The Pennsylvania State University

The Graduate School

Department of Statistics

DISTANCE-BASED MODEL-SELECTION WITH APPLICATION TO THE ANALYSIS OF GENE EXPRESSION DATA

A Thesis in

Statistics

by

Surajit Ray

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

December 2003

The thesis of Surajit Ray has been reviewed and approved* by the following:

Bruce G. Lindsay
Distinguished Professor of Statistics
Thesis Advisor
Chair of Committee

Thomas P. Hettmansperger
Professor of Statistics
Interim Head of the Department of Statistics

Francesca Chiaromonte
Assistant Professor of Statistics

Benjamin F. Pugh
Associate Professor of Biochemistry and Molecular Biology

*Signatures are on file in the Graduate School.

Abstract

Distance-based Model-Selection with application to the Analysis of Gene Expression Data

Multivariate mixture models provide a convenient method of density estimation and model based clustering as well as providing possible explanations for the actual data generation process. But the problem of choosing the number of components (g) in a statistically meaningful way is still a subject of considerable research . Available methods for estimating g include, optimizing AIC and BIC, estimating the number through nonparametric maximum likelihood, hypothesis testing and Bayesian approaches with entropy distances. In our current research we present several rules for selecting a finite mixture model, and hence g, based on estimation and inference using a quadratic distance measure.

In one methodology the goal is to find the minimal number of components that are needed to adequately describe the true distribution based on a nonparametric confidence set for the true distribution. We also present results for selecting g based on a risk analysis that includes a penalty for overfitting. Another less formal methodology is based on the concordance measure which is analogous to R^2 in regression. Moreover, we find develop diagnostics for purposes of outlier detection. These diagnostics help to distinguish between outliers and true clusters, and they provide insight into the initial values for iterative estimation of additional components.

In this dissertation we also develop tools for determining the number of modes in a mixture of multivariate normal densities. We use these criterion to select clusters which

display distinct modes. Finally we fine tune our methods to analyze gene-expression data from micro-arrays, and compare them with other competitive methods.

Table of Contents

List of Tables

List of Figures

Acknowledgements

First, I would like to thank my advisor Dr. Bruce Lindsay, for his encouragement and guidance throughout my stay at Pennstate. He was always there ready to help me and point out new directions in my research. I would also like to thank Drs. Tom Hettmansperger, Francesca Chiaromonte and Frank Pugh for the time they spent discussing my thesis. I express my deep appreciation for their efforts, questions, comments, and suggestions, as well as their unending patience. I am also grateful to all other faculty and staff members of the department, whose encouragement and support made my stay at Pennstate an enjoyable one.

Special thanks to all my teachers at Presidency College, Calcutta and Indian Statistical Institute, Calcutta who have always motivated me to do further research in Statistics. The theoretical and computational training I had in these institutions helped me tremendously in all aspects of my dissertation work.

I am extremely grateful to my parents Menoka and Amiya Kumar Ray for their love and support. They always believed in me and encouraged me in all stages of my life. Finally I would like to thank my friends, without whose support and inspiration this thesis would not have been possible.

Chapter 1

Introduction

Multivariate mixture models provide a convenient method of density estimation and model based clustering as well as providing possible explanations for the actual data generation process. But the problem of choosing the number of components (g) in a statistically meaningful way is still a subject of considerable research . Available methods for estimating g include, optimizing AIC and BIC, estimating the number through nonparametric maximum likelihood, hypothesis testing and Bayesian approaches with entropy distances. In our current research we present several rules for selecting a finite mixture model, and hence g, based on estimation and inference using a quadratic distance measure. In this chapter we will first provide the basic definitions and concepts of an arbitrary mixture of statistical distributions. Then, we will demonstrate why the problem of choosing the number of components is an important area of research. In the later part of this chapter we will give an outline of this dissertation and introduce the notations that are consistently used throughout this dissertation.

1.1 General introduction to finite mixture models

In this section we will define a finite mixture distribution. Let X be a random variable or in general a random vector taking value in the sample space $\mathcal{X}$. X is said to have a g-component mixture density with density function $f(x)$ if the density function can be written

as

$$f(x) = \sum_{j=1}^{g} \pi_j \;\; f_j(x;\lambda_j,\theta) \text{ for } x \in \mathcal{X}, \tag{1.1}$$

where, $f_j(y;\lambda_j,\theta)$'s are the component densities and π_j are mixing proportions with the restriction

$$0 \leq \pi_j \leq 1 \;\; \forall j \qquad \text{and} \qquad \sum_{j=1}^{g} \pi_j = 1. \tag{1.2}$$

Note that the component densities may have some parameter (e.g. θ in (1.1)) constant over all the component densities while some parameter (e.g. λ_j in (1.1)) may distinguish the component densities from one another. For example, we may think of a normal mixture with constant variance but different location parameter for each component. The parameter set of the mixture in equation (1.1) are often denoted in the following way

$$\theta, \begin{pmatrix} \pi_1 & \pi_2 & \dots & \pi_g \\ \lambda_1 & \lambda_2 & \dots & \lambda_g \end{pmatrix} \tag{1.3}$$

Note that in this dissertation we will not deal with infinite mixtures.

1.2 The challenge: choosing the number of components

Choosing the number of components in a statistically meaningful way is still a subject of considerable research. Mixture models are broadly used for two specific purposes. One is to give a semiparametric framework to a unknown distribution. Titterington et al. (1985) referred to this as the "indirect application" of mixture models. In this case the choice of the number of components is not usually a big issue. Referred to as the "direct application" by Titterington et al. (1985), mixture models are mainly used to provide a model-based clustering and to do so the choice of the number of components must be carefully and critically addressed. The main problem in choosing the number of components is the problem of over-fitting, which arises from the nesting of the mixture models, so that a distribution which can be well approximated

by a the mixture of g-components can also be well approximated by g_0-components, for any $g_0 > g$, in the sense that the two mixture distributions are empirically indistinguishable. Thus the choice of g should be addressed very carefully. Moreover, the choice may depend on the question asked.

Recent development in many scientific fields has produced huge datasets with high dimension and large sample size. There is a strong demand to analyze these data and to group them into appropriate clusters. Many of the existing methods for model selection are not efficient enough to use in high dimensional data. In this dissertation all our model selection methods are especially designed for high dimensional multivariate situations.

1.3 Outline of the thesis

In this section we will give an outline of the dissertation. The goal of this dissertation is to propose model selection tools to choose the number of components in a normal mixture model in a possibly high-dimensional multivariate set up. The multivariate structure of the mixture models make many available methods of model selection impractical, because of the computational complexity. Our methods of model selection are designed to overcome this problem. Along with the theory, each chapter describes the application of the proposed methods to an array of problems. In particular, we apply our model selection tools to the Iris data and the Acidity Data along with two sets of simulated dataset, all of which are described in the Appendix.

In the next chapter we will give a detailed overview of the approaches that have previously been used in choosing the number of components of a mixture model. First, we define the goals of model selection and show how the inference on the number of components may differ according to the specific goal. Widely used methods based on likelihood and information criterion cannot be used routinely as the asymptotic distribution is very complex. Bayesian methods and nonparametric methods are also discussed in the context of model se-

lection.

Chapter 3 introduces the specific problem of selecting the number of components in a mixture of multivariate normals. Generalized quadratic distance, defined using a positive definite kernel is introduced. The positive definite kernel we use is based on the normal density function. Later we use these distance to construct an array of model selection tools. Using existing theories on $\mathcal{U}$-*statistics* and utilizing the convolution properties of normal kernels we work out an unbiased estimator of the distance. These distance calculations avoid the multidimensional numerical integration by using appropriate kernel for the proposed mixing distribution. We have examined the large-sample null distributions of these quadratic distances. Based on a spectral decomposition theorem, the kernel can be decomposed into basis functions (eigen-function/eigen-value analysis). Using the above spectral decomposition, the the large sample null distribution of the distance can be written as an infinite sum of weighted chi-squared distributions. These quadratic distances can be interpreted as the L_2 distance between kernel smoothed densities. Graphical comparisons between the L_1 distance, L_2 distance and the quadratic distances are demonstrated. Finally, we develop distance based model selection rules using nonparametric confidence intervals in situations where one might wish to allow for some approximation error in model building.

The distance defined in Chapter 3 is based on the normal kernel, which in turn depends on a tuning parameter h. In Chapter 4 we will show how the choice of the tuning parameter is extremely important in designing a powerful distance. Analogy will be drawn between the tuning parameter "h" and the bin-width of a cell in the χ^2 goodness-of-fit tests. To choose h we define the "pseudo degrees of freedom", an interesting and useful single number summary of the sensitivity characteristics of the distance. It can be calculated once and for all without using the model. Furthermore, we will use the "rule of thumb" available for the choice of bin-width in χ^2 goodness-of-fit tests, to decide on an interesting range of the tuning parameter of the kernel in the quadratic distance.

In Chapter 5 we use the quadratic distance to define the concordance between two densities. Analogy is drawn between the concordance coefficients of two densities and the R^2 values in the context of regression. The concordance coefficient between a proposed model and the empirical data, describes the amount of variability in the data that has been explained by the model density considered. Future work on using the concordance coefficients as a tool for finding an interesting range of the tuning parameter h is proposed.

Chapter 6 describes the notion of risk-based model selection. Using the quadratic distance as the loss function we define the risk of selecting a proposed model that is not the true model. This risk can be decomposed into two parts: one captures the model lack of fit, while the other is strongly related to the parameter estimation cost. We introduce some novel ideas for estimating this risk and apply them to model selection problems.

In the previous three chapters we introduced the quadratic distance and used it to build model selection tools. We next use it to construct diagnostics. The unbiased estimator of distance has a natural decomposition as sum of the residuals. In Chapter 7 we define quadratic residuals and use them for outlier detection. Although the unstandardized residuals are hard to interpret, after appropriate standardization the quadratic residuals have potential as a powerful diagnostic tool. In addition to providing information about model failure, they could be used to add more components to the model. With a simulated dataset, we demonstrate the detection of outliers.

Chapter 8 examines the number of modes in a two component multivariate mixture. This chapter is not directly related to the distance based model selection tools described in Chapters 3 through 6. However, this analysis is important for interpretation of the selected model, as a mixture with many components could have only a few modes. Conditions for modality of univariate densities have been studied by many scientists; but we did not find any previous analysis for the modality of mixture of multivariate distributions. In this chapter we discover analytical solutions for the existence of multiple modes when the component

densities have the same variance covariance structure. For the unequal variance case we have created plot-based methods for detecting multimodality. We then develop the notion of modal cluster, where we cluster together fitted mixture components based on pairwise unimodality.

In Chapter 9 we apply our model selection tools to analyze gene expression data. We give a brief introduction of the dataset, the experiment and the goal of the study. Finally, we demonstrate how our model selection tools are a good choice for analyzing high dimensional data.

1.4 Notational Preliminaries

In this section we will define notation that is used consistently throughout the dissertation. Some specific notation will be introduced later when they are first used.

With respect to a vector Y, Y' is its transpose. Similarly A' is the transpose of matrix A, where as its determinant is defined by $|A|$. I denotes the identity matrix of appropriate dimension.

As for the statistical notation, $X \sim f(x)$ means the random variable X follows a distribution with density $f(x)$. More specifically $X \sim f(x;\mu,\sigma)$ means the density function $f(x)$ depends on the parameters μ and σ. In particular, $\phi(x,\mu,\sigma)$ will denote the normal density with mean μ and variance σ. The distribution of X converges in probability to the distribution of Y is denoted by $X \xrightarrow{\mathbf{P}} Y$, where as $X \xrightarrow{\mathbf{d}} Y$ denote convergence in distribution. The normal density with parameters μ and Σ is denoted by $\mathcal{N}(\mu,\Sigma)$.

We denote the real line by $\mathbb{R}$, the p-dimensional Euclidean space by $\mathbb{R}^p$, and the set of positive integers by $\mathbb{I}$. In this dissertation we deal with multivariate mixture of normals. Hence our sample space $\mathcal{X}$ defined in equation (1.1) is $\mathbb{R}^p$. For a g-component Multivariate Normal Mixture the component densities f_i's are multivariate normals.

For model selection purposes τ will denote the true underlying distribution and $\hat{M}_g$

will denote the fitted $g-$component model. $D(F,G)$ is a generic measure of distance between two probability distributions F and G. Specific distances that depend on a kernel K will be indicated by use use of a suffix as in D_K.

Chapter 2

Overview of Previous Research

Multivariate mixture models provide well known and widely used methods for density estimation, model-based clustering, and explanations for the data generation process. However, the problem of choosing the number of components of a mixture model in a statistically meaningful way is still a subject of considerable research. In this chapter we discuss some of the commonly used techniques for determining the number of components in a mixture model. It should be noted that the choice of the number of components may not be a very important issue when mixture models are used to provide a semiparametric framework to unknown distributional shapes. In this case over-estimation of the number of components is not a serious problem. In fact, when mixtures are used for distributional approximation, Leroux (1992) showed that under very mild conditions, Akaike's Information Criterion (AIC) and Bayesian Information Criterion(BIC) along with certain other penalized likelihood criterion do not underestimate the true number of components, asymptotically. However, technically speaking they are incorrectly applied because the model lacks the regularity conditions needed for the simple penalty functions based on the number of parameters (See Section 2.2.2).

Our goal is to use the mixture model to provide model-based clustering and to do so the choice of the number of components must be carefully and critically addressed. The problem of over-fitting arises from the nesting of the mixture models, so that a distribution which can be well approximated by a the mixture of $g-$components can also be well approximated by g_0-components, for any $g_0 > g$, in the sense that the two mixture distributions are empiri-

cally indistinguishable. Thus, we can only a get a lower bound on the number of components. So, the right question to ask is not "how many components are there in the mixture model?" but "what is the smallest number of the components in the mixture needed to make the model compatible with the data?". Especially, in the context of model based clustering, the answer to the latter would guarantee a reasonable explanation of the phenomenon by which the data was generated without being wasteful.

2.1 Goals of selecting the number of components

Model selection rules are driven by the specific goal they serve. So let us start this chapter with an overview of the different possible goals for selecting g.

1. **Getting the number of components right:** In this situation we want $D(\tau, \hat{M}_{\hat{g}}) \to 0$, ($D$ being some measure of distance between two probability distributions) in such a way that $\hat{g} \to g$. In other words we want our estimated g to be a consistent estimator of the number of components in the mixture model. But "getting the number of components right" is not always a realistic goal as no upper confidence limit is possible for the number of components (Donoho, 1988).

2. **Using an adequate number of components:** A more realistic approach would be to find an adequate number of components, allowing for some distributional error because we could never ensure that in nature the data was generated by a mixture of normals, or some other distribution. In this case we propose to choose g so that $D(\tau, M_{T_g}) \leq \varepsilon$, where M_{T_g} is the Kullback-Leibler best g-component density and ε is a small positive quantity.

3. **Using the g that gives the minimum risk:** Another goal of selecting the number of components is getting a model which gives us the minimum expected risk. Here we propose to select a g for which $E[D(\tau, \hat{M}_g)]$ is small. In this approach one is penalized

for estimation of parameters. If we use richer and richer model, by increasing g, the penalty for the number of parameters increases at the same time that the model fitting error decreases. This approach to model selection is extensively discussed in Chapter 6.

2.2 Approaches to Model Selection

In this section we discuss different approaches that have previously been used in estimating the number of components of a mixture model. Readers wishing to get further details should consult the specific references. Before discussing the methods let us introduce some definitions and abbreviation, which we will be using throughout the rest of this chapter and beyond.

Abbreviations and definitions

- **LRTS:** The Likelihood Ratio Test Statistic will be denoted as λ, where
$$\lambda = \frac{\hat{L_0}}{\hat{L_1}},$$
$\hat{L_i}$ being the maximized likelihood under $H_i (i = 0, 1)$.
- **NPMLE:** Nonparametric Maximum Likelihood Estimator.

2.2.1 Number of Modes

Estimating the number of components of a mixture distribution by the number of modes is one of the oldest methods based on intuition. Titterington et al. (1985) described some inferential procedures for assessing the number of modes. However, the obvious drawback of this method is that if the component densities are not sufficiently far apart the mixture distribution would still be unimodal and estimating the number of components by the number of modes would fail. Note however, that the practical interest could lie in finding components that correspond to separate modes, so that true separation occurs. Modality for the normal mixture will be investigated in further detail in Chapter 8.

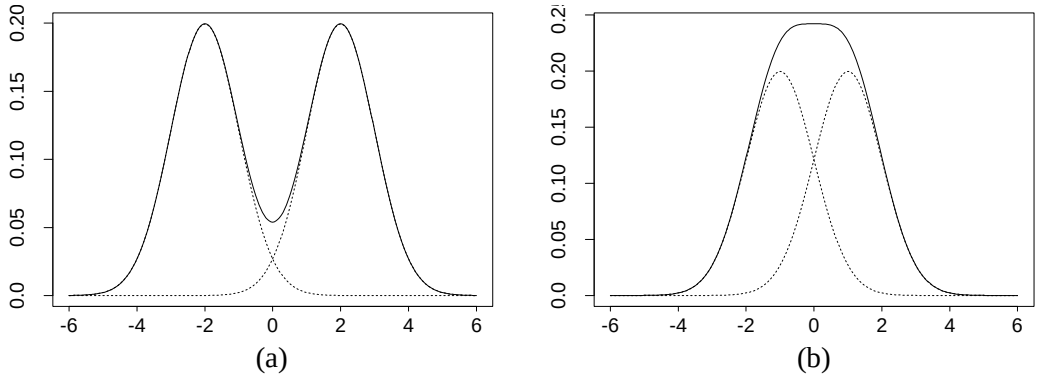

Figure 2.1: **Mixture of normals (a) means 4 standard deviation apart, (b) means 2 standard deviations apart**

We illustrate the distinction between modes and components through a simple example with the mixture of two univariate normals. Figure 2.1(a) is the mixture of two normals with equal weights, means being 4 standard deviation apart and Figure 2.1(b) has its means 2 standard deviations apart. In both the figures the dotted lines represent the component densities and the solid line is the the mixture density. Though Figure 2.1(a) produces a bimodal distribution, Figure 2.1(b) is still unimodal. Rigorous conditions for mixture of two densities displaying bimodality are discussed in Chapter 8

The above example illustrates that we cannot always infer the number of components with the help of the number of modes. Moreover visually inspecting the modes of a multivariate distribution becomes much more difficult. Conditions for bimodality in the multivariate case (both for the equal and unequal variance) are addressed in Chapter 8.

2.2.2 Likelihood based approaches

Likelihood based approaches are the most extensively used methods for testing of hypotheses. Moreover, the model selection problem can be framed as a hypothesis testing problem. One way for deciding whether the mixture model has g components is to perform a likelihood ratio hypothesis test for $H_0 : g = g_0$ *vs* $H_1 : g = g_0 + 1$. In this section we will discuss all likelihood based approaches to model selection and indicate why these tests are

hard to use for model selection in the mixture setup.

Likelihood ratio test statistic

As in any hypothesis testing problem, an obvious way of approaching the testing of the smallest number of component compatible with the data, is to use LRTS. Let us test the hypothesis that the data was generated by a mixture of g_0 components versus it was generated by g_1 components, for some $g_1 > g_0$. Framing it as a hypothesis test problem we write,

$$H_0 : g = g_0 \qquad vs \qquad H_1 : g = g_1 \tag{2.1}$$

Denoting $\hat{L}_i$ as the maximized likelihood under $H_i (i = 0, 1)$, the likelihood ratio test statistic reduces to

$$\begin{aligned} \lambda &= \frac{\hat{L_0}}{\hat{L_1}} \\ \text{or} \qquad \log \lambda &= \log(\hat{L_0}) - \log(\hat{L_1}) \end{aligned} \tag{2.2}$$

Unfortunately in the mixture model setup, the test statistic $-2 \log \lambda$ does not have the usual asymptotic null distribution of χ^2_d , d being the difference of the number of parameters under the null and the alternative. This is because the standard regularity conditions(Cramer, 1946) about the asymptotic properties of LRT are not met by this model. First of all, the null hypothesis is in the boundary of the parameter space rather than its interior (Lindsay, 1995), which does not satisfy the conditions of classical theory . Moreover there is no unique way of obtaining H_0 from H_1 (Ghosh and Sen, 1985) making H_0 a non-identifiable subset of the parameter space. This can be illustrated by the following example.

Let us consider the simple example, of testing a one component Binomial against a two component mixture of Binomials. Thus, the likelihood ratio test setting for testing this hypothesis is

$$H_0 : \quad Y \sim \text{Bin}(n, p_0) \quad \text{(1 unknown parameter)}$$

$$H_1 : \quad Y \sim \pi \ \text{Bin}(n, p_1) + (1 - \pi) \ \text{Bin}(n, p_2) \quad \text{(3 unknown parameters)}.$$

For identifiability of the distributions in the alternative hypothesis we restrict $p_1 < p_2$. In this setting the uni-component model, with arbitrary parameter p_0, can be described with many elements of the alternative parameter space. Three lines on the boundary of the parameter space give the same null distribution.

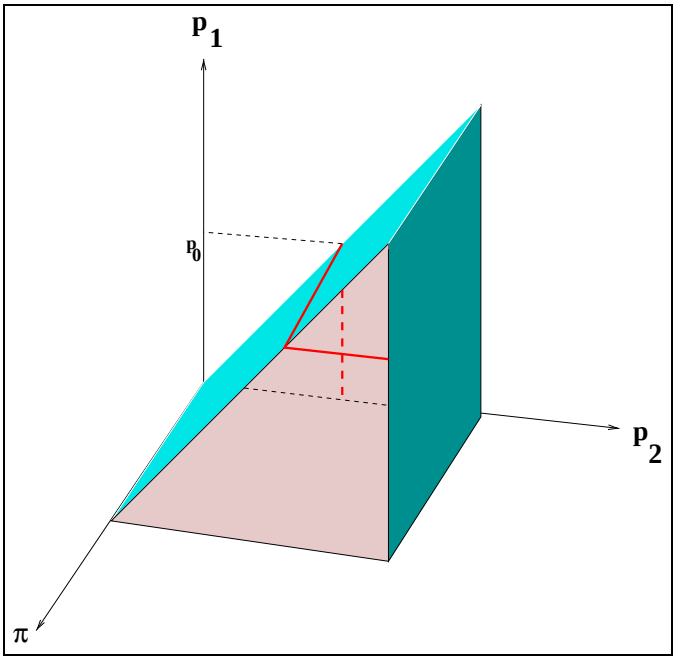

Figure 2.2: 3-D plot of the parameter space of a two component binomial density where the —– and - - - -denotes the possible ways of getting a one component model

We can get H_0 from H_1 in multiple ways.

$$\begin{pmatrix} \pi & 1-\pi \\ p_0 & p_0 \end{pmatrix} \Longleftrightarrow \begin{pmatrix} 1 & 0 \\ p_0 & p \end{pmatrix} \Longleftrightarrow \begin{pmatrix} 0 & 1 \\ p & p_0 \end{pmatrix} \tag{2.3}$$

- The line for $p_1 = p_2 = p_0$, π=anything.
- The line where $\pi = 0$, $p_2 = p_0$ and and p_1=anything.
- The line where $\pi = 0$, $p_1 = p_0$ and and p_2=anything.

In Figure 2.2 the alternative parameter values corresponding to a single null hypothesis p_o

is plotted. Thus, it can be easily seen how the standard condition needed for the asymptotic distribution of LRT's breaks down in this example.

It should be noted that, though $-2\ \log\ \lambda$ does not have a standard χ^2_d distribution, many references can be cited where the distribution of the LRT in the mixture set up has been worked out under certain special cases. The limiting distribution is complex in general making it hard to compute the critical value for the rejection or acceptance of the null hypothesis. In its most general form, the limiting distribution of $-2\ \log\ \lambda$ for normal components with unknown but identifiable mean parameters, is given by,

$$\sup\left[\frac{\mathcal{D}^+(\mu_2)}{Var\left(\mathcal{D}^+(\mu_2)\right)}\right]^2, \tag{2.4}$$

where $\mathcal{D}^+ = \max\{0, D\}$ and $\mathcal{D}^+(\mu_2)$ is a zero mean Gaussian process whose covariance kernel is a function of μ_1 under H_0. (Ghosh and Sen, 1985) derived the same asymptotic distribution for a general parametric family but under the condition of "strong identifiability".

We now discuss the asymptotic distribution of the log-likelihood ratio statistic for some special cases. It can be shown that for a mixture of two known (but general) univariate densities with unknown proportions π and $1-\pi$, the test statistic $-2\ \log\ \lambda$ follows a mixture of χ^2 distributions. For testing $H_0 : g = 1\ (\pi_1 = 1)$ vs $H_1 : g = 2\ (\pi_1 < 1)$,

$$-2\ \log\ \lambda \sim \frac{1}{2}\chi^2_0 + \frac{1}{2}\chi^2_1 \tag{2.5}$$

asymptotically under H_0, where χ^2_0, the "chi-squared with zero degrees of freedom", denotes the degenerate distribution that puts mass 1 at zero. Lindsay (1995, Section 4.2) referred to this as "chi-bar squared distribution"; that is a mixture of chi-squared distributions. Another set of special case was considered by Goffinet et al. (1992). If the component densities are unknown normals, with known mixing proportion we have the chi-bar-squared type distribution.

Bootstrapping the LRTS

One way to avoid the derivation of the complex null distribution of the test statistic is to appeal to bootstrap based calculation of the null distribution. McLachlan (1987) proposed a simple resampling based approach which would enable us to assess the *P*-value of the LRTS. For testing

$$H_0 : g = g_0 \quad vs \quad H_1 : g = g_1$$

bootstrap samples are generated from the mixture model fitted under the null hypothesis of g_0 components. The quantity $-2 \log \lambda$ is computed, one for each bootstrap sample. i.e for each bootstrap sample we fit the model under the null and the alternative hypothesis and calculate the LRTS. This process is repeated B times independently(i.e. for B-bootstrap samples) which enables us to approximate the *P*-value of the LRTS. However, for an accurate estimate of the *P*-value we need B to be large enough (Efron and Tibshirani, 1993). This is one of the drawbacks of these method, as calculation of the LRTS for a single replicate involves considerable amount of computing.

2.2.3 Information criterion based methods

The problem of model selection can be approached with methods based on information criterion. Bias-corrected log likelihood methods are commonly used for the determination of the number of components in a mixture model. In its most general form an information criterion for model selection is based on the bias-corrected log-likelihood given by

$$\log \hat{L} - b(F),$$

where b(F) is the bias corrected term. More commonly it is written in the form

$$-2 \log \hat{L} + 2C, \tag{2.6}$$

where the first term equation 2.6 is the lack of fit and the second term is the complexity of the model and we choose the model which minimizes equation 2.6. In particular AIC selects the

model which minimizes

$$-2\ \log\ \hat{L} + 2\ d, \tag{2.7}$$

where d is the total number of parameters in the model. On the other hand BIC, which has been derived within a Bayesian framework can also be applied in a non-Bayesian sense. Here, we minimize

$$-2\ \log\ \hat{L} + d\ \log\ n, \quad n \text{ being the total sample size.} \tag{2.8}$$

However, the asymptotic expansion that justify the $b(F)$ term in general, depends on the same regularity conditions as the null distribution of the LRTS, which as we have indicated, fail for tests on the number of components of the mixture models. Though in the mixture model scenario AIC tends to overestimate the correct number of components, it is often used to assess the order of a mixture model. Other information criterion based methods include *Bootstrap-Based Information Criterion (EIC)* and *Cross-Validation-Based Information Criterion* (see McLachlan and Peel, 2000, p. 205).

2.2.4 Bayesian approaches

In this subsection we discuss some approaches to the selection of number of components from a purely Bayesian perspective. Raftery (1996) took a Bayes factor based approach for choosing the model along with the number of components of the model. The Bayes factor was calculated as the ratio of the posterior to prior odds. Later Aitkin et al. (1996) used the posterior Bayes factors as a variation of the "prior" Bayes factor described by Raftery (1996). Other approaches to choose the number of components from a fully Bayesian framework was advocated by Philips and Smith (1996), and Richardson and Green (1997). In their approach the number of component g was formulated as the unknown parameter and the model selection was done by attaching a prior on g, along with other parameters for the g component model.

All the Bayesian methods for model selection in the mixture model set up depend heavily on the choice of an appropriate prior. One should note that choosing a fully non-informative prior for the component parameters is not an option in the mixture setup, as there is always a possibility that there are no observations allocated to one or more components, resulting in a divergent posterior (Wasserman, 2000). Moreover, Bayesian methods require a high computational effort even in the univariate case, making their multivariate generalization very computer intensive.

2.2.5 Approaches based on Nonparametric methods

Nonparametric methods for choosing the number of components have been considered by many statisticians. Besides the normal scores plot (Harding, 1948, Cassie, 1954) and the modified percentile plot of Fowlkes (1979), Lindsay and Roeder (1992) proposed the use of residual diagnostics for determining the number of components. Roeder (1994) also demonstrated that a mixture of two univariate normals divided by a normal density having the same mean and variance as the mixture density is always bimodal.

Böhning et al. (1992) introduced a directional derivative based approach for deciding upon the number of components in a mixture model. In this subsection we discuss this approach in detail.

Directional Derivative and Gradient function

Let $\hat{M}$ be the fitted g component mixture. The mixture maximum likelihood theorem (Böhning et al., 1992) states

(a) $\hat{M}$ is the NPMLE if and only if

$$\partial(\lambda, \hat{M}) = \frac{1}{n}\sum_{i}^{n} \frac{f(x_i, \lambda)}{f(x_i, \hat{M})} \leq 1,$$

for all λ in the parameter space.

(b) $\partial(\lambda, \hat{M}) = 0$ for all support points λ of $\hat{M}$.

Table 2.1: Accident data of Thyrion (1960) used by Simar (1976)

y_i	0	1	2	3	4	5	6	7
frequency w_i	7840	1317	239	41	14	4	4	1

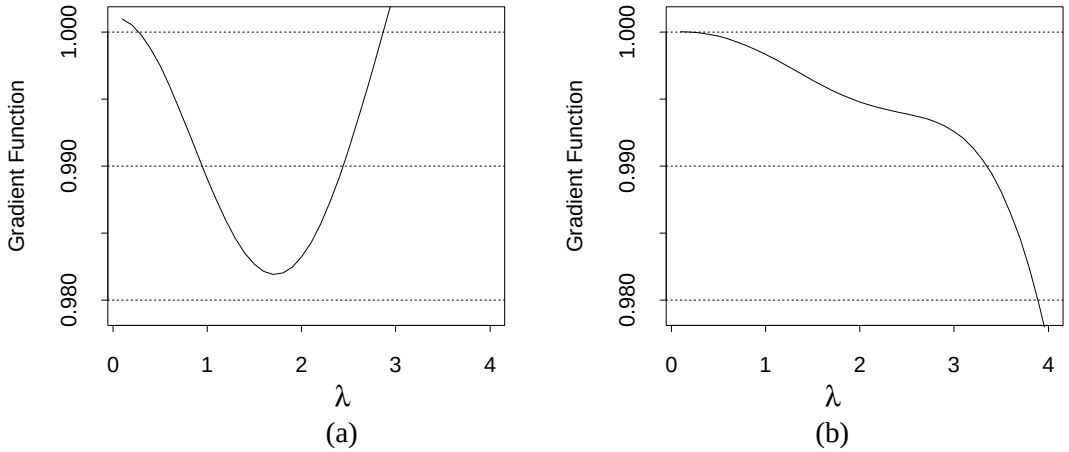

Figure 2.3: Gradient function of two sets of fit for the Simar Data

One application of this theorem is to check the optimality of a candidate mixture distribution M, which in turn can be used in determining the optimum number of components of a mixture model. It can be best illustrated by the following example.

Simar (1976) provided one of the pioneering papers of NPMLE for mixture of Poisson distributions. The study concerned the number of claims filed in a particular year by the policy holders (total n = 9461) of La Royal Belge Insurance. The data (see Table 2.1) goes back to Thyrion (1960) and has been used in many occasions in the literature.

Simar gave an estimate of a 4 component Poisson mixture model.

$$\hat{M} = \begin{pmatrix} .7600 & .2362 & .0037 & .0002 \\ .089 & .580 & 3.176 & 3.669 \end{pmatrix},$$

where the second row of $\hat{M}$ gives the mean of the components, with its corresponding propor-

tion in the first row. This fit has an associated log-likelihood of -5341.5310. This estimator has been reported (Simar, 1976) to be the NPMLE of a mixture of Poisson, but an inspection of the gradient function as given in Figure 2.3(a) shows that it attains values above 1. A fit that is closer to the NPMLE is given by a 3 component model with parameters given by

$$\hat{M} = \begin{pmatrix} .4184 & .573 & .0087 \\ 0.0000 & .3356 & 2.5454 \end{pmatrix}$$

with an associated log-likelihood of -5340.7040. Not only the likelihood is greater, the gradient function is below 1 with high accuracy, as Figure 2.3(b) demonstrates.

One problem with using the NPMLE to estimate the number of components is that the method is not guaranteed to be consistent. For example, if the true number of components is $g_0 = 1$, there is a significant probability of estimating more than one, even asymptotically (Lindsay, 1995).

2.2.6 Moment based approaches

The method of moments has been used by many researchers as an effective tool for choosing the number of components in a normal mixture model. Heckman et al. (1990), and Furman and Lindsay (1994) introduce some elegant tools based on moments to help one decide on the number of components. More recently a kurtosis-based approach has been taken by Vlassis and Likas (1999) and Vlassis et al. (1999).

2.3 Conclusion

Besides the methods discussed in this chapter, an array of other approaches for selecting the number of components have been introduced by researchers. A comprehensive account of these methods can be found in McLachlan and Peel (2000, Chapter 6). All the above methods have been used in numerous instances of model selection in the mixture model setup. AIC and BIC tend to overestimate the number of components when the true situation

has a small number of components. Still they are the most widely used methods for selecting the number of components. Many other methods, including the purely Bayesian approach have the drawback of huge increase in computational effort in multivariate cases.

Chapter 3

Generalized Quadratic Distance Based Model Selection

In this chapter we develop model selection tools based on generalized quadratic distances and discuss some of their properties. The model selection procedures based on the generalized quadratic distance that we discuss will be generic in nature and can be applied to a large number of model selection problems. Later in this chapter, we will discuss how the generalized quadratic distance can be used for selecting the number of components in a multivariate normal mixture model. The first portion of this chapter is joint work with Penn State graduate students Shu-Chuan Chen and Ke Yang.

One attractive feature of model selection tools based on generalized quadratic distances is that they depend on a nonparametric test of model fit, unlike the AIC and BIC, where one only compares models within a class of models. In our particular example of selecting the number of components in the mixture model, AIC and BIC would compare the likelihood of a 6 component model with a 5 component model, but would provide no guarantee that either fits well. On the other hand, if we determine the distribution of the generalized quadratic distances, then, we can test to see if 6 component model lies close to the true distribution in the over all distribution of the distance. Based on the distribution of the distance we can derive rules for global acceptance and rejection of any model.

We will first define the generalized quadratic distance and discuss some of its important properties. Then, we will see how this generalized quadratic distance can be interpreted as a L_2 distance in a smoothed scale. We will also discuss the asymptotic distribution of the

distance in some special cases and how to use them for our model selection purpose. Finally, it will be shown how we can develop a distance based model selection using nonparametric confidence intervals in situations where one might wish to allow for some approximation error in model building.

3.1 Generalized Quadratic Distance

In this section we will define a generalized quadratic distance between two statistical distributions. The generalized quadratic distance is based on a positive definite kernel $K(x,y)$, having certain desirable properties. Besides providing the basic definition of quadratic distance, this section will introduce several examples which are commonly used in goodness-of-fit tests. Before defining the generalized quadratic distance let us introduce some related notations and definitions.

Definition 3.1. *In its most general form, a* ***kernel function*** $K(\mathbf{x},\mathbf{y})$*, on two vector variables* $\mathbf{x}$ *and* $\mathbf{y}$*, is said to be positive definite if for any integer* N*, for any set of vectors* $\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_N$*, and any set of real numbers* $a_1,a_2,\ldots,a_N$

$$\sum_{i=1}^{N}\sum_{j=1}^{N} a_i a_j K(\mathbf{x}_i,\mathbf{x}_j) \geq 0, \tag{3.1}$$

with equality holding iff $a_i = 0\ \forall i$.

From now onwards, for notational ease, we will use unbolded-x to denote a vector as well as a variable which should be understood from the particular context. So $K(x,y)$ and $K(\mathbf{x},\mathbf{y})$ will mean the same thing. Now, we define the generalized quadratic distance based on the positive definite kernel $K(x,y)$.

Definition 3.2. *The generalized* ***quadratic distance*** *between two probability measures* F *and* G*, based on the kernel* K*,is defined by*

$$D_K(F,G) = \int\int K(x,y)d(F-G)(x)d(F-G)(y), \tag{3.2}$$

where $K(x,y)$ is a positive definite kernel function. When F and G are discrete distributions the generalized quadratic distance can be written as

$$D_K(F,G) = \sum_{x\in\mathcal{S}}\sum_{y\in\mathcal{S}} K(x,y)\Big(f(x)-g(x)\Big)\Big(f(y)-g(y)\Big), \tag{3.3}$$

where f and g are the probability mass functions of F and G respectively, and the summations are over $\mathcal{S}$, the joint support of F and G.

Note that in the discrete case the quadratic distance has the matrix representation

$$(\mathbf{f}-\mathbf{g})'K(\mathbf{f}-\mathbf{g}), \tag{3.4}$$

where

$$\begin{aligned}
\mathbf{f} &= (f(x_1), f(x_2), \ldots, f(x_n))',\\
\mathbf{g} &= (g(x_1), g(x_2), \ldots, g(x_n))',\\
K_{i,j} &= K(x_i,x_j)\ \ i,j = 1,2,\ldots,n,\\
&\quad x_1, x_2, \ldots x_n \text{ being the support points in } \mathcal{S}.
\end{aligned}$$

3.1.1 Commonly used Quadratic Distances

Particular forms of the generalized quadratic distances have been used extensively for assessing goodness-of-fit of probability models. For example, under a discrete measure, with the kernel

$$K(x,y) = \begin{cases} 1 & \text{if } x=y,\\ 0 & \text{otherwise,}\end{cases} \qquad \text{i.e.} \quad K = I(\text{Identity Matrix}), \tag{3.5}$$

the distance is given by $D_K(F,G) = \sum_x [f(x)-g(x)]^2$; this is the L_2 distance measure between two densities. To generate the Pearson's χ^2 test statistic, we would use the kernel

$$K(x,y) = \begin{cases} \frac{1}{\sqrt{g(x)g(y)}} & \text{if } x=y,\\ 0 & \text{otherwise,}\end{cases} \tag{3.6}$$

and the distance is given by $D_K(F,G) = \sum_x \frac{[g(x)-f(x)]^2}{g(x)}$.

3.2 General properties of Kernel-based distance and estimation

In this section we will discuss some general properties of kernel-based quadratic distances. Later, we will exploit some of these properties of the kernel-based quadratic distances to find an unbiased estimator of the distance. Let τ (unknown) be the true distribution and G be a candidate estimator for τ. We would like to find the distance between τ and G, i.e. we want to calculate $D_K(\tau, G)$. One way of estimating the distance $D_K(\tau, G)$ on the basis of the data $\hat{F}$ is to use the standard $\mathcal{V}$-*statistic* (von Mises, 1947) results. The $\mathcal{V}$-*statistic* estimate of $D_K(\tau, G)$ is given by

$$D_K(\tau, G) = D_K(\hat{F}, G). \tag{3.7}$$

But, it can be seen that the $E_\tau[D_K(\hat{F}, G)] \neq D_K(\tau, G)$. Instead we propose to use an unbiased estimator for the distance. To develop this, let us define the *G-centered* kernel.

