these hyperplanes, $Z_k = 1$ with probability p_k for $k = 1, \ldots, K$, $Z_k = 0$ or 1 and $\sum_{k=1}^{K} Z_k = 1$. Let $\mathbf{p} = (p_1, \ldots, p_K)'$. We denote by $\pi(\mathbf{p})$ the prior distribution for $\mathbf{p}$. We assume that, conditional on $Z_k = 1$, the signed orthogonal deviation $\mathbf{a}_k'\mathbf{X} - b_k$ has a density function $p(\cdot|\sigma_k)$ for $k = 1, \ldots, K$. Let $\pi(\sigma_1, \ldots, \sigma_K)$ be the prior distribution for $(\sigma_1, \ldots, \sigma_K)'$. Let $\mathbf{z}_1, \ldots, \mathbf{z}_n$ be the corresponding unobservable indicators for the data $\mathbf{x}_1, \ldots, \mathbf{x}_n$, where

$\mathbf{z}_i = (z_{i1}, \ldots, z_{iK})'$ for $i = 1, \ldots, n$. Let $\boldsymbol{\theta}$ be the collection of all the parameters,

$$\boldsymbol{\theta} = (\mathbf{a}_1', b_1, \sigma_1, \ldots, \mathbf{a}_K', b_K, \sigma_K, \mathbf{p}')'.$$

Formally, the posterior distribution of the unobservable cluster indicators $\mathbf{z}_1$, ..., $\mathbf{z}_n$ and $\boldsymbol{\theta}$ is

$$\begin{aligned}
&\pi(\mathbf{z}_1, \ldots, \mathbf{z}_n, \boldsymbol{\theta} | \mathbf{x}_1, \ldots, \mathbf{x}_n) \\
&\propto \left\{ \prod_{i=1}^{n} \prod_{k=1}^{K} [p_k p(\mathbf{a}_k' \mathbf{x}_i - b_k | \sigma_k)]^{z_{ik}} \right\} \pi(\mathbf{a}_1, b_1, \ldots, \mathbf{a}_K, b_K) \pi(\sigma_1, \ldots, \sigma_K) \pi(\mathbf{p}).
\end{aligned} \tag{4.2}$$

This formula is the basis for our linear clustering, in which we are primarily interested in the cluster indicators $\mathbf{z}_1, \ldots, \mathbf{z}_n$ and the elements of $\boldsymbol{\theta}$ are nuisance parameters. A Gibbs sampling scheme based on this formula is detailed in next section.

In addition, if we sum out the indicators $\mathbf{z}_1, \ldots, \mathbf{z}_n$, the marginal distribution for $\boldsymbol{\theta}$ is

$$\begin{aligned}
&\pi(\boldsymbol{\theta} | \mathbf{x}_1, \ldots, \mathbf{x}_n) \\
&\propto \left\{ \prod_{i=1}^{n} \sum_{k=1}^{K} p_k p(\mathbf{a}_k' \mathbf{x}_i - b_k | \sigma_k) \right\} \pi(\mathbf{a}_1, b_1, \ldots, \mathbf{a}_K, b_K) \pi(\sigma_1, \ldots, \sigma_K) \pi(\mathbf{p}).
\end{aligned} \tag{4.3}$$

This marginal distribution can be used if we would like to set up other sampling schemes such as Metropolis-Hastings, tempering MCMC or SMC methods. Once a MCMC sample for $\boldsymbol{\theta}$ is available, it is straightforward to sample cluster labels at each iteration. At iteration t, with $\boldsymbol{\theta}^{(t)}$, sample $\mathbf{z}_i$ from multinomial distribution

$$\Pr(\mathbf{z}_i = k) \propto p_k^{(t)} p((\mathbf{a}')_k^{(t)} \mathbf{x}_i - b_k^{(t)} | \sigma_k^{(t)}). \tag{4.4}$$

After a sequence of iterations, we have a sample $(\mathbf{z}_i^{(1)}, \ldots, \mathbf{z}_i^{(T)})$ from posterior marginal distribution $\pi(\mathbf{z}_i | \mathbf{x}_1, \ldots, \mathbf{x}_n)$, for $i = 1, \ldots, n$. The point $\mathbf{x}_i$ can be classified into the cluster with the largest frequency in the posterior sample; and the largest frequency provides a good measure of the cluster membership.

4.3 Sampling algorithms

4.3.1 Gibbs sampling

The model framework in (4.2) is conceptually straightforward; the posterior sampling, however, can be difficult, as in any Bayesian mixture modelling (Celeux et al., 2000). We present here a sampling algorithm for normal orthogonal deviations and, to make the algorithm more efficient, we integrate some parameters by using conjugate priors. For ease of presentation, we assume the data are in a 2-dimensional space.

The density functions for normal orthogonal deviations are

$$p(\mathbf{a}_k'\mathbf{x} - b_k|\sigma_k) = N(\mathbf{a}_k'\mathbf{x} - b_k; 0, \sigma_k^2), \tag{4.5}$$

for $k = 1, \ldots, K$, where $N(\cdot\,; 0, \sigma^2)$ is the normal density function with mean 0 and variance σ_k^2. The priors for $(\sigma_1^2, \ldots, \sigma_K^2)$ follow independent inverse Gamma distributions,

$$\pi(\sigma_1^2, \ldots, \sigma_K^2) = \prod_{k=1}^{K} \mathrm{IG}(\sigma_k^2; \delta_{1k}, \delta_{2k}), \tag{4.6}$$

where $\mathrm{IG}(\cdot\,; \delta_{k1}, \delta_{k2})$ is the density function of inverse Gamma with shape parameter δ_{k1} and rate parameter δ_{k2}. In the case of equal variances, $\sigma_1^2 = \ldots = \sigma_K^2 = \sigma^2$, we assume

$$\pi(\sigma^2) = \mathrm{IG}(\sigma^2; \delta_1, \delta_2), \tag{4.7}$$

For $b_1, \ldots, b_K$, we set

$$\pi(b_1, \ldots, b_K|\sigma_1^2, \ldots, \sigma_K^2) = \prod_{k=1}^{K} N(b_k; \kappa_{1k}, \sigma_k^2/\kappa_{2k}). \tag{4.8}$$

With these specifications, we consider the following special case of the posterior distribution (4.2),

$$\begin{aligned}
&\pi(\mathbf{z}_1, \ldots, \mathbf{z}_n, \boldsymbol{\theta}|\mathbf{x}_1, \ldots, \mathbf{x}_n) \\
&\propto \left\{ \prod_{i=1}^{n} \prod_{k=1}^{K} [p_k N(\mathbf{a}_k'\mathbf{x}_i - b_k; 0, \sigma_k^2)]^{z_{ik}} \right\} \\
&\times \left\{ \prod_{k=1}^{K} N(b_k; \kappa_{1k}, \sigma_k^2/\kappa_{2k}) \mathrm{IG}(\sigma_k^2; \delta_{1k}, \delta_{2k}) \right\} \pi(\mathbf{a}_1, \ldots, \mathbf{a}_K) \pi(\mathbf{p}).
\end{aligned} \tag{4.9}$$

Conditional on $\boldsymbol{\theta}$ and the data $\mathbf{x}_1, \ldots, \mathbf{x}_n$, $\mathbf{z}_1, \ldots, \mathbf{z}_n$ are independently distributed,

$$\mathbf{z}_i|(\boldsymbol{\theta}, \mathbf{x}_i) \sim M(1, (p_1 N(\mathbf{a}_1'\mathbf{x}_i - b_1; 0, \sigma_1^2), \ldots, p_K N(\mathbf{a}_K'\mathbf{x}_i - b_K; 0, \sigma_K^2))'), \tag{4.10}$$

for $i = 1, \ldots, n$, where $M(n, \mathbf{p})$ is the probability mass function for multinomial distribution with n trials and probability vector $\mathbf{p}$. Let

$$n_k = \sum_{i=1}^{n} z_{ik}, \quad \bar{\mathbf{x}}_k = \frac{1}{n_k} \sum_{i=1}^{n} z_{ik}\mathbf{x}_i, \tag{4.11}$$

and

$$A_k = \sum_{i=1}^{n} z_{ik}(\mathbf{x}_i - \bar{\mathbf{x}}_k)(\mathbf{x}_i - \bar{\mathbf{x}}_k)'. \tag{4.12}$$

The full conditional distribution of b_k is

$$\begin{aligned} b_k|\text{e.e.} &= b_k|(\mathbf{z}_1, \ldots, \mathbf{z}_n, \mathbf{a}_k, \sigma_k^2, \mathbf{x}_1, \ldots, \mathbf{x}_n) \\ &\sim N(\frac{\kappa_{1k}\kappa_{2k} + n_k\mathbf{a}_k'\bar{\mathbf{x}}_k}{n_k + \kappa_{2k}}, \frac{\sigma_k^2}{n_k + \kappa_{2k}}), \end{aligned} \tag{4.13}$$

for $k = 1, \ldots, K$, where "e.e." stands for "everything else". We integrate out $b_1, \ldots, b_K$ from (4.9) and get

$$\begin{aligned} &\pi(\mathbf{z}_1, \ldots, \mathbf{z}_n, \mathbf{a}_1, \sigma_1^2, \ldots, \mathbf{a}_K, \sigma_K^2, \mathbf{p}|\mathbf{x}_1, \ldots, \mathbf{x}_n) \\ &\propto \left\{ \prod_{k=1}^{K} p_k^{n_k} (\sigma_k^2)^{-n_k/2} (n_k + \kappa_{2k})^{-1/2} \exp(-\frac{\delta_{2k}^*}{2\sigma_k^2}) \right\} \\ &\times \left\{ \prod_{k=1}^{K} \text{IG}(\sigma_k^2; \delta_{1k}, \delta_{2k}) \right\} \pi(\mathbf{a}_1, \ldots, \mathbf{a}_K)\pi(\mathbf{p}), \end{aligned} \tag{4.14}$$

where

$$\delta_{2k}^* = \mathbf{a}_k' A_k \mathbf{a}_k + n_k(\mathbf{a}_k'\bar{\mathbf{x}}_k)^2 + \kappa_{1k}^2\kappa_{2k} - \frac{(\kappa_{1k}\kappa_{2k} + n_k\mathbf{a}_k'\bar{\mathbf{x}}_k)^2}{n_k + \kappa_{2k}}. \tag{4.15}$$

The distribution of σ_k^2 conditioning on everything else except b_k is

$$\sigma_k^2|(\mathbf{z}_1, \ldots, \mathbf{z}_n, \mathbf{a}_k, \mathbf{x}_1, \ldots, \mathbf{x}_n) \sim \text{IG}(\delta_{1k} + \frac{n_k}{2}, \delta_{2k} + \frac{\delta_{2k}^*}{2}). \tag{4.16}$$

If we assume that $\sigma_1^2 = \ldots = \sigma_K^2 = \sigma^2$, then the distribution of σ^2 conditioning on $\mathbf{z}_1, \ldots, \mathbf{z}_n, \mathbf{a}_1, \ldots, \mathbf{a}_K, \mathbf{p}$ is

$$\sigma^2 | (\mathbf{z}_1, \ldots, \mathbf{z}_n, \mathbf{a}_1, \ldots, \mathbf{a}_K, \mathbf{x}_1, \ldots, \mathbf{x}_n) \sim \text{IG}(\delta_1 + \frac{n}{2}, \delta_2 + \sum_{k=1}^{K} \frac{\delta_{2k}^*}{2}). \quad (4.17)$$

We further integrate out $\sigma_1^2, \ldots, \sigma_K^2$ from (4.14) and get

$$\begin{aligned} &\pi(\mathbf{z}_1, \ldots, \mathbf{z}_n, \mathbf{a}_1, \ldots, \mathbf{a}_K, \mathbf{p} | \mathbf{x}_1, \ldots, \mathbf{x}_n) \\ &\propto \left\{ \prod_{k=1}^{K} p_k^{n_k} (n_k + \kappa_{2k})^{-1/2} \Gamma(\delta_{1k} + \frac{n_k}{2}) (\delta_{2k} + \frac{\delta_{2k}^*}{2})^{-(\delta_{1k} + \frac{n_k}{2})} \right\} \\ &\times \pi(\mathbf{a}_1, \ldots, \mathbf{a}_K) \pi(\mathbf{p}). \end{aligned} \quad (4.18)$$

For proportions $\mathbf{p}$, we use a Dirichlet prior,

$$\pi(\mathbf{p}) = D(\mathbf{p}; \boldsymbol{\alpha}), \quad (4.19)$$

where $D(\cdot\,; \alpha)$ is the Dirichlet density function with parameter $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)'$. The full conditional distribution of $\mathbf{p}$ is

$$\mathbf{p} | \text{e.e.} = \mathbf{p} | (\mathbf{z}_1, \ldots, \mathbf{z}_n) \sim D(\boldsymbol{\alpha}^*), \quad (4.20)$$

where $\boldsymbol{\alpha}^* = (\alpha_1^*, \ldots, \alpha_K^*)'$ with $\alpha_k^* = \alpha + n_k$ for $k = 1, \ldots, K$. Therefore,

$$\begin{aligned} &\pi(\mathbf{z}_1, \ldots, \mathbf{z}_n, \mathbf{a}_1, \ldots, \mathbf{a}_K | \mathbf{x}_1, \ldots, \mathbf{x}_n) \\ &\propto \left\{ \frac{1}{\Gamma(\sum_{i=1}^{K} \alpha_k^*)} \prod_{k=1}^{K} \Gamma(\alpha_k^*) (n_k + \kappa_{2k})^{-1/2} \Gamma(\delta_{1k} + \frac{n_k}{2}) (\delta_{2k} + \frac{\delta_{2k}^*}{2})^{-(\delta_{1k} + \frac{n_k}{2})} \right\} \\ &\times \pi(\mathbf{a}_1, \ldots, \mathbf{a}_K). \end{aligned} \quad (4.21)$$

Note that $\mathbf{a}_k' \mathbf{a}_k = 1$ for $k = 1, \ldots, K$. We re-parameterize $\mathbf{a}_k$ as

$$\mathbf{a}_k = (\frac{1}{\sqrt{1 + a_k^2}}, \frac{a_k}{\sqrt{1 + a_k^2}})',$$

for $k = 1, \ldots, K$ and use normal priors for $(a_1, \ldots, a_K)'$,

$$\pi(a_1, \ldots, a_K) = \prod_{k=1}^{K} N(a_k; \nu_{1k}, \nu_{2k}^2). \quad (4.22)$$

The conditional distribution of a_k on cluster indicators is

$$\pi(a_k|\mathbf{z}_1,\ldots,\mathbf{z}_n,\mathbf{x}_1,\ldots,\mathbf{x}_n) \propto \Gamma(\delta_{1k}+\frac{n_k}{2})(\delta_{2k}+\frac{\delta^*_{2k}}{2})^{-(\delta_{1k}+\frac{n_k}{2})}N(a_k;\nu_{1k},\nu^2_{2k}). \tag{4.23}$$

From the above, discussion, a possible sampling algorithm is as follows.

Algorithm 1:

1. Initialize the algorithm with $\boldsymbol{\theta}^{(0)}$.
2. At iteration t,
   (a) Sample cluster indicators $\mathbf{z}_1^{(t)},\ldots,\mathbf{z}_n^{(t)}$ from multinomial distribution (4.10), where $\boldsymbol{\theta}$ is replaced with $\boldsymbol{\theta}^{(t-1)}$.
   (b) i. Sample $\mathbf{p}^{(t)}$ from Dirichlet distribution (4.20), where the cluster indicators are replaced with $\mathbf{z}_1^{(t)}$, ..., $\mathbf{z}_n^{(t)}$.
   ii. Sample $a_1^{(t)},\ldots,a_K^{(t)}$ using a random walk Metropolis-Hastings scheme from (4.23), where cluster indicators are replaced by $\mathbf{z}_1^{(t)}$, ..., $\mathbf{z}_n^{(t)}$, respectively.
   iii. Sample $(\sigma_1^2)^{(t)},\ldots,(\sigma_K^2)^{(t)}$ from inverse Gamma distribution (4.16), where $\mathbf{z}_1,\ldots,\mathbf{z}_n,a_1,\ldots,a_K$ are replaced by $\mathbf{z}_1^{(t)}$, ..., $\mathbf{z}_n^{(t)}$, $a_1^{(t)}$, ..., $a_K^{(t)}$, respectively; in the case of equal variances $\sigma_1^2=\ldots=\sigma_K^2=\sigma^2$, sample $(\sigma^2)^{(t)}$ from (4.17).
   iv. Sample $b_1,\ldots,b_K$ from normal distribution (4.13), with all parameters and cluster indicators replaced with those in the current iteration.