Definition 3.3. *The G-centered kernel K, denoted by $\tilde{K}^G$, is defined as*

$$\begin{aligned}&\tilde{K}^G(x,y)\\ &= K(x,y) - \int_x K(x,y)dG(x) - \int_y K(x,y)dG(y) + \int_x\int_y K(x,y)dG(x)dG(y).\end{aligned} \tag{3.8}$$

We will write the last expression as

$$K(x,y) - K(G,y) - K(x,G) + K(G,G).$$

Lemma 3.1. *Let F and G be two arbitrary distributions. Then the kernel-based quadratic distance can be written as*

$$D_K(F,G) = \int_x\int_y \tilde{K}^G(x,y)dF(x)dF(y). \tag{3.9}$$

Proof :

$$
\begin{aligned}
&\int_x\int_y \tilde{K}^G(x,y)dF(x)dF(y)\\
&= \int_x\int_y \left[K(x,y) - \int_x K(x,y)dG(x) - \int_y K(x,y)dG(y) - \int_x\int_y K(x,y)dG(x)dG(y)\right] dF(x)dF(y)\\
&= \int_x\int_y K(x,y)dF(x)dF(y) - \int_y\int_x K(x,y)dG(y)dF(x) - \int_x\int_y K(x,y)dG(x)dF(y)\\
&\quad + \int_x\int_y K(x,y)dG(x)dG(y)\\
&= \int_x\int_y K(x,y)d(F-G)(x)d(F-G)(y)\\
&= D_K(F,G).
\end{aligned}
$$

□

Equation (3.9) shows that for fixed G, $D_K(F,G)$ can be written as a $\mathcal{U}$-*functional* on F. Thus using an $\mathcal{U}$-*statistic* results from Serfling (1980) we can derive an unbiased estimator of $D_K(\tau,G)$. The result implies that, if X and Y are independent observations from τ, then

$$E_\tau[\tilde{K}^G(X,Y)] \quad = \quad D_K(\tau,G), \tag{3.10}$$

If $X_1,X_2,\ldots,X_n$ is a random sample from τ, then we can estimate the distance $D_K(\tau,G)$ by the $\mathcal{U}$-*statistic* U_n given by

$$U_n(G) \quad = \quad \sum_i\sum_{j\neq i}\frac{1}{n(n-1)}\tilde{K^G}(x_i,x_j). \tag{3.11}$$

In the following example we derive the unbiased estimator for the Pearson's χ^2 estimator.

Proposition 3.1. *The unbiased estimator of the Pearson's* χ^2 *distance is given by*

$$
\begin{aligned}
U_n(G) &= \frac{n}{n-1}\sum_x\left[\left(\frac{\left(\frac{n_x}{n}\right)^2}{g(x)}-1\right) - n_x\left(\frac{1}{g(x)}-1\right)\right]. &(3.12)\\
&= \left(\frac{1}{n-1}\right)V_n(G) - \left(\frac{n}{n-1}\right)\sum_x n_x\left(\frac{1}{g(x)}-1\right), &(3.13)
\end{aligned}
$$

where $g(x)$=$dG(x)$, n_x=number of observations belonging to the cell x, and $V_n(G)$ is the usual Pearson's χ^2 statistic

$$V_n(G) = n\sum_x \left(\frac{\left(\frac{n_x}{n}\right)^2}{g(x)} - 1 \right). \tag{3.14}$$

Proof : The Pearson's χ^2 kernel, given by equation (3.6) can also be written as

$$K(x,y) = \frac{\mathbf{I}[x=y]}{\sqrt{g(x)g(y)}}, \quad \mathbf{I} \text{ being the indicator function.}$$

Thus we have,

$$\begin{aligned}
& \tilde{K}^G(x,y) \\
= & \frac{\mathbf{I}[x=y]}{\sqrt{g(x)g(y)}} - \int_x \frac{\mathbf{I}[x=y]}{\sqrt{g(x)g(y)}} g(x) - \int_y \frac{\mathbf{I}[x=y]}{\sqrt{g(x)g(y)}} g(y) + \int_x \int_y \frac{\mathbf{I}[x=y]}{\sqrt{g(x)g(y)}} g(x)g(y) \\
= & \frac{\mathbf{I}[x=y]}{\sqrt{g(x)g(y)}} - 1 - 1 + 1 \\
= & \frac{\mathbf{I}[x=y]}{\sqrt{g(x)g(y)}} - 1
\end{aligned} \tag{3.15}$$

Thus the biased estimate of the Pearson's χ^2 is given by

$$\begin{aligned}
V_n(G) &= n \int_x \int_y \left(\frac{\mathbf{I}[x=y]}{\sqrt{g(x)g(y)}} - 1 \right) d\hat{F}(x) d\hat{F}(y) \\
&= n\sum_x \left(\frac{\left(\frac{n_x}{n}\right)^2}{g(x)} - 1 \right),
\end{aligned} \tag{3.16}$$

and the unbiased estimate U_n is such that,

$$\begin{aligned}
n(n-1)U_n &= \sum_i \sum_{j \neq i} K^G(x_i, x_j) \\
&= \sum_i \sum_j K^G(x_i, x_j) - \sum_{i=j} K^G(x_i, x_j) \\
&= n^2 \sum_x \left[\left(\frac{\left(\frac{n_x}{n}\right)^2}{g(x)} - 1 \right) - n_x \left(\frac{1}{g(x)} - 1 \right) \right].
\end{aligned} \tag{3.17}$$

□

Now,

$$E_g\left[n^2\sum_x\left(\frac{\left(\frac{n_x}{n}\right)^2}{g(x)}-1\right)\right] = n(\#\text{cells}-1) \tag{3.18}$$

and

$$\begin{aligned} E_g\left[n_x\left(\frac{1}{g(x)}-1\right)\right] &= \sum_x ng(x)\left(\frac{1}{g(x)}-1\right) \\ &= n\sum_x(1-g(x)) \\ &= n(\#\text{cells}-1). \end{aligned} \tag{3.19}$$

Adding (3.18) and (3.19) we get $E_G[U_n(G)]=0$. On the other hand the biased estimate has $E_G[V_n(G)]=(\#\text{cells}-1)$.

3.3 Consistency of estimators

In this section we will discuss the consistency properties of the estimator $D(\hat{F},M)$. It is clear from the construction of $D(\hat{F},M)$, that if the kernel $K(x,y)$ is bounded and continuous in (x,y), then the weak convergence of $\hat{F}_n$ to τ implies that

$$D(\hat{F},M) \xrightarrow{\mathbf{P}} D(\tau,M) \text{ as } n\to\infty, \tag{3.20}$$

when M is any fixed distribution. However, if M is estimated using $\hat{M}_n$, where $\hat{M}_n$ is a model fit by maximum likelihood, the statement

$$D(\hat{F}_n,\hat{M}_n) \xrightarrow{\mathbf{P}} D(\tau,\hat{M}_n) \tag{3.21}$$

is false because the right-hand side depends on n. Rather, one has

$$D(\hat{F}_n,\hat{M}_n) \xrightarrow{\mathbf{P}} D(\tau,M^*_\tau), \tag{3.22}$$

provided that the estimator $\hat{M}_n \xrightarrow{\mathbf{d}} M^*_\tau$. If $\hat{M}$ is based on maximum likelihood then M^*_τ will minimize the Kullback-Leibler distance between τ and $\mathcal{M}$, not $D(\tau,M)$, and so $D(\tau,M^*_\tau)$ will be greater than the minimum distance $\min_{M\in\mathcal{M}} D(\tau,M)=D(\tau,M_\tau)$. If we instead want to estimate

$D(\tau, M_\tau)$ we could estimate M by minimizing the empirical distance $D(\hat{F}, M)$ over the relevant parameters

$$\min_\theta D(\hat{F}, M_\theta) \quad \xrightarrow{\mathbf{P}} \quad D(\tau, M_\tau). \tag{3.23}$$

We will focus on the distance estimator $D(\hat{F}, \hat{M}_n)$ for all our model selection rules, where $\hat{M}_n$ is the MLE, because we believe potential users will find this more natural.

3.4 Choice of kernel

In this chapter, so far, we have discussed quadratic distances with arbitrary kernels. In this section we will define a natural kernel for the problem of selecting the number of components in a mixture of normals. Theoretically, any kernel could be used for this model selection problem. But, before defining the kernel, we should note that to make the distance practical for model selection the estimates should be easily computable. Also, we should note that we will mostly deal with multivariate data, potentially with a huge number of variables, so we would prefer to have a closed form for the integrals in the distance. To make an optimum choice for the kernel we will make use of the convolution properties of the normal distribution. We use the following well known result:

$$\text{if } \phi(x;\mu,\Sigma) \text{ is the density function of } X \sim \mathcal{N}(\mu,\Sigma) \tag{3.24}$$

$$\text{then } \int_y \phi(x;y,\Sigma_1)\phi(y;\mu,\Sigma_2)dy = \phi(x;\mu,\Sigma_1+\Sigma_2). \tag{3.25}$$

In other words if we define a kernel $K_\Sigma(x,y)$ by,

$$K_\Sigma(x,y) = \frac{1}{(2\pi)^{\frac{p}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x-y)'\Sigma^{-1}(x-y)\right),$$

then we have

$$\int_y K_{\Sigma_1}(x,y)K_{\Sigma_2}(y,z)dy = K_{\Sigma_1+\Sigma_2}(x,z) \tag{3.26}$$

In mathematical language, the normal kernels form an *additive semigroup* of *Hilbert-Schmidt* operators. Kernels other than the normal kernels also have this property. So, we choose the kernel in equation (3.4), which will be referred to as the "normal kernel". As it will be seen later in Section 3.9.1, the normal kernel will give a closed form for the centered-kernel and thus the distance calculation will need no numerical integration. Property (3.26) will be used to show the kernel has an explicit square root and this will be used to prove positive definiteness.

The normal kernel we will use for our model selection will have $\Sigma = h^2 I$, h being a spherical "smoothing parameter". Thus our model selection will be based on the normal kernel with only one smoothing parameter and will be defined by

$$K_h(x,y) = \frac{1}{(2\pi)^{\frac{p}{2}}|\Sigma_h|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x-y)'\Sigma_h^{-1}(x-y)\right), \text{ where } .\Sigma_h = h^2 I, \tag{3.27}$$

The choice of h is a very important issue, and will be discussed extensively in Chapter 4. Here, we should note that we could have chosen an arbitrary matrix Σ to define our kernel. But in that case we would have to choose a large number of "smoothing parameters" to define the kernel completely. However, if other smoothing "shapes" are desired, our theory could be easily extended to the kernels of the form $\Sigma = h^2 V$.

3.5 Model selection with Quadratic Distances

In this section we will describe the steps we will follow for model selection. In our case, the competing models are the probability distributions given by multivariate normal mixtures with different number of components. So our models are in the set $\mathcal{M} = \{M\}$ where the elements of the model will typically be indexed by an integer g, the number of components, so that $\mathcal{M} = \{M_g : g \in I\!I\}$. Or in other words, among the competing models $M_1 = \mathcal{N}(\mu, \Sigma), M_2, \ldots, M_l = \left\{\sum_{i=1}^{l} \pi_i \mathcal{N}(\mu_i, \Sigma_i)\right\}$, we want to decide which one is the best fit for the true distribution τ. This will be based on the distances $D(\tau, M_1^*), D(\tau, M_2^*), \ldots, D(\tau, M_l^*)$ which

in turn are estimated by $D(\hat{F},\hat{M}_1), D(\hat{F},\hat{M}_2)\ldots, D(\hat{F},\hat{M}_l))$ respectively. Here $M_1^*, M_2^*, \ldots, M_l^*$ correspond to the weak limits of the corresponding maximum likelihood estimators when τ is correct.

The model selection problem we are interested in can be written as a hypothesis testing problem as follows

$$H_0 : g = g_0 \qquad \text{vs} \qquad H_0 : g > g_0, \tag{3.28}$$

where g is the "true number of components in the mixture". Equivalently, $H_0 : \tau \in \mathcal{M}_{g_0}$, where τ is the true distribution. We will consider the use of the empirical distance $D(\hat{F},\hat{M}_g)$ or a standardized version of it as the test statistic. Ideally, to accept or reject some model we would like to have a simple null distribution for the test statistic $D(\hat{F},\hat{M}_g)$, along with a acceptance (rejection) region . For particular choices of K we can derive an asymptotic null distribution (discussed in Section 3.6), but in general it has unknown parameters. Due to this, we take a resampling based approach (Section 3.9) to estimate the null distribution of $D_K(\hat{F},\tau)$ and decide on the choice of model.

3.6 Asymptotic distribution of the generalized quadratic

Although it is impractical to derive an exact theoretical distribution for the unbiased estimator of $D_K(\tau, M)$ for an arbitrary kernel K, model M, and true distribution τ, we can develop some asymptotic results. We should also note that the *$\mathcal{U}$-statistic* estimates of $D(\tau, M)$ will have easy asymptotic distributions when $M \neq \tau$, being a normal with calculable variance, but the case of $D(\hat{F},\tau)$, which estimates zero, is more complicated. So we will need to distinguish between the asymptotic distribution for $D(\hat{F}, M)$ when $M = \tau$ and when $M \neq \tau$. In this section we will first work out the asymptotic distribution of the distance when $M = \tau$, and then we solve the asymptotic normality when $M \neq \tau$.

3.6.1 Asymptotic distribution of generalized quadratic distance under the null distribution

In this part of the thesis we will find the asymptotic distribution of the generalized quadratic distance. First we will show that positive definite kernels in two variables can be written as a infinite sum using eigenvalues and eigen-functions. We will then use this decomposition to find the asymptotic distribution. First, we will state the theorem for the general spectral decomposition of kernel based distance.

Theorem 3.2. *Let $K(x,y)$ be a real-valued $\mathfrak{B}$-measurable positive definite kernel function on a measure space $(S,\mathfrak{B},m)$ such that*

$$\int_S\int_S |K(x,y)|^2 m(dx)m(dy) < \infty. \tag{3.29}$$

Let K be the integral operator defined by the kernel $K(x,y)$:

$$(Kf)(x) = \int_s K(x,y)\,f(y)m(dy), \quad f \in L^2(S) = L^2(S,\mathfrak{B},m).$$

Then $K(x,y)$ can be written as

$$K(x,y) = \sum_{j=1}^{\infty} \lambda_j f_j(x)\, f_j(y), \tag{3.30}$$

where λ_j's and $f_j(x)$'s are eigenvalues and corresponding normalized eigen-vectors of K, that is

$$(Kf_j)(x) = \int_S K(x,y)\,f_j(y)m(dy) = \lambda_j f_j(x), \;\; ||f_j||^2 = \int_S f_j^2(y)m(dy) = 1.$$

The series in (3.30) converges strongly to K, that is

$$\lim_{n\to\infty}\int_S \left(\int_S K(x,y)g(y)m(dy) - \sum_{j=1}^{n}\int_S \lambda_j f_j(x)\,f_j(y)g(y)m(dy)\right)^2 m(dx) = 0, \; \forall g \in L^2(S).$$

Moreover, $\lambda_j \geq 0$ since K is positive definite.

Here we will give a background of the above theorem instead of the detailed proof. A kernel $K(x,y)$ that satisfies (3.29) is said to be of the *Hilbert-Schmidt type*. It can be shown

that K is compact as an operator $\in L(L^2(S), L^2(S))$. If $K(x,y)$ is real-valued and symmetric, K is a self-adjoint operator, or say, symmetric transformation, that is $(Kf, g) = (f, Kg)$. The decomposition of $K(x,y)$ given in equation (3.30) corresponds to the spectral decomposition for a *compact, self-adjoint operator*. More about the operator's spectral decomposition can be found in Yosida (1980) and Riesz and Sz.-Nagy (1990).

This theorem requires only a weak assumption (equation (3.29)), which is satisfied by the normal kernel, as

$$\sup_{x,y} K_h^2(x,y) = \frac{1}{(2\pi h)^p} < \infty,$$

and so we have

$$\int\int |K(x,y)|^2 d\tau(x) d\tau(y) < \sup_{x,y} K_h^2(x,y) \int\int d\tau(x) d\tau(y) = \sup_{x,y} K_h^2(x,y) < \infty.$$

Under the theorem of spectral decomposition we have

$$\sum_{j=1}^{\infty} \lambda_j^2 = \int_S \int_S |K(x,y)|^2 m(dx) m(dy) < \infty.$$

But many kernel functions satisfy even a stronger condition, that is $\sum_{j=1}^{\infty} \lambda_j < \infty$. The operators defined by those kernel are called nuclear. Now,

$$\sum_{j=1}^{\infty} \lambda_j = \int K(x,x) m(dy), \tag{3.31}$$

and in fact the normal kernel is nuclear as

$$K(x,x) = \frac{1}{(2\pi)^{\frac{p}{2}} |\Sigma_h|^{\frac{1}{2}}} < \infty \quad \forall x,$$

which in turn implies

$$\begin{aligned} \int K(x,x) d\tau(x) &= \frac{1}{(2\pi)^{\frac{p}{2}} |\Sigma_h|^{\frac{1}{2}}} d\tau(x) \\ &= \frac{1}{(2\pi)^{\frac{p}{2}} |\Sigma_h|^{\frac{1}{2}}} < \infty \end{aligned}$$

Next we will use the nuclear property of kernels to find the asymptotic distribution of kernel based distances.

Theorem 3.3. *Let $X_1, X_2, ..., X_n, ...$ be a sequence of random variables such that $\forall n$,*

$$X_n = \sum_{i=1}^{\infty} \lambda_i Z_{ni}^2$$

where

$$\sum_{i=1}^{\infty} \lambda_i < \infty, \quad \lambda_i \geq 0, \quad E(Z_{ni}^2) = 1, E(Z_{ni}) = 0, \forall i, \forall n, \quad \text{and } E(Z_{ni}Z_{nj}) = 0, \forall i \neq j, \forall n.$$

Moreover we assume that for every finite $k = 1, 2, 3, ...$

$$\begin{pmatrix} Z_{n1} \\ Z_{n2} \\ \vdots \\ Z_{nk} \end{pmatrix} \xrightarrow{d} N(0, I_k), \text{ as } n \to \infty.$$

Then X_n satisfies

$$X_n \xrightarrow{d} \sum_{i=1}^{\infty} \lambda_i Z_i^2,$$

where Z_i's are independent $N(0,1)$.

Proof : Let us decompose $X_n = \sum_{i=1}^{\infty} \lambda_i Z_{ni}^2$ into two parts, the partial sum

$$S_{nk} = \sum_{i=1}^{k} \lambda_i Z_{ni}^2$$

and the remainder sum

$$R_{nk} = \sum_{i=k+1}^{\infty} \lambda_i Z_{ni}^2.$$

By Tchebysheff's inequality, for each $\delta > 0$,

$$\begin{aligned} P(R_{nk} > \delta) &\leq \frac{E(R_{nk})}{\delta} \\ &= \frac{E(\sum_{i=k+1}^{\infty} \lambda_i Z_{ni}^2)}{\delta} \\ &= \frac{\sum_{i=k+1}^{\infty} \lambda_i}{\delta}. \end{aligned} \tag{3.32}$$

Since $\sum_{i=1}^{\infty}\lambda_i < \infty$, it follows from this inequality that for any given $\varepsilon > 0$ and $\delta > 0$, we can choose $k_0 = k_0(\varepsilon,\delta)$ such that $\forall n$

$$P(R_{nk} > \delta) \quad < \quad \varepsilon, \text{ for all } k > k_0. \tag{3.33}$$

Fix $t \geq 0$. We now show that $\lim_{n\to\infty} P(X_n \leq t) = P(\sum_{i=1}^{\infty}\lambda_i Z_i^2 \leq t)$ for every t.

First, because $X_n \leq t \Rightarrow S_{nk} \leq t\ \forall k$,

$$P(X_n \leq t) \quad \leq \quad P(S_{nk} \leq t), \forall k. \tag{3.34}$$

Also,

$$\begin{aligned}
P(X_n \leq t) &\geq P(X_n \leq t \cap R_{nk} \in [0,\delta]) \\
&= P(S_{nk} + R_{nk} \leq t \cap R_{nk} \in [0,\delta]) \\
&\geq P(S_{nk} \leq t-\delta \cap R_{nk} \in [0,\delta]) \text{ [nested set]} \\
&= P(S_{nk} \leq t-\delta) - P(S_{nk} \leq t-\delta, R_{nk} > \delta) \\
&\geq P(S_{nk} \leq t-\delta) - P(R_{nk} > \delta) \\
&\geq P(S_{nk} \leq t-\delta) - \varepsilon\ ,\ \forall k \geq k_0.
\end{aligned} \tag{3.35}$$

Thus, from (3.34) and (3.35), we have for all n and for all $k \geq k_0(\varepsilon,\delta)$

$$P(S_{nk} \leq t-\delta) - \varepsilon \leq P(X_n \leq t) \leq P(S_{nk} \leq t).$$

Now, letting $n \to \infty$, since $S_{nk} \to S_k = \sum_{i=1}^{k}\lambda_i Z_i^2$, by the normality assumption, for all $k \geq k_0$ we have

$$\begin{aligned}
P(S_k \leq t-\delta) - \varepsilon &\leq \liminf_{n\to\infty} P(X_n \leq t) \\
&\leq \limsup_{n\to\infty} P(X_n \leq t) \\
&\leq P(S_k \leq t).
\end{aligned} \tag{3.36}$$

Also, since $\{S_k \le t\} \supset \{S_{k+1} \le t\}$, so the sets are decreasing in k, for any t, we have

$$\begin{aligned} \lim_{k\to\infty} P(S_k \le t) &= P(\lim_{k\to\infty}(S_k \le t)) && (3.37)\\ &= P(S_\infty \le t), \quad \text{where } S_\infty = \sum_{i=1}^{\infty} \lambda_i Z_i^2. && (3.38) \end{aligned}$$

Therefore, from (3.36), letting $k \to \infty$,

$$\begin{aligned} P(S_\infty \le t-\delta) - \varepsilon &\le \liminf_{n\to\infty} P(X_n \le t)\\ &\le \limsup_{n\to\infty} P(X_n \le t)\\ &\le P(S_\infty \le t). && (3.39) \end{aligned}$$

The above result is true for all $\varepsilon > 0$ and $\delta > 0$, so we can let $\varepsilon \to 0$, to obtain

$$\begin{aligned} P(S_\infty \le t-\delta) &\le \liminf_{n\to\infty} P(X_n \le t)\\ &\le \limsup_{n\to\infty} P(X_n \le t)\\ &\le P(S_\infty \le t). && (3.40) \end{aligned}$$

Finally, letting $\delta \to 0$,

$$\begin{aligned} P(S_\infty < t) &\le \liminf_{n\to\infty} P(X_n \le t)\\ &\le \limsup_{n\to\infty} P(X_n \le t)\\ &\le P(S_\infty \le t). && (3.41) \end{aligned}$$

However S_∞ has a continuous distribution, so

$$\lim_{n\to\infty} P(X_n \le t) = P(S_\infty \le t)$$

as needed to finish the proof. □

Next, using the spectral decomposition, the estimate of the distance can be written as follows,

$$\begin{aligned} D_K(\hat{F},\tau) &= \int\int \tilde{K}(x,y)d\hat{F}(x)d\hat{F}(y) \\ &= \int\int \sum_{j=1}^{\infty} \lambda_j f_j(x) f_j(y) d\hat{F}(x)d\hat{F}(y) \\ &= \int\int \sum_{j=1}^{\infty} \lambda_j f_j(x) f_j(y) d\hat{F}(x)d\hat{F}(y) \\ &= \sum_{j=1}^{\infty} \lambda_j \left(\int_x f_j(x)d\hat{F}(x)\right)\left(\int_y f_j(y)d\hat{F}(y)\right) \\ &= \sum \lambda_{j=1}^{\infty} \left[\bar{f_j}\right]^2 \\ &\quad \text{where } \bar{f_j} = \frac{1}{n}\sum_i f_j(x_i) \end{aligned} \tag{3.42}$$

The f_j's are mean 0 and $f_j \perp f_k$ for all $j \neq k$. Hence $Z_{nj} = \bar{f_j}$ satisfy the assumptions of Theorem 3.3.

Now, using Theorem 3.3 we have

$$D_K(\hat{F},\tau) \xrightarrow{d} \sum_{i=1}^{\infty} \lambda_i Z_i^2 \tag{3.43}$$

where $Z_i s$ are independent $N(0,1)$. Thus, the, the limiting distribution is $\sum \lambda_i \chi_i^2(1)$ where the χ_i^2 are independent and each $\chi_i^2(1)$ is a one degree of freedom chi-square variable.

3.6.2 Asymptotic distribution of generalized quadratic distance under the alternative

Now we will show that when $M \neq \tau$ the asymptotic distribution of the distance is normal. Using the standard results of of asymptotic properties of U-statistics (Hoeffding, 1948) it can be shown that the distance is asymptotically normal if the following conditions are satisfied.

Theorem 3.4. *(Hoeffding, 1948) If* $E_\tau[\tilde{K}^2(X,Y)] < \infty$ *and* $\psi_1 = Var_y[E_x(\tilde{K}(X,Y))] > 0$ *then*

$$\sqrt{n}U_n \xrightarrow{\mathbf{d}} \mathcal{N}(0, 4\psi_1) \tag{3.44}$$

For the quadratic kernel

$$E_x(\tilde{K}(X,Y)) = K(\tau,y) - K(\tau,M) - K(M,y) + K(M,M) \tag{3.45}$$

Thus when $M = \tau$, $E_x(\tilde{K}(X,y)) = 0$, implying $\psi_1 = 0$ and so asymptotic normality does not hold in this case. On the other hand, since K is a strictly positive definite kernel if $M \neq \tau$ we will have then $D_K(\tau, M) \neq 0$. Thus if $M \neq \tau$ we always have $\psi_1 > 0$. We have already shown that for our kernel, $E_\tau[\tilde{K}^2(X,Y)] < \infty$. Thus when $M \neq \tau$ we have

$$\sqrt{n}U_n \xrightarrow{\mathbf{d}} \mathcal{N}(0, 4\psi_1) \tag{3.46}$$

where the explicit form of ψ_1 will be computed in Section 3.12.

3.7 Interpretation of generalized quadratic distance as an L_2 distance in smoothed scale

A kernel based quadratic distance between two statistical distributions has a nice interpretation as the L_2 distance between kernel-smoothed version of their densities. In this section we will see how this can be done. We will also provide a comparison between kernel-based densities with different smoothing parameters, the L_2 distance, and the L_1 distance.

3.7.1 L_2 distance in smoothed scale

We have defined the kernel-based distance on positive definite kernels. Analogous to the positive definite matrices, associated with every positive definite kernel there is a square-root kernel. If the positive definite kernel has a spectral decomposition

$$K(x,y) = \sum_{j=1}^{\infty} \lambda_j f_j(x) f_j(y),$$

the square root kernel, $L(x,y)$ can be written as

$$L(x,y) = \sum_{j=1}^{\infty} \sqrt{\lambda_j} f_j(x) f_j(y), \tag{3.47}$$

and it satisfies the relation

$$K(x,y) \quad = \quad \int_w L(x,w)L(w,y)dw. \tag{3.48}$$

For example, if K_h is the normal kernel given in (3.4), then the square root kernel is given by

$$L(x,y) = K_{\frac{h}{\sqrt{2}}}(x,y) = \frac{1}{(2\pi)^{\frac{p}{2}}\left(\frac{h}{\sqrt{2}}\right)^p}\exp\left(-\frac{1}{2}\frac{(x-y)'(x-y)}{\frac{h^2}{2}}\right). \tag{3.49}$$

The above follows from the properties of normal convolution.

In general, using the relation of equation (3.48) the distance distance between two measures F and G based on the kernel K, given by $D_K(F,G)$ can be written as

$$\begin{aligned} D_K(F,G) \quad &= \int_x\int_y K(x,y)d(F-G)(x)d(F-G)(y) \\ &= \int_x\int_y\left(\int_w L(x,w)L(w,y)\right)d(F-G)(x)d(F-G)(y)dw \\ &= \int_w\left(\int_x L(x,w)d(F-G)(x)\right)^2 dw \\ &= \int_w\left(f^*(w)-g^*(w)\right)^2 dw \end{aligned} \tag{3.50}$$

where $f^*(w) = \int L(x,w)dF(x)$ and $g^*(w) = \int L(x,w)dG(x)$. Note that for the normal kernel, (3.50) implies that it is positive definite. We can interpret f^* and g^* as "kernel smoothed" densities for F and G. Note that, even though the original distributions are discrete, their kernel-smoothed versions are always continuous and so we can find the distance between a discrete and a continuous distribution with the help of generalized quadratic distance. Thus, generalized quadratic distances can be easily used to find the distance between an empirical density, which is inherently discrete, and its fitted continuous density.

3.7.2 Example: Galaxy Data

Mixture analysis of the galaxy data was introduced by Roeder (1990). the dataset consists of velocity of 82 galaxies moving away from our galaxy. These 82 galaxies are

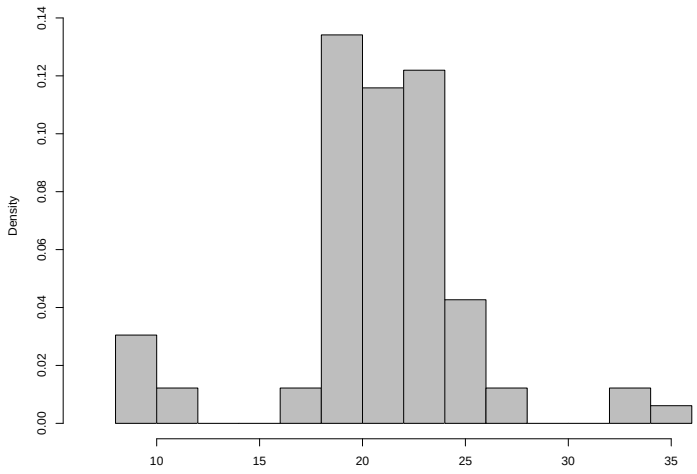

Figure 3.1: **Histogram of Galaxy Data.**

believed to belong to some distinct clusters. This data set has also been analyzed by several authors. In this example we fit a five component normal mixture to the galaxy data. The best 5 component fit to the galaxy data is given by

$$G = .08\ \ \mathcal{N}(9.71, 0.18) + .02\ \ \mathcal{N}(16.13, 0.01) + .4\ \ \mathcal{N}(19.79, .45) \\ + .42\ \ \mathcal{N}(22.92, 1.44) + .02\ \ \mathcal{N}(26.98, 0.01) + .04\ \ \mathcal{N}(33.04, 0.85). \quad (3.51)$$

This was given by McLachlan and Peel (2000). Also, for the normal kernel, the square root of K is another normal kernel, so $f^*(w) = \int L(x,w)d\hat{F}(x)$ is just a kernel smoothed density estimator of $\hat{F}$ and the $g^*(w) = \int L(x,w)dG(x)$ is the kernel smoothed version of the 5 component normal. Thus, the kernel-based distance between $\hat{F}$ and G can be easily interpreted as the L_2 distance between f^* and g^*.

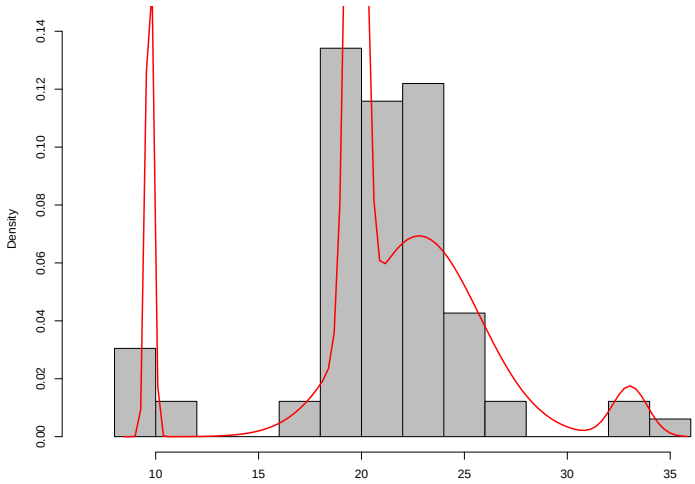

Figure 3.2: **Histogram of Galaxy Data with the 5 component fitted mixture of normals given by** (3.51)**.**

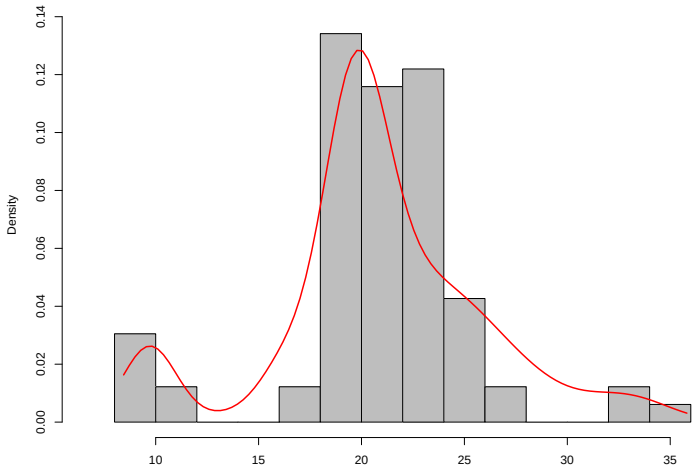

Figure 3.3: **Histogram of Galaxy Data and the smoothed 5 component mixture of normals.**

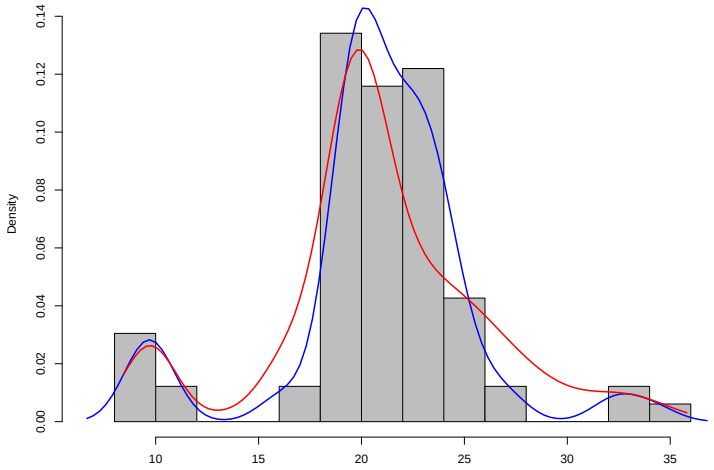

Figure 3.4: Histogram of Galaxy Data, the smoothed empirical density (—) and the smoothed 5 component mixture of normals(—).

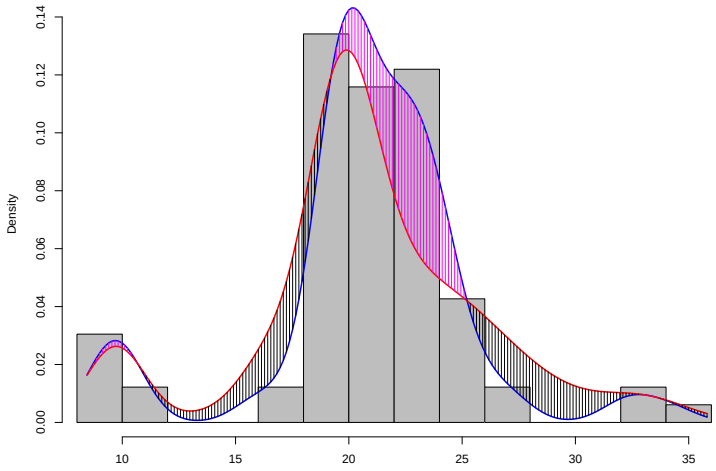

Figure 3.5: Histogram of Galaxy Data, the smoothed empirical density and the smoothed 5 component mixture of normals. The difference between the two densities is the shaded region.

Figures 3.1 to 3.5 illustrate the L_2 interpretation with the help of some graphical displays. Figure 3.1 displays the probability histogram of the galaxy data. Figure 3.2 displays the smoothed fitted 5 component normal (smoothing parameter h=.5) overlaid on the histogram of the galaxy data. Figure 3.4 displays the smoothed fitted 5 component normal(—) and the smoothed empirical density(—) overlaid on the histogram. Figure 3.4 displays smoothed fitted 5 component normal, the smoothed empirical density, and the their difference. The integral of the squared difference is the generalized quadratic distance.

3.8 Comparison of Generalized Quadratic distance with other measures of distance

In this section we will compare the theoretical values of the generalized quadratic distance, the L_2 distance and the L_1 distance for the mixture of two multivariate normals. First, we calculate the theoretical values of the generalized quadratic distance. The L_2 distance will be obtained as a special case of the generalized quadratic distance. Finally, we calculate the L_1 distance and compare them all.

3.8.1 Quadratic distance between two mixture of normals

Proposition 3.2. *Let f_1 and f_2 be two multivariate normal mixture densities with*

$$f_1(x) = \sum_{l=1}^{g_1} \pi_{1l}\phi(x,\mu_{1l},V_{1l}) \quad \textit{and} \quad f_2(x) = \sum_{k=1}^{g_2} \pi_{2k}\phi(x,\mu_{2k},V_{2k}) \tag{3.52}$$

Then the generalized quadratic distance based on kernel K_Σ is given by

$$D_K(F_1,F_2) = \sum_{i=1}^{g_1}\sum_{j=1}^{g_1}\pi_{1i}\pi_{1j}K_{W_{ij}^{11}}(\mu_{1i},\mu_{1j}) + \sum_{i=1}^{g_2}\sum_{j=1}^{g_2}\pi_{2i}\pi_{2j}K_{W_{ij}^{22}}(\mu_{2i},\mu_{2j}) - 2\sum_{i=1}^{g_1}\sum_{j=1}^{g_2}\pi_{1i}\pi_{2j}K_{W_{ij}^{12}}(\mu_{1i},\mu_{2j}), \tag{3.53}$$

where $W_{ij}^{lk} = \Sigma + V_{li} + V_{kj}, \quad i = 1,2,\ldots,g_1, \quad j = 1,2,\ldots,g_2.$

Proof : By equation (3.50) we have,

$$\begin{aligned} D_K(F_1,F_2) &= \int_x (f_1^*(x) - f_2^*(x))^2 dx \\ &= \int_x \left[f_1^*(x)^2 - 2 f_1^*(x) f_2^*(x) + f_2^*(x)^2 \right] dx \end{aligned} \tag{3.54}$$

where $f_1^*(x) = \int_w K_{\frac{\Sigma}{2}}(x,w)\phi(w)dw$. Now using the convolution properties of the normal,

$$f_1^*(x) = \int_w K_{\frac{\Sigma}{2}}(x,w) f_1(w) dw = \sum_{i=1}^{g_1} \pi_{1i} K_{\frac{\Sigma}{2}+V_{1i}}(x,\mu_{1i}), \tag{3.55}$$

which implies,

$$\begin{aligned} \int_x (f_1^*(x))^2 dx &= \int_x \left(\sum_{i=1}^{g_1} \pi_{1i} K_{\frac{\Sigma}{2}+V_{1i}}(x,\mu_{1i}) \right) \left(\sum_{j=1}^{g_1} \pi_{1j} K_{\frac{\Sigma}{2}+V_{1j}}(x,\mu_{1j}) \right) dx \\ &= \sum_{i=1}^{g_1}\sum_{j=1}^{g_1} \pi_{1i}\pi_{1j} K_{W_{ij}^{11}}(\mu_{1i},\mu_{1j}). \end{aligned} \tag{3.56}$$

Similarly, $$\int_x (f_2^*(x))^2 dx = \sum_{i=1}^{g_2}\sum_{j=1}^{g_2} \pi_{2i}\pi_{2j} K_{W_{ij}^{22}}(\mu_{2i},\mu_{2j}) \tag{3.57}$$

and $$\int_x f_1^*(x) f_2^*(x) dx = \sum_{i=1}^{g_1}\sum_{j=1}^{g_2} \pi_{1i}\pi_{2j} K_{W_{ij}^{12}}(\mu_{1i},\mu_{2j}). \tag{3.58}$$

Putting the values from equations (3.56), (3.57) and (3.58) in equation (3.54) we have the proof of the proposition. □

3.8.2 Comparison with other distance

To compare the quadratic distance with other distances we look at the simplest case of distance between two one component univariate normals. The quadratic distance between $f_1 \sim \mathcal{N}(\mu_1,\sigma_1)$ and $f_2 \sim \mathcal{N}(\mu_2,\sigma_2)$ can be simplified to

$$\frac{1}{\sqrt{2\pi}} \left(\frac{1}{\sqrt{2\sigma_1^2+h^2}} + \frac{1}{\sqrt{2\sigma_2^2+h^2}} - 2\frac{\exp\left(-\frac{1}{2}\frac{(\mu_1-\mu_2)^2}{(\sigma_1^2+\sigma_2^2+h^2)}\right)}{\sqrt{\sigma_1^2+\sigma_2^2+h^2}} \right).$$

But we should note that quadratic distances are not scale invariant. So to make the distances comparable they are scaled. A scale invariant version of $D_k(F_1,F_2)$, denoted by $D_k^*(F_1,F_2)$, is

defined as,

$$\begin{aligned} D_k^*(F_1,F_2) &= \frac{D_k(F_1,F_2)}{\int_x f_1^{*^2}(x)dx+\int_x f_2^{*^2}(x)dx} \\ &= \frac{\int_x f_1^*(x)^2-2f_1^*(x)f_2^*(x)+f_2^*(x)^2dx}{\int_x f_1^{*^2}(x)dx+\int_x f_2^{*^2}(x)dx} \\ &= 1-2\frac{\int_x f_1^*(x)f_2^*(x)dx}{\int_x f_1^{*^2}(x)dx+\int_x f_2^{*^2}(x)dx} \\ &= 1-2\frac{\sum_{i=1}^{g_1}\sum_{j=1}^{g_2}\pi_{1i}\pi_{2j}K_{W_{ij}^{12}}(\mu_{1i},\mu_{2j})}{\sum_{i=1}^{g_1}\sum_{j=1}^{g_1}\pi_{1i}\pi_{1j}K_{W_{ij}^{11}}(\mu_{1i},\mu_{1j})+\sum_{i=1}^{g_2}\sum_{j=1}^{g_2}\pi_{2i}\pi_{2j}K_{W_{ij}^{22}}(\mu_{2i},\mu_{2j})}. \end{aligned} \tag{3.59}$$

This scaling will return in Chapter 5, where we define and discuss "discordance". Thus in the univariate one component case the scaled distance will be

$$1-2\frac{\frac{\exp\left(-\frac{1}{2}\frac{(\mu_1-\mu_2)^2}{(\sigma_1^2+\sigma_2^2+h^2)}\right)}{\sqrt{\sigma_1^2+\sigma_2^2+h^2}}}{\frac{1}{\sqrt{2\sigma_1^2+h^2}}+\frac{1}{\sqrt{2\sigma_2^2+h^2}}}.$$

It can be easily noticed that if $f_1=f_2$, implying $\mu_1=\mu_2$ and $\sigma_1=\sigma_2$, the distance is 0. Also if $(\mu_1-\mu_2)^2\to\infty$ the distance goes to 1. In general, for fixed (σ_1^2,σ_2^2) the distance between densities is a increasing function of the distance between their respective means. Note that the L_2 distance corresponds to the generalized quadratic distance with smoothing parameter being $h=0$. Thus the scaled L_2 distance between f_1 and f_2 is

$$L_2(F_1,F_2)=1-2\frac{\frac{\exp\left(-\frac{1}{2}\frac{(\mu_1-\mu_2)^2}{(\sigma_1^2+\sigma_2^2)}\right)}{\sqrt{\sigma_1^2+\sigma_2^2}}}{\frac{1}{\sqrt{2\sigma_1^2}}+\frac{1}{\sqrt{2\sigma_2^2}}}. \tag{3.60}$$

In general the L_1 distance between f_1 and f_2 is given by

$$L_1(F_1,F_2) = \frac{1}{2}\int\left|f_1(x)-f_2(x)\right|dx.$$

For $f_1\sim\mathcal{N}(\mu_1,\sigma)$ and $f_2\sim\mathcal{N}(\mu_2,\sigma)$ the L_1 distance is

$$L_1(F_1,F_2) = 1-2\Phi\left(-\frac{|\mu_2-\mu_1|}{2}\right), \tag{3.61}$$

where $\Phi(x)$is the normal probability integral. Note that the L_1 and the scaled L_2 and the scaled quadratic distance all lie between 0 and 1, and attain the value 1 for $|\mu_2 - \mu_1| \longrightarrow \infty$.

The L_1 distance is important to our consideration because of its properties. Although, the Kullback-Leibler distance is often used as the standard statistical distance due to its role in asymptotic estimation, the L_1 distance has several nice statistical interpretations. First, it equals $\sup_A |P_1(A) - P_2(A)|$, where P_1 and P_2 are the probability measures associated with f_1 and f_2, and A is a arbitrary Borel set. Secondly, if we were to test f_1 vs f_2 using the Neyman-Pearson lemma, the L_1 distance equals to the maximal value of the "power minus size" among such tests. (See Lindsay and Markatou (2003))

Figures 3.6 and 3.7 display the square root of the scaled quadratic distance, for different values of h, along with the square root of scaled L_2 distance and the L_1 distance, when F and G are univariate normals with same variance. The x-axis is the difference in means of the two normals and the y-axis gives the distance. The distances in Figure 3.6 are plotted against means differing at most by 10 units, whereas in Figure 3.6 the plot is shown for a smaller range of x. As expected L_1, the L_2 and the quadratic distances with reasonable h's reaches the asymptotic value of 1 very fast. However, the quadratic distance with h=10 is very slow in approaching the asymptote of 1. In essence, the smoothing has been so extreme as to loose all sensitivity. If we plotted using larger values of the difference in means, then we would have found that all the distances approach the value 1. But, initially, especially for large values of h, it is very difficult to detect the asymptote as the curves have too little growth. Thus, we can see how the choice of h makes the distance more or less sensitive, and how the scaled quadratic distance compares with the scaled L_2 distance and L_1 distance. We also notice that for $h \leq 1$ the square root of the scaled quadratic distances are larger than or equal to the L_1 distance for the whole range of $(\mu_2 - \mu_1)$, whereas the distances are smaller than or equal to the L_1 distance for $h \geq 1$.

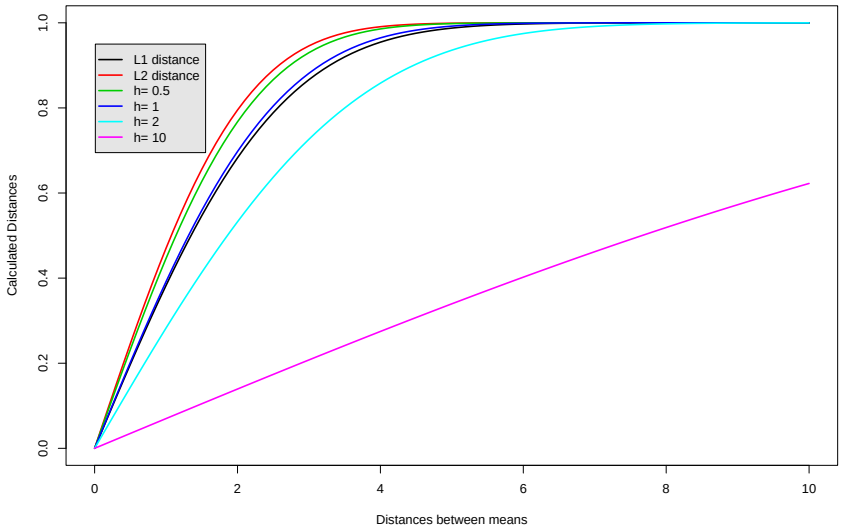

Figure 3.6: **Comparison of quadratic distance with various smoothing parameters, L_2 and L_1 distance. The distance is a function of the difference of means**

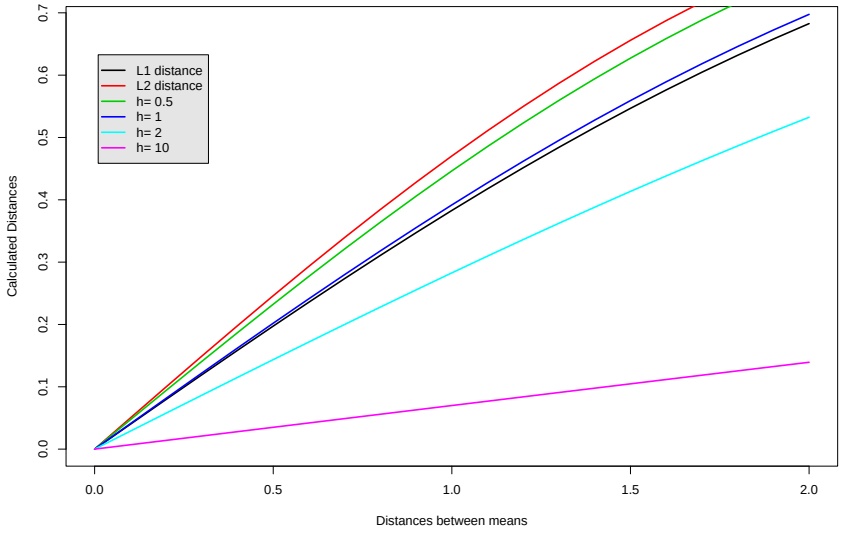

Figure 3.7: **Comparison of quadratic distance with various smoothing parameters, L_2 and L_1 distance. The distance is a function of the difference of means less than 2 standard deviations**

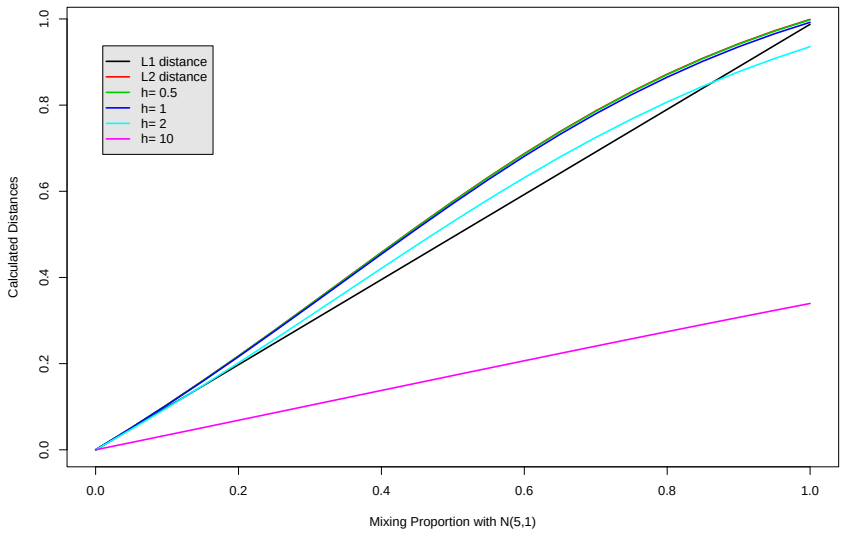

Figure 3.8: **Comparison of quadratic distance with various smoothing parameters, L_2 and L_1 distance. The distance is a function of the mixing proportion ε of the mixing distribution $(1-\varepsilon)\mathcal{N}(0,1)+\varepsilon\ \mathcal{N}(5,1)$**

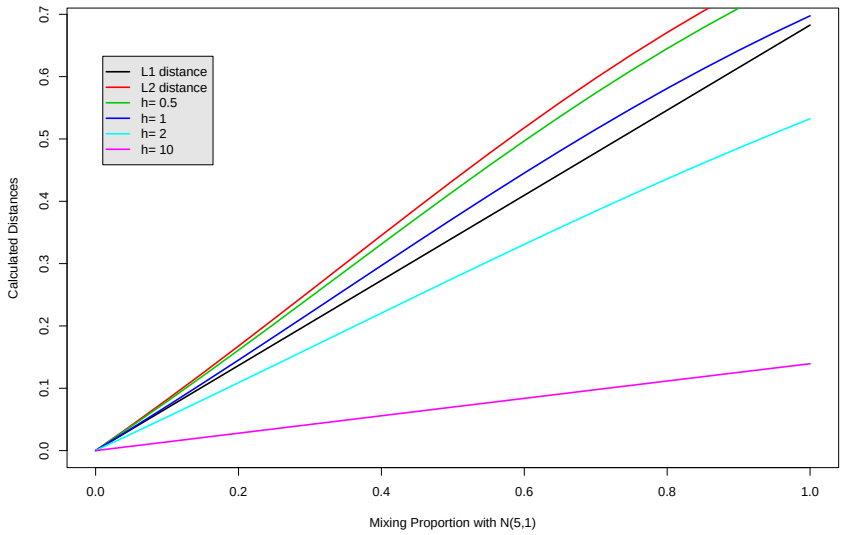

Figure 3.9: **Comparison of quadratic distance with various smoothing parameters, L_2 and L_1 distance. The distance is a function of the mixing proportion ε of the mixing distribution $(1-\varepsilon)\mathcal{N}(0,1)+\varepsilon\ \mathcal{N}(2,1)$**

Next, we compare the various distances in a different way. Let f_1 remain the same, that is $f_1 \sim \mathcal{N}(\mu, 1)$, but now let f_2 be a mixture of normals with means μ and $\mu+5$ and mixing proportion ε. Thus f_2 is given by

$$f_2 \sim (1-\varepsilon)\mathcal{N}(\mu,\sigma) + \varepsilon \ \ \mathcal{N}(\mu+5,\sigma).$$

For these comparison we chose $\mu = 0$ and $\sigma = 1$.We calculate the scaled L_2 distance and quadratic distance with different h's using the formula in equation (3.59). The L_1 distance reduces to $\varepsilon\left(1-2\Phi\left(-\frac{5}{2}\right)\right)$. Figure 3.8 gives the various distances as a function of the mixing proportion ε. Again, we can see how, with large values of h, one fails to detect the difference between f_2 and f_1, but smaller values are more sensitive. At $\varepsilon = 1$ we have $\mathcal{N}(0,1)$ *vs* $\mathcal{N}(5,1)$ and all reasonable h give 1 to eyeball accuracy. To observe the behavior of the distances when the mixing distributions are not far apart we next take

$$f_2 \sim (1-\varepsilon)\mathcal{N}(\mu,1) + \varepsilon \ \ \mathcal{N}(\mu+2,1).$$

Figure 3.9 gives the various distances as a function of the mixing proportion ε for the mixing distributions, with means two standard deviation apart. Again, we see that for $h \leq 1$ the square root of the scaled quadratic distances are larger than or equal to the L_1 distance for $\varepsilon \in [0,1]$, whereas the distances are smaller than or equal to the L_1 distance for $h \geq 1$.

3.9 Resampling based nonparametric acceptance region

In this section we will take a resampling-based approach to derive an approximate null distribution for $D_K(\hat{F},\tau)$ and will suggest an acceptance rule for selecting a model. In Section 3.6 we showed how we find the asymptotic distribution of generalized quadratic distance. But for a normal kernel the spectral decomposition is hard to find and the asymptotics may be suspect. So, instead we use a nonparametric estimator of the null distribution. As τ is unknown, we use bootstrap methodology and mimic the distribution of $D_K(\hat{F},\tau)$ with the dis-

tribution of $D_K(\hat{F}^*, \hat{F})$ where $\hat{F}^*$ has the bootstrap distribution corresponding to resampling from $\hat{F}$, i.e. bootstrap samples from the data.

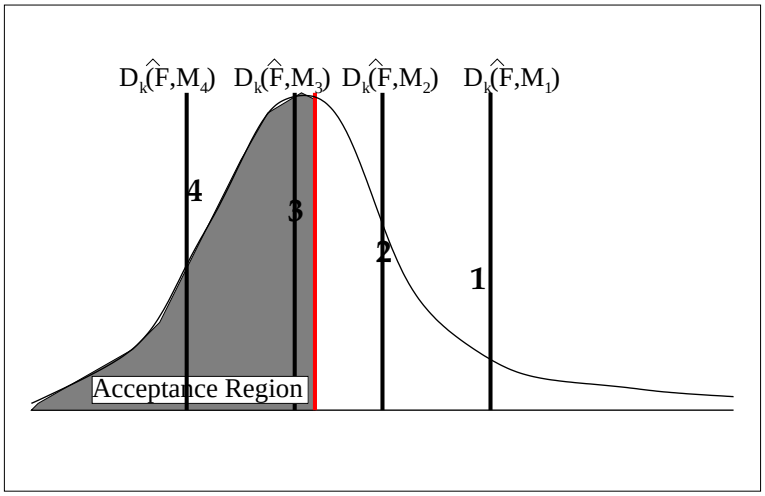

Figure 3.10: Diagram showing the histogram of $D_K(\hat{F}^*, \hat{F})$ along with the relative position of the quadratic distance of several models with different number of components and the acceptance region of the model selection rule

Next we need to decide upon an acceptance rule. Donoho (1988) showed that for the testing for the number of components in a mixture model, also known as mixture complexity, one can only construct a lower confidence limit, but not an upper confidence limit for the number of components. Intuitively, if we are given any g-component mixture we can find a $(g+100)$ component mixture that is arbitrarily close in distribution. So, we should reasonably expect only to make one sided tests, and conclude only that g is larger than some lower limit. However we also wish to be parsimonious, so if we are selecting a number of components, we select the smallest g that is acceptable. Thus we propose the following acceptance rule.

Let $\hat{M}_1, \hat{M}_2, \ldots, \hat{M}_g$ be the fitted normal mixtures with $1, 2 \ldots, g$ components respectively. The number of components selected will be the smallest g such that

$$D(\hat{F}, \hat{M}_g) \leq c_{.5}, \qquad \text{where } P(D(\hat{F}^*, \hat{F}) \leq c_{.5}) = .5 \tag{3.62}$$

Based on Donoho's results, we expect our estimator of g to be consistent.

Conjecture: $\lim_{n\to\infty} P(\hat{g}_n \leq g_0) = .5$,

so that $\hat{g}$ provides a valid 50% lower limit for g_0.

3.9.1 Estimation of distance under the normal kernel

In this section we will formulate estimation of the distance $D_K(\tau, M_g^*)$, M_g^* being the asymptotically best-fitting mixture normal density in g components, when using maximum likelihood. With the normal kernel given by (3.4), the quadratic distance $D_K(\hat{F}, M_g)$ can be written explicitly in an expression involving the data-points and the mixed normal density M_g.

Using the convolution properties of normal we can show that the centered kernel with respect to the normal density M where $dM = \sum_i^g \phi(x; \mu_i, V_i)$ is given by

$$\begin{aligned} \tilde{K}^M(x,y) &= K(x,y) - K(x,M) - K(M,y) + K(M,M) \\ &= K(x,y) - \sum_{i=1}^{g} \pi_i K_{\Sigma_h + V_i}(x, \mu_i) - \sum_{i=1}^{g} \pi_i K_{\Sigma_h + V_i}(\mu_i, y) + \sum_{i=1}^{g}\sum_{j=1}^{g} \pi_i K_{W_{ij}}(\mu_i, \mu_j) \\ &\quad , \end{aligned} \tag{3.63}$$

where $W_{ij} = \Sigma_h + V_i + V_j$

Thus using equation (3.11), the unbiased estimator of distance is

$$\begin{aligned} U_n(M_g) &= \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} \tilde{K}(x_i, x_j) \\ &= \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{j<i} \tilde{K}(x_i, x_j), \end{aligned} \tag{3.64}$$

where $\tilde{K}$ is given by (3.63). We can now recall one important reason for using the normal kernel: it leads to explicit integrals for $K(x,M)$ and $K(M,M)$ and so we have avoided numerical integration in calculating our distance.

Now if M_g is estimated by $\hat{M}_g$ using maximum likelihood, the resulting estimator $U_n(\hat{M}_g)$ should be consistent for $D(\tau, M_g^*)$; however, it is not necessarily unbiased for this target parameter. However, we found that using the *U-statistic* estimator gives much more satisfactory results than using $D(\hat{F}, \hat{M}_g)$.

3.10 Choice of tuning parameter

When calculating the distances for a specific model one is faced with specifying the "tuning parameter" h. As it turns out the best "tuning parameter" depends on the number of observations, the dimensionality of the data, and the hypotheses of interest. This tuning parameter selection will determine the tradeoff between our sensitivity to hypotheses of interest and the "noise" (variability) that arises from being sensitive towards too rich a class of alternatives. More insight on the choice of h is provided in Chapters 4 and 5 where we discuss the concept of "pseudo degrees of freedom", and the concordance and discordance coefficients.

3.11 Using "detectable distance" instead of the raw distance

In this section we introduce the idea of "detectable distance" to make the distances on a range of h comparable.

The idea here is that when the truth is not the model, the estimated distances are asymptotically normal; i.e. if $D(\tau, G) \neq 0$, then

$$\sqrt{n}\left(D(\hat{F}, G) - D(\tau, G)\right) \rightarrow \mathcal{N}(0, \sigma_h^2(\tau)) \tag{3.65}$$

$$\text{or} \quad \sqrt{n}\frac{\left(D(\hat{F}, G) - D(\tau, G)\right)}{\sigma_h(\tau)} \rightarrow \mathcal{N}(0, 1). \tag{3.66}$$

Let us call $\dfrac{D(\tau, G)}{\sigma_h(\tau)}$ the *"detectable distance"*. It is clear that dividing the raw distance by the true standard deviation puts them all in a standard scale which we might call the *"signal-to-noise-ratio"*. The "detectable distance", $\dfrac{D(\tau, G)}{\sigma_h(\tau)}$ can be estimated by $\dfrac{D(\hat{F}, G)}{\sigma_h(\hat{F})}$ or $\dfrac{U_n(G)}{\sigma_h(G)}$. In the following section we will detail the calculation of the variance $\sigma_h^2(\tau)$.

When the truth equals the model, one cannot appeal to asymptotic normality anymore, but we might hope that the standardized "null" distributions are now more similar, and so the modified bootstrap distribution is estimating a more stable distribution now. In fact, one would now bootstrap the ratios $\dfrac{D(\hat{F}^*, \hat{F})}{\sigma(\hat{F}^*)}$, where $\sigma(\hat{F}^*)$ can be computed explicitly.

3.12 Variance calculation

In this section we will derive the exact expression of the variance of the unbiased estimator $U_n(G)$, given in equation (3.11), for any arbitrary distribution G. For notational ease and compactness let us introduce the following,

$$\begin{aligned} K_{ij} &= K(x_i, x_j) \\ K_{i\tau} &= \int_y K(x_i, y) d\tau(y) \\ K_{\tau G} &= \int_x \int_y K(x, y) d\tau(x) dG(y) \\ \Delta_i &= (K_{i\tau} - K_{iG}) - (K_{\tau\tau} - K_{\tau G}). \end{aligned}$$

Proposition 3.3. *The exact variance of $U_n(G)$ is given by*

$$Var(U_n(G)) = \frac{2}{n(n-1)} E_\tau \left[\tilde{K}^\tau(x_i, x_j)\right]^2 + \frac{4}{n} E_\tau[\Delta_i^2]. \tag{3.67}$$

Proof :

$$\begin{aligned} & (n(n-1))^2 Var_\tau U_n(G) \\ = \; & (n(n-1))^2 E_\tau [U_n(G) - E_\tau[U_n(G]]^2 \\ = \; & E_\tau \left[\sum_{i=1}^{n} \sum_{j \neq i}^{n} (K_{ij} - K_{iG} - K_{Gj} + K_{GG}) - (K_{\tau\tau} - K_{\tau G} - K_{G\tau} + K_{GG}) \right]^2. \end{aligned}$$

We next re-center K about τ to obtain terms that will later be orthogonal. Thus

$$\begin{aligned}
& (n(n-1))^2 Var_\tau U_n(G) \\
= \; & E_\tau \left[\sum_{i=1}^{n} \sum_{j \neq i}^{n} (\tilde{K}_{ij}^{\tau} - K_{iG} - K_{Gj} - 2K_{\tau\tau} + K_{\tau G} + K_{G\tau} + K_{i\tau} + K_{\tau j} \right]^2 \\
= \; & E_\tau \left[\sum_{i=1}^{n} \sum_{j \neq i}^{n} \tilde{K}_{ij}^{\tau} + 2(n-1) \sum_{i=1}^{n} (K_{i\tau} - K_{iG} - K_{\tau\tau} + K_{\tau G}) \right]^2 \\
= \; & E_\tau \left[\sum_{i=1}^{n} \sum_{j \neq i}^{n} \tilde{K}_{ij}^{\tau} + 2(n-1) \sum_{i=1}^{n} \Delta_i \right]^2 \\
& \quad \text{where } \Delta_i = (K_{i\tau} - K_{iG}) - (K_{\tau\tau} - K_{\tau G}) \\
= \; & E_\tau \left[\sum_{i=1}^{n} \sum_{j \neq i}^{n} \tilde{K}_{ij}^{\tau} \right]^2 + 4(n-1)^2 E_\tau \left[\sum_{i=1}^{n} \Delta_i^2 \right] \\
& \quad \text{because} E\left[(\tilde{K}_{ij}^{\tau}) \Delta_{i'} \right] = 0 \quad \forall \quad i \neq j, i'. \\
= \; & E_\tau \left[\sum_{i=1}^{n} \sum_{j \neq i}^{n} \sum_{k=1}^{n} \sum_{k \neq l}^{n} \tilde{K}_{ij}^{\tau} \tilde{K}_{kl}^{\tau} \right] + 4(n-1)^2 E_\tau \left[\sum_{i=1}^{n} \Delta_i^2 \right] \\
= \; & 2n(n-1) E_\tau \left[\tilde{K}_{ij}^{\tau} \right]^2 + 4n(n-1)^2 E_\tau [\Delta_i^2]. \\
& \quad \text{since } E[\tilde{K}_{ij}^{\tau} \tilde{K}_{kl}^{\tau}] = 0 \text{ unless } i = k \text{ and } j = l \text{ or vice versa.} \\
= \; & 2n(n-1) E_\tau \left[\tilde{K}^{\tau}(x_i, x_j) \right]^2 + 4(n-1)^2 n E_\tau [\Delta_i^2].
\end{aligned}$$

Thus the variance of $U_n(G)$ is given by

$$Var(U_n(G)) = \frac{2}{n(n-1)} E_\tau \left[\tilde{K}^{\tau}(x_i, x_j) \right]^2 + \frac{4}{n} E_\tau [\Delta_i^2].$$

□

Remark: Notice that the two terms in the variance expression have orders of magnitude $\frac{1}{n^2}$ and $\frac{1}{n}$ respectively, so that asymptotically the second term dominates for any fixed pair τ and G with $\tau \neq G$. However as τ and G become closer, the magnitude of $E(\Delta_i^2)$ decreases. In this dissertation, where we seek models close to the truth τ, both terms will be relevant. In fact when $G = \tau$, i.e, when the model is correct, the second term in the variance expression equals

to zero as $\Delta_i^2 = 0, \ \forall \ i$. Thus

$$Var(U_n|G) = O\left(\tfrac{1}{n^2}\right) \implies Var(n\hat{d}|G) = O(1),$$
$$Var(U_n|\tau) = O\left(\tfrac{1}{n}\right) \implies Var(\sqrt{n}\hat{d}|G) = O(1).$$

3.12.1 Estimation of variance

An unbiased estimator of $E_\tau\left(\tilde{K}_\tau(x_i,x_j)\right)^2$ is $\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\left[\tilde{K}^2_{\hat{F}}(x_i,x_j)\right]$ and a consistent estimator of $E_\tau(\Delta_i^2)$ is $\frac{1}{n}\sum_{i=1}^{n}\hat{\Delta}_{\hat{F}}(x_i)$, where $\hat{\Delta}_{\hat{F}}(x_i) = (K_{i\hat{F}} - K_{iG} - K_{\hat{F}\hat{F}} + K_{\hat{F}G})$. Thus an estimator of the variance of $U_n(G)$ is given by

$$\widehat{Var(U_n(G))} = \frac{1}{(n(n-1)^2)}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\left[\tilde{K}^2_{\hat{F}}(x_i,x_j)\right] + \frac{4}{n^2}(K_{i\hat{F}} - K_{iG} - K_{\hat{F}\hat{F}} + K_{\hat{F}G}). \qquad (3.68)$$

3.13 Summary

In short an algorithm for choosing the number of components of a Multivariate Normal Mixture model using the nonparametric confidence interval is given in Figure 3.11.

3.14 Results

In this section we present some results based on the model selection criterion discussed in this chapter. Description of four datasets discussed below can be found in Appendix I.

Now we give a general description of the figures described in this section. The figures consists of the relative position of the distance of a g component model fitted to the respective data, overlaid on a histogram obtained from 1000 bootstrap samples of the distance. In each plot the median ($C_{.5}$) of the histogram is a shown as a vertical line in red. That means the acceptance region is to the left of the red line. Also the distance of a particular model, is

- **Choose h:** Choose several values of the tuning parameter h based on a pseudo degrees of freedom analysis.
- **Histogram:** Estimate the scaled distribution of the $D(\hat{F},\tau)$ by the histogram of the distances obtained from the bootstrap samples, i.e. the histogram of $\frac{D(\hat{F}^*,\hat{F})}{\widehat{\sigma(\hat{F}^*)}}$ for different bootstrap samples $\hat{F}^*$.
- **Model Fitting:** Estimate the parameters of the Multivariate Mixture Normals $\hat{M}_1,\ldots,\hat{M}_g$, where $\hat{M}_g$ is the fitted mixture with g components.
- **Model Distance:** Obtain the estimated distances $\frac{D(\hat{F},\hat{M}_1)}{\sigma_h(\hat{F})}, \frac{D(\hat{F},\hat{M}_2)}{\sigma_h(\hat{F})},\ldots,\frac{D(\hat{F},\hat{M}_l)}{\sigma_h(\hat{F})}$.
- **Model Selection:** g_0, our choice of the number of components is the first g that falls within the acceptance region described in Section 3.9.

Figure 3.11: **Algorithm for Model Selection based on the nonparametric confidence interval**

denoted by a vertical black line, with the number of components imprinted on it (e.g. see Figure 3.12).

For all the four datasets analyzed in this section, the choice of the tuning parameter was done on the basis of the theory in Chapter 4 (in particular, see Section 4.6). For the Iris data we used the tuning parameter $h = 0.5$. On the basis of our decision rule, we infer from Figure 3.12 that the data has 5 mixture components, as $g = 5$ is the first model whose distance is less than the median of the bootstrap distribution. Note that, though the Iris data was collected from 3 different species namely Setosa, Virginica and Versicolor, many analyses including the aural approach of Wilson (1982) show that there could be subspecies within the two species making a total of 5 subclusters. But, if we select $h = 0.8$ (Figure 3.13) our conclusion would be 3 components as we have smoothed the data more compared to choosing $h = 0.5$, and thus we ignore the the finer subclusters. From Figure 3.12 and Figure 3.13 we can appreciate the fact that the generalized quadratic distance can analyze a data at different scales, revealing clusters, superclusters and subclusters. In addition we can see that the two

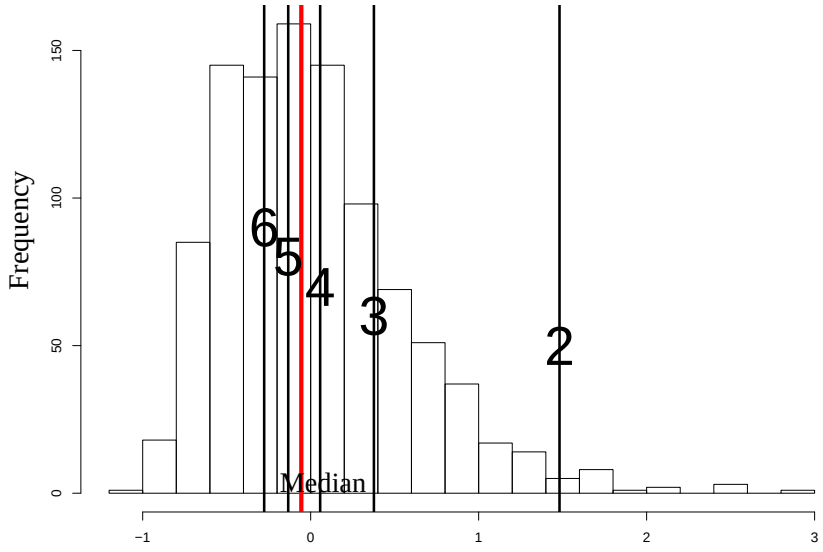

Figure 3.12: **Confidence Set decision for the Iris data with h=.5**

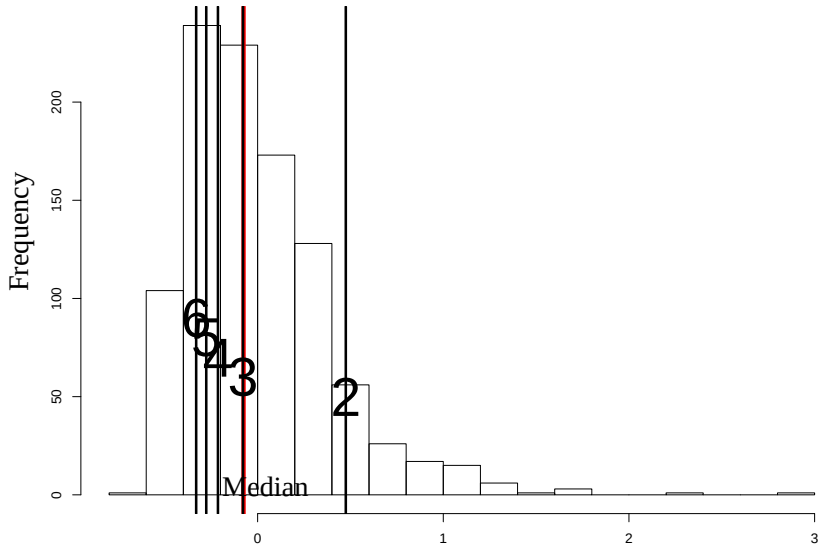

Figure 3.13: **Confidence Set decision for the Iris data with h=.8**

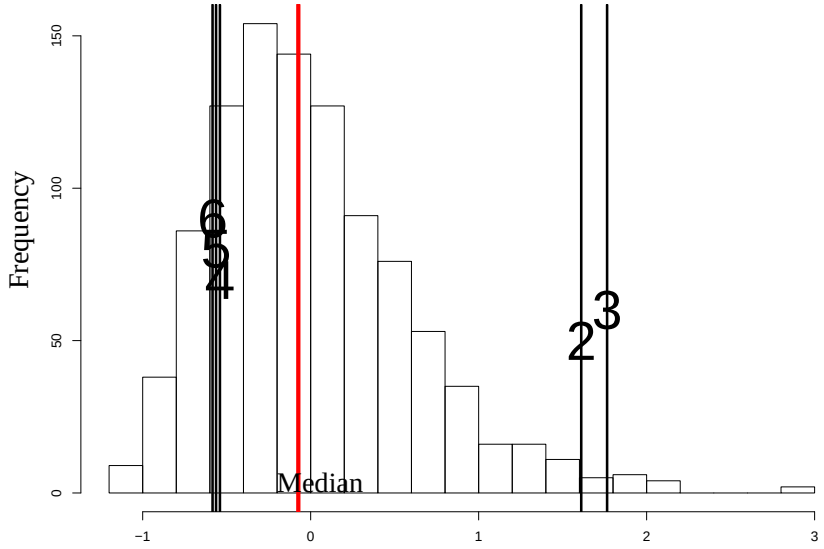

Figure 3.14: **Confidence Set decision for the Simulated Data 1 with** h**=.5**

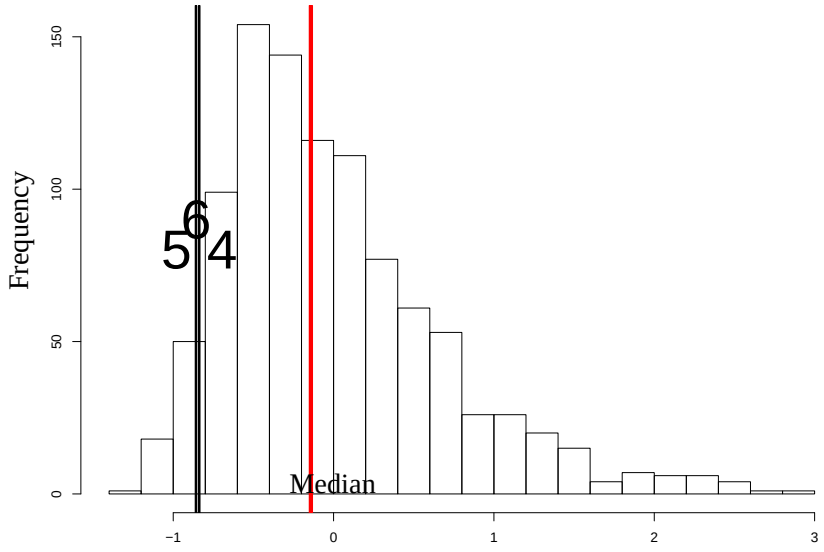

Figure 3.15: **Confidence Set decision for the Simulated Data 2 with** h**=.5**

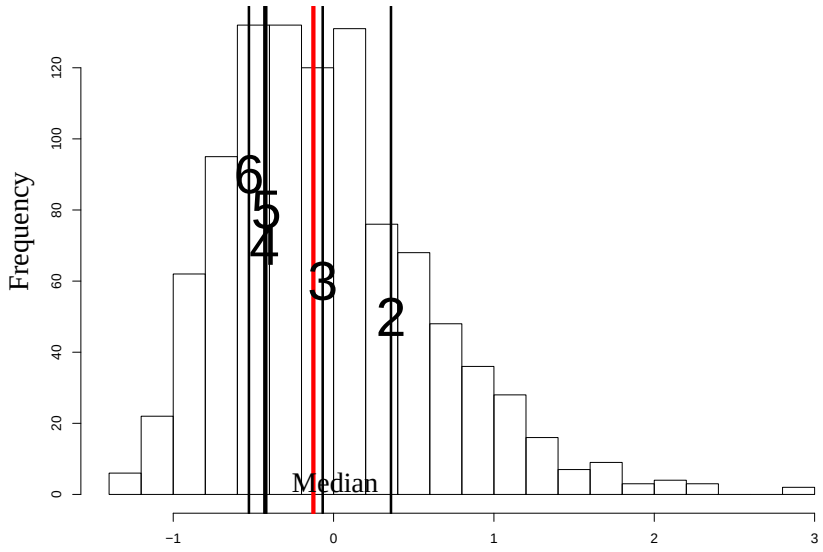

Figure 3.16: **Confidence Set decision for the Acidity data with h=.2**

component model shows definite inadequacy compared to three or more.

Figures 3.14 and 3.15 represent the analysis of the two simulated datasets 1 and 2 (see Appendix). Observing Figures 3.14 and 3.15, we would select the right number of components, i.e. 4, and see that the evidence is pretty strong. Finally in Figure 3.16, which is the univariate acidity dataset, we used $h = 0.05$, and would choose 4 components. In this case there is no known answer. It also appears that three components provide a good fit.

3.15 Conclusion

Generalized quadratic distances are a tool for measuring the distance between two arbitrary distributions. They can be used to create a global model selection tool. One advantage of using these distances is that they can be used when one distribution is continuous and the other discrete.

Since the empirical distances can be re-written as $\mathcal{U}$-$functionals$, we can use all the

standard results on $\mathcal{U}$-*statistics*. We can also develop a simple asymptotic theory. Moreover, using the product closure properties of the kernel, we avoid multidimensional numerical integration in calculating the distance. This speed in calculation enables us to construct a nonparametric confidence interval for detectable distance. In Chapters 5 and 6 we will use the quadratic distance to construct other model selection tools.

Chapter 4

Pseudo Degrees of Freedom

The generalized quadratic distances described earlier in Chapter 3 depend on the selection of a kernel K which, in turn, depends on a tuning parameter, "h". Furthermore, since we do not have a homogeneous distance measure (i.e., one based on transformation to a uniform scale), signal-to-noise analyses will depend on the kernel, the true state of nature τ, and the directions of its deviation M (proposed model) being considered. In order to better understand and control the behavior of the distance we now develop data-based tools for use in selecting the tuning parameter h. An analogy can be drawn between the tuning parameter "h" and the bin-width of a cell in the χ^2 goodness-of-fit tests. The choice of the parameter is a very important in designing a powerful distance between two distributions. In this chapter we propose a simple summary statistic, the "Pseudo Degrees of Freedom" ($pDOF$), to help us decide on the range of the tuning parameter.

The idea of the "degrees of freedom" as‘ a summary statistic for the level of smoothing has been used in other scientific literature. For choosing the tuning parameter Kou and Efron (2002) appealed to the "ideal degrees of freedom", which was developed based on a similar theory. Gu (1998) and Hall and Johnstone (1992) also refer to degrees of freedom for choosing tuning parameters. The idea of "stochastic degrees of freedom", which is most similar to the quantity used here, was used by Yoo and Stark (2003) in the context of covariance kernel for Gaussian processes.

4.1 Motivation

As mentioned before, the problem of choosing the tuning parameter "h" is analogous to choosing the bin-width of each cell (or, equivalently choosing the number of cells) in a χ^2 goodness-of-fit test. If the number of cells is small, then the test may be unable to detect important discrepancies between two distributions because too much has been "smoothed out". On the other hand, if the number of cells is too large the test statistic is much more variable, and so power of the test is lost against many alternatives. (See Kallenberg et al., 1985).

For a χ^2 goodness-of-fit test, the degrees of freedom is equal to the (# of cells)-1. Rough rules of thumb for selecting the number of cells is that the degrees of freedom should be more than 5 and less than $n/5$, n being the total number of observations. So to choose an interesting range of the tuning parameter "h" of the quadratic kernels, we will determine a natural extension of the "degrees of freedom", and then use a rule of thumb for selecting the appropriate value.