4.3.2 Metropolis-Hastings sampling

Alternatively, we can sum out the indicators $\mathbf{z}_1,\ldots,\mathbf{z}_n$ and sample $\boldsymbol{\theta}$ from the marginal posterior distribution $\pi(\boldsymbol{\theta}|\mathbf{x}_1,\ldots,\mathbf{x}_n)$ in (4.3). Standard random walk Metropolis-Hastings algorithms can be used directly to sample from the marginal posterior distribution $\pi(\boldsymbol{\theta}|\mathbf{x}_1,\ldots,\mathbf{x}_n)$. However, we found that it is advantageous to reparameterize $\boldsymbol{\theta}$ so that the constraints such as $\sum_{k=1}^K p_k = 1$ and $\mathbf{a}'\mathbf{a} = 1$ are implicitly incorporated (see Section 4.5). Since the proposal distribution and the posterior distribution have the same support, the sampling schemes on the transformed parameters are more efficient. Suppose we reparameterize $\boldsymbol{\theta}$ as

$$\boldsymbol{\theta} = \phi(\boldsymbol{\eta}),$$

such that ϕ is a one-to-one map and that the domain of $\boldsymbol{\eta}$ is $\Re^q$, where q is the dimension of η. Let J be the Jacobian

$$J = \frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\eta}},$$

with the understanding that only q free components of $\boldsymbol{\theta}$ are used. Then the marginal posterior distribution of $\boldsymbol{\eta}$ is

$$\pi(\boldsymbol{\eta}|\mathbf{x}_1, \ldots, \mathbf{x}_n) = \pi(\boldsymbol{\theta}|\mathbf{x}_1, \ldots, \mathbf{x}_n)|J|. \tag{4.24}$$

To sample $\boldsymbol{\eta}$, we first run a Gibbs sampler, within which a random walk Metropolis step is used if a full conditional distribution is not of an explicit form.

Algorithm 2.0:

1. Initialize the algorithm with $\boldsymbol{\eta}^{(0)}$.
2. At iteration t, update $\eta_1^{(t-1)}, \ldots, \eta_q^{(t-1)}$ sequentially. For the jth component,
   (a) Sample ν from $N(\eta_j^{(t-1)}, s_j^2)$.
   (b) Compute $r = \min\{1, \frac{\pi(\eta_1^{(t)}, \ldots, \eta_{j-1}^{(t)}, \nu, \eta_{j+1}^{(t-1)}, \ldots, \eta_q^{(t-1)}|\mathbf{x}_1, \ldots, \mathbf{x}_n)}{\pi(\eta_1^{(t)}, \ldots, \eta_{j-1}^{(t)}, \eta_j^{(t-1)}, \eta_{j+1}^{(t-1)}, \ldots, \eta_q^{(t-1)}|\mathbf{x}_1, \ldots, \mathbf{x}_n)}\}$.
   (c) Update $\eta_j^{(t)} = \nu$ with probability r.

After a burn-in period, we can get a crude estimate $\hat{\Sigma}$ of the covariance matrix of $\boldsymbol{\eta}$. In the second stage of the algorithm, we update $\boldsymbol{\eta}$ using a Metropolis-Hastings sampler in one block.

Algorithm 2:

1. Initialize the algorithm with $\boldsymbol{\eta}^{(0)}$.
2. At iteration t,
   (a) Sample $\boldsymbol{\nu}$ from $N(\boldsymbol{\eta}^{(t-1)}, s^2\hat{\Sigma})$.
   (b) Compute $r = \min\{1, \frac{\pi(\boldsymbol{\nu}|\mathbf{x}_1, \ldots, \mathbf{x}_n)}{\pi(\boldsymbol{\eta}^{(t-1)}|\mathbf{x}_1, \ldots, \mathbf{x}_n)}\}$.
   (c) Update $\boldsymbol{\eta}^{(t)} = \boldsymbol{\nu}$ with probability r.

Here we use an *adaptive* strategy; see page 307, Gelman et al. (2004).

4.4 Australia rock crab data

Now we consider the rock crab data set of Campbell and Mahon (1974) on the genus *Leptograpsus*. We focus on the 100 blue crabs, 50 of which are males and 50 are females. Each crab has five measurements, FL, the width of frontal lip, RW, the rear width, CL, the length along the midline, CW, the maximum width of the carapace, and BD, the body depth in mm. As in Peel and McLachlan (2000), we are interested in the classification of male and female crabs. Peel and McLachlan (2000) had 18 crabs misclassified applying a mixture of two t distributions using all the five measurements. Using a mixture of two normal distributions resulted in one more crab being misclassified.

Inspecting all pairwise scatter plots of the data set, RW and CL display two fairly well separated lines as in Figure 4.3. We shall classify these crabs using only RW and CL. The number of clusters is set to 2. We run the Gibbs sampler "Algorithm 1" based on (4.2).

To initialize the sampler, we adopt the strategy of Van Aelst et al. (2006). We randomly select $K = 2$ mutually exclusive subsets of $d + 1 = 3$ observations from $\mathbf{x}_1, \ldots, \mathbf{x}_n$. For $k = 1, \ldots, K$, we compute the maximum partial likelihood estimates $\mathbf{a}_k^{(0)}$, $b_k^{(0)}$ and $(\sigma_k^2)^{(0)}$ from the kth subset of $d + 1$ observations (see Subsection 2 of Yan et al. (2008)). The proportions $p_k^{(0)}$ are all set as $1/K$. The initial values for $\boldsymbol{\theta}$ are then $\boldsymbol{\theta}^{(0)} = ((\mathbf{a}_1')^{(0)}, b_1^{(0)}, (\sigma_1^2)^{(0)}, \ldots, (\mathbf{a}_K')^{(0)}, b_K^{(0)}, (\sigma_K^2)^{(0)}, p_1^{(0)}, \ldots, p_K^{(0)})'$. We then monitor the evolution of cluster indicators $\mathbf{z}_1, \ldots, \mathbf{z}_n$. If the number of points in one cluster has fallen below $d + 1$, the sampler will get stuck and never recover from it; we then re-initialize the sampler with the above strategy.

With the above initialization strategy, we actually impose a constraint on the distribution of cluster labels such that any allocation leading to less than $d + 1$ points for a cluster is not allowed, i.e., has a probability of zero. As a result, we can then specify the prior distributions of component parameters with vague priors. Specifically, we set the priors as follows.

$$a_k \sim N(0, 100000^2), \pi(b_k) \propto 1, \sigma_k^2 \sim \text{IG}(0.0001, 0.0001), p_k \propto 1, \text{for } k = 1, 2.$$

To alleviate the effect of autocorrelations, we keep only one simulated point for every ten iterations. Hereafter the iteration numbers are based on the "thinned" samples.

Figure 4.4 displays the evolution of sampled values of $\boldsymbol{\theta}$ for $11,000$ iterations. Figure 4.5 displays the evolution of log unnormalized posterior density values of sampled values for ($\boldsymbol{\theta}$, $\mathbf{z}_1$, ..., $\mathbf{z}_n$) for $11,000$ iterations.

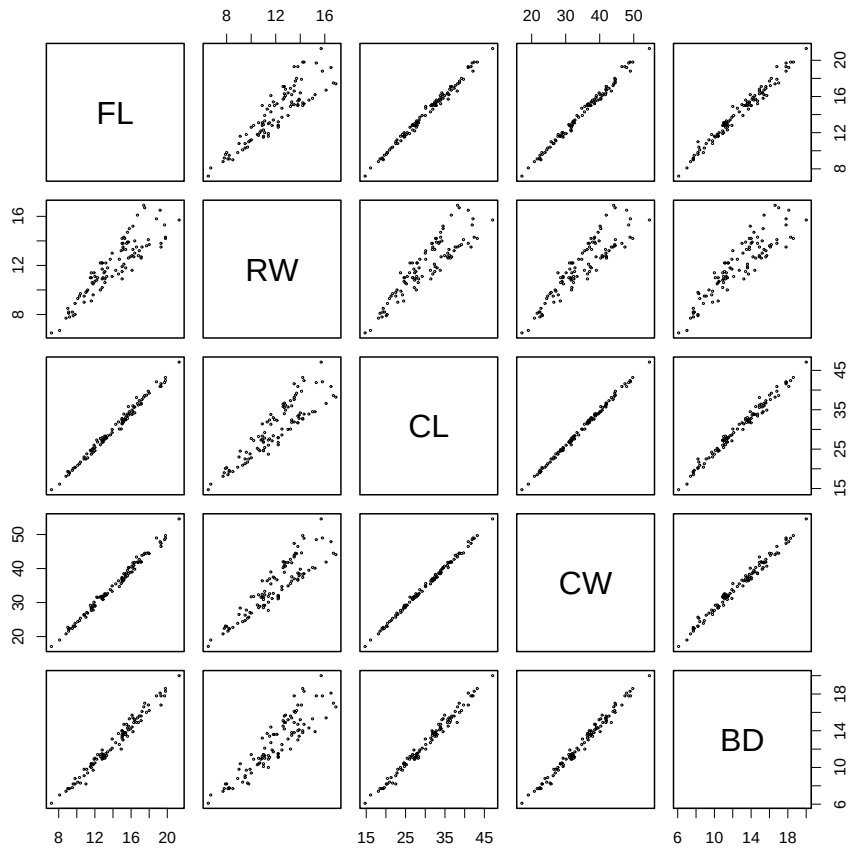

Figure 4.3: Pairwise scatter plots of the blue crab data.

From these two figures we found no reason to suspect that the sampler has reached a stationary distribution. Figure 4.6 displays the autocorrelations of sampled values for $\boldsymbol{\theta}$ for $11,000$ iterations. The autocorrelations decay fast, especially for variances σ_1^2 and σ_2^2 and proportions p_1 and p_2.

Further checking the density curves in Figure 4.7 of sampled values for $\boldsymbol{\theta}$ for $10,000$ iterations after a burn-in of 1000 iterations, we observe that all the estimated density curves are roughly unimodal. There is no obvious reason for the existence of minor modes. In Figure 4.8, we demonstrate the isolation of the two modes of the marginal posterior distribution of $\boldsymbol{\theta}$. One mode is estimated by maximum *a posteriori* estimation; the other one is obtained by permutation of cluster indices. The horizontal axis t ranges from 0 to 1, the vertical axis is the unnormalized marginal posterior density value of $(1-t)\hat{\boldsymbol{\theta}}_1 + t\hat{\boldsymbol{\theta}}_2$, where $\hat{\boldsymbol{\theta}}_1$ and $\hat{\boldsymbol{\theta}}_2$ are the two estimated modes. From this plot, we can see that the two modes are very peaky and isolated. It is then understandable that label-switching does not occur. While Gibbs sampling is generally criticized for not traversing the whole support of the posterior distribution for $\boldsymbol{\theta}$, it is reasonable to think that it has fully explored the support around one isolated mode in this data set.

Figure 4.9 shows the posterior probabilities of being classified into Class 1 of the 100 crabs estimated from $10,000$ iterations using the Gibbs sampler Algorithm 1, after a burn-in of 1000 iterations. If we were to classify all the crabs, i.e., classifying a crab to the class with larger posterior probability, five crabs are misclassified, as indicated in the upper-left corner of Figure 4.9. The number of misclassification is the same as that of the partial likelihood approach in Chapter 3; this result is expected as we used essentially noninformative priors in this example. When informative priors are necessary, as in our next example, the Bayesian approach does have an advantage.

We have also run the adaptive Metropolis-Hastings algorithm for the simulation of marginal posterior distribution for $\boldsymbol{\theta}$ and get very similar results. It runs much slower than the Gibbs sampler though.

4.5 Taqman single nucleotide polymorphism genotyping data

Now we analyze the SNP genotyping data in Figure 4.1 using the linear clustering strategy. We fit a mixture of five components to the SNP genotyping data, to represent the three genotype clusters, the negative controls, and the failed samples, respectively. The three genotype components are modelled

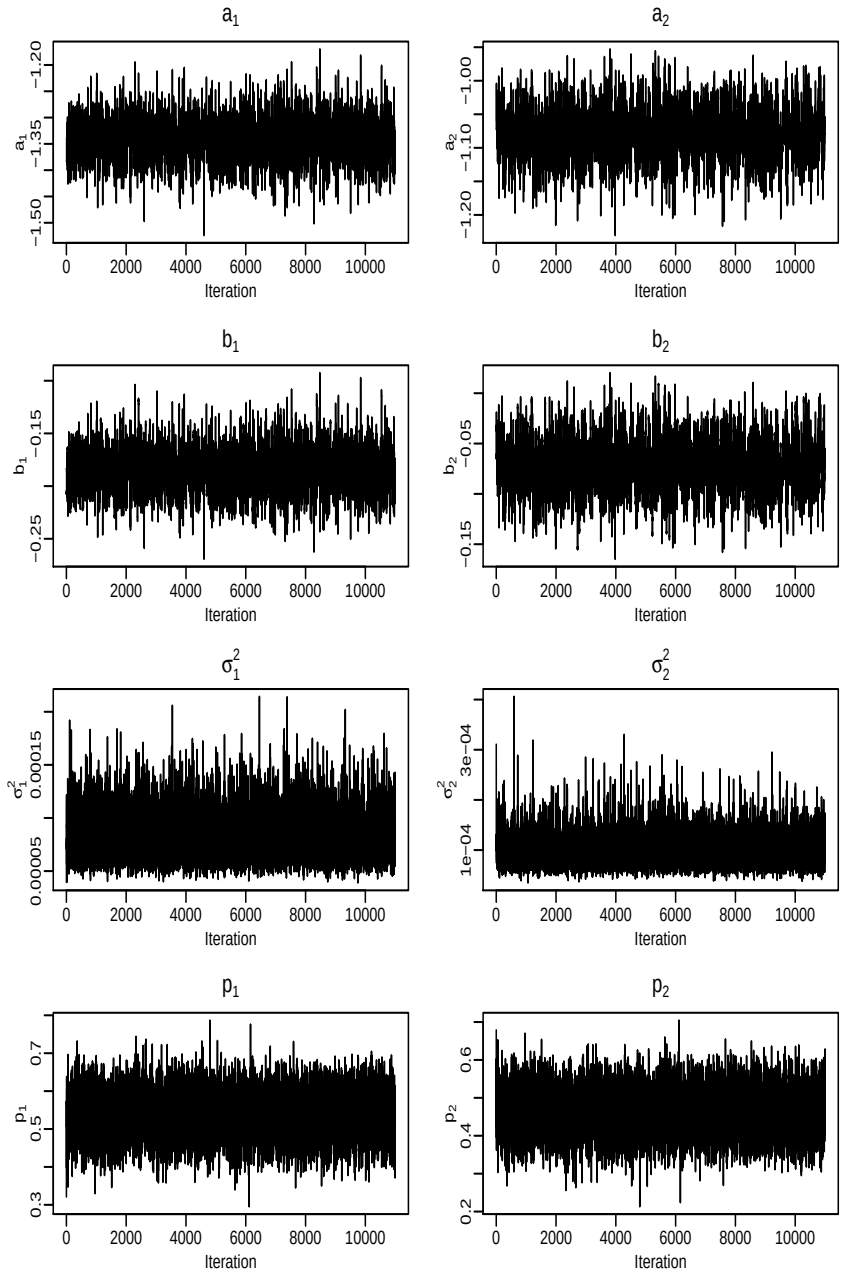

Figure 4.4: Evolution of sampled values for $\boldsymbol{\theta}$ for $11,000$ iterations using Algorithm 1 for blue crab data.

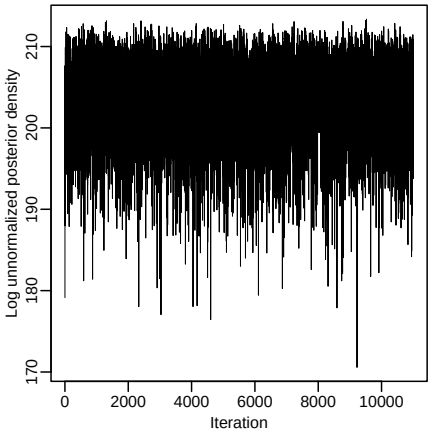

Figure 4.5: Evolution of log unnormalized posterior density of sampled values for ($\boldsymbol{\theta}$, $\mathbf{z}_1$, ..., $\mathbf{z}_n$) for 11,000 iterations using Algorithm 1 for blue crab data.

linearly. The fourth component is modelled with a bivariate distribution with diagonal covariance matrix, and the fifth is a uniform component for possible outlying points.

In the model fitting, we use the known labels for the negative controls. All the negative controls are assigned to the fourth component *a priori*, as are all points with both signals less than the corresponding maximum intensities of negative controls. Denote the set of these points by N.