It should be noted that here we want to define the degrees of freedom in a multivariate situation. Thus, instead of implicitly defining the length of each interval (bin-width), we are in effect defining higher dimensional bins (hypercubes or hyperballs). An advantage of our method is that we only have to define the tuning parameter "h" in the quadratic kernel K, not the location and size of bins. That is, our method requires no selection of the number of bins, but rather the effective bin-width (h).

4.2 Definition and properties of *pDOF*

In this section a formal definition of the *pDOF* will be provided. We will also discuss some important properties of the *pDOF*.

From the spectral decomposition theory described in Chapter 3 we know that the null

distribution of the generalized quadratic distance $D_K(F,G)$ is such that,

$$n\, D_K(\hat{F},\tau) \xrightarrow{\mathbf{d}} \sum_{i=1}^{\infty} \lambda_i \chi^2_{i(1)}.$$

One could approximate the null distribution of the distance D, which is a infinite sum of weighted $\chi^2_{(1)}$, by a scaled χ^2 distribution by matching the first two moment (Satterthwaite, 1941). Let the approximate scaled χ^2 distribution be given by $c\chi^2_{pDOF}$, where the degrees of freedom is $pDOF$. Matching the first two moments we have

$$E[c\chi^2_{pDOF}] = E[D] = \sum_{i=1}^{\infty} \lambda_i \quad \Longrightarrow \quad c\ \ pDOF = \sum_{i=1}^{\infty} \lambda_i, \tag{4.1}$$

$$V[c\chi^2_{pDOF}] = V[D] = 2\sum_{i=1}^{\infty} \lambda_i^2 \quad \Longrightarrow \quad c^2\ \ pDOF = \sum_{i=1}^{\infty} \lambda_i^2. \tag{4.2}$$

Solving equations 4.1 and 4.2 we get,

$$c = \frac{\sum_{i=1}^{\infty} \lambda_i^2}{\sum_{i=1}^{\infty} \lambda_i}, \qquad pDOF = \frac{\left(\sum_{i=1}^{\infty} \lambda_i\right)^2}{\sum_{i=1}^{\infty} \lambda_i^2}. \tag{4.3}$$

Also from the spectral decomposition of the kernel K with respect to τ, we have the following

$$\int_x \tilde{K}^{\tau}(x,x)d\tau(x) = \sum_{i=1}^{\infty} \lambda_i \text{ and } \int_x \int_y \left(\tilde{K}^{\tau}(x,y)\right)^2 d\tau(x)d\tau(y) = \sum_{i=1}^{\infty} \lambda_i^2. \tag{4.4}$$

Thus we define the summary statistic $pDOF$ by,

$$pDOF = \frac{\left(\int \tilde{K}^{\tau}_h(x,x)\ d\tau(x)\right)^2}{\int\int \left(\tilde{K}^{\tau}(x,y)\right)^2\ d\tau(x)d\tau(y)}, \tag{4.5}$$

where $\tilde{K}_h(x,y)$ is the centered kernel defined in Chapter 3, which is

$$\tilde{K}^{\tau}_h(x,y) = K_h(x,y) - K_h(\tau,y) - K_h(x,\tau) + K_h(\tau,\tau), \tag{4.6}$$

$$\text{where } K_h(x,y) = \frac{1}{(2\pi)^{\frac{p}{2}}\ |\Sigma_h|^{\frac{1}{2}}} \exp\left(-\tfrac{1}{2}(x-y)'\Sigma_h^{-1}(x-y)\right)$$

and $\Sigma_h = h^2 I$, h being the "tuning parameter". In this Chapter, from now onwards we will suppress the symbol τ in $\tilde{K}^{\tau}$. So unless otherwise specified $\tilde{K}$ will mean $\tilde{K}^{\tau}$.

Note that if the asymptotic distribution of the distance D is exactly chi-squared, which would happen when q eigenvalues equal to 1 and the rest are equal to zero, then the *pDOF* is exactly the degrees of freedom of the χ^2 distance $D(\hat{F},\tau)$. Otherwise, the *pDOF* equals the approximate degrees of freedom used by Satterthwaite (1946) to create approximate test regions.

Notice also that *pDOF* is an invariant measure under scale multiplication of D. It is a simple and invariant natural summary of the overall concentration of the eigenvalues. If we let $\pi_j = (\lambda_j / \sum_k \lambda_k)$, then it can be written as

$$pDOF = \left(\sum \pi_j^2\right)^{-1}.$$

Expressed in this form we can see that it approaches infinity as the eigenvalues become more equal. In the models we study here, the *pDOF* will go to infinity as the bandwidth h shrinks, corresponding intuitively to having a chi-squared statistic with more and more data cells. We shall return to this intuition later in the Chapter.

4.3 Theoretical Calculation of the *pDOF*

In this section we calculate *pDOF* for our normal kernel when the true distribution τ is a mixture of normals.

Now for our normal kernel $K_h(x,y)$, with true distribution τ being a mixture of normals given by $\tau = \sum_{l=1}^{g} \pi_l \mathcal{N}(\mu_l, V_l)$, we have the following calculations. The numerator of

equation (4.5) can be simplified in the following way.

$$\begin{aligned}
\int \tilde{K}_h(x,x)\ d\tau(x) &= \int K_h(x,x)\ d\tau(x) - \int K_h(\tau,x)\ d\tau(x) - \int K_h(x,\tau)\ d\tau(x) + \int K_h(\tau,\tau)\ d\tau(x) \quad (4.7)\\
&= \int K_h(x,x)\ d\tau(x) - K_h(\tau,\tau)\\
&\quad \text{as } \int K_h(\tau,x)\ d\tau(x) = \int K_h(x,\tau)\ d\tau(x) = \int_x \left(\int_y K_h(x,y) d\tau(y)\right) d\tau(x) = K_h(\tau,\tau)\\
&= \int \frac{1}{(\sqrt{2\pi}h)^p}\ d\tau(x) - K_h(\tau,\tau) \quad (4.8)\\
&= \frac{1}{(\sqrt{2\pi}h)^p} - \sum_{l=1}^{g}\sum_{k=1}^{g} \pi_l \pi_k c_{lk},
\end{aligned}$$

where

$$c_{lk} = \frac{1}{(2\pi)^{\frac{p}{2}} |V_l + V_k + \Sigma_h|^{\frac{1}{2}}} \exp\left(-\frac{(\mu_i - \mu_j)'(V_l + V_k + \Sigma_h)^{-1}(\mu_i - \mu_j)}{2}\right) = K_{W_{lk}}(\mu_l, \mu_k)$$

and $W_{lk} = \Sigma_h + V_l + V_k$.

Next, applying the convolution properties of normal and using the result from equation (3.63) we calculate the theoretical value of the denominator, $\int\int \tilde{K}_h^2(x,y)\ d\tau(x)d\tau(y)$. through the following steps. We have, first,

$$\tilde{K}_h^2(x,y) = [K_h(x,y) - K_h(\tau,y) - K_h(x,\tau) + K_h(\tau,\tau)]^2. \quad (4.9)$$

Integrating the above we get,

$$\begin{aligned}
&\int_x\int_y \tilde{K}_h^2(x,y)\ d\tau(x)d\tau(y)\\
= \quad &\int_x\int_y K_h^2(x,y)\ d\tau(x)d\tau(y) + \int_x K_h^2(x,\tau)\ d\tau(x) + \int_y K_h^2(\tau,y)\ d\tau(y) + K_h^2(\tau,\tau)\\
&-2\int_x\int_y K_h(x,y)K_h(x,\tau)\ d\tau(x)d\tau(y) - 2\int_x\int_y K_h(x,y)K_h(\tau,y)\ d\tau(x)d\tau(y)\\
&+2K_h(\tau,\tau)\int_x\int_y K_h(x,y)\ d\tau(x)d\tau(y) + 2\int_x\int_y K_h(x,\tau)K_h(\tau,y)\ d\tau(x)d\tau(y)\\
&-2K_h(\tau,\tau)\int_x K_h(x,\tau)\ d\tau(x) - 2K_h(\tau,\tau)\int_y K_h(\tau,x)\ d\tau(y). \quad (4.10)
\end{aligned}$$

Also, the following identities are always true. The terms $\int_x K_h^2(x,\tau)\ d\tau(x)$, $\int_y K_h^2(\tau,y)\ d\tau(y)$, $\int_x\int_y K_h(x,y)K_h(x,\tau)\ d\tau(x)d\tau(y)$, and $\int_x\int_y K_h(x,y)K_h(\tau,y)\ d\tau(x)d\tau(y)$, are all equal to

$\int_x \int_y K_h(x,y)K_h(z,y)\, d\tau(x)d\tau(y)d\tau(z)$ and $K_h(\tau,\tau)\int_x \int_y K_h(x,y)\, d\tau(x)d\tau(y)$, $K_h(\tau,\tau)\int_x K_h(x,\tau)\, d\tau(x)$, $K_h(\tau,\tau)\int_y K_h(\tau,x)\, d\tau(y)$, $\int_x \int_y K_h(x,\tau)K_h(\tau,y)\, d\tau(x)d\tau(y)$, and $K_h^2(\tau,\tau)$ are all equal to $\left(\int_x \int_y K_h(x,y)\, d\tau(x)d\tau(y)\right)^2$.

Using the above set of identities, equation (4.10) reduces to

$$\int_x \int_y K_h^2(x,y)\, d\tau(x)d\tau(y) - 2\int_x \int_y K_h(x,y)K_h(z,y)\, d\tau(x)d\tau(y)d\tau(z) + \left(\int_x \int_y K_h(x,y)\, d\tau(x)d\tau(y)\right)^2. \tag{4.11}$$

We can calculate the first term of equation (4.11) by first showing

$$\begin{aligned} K_h^2(x,y) &= \left[\frac{1}{(\sqrt{2\pi}h)^p}\exp\left(-\frac{1}{2}\frac{(x-y)'(x-y)}{h^2}\right)\right]^2 \\ &= \frac{1}{(\sqrt{2\pi}h)^{2p}}\exp\left(-\frac{(x-y)'(x-y)}{h^2}\right) \\ &= \frac{1}{(2\sqrt{\pi}h)^p}\frac{1}{(\sqrt{2\pi}\frac{h}{\sqrt{2}})^p}\exp\left(-\frac{1}{2}\frac{(x-y)'(x-y)}{(\frac{h}{\sqrt{2}})^2}\right) \\ &= \frac{1}{(2\sqrt{\pi}h)^p}K_{\frac{h}{\sqrt{2}}}(x,y), \end{aligned} \tag{4.12}$$

and then integrating the above we have,

$$\int_x \int_y K_h^2(x,y)\, d\tau(x)d\tau(y) = \frac{1}{(2\sqrt{\pi}h)^2}\sum_{l=1}^{g}\sum_{k=1}^{g}\pi_l\pi_k c_{lk}^*, \tag{4.13}$$

where

$$c_{lk}^* = \frac{1}{(2\pi)^{\frac{p}{2}}|V_l+V_k+\Sigma/2|}\exp\left(-\frac{(\mu_l-\mu_k)'(V_l+V_k+\Sigma/2)^{-1}(\mu_l-\mu_k)}{2}\right) = K_{W_{lk}}(\mu_l,\mu_k)$$

and $W_{lk}^* = \frac{\Sigma_h}{2}+V_l+V_k$. The third term of equation (4.11) is the square of the second term in equation (4.8). The middle term $\int_x \int_y K_h(x,y)K_h(z,y)\, d\tau(x)d\tau(y)d\tau(z)$ is tedious to calculate. So, we next simplify to the case where τ is a single component normal. For a one component p-variate Normal with variance V, $pDOF$ reduces to

$$pDOF = \frac{\left(|\Sigma_h|^{-\frac{1}{2}} - |\Sigma_h+2V|^{-\frac{1}{2}}\right)^2}{|\Sigma_h|^{-\frac{1}{2}}|\Sigma_h+4V|^{-\frac{1}{2}} - 2|\Sigma_h+V|^{-\frac{1}{2}}|\Sigma_h+3V|^{-\frac{1}{2}} + |\Sigma_h+2V|^{-1}}. \tag{4.14}$$

For $V = I$, i.e. in case of a standard normal,

$$pDOF = \frac{\left(h^{-p} - (h^2+2)^{-\frac{p}{2}}\right)^2}{h^{-p}(h^2+4)^{-\frac{p}{2}} - 2(h^2+1)^{-\frac{p}{2}}(h^2+3)^{-\frac{p}{2}} + (h^2+2)^{-p}}. \tag{4.15}$$

From this calculation it can be seen that $pDOF \to \infty$ as $h \to 0$. Also, for small h equation (4.15) can be reduced to,

$$\begin{aligned} pDOF &= \frac{\frac{1}{h^{2p}}\left(1 - \left(\frac{h^2}{h^2+2}\right)^{\frac{p}{2}}\right)^2}{\frac{1}{h^{2p}(h^2+4)^{\frac{p}{2}}}\left(1 - 2\frac{h^{2p}(h^2+2)^{\frac{p}{2}}}{(h^2+1)^{\frac{p}{2}}(h^2+3)^{\frac{p}{2}}} + \frac{h^{2p}(h^2+2)^{\frac{p}{2}}}{(h^2+2)^p}\right)} \\ &\approx \left(\frac{h^2+4}{h^2}\right)^{\frac{p}{2}} \qquad \text{for small } h \\ &\approx \left(\frac{2}{h}\right)^p. \end{aligned} \tag{4.16}$$

Here, we might note that the *pDOF* increases exponentially in p, the dimension of the data, for fixed h, so that much larger values of h are needed in higher dimensions to obtain the same *pDOF*. In the standard normal for h^2 small, the approximation $pDOF \approx (2/h)^p$ can be a useful guide, as we show later. Note that this approximation has a natural interpretation. If we think of the standard normal density as being $4\sigma = 4$ standard deviations wide, and the normal kernel, standard deviation h as being a window of effective width $2h$, then $2/h$ equal to the "number of effective bins."

4.4 Estimating the Pseudo Degrees of freedom

For a given dataset, one can estimate *pDOF* empirically as follows. For a particular h, we estimate $\int K_h^\tau(x,x)\, d\tau(x)$ by $\frac{1}{n}\sum_i^n \tilde{K}_h^{\hat{F}}(x_i,x_i)$, where $\tilde{K}_h^{\hat{F}}(x_i,x_j)$ is the centered kernel, centered using the empirical distribution. The explicit form of the centered kernel is given by is given by,

$$\tilde{K}_h(x_i,x_j) = K_h(x_i,x_j) - \frac{1}{n}\sum_i K_h(x_i,x_j) - \frac{1}{n}\sum_i K_h(x_i,x_j) + \frac{1}{n^2}\sum_i\sum_j K_h(x_i,x_j)$$

Moreover, using the U-Statistics results, $\int\int \left(K_h^{\tau}(x,y)\right)^2 \, d\tau(x) d\tau(y)$ can be estimated by

$$\frac{2}{n(n-1)} \sum_{i}^{n} \sum_{j<i} \left(\tilde{K}_h^{\hat{F}}(x_i, x_j)\right)^2.$$

Thus, we will use the estimator

$$\widehat{pDOF} = \frac{\frac{1}{n}\sum_i^n \tilde{K}_h(x_i, x_i)}{\frac{2}{n(n-1)}\sum_i^n \sum_{j<i} \left(\tilde{K}_h^{\hat{F}}(x_i, x_j)\right)^2} \quad . \tag{4.17}$$

4.5 Preliminary ideas on selecting *pDOF*

In this section we will propose some preliminary rules for selecting *pDOF* and hence h. If for a given h the *pDOF* is smaller than 5, we will consider it likely to be too much smoothing, whereas a $pDOF > n/5$ would mean too little smoothing. Thus we should choose a reasonable range in between to look for signals at different bandwidths. Further research on the choice of degrees of freedom is needed.

4.6 Results

In addition to the four datasets that were discussed in the Chapter 3, we also use two other simulated datasets, both from a uni-component multivariate normal, one having uncorrelated variables and the other one having correlated variables. These calculations were done in order to compare the estimator $\widehat{pDOF}$ with the theoretical *pDOF* in section 4.3. Simulated dataset 3 is a sample of size 160 generated from a multivariate normal with

$$\mu = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} \quad \text{and} \quad V = I_4 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \tag{4.18}$$

and the Simulated Dataset 4 consists of 160 samples generated from multivariate normal with

$$\mu = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} \quad \text{and} \quad V = \begin{pmatrix} 1 & 0.44 & 0.44 & 0.44 \\ 0.44 & 1 & 0.44 & 0.44 \\ 0.44 & 0.44 & 1 & 0.44 \\ 0.44 & 0.44 & 0.44 & 1 \end{pmatrix}. \tag{4.19}$$

Table 4.1: **Pseudo degrees of freedom for Uni-component Multivariate Normal Uncorrelated Dataset**

p=1			
h	$\widehat{pDOF}$	*pDOF*	*pDOF**
0.1	20.12	19.89	20.0
0.2	10.09	9.98	10.0
0.5	4.16	4.21	4.0
1.0	2.30	2.43	2.0
2.0	1.46	1.56	1.0
4.0	1.13	1.17	0.5

p=2			
h	$\widehat{pDOF}$	*pDOF*	*pDOF**
0.1	361.3	403.66	400.00
0.2	98.7	103.65	100.00
0.5	19.4	19.56	16.00
1.0	7.21	7.27	4.00
2.0	3.60	3.68	1.00
4.0	2.44	2.48	0.25

p=3			
h	$\widehat{pDOF}$	*pDOF*	*pDOF**
0.1	6075.76	8040.79	8000.00
0.2	1070.02	1025.32	1000.00
0.5	82.42	78.44	64.00
1.0	17.23	17.12	8.00
2.0	6.56	6.55	1.00
4.0	3.93	3.93	0.12

p=4			
h	$\widehat{pDOF}$	*pDOF*	*pDOF**
0.1	24183.25	160833.16	160000
0.2	7609.33	10231.19	10000
0.5	324.47	310.73	256
1.0	37.64	37.45	16
2.0	10.48	10.44	1
4.0	5.57	5.53	6.00e-02

First, for simulated dataset 3 and simulated dataset 4, we calculate the theoretical *pDOF* and the estimated *pDOF* for different *h* and for different dimension of data. For the simulated dataset 3, which comes from a standard normal, the theoretical *pDOF* and $\widehat{pDOF}$ are also compared with the approximation given by equation (4.16). For the other four datasets

given in Appendix I, we only calculate $\widehat{pDOF}$. For the multidimensional datasets (Iris and Simulated 1 to 4) we tabulate the values for all dimensions to inspect the change in the *pDOF* when the number of dimension changes. For example for the IRIS dataset, p=1 means the first variables (Petal width), p=2 means the first two variables (Petal width, Petal length) and so forth. Also, to choose the appropriate h for each dataset and for each subsets of variable we calculated the *pDOF* for a range of h.

In Table 4.1, we first compare the estimated *pDOF*, denoted by $\widehat{pDOF}$ (Column 2) with the theoretical *pDOF* (Column 3), given by equation (4.15) for the uni-component multivariate uncorrelated dataset. For each p and h, *pDOF* and $\widehat{pDOF}$ are quite close. Thus, we can conclude that our estimator appears to work well in the uncorrelated case. Moreover, we should also observe that, for a fixed number of variables, as $h \uparrow$, $pDOF \downarrow$. Also, we should notice the change in the value of *pDOF* for changes in dimensionality of the data. As the dimension goes up the value of the *pDOF* goes up exponentially. So, if we want to keep the degrees of freedom fixed, we must choose much larger h when we include more variables.

Now, let us discuss how we select an interesting range of h based on Table 4.1. We will follow the same selection criterion for the other tables. Here, our dataset consists of 160 observations. So, by the threshold values proposed in Section 4.5 we would be interested in the range of values of *pDOF* somewhere between 5 and 32. Thus considering $p = 1$ we should choose $h = 0.2$ or $h = 0.1$, as the corresponding *pDOF* values are 9.98 and 19.89. If we do further computation we could refine 0.1 to an even smaller value of h. For example, for $h = .07$ we have $pDOF = 28.42937$. Going by the same rule, our choice of h for dimension 2, 3 and 4 would be somewhere around .5, 1 and 2 respectively.

Next, we compare the the actual *pDOF* with the approximate *pDOF*, denoted by $pDOF^*$, which is $(2/h)^p$. Columns 4 in Table 4.1 gives this approximation. As mentioned earlier this approximation works well only for small h. For $p = 1$, i.e. the approximation works very well for $h < .2$. It even works quite well for $h = 0.5$ and $h = 1$. Comparing the

Table 4.2: Pseudo degrees of freedom for Uni-component Multivariate Normal Correlated Dataset

p=1		
h	$\widehat{pDOF}$	pDOF
0.1	20.64	19.89
0.2	10.27	9.98
0.5	4.20	4.21
1.0	2.36	2.43
2.0	1.51	1.56
4.0	1.16	1.17

p=2		
h	$\widehat{pDOF}$	pDOF
0.1	382.82	362.18
0.2	103.30	93.34
0.5	18.81	17.85
1.0	6.69	6.64
2.0	3.28	3.30
4.0	2.19	2.15

p=3		
h	$\widehat{pDOF}$	pDOF
0.1	9597.11	6154.00
0.2	1151.09	789.80
0.5	74.81	62.82
1.0	15.12	14.21
2.0	5.53	5.35
4.0	3.22	3.04

p=4		
h	$\widehat{pDOF}$	pDOF
0.1	24905.90	102067.43
0.2	6229.29	6545.65
0.5	243.00	211.37
1.0	30.07	27.92
2.0	8.10	7.83
4.0	4.06	3.89

the other 3 sub-tables of Table 4.1 for $p = 2,3,4$ we still find that the $pDOF^*$ is a very good approximation for $pDOF$, for small values of h. However, it can be a severe underestimate at some reasonable degrees of freedom in higher dimension.

Since we have discussed the interpretation of Table 4.1 in detail we briefly note our conclusions from the $pDOF$ values of the other table. For Table 4.2 we have the estimated and the actual $pDOF$. According to the rule of thumb, for 160 observations from our grid of h values we would have chosen $h = .1$ for $p = 1$, $h = .5$ for $p = 2$, $h = 1$ for $p = 3$ $h = 1$ for $p = 4$. Also the estimates (Column 2) are quite close to the actual values (Column 3) for all p

and for reasonable h's.

Table 4.3: Pseudo degrees of freedom for Iris Dataset

p=1	
h	pDOF
0.1	19.80
0.2	9.80
0.5	3.97
1.0	2.22
2.0	1.44
4.0	1.13

p=2	
h	pDOF
0.1	202.10
0.2	75.18
0.5	15.50
1.0	6.01
2.0	3.30
4.0	2.37

p=3	
h	pDOF
0.1	723.93
0.2	161.49
0.5	21.53
1.0	6.61
2.0	3.18
4.0	2.23

p=4	
h	pDOF
0.1	2508.06
0.2	325.77
0.5	31.07
1.0	8.03
2.0	3.26
4.0	2.12

Table 4.4: Pseudo degrees of freedom for Generated Dataset 1

p=1	
h	pDOF
0.1	7.48
0.2	3.85
0.5	1.61
1.0	1.17
2.0	1.05
4.0	1.02

p=2	
h	pDOF
0.1	61.32
0.2	19.00
0.5	5.51
1.0	2.56
2.0	1.56
4.0	1.25

p=3	
h	pDOF
0.1	379.09
0.2	66.72
0.5	10.67
1.0	3.53
2.0	1.76
4.0	1.34

p=4	
h	pDOF
0.1	2552.86
0.2	326.50
0.5	27.68
1.0	7.59
2.0	3.06
4.0	1.90

For the Iris dataset we have 4 variables and 150 observations. So, among the values Table 4.3, we might choose $h = .5$ when $p = 4$, which gives $pDOF = 31.07$. Drawing a parallel to the chi-squared statistic, this implies that on an average each cell will have around five observations. Similarly for the simulated datasets 1 and 2 ($n = 160, p = 4$) we would go for $h = .5$ (see Tables 4.4 and 4.5). For the acidity dataset ($p = 1$) we evaluate the $pDOF$ for a greater range of h including some smaller values. Our best choice from the table would be $h = .05$, as in this dataset for $h = .5$ the $pDOF$ is too small (3.47).

Table 4.5: Pseudo degrees of freedom for Generated Dataset 2

p=1	
h	pDOF
0.1	4.90
0.2	3.54
0.5	2.93
1.0	1.92
2.0	1.27
4.0	1.07

p=2	
h	pDOF
0.1	26.26
0.2	11.41
0.5	4.90
1.0	2.83
2.0	1.82
4.0	1.56

p=3	
h	pDOF
0.1	114.01
0.2	28.02
0.5	6.69
1.0	3.90
2.0	3.00
4.0	2.58

p=4	
h	pDOF
0.1	766.21
0.2	98.69
0.5	12.01
1.0	4.88
2.0	3.34
4.0	2.81

Table 4.6: Pseudo degrees of freedom for Acidity Dataset

p=1	
h	$\widehat{pDOF}$
0.001	251.44
0.01	146.78
0.05	29.79
0.1	15.11
0.2	7.77
0.5	3.06
1.0	1.66
2.0	1.20
4.0	1.06

4.7 Conclusion

The estimated pseudo degrees of freedom, $\widehat{pDOF}$ provides an interesting and useful single number summary of the sensitivity characteristics of the distance. It can be calculated once and for all without using the model. Even though more research needs to be done on

the choice of *pDOF*, following our "rule of thumb" we can get a useful range of the tuning parameter. In general, we observed that when we include more variables we have to make h larger. Observing the difference in the values of the *pDOF* of Simulated dataset 1 and 2, we can conclude that the selection of h also depends on the variability structure and amount of separation between the components.

Again, we should note that, our main goal is not to approximate the asymptotic distribution of $nD_K(\hat{F}, \tau)$ by $c\,\chi^2_{pDOF}$. Equation (4.3) could also be used to find a solution for c, and then one could approximate $\sum_{i=1}^{\infty} \lambda_i \chi^2_{i(1)}$. However, in this dissertation the *pDOF* will only be used as a tool for selection of an interesting range of the "tuning parameter" h.

Chapter 5

Concordance and Discordance based Analysis

In this chapter we will introduce a concordance/discordance based analysis of the quadratic distance. The idea of concordance and discordance has been used by Lin (1989) to evaluate reproducibility in assay validation. In general, the concordance curves, which have R^2 like properties, determine the amount of variability in the empirical density explained by the models. The concordance coefficient will be examined as an informal model selection criterion. The sensitivity of the concordance coefficients for a particular "smoothing parameter", h can also be used to select an interesting range of h.

5.1 Definition of Concordance/Discordance

Definition 5.1. *As a density distance measure the* ***discordance*** *between two probability densities f and g is defined as*

$$\delta(f,g) = \frac{\int (f(x)-g(x))^2\,dx}{\int f^2(x)dx + \int g^2(x)dx} \tag{5.1}$$

and the corresponding ***concordance*** *is*

$$C(f,g) = 1-\delta(f,g) = \frac{2\int f(x)g(x)dx}{\int f^2(x)dx + \int g^2(x)dx}. \tag{5.2}$$

Notice that $\delta(f,g)$ is the scaled version of the "L_2" distance between f and g discussed in Section 3.8. The scaling gives $\delta(f,g)$ a natural range of $[0,1]$ as we see in the next lemma.

Lemma 5.1. *The discordance and concordance coefficients lie between 0 and 1 and reaches their extreme values when*

- $f = g \iff C = 1$ *and* $\delta = 0$
- $f \perp g$ *(i.e. completely different support sets)* $\iff C = 0$ *and* $\delta = 1$

Proof : First of all both the the numerator and the denominator of δ are positive quantities, so we have $\delta \geq 0$. Also

$$\begin{aligned} \delta = \frac{\int (f-g)^2 dx}{\int f^2 dx + \int g^2 dx} &= \frac{\int f^2 dx + \int g^2 dx - 2\int fg dx}{\int f^2 dx + \int g^2 dx} \\ &\leq 1 \quad \text{since for} \quad a,b>0, \qquad a^2+b^2 \geq (a-b)^2. \end{aligned}$$

Thus $0 \leq \delta \leq 1$. Moreover

$$\begin{aligned} f = g &\implies \int (f-g)^2 = 0 \implies \delta = 0 \\ f \perp g &\implies \int fg dx = 0 \implies \delta = 1 \end{aligned}$$

□

5.2 Concordance Correlation in the choice of g in the Mixture Model

As before, let us choose the quadratic distance with the normal kernel K, where

$$K(x,y) = \frac{1}{(2\pi)^{\frac{p}{2}} |\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x-y)'\Sigma^{-1}(x-y) \right)$$

and $\Sigma = h^2 I$, h being the "smoothing bandwidth parameter".

We extend the concordance/discordance coefficients to our model selection criterion. Based on the density interpretation discussed in Section 3.7, the generalized quadratic distance

between two densities F and G can be written as an L_2 distance on the kernel smoothed densities. If we write the concordance and discordance in terms of $f^*(w) = \int L(x,w)dF(x)$ and $g^*(w) = \int L(x,w)dG(x)$, we can define a kernel smoothed concordance coefficient between F and G as the concordance between f^* and g^*.

$$\begin{aligned} C(F,G) &= 1 - \frac{\int (f^* - g^*)^2 dx}{\int (f^*)^2 dx + \int (g^*)^2 dx} \\ &= 1 - \frac{\int K(x,y)\ d(F-G)(x)\ d(F-G)(y)}{\int K(x,x)\ dF(x)\ dF(x) + \int K\ dG(x)\ dG(x)} \qquad (5.3) \\ &= 1 - \frac{D_K(F,G)}{\int K(x,x)\ dF(x)\ dF(x) + \int K\ dG(x)\ dG(x)}. \qquad (5.4) \end{aligned}$$

Notice that $C(F,G)$ is a function of the smoothing parameter h because K is. For notational convenience we will use the following notation throughout this chapter:

$$K(F,G) = \int K(x,y)\ dF(x)\ dG(y).$$

so that equation (5.4) becomes

$$C(F,G) = 1 - \frac{D_K(F,G)}{K(F,F) + K(F,G)} \qquad (5.5)$$

5.3 Estimation

We now consider the estimation of $C(F,M)$, where F is the true model and M is a fitted model. For estimating the concordance coefficient we examine both the unbiased and biased estimators. The biased estimator is based on the *$\mathcal{V}$-statistic* (von Mises, 1947), while the unbiased estimator is based on the *$\mathcal{U}$-statistic* results (Serfling, 1980). We will estimate each of the quantities $D_K(F,M)$, $K(F,F)$, and $K(M,M)$ separately and combine them to get an estimator of the concordance coefficient $C(F,M)$. Using the results from Subsection 3.9.1, for a particular multivariate normal mixture model, the quantity $K(M,M)$ can be calculated

as,

$$\widehat{K(M,M)} = \sum_{l=1}^{g}\sum_{k=1}^{g}\pi_l\pi_k c_{lk}, \tag{5.6}$$

$$\text{where } c_{lk} = \frac{1}{(2\pi)^{\frac{p}{2}}|V_l+V_k+\Sigma|}\exp\left(-\frac{(\mu_i-\mu_j)'(V_l+V_k+\Sigma)^{-1}(\mu_i-\mu_j)}{2}\right).$$

From now onwards we will denote the biased estimates with a superscript " ^{b}" and the unbiased estimates with a superscript " ^{u}". From equation (3.7) biased estimates of $D_K(\hat{F},M)$ and K_{FF} are given by

$$\widehat{D_K(F,M)}^b = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\tilde{K}(x_i,x_j), \tag{5.7}$$

$$\widehat{K(F,F)}^b = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}K(x_i,x_j). \tag{5.8}$$

From equation (3.64) Unbiased estimates of $D_K(\hat{F},M)$ and $K(F,F)$ are given by

$$\widehat{D_K(F,M)}^u = \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq 1}^{n}\tilde{K}(x_i,x_j), \tag{5.9}$$

$$\widehat{K(F,F)}^u = \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq 1}^{n}K(x_i,x_j). \tag{5.10}$$

Thus the biased estimator of the concordance coefficient is given by

$$\widehat{C^b} = 1 - \frac{\widehat{D_K(\hat{F},M)}^b}{\widehat{K(F,F)}^b + \widehat{K(M,M)}}, \tag{5.11}$$

and the corresponding unbiased estimator is given by

$$\widehat{C^u} = 1 - \frac{\widehat{D_K(\hat{F},M)}^u}{\widehat{K(F,F)}^u + \widehat{K(M,M)}}. \tag{5.12}$$

Note that value of concordance coefficient of the unbiased estimator $\widehat{C^u}$ may be greater than 1, due to the fact that $\widehat{D_K(\hat{F},M)}^u$ can take on negative values.

5.4 Interpreting the concordance curves

Concordance curves can be used as an informal tool for model selection. First let us draw some analogy between the concordance value and the R^2 value used in regression analysis.

The R^2 value, in the context of regression, is interpreted as the proportion of variability of the dependent variable explained by the regression model. Likewise, the concordance value between a particular model and the empirical distribution, describes the amount of variability in the data density that has been explained by the model density considered. A concordance value of 'zero' can be interpreted as the model having no explanatory or predictive power, whereas a concordance value of 'one' means that the model density is an exact fit with the data density.

A parallel can be drawn between the selection of a number of components using the concordance values and selection of a subset of variables using the R^2 values, assuming there exists a large number of covariates. Addition of a new variable to the existing subset will always increase the R^2 value. Thus in order to check whether we should include a new variable in the subset of covariates, we might design a stopping rule based on either a target value of R^2 for the new subset, or the increase in value of R^2 due to using the new variable. The design of a good stopping rule depends on inferential objectives and subject matter knowledge.

We can use the idea of 'subset selection' for selecting the number of components in a mixture model. We have already noticed that a richer model (i.e, a model with more components) always fits the data better. Let us denote all g component models by M_g, and $\hat{M}_g$ the best g-component fit. Thus, irrespective of the value of the smoothing parameter, $D_K(F, \hat{M}_{g_0}) \geq D_K(F, \hat{M}_{g_0+1})$ for any g_0. Though a formal proof is not yet available, it has also been observed that, empirically, $K(\hat{M}_g, \hat{M}_g) < K(\hat{M}_{g_0+1}, \hat{M}_{g_0+1})$ for any g_0 (see Table 5.1). Thus, we expect $C(F, \hat{M}_{g_0}) \leq C(F, \hat{M}_{g_0+1})$ for any g_0. So, for a particular h, we can design a stopping rule for including extra components based on the concordance values and use it is an model selection criterion.

Next, we discuss Figure 5.1 in detail and show how the concordance values can be used as a tool for model selection. Though we present both the unbiased and biased concordance curves we will base our model selection results on the unbiased version. It has already

Table 5.1: **Calculation of the unbiased estimates of the concordance values for the Iris data set, with $h = 0.5$, and g ranging from 1 to 6.**

g	$\widehat{K(F,F)^u}$	$\widehat{K(M,M)}$	$\widehat{D_K(F,M)^u}$	$\widehat{C^u(F,M)}$
1	0.029	0.0180	0.0132	0.719
2	0.029	0.0265	0.0035	0.937
3	0.029	0.0259	0.0007	0.989
4	0.029	0.0264	0.0001	0.998
5	0.029	0.0269	-0.0002	1.005
6	0.029	0.0280	-0.0006	1.011

been observed that $h = 0.5$ is a reasonable smoothing parameter for the Iris data . The concordance values for $g = 1, 2, 3, 4, 5, 6$ are $.719, .937, .989, .998, 1, 1.01$ respectively. If we had set our stopping rule as the absolute value of the concordance being greater than .98, then we would have selected the model with 3 components. On the other hand if our stopping rule was set at concordance$>$ .9 we would have gone with the two component model. Also, if we had set the stopping rule as the increase in concordance value being more than .01, we would have selected the model with 3 components.

All these rules of model selection based on the concordance are ad-hoc. For this reason we call it an informal method of model selection. The method can be formalized if we derive the null distribution of the concordance values and then form a rejection region. Then we can associate a confidence value with the rejection or acceptance of a model that fits the data. Further research is needed in this area.

5.5 Results

In this section we will use the concordance values to select the number of components in the other three datasets given in the Appendix. For the Simulated Dataset 1 and 2 we have

already seen that $h = 0.5$ is a good choice for the "smoothing parameters", whereas for the acidity dataset we chose $h = 0.05$. We will base our model selection on the unbiased estimates of the concordance values for each of these dataset.

Looking at the $h = 0.5$ curve in Figure 5.3 we would decide on the number of components being 4 with the estimated concordance coefficient $\hat{C} = 1.0189$ (unbiased estimate). One interesting thing to notice here is that if we had based our model selection strictly on the relative increase in concordance, we might have decided on only 2 components, as there is not much increase in the concordance value from 2 to 3 components. However, when using concordance we would know $\hat{C} = .8976$ for three components. So we would look for a better fit, although we we would still know that a great improvement of fit occurred going from 1 to 2 component.

For the acidity dataset (Figure 5.7), looking at the concordance values for $h = .05$ we might choose 3 components with $\hat{C} = 1.0156$. If we had used the $h = .5$ curve in this Figure we would have selected 2 components with $\hat{C} = 1.00114$. Since $h = .5$ for the acidity dataset is "over-smoothing", some of the local variation is obscured and so we choose 2 components instead of 3. On the other hand even at $h = 0.5$, most of the fit improvement occurs going from 1 to 2 components

5.6 Using Concordance to choose the smoothing parameter

In this section we will discuss how the sensitivity of the concordance curve might be used to select an interesting range of h. Closely examining the concordance curves for the four datasets we can easily observe that large or small values of h produce a less responsive curve, while the middle values show significant increase with the increase of the number of components.

For example, let us examine the concordance curves of the Iris dataset (see Figure 5.2.

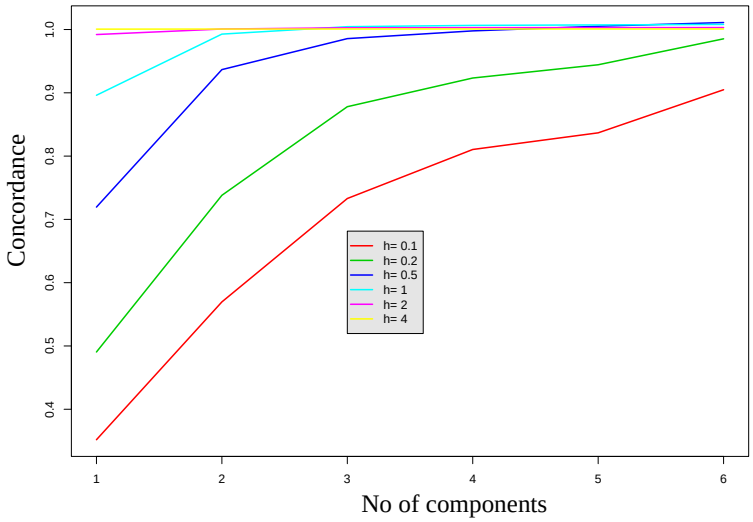

Figure 5.1: **Concordance values of the iris data for different h with unbiased estimates**

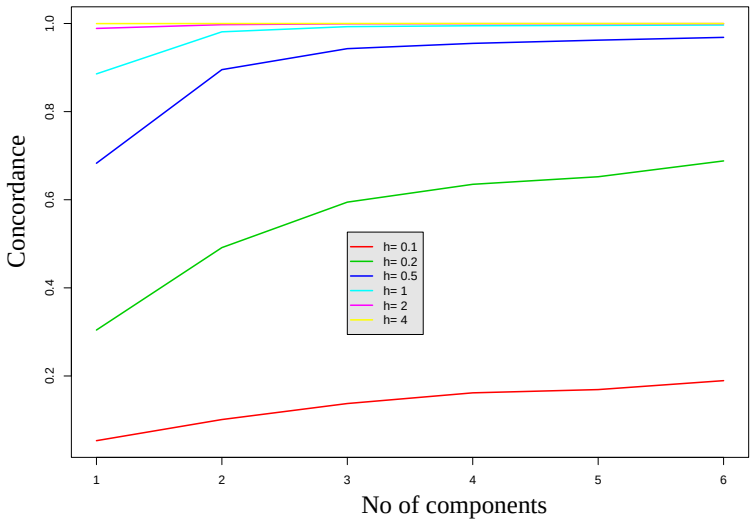

Figure 5.2: **Concordance values of the iris data for different h with biased estimates**

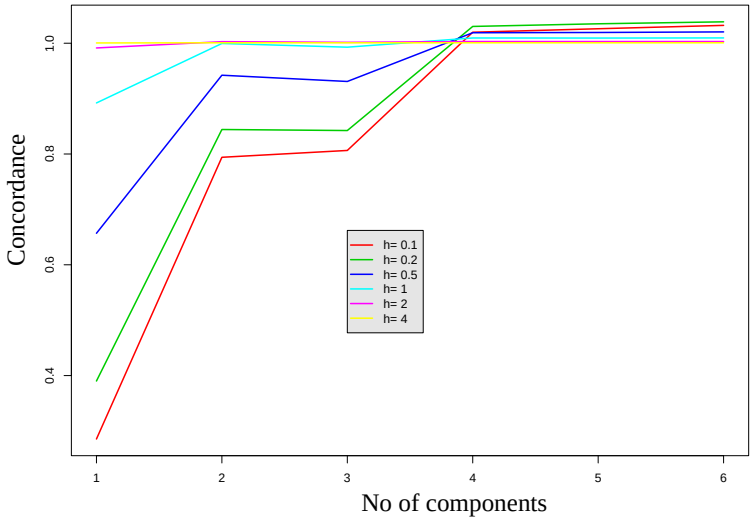

Figure 5.3: Concordance values of the Simulated dataset 1 for different h with unbiased estimates

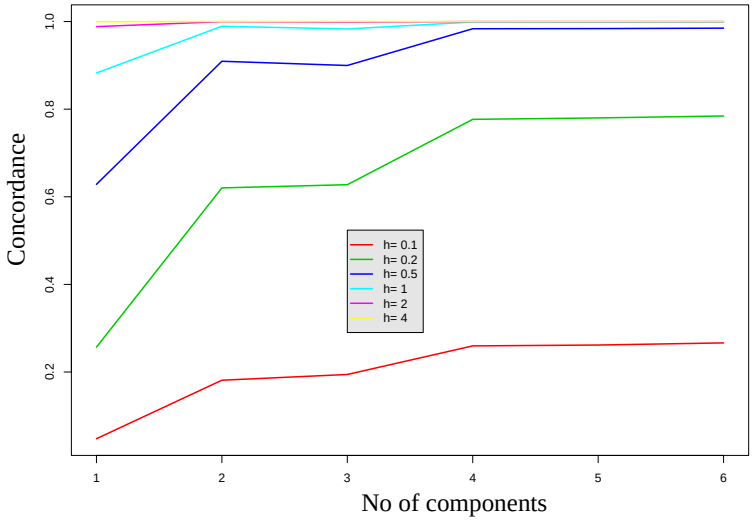

Figure 5.4: Concordance values of the Simulated dataset 1 for different h with biased estimates

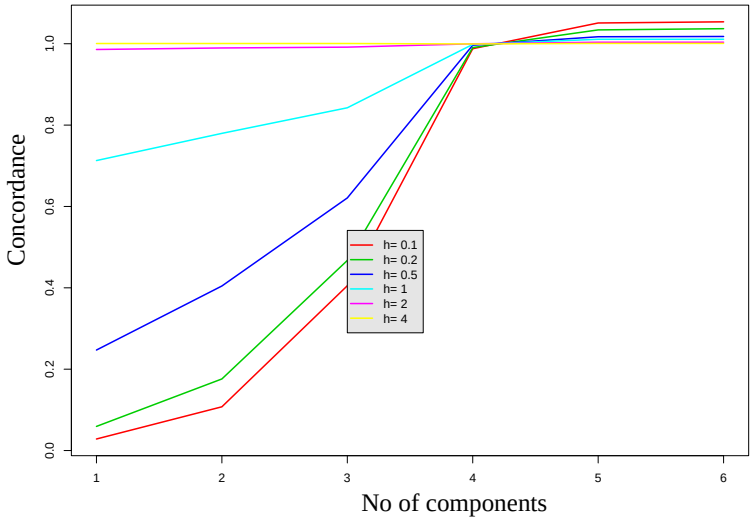

Figure 5.5: Concordance values of the Simulated dataset 2 for different h with unbiased estimates

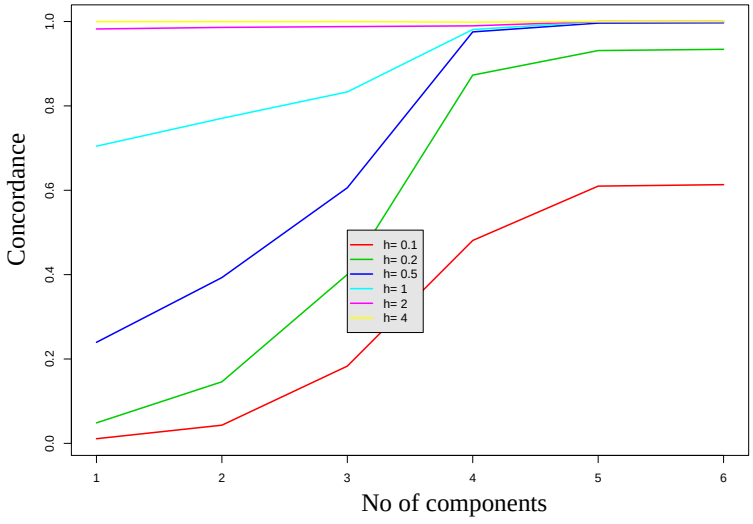

Figure 5.6: Concordance values of the Simulated dataset 2 for different h with biased estimates

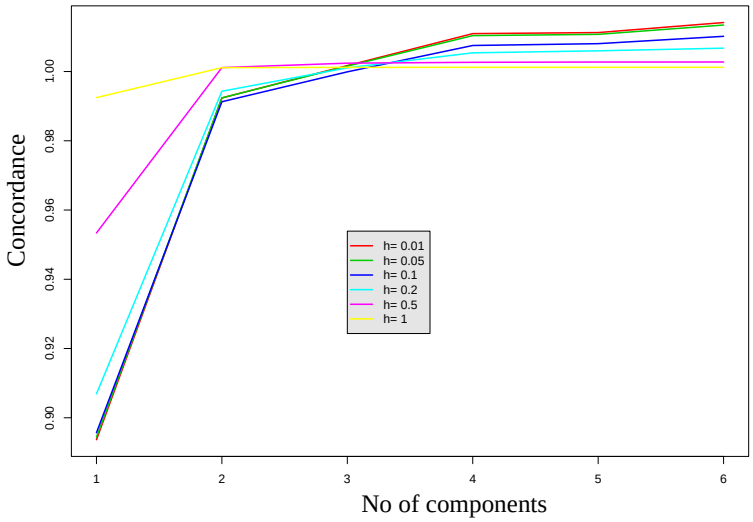

Figure 5.7: **Concordance values of the acidity data for different h with unbiased estimates**

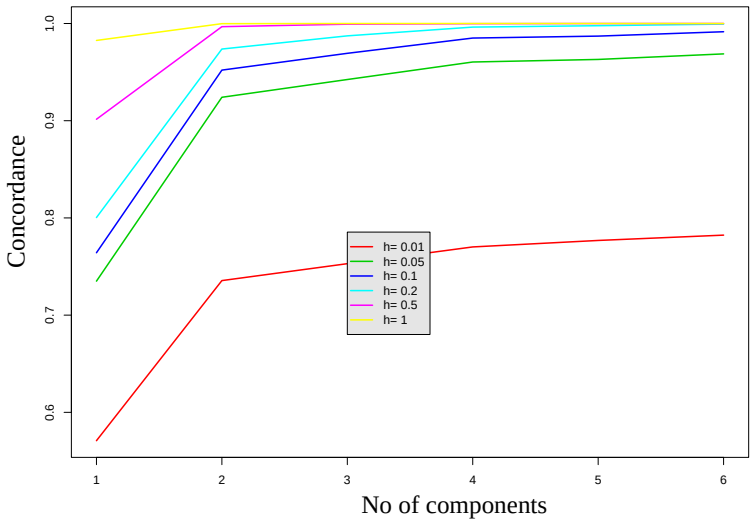

Figure 5.8: **Concordance values of the acidity data for different h with biased estimates**

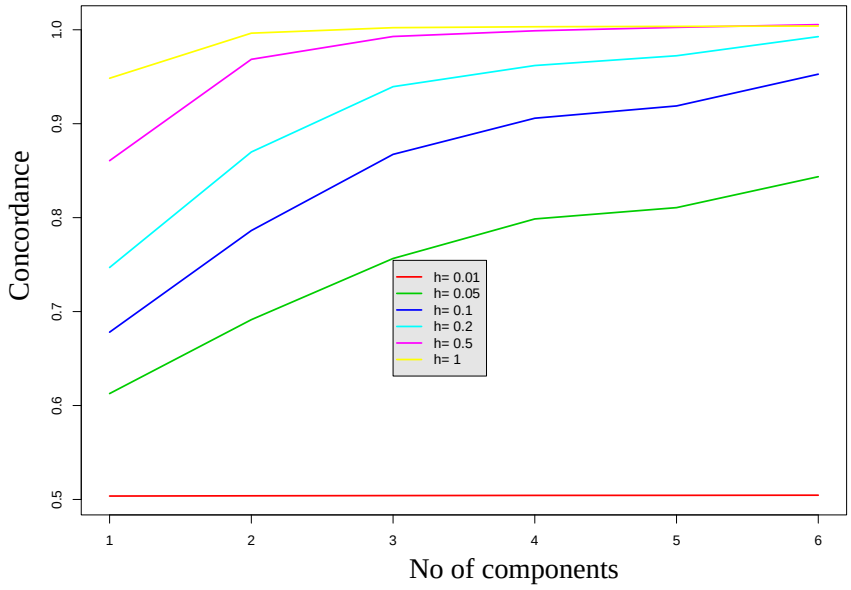

Figure 5.9: **Concordance values of the iris data for a different set of h with unbiased estimates**

Here we should note that this h is proportional to the unit variances of the scaled dataset, because whenever we calculate the concordance coefficient we scale the data to make each of the variables have unit variance. On this scale $h \geq 1$ is never a good choice as we can see in all the concordance curves. Intuitively, $h \geq 1$ uses a kernel which has a larger variance than the data, thus over smoothing it. From the concordance curves, this phenomenon can be observed from the fact that for almost all $h > 1$ the concordance coefficients are nearly 1 even with very few or even 1 component. This implies that for these h the concordance coefficients are not sensitive for different components. So most often our choice will be limited to $h < 1$.

On the other hand, if using very small h we would be forced to model even the "noise" by using more and more components, thus not achieving large values for the concordance coefficients even after fitting a considerable number of components. This can be observed from the concordance values for $h = .1$. To further illustrate this phenomenon another graph (Figure 5.9) is presented for the Iris data with smaller values of h.

It can easily be seen that for $h = .1$ and $h = .05$ the concordance curves for the iris dataset are much less responsive. For $h = .01$ the concordance curve is almost flat. Similar phenomenon can be observed for other datasets.

Any $h > 1$ produced a flat curve for the two generated datasets. However, as the clusters in the generated datasets were really well separated, the lower limit of h for which we still had substantial change in the concordance curve was lower than for the Iris dataset. In fact the concordance curves displayed strong slopes until $h = .01$. Though, the Iris dataset and the two generated datasets are of the same dimension, the range of h for which the concordance curve displayed a strong slope varied, because of the difference in the strength of the signal.

For the acidity dataset as observed before the concordance curves starts to display an interesting slope only when $h \geq .1$. This is because of the dimension of the data being 1.

All the observations so far indicate that there exists an interesting range of h beyond which the concordance curve looks almost flat and so is non-informative.

So, the first and foremost criterion of choosing an interesting range would be to find the range of h for which the shape of the concordance curve looks informative about questions of interest. In future research we intend to investigate the relationship between the choice of smoothing parameter h, the sensitivity of the distance to questions of interest, and the pseudo degrees of freedom.

5.7 Conclusion

The concordance coefficient provide a great deal of information about the fit of a model. It has many attractive properties; we believe it has a very good potential to be used as an model selection tool. Importantly, it is a non decreasing function of the number of components, and theoretically it lies between 0 and 1. Concordance values at different h helps us analyze the data at different levels of smoothness. But, the main hindrance in using this

method as a formal model selection tool is the same problem as for R^2, choosing the stopping rule. As mentioned earlier, this needs to be investigated in detail.

Chapter 6

Risk-based model selection

In this chapter we will take yet a different approach to the problem of model selection. We will estimate the statistical risk of the models and choose the model with minimum risk. To do so, we define a loss function for an estimator $\hat{M}$ based on a distance $D(\tau, \hat{M})$ and then let the risk equal the expected value of the loss function. For a particular dataset, model selected by the risk analysis need not be the same as the models select by other methods described in the dissertation. The reason is that the objective of risk-based model selection is somewhat different from the other methods.

6.1 Motivation

As we have already noticed, the problem of selecting the right number of components of the mixture distribution is hard because the fit gets better with more and more components. But, we should also observe that, as we fit more and more components the number of parameters to be estimated increases, too. AIC, BIC and other penalized likelihood based methods for model selection incorporate the above idea by penalizing the likelihood by some function of the number of parameters to be estimated. In this chapter we will marry the idea of quadratic distances with the notion of parameter estimation cost by assigning a measure of risk for each competing model. Again, let τ be the true distribution. Let us also define a set of competing models given by $\mathcal{M}$. For example, in our case $\mathcal{M}$ will be the set of all normal mixture model with a finite number of components. Note that, unlike the likelihood based penalized

methods, in our risk analysis based model selection we will not define the penalty function for parameters explicitly, rather the distance inherently takes care of the parameter estimation cost.

6.1.1 Generalized Quadratic Distance as the Loss function

We next define the loss function using the quadratic distance introduced in Chapter 3. Let $D(\tau, \hat{M})$ be the loss incurred when using estimator $\hat{M}$ to estimate τ. This in turn defines a risk function

$$R_n(\tau) = E_\tau[D(\tau, \hat{M})], \tag{6.1}$$

where n is the sample size used to estimate M.

Defining the distance as the loss function has many desirable properties. We can use all the properties of estimation of distance to calculate the estimated risk of model. In particular, $\mathcal{U}-$statistics results can be applied to find an unbiased estimator of the risk. The estimation of the risk will be discussed in detail in Section 6.3

Furthermore, the risk has a very attractive interpretation. It can be decomposed explicitly into a model lack-of-fit item plus a parameter estimation cost. This is discussed in detail in the following section.

6.2 Decomposition of the Risk

In this section we will illustrate how the total risk $E_\tau[D(\tau, \hat{M})]$ can be broken into two parts, one attributable to the lack-of-fit of the model class and the other part evaluating the parameter estimation cost. Let $D(\tau, \mathcal{M}) = \inf_{M \in \mathcal{M}} D(\tau, M)$ be called the Model lack-of-fit or model building error. We assume the infimum is attained at $M = M_\tau$. The model lack-of-fit is zero if the model class is correct and otherwise is an intrinsic error occurring from using the incorrect model class for τ.

However in practice we will have $\hat{M}$, where $\hat{M}$ is an estimator of the model M based on $\hat{F}$, i.e. the data. We can decompose the loss function when using $\hat{M}$, into two parts

$$D(\tau,\hat{M}) \quad = \quad D(\tau,M_\tau) + \left[D(\tau,\hat{M}) - D(\tau,M_\tau)\right], \tag{6.2}$$

where $D(\tau,M_\tau)$ is the *model lack-of-fit* and $\left[D(\tau,\hat{M}) - D(\tau,M_\tau)\right]$ is a positive term because M_τ is the distance minimizer. Taking expectations of equation (6.2) we get

$$E_\tau[D(\tau,\hat{M})] \quad = \quad D(\tau,M_\tau) + E_\tau\left[D(\tau,\hat{M}) - D(\tau,M_\tau)\right], \tag{6.3}$$

where $E_\tau\left[D(\tau,\hat{M}) - D(\tau,M_\tau)\right]$ represents the parameter estimation cost. We expect that its magnitude is strongly related to the number of parameters estimated. We will investigate this feature further in our simulations.

Thus $E_\tau[D(\tau,\hat{M})]$, the risk associated with using model estimation method $\hat{M}$ is the sum of the model error and the mean parameter estimation cost.

6.3 Estimation

In this section we will use the results on distance to calculate the estimated risk of a fitted model. Let us first introduce some notation which will be used through out the rest of this chapter.

- $\widehat{M_{<i,j>}}$ is the fitted model M whose estimates are calculated on the basis of all but observations X_i and X_j. In general, $\widehat{M_{<\mathcal{S}>}}$ is the fitted model with the estimates calculated leaving out those X whose indices are in $\mathcal{S}$, where $\mathcal{S} \subset \{1,2,\ldots,n\}$. We will denote a subset $\mathcal{S}$ containing exactly n_1 elements by $\mathcal{S}_{n_1}$.

- The delete-n_1 risk is based on $n-n_1$ data points, denoted as $R_{n-n_1}(\tau)$, and is given by

$$R_{n-n_1}(\tau) = E_\tau\left[D(\tau,\widehat{M_{<\mathcal{S}>}})\right].$$

Now let us start with the definition of the "leave-two-out estimator",

$$\hat{R}_{n-2}(\tau) = \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{j<i} \widehat{\tilde{K}^{M(i,j)}}(x_i, x_j). \tag{6.4}$$

It can be easily seen that, $\hat{R}_{n-2}(\tau)$ is an unbiased estimator of $R_{n-2}(\tau)$.

An intuitive idea behind the "leave-two-out" estimator is that we expect the risk to decrease as long as the the fitted distribution, leaving the two out, also fits those two observations well, but the risk should start increasing when we fit too many components because the fitted distribution will be able to fit the $(n-2)$ observations very well, but at the price of poorly fitting the omitted observations. Since we would expect $R_{n-2}(\tau)$ to be close to $R_n(\tau)$ for large n, we will use this estimator to assess our risk in using our model $\hat{M}$.

One of the difficulties of implementing this method is that each time we leave out two or more observations, we have to recompute all the parameters based on the remaining observations. Although, we can start the EM algorithm, for the new parameter estimates, using as the starting values the estimates based on all the n observations, the whole process becomes computationally expensive, especially, as the number of components increases. Moreover, in the leave-two-out estimation, the point estimates change so little that the difference in distance before and after deletion was hard to measure accurately. Finally the leave-two-out estimator requires n^2 distance calculations.

So, for practical purposes, we implemented the following strategy. We deleted n_1 observations, then calculated the model estimate $\widehat{M_{<\mathcal{S}_{n_1}>}}$, and took an average of the risks over the repeated random deletions of size n_1. Taking the average over Q sets of deletions the estimator can be written as

$$\hat{R}_{n-n_1} = \frac{1}{Q} \sum_{q=1}^{Q} \left(\frac{1}{n_1(n_1-1)} \sum_{i,j \in \mathcal{S}_q} \sum_{j \neq i} \widehat{K^{M_{<\mathcal{S}_q>}}}(x_i, x_j) \right). \tag{6.5}$$

Note that $\hat{R}_{n-n_1}$, like $\hat{R}_{n-2}$ is still an unbiased estimate of R_{n-n_1}, but with a lower precision. In our calculations we have used $Q = 20$ and $n_1 = 30$ (about 20% of the number of observations).

Along with the estimates of risk for specified models, we would also like to have a significance level of the increase or decrease in risk from model A to model B. One may think that the significance values can be calculated as the paired difference of the Q samples from each model, where the p-value is calculated with respect to t with $Q-1$ degrees of freedom (paired-sample t test). But it should be noted that the Q samples are correlated. So, for now we avoid calculating the significance values and base our conclusions on the absolute value of the changes in risk.

6.4 Results

In this section we discuss examples of model selection based on the risk analysis criterion. Description of the 4 datasets discussed below can be found in Appendix I.

Based on the *pDOF* analysis in Chapter 4 the risk-based model selection are done taking $h = 0.5$ for the Iris data and the two simulated data. For the Iris data (see Figure 6.1) a decision based on minimum risk would suggest 3 components, as the risk is clearly minimized for 3 components. Similarly, for the simulated dataset 1 (see Figure 6.2), we would go for 4 components, as there is no significant improvement in the of risk for 5 and 6 components. For the simulated dataset 2, where we have four distinct clusters our decision method gives a strong indication to go for 4 components (see Figure 6.3). For the acidity dataset with $h = .05$, we would choose a 3 component model (see Figure 6.4).

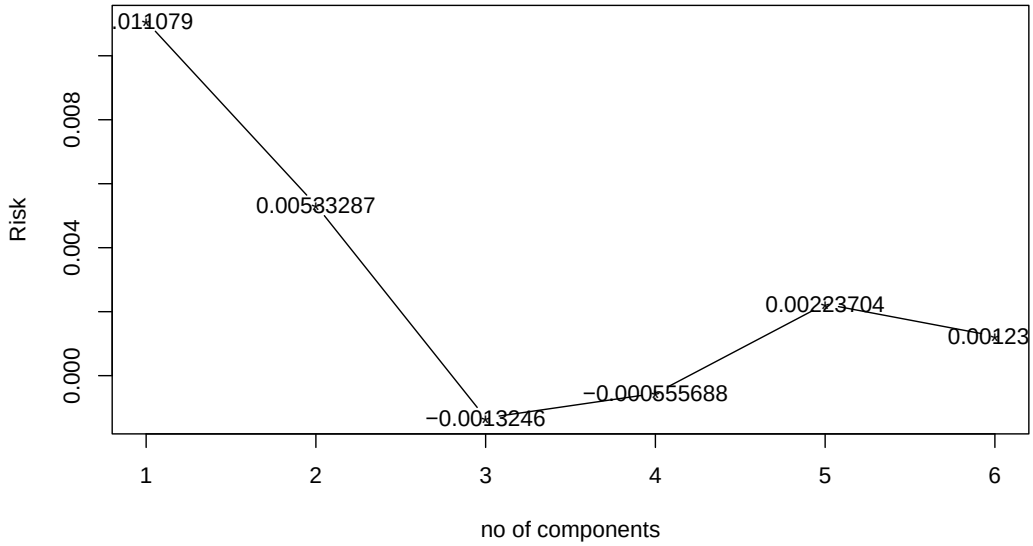

Figure 6.1: Risk analysis of the Iris data with h=.5

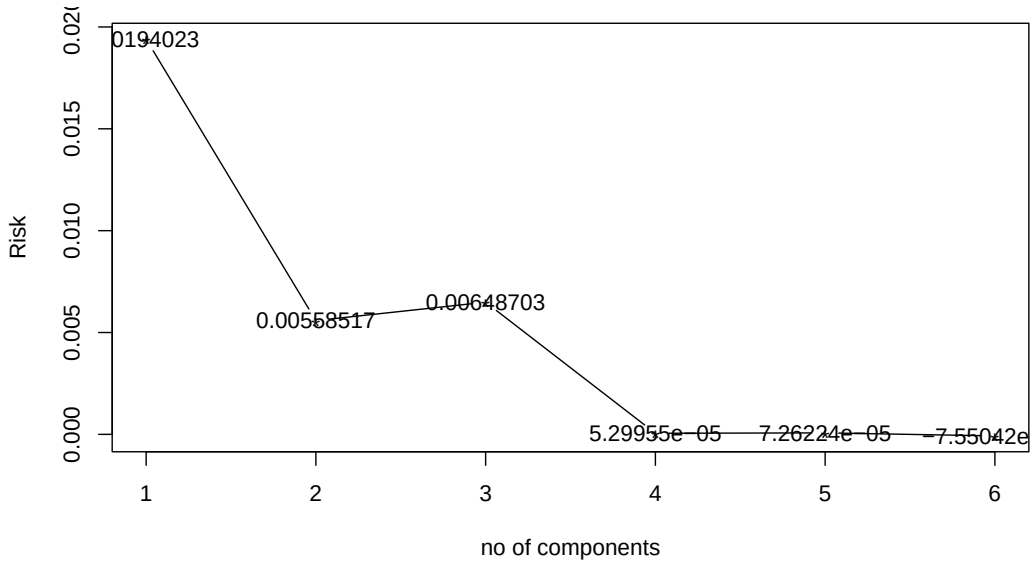

Figure 6.2: Risk analysis of the Simulated Dataset 1 with h=.5

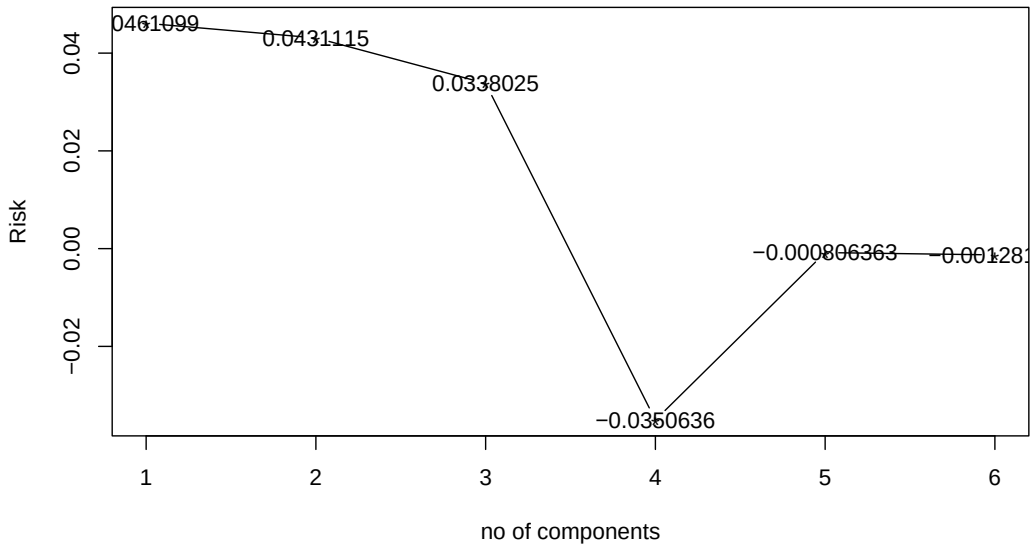

Figure 6.3: Risk analysis of the Simulated Dataset 2 with h=.5

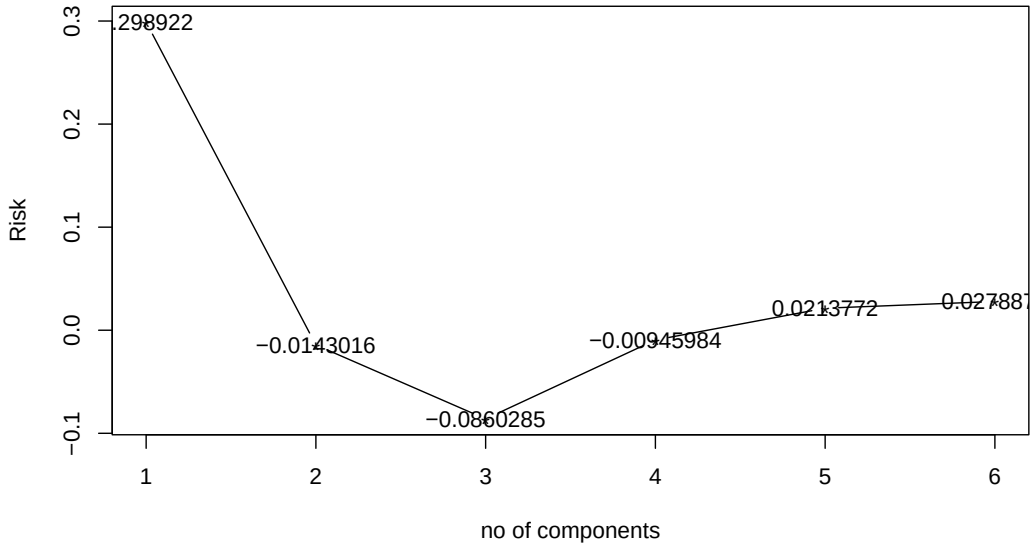

Figure 6.4: Risk analysis of the Acidity data with h=.05

6.5 Conclusion

The risk of a model can be used as a very important and general tool in selecting a model. As mentioned earlier, we should not compare the results of risk analysis with other model selection methods described in this dissertation, as the optimizing criteria are not the same. Here the best model is chosen based on the optimum balance between the model lack-of-fit and the parameter estimation cost. In the other methods, we wish only to get close to the true τ without regard to the parameter cost. Note that, like the other distance based methods, the risk based analysis also allows one to "play" with the "smoothing parameter" h, in the analysis of the model which could result in different choices for g.

Future research in this topic could include the selection of the optimum deletion parameter (n_1). Further research should also be done for calculating a significance level for the change in risk from model to model.

Chapter 7

Residual Analysis through Quadratic distance

In this chapter we will discuss some initial ideas about how the unbiased estimator of the quadratic distance could be used as diagnostic tool for outlier detection. We believe that these outliers could be used for detection of extra components as well as indicating the locations of the needed components. The most attractive feature of these residuals is that we do not have to calculate them explicitly; rather they are a natural outcome of the distance estimation. At this time, issues regarding standardization and determining a cutoff value for extreme values remain to be rigorously worked out. However, we believe that the generalized quadratic residuals we describe next could prove to be an important diagnostic tool in model assessment.

7.1 Motivation

Our definition of model residuals arises from a re-expression of the unbiased estimator of the quadratic distance. Using the *$\mathcal{U}$-statistics* results found in Chapter 3 we showed that the unbiased estimator of the distance $D_K(\tau, G)$ can be written as,

$$U_n(G) = \frac{1}{n(n-1)} \sum_i \sum_{j \neq i} \tilde{K^G}(x_i, x_j), \tag{7.1}$$

where $X_1, X_2, \ldots, X_n$ are random samples from the distribution τ. The above equation can be re-written as

$$\begin{aligned} U_n(G) &= \frac{1}{n}\sum_i \left[\frac{1}{n-1}\sum_{j\neq i}\tilde{K^G}(x_i,x_j)\right], \\ &= \frac{1}{n}\sum_i r(x_i) \end{aligned} \tag{7.2}$$

where $r(x_i) = \frac{1}{n-1}\sum_{j\neq i}\tilde{K^G}(x_i,x_j)$ will be called the *quadratic residual* at the data point X_i. From the expression in equation (7.2) the total distance $U_n(G)$ can be thought as the sum of residuals of all the data points. Since large values of the estimated distance provide evidence against G, it might be anticipated that $r(x_i)$ should somehow represent the contribution of the distance coming from observation $X_i = x_i$

To understand how $r(x_i)$ serves as a measure of the effect of an observation on the estimated distance one can ask the question: Given the rest of the dataset, how does the estimated distance change if we add X_i? Without loss of generality, let $i = n$, and delete X_n. Then

$$\begin{aligned} U_n(G) - U_{n-1}(G) &= \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\tilde{K^G}(x_i,x_j) - \frac{1}{(n-1)(n-2)}\sum_{i=1}^{n-1}\sum_{j\neq i}^{n-1}\tilde{K^G}(x_i,x_j) \\ &= \left[\frac{1}{n(n-1)} - \frac{1}{(n-1)(n-2)}\right]\sum_{i=1}^{n-1}\sum_{j\neq i}^{n-1}\tilde{K^G}(x_i,x_j) + \frac{1}{n(n-1)}2\sum_{i=1}^{n-1}\tilde{K^G}(x_i,x_n) \\ &= -\frac{2}{n}U_{n-1}(G) + \frac{2}{n}r(x_n). \end{aligned} \tag{7.3}$$

That is, the change in estimated distance is a function of the new data point only through the magnitude of $r(x_n)$. As an aside, from equation (7.3) we can get the following recursive formula for calculating the distance,

$$U_n(G) = \frac{n-2}{n}U_{n-1}(G) + \frac{2}{n}r(x_n). \tag{7.4}$$

To understand the structure of the residuals further, we decompose $r(x_i)$ as follows

$$\begin{aligned} r(x_i) &= \frac{1}{n-1}\sum_{j\neq i}\tilde{K^G}(x_i,x_j) \\ &= \frac{1}{n-1}\sum_{j\neq i}[K(x_i,x_j)-K(x_i,G)-K(G,x_j)+K(G,G)] \\ &= \frac{1}{n-1}\sum_{j\neq i}K(x_i,x_j)-K(x_i,G)-\frac{1}{n-1}\sum_{i\neq j}K(G,x_j)+K(G,G) \\ &= \frac{1}{n-1}\sum_{j\neq i}K(x_i,x_j)-\left(1-\frac{1}{n-1}\right)K(x_i,G)-\frac{1}{n-1}\sum_{j=1}^{n}K(G,x_j)+K(G,G). \end{aligned} \tag{7.5}$$

Now, the last two terms in equation (7.5) remain constant over all i. Moreover, as $n\to\infty$, when G is correct, their sum goes to zero. So finally, the variability in the values of $r(x_i)$ over the sample can be attributed to the first two terms of equation (7.5).

Let $f^*_{<i>}(t)=\frac{1}{n-1}\sum_{j\neq i}K(t,x_j)$. Then $f^*_{<i>}(t)$ is a kernel density estimator for τ based on the data with X_i deleted. And the first term in equation (7.5) is $f^*_{<i>}(x_i)$. This value objectively defines how well the i^{th} data point agrees with the rest of the empirical distribution. On the other hand, $K(G,t)=g^*(t)$ corresponds to a kernel smoothed density for G, so $g^*(x_i)$ represents the smoothed null density at the observation. Here we can draw a similarity with the Pearson's χ^2 test in that $r(x_i)$ nearly has the form of observed minus expected frequency. Note, that the last two terms in equation (7.5) are required to make the overall distance unbiased. So, the residuals can be re-written as

$$r(x_i) = f^*_{<i>}(x_i)-\left(1-\frac{1}{n-1}\right)g^*(x_i)-c, \quad \text{for } i=1,2,\ldots,n, \tag{7.6}$$

where $c=\frac{1}{n-1}\sum_{j=1}^{n}K(G,x_j)+K(G,G)$ and $f^*_{<i>}(x_i)$ and $g^*(x_i)$ already described in the previous paragraph. For identifying extreme values we could as well work with only

$$\tilde{r}(x_i)=f^*_{<i>}(x_i)-\left(1-\frac{1}{n-1}\right)g^*(x_i), \quad \text{for } i=1,2,\ldots,n \tag{7.7}$$

instead of $r(x_i)$.

7.2 Standardizing the residuals

Unfortunately, without standardization these residuals may not reflect the "surprise" that one has due to the observation. Indeed we have found that the raw residuals are not reliable at detecting outliers. This is still a subject of ongoing research, about which we have the following comments.

Viewed as a random function of the data, the $r(x_i)$ are exchangeable in distribution, and so they have constant variance. They are also correlated with each other. If the spectral decomposition corresponding to the kernel $\tilde{K}^G$ under G were available, we might use the orthogonal eigen-functions as diagnostics for sources of model lack of fit.

Since such a decomposition is not feasible, we might instead ask the question as follows: Given the data points $X_1, X_2, \ldots, X_{n-1}$ and the model G, is the new data point X_n "surprising" in the sense of the magnitude of its change in the estimated distance. In doing this we seek a conditional distribution under model G, with $X_1, X_2, \ldots, X_{n-1}$ fixed at $x_1, x_2, \ldots, x_{n-1}$.

We first note that the conditional mean of $r(x_n)$ is zero under G:

$$E[r(X_n)|x_1, x_2, \ldots, x_{n-1}] = \frac{1}{n-1}\sum_{i=1}^{n-1} E[\tilde{K}^G(X_n, x_i)] = 0 \tag{7.8}$$

Secondly, the conditional variance has the form

$$\sigma^2_{r,n} = \frac{1}{(n-1)^2} E_G\left[\sum_{a=1}^{n-1}\sum_{b=1}^{n-1} \tilde{K}^G(X_n, x_a)\tilde{K}^G(X_n, x_b)\right] \tag{7.9}$$

$$= \frac{1}{(n-1)^2}\sum_{a=1}^{n-1}\sum_{b=1}^{n-1} E_G\left[\tilde{K}^G(X_n, x_a)\tilde{K}^G(X_n, x_b)\right]. \tag{7.10}$$

If we calculate $h(t_1, t_2) = E_G\left[\tilde{K}^G(X, t_1)\tilde{K}^G(X, t_2)\right]$, then

$$\sigma^2_{r,n} = \frac{1}{(n-1)^2}\sum_{a=1}^{n-1}\sum_{b=1}^{n-1} h(x_a, x_b). \tag{7.11}$$

This leads to the standardized residuals

$$r^*_0(x_i) = \frac{r(x_i)}{\sigma_{r,n}}. \tag{7.12}$$

An alternative analysis would be to condition on $X_n = x_n$. In this case $\sum_{i=1}^{n-1} \tilde{K^G}(x_n, x_i)$ is an *i.i.d.* sum. It is also conditionally mean zero and so the conditional variance under G is

$$\begin{aligned} \sigma^2_{1,X_n} &= E_G\left[\left(\frac{1}{n-1}\sum_{i=1}^{n-1}\tilde{K^G}(x_n,x_i)\right)^2\right] && (7.13)\\ &= \frac{1}{n-1}E_G\left[\left(\tilde{K^G}(x_n,x_i)\right)^2\right]. && (7.14)\end{aligned}$$

These variances do depend on X_n and thus provide local standardizing. They can be calculated as follows

$$\begin{aligned} \sigma^2_{1,X} &= E\left[\left(\tilde{K_G}(x,X)\right)^2\right]\\ &= E\left[\left(K(x,X)-K(x,G)-K(G,X)+K(G,G)\right)^2\right]\\ &= E[(K(x,X)-K(x,G))^2]+E[(K(G,X)-K(G,G))^2]-\\ &\quad 2E[(K(x,X)-K(x,G))(K(G,X)-K(G,G))]\\ &= Var_G[K(x,X)]+Var_G[g^*(X)]-2Cov_G[K(x,X),g^*(X)]. && (7.15)\end{aligned}$$

Through rigorous calculation it can be shown that for G being a p-component univariate normal with mean μ and variance V

$$\begin{aligned} \sigma^2_{1,X} &= \frac{1}{(2\sqrt{\pi})^p}\Big(\frac{1}{|\Sigma_h|^{\frac{1}{2}}}K_{\Sigma_h+V}(X,\mu)-\frac{1}{|\Sigma_h+V|^{\frac{1}{2}}}K_{\Sigma_h/2+V/2}(X,\mu)\\ &\quad +\frac{1}{|\Sigma_h+V|^{\frac{1}{2}}}K_{\Sigma_h/2+3V/2}(X,\mu)-\frac{1}{|\Sigma_h+2V|^{\frac{1}{2}}}K_{\Sigma_h/2+V}(X,\mu)\Big)\\ &\quad -2\Big(\frac{1}{|\Sigma_h|^{\frac{1}{2}}|\Sigma_h+2V|^{\frac{1}{2}}\left|\left(\Sigma_h^{-1}+(\Sigma_h+2V)^{-1}\right)^{-1}\right|^{\frac{1}{2}}}K_{\Sigma_h+\left(\Sigma_h^{-1}+(\Sigma_h+V)^{-1}\right)}(X,\mu)\\ &\quad -\frac{1}{|\Sigma_h|^{\frac{1}{2}}|\Sigma_h+V|^{\frac{1}{2}}\left|\left((\Sigma_h+V)^{-1}+(\Sigma_h+2V)^{-1}\right)^{-1}\right|^{\frac{1}{2}}}K_{((\Sigma_h+V)^{-1}+(\Sigma_h+2V)^{-1})}(X,\mu)\Big). \end{aligned} \quad (7.16)$$

From this point one can consider either using estimates of variance by using the sample variance $s^2_{1X_n}$ of $\tilde{K}(x_n,X_1),\tilde{K}(x_n,X_1),\ldots,\tilde{K}(x_n,X_{n-1})$ or by carrying out the above calculation (7.15), which can be done explicitly in the normal mixture. This yields two forms for the

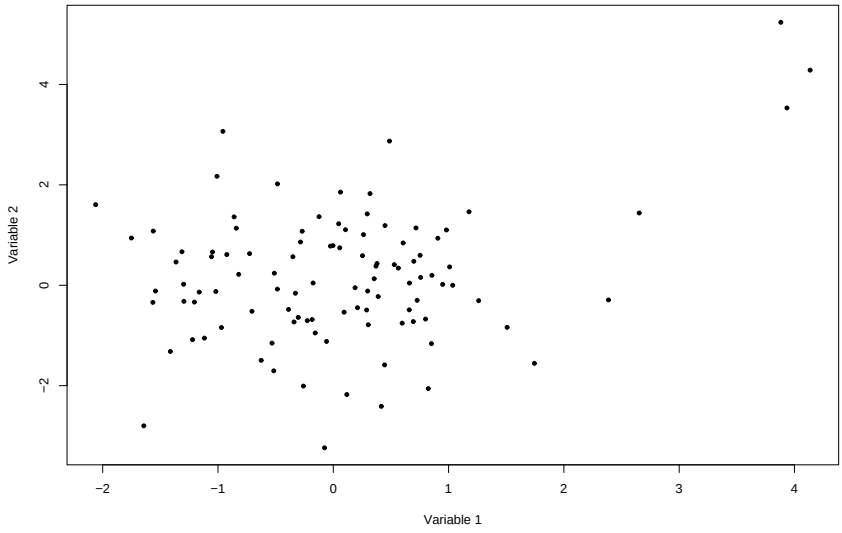

Figure 7.1: **Plot of 100 sample from f_1 and 3 sample from f_2**

residuals,

$$r_1^*(x_i) = \frac{r_1(x_i)}{s_{1X_i}} \tag{7.17}$$

$$\text{and} \quad r_2^*(x_i) = \frac{r_1(x_i)}{\sigma_{1X_i}}. \tag{7.18}$$

7.3 Results

In this section we will use standardized residuals to detect outliers in a synthetic data. To demonstrate the idea through plots we examine a bivariate data set. Let f_1 and f_2 be two bivariate normals with mean vector $\mu_1 = (0,0)'$ and $\mu_2 = (4,4)'$ respectively, and with common variance being the identity matrix. The first 100 observations were generated from f_1 and 3 were generated from f_2. Thus, the dataset has 103 points, among which the last three are possible outliers. A plot of the 103 data points is given in Figure 7.1.