Hence, the modified mixture likelihood for the problem is

$$\begin{aligned} & L(\boldsymbol{\theta}|\mathbf{x}_1, \ldots, \mathbf{x}_n) \\ = & \prod_{i \notin N} \left\{ \sum_{k=1}^{3} p_k \frac{1}{\sigma_k} t_2\left(\frac{x_{i1} + a_k x_{i2}}{\sigma_k \sqrt{1 + a_k^2}} - \frac{b_k}{\sigma_k}\right) + p_4 t_2(\mathbf{x}_i; \boldsymbol{\mu}, \Sigma) + p_5 \mathrm{U}(\mathbf{x}_i; R) \right\} \\ \times & \prod_{i \in N} t_2(\mathbf{x}_i; \boldsymbol{\mu}, \Sigma), \end{aligned} \tag{4.25}$$

where t_2 is the density for Student's t distribution with 2 degrees of freedom, $\Sigma = \begin{bmatrix} \sigma_{11} & 0 \\ 0 & \sigma_{22} \end{bmatrix}$ and U is the uniform distribution on the rectangular region R given by $(\min(x_{i1}), \max(x_{i1})) \times (\min(x_{i2}), \max(x_{i2}))$. For robustness consideration, we use Student's t distribution for orthogonal deviation of a point from a line and also for the elliptical distribution for negative control

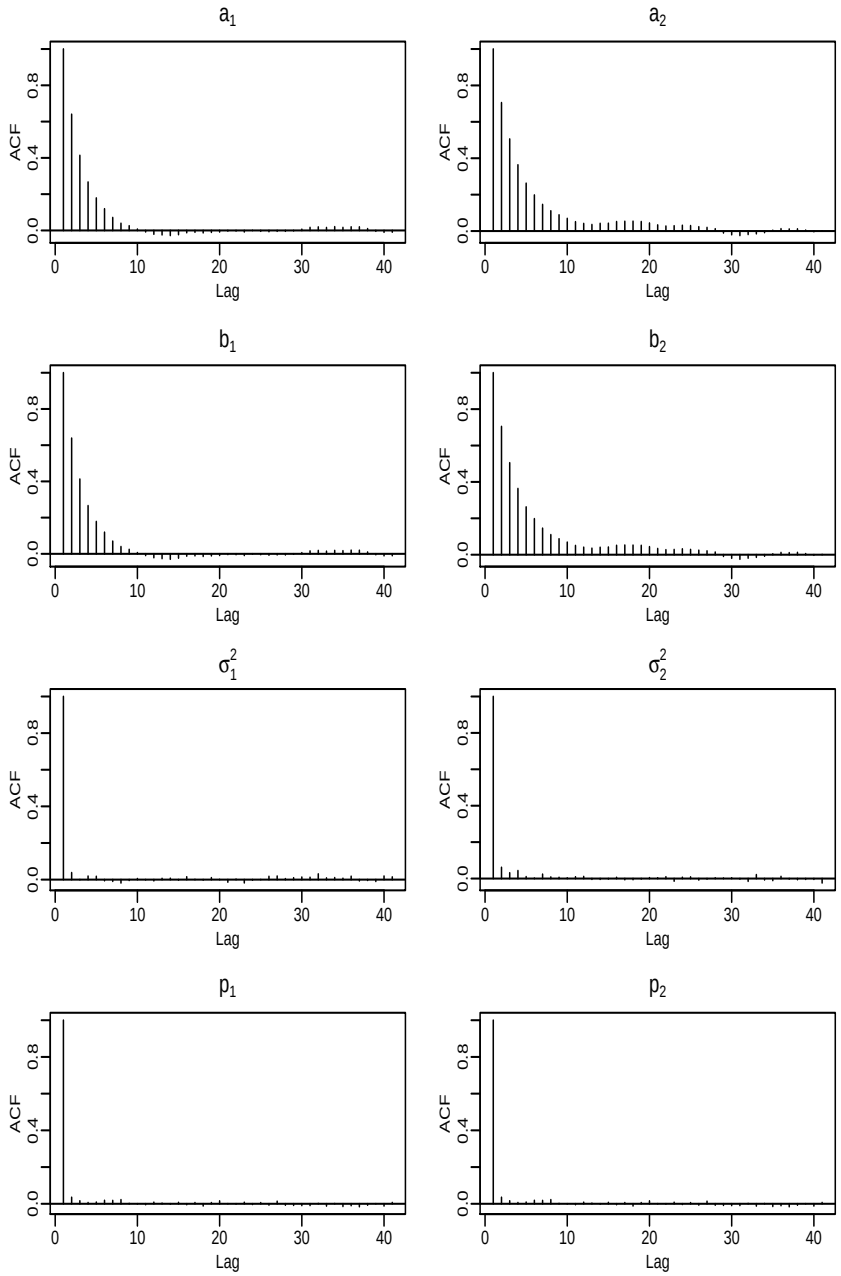

Figure 4.6: Autocorrelations of sampled values for $\boldsymbol{\theta}$ for $11,000$ iterations using Algorithm 1 for blue crab data.

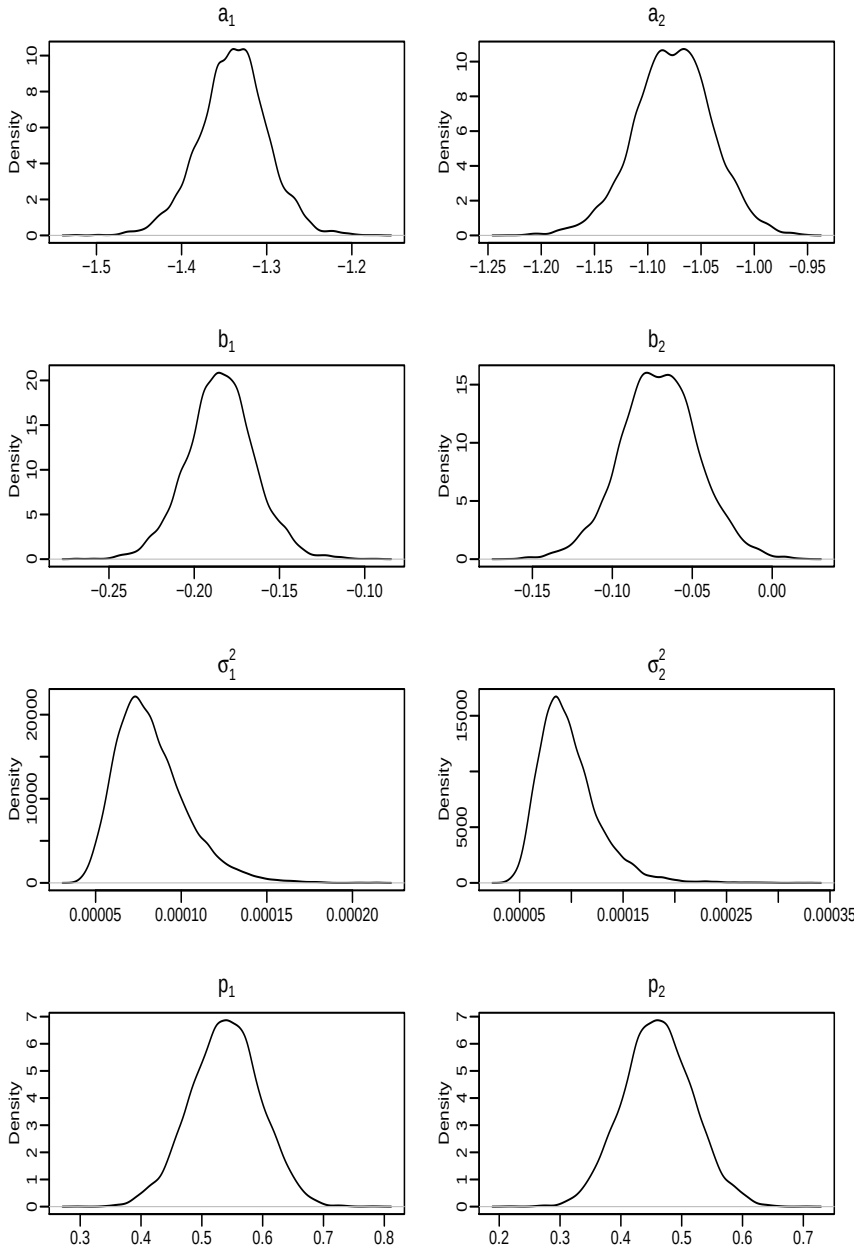

Figure 4.7: Density curves of sampled values for $\boldsymbol{\theta}$ for $10,000$ iterations using Algorithm 1, after a burn-in of 1000 iterations for blue crab data.

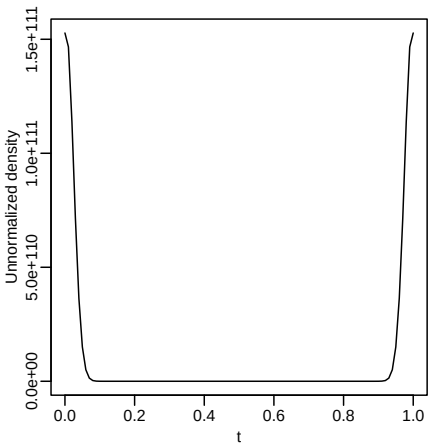

Figure 4.8: Unnormalized marginal posterior density values of $\boldsymbol{\theta}$ along a line segment $(1-t)\hat{\boldsymbol{\theta}}_1 + \hat{\boldsymbol{\theta}}_2$ connecting the two estimated modes $\hat{\boldsymbol{\theta}}_1$ and $\hat{\boldsymbol{\theta}}_2$. One mode is estimated by maximum *a posteriori* estimation from $10,000$ iterations using Algorithm 1, after a burn-in of 1000 iterations and the other is obtained by permutation. The blue crab data is used.

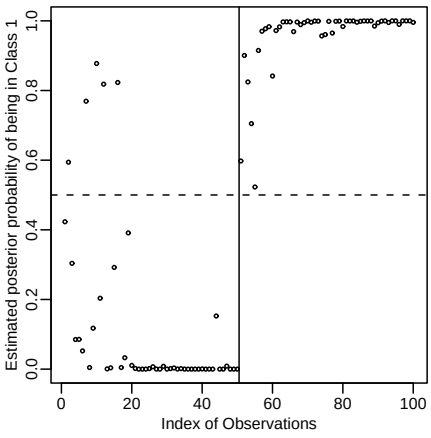

Figure 4.9: Estimated posterior probability of being classified into Class 1 of the 100 crabs from $10,000$ iterations using Algorithm 1, after a burn-in of 1000 iterations. The solid vertical line separates these crabs by their true classes and the horizontal dashed line corresponds to a probability of 0.5.

points. This mix of linear, elliptical and uniform components demonstrates the flexibility of our modelling approach.

Based on the subject matter knowledge, we add various constraints in the specification of priors. For the slopes $-1/a_k$ of lines,

$$0 < -\frac{1}{a_1} < -\frac{1}{a_2} < -\frac{1}{a_3}, \quad -\frac{1}{a_3} + \frac{1}{a_2} > -\frac{1}{a_2} + \frac{1}{a_1}, \tag{4.26}$$

which maintains the order of positive slopes of the three genotype clusters and requires that the "gap" between the two lines of the heterozygotic clusters and the variant allele homozygotic clusters cannot be too small relative to the "gap" between the lines of the wild type allele homozygotic clusters and the heterozygotic clusters. These constraints on slopes are natural for SNP genotyping data. To implement Algorithm 2, we reparameterize (a_1, a_2, a_3) into (η_1, η_2, η_3),

$$\begin{cases} a_1 &= -1/\exp(\eta_1), \\ a_2 &= -1/(\exp(\eta_1) + \exp(\eta_2)), \\ a_3 &= -1/(\exp(\eta_1) + 2\exp(\eta_2) + \exp(\eta_3)). \end{cases} \tag{4.27}$$

It is obvious that the constraint (4.26) is satisfied.

In the case of few points in the variant allele homozygotic cluster, the orthogonal variance σ_3^2 may not be estimated reliably. Hence we require

$$\sigma_3^2 = \sigma_1^2, \tag{4.28}$$

and then apply a log transformation to (σ_1^2, σ_2^2),

$$\sigma_1^2 = \exp(\eta_7), \quad \sigma_2^2 = \exp(\eta_8), \tag{4.29}$$

For variances $(\sigma_{11}, \sigma_{22})$, a log transformation is applied,

$$\sigma_{11} = \exp(\eta_{11}), \quad \sigma_{22} = \exp(\eta_{12}). \tag{4.30}$$

For the proportions, we add the constraint

$$p_3 < p_1, \quad p_3 < p_2, \tag{4.31}$$

which requires the proportion of variant allele homozygotic points is less than that of the other two genotype clusters. We reparameterize the proportions $\mathbf{p}$ first with $\mathbf{q}$ for constraint (4.31),

$$\begin{cases} p_1 &= q_1 + q_3/3, \\ p_2 &= q_2 + q_3/3, \\ p_3 &= q_3/3, \\ p_4 &= q_4, \\ p_5 &= q_5. \end{cases} \tag{4.32}$$

and then apply a logit transformation to $\mathbf{q}$,

$$\begin{cases} q_1 = \exp(\eta_{13})/(1 + (\exp(\eta_{13}) + \ldots + \exp(\eta_{16}))), \\ \ldots, \\ q_4 = \exp(\eta_{16})/(1 + (\exp(\eta_{13}) + \ldots + \exp(\eta_{16}))). \end{cases} \tag{4.33}$$

With the constraints (4.26), (4.28) and (4.31), we can fully identify the five components of the mixture. Label-switching is effectively prevented by these informative constraints. In addition, conforming to these constraints, noninformative priors cause no problem to guarantee a proper posterior distribution.

From (4.27), (4.29), (4.30) and (4.33), we obtain a one-to-one map from

$$\theta = (a_1, a_2, a_3, b_1, b_2, b_3, \sigma_1^2, \sigma_2^2, \sigma_3^2, \mu_1, \mu_2, \sigma_{11}, \sigma_{22}, p_1, \ldots, p_5)$$

to

$$\eta = (\eta_1, \eta_2, \eta_3, b_1, b_2, b_3, \eta_7, \eta_8, \mu_1, \mu_2, \eta_{11}, \eta_{12}, \eta_{13}, \eta_{14}, \eta_{15}, \eta_{16}),$$

in which the constraints (4.26), (4.28) and (4.31) are satisfied. The absolute value of the Jacobian of the transformation is

$$\begin{aligned} |J| &= \left| \frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\eta}} \right| \\ &= |a_3^2(a_1 - a_2)(a_2 - 2a_1)/a_1| \sigma_1^2 \sigma_2^2 \sigma_{11} \sigma_{22} (p_1 - p_3)(p_2 - p_3) p_3 p_4 p_5. \end{aligned} \tag{4.34}$$

We set the priors as follows.

$$\eta_k \sim N(0, 10^{10}), \;\; b_k \sim N(0, 10^{-6}), \;\; \sigma_k^2 \sim \mathrm{IG}(0.0001, 0.0001), \text{ for } k = 1, 2, 3,$$

and

$$\pi(\mu, \sigma_{11}, \sigma_{22}) \propto \frac{1}{\sigma_{11}\sigma_{22}}, \;\; \pi(\mathbf{q}) \propto 1.$$

The priors for b_1, b_2, b_3 is strong with the implication that all three lines should roughly pass through the origin; all priors for others parameters are vague except the constraints specified above.

The five-component mixture model is applied to the plate shown in Figure 4.1. For an algorithm to work in this relatively high dimensional problem, it is essential that the initial state be close to the substantial support of the posterior distribution. We first run the "optim" function in the R language multiple times, which is initialized with standard normal $N(\mathbf{0}, I_{16})$. Then we initialize the algorithm with the solution with the largest posterior density, say $\boldsymbol{\eta}^{(0)}$.

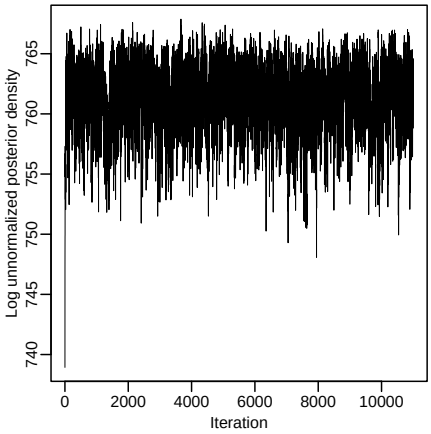

Figure 4.10: Evolution of log unnormalized posterior density of sampled values for $\boldsymbol{\eta}$ for $11,000$ iterations using Algorithm 2 for the SNP genotyping data.