In Figure 7.2 data points denoted by "$\star$" denote the 10 most extreme residuals using the standardized residuals r_1^*. The next ten extreme residuals are denoted by "$\star$". Figure 7.3

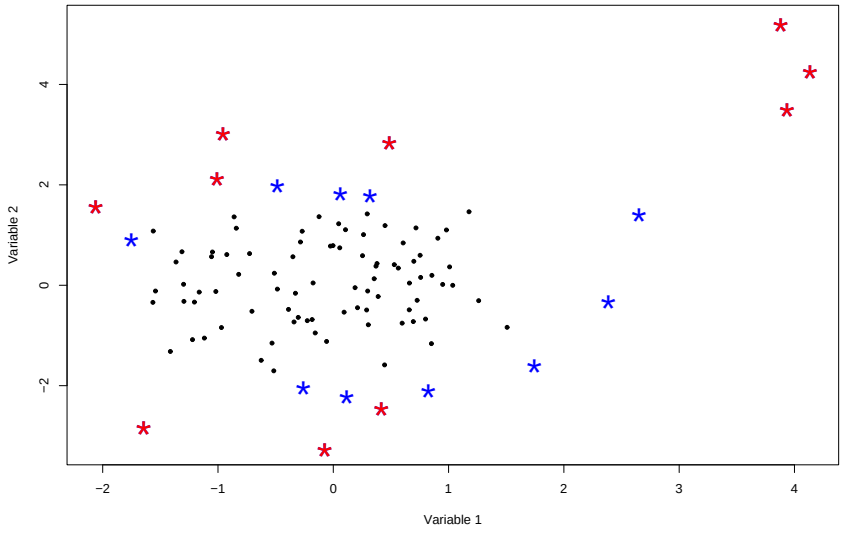

Figure 7.2: **Detecting outliers based on the standardized residuals** $r_1^*(x_i)$

gives the histogram for the standardized residuals. displays the outliers for standardized residuals $r_1^*(x_i)$. Here, we observe that the three data points from f_2 are not just classified among the 10 more extreme values of the residuals, they are outliers in the residual distribution (see Figure 7.3). In the histogram, the three bins with frequency one are the three outliers. In fact, the magnitude of the three observations from the distribution f_2 was extremely high. These 10 most extreme residuals also include all the outlying points of the distribution f_1. Table 7.1 gives the value of the 10 largest standardized residuals, using the $r_1^*(x_i)$ values. For a comparative study, we also give the histogram of the raw residuals in Figure 7.4. From the histogram it is clear that the raw residuals would have been unable to detect the outliers.

From the magnitude of the residuals some "ad hoc" rule of of outlier detection can be proposed. For now, we propose plotting the residual distribution to determine if there are very unusual points.

A further note: although we identify individual points as outliers, we are actually

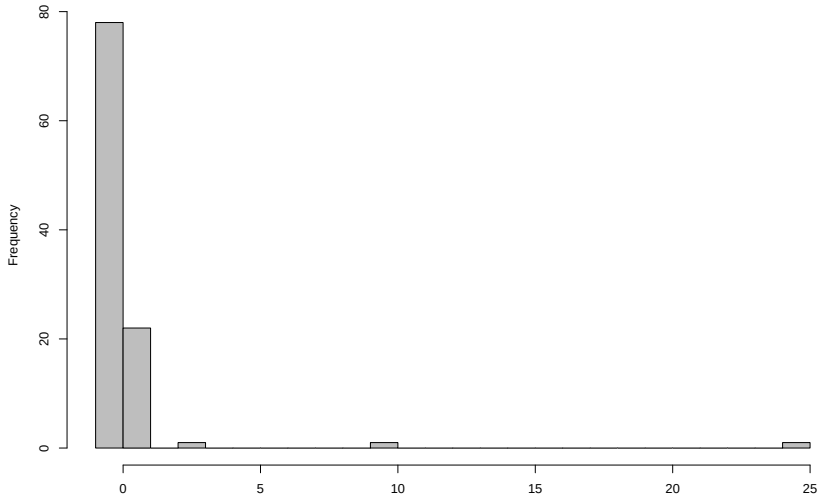

Figure 7.3: **Histogram of the standardized residuals** r_1^*

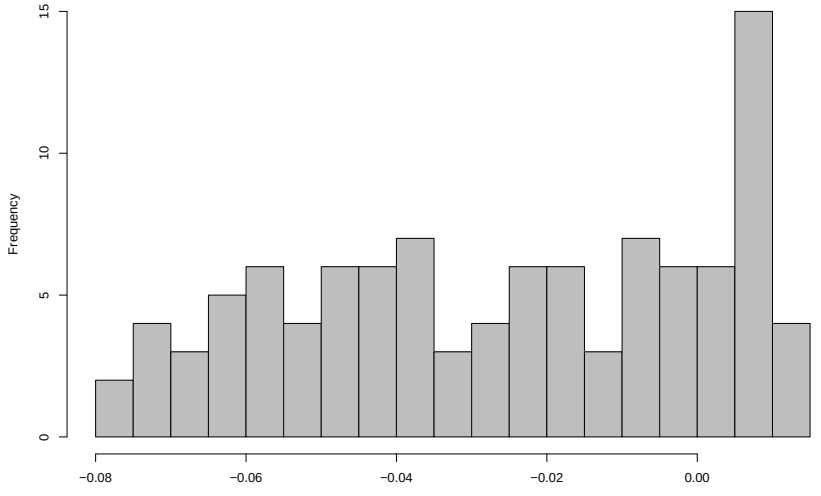

Figure 7.4: **Histogram of the raw residuals** *r*

Table 7.1: **Residual analysis of the simulated data with 3 outliers, using standardized residuals (only 10 largest values are shown)**

Data Coordinates		r	$\sigma^2_{1,X}$	r_1	Rank
x	y				(Largest to smallest)
3.881	5.238	0.00293	1.458406e-08	24.3069	1
4.135	4.285	0.00372	1.564925e-07	9.4168	2
3.935	3.532	0.00321	2.237121e-06	2.1480	3
-0.957	3.066	0.00747	3.687338e-03	0.1230	4
-0.076	-3.237	0.00714	3.405210e-03	0.1224	5
-1.644	-2.800	0.00582	3.323357e-03	0.1010	6
0.487	2.871	0.00911	8.670933e-03	0.0978	7
-2.061	1.605	0.01039	1.889262e-02	0.0755	8
-1.009	2.171	0.01271	3.168819e-02	0.0714	9
0.416	-2.411	0.01045	2.806342e-02	0.0623	10

looking at the model discrepancies in the neighborhood of the point, where the point itself is excluded. Thus we would only find outlier clusters of two or more, not singletons, using this method.

7.4 Conclusion

Residual analysis is a natural outcome of the distance estimation process for the quadratic distance. We should note that the outliers considered here also depend on the value of the "smoothing parameter" of the kernel, so the same data point can be an extreme outlier for some value of h, but at some other value h' the residual of that data point might not be significantly large. In the particular case of finding the number of components the residual analysis may be used to find and explore further clusters to be added to the model. That is, if we find a several outliers together we may want to check whether we can find an extra

component consisting of the close knit outliers. We would also get an idea of the location of the component. Residual analysis in the generalized quadratic distance seems quite promising. Further research needs to be done to standardize the residuals and find cut-off values to determine outliers.

Chapter 8

Detection Number of Modes in Two Component Mixture

Modality is one objective way to define what we mean by distinct clusters. As we have discussed in Chapter 2, even though the number of modes and the number of components may not be the same, the number of modes provides further insight into the distribution of a non-homogeneous population. Considerable work has been done for the detection of bimodality in univariate models with specific distributions. Helguero (1904) determined necessary and sufficient conditions for bimodality in the mixture of univariate normals with equal variances and mixing proportions. Later, conditions for bimodality in the mixture of univariate normal distribution with unequal variance and unequal mixing proportion was studied by Eisenberger (1964), Behboodian (1970) and Robertson and Fryer (1969). Kakiuchi (1981) and Kemperman (1991) addressed conditions for bimodality using non-normal component densities.

In this chapter we will investigate modality conditions for a mixture of multivariate normal densities. To our knowledge, no other previous results are available on this topic. Bimodality in the multivariate situation is harder to describe in terms of the first and second order differentials. We propose a new method for detecting modal structure in a mixture of multivariate normals. For the equal variance matrix case we have an "if and only if" condition for bimodality which depends on the mean vectors and the common covariance matrix. For the unequal variance matrix case we develop plotting methods which are guaranteed to detect the number of modes. However, simple algebraic conditions for bimodality for a mixture of

multivariate normal, when the component densities have unequal variances, are not given. In the univariate case such formulas exist; we doubt their existence in the multivariate case.

In this chapter determination of modality refers to determination of modality in the context of the two component multivariate (say p-variate) normal distribution. In other words we will determine the modality of the density $f_{\mathbf{X}}(\mathbf{x})$, where

$$\mathbf{X} \sim \pi\mathcal{N}(\mu_1, \Sigma_1) + (1-\pi)\mathcal{N}(\mu_0, \Sigma_0), \tag{8.1}$$

where μ_i's and Σ_i's are the means and variances of the component i, $i = 1, 2$ and π is the proportion of the 1^{st} component. Denoting the the density of a multivariate normal with mean μ and variance Σ, by $\phi(\mathbf{x}; \mu, \Sigma)$ we can write $f_X(x)$ as,

$$f_{\mathbf{X}}(\mathbf{x}) = \pi\phi(\mathbf{x}; \mu_1, \Sigma_1) + (1-\pi)\phi(\mathbf{x}; \mu_0, \Sigma_0) \tag{8.2}$$

The work on generalization to more than two components is still in progress but the answer is bound to be significantly more complex.

One way of detecting bimodality in the univariate case is by visual inspection of a density histogram. But, for the multivariate case visual inspection is not a good option. Here, we present an example to illustrate why the detection of bimodality, from density plots is difficult, even in the bivariate case. Figure 8.1 displays the density function of a mixture of normal with the means $\mu_1 = (-1, -1)'$, $\mu_1 = (1, 1)'$ and common variance $\Sigma = I_2$. For detecting the bimodality in the density, visually inspecting the 3-dimensional plot is possible, but not easy. So, we look at the marginal distribution of only one variable. The distribution along the x and y axis is the same and is given by Figure 8.2. The density along the x and y axis is not bimodal. But if we look at the distribution along the line $x = y$ (Figure 8.3), we can easily detect the bimodality. Visual inspection will be even more difficult in greater than two dimensions. Nor is it clear that multivariate bimodality will always be evident form a marginal plot.

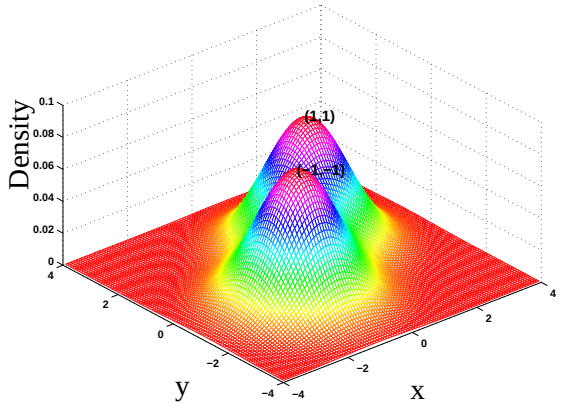

Figure 8.1: **Density plot of the mixture of two bivariate normals with means $\mu_1 = (-1,-1)'$, $\mu_1 = (1,1)'$ and common variance $\Sigma = I_2$.**

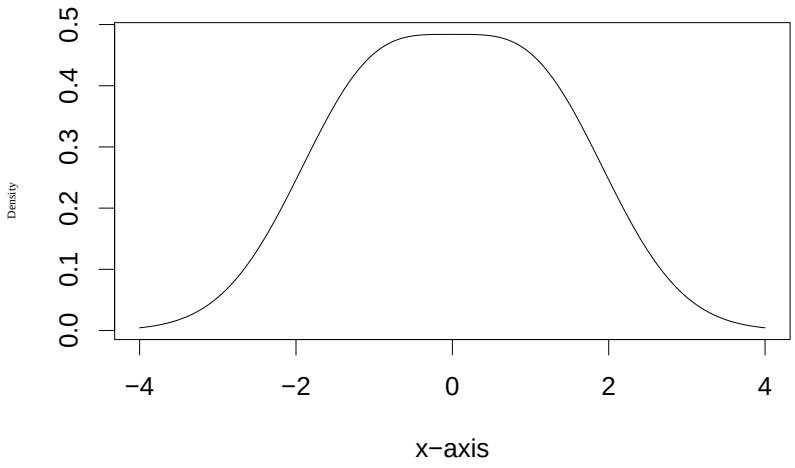

Figure 8.2: **Marginal distribution of the mixture of two bivariate normals along the axes**

In this chapter we will show that there exists a one dimensional curve in $\mathbf{X}$ such that, in order to detect all the modes of (8.2), it is necessary and sufficient to determine the modes of the density along that curve.

The layout of the chapter is as follows. First, we determine conditions for bimodality in a mixture of multivariate normals when the component variances are equal. It will be seen

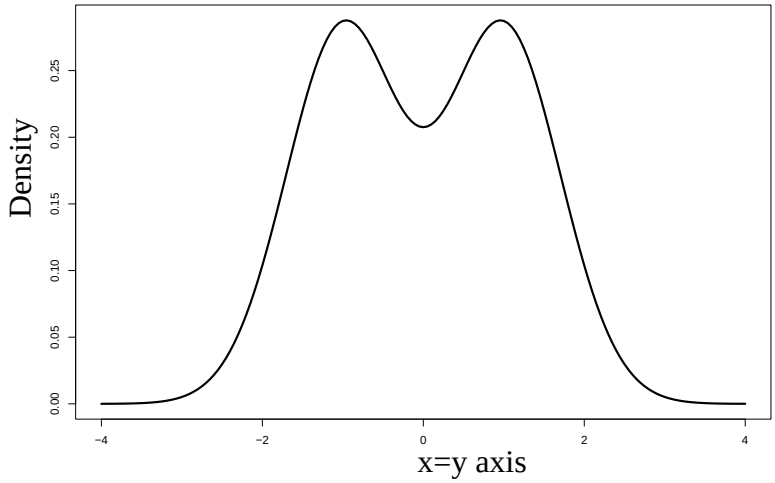

Figure 8.3: **Distribution of the mixture of two bivariate normals along the "axis of maximum separation"**

that the condition for bimodality of the multivariate distribution can be reduced to the univariate condition of bimodality for the density of a specific linear combination of the variables. We will denote this linear combination as the "axis of maximum separation".

Then we proceed to the case of unequal variances. We derive several plots one can use to detect bimodality. They are based on reducing the p-dimensional problem to a one dimensional problem by constructing an explicit "ridge-line" curve along which all the modes must occur. In the unequal variance case the modes can be surely detected if we can detect modality along the ridge-line curve.

8.1 Detection of Bimodality: The equal variance case

In this section we will derive conditions for bimodality for a mixture of multivariate normals where the components have equal variance. We will derive the "axis of maximum separation" and the inference on the modality of the multivariate distribution will be based on the modality of the distribution of the univariate distribution over the "axis of maximum

separation".

8.1.1 The "axis of maximum separation" for a multivariate normal mixture

A mixture of multivariate normals is bimodal if and only if it is bimodal along some line. By this we mean that, if a mixture of multivariate normals is bimodal, with modes at $\mathbf{X}_1$ and $\mathbf{X}_0$, then the density is bimodal when calculated as a function of α along the line $\alpha\mathbf{X}_0 + \bar{\alpha}\mathbf{X}_1$. It turns out that if $\Sigma_1 = \Sigma_0$, then we will show there is also a vector $\mathbf{V}$, corresponding to a change of axes, such that $\mathbf{V}'\mathbf{X}$ has a bimodal univariate density. If $\Sigma_1 = \Sigma_0$, then our density of interest is

$$f_{\mathbf{X}}(\mathbf{x}) = \pi\phi(\mathbf{x};\mu_1,\Sigma) + (1-\pi)\phi(\mathbf{x};\mu_0,\Sigma). \tag{8.3}$$

Let $\mathbf{Y} = \Sigma^{-\frac{1}{2}}(\mathbf{X} - \mu_1)$. Then,

$$\mathbf{Y} \sim \pi\mathcal{N}(0,I) + (1-\pi)\mathcal{N}(\mu_y,I) \quad \text{where } \mu_y = \Sigma^{-\frac{1}{2}}(\mu_0 - \mu_1). \tag{8.4}$$

Let B be an orthonormal matrix of size p with the first row γ_1' being proportional to μ_y. Thus by Gram-Schmidt orthogonalization we can have B, such that

$$B = \begin{bmatrix} \gamma_1' \\ \gamma_2' \\ \vdots \\ \gamma_p' \end{bmatrix} \tag{8.5}$$

$$\text{where } \gamma_1 = \frac{\mu_y}{\|\mu_y\|},\ \gamma_i'\gamma_j = 0\ \ \forall i \neq j, \text{ and } \gamma_i'\gamma_i = 1\ \ \forall i.$$

Thus

$$\mathbf{Z} = B\mathbf{Y} \sim= B\Sigma^{-\frac{1}{2}}(\mathbf{X} - \mu_1) = (\pi\mathcal{N}(0,I) + (1-\pi)\mathcal{N}(\mu_y^*,I), \tag{8.6}$$

$$\text{where } \mu_y^* = B\mu_y = \begin{pmatrix} \|\mu_y\| \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

Lemma 8.1. *If* $\mathbf{Z} = (Z_1, Z_2, \ldots, Z_p)$ *is a random variable such that*

$$\begin{pmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_p \end{pmatrix} \sim \pi\mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, I_p\right) + (1-\pi)\mathcal{N}\left(\begin{pmatrix} c \\ 0 \\ \vdots \\ 0 \end{pmatrix}, I_p\right), \tag{8.7}$$

then $Z_1, Z_2, \ldots, Z_p$ *are statistically independent.*

Proof : The moment generating function (MGF) of $\mathbf{Z}$ denoted by $M_{\mathbf{Z}}(t)$ is

$$\begin{aligned} M_{\mathbf{Z}}(t) &= E[exp(t'\mathbf{Z})] \\ &= \pi \exp\left(\frac{1}{2}t'It\right) + (1-\pi)\, exp\left(t_1c + \frac{1}{2}t'It\right) \\ &= \pi \exp\left(\frac{1}{2}\sum_{i=1}^{p} t_i^2\right) + (1-\pi)\, exp\left(t_1c + \frac{1}{2}\sum_{i=1}^{p} t_i^2\right) \\ &= \exp\left(\frac{1}{2}\sum_{i=2}^{p} t_i^2\right)\left[\pi \exp\left(\frac{1}{2}t_1^2\right) + (1-\pi)\, \exp\left(t_1c + \frac{1}{2}t_1^2\right)\right] \\ &= M_{Z_1}(t_1)M_{Z_2}(t_0)\ldots M_{Z_p}(t_p). \end{aligned} \tag{8.8}$$

Thus $Z_1, Z_2, \ldots, Z_p$ are statistically independent. □

Moreover, $Z_2, Z_3, \ldots, Z_p$ are independent one component normals, and Z_1 is the only variable which is a mixture of two normals and corresponds to the axis of maximal separation. So if we can show that the modality of the original random variable $\mathbf{X}$ is preserved under the transformation $\mathbf{Z} = B\Sigma^{\frac{1}{2}}(\mathbf{X} - \mu_1)$, then we have the following if and only if condition, " $\mathbf{X}$ has g modes iff Z_1 has g modes".

Lemma 8.2. *Let A be any positive definite* $p \times p$ *matrix. A r.v.* $\mathbf{X}$ *has g modes to its density* $f_{\mathbf{X}}(\mathbf{x})$ *iff* $\mathbf{Z} = A\mathbf{X}$ *has has g modes to its density* $f_{\mathbf{Z}}(\mathbf{z})$.

Proof : Let $x_1, x_2, \ldots, x_g$ be the g distinct modes of the density $f_{\mathbf{X}}(\mathbf{x})$. Thus, we have,

$$\frac{\partial}{\partial \mathbf{x}} f(\mathbf{x}_i) = 0 \text{ for } i = 1, 2, \ldots g, \tag{8.9}$$

$$\frac{\partial^2}{\partial \mathbf{x}\partial \mathbf{x}'} f(\mathbf{x}_i) \text{ is a negative definite matrix for } i = 1, 2, \ldots g. \tag{8.10}$$

The density function of $\mathbf{Z}$ is given by

$$f_{\mathbf{Z}}(\mathbf{z}) = f_{\mathbf{X}}(\mathbf{x})\frac{1}{|A|} \tag{8.11}$$

Therefore,

$$\begin{aligned}\frac{\partial}{\partial \mathbf{z}} f_{\mathbf{Z}}(\mathbf{z}) &= \left[\frac{\partial}{\partial \mathbf{x}} f_{\mathbf{X}}(\mathbf{x})\right]\frac{\partial \mathbf{x}}{\partial \mathbf{z}}\frac{1}{|A|} \\ &= \frac{A^{-1}}{|A|}\left[\frac{\partial}{\partial \mathbf{x}} f_{\mathbf{X}}(\mathbf{x})\right]\end{aligned} \tag{8.12}$$

and

$$\frac{\partial^2}{\partial \mathbf{z}\partial \mathbf{z}'} f_{\mathbf{Z}}(z) = \frac{A^{-1}}{|A|}\left[\frac{\partial^2}{\partial \mathbf{x}\partial \mathbf{x}'} f_{\mathbf{X}}(\mathbf{x})\right]A^{-1}. \tag{8.13}$$

Note that A is positive definite implies A^{-1} is also positive definite. Equations (8.9) and (8.12) $\Longrightarrow \frac{\partial}{\partial \mathbf{z}} f_{\mathbf{Z}}(\mathbf{z}) = 0$ for $\mathbf{z} = \mathbf{z}_i = A\mathbf{x}_i$, $i = 1, 2, \ldots g$ and equations (8.9) and (8.12) $\Longrightarrow \frac{\partial^2}{\partial \mathbf{z}\partial \mathbf{z}'} f_{\mathbf{Z}}(\mathbf{z})$ is negative definite for $\mathbf{z} = \mathbf{z}_i = A\mathbf{x}_i$, $i = 1, 2, \ldots g$. Also as A is positive definite $\mathbf{x}_i \neq \mathbf{x}_j \Longrightarrow \mathbf{z}_i \neq \mathbf{z}_j$ for all $i \neq j$. Thus if $\mathbf{X}$ has a g modal density so does $\mathbf{Z} = A\mathbf{X}$. The "only if" condition can be verified trivially since $X = A^{-1}Z$ and A^{-1} is also a positive definite matrix. □

We know that Z_1, being the mixture of two univariate normal densities, can have at most two modes (Behboodian, 1970). Thus using lemma 8.1 and 8.2 we can conclude that $\mathbf{X}$ has at most a bimodal density and is bimodal iff Z_1 has a bimodal density. Now,

$$\begin{aligned} Z_1 &= 1^{st} \text{ element of } B\Sigma^{-\frac{1}{2}}(\mathbf{X} - \mu_1) \\ &= \gamma_1'\Sigma^{-\frac{1}{2}}(\mathbf{X} - \mu_1) \\ &= \frac{(\mu_0 - \mu_1)\Sigma^{-1}(\mathbf{X} - \mu_1)}{(\mu_0 - \mu_1)\Sigma^{-1}(\mu_0 - \mu_1)}. \end{aligned} \tag{8.14}$$

Thus the variable of maximum separation is $(\mu_0 - \mu_1)\Sigma^{-1}(\mathbf{X} - \mu_1)$.

8.1.2 Conditions for bimodality of Multivariate Mixture

Before deriving the conditions for bimodality in the multivariate situation, we would like to mention the bimodality condition for the mixture of univariate normals.

Theorem 8.3. *(Helguero, 1904). Let X be the mixture of two univariate normal random variable with mean μ_1 and μ_0, ($\mu_1 \neq \mu_0$), equal variance σ^2 and equal mixing proportion. The distribution of X is bimodal iff*

$$\frac{|\mu_0 - \mu_1|}{\sigma} \geq 2. \tag{8.15}$$

Thus, based on theorem 8.3, lemmas 8.1 and 8.2 we have the following corollary,

Corollary 8.1. *Let $\mathbf{X}$ be the mixture of two multivariate normals, with unequal means μ_1 and μ_0, equal variances Σ, and equal mixing proportions $\pi = 1 - \pi = .5$ The density of $\mathbf{X}$ is bimodal iff*

$$(\mu_1 - \mu_0)' \Sigma^{-1} (\mu_1 - \mu_0) > 4. \tag{8.16}$$

The known modality conditions for univariate mixtures with arbitrary mixing proportions (see Theorem 8.4 with $\sigma_1 = \sigma_0$) can be applied to the multivariate case again by using the density of the marginal univariate distribution over the "axis of maximum separation".

8.2 Detection of bimodality: the unequal variance case

Deriving the conditions for bimodality in the unequal variance case is a challenging subject of research. Univariate conditions of bimodality in the case of normal mixtures are discussed in Eisenberger (1964), Behboodian (1970). Robertson and Fryer (1969) and recently Schilling et al. (2002) derived the following condition for bimodality in the univariate case:

Theorem 8.4. *Let $X \sim \pi\mathcal{N}(\mu_1, \sigma_1^2) + (1-\pi)\mathcal{N}(\mu_0, \sigma_0^2)$ and $r = \dfrac{\sigma_1^2}{\sigma_0^2}$.*

- *Then X has a unimodal distribution for all $0 < \pi < 1$, iff*

$$|\mu_1 - \mu_0| \leq S(r)(\sigma_1 + \sigma_0) \tag{8.17}$$

$$\text{where } S(r) = \frac{\sqrt{-2 + 3r + 3r^2 - 2r^3 + 2(1 - r + r^2)^{\frac{3}{2}}}}{\sqrt{r}(1 + \sqrt{r})}.$$

- *For* $|\mu_1 - \mu_0| \geq S(r)(\sigma_1 + \sigma_0)$, *X is bimodal iff* $\pi \in (\pi_1, \pi_0)$ *where,*

$$\pi_i^{-1} = 1 + \frac{r^{\frac{3}{2}} y_i}{\mu - y_i} \exp\left\{ -\frac{1}{2} y_i^2 + \frac{1}{2}\left(\frac{y_i - \mu}{\sqrt{r}}\right)^2 \right\} \text{ for } i = 1,2 \text{ and } \mu = \frac{\mu_0 - \mu_1}{\sigma} \quad (8.18)$$

and y_1 *and* y_0 *are the solutions of the cubic equation*

$$(r-1)y^3 - \mu(r-2)y^2 - \mu^2 y + \mu r = 0$$

with $0 < y_1 < y_0 < \mu$.

Otherwise, X is unimodal.

In the multivariate situation with unequal variances the problem of determining modality becomes more difficult. Earlier we argued that if a multivariate distribution has two modes there exists a line (the line between modes) along which the joint density is bimodal. However, the converse is not necessarily true. However, we now show how to construct a one dimensional curve along which the converse is true.

8.2.1 The X-modality curve

Theorem 8.5. *Let* $f_{\mathbf{X}}(\mathbf{x})$ *be the mixture of two multivariate normal densities, namely*

$$f_{\mathbf{X}}(\mathbf{x}) = \pi\phi(\mathbf{x}; \mu_1, \Sigma_1) + \bar{\pi}\phi(\mathbf{x}; \mu_0, \Sigma_0).$$

Then all of $f_{\mathbf{X}}(\mathbf{x})$ *'s critical values, and hence modes, lie along the curve in* α *defined by*

$$x_\alpha = \left[\alpha\Sigma_1^{-1} + \bar{\alpha}\Sigma_0^{-1}\right]^{-1}\left[\alpha\Sigma_1^{-1}\mu_1 + \bar{\alpha}\Sigma_0^{-1}\mu_0\right], \quad (8.19)$$

where $\alpha \in [0,1]$ *and* $\bar{\alpha} = 1 - \alpha$.

Proof : Suppose that $\nabla f_{\mathbf{X}}(\mathbf{x}^*) = 0$ so $\mathbf{x}^*$ is a critical point. Then we have

$$0 = \pi\phi(\mathbf{x}^*; \mu_1, \Sigma_1)\frac{\nabla\phi(\mathbf{x}^*; \mu_1, \Sigma_1)}{\phi(\mathbf{x}^*; \mu_1, \Sigma_1)} + \bar{\pi}\phi(\mathbf{x}^*; \mu_0, \Sigma_0)\frac{\nabla\phi(\mathbf{x}^*; \mu_0, \Sigma_0)}{\phi(\mathbf{x}^*; \mu_0, \Sigma_0)} \quad (8.20)$$

Let $\alpha = \dfrac{\pi\phi(\mathbf{x}^*;\mu_1,\Sigma)}{\pi\phi(\mathbf{x}^*;\mu_1,\Sigma)+\bar{\pi}\phi(\mathbf{x}^*;\mu_0,\Sigma)}$, which is between 0 and 1. Note further that

$$\frac{\nabla\phi(\mathbf{x}^*;\mu,\Sigma)}{\phi(\mathbf{x}^*;\mu,\Sigma)} = -\Sigma^{-1}(\mathbf{X}^*-\mu). \tag{8.21}$$

Thus, we have from equation (8.20), that for every critical value $\mathbf{x}^*$ there exists an α such that

$$\alpha\Sigma_1^{-1}(\mathbf{x}^*-\mu_1)+\bar{\alpha}\Sigma_0^{-1}(\mathbf{x}^*-\mu_0)=0. \tag{8.22}$$

Solving this equation for $\mathbf{x}^*$ gives the theorem.

□

In general, as α varies from 0 to 1, $\mathbf{X}_\alpha$ is a curve from μ_0 to μ_1 along which the modes and saddle points must occur. We will call this line the **X**-modality curve. It could also be called the "ridge-line curve" because of its similarity to a mountain ridge-line on which the saddles and the peaks occur. Therefore, to check whether the distribution of **X** is multimodal or unimodal, we can focus our attention 'to the density on the ridge-line $\mathbf{X}_\alpha$.

This ridge-line can be given a second interpretation. Consider any contour $\{\mathbf{X} : \phi_1(\mathbf{x}) = c\}$ of the component ϕ_1. This forms an ellipse. Provided μ_0 is not inside the ellipse, there exists constant d such that the ellipse of the other component $\{\mathbf{X} : \phi_0(\mathbf{x}) = d\}$ just touches the first ellipse. One can show that the point $\tilde{\mathbf{X}}$ they have in common is necessarily a point on the ridge-line, and that all points on the ridge-line have this characteristic. The interpretation of such point is that, no matter which direction one heads from it, one of $\{\phi_1(\mathbf{x}),\phi_0(\mathbf{x})\}$ increases and the other decreases, so that it is not possible to go in any direction in which both decrease.

8.2.2 Plots for detecting modality on the basis of the X-modality curve

Our next problem is to develop some diagnostic tools for the modality analysis. We will focus on plotting methods. We will also treat the component parameters as fixed in each

analysis, and seek to determine the modality as a function of π. Based on the structure of the **X**-modality curve we define the "density curvature plot" to be the plot of $(\phi_1(x_\alpha), \phi_0(x_\alpha))$ as curve in α (See Figure 8.4). Note that, the density value $\pi\phi_1 + \bar{\pi}\phi_0$ is the inner product of $(\pi, \bar{\pi})$ and (ϕ_1, ϕ_0). And the local maxima or the modes correspond to $(\pi, \bar{\pi}) \perp (\phi_1', \phi_0')$, that is the inner product of $(\pi, \bar{\pi})$ and (ϕ_1', ϕ_0') being 0, where ϕ_i' refers to the α-derivative of ϕ_i. That is, $(\pi, \bar{\pi})$ is orthogonal to the tangent vector of the curve at the critical points in α. This means that the values of π with multiple modes must have multiple tangent points.

To calculate $\phi_1(x_\alpha)$ we need to calculate the exponent. Let $\Sigma_\alpha = \alpha\Sigma_1 + \bar{\alpha}\Sigma_0$. Then

$$\begin{aligned}
&(x_\alpha - \mu_1)'\Sigma_1^{-1}(x_\alpha - \mu_1) \\
&= \left[\Sigma_\alpha^{-1}\left[\alpha\Sigma_1^{-1}\mu_1 + \bar{\alpha}\Sigma_0^{-1}\mu_0\right] - \mu_1\right]'\Sigma_1^{-1}\left[\Sigma_\alpha^{-1}\left[\alpha\Sigma_1^{-1}\mu_1 + \bar{\alpha}\Sigma_0^{-1}\mu_0\right] - \mu_1\right] \\
&= \left[\Sigma_\alpha^{-1}\bar{\alpha}\Sigma_0^{-1}(\mu_1 - \mu_0)\right]'\Sigma_1^{-1}\left[\Sigma_\alpha^{-1}\bar{\alpha}\Sigma_0^{-1}(\mu_1 - \mu_0)\right] \\
&= (\mu_1 - \mu_0)'\underbrace{\bar{\alpha}\Sigma_0^{-1}\Sigma_\alpha^{-1}\Sigma_1^{-1}\Sigma_\alpha^{-1}\bar{\alpha}\Sigma_0^{-1}}_{\alpha^2 W_{1\alpha}.}(\mu_1 - \mu_0). \qquad (8.23)
\end{aligned}$$

Similarly,

$$\begin{aligned}
(x_\alpha - \mu_0)'\Sigma_0^{-1}(x_\alpha - \mu_0) \quad &= \quad (\mu_1 - \mu_0)'\alpha^2 W_{2\alpha}(\mu_1 - \mu_0), \qquad (8.24) \\
&\text{where } W_{2\alpha} = \Sigma_1^{-1}\Sigma_\alpha^{-1}\Sigma_0^{-1}\Sigma_\alpha^{-1}\Sigma_1^{-1}.
\end{aligned}$$

We would later find it useful to have the square roots of W_1 and W_0. Note that

$$\left(\Sigma_0^{-1}\Sigma_\alpha^{-1}\Sigma_1^{-\frac{1}{2}}\right)^{-1} = \Sigma_1^{\frac{1}{2}}\Sigma_\alpha\Sigma_0 = \alpha\Sigma_1^{-\frac{1}{2}}\Sigma_0 + \bar{\alpha}\Sigma_1^{\frac{1}{2}} = P_{1\alpha} \text{ (say)}$$

$$\implies \Sigma_0^{-1}\Sigma_\alpha^{-1}\Sigma_1^{-\frac{1}{2}} = \left(\alpha\Sigma_1^{-\frac{1}{2}}\Sigma_0 + \bar{\alpha}\Sigma_1^{\frac{1}{2}}\right)^{-1} \implies W_{1\alpha} = P_{1\alpha}^{-2}. \qquad (8.25)$$

Also taking $P_{2\alpha} = \alpha\Sigma_0^{\frac{1}{2}} + \bar{\alpha}\Sigma_1\Sigma_0^{-\frac{1}{2}}$ we have $W_{2\alpha} = P_{2\alpha}^{-2}$.

We claim that if this curve has no sign changes in its curvature, then we get a unimodal density for all values of π because $(\pi, \bar{\pi})$ cannot be orthogonal to (ϕ_1', ϕ_0') more than once.

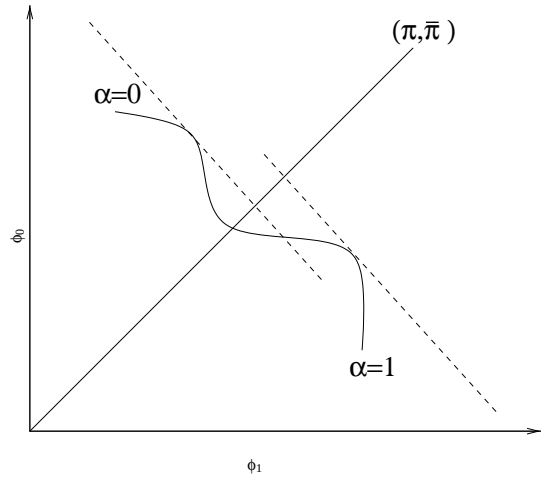

Figure 8.4: **Hypothetical density Curvature plot** $(\phi_1(x_\alpha), \phi_0(x_\alpha))$ **of a bimodal density**

Also, the signed change of this curvature at point α is given by

$$\frac{\phi_0''(\alpha)\phi_1'(\alpha) - \phi_1''(\alpha)\phi_0'(\alpha)}{c} \quad \text{where } c \text{ is a positive function of } \alpha\,. \tag{8.26}$$

Hence, we can use the numerator of equation (8.26) to determine the sign changes. For example, we can use the curvature function $K(\alpha)$ defined by

$$K(\alpha) = \frac{\phi_0''(\alpha)\phi_1'(\alpha)}{\phi_1(\alpha)\phi_0(\alpha)} - \frac{\phi_1''(\alpha)\phi_0'(\alpha)}{\phi_1(\alpha)\phi_0(\alpha)} \tag{8.27}$$

to determine whether the distribution is multimodal or unimodal. If the function $K(\alpha)$ changes its sign in the range $\alpha \in [0,1]$, then the density can be multimodal, depending on the value of π. In fact, if the density is bimodal $K(\alpha)$ will have exactly two sign change $(+,-,+)$ and the two potential modes will lie on either side of the two points of sign change. In particular, we can show that in the equal variance case $K(\alpha)$ is a quadratic equation and for the proportional variance case ($\Sigma_1 = \sigma_1^2 V, \Sigma_0 = \sigma_0^2 V$), $K(\alpha)$ is a cubic in α. These polynomials agree with the ones used to determine univariate modality along the axis of maximal separation.

It turns out that the curvature plots are a bit hard to use to determine the range of π where multi modality occurs, so we develop a second plotting method. If α is a critical value

it satisfies $\pi\phi_1' + \bar{\pi}\phi_0' = 0$. Now set

$$\frac{\phi_0'(\alpha)}{-\phi_1'(\alpha)} = \gamma(\alpha). \tag{8.28}$$

Note, that if α is a critical value, $\gamma(\alpha) = \frac{\pi}{\bar{\pi}}$. Further, note that

$$\gamma'(\alpha) = -\frac{\phi_0''(\alpha)\phi_1'(\alpha) - \phi_1''(\alpha)\phi_0'(\alpha)}{(\phi_1'(\alpha))^2}, \tag{8.29}$$

which has the opposite sign to the curvature function $K(\alpha)$, given in equation. 10.1. The zeroes in the curvature will be critical values of $\gamma(\alpha)$.

We can further describe $\gamma(\alpha)$ on $\alpha \in [0,1]$, by noticing that that $\phi_1 \uparrow$ on range $\alpha \in [0,1)$ beacuse $\phi_1' > 0$; also $\phi_1' = 0$ at $\alpha = 1$. Simultaneously, $\phi_0 \downarrow$ on range $\alpha \in [0,1)$ as $\phi_0' < 0$ and $\phi_0' = 0$ at $\alpha = 0$. Figure 8.5 shows an example of $\gamma(\alpha)$ when $f_{\mathbf{X}}(\mathbf{x})$ has a bimodal density.