We first run Algorithm 2.0, which is relatively slow, to get a rough estimate of the covariance matrix of $\boldsymbol{\eta}$. Next we run the faster Algorithm 2. To alleviate the effect of autocorrelations, we keep only one simulated point for every ten iterations. Hereafter the iteration numbers are based on the "thinned" samples.

Figure 4.10 displays the evolution of log unnormalized posterior density values of sampled values for $\boldsymbol{\eta}$ for $11,000$ iterations. Figure 4.11 displays the evolution of sampled values of $\boldsymbol{\theta}$ for $11,000$ iterations. The first three panels are for slopes of the three lines $-1/a_1$, $-1/a_2$ and $-1/a_3$. From these two figures we conclude empirically that the sampler has reached a stationary distribution. Figure 4.12 displays the autocorrelations of sampled values for $\boldsymbol{\theta}$ for $11,000$ iterations. There are high autocorrelations in most parameters indicating that the sampler is not very efficient; more efficient transformations/samplers shall be investigated in future research.

Further checking the density curves in Figure 4.13 of sampled values for $\boldsymbol{\theta}$ for $11,000$ iterations, we observe that all the estimated density curves are roughly unimodal.

The clustering results of this plate is displayed in Figure 4.14 where each point is classified into the cluster with the largest posterior membership probability. We note that the Bayesian method does identify several points into the variant allele homozygous genotype clusters (represented by "4" in Figure 4.14). Quite a few points are classified as background noise

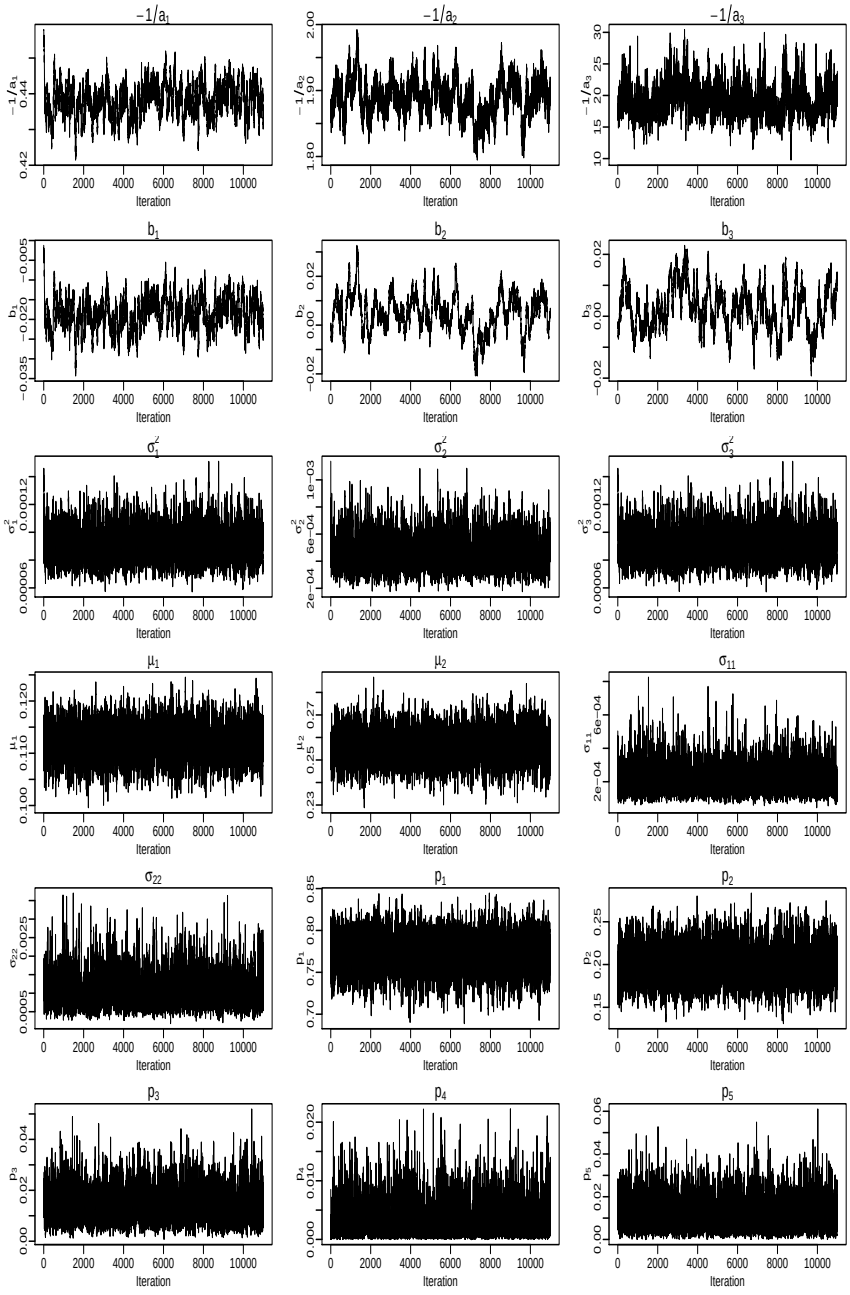

Figure 4.11: Evolution of sampled values for $\boldsymbol{\theta}$ for $11,000$ iterations using Algorithm 2 for the SNP genotyping data. The first three panels are for slopes.

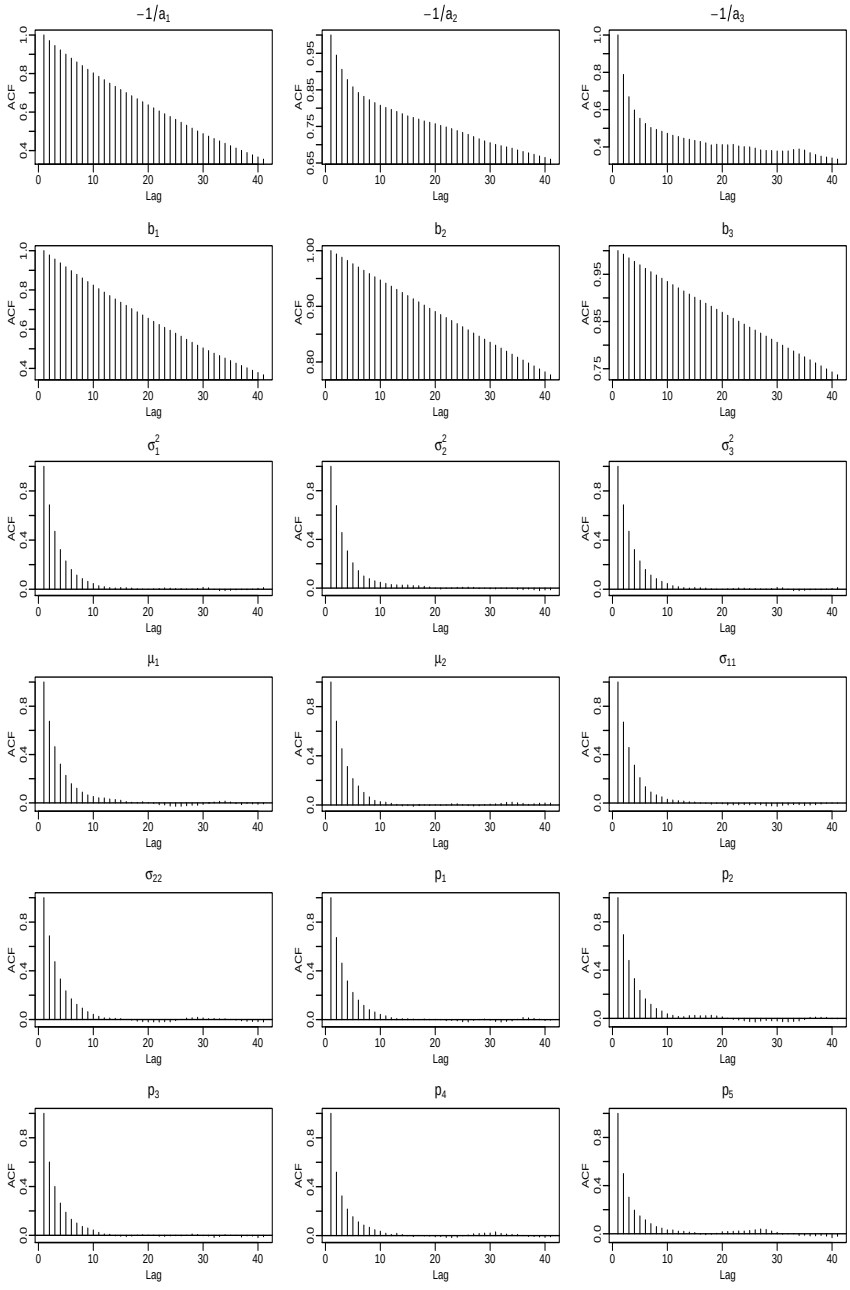

Figure 4.12: Autocorrelations of sampled values for $\boldsymbol{\theta}$ for $11,000$ iterations using Algorithm 2 for the SNP genotyping data.

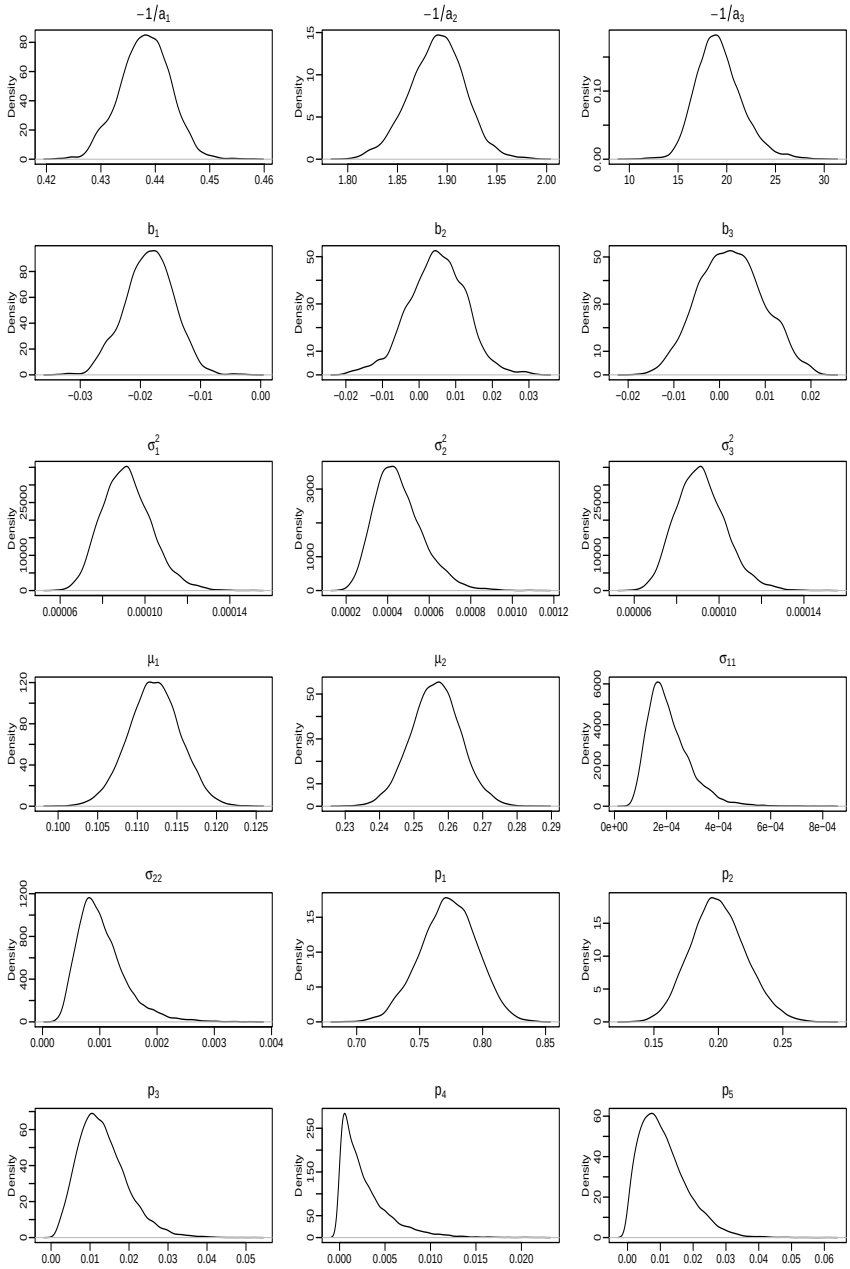

Figure 4.13: Density curves of sampled values for $\boldsymbol{\theta}$ for $11,000$ iterations using Algorithm 2 for the SNP genotyping data.

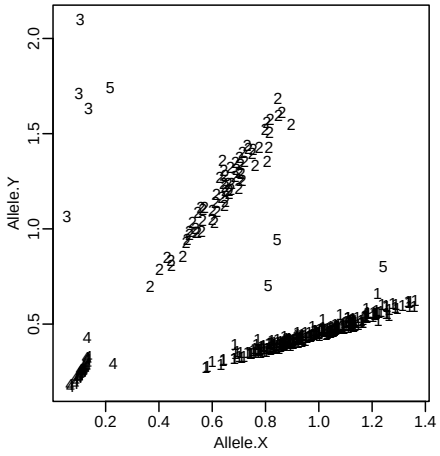

Figure 4.14: Clustering results of the SNP genotyping data in which points are classified into the cluster with the largest posterior membership probability.

(represented by "5"); this is conservative because the orthogonal variation in the wild-type allele homozygous genotype cluster XX is very small and the orthogonal variation of cluster YY is linked to that of cluster XX.

By depleting points in the YY cluster, we have observed that the Bayesian method also works if only one point is present in the YY cluster, which is due to the informative priors and constraints. Some other plates with sparse clusters are also analyzed using the Bayesian approach and the clustering results are satisfactory; the clustering results are omitted. For plates without sparse clusters, the effect of the above prior specification is minimal, we usually obtain similar clustering results to that of the partial likelihood approach in Chapter 3. Figure 4.15 shows the clustering results of the Bayesian approach for the plate analyzed in Chapter 3.

4.6 Discussion

We have proposed a Bayesian approach to linear clustering. It is flexible in the sense that only the deviation from a hyperplane is modelled parametrically; the position on the hyperplane is not modelled. The advantage of this approach is illustrated in the genotyping example, where the distribution along the line is complex and difficult to model. Furthermore, as was also illustrated in the SNP genotyping, we can incorporate elliptical clusters as

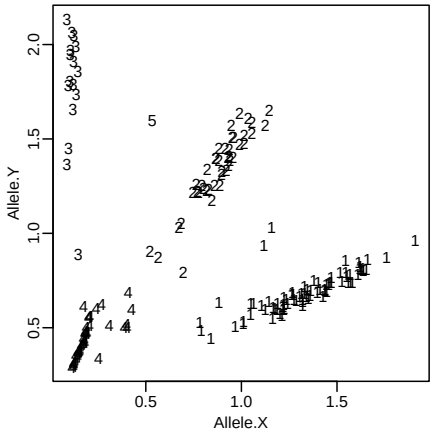

Figure 4.15: Clustering results of the Bayesian approach for a plate without a sparse cluster.

necessary and a background cluster, borrowing from standard model-based clustering.

In our examples, label-switching is prevented either by a Gibbs sampler applied to a posterior distribution with isolated modes or by informative priors. In more general situations, we may need the ideas of tempering MCMC or Sequential Monte Carlo to explore the whole support of the posterior distribution and to deal with the label-switching problem.

In this paper, the number of linear clusters are assumed known. In the situation of unknown number of clusters, our first thought is to investigate the feasibility of the Reversible Jump MCMC of Richardson and Green (1997). This may imply heavy computational burden. A related problem is the scalability of the Bayesian approach to large datasets and high dimensions. We leave these problems for further research.

Bibliography

Anderson, T. (2003). *An introduction to multivariate statistical analysis.* Wiley Series in Probability and Statistics.

Campbell, N. and R. Mahon (1974). A multivariate study of variation in two species of rock crab of genus Leptograpsus. *Australian Journal of Zoology 22*(3), 417–425.

Celeux, G., M. Hurn, and C. Robert (2000). Computational and Inferential Difficulties with Mixture Posterior Distributions. *Journal of the American Statistical Association 95*(451).

Fraley, C. and R. AE (2006). mclust Version 3 for R: Normal Mixture Modeling and Model-based Clustering. Technical report, Technical Report 504, University of Washington, Department of Statistics.

Fujisawa, H., S. Eguchi, M. Ushijima, S. Miyata, Y. Miki, T. Muto, and M. Matsuura (2004). Genotyping of single nucleotide polymorphism using model-based clustering. *Bioinformatics 20*, 5.

Fuller, W. (1987). Measurement error models. *Wiley Series in Probability and Mathematical Statistics, New York: Wiley, 1987*.

Gelman, A., J. Carlin, H. Stern, and D. Rubin (2004). *Bayesian data analysis*. Boca Raton, FL: Chapman and Hall/CRC.

Harrington, J. (2007). *lga: Tools for linear grouping analysis (LGA)*. R package version 1.0-0.

Kang, H., Z. Qin, T. Niu, and J. Liu (2004). Incorporating Genotyping Uncertainty in Haplotype Inference for Single-Nucleotide Polymorphisms. *The American Journal of Human Genetics 74*(3), 495–510.

Olivier, M., L. Chuang, M. Chang, Y. Chen, D. Pei, K. Ranade, A. de Witte, J. Allen, N. Tran, D. Curb, et al. (2002). High-throughput genotyping of single nucleotide polymorphisms using new biplex invader technology. *Nucleic Acids Research 30*(12), e53.

Peel, D. and G. McLachlan (2000). Robust mixture modelling using the t distribution. *Statistics and Computing 10*(4), 339–348.

Ranade, K., M. Chang, C. Ting, D. Pei, C. Hsiao, M. Olivier, R. Pesich, J. Hebert, Y. Chen, V. Dzau, et al. (2001). High-Throughput Genotyping with Single Nucleotide Polymorphisms. *Genome Research 11*(7), 1262–1268.

Richardson, S. and P. Green (1997). On Bayesian Analysis of Mixtures with an Unknown Number of Components. *Journal of the Royal Statistical Society. Series B (Methodological) 59*(4), 731–792.

Turner, T. (2006). *mixreg: Functions to fit mixtures of regressions.* R package version 0.0-2.

Van Aelst, S., X. Wang, R. Zamar, and R. Zhu (2006). Linear grouping using orthogonal regression. *Computational Statistics and Data Analysis 50*(5), 1287–1312.

Yan, G., W. Welch, and R. Zamar (2008). A likelihood approach to linear clustering. Preprint.

Chapter 5

Consistency and asymptotic normality in a partial likelihood approach to linear clustering

5.1 Introduction

By linear clustering, we mean detecting linearly shaped clusters in a data set. We proposed in Yan et al. (2008) a parsimonious partial likelihood approach to linear clustering. Assume that the data $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are drawn by a random mechanism, represented by random vector $\mathbf{X}$ in a d-dimensional space which we do not model fully. Assume that The data lie around K hyperplanes $\{\mathbf{x} : \mathbf{a}_k'\mathbf{x} = b_k\}$, $k = 1, \ldots, K$. Let $\mathbf{Z} = (Z_1, \ldots, Z_K)'$ be a random vector indicating these hyperplanes and $Z_k = 1$ with probability p_k for $k = 1, \ldots, K$. Let $\mathbf{p} = (p_1, \ldots, p_K)'$. We assume that, conditional on $Z_k = 1$,

$$\mathbf{a}_k'\mathbf{X} - b_k \sim N(0, \sigma_k^2), \quad k = 1, \ldots, K.$$

Let $\mathbf{z}_1, \ldots, \mathbf{z}_n$ be the corresponding unobservable indicators for the data $\mathbf{x}_1, \ldots, \mathbf{x}_n$. Let $\boldsymbol{\kappa}$ be the collection of component parameters, $\boldsymbol{\kappa} = (\mathbf{a}_1', b_1, \sigma_1^2, \ldots, \mathbf{a}_K', b_K, \sigma_K^2)'$, and $\boldsymbol{\theta} = (\boldsymbol{\kappa}', \mathbf{p}')'$. The indicators $\mathbf{z}_1, \ldots, \mathbf{z}_n$ can be regarded as realizations of random vectors $\mathbf{Z}_1, \ldots, \mathbf{Z}_n$, which in turn are an independent and identically distributed sample from $\mathbf{Z}$. After integrating out the indicators $\mathbf{z}_1, \ldots, \mathbf{z}_n$, the partial likelihood function for parameters

A version of this chapter will be submitted for publication. Authors: Guohua Yan, William J. Welch and Ruben H. Zamar.

$\boldsymbol{\theta}$ is

$$L(\boldsymbol{\theta}|\mathbf{x}_1,\ldots,\mathbf{x}_n)=\prod_{i=1}^{n}\sum_{k=1}^{K}p_kN(\mathbf{a}_k'\mathbf{x}_i-b_k;0,\sigma_k^2). \tag{5.1}$$

As in the usual normal mixture model, the partial likelihood function (5.1) is unbounded: when a cluster consists only of points lying on a hyperplane, the contribution of each of these points to the partial likelihood tends to infinity. The infinity occurs on the boundary of the parameter space. We adopt the constraint of Hathaway (1985). Specifically, the following constraint is imposed on the standard deviations,

$$\min_{1\le i\ne j\le K}(\sigma_i/\sigma_j)\ge c>0, \tag{5.2}$$

where c is a known constant determined *a priori*. In this article, constraint (5.2) is assumed whenever we refer to the partial likelihood function (5.1).

The partial likelihood function (5.1) naturally brings the clustering problem into a finite mixture model framework. An EM algorithm can be used to maximize (5.1); once an maximum partial likelihood estimate $\hat{\boldsymbol{\theta}}$ is obtained, data point $\mathbf{x}_i$ can be assigned to the component with the largest posterior probability. The probabilities are given by

$$\hat{w}_{ik}=\frac{\hat{p}_kN(\hat{\mathbf{a}}_k'\mathbf{x}_i-\hat{b}_k;0,\hat{\sigma}_k^2)}{\sum_{k=1}^{K}\hat{p}_kN(\hat{\mathbf{a}}_k'\mathbf{x}_i-\hat{b}_k;0,\hat{\sigma}_k^2)},\ i=1,\ldots,n;\ k=1,\ldots,K, \tag{5.3}$$

which also serve as a measure of uncertainty of classifying data point $\mathbf{x}_i$.

We shall investigate the asymptotic properties of solutions to the partial likelihood function (5.1). As (5.1), regarded as a function of $\mathbf{x}$, is not a density function for $\mathbf{X}$, classical results of asymptotics on maximum likelihood estimators cannot be used directly. We borrow ideas in the formulation and in the proofs of results from García-Escudero et al. (2007), Wald (1949), Redner (1981) and Hathaway (1983, 1985).

The rest of this article is organized as follows. Section 5.2 discusses the population counterpart of the partial likelihood function (5.1) and establishes the existence of its maximum. Section 5.3 proves the consistency of a maximum of the partial likelihood function (5.1) to the maximum of the population counterpart. The asymptotic normality of a solution of the partial likelihood function (5.1) is investigated in section 5.4.

5.2 Population version of the objective function

To motivate the population version of the objective function (5.1), we first review the connection between maximum likelihood estimation and the Kullback-Leibler divergence. See for example Kullback and Leibler (1951) and Eguchi and Copas (2006).

Suppose that probability distributions P and Q are absolutely continuous with respect to the Lebesgue measure λ. Let $p(x)$ and $q(x)$ be the density functions (Radon-Nikodym derivatives) of P and Q respectively with respect to λ. Then the Kullback-Leibler divergence from P to Q is defined as

$$D_{\mathrm{KL}}(P\|Q) = \int \log \frac{p(x)}{q(x)} dP(x).$$

Given a random sample $x_1, x_2, \ldots, x_n$ from the underlying distribution P, let P_n be the empirical distribution. Now let Q_θ be a statistical model having density $f(x;\theta)$ with respect to λ, where θ is a collection of unknown parameters. The Kullback-Leibler divergence from P to Q_θ is

$$D_{\mathrm{KL}}(P\|Q_\theta) = \int [\log p(x) - \log f(x;\theta)] dP(x).$$

The empirical version of $D_{\mathrm{KL}}(P||Q_\theta)$ is

$$\begin{aligned} D_{\mathrm{KL}}(P_n\|Q_\theta) &= \frac{1}{n}\sum_{i=1}^{n} [\log(1/n) - \log f(x_i;\theta)] \\ &= \log(1/n) - \frac{1}{n}\sum_{i=1}^{n} \log f(x_i;\theta). \end{aligned}$$

Apart from the factor $1/n$, the second term is just the log likelihood function. Hence, maximizing the likelihood function is equivalent to minimizing the Kullback-Leibler divergence $D_{\mathrm{KL}}(P_n\|Q_\theta)$. If $P = Q_{\theta_0}$ for some θ_0, then a maximum likelihood estimator is strongly consistent under some regular conditions, i.e., it converges almost surely to $\arg\min_\theta D_{\mathrm{KL}}(P\|Q_\theta)$.

Back to the linear clustering setting, let

$$f(\mathbf{x};\boldsymbol{\theta}) = \sum_{k=1}^{K} p_k f_k(\mathbf{x};\boldsymbol{\theta}), \tag{5.4}$$

where

$$f_k(\mathbf{x};\boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{(\mathbf{a}_k'\mathbf{x} - b_k)^2}{2\sigma_k^2}\right),$$

and $\mathbf{x}$ is a generic point in $\Re^d$. We still use P to denote the underlying distribution of the data and use $p(\mathbf{x})$ to denote the density of P with respect to the Lebesgue measure λ^d. Although $f(\mathbf{x};\boldsymbol{\theta})$ cannot be a density as a function of $\mathbf{x}$, we can nevertheless define

$$D_{\mathrm{KL}}(P\|f(\cdot;\boldsymbol{\theta})) = \int [\log p(\mathbf{x}) - \log f(\mathbf{x};\boldsymbol{\theta})] dP(\mathbf{x}),$$

or equivalently define

$$g(\boldsymbol{\theta}) = \int \log f(\mathbf{x};\boldsymbol{\theta}) dP(\mathbf{x}), \tag{5.5}$$

and show that the maximum partial likelihood solution of (5.1) converges to the set

$$\{\boldsymbol{\theta}_0 : g(\boldsymbol{\theta}_0) = \max_{\boldsymbol{\theta}} g(\boldsymbol{\theta})\}.$$

With constraint (5.2), The parameter space is

$$\Theta_c = \left\{ \begin{array}{l} \boldsymbol{\theta} = (\mathbf{a}_1, b_1, \sigma_1^2, \ldots, \mathbf{a}_K, b_K, \sigma_K^2, p_1, \ldots, p_K) : \\ 0 < p_k < 1, \mathbf{a}_k'\mathbf{a}_k = 1, -\infty < b_k < \infty, \sigma_k > 0, k = 1, \ldots, K. \\ \sum_{k=1}^K p_k = 1, \min_{i,j} \sigma_i/\sigma_j \geq c > 0. \end{array} \right\}.$$

We show in this section that the supremum of g is finite and attainable. First we prove the following lemma.

Lemma 5.1 *Let P be an absolutely continuous probability measure with finite second moments. Let $0 < c < 1$, if $c \leq \sigma_k \leq 1/c$, for all $k = 1, \ldots, K$. Let*

$$s(\boldsymbol{\kappa}) \equiv \int \min_k \frac{(\mathbf{a}_k'\mathbf{x} - b_k)^2}{\sigma_k^2} dP(\mathbf{x}).$$

Then $s(\boldsymbol{\kappa})$ attains its infimum $s_0 > 0$ under constraint (5.2) at some $\boldsymbol{\kappa}_0$.

Proof It is obvious that $0 < s(\boldsymbol{\kappa}) < \infty$. First we show that it is possible to bound b_k, $k = 1, \ldots, K$ and the infimum is not missed.

Suppose that there is a positive number $r < |b_k|$, for $k = 1, \ldots, K$. Let $r_0 > 0$ such that $\mathrm{Prob}(|\mathbf{X}| \leq r_0) > 0$. Then when $r > r_0$,

$$s(\boldsymbol{\theta}) \geq c^2(r - r_0)^2 \mathrm{Prob}(|\mathbf{X}| \leq r_0) \to \infty, \quad \text{as } r \to \infty.$$

Now without loss of generality, we assume that $|b_k| \leq v$ for some v and for $k = 1, \ldots, K-1$. Let $r = c^2(|b_K| - v)$. Then when $|b_K|$ is getting large,

$$\begin{aligned} s(\boldsymbol{\kappa}) &\geq \int_{|\mathbf{x}|\leq r} \min_{1\leq k\leq K} \frac{(\mathbf{a}_k'\mathbf{x} - b_k)^2}{\sigma_k^2} dP(\mathbf{x}) \\ &= \int_{|\mathbf{x}|\leq r} \min_{1\leq k\leq K-1} \frac{(\mathbf{a}_k'\mathbf{x} - b_k)^2}{\sigma_k^2} dP(\mathbf{x}) \\ &\to \int \min_{1\leq k\leq K-1} \frac{(\mathbf{a}_k'\mathbf{x} - b_k)^2}{\sigma_k^2} dP(\mathbf{x}), \qquad \text{as } |b_K| \to \infty \\ &\geq \int \min_{1\leq k\leq K} \frac{(\mathbf{a}_k'\mathbf{x} - b_k)^2}{\sigma_k^2} dP(\mathbf{x}), \end{aligned}$$

where the $|b_K|$ in the last term is arbitrary number bounded from below by v.

Since

$$\min_k \frac{(\mathbf{a}_k'\mathbf{x} - b_k)^2}{\sigma_k^2} \leq \frac{2(|\mathbf{x}|^2 + v^2)}{c^2},$$

by Lebesgue's dominated convergence theorem, s is continuous. Now it is constrained on a compact set, the infimum of s is attainable for some $\boldsymbol{\kappa}_0$ as $s_0 = s(\boldsymbol{\kappa}_0) > 0$. ∎

Theorem 5.1 *Let P be an absolutely continuous probability measure with finite second moments. Then the supremum of g, which is defined in (5.5), over Θ_c is finite and attainable.*

Proof Let $\sigma_k^2 = \tau_k^2\sigma^2$ for $k = 1, \ldots, K$. For every $\mathbf{x}$ and $\boldsymbol{\theta}$, we have

$$\log f(\mathbf{x}, \boldsymbol{\theta}) \leq \log \max_k f_k(\mathbf{x}, \boldsymbol{\theta}) = \max_k \log f_k(\mathbf{x}, \boldsymbol{\theta}).$$

Therefore,

$$\begin{aligned} g(\boldsymbol{\theta}) &\leq \int \max_k \log f_k(\mathbf{x}, \boldsymbol{\theta}) dP(\mathbf{x}) \\ &= \int -\min_k \left\{ \frac{1}{2}\log(2\pi\tau_k^2\sigma^2) + \frac{(\mathbf{a}_k'\mathbf{x} - b_k)^2}{2\tau_k^2\sigma^2} \right\} dP(\mathbf{x}) \\ &\leq -\frac{1}{2}\log(2\pi c^2) - \log\sigma - \frac{1}{2\sigma^2}\left\{ \int \min_k \frac{(\mathbf{a}_k'\mathbf{x} - b_k)^2}{\tau_k^2} dP(\mathbf{x}) \right\} \\ &\leq -\frac{1}{2}\log(2\pi c^2) - \log\sigma - \frac{s_0}{2\sigma^2}, \end{aligned}$$

where Lemma 5.1 is used. Clearly, the last term attains its maximum at $\sigma^2 = s_0$ and it goes to $-\infty$ as $\sigma \to 0$ or ∞. Take an arbitrary $\boldsymbol{\theta}_0$, there exist positive numbers s_1 and s_2 such that $g(\boldsymbol{\theta}) < g(\boldsymbol{\theta}_0)$ uniformly for $\boldsymbol{\theta}$ such that $\sigma < s_1/c$ or $\sigma > cs_2$.

Now if we bound $s_1/c \leq \sigma \leq cs_2$, we have

$$\begin{aligned}
\log f(\mathbf{x}, \boldsymbol{\theta}) &\geq \log \min_k f_k(\mathbf{x}, \boldsymbol{\theta}) \\
&= \min \log f_k(\mathbf{x}, \boldsymbol{\theta}) \\
&= \min(-\frac{1}{2}\log(2\pi\tau_k^2\sigma^2) - \frac{(\mathbf{a}_k'\mathbf{x} - b_k)^2}{2\tau_k^2\sigma^2}) \\
&\geq -\frac{1}{2}\log(2\pi s_2^2) - \frac{2(|\mathbf{x}|^2 + \max_k |b_k|^2)}{2s_1^2},
\end{aligned}$$

and

$$\log f(\mathbf{x}, \boldsymbol{\theta}) \leq \log \max_k f_k(\mathbf{x}, \boldsymbol{\theta}) \leq -\frac{1}{2}\log(2\pi s_1^2).$$

By Lebesgue's dominated convergence theorem, g is continuous.