Given the curve $\gamma(\alpha)$, and densities f_1 and f_0, one can determine the range of values of $\frac{\pi}{\bar{\pi}}$ that give the multimodality. For any given $\frac{\pi}{\bar{\pi}} = y_0$ draw a horizontal line $y = y_0$. If it crosses $\gamma(\alpha)$ once, there is a single critical value for the density corresponding to single mode. If it crosses the three times, then there exists three critical points, corresponding to mode, antimode, and mode, respectively.

Also from equation (8.28) we have $\pi = \dfrac{\phi_0'(\alpha)}{\phi_0'(\alpha) - \phi_1'(\alpha)}$ at any critical value. A direct idea of the range of π for which the density is multiimodal can be obtained by plotting $\gamma_1(\alpha) = \dfrac{\phi_0'(\alpha)}{\phi_0'(\alpha) - \phi_1'(\alpha)}$. Note that $\gamma_1(\alpha)$ also has the change of signs at the same place as we have

$$\begin{aligned} \gamma_1'(\alpha) &= \frac{\phi_0''(\alpha)\left(\phi_0'(\alpha) - \phi_1'(\alpha)\right) - \left(\phi_0'' - \phi_1''(\alpha)\right)\phi_0'(\alpha)}{\left(\phi_0'(\alpha) - \phi_1'(\alpha)\right)^2} \\ &= -\frac{\phi_0''(\alpha)\phi_1'(\alpha) - \phi_1''(\alpha)\phi_0'(\alpha)}{\left(\phi_0'(\alpha) - \phi_1'(\alpha)\right)^2}. \end{aligned} \tag{8.30}$$

So we can observe the change of curvature and the range of π solutions from both $\gamma(\alpha)$ and $\gamma_1(\alpha)$. (See Figure 8.6). But, for $\gamma_1(\alpha)$ we can directly determine the range of π for which the

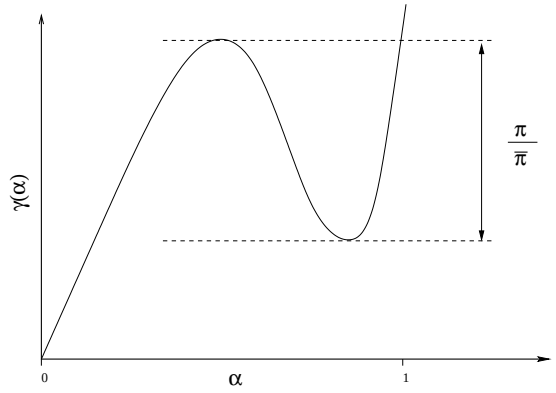

Figure 8.5: **Hypothetical curvature plot ($\gamma(\alpha)$) of a bimodal density**

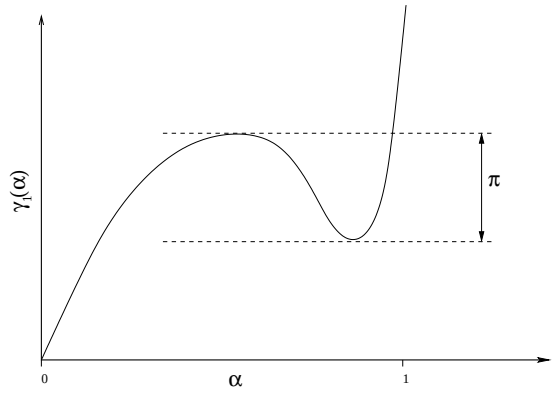

Figure 8.6: **Hypothetical curvature plot ($\gamma_1(\alpha)$) of a bimodal density**

distribution is multimodal. Another advantage of using $\gamma_1(\alpha)$ is the plots are visually more comparable over different densities because $\gamma_1(\alpha) \in [0,1]$.

8.3 Example: Bimodality of bivariate normals

Now, let us see how well the three types of plots $\phi_1(\alpha)$ *vs* $\phi_0(\alpha), \gamma(\alpha)$ and $\gamma_1(\alpha)$. capture the modality structure for some specific bivariate mixtures of normals. In this section we investigates the plots for detection of modality for various selection of mean and variance structure of an arbitrary mixture two component bivariate normals.

We will first give four sets of parameters which will define four different mixtures.

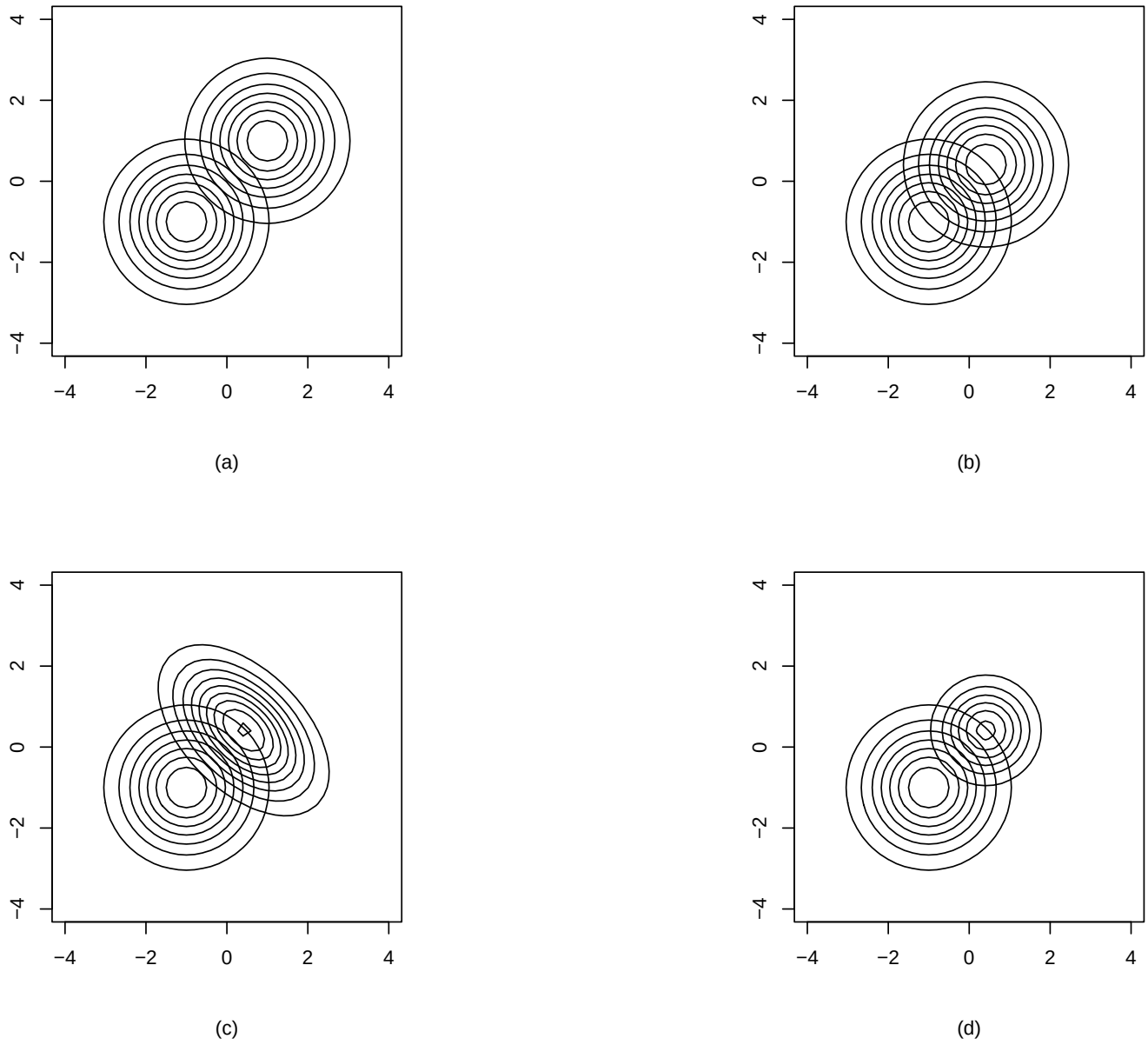

Figure 8.7: **Contour plot of mixtures of bivariate normals for four different sets of parameters described in** (8.31)

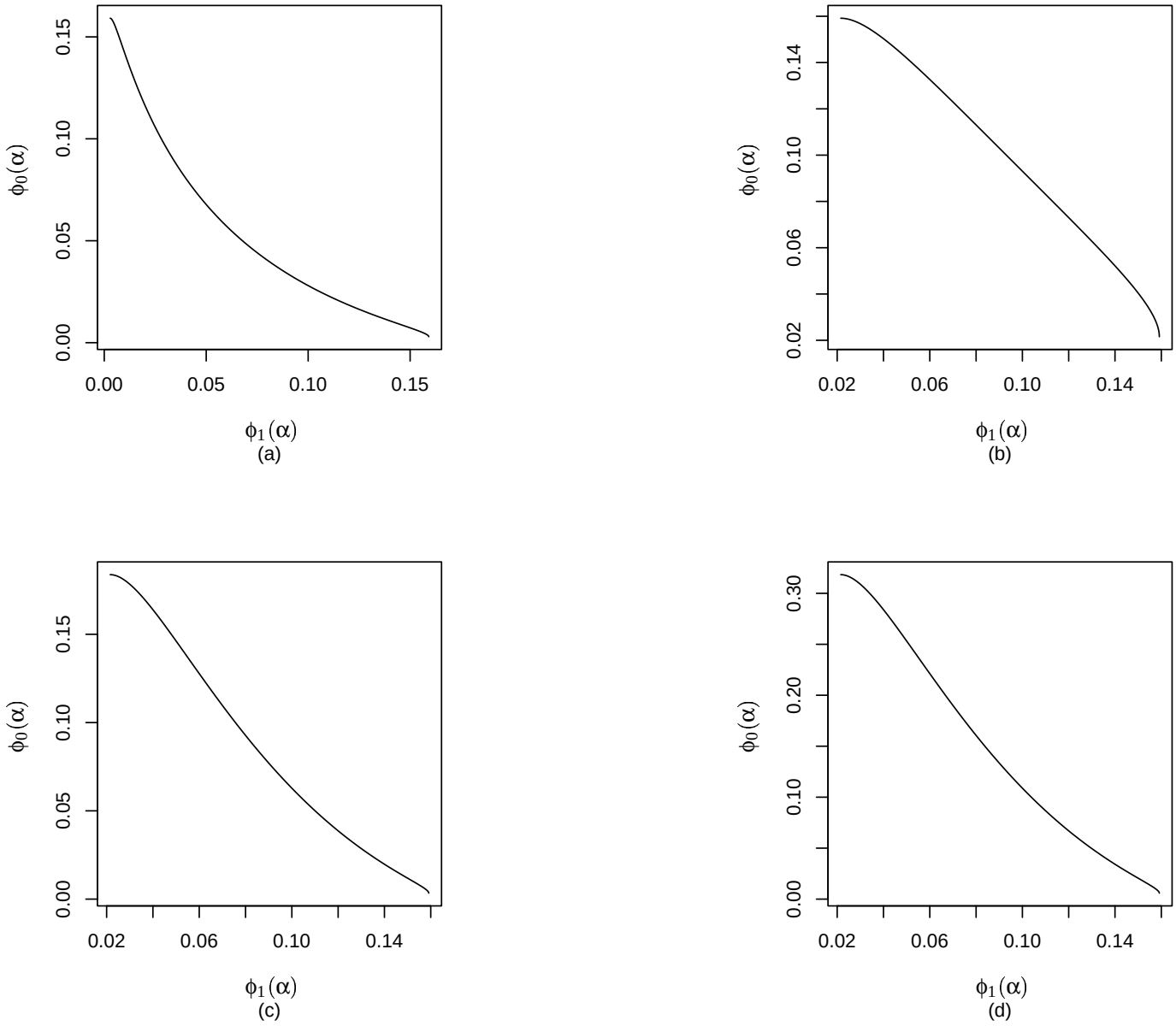

Figure 8.8: Plot of $\phi_0(\alpha)$ *vs* $\phi_1(\alpha)$ for the four sets of mixtures

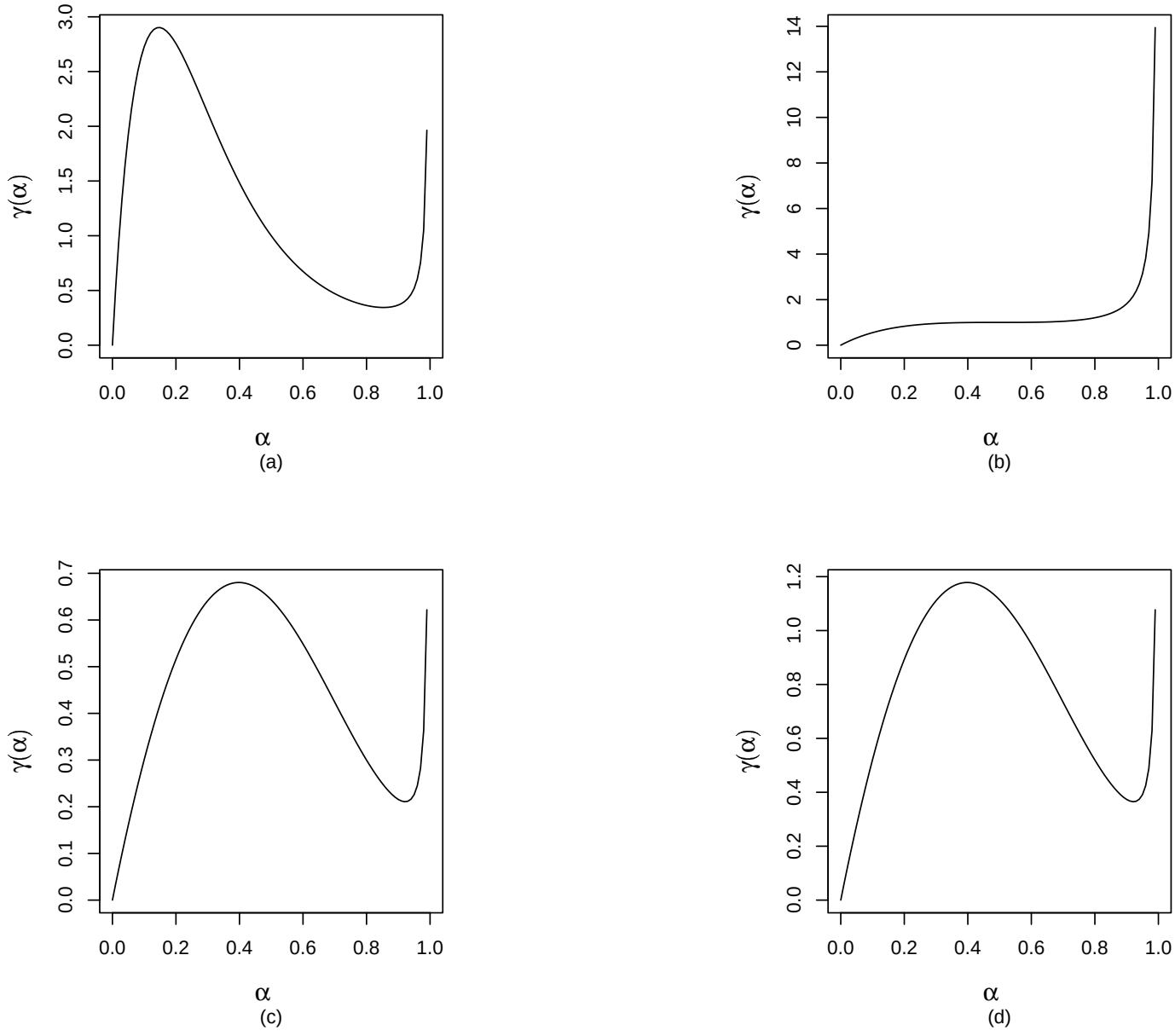

Figure 8.9: Plot of $\gamma(\alpha)$ for the four sets of mixtures

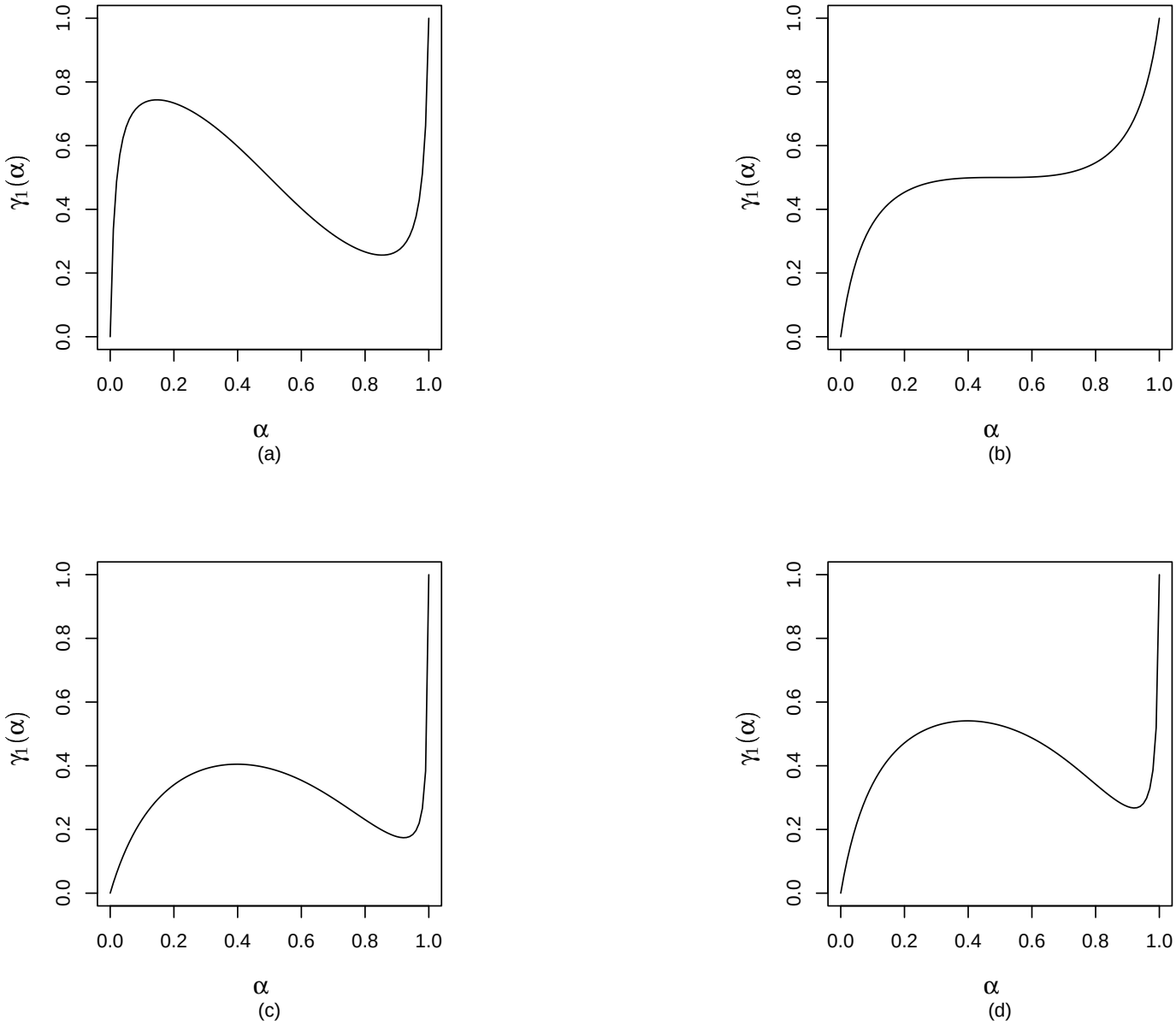

Figure 8.10: **Plot of $\gamma_1(\alpha)$ for the four sets of mixtures**

The contour plots of the component densities of these mixture densities will be displayed in Figure 8.7. Figures 8.8, 8.9 and 8.10 are the $\phi_1(\alpha)$ *vs* $\phi_1(\alpha), \gamma(\alpha)$ and $\gamma_1(\alpha)$ plots, in the order mentioned, of these four bivariate mixtures. Parameters of four sets of mixtures are given below.

(a) $\mu_1 = \begin{pmatrix} -1 \\ -1 \end{pmatrix}, \Sigma_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \mu_0 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \Sigma_0 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$

(b) $\mu_1 = \begin{pmatrix} -1 \\ -1 \end{pmatrix}, \Sigma_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \mu_0 = \begin{pmatrix} \sqrt{2}-1 \\ \sqrt{2}-1 \end{pmatrix}, \Sigma_0 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$

(c) $\mu_1 = \begin{pmatrix} -1 \\ -1 \end{pmatrix}, \Sigma_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \mu_0 = \begin{pmatrix} \sqrt{2}-1 \\ \sqrt{2}-1 \end{pmatrix}, \Sigma_0 = \begin{pmatrix} 1 & -0.5 \\ -0.5 & 1 \end{pmatrix}$

(d) $\mu_1 = \begin{pmatrix} -1 \\ -1 \end{pmatrix}, \Sigma_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \mu_0 = \begin{pmatrix} \sqrt{2}-1 \\ \sqrt{2}-1 \end{pmatrix}, \Sigma_0 = \begin{pmatrix} 0.5 & 0 \\ 0 & 0.5 \end{pmatrix}$

Mixture (a) is the mixture of two bivariate normals with a common variance and as $(\mu_1 - \mu_0)'\Sigma^{-1}(\mu_1 - \mu_0) > 4$. Using the condition of Theorem 8.1 we can infer that there exist some mixing proportion π for which the density of the mixture is bimodal. In fact, the existence of bimodality is reflected by Figures 8.8(a) 8.9(a) and 8.10(a). Moreover, from Figure 8.10(a) we can see that the range of π for which the density is bimodal is roughly $(.3, .7)$ and otherwise it is unimodal.

For the mixture given by the parameters in (b) we have common variance $\Sigma = I$ and $(\mu_1 - \mu_0)'\Sigma^{-1}(\mu_1 - \mu_0) = 4$. Thus, the density cannot be bimodal for any value of π. The plots of Figure 8.8(b) shows no change of curvature. Also Figures 8.9(b) and 8.9(b) shows that the functions γ and γ_1, respectively, are non-decreasing. This implies that the there is no value of π for which we will have a bimodal density.

Now we will move on to (c) and (d) with unequal variances, in which case we cannot

infer modality using Theorem 8.1. In fact, mixture (c) has the same mean as mixture (b) but a different variance structure. Figure 8.8(c) displays a change of curvature implying the presence of bimodality. Also, Figures 8.9(c) and 8.9(c) shows that mixtures given by the parameters in (c) will have a bimodal density for values of π roughly in the range $(.2, .4)$. In mixture (d) the variances are proportional. Here Figures 8.8(d), 8.9(d) and 8.10(d) detect the possibly bimodal structure of the density and show that the range is about $(.25, .55)$.

Among the three plots, we believe that the γ_1 plots are the most informative and easy to use. The $(\phi_1(\alpha), \phi_1(\alpha))$-plot detects the bimodal structure through the change of curvature, but it is not easily detectable from the plots. As we can only plot the curve for a finite set of values, we may not be able to detect the minute change in curvature for a certain level of precision. Thus for all practical purposes we would depend on either $\gamma(\alpha)$ or $\gamma_1(\alpha)$ plots. As mentioned earlier $\gamma_1(\alpha)$ is better for two reasons:

- The $\gamma_1(\alpha)$ plots are comparable to one another as they always lie in the same scale (γ_1 values are always between 0 and 1).
- We can directly get the range of π for which the distribution is bimodal.

8.4 Results

We applied the modality criterion to detect clusters in the four datasets we analyzed in the previous chapters. After getting the parameter estimates for a *g* component fit, we compare the components pairwise to see whether they display bimodality. If the mixture of two components, with weights proportional to fitted weights, is unimodal we will say the the components are *linked*. Intuitively, we might think of the linkage as defining a single modal cluster. One could go further and consider the set of linkages in a modal fit as describing its overall structure. For equal variance estimates, we used the analytical formula given in Section 8.1. For the unequal variance case we used the graphical plot $\gamma_1(\alpha)$ described in

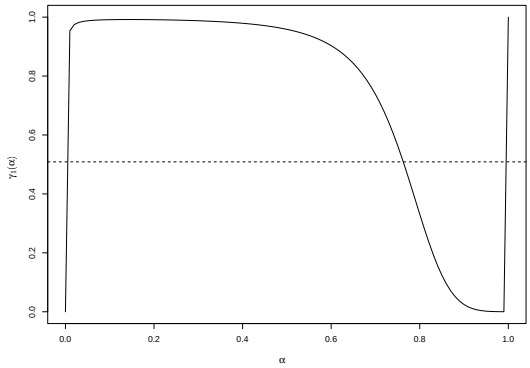

Figure 8.11: **Plot of $\gamma_1(\alpha)$ for the mixture of components 2 and 3 in the 3 components fit of the Iris data**

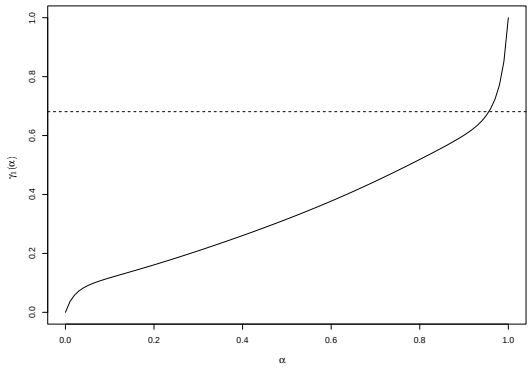

Figure 8.12: **Plot of $\gamma_1(\alpha)$ for the mixture of components 1 and 3 in the 5 components fit of the Iris data**

Section 8.2. Here we present the results for the Iris data set. Fits with 1 through 6 components were tried. Each pairwise comparisons for the 2,3 and 4 component fits indicated no linkage. But when we fit 5 components, components 1 and 3 were linked, suggesting that we might merge components 1 and 3, and so we will have 4 distinct modal clusters in the dataset.

For the unequal variance estimates we appeal to the pairwise $\gamma_1(\alpha)$ plot. For the 2 components in the pair, if the horizontal line at the mixing proportion π (denoted by the - - - - line Figure 8.12 and 8.11) cuts the $\gamma_1(\alpha)$ curve three times in its range [0,1] then the two components have distinct modes. Again comparing different pairs we find that all pairs up till

the 4 component fit displays two distinct modes. See for one example, Figure 8.12. But for the 5 component fit, components 1 and 3 were linked. See Figure 8.11. Thus going by the rule, "distinct modes correspond to distinct clusters", we would suggest having 4 modal clusters.

8.5 Conclusion

Determining conditions for bimodality in the multivariate normal mixture problem is still a challenging issue. We derived an analytical condition for the equal variance case. For the unequal variance case, we have created plots that display a variety of information and can easily detect bimodality. These methods can be used in clustering by doing pairwise comparisons of the different components in a multi component fit and merging components which are linked. In this way we would arrive at a more parsimonious description, as well as a greater understanding of how "clustered" the clusters actually are.

This leaves several open questions. First, is it possible to prove from these results that there exists at most two modes in the two component mixture? This will involve a detailed analysis of the curvature function $K(\alpha)$. Secondly, is it possible to arrive at analytical solutions to zeroes in the curvature function which in turn give explicit calculations for the range of π? Of this we are doubtful. However, this seems not so important as we believe that one can produce an elementary numerical algorithm which would quickly and reliably find these points.

Next, pairwise linkages define a graphical relationship between components. For example, for the three components we could have component 1 linked to 2 and 2 linked to 3, but not 1 to 3. this suggests a ridge-line structure, with some shape. However, if 1, 2, 3 are all pairwise linked, the structure would see more like a single mass. Can we use graph theory in a fruitful way to describe structure?

Finally at this time we do not know the relationship between pairwise modality and

overall modality. As a clustering tool, however, the pairwise method has distinct advantages. For example, if we find that a three components has two modes, it is not so obvious how to identify the three components with two clusters.

Chapter 9

Application: Analysis of Gene Expression Data

In this chapter we will apply our model selection methodology to one of the most challenging areas of statistical and scientific research, gene-expression data. Despite the information made available by ongoing research on sequencing the human genome (structural genomics), we lack a full understanding of how our genes are properly turned on or off so as to maintain a healthy body (functional genomics). Many diseases including cancer, genetic diseases and other infectious diseases are a direct consequence of mis-expression by the genes. To examine how gene regulatory proteins assemble a gene and regulate its expression we need to know the expression levels of thousands of genes at the same conditions. These rich data are made available by micro-array experiments that are performed over a set of conditions. But the dimension and complexity of raw gene expression data obtained by oligonucleotide chips, spotted arrays, or whatever technology is used, create challenging data analysis and data management problems. In a limited way these challenges can be met by existing software systems and analysis methods in the hands of end users.

Microarray data can be analyzed using several approaches. Clustering methods (i.e. unsupervised learning) are widely used and have the ability to uncover coordinated expression patterns from a collection of microarrays (e.g., Eisen et al., 1998, Getz et al., 2000, Tibshirani et al., 2002, 2001, Dudoit et al., 2002, Kerr et al., 2002). The analysis of gene-expression data using clustering techniques has an important role to play in the discovery, validation and understanding of various classes and subclasses of disease. Because of the high levels of noise

inherent in this technology, as well as in the cell itself, it is desirable to carry out the analysis of microarray data within a statistical framework. Widely used methods of clustering such as k-means clustering, hierarchical clustering and other agglomerative and divisive algorithms do not utilize the inherent statistical structure of the data. These methods can be pulled under the broad category of model-free methods. We will address the problem of clustering using a model based approach.

Model based methods provide a stable and useful way of clustering. It also gives a clear definition that a cluster is a subpopulation with a certain distribution, and several statistical methods can be applied to address the problem of choosing of number of clusters in an objective way. Finite mixtures of distributions provide a flexible as well as rigorous approach to modeling various random phenomenon. In this section we will use the finite mixture of normals to cluster genes on the basis of experimental conditions (treatment) and cluster conditions on the basis of genes. The later is a non-standard problem in parametric (model-based) clustering because the dimension of the feature space (the number of genes) is typically much greater than the number of conditions ($2-100$ conditions versus 10^3-10^4 genes). In this dissertation we will mostly study the first type, but will discuss the second type of clustering and show how mixture of multivariate normals can solve the issue of dimensionality.

9.1 Description of the dataset

In this section a short description of the data will be provided. For more details on the data please see Chitikila et al. (2002). The experiments were done in the Pugh Lab, Pennsylvania State University (`http://www.bmb.psu.edu/faculty/pugh/lab/lab.html`). As the research is an ongoing one, new experiments on various other conditions are in progress. Here I provide a very short and non-technical description of the experimental data.

9.1.1 Notations and abbreviations

First, let us introduce some notations and abbreviations that we will use through out this chapter

- TATA Box: A conserved AT-rich septa-mer found between 250 bp to 50 bp before the start point of each eukaryotic RNA polymerase II transcript unit; it may be involved in positioning the enzyme for correct initiation.
- TBP: TATA Binding Protein: The proteins which bind to the TATA-Box and initiate the transcription. This is required for the expression of nearly all genes.
- TAND: A domain(region) of TAFI protein.
- ΔTAND : A strain of yeast protein TAFI, removing the TAND region.

9.1.2 The Biology

The TATA-binding protein binds to the TATA box. Though actually a long string, the TBP is intertwined in such a way that it forms a concave surface and a convex surface. The concave surface is supposed to attach with the TATA box, whereas the convex surface may attach to some other protein. Thus, a mutation in the concave surface may inhibit the effect of the TBP, whereas a mutation the convex surface may affect the interaction of of TBP and the other protein and thus have an impact on the transcription. In our experimental example the proteins in the concave region (positions 71, 161,69) are altered, but we also have an experiment with a mutation in the 182^{th} position, the location of interaction with another protein NC_2. We reproduce a figure (Figure 9.1.2) from Chitikila et al. (2002) to show the relative position of the mutation and the protein binding positions.

The interplay of the TAFI and the TBP protein is still not completely understood. What we do know is that the TBP binds with the TAND region/domain of TAF1. Some models proposed in Chitikila et al. (2002) are provided in Figure 9.1.2.

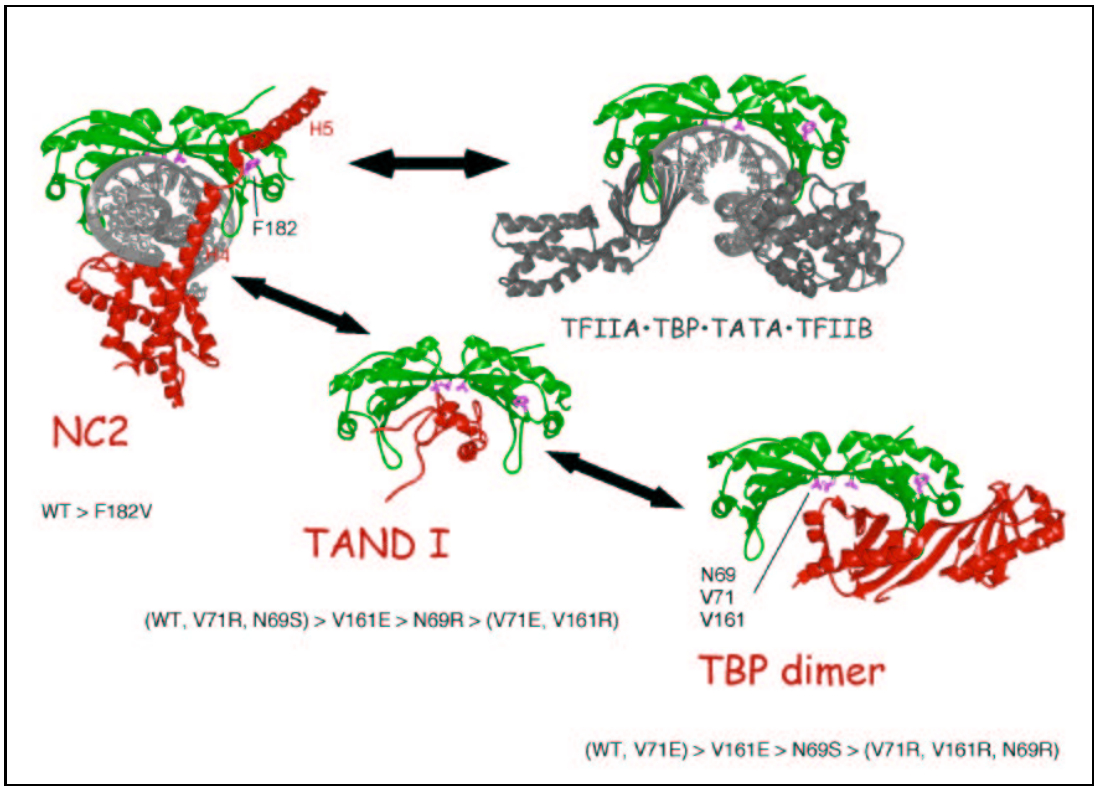

Figure 9.1: (Reproduced from Chitikila et al. (2002)). Structures of TBP interaction with NC_2 and TAND 1 domain of TAFI.

9.1.3 The experiment

An experiment was conducted to address whether TBP and TBP-TAND interactions represents distinct mechanisms. In the experiment there are two yeast strain WT and ΔTAND , where WT is the wild type and ΔTAND is a strain which has a mutation in a certain protein of interest (namely TAF145). Codes V71R, V161R, N69R, N69S, V161E, and V71E are for the mutant versions obtained by changing particular base pairs in the protein; e.g. for V161E one changes protein V to protein E in the 161^{st} position. In addition to the mutant versions, we have the WT and the null experimental condition. Observing the expressions of genes under each version of TBP mutants for each of the yeast strain (WT or ΔTAND), we have a total of 16 experimental conditions. Moreover under WT for the null condition and V161R

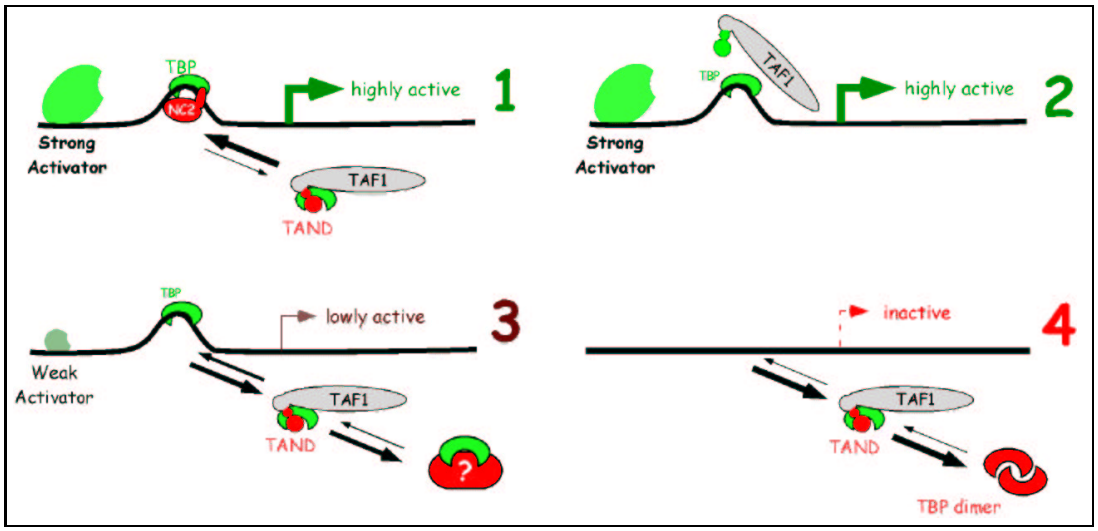

Figure 9.2: (Reproduced from Chitikila et al. (2002)). Proposed models for the interplay of TBP effectors in regulating the genes identified in the 4 clusters identified by k-means.

we have the data from two researchers (coded by null-K, null-L and V161R-K, V161R-L). Similarly under ΔTAND for the V161R mutant, we have data from those two researchers. We also have a negative control for the WT strain from the mutant version F182V. As the 182^{th} position does not interact with the ΔTAND domain, the F182V mutation should not affect the TAF1-TBP interaction. The eventual goal is to have an expanded study of more interactions with different mutant versions of the protein.

To summarize we have 20 conditions. 11 under WT and 9 under ΔTAND .

- **WT:** WT, null-K, null-L,V161R-K, V161R-L, V71R, N69R, N69S, V161E, V71E, F182V.
- **ΔTAND :** WT, null, V161R-K, V161R-L, V71R, N69R, N69S, V161E, V71E.

In our analysis, for each condition and each gene we have one value the log (base 2) ratio of the test signal to the reference signal. The log base 2 is used in the gene expression literature

as it gives an easy interpretation to the data in terms as fold changes in expression level. To neutralize the dye effect, dyes were switched and the same experiment was repeated. Thus, actually there two replicates for each set of experimental conditions. But we only consider their average log-ratio in our analysis. The final data consists of 6226 rows (6188 genes covering 99.4% of the *S. cerevisiae* strain S288C and 38 control) over 20. During filtering, which will be discussed in the following subsection, some spots were flagged which gave rise to missing data.

9.1.4 Preprocessing

Preprocessing of the data is described in the details in Chitikila et al. (2002)(See Section on Statistical Filtering). Here we provide a short description of the statistical filtering. The $\log_2$ ratios of the gene expression from a single experiment (test vs. reference) were normalized by mode centering. Finally the mode centered data were filtered in the following way. The significant fold changes were selected if it met all the below mentioned criterion. [1]

1. Raw gene expression intensities were greater than one standard deviation above local background in both the test and reference samples in both replicates.
2. Ratios changed in the same direction in each replicate.
3. Ratios in each replicate were greater than two standard deviations above 1.0
4. p-values of the arithmetic average of the $\log_2$ ratios were < 0.005.
5. Fold changes in gene expression level were > 1.5.

After applying the filters several spots were flagged; so, now the array of 6226 rows and 20 columns have several missing values.

[1] Reproduced from Chitikila et al. (2002)

9.2 Proposed Analysis

The analysis of this gene-expression data can be done in a number of ways. Here we mainly deal with the number of clusters or number of distinct patterns of expression. Formulating the problem in terms of multivariate mixture of normals we denote the genes by $Y_1, Y_2, \ldots, Y_n$ where each Y_i is a p-dimensional vector. Thus, in this we initial have p=20 and n= 6226. Several issues were rigorously checked before applying the model selection tools . Though, the number of genes decreased when we included on those genes whose expression changed significantly, still the number of genes were quite high. The other issue was to deal with the large number of variables. And the last but not the least is the issue of tackling missing values. In our model fitting and model selection process we also keep the option of incorporating known structures of the data.