If $b_k \to \pm\infty$ for some k, then by Fatou's lemma, we have

$$\limsup_{b_k \to \pm\infty} g(\boldsymbol{\theta}) \leq \int \limsup_{b_k \to \pm\infty} \log f(\mathbf{x}, \boldsymbol{\theta}) dP(\mathbf{x}) = \int \log \sum_{k' \neq k} p_{k'} f_{k'}(\mathbf{x}, \boldsymbol{\theta}) dP(\mathbf{x}).$$

Therefore, there exist real numbers t_1 and t_2 such that we would not miss the supremum of g, if we bound $\boldsymbol{\theta}$ in the compact set

$$\{\boldsymbol{\theta} : s_1 \leq \sigma_k^2 \leq s_2, t_1 \leq b_k \leq t_2, \text{ for all } k = 1, \ldots, K\}.$$

Since g is continuous, its supremum is attainable. ∎

Remark In the proof it is not ruled out that some p_k may be 0 for g to attain its maximum.

5.3 Consistency

First we prove that there exists a global maximizer of the partial likelihood function (5.1) over Θ_c.

Theorem 5.2 *Let $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ be a set of points in $\Re^d$ that cannot fit in K hyperplanes. That is, there do not exist $\{(\mathbf{a}_k, b_k) : k = 1, \ldots, K\}$ such that, for every $\mathbf{x}_i$, $\mathbf{a}_k'\mathbf{x}_i = b_k$ for some k (which may depend on i). Let*

$$l(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log f(\mathbf{x}_i, \boldsymbol{\theta}).$$

Then a constrained global maximizer of $l(\boldsymbol{\theta})$ over Θ_c exists.

Proof Let $B(0,r) = \{\mathbf{x} : |\mathbf{x}| \le r\}$ be a ball that contains all the points $\mathbf{x}_1, \ldots, \mathbf{x}_n$. For any $\boldsymbol{\theta}$, if there is some b_k such that $|b_k| > r$, then $(\mathbf{a}_k'\mathbf{x}_i - b_k)^2 \ge (\mathbf{a}_k'\mathbf{x}_i - r\text{sgn}(b_k))^2$. So $l(\boldsymbol{\theta})$ is not decreased if b_k is replaced with $r\text{sgn}(b_k)$ whenever $|b_k| \ge r$. So we can bound $\boldsymbol{\theta}$ such that $|b_k| \le r$, for all $k = 1, \ldots, K$.

Let

$$s_0 = \min_{\{(\mathbf{a}_1, b_1, \ldots, \mathbf{a}_K, b_K): \mathbf{a}_k'\mathbf{a}_k = 1, |b_k| \le r\}} \max_i \min_k (\mathbf{a}_k'\mathbf{x}_i - b_k)^2.$$

Since the points $\mathbf{x}_1, \ldots, \mathbf{x}_n$ cannot fit in K hyperplanes, we have $s_0 > 0$. There exists a point $\mathbf{x}_j$ such that $(\mathbf{a}_k'\mathbf{x}_j - b_k)^2 \ge s_0$ for all $k = 1, \ldots, K$. Note also that $1/\sigma_k \le 1/(c\sigma_1)$ and that $1/\sigma_k \ge c/\sigma_1$, we have

$$\begin{aligned} l(\boldsymbol{\theta}) &\le \sum_{i \ne j} \log(\sum_k p_k \frac{1}{\sqrt{2\pi}\sigma_k}) + \log(\sum_k p_k \frac{1}{\sqrt{2\pi}\sigma_k} \exp(-\frac{s_0}{2\sigma_k^2})) \\ &\le (n-1)\log(\sum_k p_k \frac{1}{\sqrt{2\pi}c\sigma_1}) + \log(\sum_k p_k \frac{1}{\sqrt{2\pi}c\sigma_1} \exp(-\frac{cs_0}{2\sigma_1^2})) \\ &= -n\log(\sqrt{2\pi}c\sigma_1) - \frac{cs_0}{2\sigma_1^2}. \end{aligned}$$

The last term tends to $-\infty$ as σ_1^2 tends to zero. If some σ_k^2 tends to zero, so do all other σ_k^2s. Thus, there exists a positive number s_1 such that $l(\boldsymbol{\theta}) \le l(\boldsymbol{\theta}_0)$ whenever a $\sigma_k^2 < s_1$.

On the other hand,

$$l(\boldsymbol{\theta}) \le \sum_i \log(\sum_k p_k \frac{1}{\sqrt{2\pi}\sigma_k}) \le n\log(\max_k \frac{1}{\sqrt{2\pi}\sigma_k}) = -n \min_k \log(\sqrt{2\pi}\sigma_k).$$

If some σ_k^2 tends to ∞, so do all other σ_k^2's and then $l(\boldsymbol{\theta})$ decreases to $-\infty$. Similarly, there is a positive number s_2 such that $l(\boldsymbol{\theta}) \le l(\boldsymbol{\theta}_0)$ whenever a $\sigma_k^2 > s_2$.

Therefore, we can bound $\boldsymbol{\theta}$ in a compact set

$$\{\boldsymbol{\theta} : |b_k| \le r, s_1 \le \sigma_k^2 \le s_2, \text{ for all } k = 1, \ldots, K\}.$$

As l is continuous, a global maximizer over Θ_c exists. ∎

Corollary 5.1 *Let P be an absolutely continuous probability measure with finite second moments. Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be a sample from P. Then a constrained global maximizer of $l(\boldsymbol{\theta}|\mathbf{X}_1, \ldots, \mathbf{X}_n)$ exists almost surely if $n \ge Kd + 1$.*

Proof As P is absolute continuous, $\mathbf{X}_1, \ldots, \mathbf{X}_n$ are distinct with probability 1. We need Kd points to determine K hyperplanes. Each of the remaining points lies on one of these hyperplanes with probability 0, since the set of points in K hyperplanes has zero probability. ∎

To prove the consistency result, we need some preliminary results.

Lemma 5.2 *If the probability measure P has finite second moments, then $\int |\log f(\mathbf{x}, \boldsymbol{\theta})| dP(\mathbf{x}) < \infty$.*

Proof

$$
\begin{aligned}
&\int |\log f(\mathbf{x}, \boldsymbol{\theta})| \, dP(\mathbf{x}) \\
=&\int [\log f(\mathbf{x}, \boldsymbol{\theta})]^+ \, dP(\mathbf{x}) + \int [\log f(\mathbf{x}, \boldsymbol{\theta})]^- dP(\mathbf{x}) \\
\leq&\int_{\{\mathbf{x}: f(\mathbf{x}, \boldsymbol{\theta}) \geq 1\}} [\log \max_k f_k(\mathbf{x}, \boldsymbol{\theta})]^+ dP(\mathbf{x}) \\
&+ \int_{\{\mathbf{x}: f(\mathbf{x}, \boldsymbol{\theta}) < 1\}} [\log \min_k f_k(\mathbf{x}, \boldsymbol{\theta})]^- dP(\mathbf{x}) \\
\leq&\int_{\{\mathbf{x}: f(\mathbf{x}, \boldsymbol{\theta}) \geq 1\}} \sum_k [\log f_k(\mathbf{x}, \boldsymbol{\theta})]^+ dP(\mathbf{x}) \\
&+ \int_{\{\mathbf{x}: f(\mathbf{x}, \boldsymbol{\theta}) < 1\}} \sum_k [\log f_k(\mathbf{x}, \boldsymbol{\theta})]^- dP(\mathbf{x}) \\
&\\
=&\sum_k \int |\log f_k(\mathbf{x}, \boldsymbol{\theta})| dP(\mathbf{x}) \\
\leq&\sum_k [|\log(\sqrt{2\pi}\sigma_k)| + \int \frac{(\mathbf{a}_k' \mathbf{x} - b_k)^2}{2\sigma_k^2} dP(\mathbf{x})] \\
<&\infty.
\end{aligned}
$$

∎

Lemma 5.3 *Let*

$$
w(\mathbf{x}, \boldsymbol{\theta}, \rho) = \sup_{\{\boldsymbol{\theta}': |\boldsymbol{\theta}' - \boldsymbol{\theta}| \leq \rho\}} f(\mathbf{x}, \boldsymbol{\theta}').
$$

Then

$$
\int [\log w(\mathbf{x}, \boldsymbol{\theta}, \rho)]^+ dP(\mathbf{x}) < \infty.
$$

Proof In fact,

$$\begin{aligned} w(\mathbf{x}, \boldsymbol{\theta}, \rho) &\leq \sup_{\{\boldsymbol{\theta}' : |\boldsymbol{\theta}' - \boldsymbol{\theta}| \leq \rho\}} \sum_k f_k(\mathbf{x}, \boldsymbol{\theta}) \\ &\leq \sup_{\{\boldsymbol{\theta}' : |\boldsymbol{\theta}' - \boldsymbol{\theta}| \leq \rho\}} \sum_k \frac{1}{\sqrt{2\pi}\sigma_k} \\ &\equiv M(\boldsymbol{\theta}, \rho). \end{aligned}$$

Therefore,

$$\int [\log w(\mathbf{x}, \boldsymbol{\theta}, \rho)]^+ dP(\mathbf{x}) \leq \int [\log M(\boldsymbol{\theta}, \rho)]^+ dP(\mathbf{x}) < \infty.$$

∎

Lemma 5.4

$$\lim_{\rho \to 0} \int \log w(\mathbf{x}, \boldsymbol{\theta}, \rho) dP(\mathbf{x}) = \int \log f(\mathbf{x}, \boldsymbol{\theta}) dP(\mathbf{x}).$$

Proof Since $\log w(\mathbf{x}, \boldsymbol{\theta}, \cdot)$ is an increasing function of ρ, as $\rho \to 0$, $[\log w(\mathbf{x}, \boldsymbol{\theta}, \rho)]^-$ increases. By the monotone convergence theorem,

$$\lim_{\rho \to 0} \int [\log w(\mathbf{x}, \boldsymbol{\theta}, \rho)]^- dP(\mathbf{x}) = \int [\log f(\mathbf{x}, \boldsymbol{\theta})]^- dP(\mathbf{x}).$$

When ρ is sufficiently small, $[\log w(\mathbf{x}, \boldsymbol{\theta}, \rho)]^+$ is dominated by $[\log w(\mathbf{x}, \boldsymbol{\theta}, \rho_0)]^+$ for some ρ_0. And the latter is integrable by Lemma 5.3. By Lebesgue's dominated convergence theorem,

$$\lim_{\rho \to 0} \int [\log w(\mathbf{x}, \boldsymbol{\theta}, \rho)]^+ dP(\mathbf{x}) = \int [\log f(\mathbf{x}, \boldsymbol{\theta})]^+ dP(\mathbf{x}).$$

Notice that $\int [\log w(\mathbf{x}, \boldsymbol{\theta}, \rho)]^+ dP(\mathbf{x})$ and $\int [\log f(\mathbf{x}, \boldsymbol{\theta})]^+ dP(\mathbf{x})$ are finite, the lemma is proved. ∎

Note that points in Θ_c are not identifiable for $f(\mathbf{x}; \cdot)$. The function $f(\mathbf{x}; \cdot)$ remains the same if we permutate the labels 1, ..., K; p_{k_1} and p_{k_2} are not identifiable if $(\mathbf{a}_{k_1}, b_{k_1}, \sigma_{k_1}^2) = (\mathbf{a}_{k_2}, b_{k_2}, \sigma_{k_2}^2)$. Thus the consistency result is in a quotient topology space. Let $\sim$ be an equivalent relation on Θ_c such that $\boldsymbol{\theta}_1 \sim \boldsymbol{\theta}_2$ if and only if $f(\mathbf{x}; \boldsymbol{\theta}_1) = f(\mathbf{x}; \boldsymbol{\theta}_2)$ almost surely in P. Denote by Θ_c^q the quotient topological space consisting of all equivalent classes of $\sim$. For a point $\boldsymbol{\theta}_0$ that maximizes $\int \log f(\mathbf{x}; \boldsymbol{\theta}) dP(\mathbf{x})$, its equivalent class in Θ_c^q is denoted by $\boldsymbol{\theta}_0^q$.

Theorem 5.3 *Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be a sample from P which is absolutely continuous with finite second moments. Let B be a compact subset of Θ_c that contains C_0. Let $\hat{\boldsymbol{\theta}}^{(n)}$ be a global maximizer of $l(\boldsymbol{\theta}|\mathbf{X}_1, \ldots, \mathbf{X}_n)$ on B. Then $\hat{\boldsymbol{\theta}}^{(n)} \to \boldsymbol{\theta}_0^q$ almost surely in the topological space B^q.*

Proof Let ω be a closed subset of B which does not intersect with C_0. For each point $\boldsymbol{\theta}$ in ω, we associate a positive value $\rho_{\boldsymbol{\theta}}$ such that

$$E \log w(\mathbf{X}, \boldsymbol{\theta}, \rho_{\boldsymbol{\theta}}) < E \log f(\mathbf{X}, \boldsymbol{\theta}_0),$$

where $\boldsymbol{\theta}_0$ is a point in C_0. The existence of such a $\rho_{\boldsymbol{\theta}}$ follows from Lemma 5.4 and the definition of C_0 (C_0 exists because P has finite second moments, by Theorem 5.1; $\hat{\boldsymbol{\theta}}^{(n)}$ exists because P is absolutely continuous, by Theorem 5.2).

Since B is compact, its closed subset ω is also compact. There exists a finite number of points $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_h$ in ω such that

$$\omega \subset \bigcup_{j=1}^{h} \{\boldsymbol{\theta}' : |\boldsymbol{\theta}' - \boldsymbol{\theta}_j| \leq \rho_{\boldsymbol{\theta}_j}\}.$$

Then

$$0 \leq \sup_{\boldsymbol{\theta} \in \omega} f(\mathbf{x}_1, \boldsymbol{\theta}) \ldots f(\mathbf{x}_n, \boldsymbol{\theta}) \leq \sum_{j=1}^{h} w(\mathbf{x}_1, \boldsymbol{\theta}_j, \rho_{\boldsymbol{\theta}_j}) \ldots w(\mathbf{x}_n, \boldsymbol{\theta}_j, \rho_{\boldsymbol{\theta}_j}).$$

By the strong law of large numbers,

$$\text{Prob}\left\{ \lim_{n \to \infty} \sum_{i=1}^{n} [\log w(\mathbf{X}_i, \boldsymbol{\theta}_j, \rho_{\boldsymbol{\theta}_j}) - \log f(\mathbf{X}_i, \boldsymbol{\theta}_0)] = -\infty \right\} = 1,$$

for $j = 1, \ldots, h$. That is,

$$\text{Prob}\left\{ \lim_{n \to \infty} \frac{w(\mathbf{X}_1, \boldsymbol{\theta}_j, \rho_{\boldsymbol{\theta}_j}) \ldots w(\mathbf{X}_n, \boldsymbol{\theta}_j, \rho_{\boldsymbol{\theta}_j})}{f(\mathbf{X}_1, \boldsymbol{\theta}_0) \ldots f(\mathbf{X}_n, \boldsymbol{\theta}_0)} = 0 \right\} = 1,$$

for $j = 1, \ldots, h$. Therefore, we have

$$\text{Prob}\left\{ \lim_{n \to \infty} \frac{\sup_{\boldsymbol{\theta} \in \omega} f(\mathbf{X}_1, \boldsymbol{\theta}) \ldots f(\mathbf{X}_n, \boldsymbol{\theta})}{f(\mathbf{X}_1, \boldsymbol{\theta}_0) \ldots f(\mathbf{X}_n, \boldsymbol{\theta}_0)} = 0 \right\} = 1. \tag{5.6}$$

Denote $|\boldsymbol{\theta} - C_0| \equiv \min_{\boldsymbol{\theta}_0 \in C_0} |\boldsymbol{\theta} - \boldsymbol{\theta}_0|$. The minimum is attainable since C is a closed set in B and hence compact. We need only to prove that

all limit points $\boldsymbol{\theta}^*$ of the sequence $\hat{\boldsymbol{\theta}}^{(n)}$ are in C_0. If not, there exists a limit point $\boldsymbol{\theta}^*$ and an $\epsilon > 0$ such that $|\boldsymbol{\theta}^* - C_0| \geq \epsilon$. This implies that there are infinitely many $\hat{\boldsymbol{\theta}}^{(n)}$ that lie in $\omega_\epsilon \equiv \{\boldsymbol{\theta} : |\boldsymbol{\theta} - C_0| \geq \epsilon\}$. Thus $f(\mathbf{x}_1, \hat{\boldsymbol{\theta}}^{(n)}) \ldots f(\mathbf{x}_n, \hat{\boldsymbol{\theta}}^{(n)}) \leq \sup_{\boldsymbol{\theta} \in \omega_\epsilon} f(\mathbf{x}_1, \boldsymbol{\theta}) \ldots f(\mathbf{x}_n, \boldsymbol{\theta})$ for infinitely many n. Since $f(\mathbf{x}_1, \hat{\boldsymbol{\theta}}^{(n)}) \ldots f(\mathbf{x}_n, \hat{\boldsymbol{\theta}}^{(n)}) \geq f(\mathbf{x}_1, \boldsymbol{\theta}_0) \ldots f(\mathbf{x}_n, \boldsymbol{\theta}_0)$. We have

$$f(\mathbf{x}_1, \boldsymbol{\theta}_0) \ldots f(\mathbf{x}_n, \boldsymbol{\theta}_0) \leq \sup_{\boldsymbol{\theta} \in \omega_\epsilon} f(\mathbf{x}_1, \boldsymbol{\theta}) \ldots f(\mathbf{x}_n, \boldsymbol{\theta}),$$

for infinitely many n. Since ω_ϵ is a closed set in B, this is an event with probability zero according to equation 5.6. This completes the proof. ∎

In next step, we shall show that limit points of the constrained global maximizer $\hat{\boldsymbol{\theta}}^{(n)}$ over Θ_c are almost surely in a compact space and hence consistency follows from Theorem 5.3.

Lemma 5.5 *Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be a sample of P which is absolutely continuous with finite second moments. Let $\hat{\boldsymbol{\theta}}^{(n)}$ is a constrained global maximizer of $l(\boldsymbol{\theta}|\mathbf{X}_1, \ldots, \mathbf{X}_n)$ over Θ_c. Then there exist positive numbers s_1 and s_2 such that*

$$\text{Prob}\{\liminf_{n\to\infty} \min_k (\hat{\sigma}_k^{(n)})^2 \geq s_1\} = 1, \tag{5.7}$$

and

$$\text{Prob}\{\limsup_{n\to\infty} \max_k (\hat{\sigma}_k^{(n)})^2 \leq s_2\} = 1. \tag{5.8}$$

Proof First we prove Equation (5.7). Let $\hat{\sigma}_k^{(n)} = \hat{\tau}_k^{(n)} \hat{\sigma}_1^{(n)}$ for $k = 1, \ldots, K$. Then $\hat{\tau}_k^{(n)} \in [c, 1/c]$. We write

$$h(\sigma) = l(\hat{p}_1^{(n)}, \ldots, \hat{p}_K^{(n)}, \hat{\mathbf{a}}_1^{(n)}, \ldots, \hat{\mathbf{a}}_K^{(n)}, \hat{b}_1^{(n)}, \ldots, \hat{b}_K^{(n)}, \hat{\tau}_1^{(n)}\sigma, \ldots, \hat{\tau}_K^{(n)}\sigma).$$

Then it satisfies

$$\frac{dh(\sigma)}{d\sigma}|_{\hat{\sigma}_1^{(n)}} = 0.$$

This gives rise to

$$\sum_{i=1}^{n} \frac{-\frac{1}{\hat{\sigma}_1^{(n)}} f(\mathbf{X}_i, \hat{\boldsymbol{\theta}}^{(n)}) + \frac{1}{(\hat{\sigma}_1^{(n)})^3} \sum_{k=1}^{K} \frac{((\hat{\mathbf{a}}_k^{(n)})'\mathbf{X}_i - \hat{b}_k^{(n)})^2}{(\hat{\tau}_k^{(n)})^2} \hat{p}_k^{(n)} f_k(\mathbf{X}_i, \hat{\boldsymbol{\theta}}^{(n)})}{f(\mathbf{X}_i, \hat{\boldsymbol{\theta}}^{(n)})} = 0.$$

Solving for $\hat{\sigma}_1^{(n)}$ yields

$$(\hat{\sigma}_1^{(n)})^2 = \frac{1}{n}\sum_{i=1}^{n} \frac{\sum_{k=1}^{K} \frac{((\hat{\mathbf{a}}_k^{(n)})'\mathbf{X}_i - \hat{b}_k^{(n)})^2}{(\hat{\tau}_k^{(n)})^2} \hat{p}_k^{(n)} f_k(\mathbf{X}_i, \hat{\boldsymbol{\theta}}^{(n)})}{f(\mathbf{X}_i, \hat{\boldsymbol{\theta}}^{(n)})}.$$

Then, we have

$$(\hat{\sigma}_1^{(n)})^2 \geq c^2 \frac{1}{n}\sum_{i=1}^{n} \min_{1\leq k\leq K}((\hat{\mathbf{a}}_k^{(n)})'\mathbf{X}_i - \hat{b}_k^{(n)})^2. \tag{5.9}$$

We now show that for $n_0 \geq Kd+1$,

$$s(n_0) = E\{\inf_{\{(\mathbf{a}_1,\ldots,\mathbf{a}_K,b_1,\ldots,b_K):\mathbf{a}_k'\mathbf{a}_k=1\}} \sum_{i=1}^{n_0} \min_{1\leq k\leq K}(\mathbf{a}_k'\mathbf{X}_i - b_k)^2\} > 0. \tag{5.10}$$

In fact,

$$\begin{aligned}
&\int \inf_{\{(\mathbf{a}_1,\ldots,\mathbf{a}_K,b_1,\ldots,b_K):\mathbf{a}_k'\mathbf{a}_k=1\}} \sum_{i=1}^{n_0} \min_{1\leq k\leq K}(\mathbf{a}_k'\mathbf{x}_i - b_k)^2 dP^{n_0}(\mathbf{x}) \\
=&\int \inf_{\{(\mathbf{a}_1,\ldots,\mathbf{a}_K,b_1,\ldots,b_K):\mathbf{a}_k'\mathbf{a}_k=1,|b_k|\leq r(\mathbf{x}_1,\ldots,\mathbf{x}_{n_0})\}} \sum_{i=1}^{n_0} \min_{1\leq k\leq K}(\mathbf{a}_k'\mathbf{x}_i - b_k)^2 dP^{n_0}(\mathbf{x}) \\
=&\int \min_{\{(\mathbf{a}_1,\ldots,\mathbf{a}_K,b_1,\ldots,b_K):\mathbf{a}_k'\mathbf{a}_k=1,|b_k|\leq r(\mathbf{x}_1,\ldots,\mathbf{x}_{n_0})\}} \sum_{i=1}^{n_0} \min_{1\leq k\leq K}(\mathbf{a}_k'\mathbf{x}_i - b_k)^2 dP^{n_0}(\mathbf{x}) \\
\equiv&\int s_0(\mathbf{x}_1,\ldots,\mathbf{x}_{n_0}) dP^{n_0}(\mathbf{x}) > 0,
\end{aligned}$$

since $s_0(\mathbf{X}_1,\ldots,\mathbf{X}_{n_0}) > 0$ almost surely from Corollary5.1 and the argument in Theorem 5.2. By the strong law of large numbers, we have

$$\frac{1}{m}\sum_{j=1}^{m}\left\{\inf_{\{(\mathbf{a}_1,\ldots,\mathbf{a}_K,b_1,\ldots,b_K):\mathbf{a}_k'\mathbf{a}_k=1\}} \sum_{i=1}^{n_0} \min_{1\leq k\leq K}(\mathbf{a}_k'\mathbf{X}_{(j-1)n_0+i} - b_k)^2\right\} \to s(n_0),$$

with probability 1. Let $s_1 = c^2 s(n_0, P)/n_0$. Then by equation (5.9),

$$\begin{aligned}
&\liminf_{n\to\infty}(\hat{\sigma}_1^{(n)})^2 \\
\geq& c^2 \liminf_{m\to\infty}\{\frac{1}{mn_0}\inf_{\{(\mathbf{a}_1,\ldots,\mathbf{a}_K,b_1,\ldots,b_K):\mathbf{a}_k'\mathbf{a}_k=1\}} \sum_{i=1}^{mn_0} \min_{1\leq k\leq K}(\mathbf{a}_k'\mathbf{X}_i - b_k)^2\} \\
\geq& \frac{c^2}{n_0}\lim_{m\to\infty}\frac{1}{m}\sum_{j=1}^{m}\{\inf_{\{(\mathbf{a}_1,\ldots,\mathbf{a}_K,b_1,\ldots,b_K):\mathbf{a}_k'\mathbf{a}_k=1\}} \sum_{i=1}^{n_0} \min_{1\leq k\leq K}(\mathbf{a}_k'\mathbf{X}_{(j-1)n_0+i} - b_k)^2\} \\
=& s_1,
\end{aligned}$$

with probability 1. The same s_1 serves as a lower bound of $\liminf_n(\hat{\sigma}_k^{(n)})^2$ for $k = 2, \ldots, K$ as well. Noticing that s_1 does not depend on a set of observations $\mathbf{x}_1, \ldots, \mathbf{x}_n$, we have proved Equation (5.7).

Now we prove Equation (5.8). Let $\Theta_c^s = \{\boldsymbol{\theta} : \boldsymbol{\theta} \in \Theta_c, \sigma_k^2 > s \text{ for some } k\}$. Define

$$q(\mathbf{x}, s) = \sup_{\boldsymbol{\theta} \in \Theta_c^s} f(\mathbf{x}, \boldsymbol{\theta}).$$

Then $q(\mathbf{x}, s) \leq \frac{1}{\sqrt{2\pi sc}}$. By Lemma 5.2, $E|\log f(\mathbf{X}, \boldsymbol{\theta}_0)| < \infty$. There exists a positive number s_2, such that

$$E(\log(q(\mathbf{X}, s_2))) < E(\log f(\mathbf{X}, \boldsymbol{\theta}_0).$$

By the strong law of large numbers,

$$\text{Prob}\{\lim_{n\to\infty} \sum_{i=1}^{n} [\log q(\mathbf{X}_i, s_2) - \log f(\mathbf{X}_i, \boldsymbol{\theta}_0)] = -\infty\} = 1.$$

This implies that

$$\text{Prob}\{\lim_{n\to\infty} \frac{\sup_{\boldsymbol{\theta} \in \Theta_c^{s_2}} f(\mathbf{X}_1, \boldsymbol{\theta}) \ldots f(\mathbf{X}_n, \boldsymbol{\theta})}{f(\mathbf{X}_1, \boldsymbol{\theta}_0) \ldots f(\mathbf{X}_n, \boldsymbol{\theta}_0)} = 0\} = 1.$$

Equation 5.8 follows immediately, since $\hat{\boldsymbol{\theta}}^{(n)}$ always satisfies

$$\frac{f(\mathbf{X}_1, \hat{\boldsymbol{\theta}}^{(n)}) \ldots f(\mathbf{X}_n, \hat{\boldsymbol{\theta}}^{(n)})}{f(\mathbf{X}_1, \boldsymbol{\theta}_0) \ldots f(\mathbf{X}_n, \boldsymbol{\theta}_0)} \geq 1.$$

∎

Now we prove the main consistency result.

Theorem 5.4 *Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be a sample from an absolutely continuous probability measure P with finite second moments. Let $\hat{\boldsymbol{\theta}}^{(n)}$ be a global maximizer of $l(\boldsymbol{\theta} \,|\mathbf{X}_1, \ldots, \mathbf{X}_n)$ over Θ_c. Then $\hat{\boldsymbol{\theta}}^{(n)} \to \boldsymbol{\theta}_0^q$ almost surely in the topological space Θ_c^q.*

Proof From Lemma 5.5, we need only to consider the subspace of Θ_c,

$$\Theta_c^s = \{\boldsymbol{\theta} \in \Theta_c : s_1 \leq \sigma_k^2 \leq s_2, \text{ for all } k = 1, \ldots, K\},$$

where s_1, s_2 are positive numbers determined in Lemma 5.5.

Since Θ_c^s is not compact, Theorem 5.3 cannot be used directly. We shall use the compactification device in Hathaway (1985) and also in Kiefer and Wolfowitz (1956). In the space Θ_c^s, define the metric

$$\delta(\boldsymbol{\theta}, \boldsymbol{\theta}') = \sum_i |\arctan \boldsymbol{\theta}_i - \arctan \boldsymbol{\theta}_i'|,$$

where $|\cdot|$ is the Euclidean distance and $\boldsymbol{\theta}_i$ and $\boldsymbol{\theta}_i'$ are components of $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$ respectively. Let $\bar{\Theta}_c^s$ be the set of Θ_c^s along with all its limit points. Then

$$\bar{\Theta}_c^s = \left\{ \begin{array}{l} \boldsymbol{\theta} : 0 \le p_k \le 1, \sum_{k=1}^K p_k = 1, \min_{i,j} \sigma_i/\sigma_j \ge c > 0, \\ \mathbf{a}_k'\mathbf{a}_k = 1, -\infty \le b_k \le \infty, s_1 \le \sigma_k \le s_2, k = 1, \ldots, K. \end{array} \right\}$$

is compact. Since $f(\mathbf{x}, \cdot)$ is continuous on Θ_c^s, it can be extended to $\bar{\Theta}_c^s$ as

$$f(\mathbf{x}, \boldsymbol{\theta}) = \sum_k p_k I(\infty < b_k < \infty) f_k(\mathbf{x}, \boldsymbol{\theta}).$$

We have shown $g(\boldsymbol{\theta}) = E(\log f(\mathbf{x}, \boldsymbol{\theta}))$ is continuous on Θ_c^s. It is continuous on $\bar{\Theta}_c^s$ as well. To see this, let $\boldsymbol{\theta}^{(n)}$ be a sequence tending to $\boldsymbol{\theta}$. If all $b_k = \pm\infty$, or all $p_k = 0$ whenever $b_k \ne \pm\infty$, then $f(\mathbf{x}, \boldsymbol{\theta}) = 0$ and $g(\boldsymbol{\theta}) = -\infty$. In this case,

$$\begin{aligned} g(\boldsymbol{\theta}^{(n)}) &= \int \log f(\mathbf{x}, \boldsymbol{\theta}^{(n)}) dP(\mathbf{x}) \\ &\le \int \max_{\{k:p_k>0\}} \log f_k(\mathbf{x}, \boldsymbol{\theta}^{(n)}) dP(\mathbf{x}) \\ &\to \int -\min_{\{k:p_k>0\}} \left[\frac{1}{2} \log(2\pi\sigma_k^2) + \frac{(\mathbf{a}_k'\mathbf{x} - b_k)^2}{2\sigma_k^2} \right] dP(\mathbf{x}) \\ &= -\infty. \end{aligned}$$

If there is some k such that $p_k \ne 0$ and $b_k \ne \pm\infty$. Then

$$\log f(\mathbf{x}, \boldsymbol{\theta}^{(n)}) \le \log \max f_k(\mathbf{x}, \boldsymbol{\theta}^{(n)}) \le -\frac{1}{2} \log(2\pi s_1),$$

and

$$\begin{aligned} \log f(\mathbf{x}, \boldsymbol{\theta}^{(n)}) &\ge \log \min_{\{k:p_k>0, b_k\ne\pm\infty\}} p_k^{(n)} f_k(\mathbf{x}, \boldsymbol{\theta}^{(n)}) \\ &\ge \log \min_{\{k:p_k>0, b_k\ne\pm\infty\}} p_k f_k(\mathbf{x}, \boldsymbol{\theta}) - \epsilon, \end{aligned}$$

for some $\epsilon > 0$ when n is sufficiently large. By Lemma 5.2, the latter term is integrable. By Lebesgue's dominated convergence theorem, g is continuous at $\boldsymbol{\theta}$.

Lemmas 5.3 and 5.4 are easily seen to hold and hence we can repeat literally the proof of Theorem 5.3. This completes the proof. ∎

5.4 Asymptotical normality

The global maximizer $\hat{\boldsymbol{\theta}}^{(n)}$ converges in the above quotient topological space. There is hence a subsequence, still denoted by $\hat{\boldsymbol{\theta}}^{(n)}$, which converges to a point $\boldsymbol{\theta}_0$ in C_0 in the original space. If $\boldsymbol{\theta}_0$ is an interior point of Θ_c, then we can expand $l'(\hat{\boldsymbol{\theta}}^{(n)})$ about $\boldsymbol{\theta}_0$,

$$\begin{aligned} & l'(\hat{\boldsymbol{\theta}}^{(n)}) \\ =& l'(\boldsymbol{\theta}_0) + l''(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}}^{(n)} - \boldsymbol{\theta}_0) \\ & + \frac{1}{2}[(\hat{\boldsymbol{\theta}}^{(n)} - \boldsymbol{\theta}_0)^T l_1'''(\boldsymbol{\theta}^*)(\hat{\boldsymbol{\theta}}^{(n)} - \boldsymbol{\theta}_0)], \ldots, (\hat{\boldsymbol{\theta}}^{(n)} - \boldsymbol{\theta}_0)^T l_s'''(\boldsymbol{\theta}^*)(\hat{\boldsymbol{\theta}}^{(n)} - \boldsymbol{\theta}_0)]^T, \end{aligned} \tag{5.11}$$

where s is the dimension of $\boldsymbol{\theta}$ and $\boldsymbol{\theta}^*$ is a point on the line segment connecting $\hat{\boldsymbol{\theta}}^{(n)}$ and $\boldsymbol{\theta}_0$; l' is an s-dimensional vector of first derivatives of l, l'' and l_j''' are $s \times s$ matrices of second and third derivatives of l. For ease of notation, we do not differentiate which components of $\mathbf{p}$ and $\mathbf{a}_k$ are free parameters and which are not; it is assumed that the constraints $\sum p_k = 1$ and $\mathbf{a}_k'\mathbf{a}_k = 1$ are taken care in the calculation of these derivatives.

The left-hand side of (5.11) is $\mathbf{0}$, because $\hat{\boldsymbol{\theta}}^{(n)}$ satisfies the first order conditions.

Let

$$v_0(\boldsymbol{\theta}_0) = E[((\log f(\mathbf{X}, \boldsymbol{\theta}_0))'].$$

Then $v_0(\boldsymbol{\theta}_0) = g'(\boldsymbol{\theta}_0) = 0$ by Lebesgue's dominated convergence theorem.

Let

$$v_1(\boldsymbol{\theta}_0) = E[((\log f(\mathbf{X}, \boldsymbol{\theta}_0))')^2].$$

It is straightforward to verify that all the entries of $v_1(\boldsymbol{\theta}_0)$ are finite if the underlying distribution P has finite fourth moments. Then by the central limit theorem, the first derivative $l'(\boldsymbol{\theta}_0)/\sqrt{n}$ is asymptotical normal,

$$\frac{1}{\sqrt{n}} l'(\boldsymbol{\theta}_0) \xrightarrow{L} N(0, v_1(\boldsymbol{\theta}_0)).$$

Let

$$v_2(\boldsymbol{\theta}_0) = E[(\log f(\mathbf{X}, \boldsymbol{\theta}_0))''],$$

and

$$v_3(\boldsymbol{\theta}_0) = E[(\log f(\mathbf{X}, \boldsymbol{\theta}_0))'''].$$

Again, it is straightforward but tedious to verify that all the entries of $v_2(\boldsymbol{\theta}_0)$ are finite if P has finite fourth moments and that all the entries of $v_3(\boldsymbol{\theta}_0)$ are finite if P has finite sixth moments.

By the strong law of large numbers, the second derivative $l''(\boldsymbol{\theta}_0)/n$ tends to a constant matrix,

$$\frac{1}{n} l''(\boldsymbol{\theta}_0) \to v_2(\boldsymbol{\theta}_0),$$

and the third derivative l''' is bounded entry-wise. So we have the following

Theorem 5.5 *Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be a sample from an absolutely continuous probability measure P with finite sixth moments. Let $\hat{\boldsymbol{\theta}}^{(n)}$ be a subsequence of global maximizer of $l(\boldsymbol{\theta}|\mathbf{X}_1, \ldots, \mathbf{X}_n)$ over Θ_c, which tends to an interior point $\boldsymbol{\theta}_0$. Then*

$$\sqrt{n}(\hat{\boldsymbol{\theta}}^{(n)} - \boldsymbol{\theta}_0) \xrightarrow{L} N(0, v(\boldsymbol{\theta}_0)),$$

where $v(\boldsymbol{\theta}_0) = [v_2(\boldsymbol{\theta}_0)]^+ v_1(\boldsymbol{\theta}_0)[v_2(\boldsymbol{\theta}_0)]^+$ and A^+ is the Moore-Penrose inverse of matrix A.

In equation (5.3), $\hat{w}_{ik}$ is a function of $\mathbf{x}_i$ and $\hat{\boldsymbol{\theta}}$. Denote $\hat{w}_{ik} = h_k(\mathbf{x}_i, \hat{\boldsymbol{\theta}})$, $k = 1, \ldots, K$. By the Delta method, we have the following

Corollary 5.2 *Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be a sample from an absolutely continuous probability measure P with finite sixth moments. Let $\hat{\boldsymbol{\theta}}^{(n)}$ be a subsequence of global maximizer of $L(\boldsymbol{\theta}|\mathbf{X}_1, \ldots, \mathbf{X}_n)$ over Θ_c, which tends to an interior point $\boldsymbol{\theta}_0$. Let $\mathbf{x}$ be a data point. Then*

$$\sqrt{n}(h_k(\mathbf{x}, \hat{\boldsymbol{\theta}}^{(n)}) - h_k(\mathbf{x}, \boldsymbol{\theta}_0)) \xrightarrow{L} N(0, [h_k^{(0)}(\mathbf{x}, \boldsymbol{\theta}_0)]' v(\boldsymbol{\theta}_0)[h_k^{(0)}(\mathbf{x}, \boldsymbol{\theta}_0)]),$$

for $k = 1, \ldots, K$, where $h_k^{(0)}(\mathbf{x}, \boldsymbol{\theta}_0) = \frac{\partial h_k(\mathbf{x}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}$.

Using this corollary, we can build approximate confidence intervals for w_{ik} by replacing $\boldsymbol{\theta}_0$ with $\hat{\boldsymbol{\theta}}$ and hence evaluate the clustering of a data set.

Bibliography

Eguchi, S. and J. Copas (2006). Interpreting Kullback–Leibler divergence with the Neyman–Pearson lemma. *Journal of Multivariate Analysis 97*(9), 2034–2040.

García-Escudero, L., A. Gordaliza, R. San Martín, S. van Aelst, and R. Zamar (2007). Robust linear clustering. Preprint.

Hathaway, R. (1983). *Constrained maximum-likelihood estimation for a mixture of m univariate normal distributions*. Ph. D. thesis, Rice University.

Hathaway, R. (1985). A Constrained Formulation of Maximum-Likelihood Estimation for Normal Mixture Distributions. *The Annals of Statistics 13*(2), 795–800.

Kullback, S. and R. Leibler (1951). On Information and Sufficiency. *The Annals of Mathematical Statistics 22*(1), 79–86.

Redner, R. (1981). Note on the Consistency of the Maximum Likelihood Estimate for Nonidentifiable Distributions. *The Annals of Statistics 9*(1), 225–228.

Wald, A. (1949). Note on the Consistency of the Maximum Likelihood Estimate. *The Annals of Mathematical Statistics 20*(4), 595–601.

Yan, G., W. Welch, and R. Zamar (2008). A likelihood approach to linear clustering. Preprint.

Chapter 6

Future Work

In this thesis work, we have developed a SNP genotype calling algorithm based on the linear grouping algorithm (Van Aelst et al., 2006), proposed a flexible model-based approach to linear clustering and introduced a Bayesian approach to linear clustering with specific relevance to the SNP genotyping problem. In this chapter, we briefly describe a few possible directions for future research.

6.1 Robustness consideration

Robustness to outliers is desirable in linear clustering, as the assumption of normal deviations around hyperplanes is sensitive to large deviations in the orthogonal direction. In addition to the inclusion of a uniform background cluster (Banfield and Raftery, 1993), one option would be to use a heavier tailed distribution, for example, Student's t distribution with small degrees of freedom or with degrees of freedom depending on the data. In the partial likelihood approach, this would adapt Peel and McLachlan (2000)'s EM algorithm for t mixture models from the elliptical context to the linear clustering setting. The adaptation is straightforward but computationally more expensive. Further ideas include estimating the component covariance matrices in the M-step in a robust way, for example, trimming off some points as done by García-Escudero et al. (2007). In the Bayesian approach, we have already used a Student's t distribution with small degrees of freedom.

6.2 Asymptotics

For the partial likelihood approach, we produced some asymptotic results in the case of normal orthogonal deviations. For more general cases, such as Student's t distribution, these results have not been established. We shall study the asymptotic evaluation in more general cases.

6.3 Model Extension

With $\mathbf{a}'\mathbf{x} = b$, we are specifying a hyperplane in $d-1$ dimensions. With little effort, this could be generalized to a mixture of partial likelihoods, each of which specifies a hyperplane of dimension $q < d$,

$$l(\boldsymbol{\kappa}, \mathbf{p}|\mathbf{x}_{1:n}) = \prod_{i=1}^{n} \sum_{k=1}^{K} p_k N(A_k'\mathbf{x}_i - \mathbf{b}_k; \mathbf{0}, \Sigma_k), \tag{6.1}$$

where A is of dimension $d \times (d-q)$, b is a vector of dimension $d-q$, and Σ_k is a $(d-q) \times (d-q)$ covariance matrix for the deviation from the hyperplane. In the extreme case of a 0-dimension hyperplane, which is a point, we have the usual mixture of multivariate normal distributions. A mixture of components with various dimensions could be considered.

6.4 Variable/model selection

For a large, high dimensional dataset, efficient computation is essential in order to uncover linear patterns. In addition to develop more efficient algorithms, one idea is to use a subset of the data and/or a subset of variables. To this end, we are interested in methods to screen variables as well as to determine the number of linear structures.

6.5 Bayesian computation

In our examples in the Bayesian approach, label-switching is prevented either by a Gibbs sampler applied to a posterior distribution with isolated modes or by informative priors. In more general situations, we may need the ideas of tempering MCMC or Sequential Monte Carlo to explore the whole support of the posterior distribution and deal with the label-switching problem.

In addition, the number of linear clusters are assumed known in our approach. In the situation of unknown number of clusters, our first thought is to investigate the feasibility of the Reversible Jump MCMC of Richardson and Green (1997). This may imply heavy computational burden. A related problem is again the scalability of the Bayesian approach to large datasets and high dimensions. We leave these problems for further research.

Bibliography

Banfield, J. and A. Raftery (1993). Model-Based Gaussian and Non-Gaussian Clustering. *Biometrics 49*(3), 803–821.

García-Escudero, L., A. Gordaliza, R. San Martín, S. van Aelst, and R. Zamar (2007). Robust linear clustering. Preprint.

Peel, D. and G. McLachlan (2000). Robust mixture modelling using the t distribution. *Statistics and Computing 10*(4), 339–348.

Richardson, S. and P. Green (1997). On Bayesian Analysis of Mixtures with an Unknown Number of Components. *Journal of the Royal Statistical Society. Series B (Methodological) 59*(4), 731–792.

Van Aelst, S., X. Wang, R. Zamar, and R. Zhu (2006). Linear grouping using orthogonal regression. *Computational Statistics and Data Analysis 50*(5), 1287–1312.

Appendix A

Glossary of some genetic terms

(Extracted from the website http://www.genome.gov/glossary.cfm.)

adenine (A) One of the four bases in DNA that make up the letters ATGC, adenine is the "A". The others are guanine, cytosine, and thymine. Adenine always pairs with thymine.

allele One of the variant forms of a gene at a particular locus, or location, on a chromosome. Different alleles produce variation in inherited characteristics such as hair color or blood type. In an individual, one form of the allele (the dominant one) may be expressed more than another form (the recessive one).

base pair Two bases which form a "rung of the DNA ladder." A DNA nucleotide is made of a molecule of sugar, a molecule of phosphoric acid, and a molecule called a base. The bases are the "letters" that spell out the genetic code. In DNA, the code letters are A, T, G, and C, which stand for the chemicals adenine, thymine, guanine, and cytosine, respectively. In base pairing, adenine always pairs with thymine, and guanine always pairs with cytosine.

chromosome One of the thread-like "packages" of genes and other DNA in the nucleus of a cell. Different kinds of organisms have different numbers of chromosome. Humans have 23 pairs of chromosomes, 46 in all: 44 autosomes and two sex chromosomes. Each parent contributes one chromosome to each pair, so children get half of their chromosomes from their mothers and half from their fathers.

cytosine One of the four bases in DNA that make up the letters ATGC, cytosine is the "C". The others are adenine, guanine, and thymine. Cytosine always pairs with guanine.

deoxyribonucleic acid (DNA) The chemical inside the nucleus of a cell that carries the genetic instructions for making living organisms.

diploid The number of chromosomes in most cells except the gametes. In humans, the diploid number is 46.

dominant A gene that almost always results in a specific physical characteristic, for example, a disease, even though the patient's genome possesses only one copy. With a dominant gene, the chance of passing on the gene (and therefore the disease) to children is 50-50 in each pregnancy.

gene The functional and physical unit of heredity passed from parent to offspring. Genes are pieces of DNA, and most genes contain the information for making a specific protein.

gene amplification An increase in the number of copies of any particular piece of DNA. A tumor cell amplifies, or copies, DNA segments naturally as a result of cell signals and sometimes environmental events.

gene expression The process by which proteins are made from the instructions encoded in DNA.

genetic code (ATCG) The instructions in a gene that tell the cell how to make a specific protein. A, T, G, and C are the "letters" of the DNA code; they stand for the chemicals adenine, thymine, guanine, and cytosine, respectively, that make up the nucleotide bases of DNA. Each gene's code combines the four chemicals in various ways to spell out 3-letter "words" that specify which amino acid is needed at every step in making a protein.

genetic marker A segment of DNA with an identifiable physical location on a chromosome and whose inheritance can be followed. A marker can be a gene, or it can be some section of DNA with no known function. Because DNA segments that lie near each other on a chromosome tend to be inherited together, markers are often used as indirect ways of tracking the inheritance

pattern of a gene that has not yet been identified, but whose approximate location is known.

genome All the DNA contained in an organism or a cell, which includes both the chromosomes within the nucleus and the DNA in mitochondria.

genotype The genetic identity of an individual that does not show as outward characteristics.

guanine One of the four bases in DNA that make up the letters ATGC, guanine is the "G". The others are adenine, cytosine, and thymine. Guanine always pairs with cytosine.

haploid The number of chromosomes in a sperm or egg cell, half the diploid number.

Haplotype It refers to a set of SNPs found to be statistically associated on a single chromatid. With this knowledge, the identification of a few alleles of a haplotype block unambiguously identifies all other polymorphic sites in this region. Such information is most valuable to investigate the genetics behind common diseases and is collected by the International HapMap Project (http://en.wikipedia.org/wiki/Haplotype).

Hardy-Weinberg equilibrium (HWE) It states that, under certain conditions, after on generation of random mating, the genotype frequencies at a single gene locus will become fixed at a particular equilibrium value. It also specifies that those equilibrium frequencies can be represented as a simple function of the allele frequencies at that locus.(http://en.wikipedia.org /wiki /Hardy-Weinberg _ equilibrium).

heterozygous Possessing two different forms of a particular gene, one inherited from each parent.

homozygous Possessing two identical forms of a particular gene, one inherited from each parent.

linkage The association of genes and/or markers that lie near each other on a chromosome. Linked genes and markers tend to be inherited together.

linkage disequilibrium (LD) The non-random association of alleles at two or more loci on a chromosome. It describes a situation in which some combinations of alleles or genetic markers occur more or less frequently in a population than would be expected from a random formation of haplotypes from alleles based on their frequencies. In population genetics, linkage disequilibrium is said to characterize the haplotype distribution at two or more loci (http://en.wikipedia.org/wiki/Linkage_disequilibrium).

locus The place on a chromosome where a specific gene is located, a kind of address for the gene. The plural is "loci," not "locuses."

non-coding DNA The strand of DNA that does not carry the information necessary to make a protein. The non-coding strand is the mirror image of the coding strand and is also known as the antisense strand.

nucleotide One of the structural components, or building blocks, of DNA and RNA. A nucleotide consists of a base (one of four chemicals: adenine, thymine, guanine, and cytosine) plus a molecule of sugar and one of phosphoric acid.

phenotype The observable traits or characteristics of an organism, for example hair color, weight, or the presence or absence of a disease. Phenotypic traits are not necessarily genetic.

polymerase chain reaction (PCR) A fast, inexpensive technique for making an unlimited number of copies of any piece of DNA. Sometimes called "molecular photocopying," PCR has had an immense impact on biology and medicine, especially genetic research.

polymorphism A common variation in the sequence of DNA among individuals.

primer A short oligonucleotide sequence used in a polymerase chain reaction.

probe A piece of labeled DNA or RNA or an antibody used to detect the function of a gene.

single nucleotide polymorphism Common, but minute, variations that occur in human DNA at a frequency of one every 1,000 bases. These variations can be used to track inheritance in families. SNP is pronounced "snip".

thymine One of the four bases in DNA that make up the letters ATGC, thymine is the "T". The others are adenine, guanine, and cytosine. Thymine always pairs with adenine.

wild-type allele The allele designated as the standard ("normal") for a strain of organism.