The most attractive feature of our model selection tools is that it can work very efficiently with high dimensional data. Most other, clustering methods including K-means, are based on the calculation of distance between two data points. These dissimilarity (distance) matrix can be thought as a multidimensional scaling; intuitively, it loses most of the information present in the full 20 dimensions in the process of conversion to a matrix of distances. Moreover, the choice of the distance measure is always a critical issue for the construction of the dissimilarity matrix. In our analysis, we can effectively use all the variables to select the number of components in the multivariate mixture model. Just the same, we can apply other variable selection tools to reduce the dimensionality if we feel some variables are statistically redundant.

Moreover, missing data, or the flagged spots, can be modeled and included in the EM algorithm to find a best solution by imputing the missing value. This will help us include the genes with some missing values which would have been discarded otherwise.

In the next section we present the results obtained from analyzing the gene expression data.

9.3 Results

After filtering we chose genes with less than 2 missing values. This reduced the number of genes with significant changes in expression level to 2358. Figure 9.3 gives a k-means clustering of the 2358 genes over 4 clusters. How the 4 clusters were obtained are explained in Chitikila et al. (2002). Before explaining the figure let us give a general description of Figures 9.3 and 9.3. The rows are genes and the columns are the experimental conditions. The green indicates a decreased expression, red indicates a increased expression and black represents no change. The missing values are denoted by grey. The plots were drawn using the Treeview (Eisen et al., 1998) available at `http://rana.lbl.gov/EisenSoftware.htm`.

Here we present result for a risk analysis of the gene expression data. Again, the risk analysis was done with 2358 genes over 20 conditions. We choose 200 observations for deletion (see 6) for calculating the risk. Moreover, as we have dimension 20, we chose $h = 2$ (see Table 9.1).

Table 9.1: **Table Pseudo degrees of freedom for the gene expression data with 20 dimension**

h	$\widehat{pDOF}$
0.1	7.885e+06
0.2	7.885e+06
0.5	3.246e+06
1	4659
2	95.91
4	16.81

From the risk analysis (Figure 9.4) at this level of smoothness we see a clear indication that there are 5 components. Also we see a slight dip of the risk when we have 10 components. A logical explanation of the 10 components may be that the 10 components are

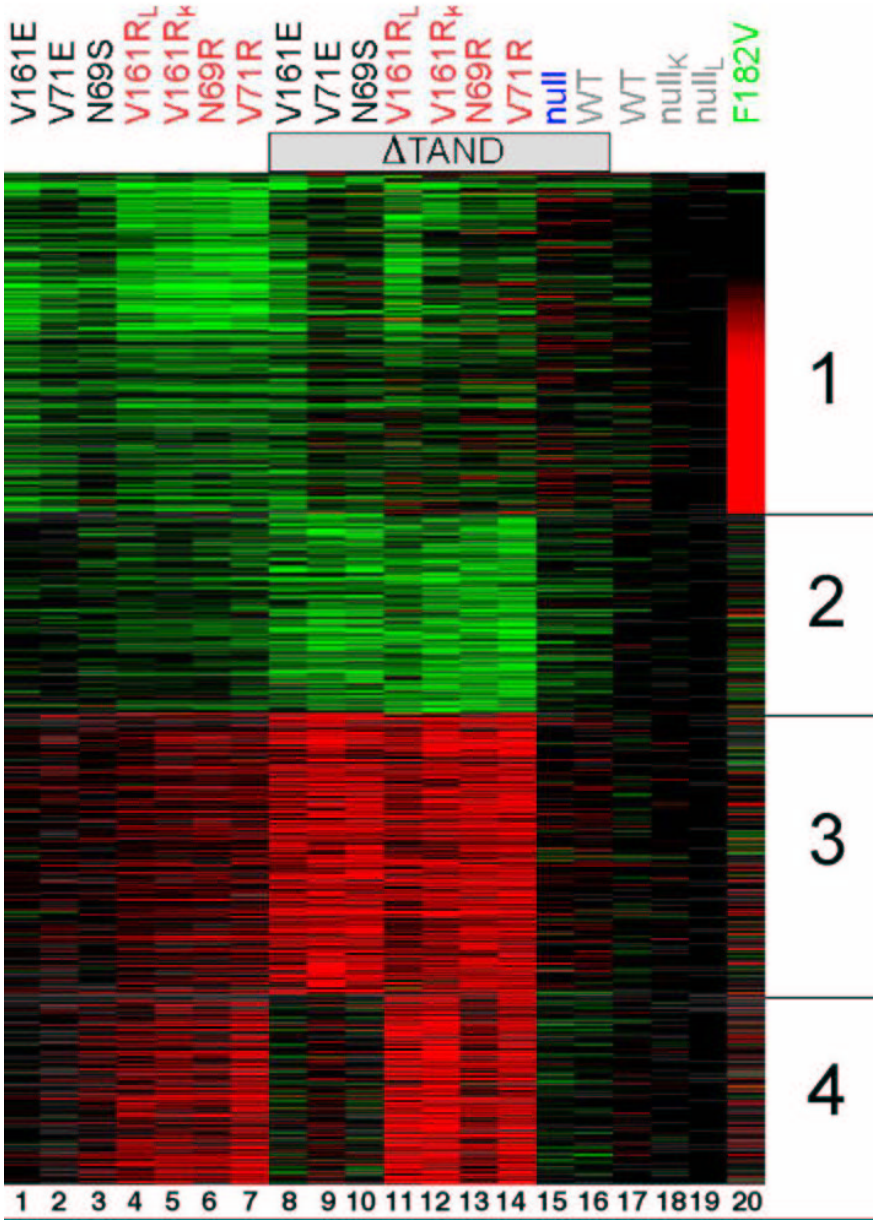

Figure 9.3: (Reproduced from Chitikila et al. (2002)). K-means cluster of gene expression data with 4 clusters

finer classification of the 5 main components. Our, analysis over a large range of h gave us 5 clusters. So our overall conclusion would be that the data set has 5 clusters.

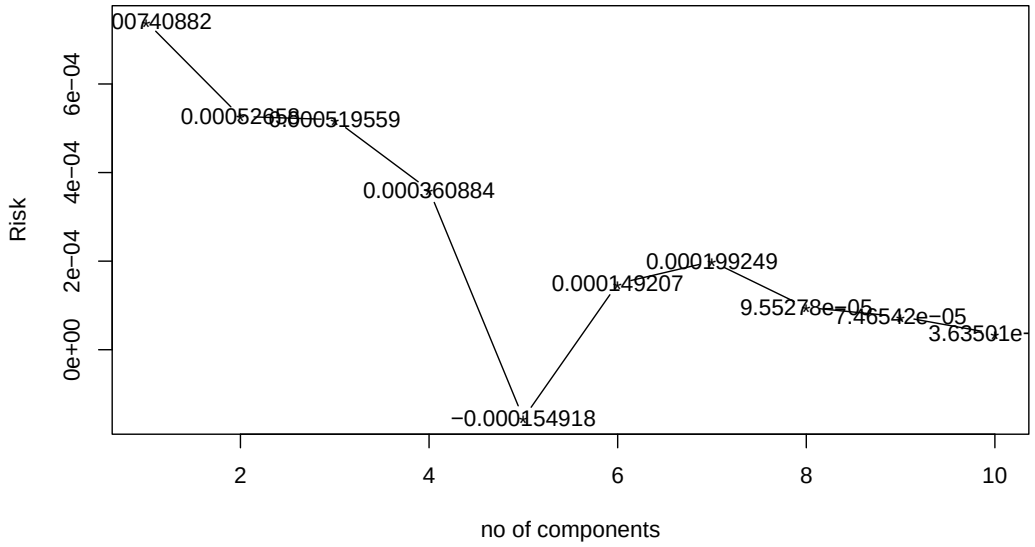

Figure 9.4: **Risk analysis of the gene expression data with h=.7**

The multivariate normals were fitted by an EM algorithm. After getting the posterior probabilities for the 5 clusters we did a hard clustering by assigning the gene to the class with maximum posterior probabilities. The 5 clusters are given in Figure 9.3.

We found that if we ignore the variation in column F182V, then 4 components may be reasonable. Several other analyses are in progress. Some analyses are being done with fewer variables or linear combinations of variables.

Another natural outcome of applying the multivariate mixture model to fit the gene expression is that it also provides a mean pattern μ_i for the i^{th} cluster. Thus comparing the difference between clusters will be more objective. How tight the clusters are can be determined from their estimated variances. We can determine overlap by examining the modality of neighbors. The estimated $\pi_i's$ will provide information about the proportion of gene in each

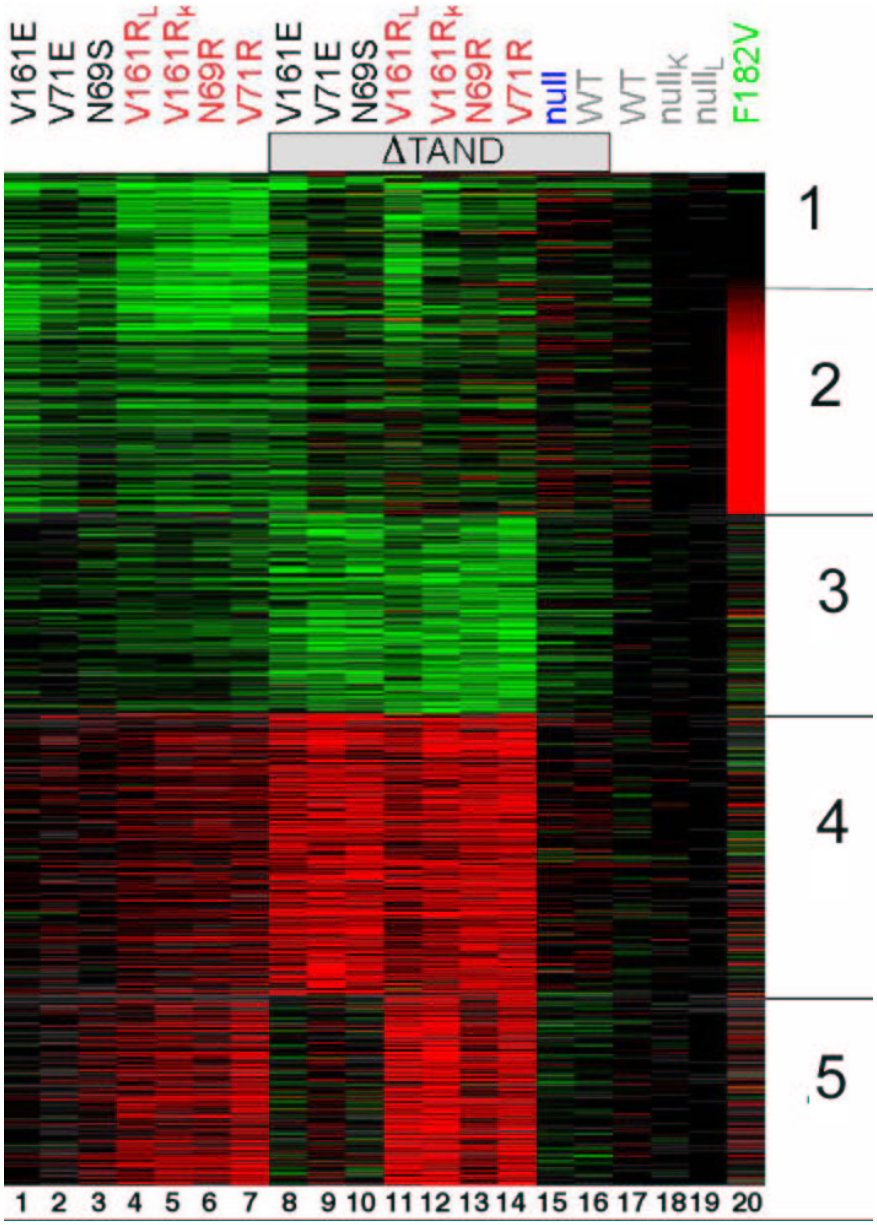

Figure 9.5: Cluster of gene expression data with 5 clusters from the risk analysis

cluster. This may help us delete small clusters from further analysis.

A related biological question would be to assess whether one cluster is the mirror image of another cluster. We may propose an ad-hoc testing criterion that cluster i would be called the mirror image of cluster j another if $\mu_i = -\mu_j$. Statistical tests depending on the parameter estimates of the multivariate mixture models can be developed easily to check the above hypothesis.

9.4 Biological significance

The biological significance of the above clustering is still under study. But, from the clustering it is very clear that there is interaction between the TBP and the TAND domain of the TAF1 protein. In almost all the clusters it can be observed that the same TBP mutants have different expression levels in the WT and ΔTAND strains.

Especially, observing the set of genes in clusters 1,3, and 4 in Figure 9.3 it can be inferred that the presence and absence of the TAND region affects the expression level significantly. As expected, the WT and the null column displays no significant changes in expression levels over all the clusters as expected. Also the changes in F182V can be clearly seen to have no relation with the changes in the expression levels of the mutants in WT and ΔTAND .

9.5 Conclusion

Clustering high-dimensional data is a challenging task. Efficient methods designed for clustering univariate data may not be useful for high-dimension. Moreover, as pointed out before, clustering methods based on dissimilarity matrices often ignores the rich data available in high-dimension as it bases the clustering on a reduction of the data.

Our methods of model selection are very efficient for clustering high-dimensional data. It can use all the structural information about the relation of the variables with one

another and produce an efficient model selection method. Moreover, it avoids numerical integration while calculating the distance of the proposed model from the data, thus making computing time manageable. Another attractive feature of these model selection methods based on quadratic distance is the smoothing parameter, *h*. Varying *h* we can observe finer clusters which may be attributed to subclasses of bigger clusters. Also, as mentioned before, missing data, which is a common occurrence in array data, can be handled easily.

All these attractive features, make distance based model selection a highly appropriate clustering method for gene expression data. Further research will be on finding modal clusters and including sequence information in clustering.

Chapter 10

Discussion

10.1 Conclusions

In this dissertation we have introduced generalized quadratic distances and used them as a model selection tool. The model selection tools are general in nature and can be used for a variety of problems. The distance was defined using a positive definite kernel with a smoothing parameter. By varying the parameter we could analyze the model fits at different scales, revealing clusters, superclusters and subclusters. One of the greatest problems of available multivariate model selection tools, is the enormous calculation involved. We bypass this problem by choosing an appropriate kernel and thus getting a closed form for the quadratic distance.

By using the spectral decomposition of positive definite kernels, we derived the null distribution of the distance. We can readily use this for model selection purposes. To summarize, the following are the definite advantages of using quadratic distance for model selection:

- Calculation of the distance does not require multidimensional, integration making these distance a natural choice for model selection tools in the multivariate scenario.
- Varying the tuning parameter h allows us to construct a set of distances, and thus analyze the data at different levels of smoothness.
- Asymptotic distributions of these distances can be easily worked out.

- As a natural outcome for the distance we get the residuals of each data point which can be used for diagnostics.

Choosing an appropriate range of h for the distances was the subject of discussion in Chapter 4. We designed a summary statistic, the pseudo degrees of freedom and used it to find a suitable range of h. One attractive feature of this statistic is that it can be calculated once and for all because it depends only on the data.

In the next two chapters after 4 we used the quadratic distance to design model selection tools. Chapter 5 introduced the concordance coefficient, a scaled measure of discrepancy between two densities. This measure lies between 0 and 1 so as a distance it has an interpretable magnitude. Based on drawing an analogy to the subset selection in regression using the R^2 coefficient, we used the concordance coefficients to select the number of components in the mixture distribution of a proposed model.

In Chapter 6 we used the distance as a loss function and developed a risk based model selection tool. Unlike AIC and BIC, where we have to define the penalty term explicitly on the number of parameters estimated, our risk function inherently incorporates a term which is designed to capture the parameter estimation cost of a model.

In Chapter 3 it was observed that the distance can naturally be decomposed into sum of residuals. Using these residuals with appropriate standardization, in Chapter 7 we developed diagnostic methods for outlier detection.

Next we took up the issue of determining number of modes in a mixture of multivariate normals. This chapter is indirectly related to the distance based methods, though through the modality conditions one can assess the degree of separation of fitted mixture components. If the fitted mixture of two components display a unimodal density then we might merge them into one single “modal cluster”. Using the existing conditions of bimodality in the univariate case for the equal variance case, we devised analytic conditions in the multivariate situations.

For the unequal variance case we designed plots which display modality properties in a clear and easy-to-interpret fashion.

The last chapter deals with the application of our model selection tools to the analysis of gene expression. This is an area of application demanding analysis of high dimensional data with large sample sizes. Most of the usual clustering methods ignore the rich information available from all the variables and perform the clustering only on a summary of the data. Our methods are designed to deal with the high dimension while retaining the low computational complexity that makes the methods practical to use. Moreover, most clustering algorithms are designed for a fixed number of clusters. Methods for comparing the clustering results using different numbers of clusters are not obvious. In contrast, our methods does the clustering for a user-specified range on the number of clusters. Along with the clustering our model selection tools gives a method of comparison among the different sets of clusters so one can choose the best one according to the question asked.

10.2 Future Work

In this dissertation we proposed the distance based model selection and detection of modality in the multivariate mixture situation. A number of issues in both these broad topics need to be investigated further. Here we discuss some of the possible extensions of our methods.

10.2.1 Combining the results from different h

As mentioned before sliding the tuning parameter in the kernel we can analyze the data at a number of smoothing scales. Future work along these lines may be to develop theories to choose the number of subclusters and superclusters along with the number of clusters. One problem that is strongly related to the above is to find out the suitable range of h (discussed in Chapter 4) and then divide into sub ranges, which we can then relate to regions of h giving

subclusters, clusters and superclusters. This should be a very useful approach to discover all the structure in the data, not just analyzing it from one level.

10.2.2 Standardization of residuals and distribution of standardized residuals

The raw residuals, which are the natural outcome of the distance estimates, are very informative, though by themselves they cannot be used as a diagnostic tool. We need to standardize them properly to use them to detect outliers. The proper standardization is not obvious. In this dissertation we proposed a few standardizations; we expect to do future research on the standardization. Moreover, the distribution of the standardized residuals is another important issue. To classify some values of the residuals as outliers we need to use the distribution of the residuals, to find a cutoff value.

10.2.3 Analytical conditions for multi modality in the unequal variance case

In this thesis we have resolved the issue of existence of more than one mode in mixture of multivariate normals with equal component variances, by providing analytical conditions. For the unequal variance case we provided some informative plots. The curvature function $K(\alpha)$, defined by

$$K(\alpha) = \frac{\phi_0''(\alpha)\phi_1'(\alpha)}{\phi_1(\alpha)\phi_0(\alpha)} - \frac{\phi_1''(\alpha)\phi_0'(\alpha)}{\phi_1(\alpha)\phi_0(\alpha)}, \tag{10.1}$$

could be studied further to determine if there are analytical conditions in the unequal variance case, and to confirm that there are at most two modes to a mixture of 2 multivariate normals. Moreover, more work should be done on using modality ideas to create modal clusters.

10.2.4 Analysis of gene expression data

Application of the model selection tools to the gene expression data has several possible extensions. The mean and variance structure could be modeled through known biological information. Missing data could also be modeled using the existing data and other biologi-

cal knowledge. In general, besides clustering the genes, one may ask the question of how to cluster the conditions. This is a non-standard problem in parametric (model-based) clustering because the dimension of the feature space (the number of genes) is typically much greater than the number of conditions ($2-100$ conditions versus 10^3-10^4 genes). Future work on extending our methods to classify conditions is underway. Finally, we would like an appropriate application to try out ideas and methods for combining results from distances based on different *h*, in the context of gene expression data.

APPENDIX

Here we provide the description of the dataset we have analyzed in different chapters. Complete datasets are available at http://www.stat.psu.edu/ ~surajit/mixture/.

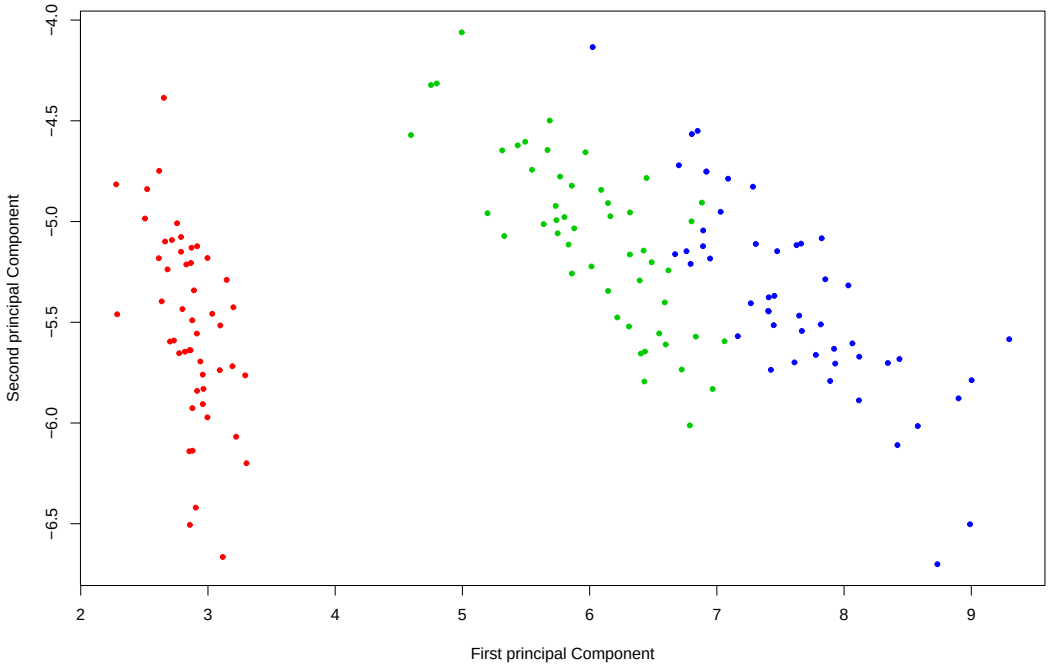

Figure 1: **Projection of Iris Data on the first two principal components**

.1 Iris Data

This is the dataset made famous by Fisher, who used it to illustrate principles of discriminant analysis. Data on 4 variables namely *Petal width, Petal length, Sepal width, and Sepal length* were collected on flowers of 3 species species: Setosa, Verginica, Versicolor. Each species has 50 observations. So the whole dataset consists of 150 observations on 4 dimensions. First we present a plot of the 4 dimensional iris data on the first two principal components (Figure 1), where the three species are coded by different colors.

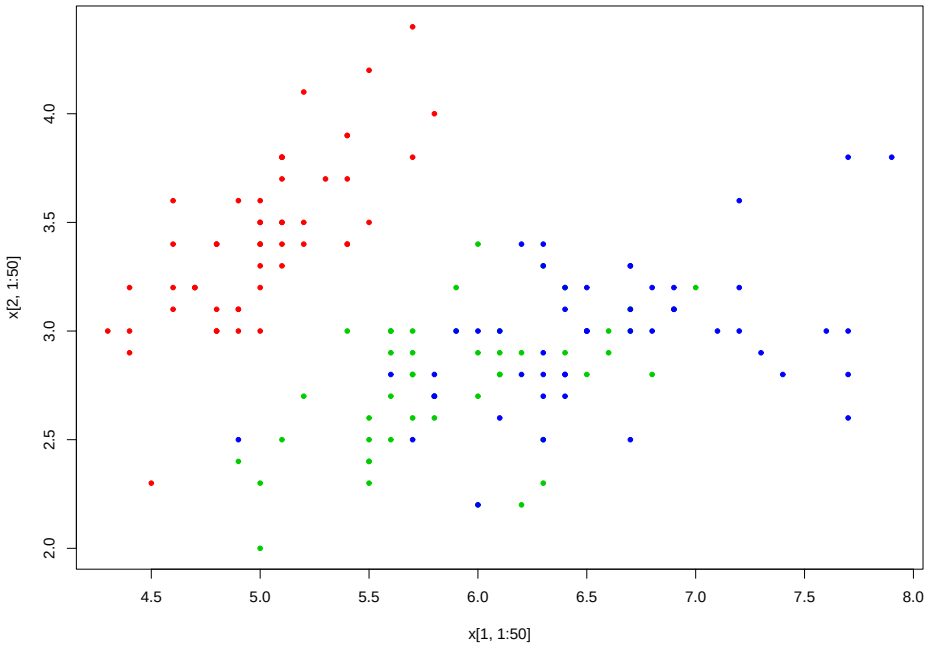

Figure 2: Plot of Iris Data on the variables Sepal length and Sepal width

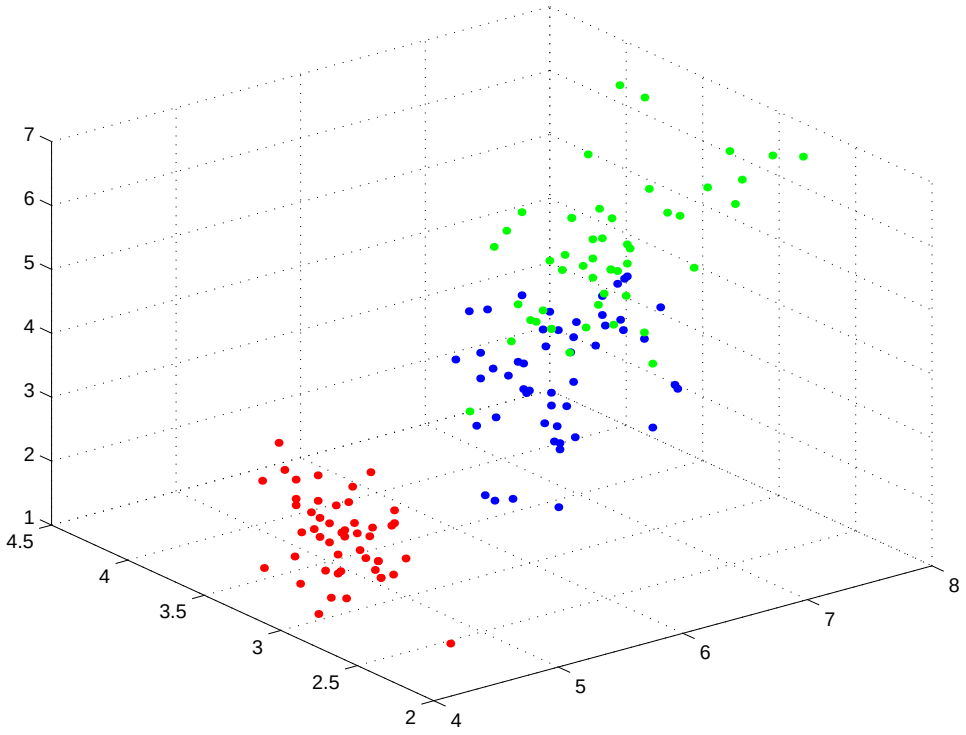

Figure 3: 3D-Plot of Iris Data on the variables Sepal length, Sepal width and Petal length

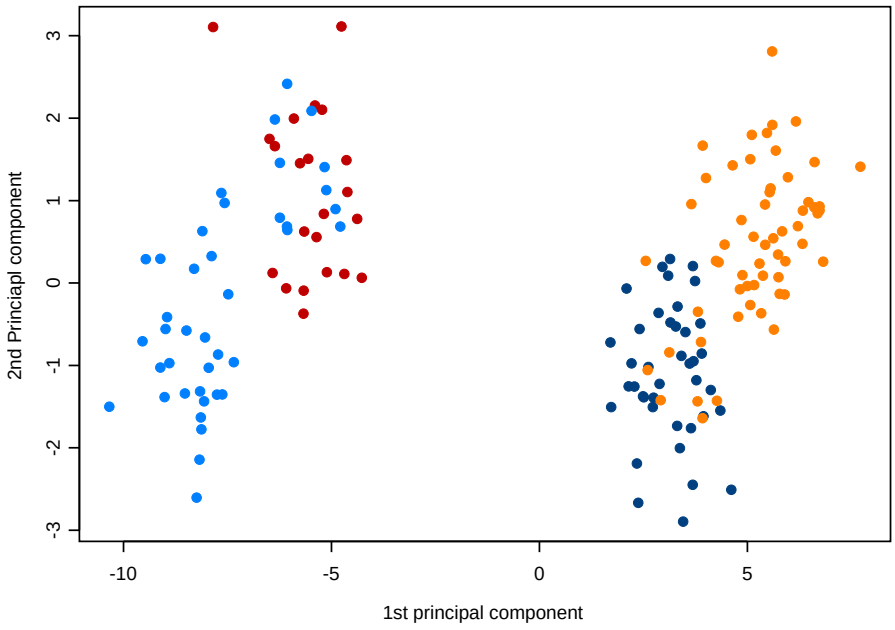

Figure 4: **Projection of simulated dataset 1 on the first two principal components**

.2 Simulated Dataset 1

This data was simulated as a sample of size 160 from a mixture of 4 multivariate (4 dimension) normals with equal mixing proportions and variance. The mean structure of the 4 components were such that, component 1 is very close to component 2 and components 3 is very close to component 4, but components 1 and 2 are far from component 3 and 4. Figure 4 gives a projection of the data cloud in four dimensions, projected on the first two principal components.

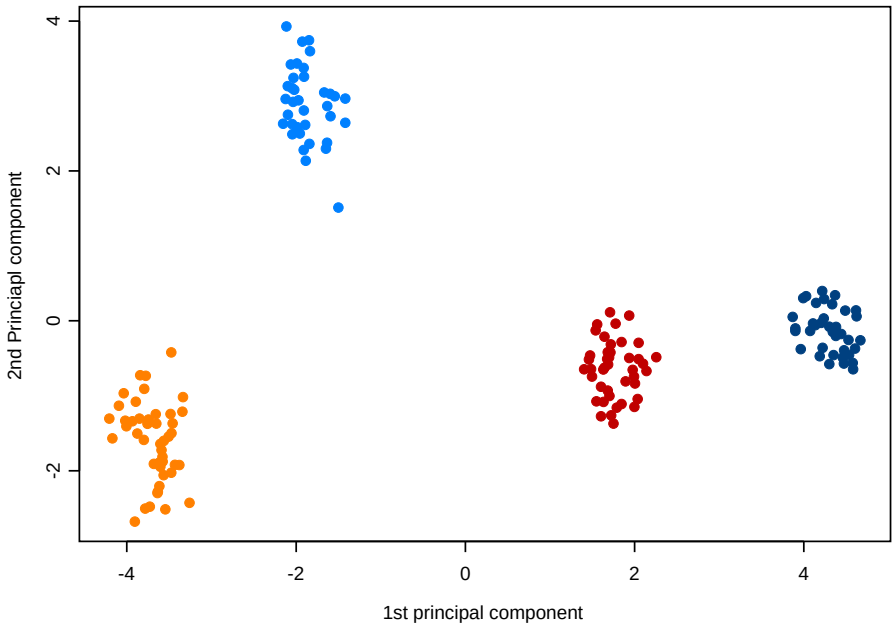

Figure 5: **Projection of simulated dataset 2 on the first two principal components**

.3 Simulated Dataset 2

This dataset was also simulated from a mixture of 4 multivariate (4 dimension) normals with equal mixing proportions and variance. Here too we have 160 samples. But, unlike the simulated dataset 1, the means of the 4 components in this dataset are further apart. Figure 5 gives a projection of the data cloud in four dimensions, projected on the first two principal components.

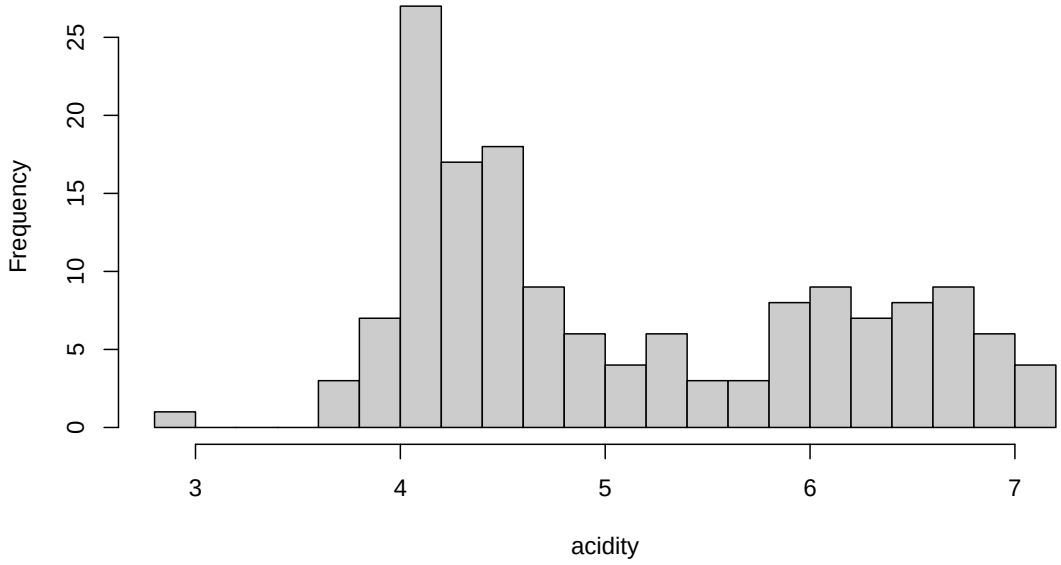

Figure 6: **Histogram of Acidity data**

.4 Acidity Data example

This data set concerns an acidity index measured in a sample of 155 lakes in the Northeastern United States and has been previously analyzed as a mixture of gaussian distributions on the log scale by Crawford et al. (1992). Figure 6 gives the histogram of the acidity dataset.

Bibliography

Aitkin, M., Finch, S., Mendell, N., and Thode, H. (1996). A new test for the presence of a normal mixture distribution based on the posterior Bayes factor. *Statistics and Computing*, 6:121–125. 16

Behboodian, J. (1970). On the modes of a mixture of two normal distributions. *Technometrics*, 12:131–139. 106, 112, 113

Böhning, D., Schlattmann, P., and Lindsay, B. (1992). Computer-assisted analysis of mixtures (C.A.MAN): Statistical algorithms. *Biometrics*, 48:283–303. 17

Cassie, R. (1954). Some uses of probability paper in the analysis of size frequency distribution. *Australian Journal of Marine and Freshwater Research*, 5:513–522. 17

Chitikila, C., Huisinga, K. L., Irvin, J. D., Basehoar, A. D., and Pugh, B. F. (2002). Interplay of tbp inhibitors in global transcriptional control. *Molecular Cell*, 10:871–882. 130, 131, 132, 133, 134, 136, 137

Crawford, S. L., DeGroot, M. H., Kadane, J. B., and Small, M. J. (1992). Modeling lake-chemistry distributions: Approximate Bayesian methods for estimating a finite-mixture model. *Technometrics*, 34:441–453. 152

Donoho, D. L. (1988). One-sided inference about functionals of a density. *The Annals of Statistics*, 16:1390–1420. 9, 49

Dudoit, S., Fridlyand, J., and Speed, T. P. (2002). Comparison of discrimination methods for

the classification of tumors using gene expression data. *Journal of the American Statistical Association*, 97(457):77–87. 129

Efron, B. and Tibshirani, R. (1993). *An introduction to the bootstrap*. Chapman & Hall Ltd. 15

Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *PNAS*, 95(25):14863–14868. 129, 136

Eisenberger, I. (1964). Genesis of bimodal distributions. *Technometrics*, 6:357–363. 106, 113

Fowlkes, E. B. (1979). Some methods for studying the mixture of two normal (lognormal) distributions. *Journal of the American Statistical Association*, 74:561–575. 17

Furman, W. D. and Lindsay, B. G. (1994). Testing for the number of components in a mixture of normal distributions using moment estimators. *Computational Statistics and Data Analysis*, 17:473–492. 19

Getz, G., Levine, E., and Domany, E. (2000). Coupled two-way clustering analysis of gene microarray data. *PNAS*, 97(22):12079–12084. 129

Ghosh, J. K. and Sen, P. K. (1985). On the asymptotic performance of the log likelihood ratio statistic for the mixture model and related results. In *Proceedings of the Berkeley conference in honor of Jerzy Neyman and Jack Kiefer (Vol. 2)*, pages 789–806. 12, 14

Goffinet, B., Loisel, P., and Laurent, B. (1992). Testing in normal mixture models when the proportions are known. *Biometrika*, 79:842–846. 14

Gu, C. (1998). Reply to comments on "Model indexing and smoothing parameter selection in nonparametric function estimation". *Statistica Sinica*, 8:638–646. 60

Hall, P. and Johnstone, I. (1992). Empirical functionals and efficient smoothing parameter selection (Disc: p509-531). *Journal of the Royal Statistical Society, Series B, Methodological*, 54:475–509. 60

Harding, J. (1948). The use of probability paper for the graphical analysis of polymodal frequency distributions. *Journal of Marine Biollogical Association*, 28:141–153. 17

Heckman, J. J., Robb, R., and Walker, J. R. (1990). Testing the mixture of exponentials hypothesis and estimating the mixing distribution by the method of moments. *Journal of the American Statistical Association*, 85:582–589. 19

Helguero, F. d. (1904). Sui massimi delle curve dimorfiche,. *Biometrika*, 3:85–98. 106, 113

Hoeffding, W. (1948). A class of statistics with asymptotically normal distribution. *Annals of Mathematical Statistics*, 19:293–325. 36

Kakiuchi, I. (1981). Unimodality conditions of the distribution of a mixture of two distributions. *Kobe University Mathematics Seminar Notes*, 9:315–32w5. 106

Kallenberg, W. C. M., Oosterhoff, J., and Schriever, B. F. (1985). The number of classes in chi-squared goodness-of-fit tests. *Journal of the American Statistical Association*, 80:959–968. 61

Kemperman, J. H. B. (1991). Mixtures with a limited number of modal intervals. *The Annals of Statistics*, 19:2120–2144. 106

Kerr, M. K., Afshari, C. A., Bennett, L., Bushel, P., Martinez, J., Walker, N. J., and Churchill, G. A. (2002). Statistical analysis of a gene expression microarray experiment with replication. *Statistica Sinica*, 12(1):203–217. 129

Kou, S. C. and Efron, B. (2002). Smoothers and the C_p, generalized maximum likelihood, and extended exponential criteria: A geometric approach. *Journal of the American Statistical Association*, 97(459):766–782. 60

Leroux, B. G. (1992). Consistent estimation of a mixing distribution. *The Annals of Statistics*, 20:1350–1360. 8

Lin, L. I.-K. (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, 45:255–268. 74

Lindsay, B. and Markatou, M. (2003). Statistical distancse: A global framework for inference. Book Manuscript. 45

Lindsay, B. G. (1995). *Mixture models: Theory, geometry and applications*. Institute of Mathematical Statistics. 12, 14, 19

Lindsay, B. G. and Roeder, K. (1992). Residual diagnostics for mixture models. *Journal of the American Statistical Association*, 87:785–794. 17

McLachlan, G. and Peel, D. (2000). *Finite Mixture Models*. John Wiley and Sons Inc. 16, 19, 39

McLachlan, G. J. (1987). On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture. *Applied Statistics*, 36:318–324. 15

Philips, D. and Smith, A. (1996). Bayesian model comparison via jump diffusions. In *Markov chain Monte Carlo in practice*, Gilks, W. R., Richardson, S. and Spiegelhalter, D. J. (EdS). London: Chapman & Hall Ltd , pages 115–130. 16

Raftery, A. (1996). Hypothesis testing and model selection. In *Markov chain Monte Carlo in practice*, Gilks, W. R., Richardson, S. and Spiegelhalter, D. J. (EdS). London: Chapman & Hall Ltd , pages 115–130. 16

Richardson, S. and Green, P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components (Disc: p758-792) (Corr: 1998V60 p661). *Journal of the Royal Statistical Society, Series B, Methodological*, 59:731–758. 16

Riesz, F. and Sz.-Nagy, B. (1990). *Functional Analysis*. Dover. 32

Robertson, C. A. and Fryer, J. G. (1969). Some descriptive properties of normal mixtures. *Skandinavisk Aktuarietidskrift*, 69:137–146. 106, 113

Roeder, K. (1990). Density estimation with confidence sets exemplified by superclusters and voids in the galaxies. *Journal of the American Statistical Association*, 85:617–624. 38

Roeder, K. (1994). A graphical technique for determining the number of components in a mixture of normals. *Journal of the American Statistical Association*, 89:487–495. 17

Satterthwaite, F. W. (1941). Synthesis of variance. *Psychometrika*, 6:309–316. 62

Satterthwaite, F. W. (1946). An approximate distribution of estimates of variance components. *Biometrics Bull.*, 2:110–114. 63

Schilling, M. F., Watkins, A. E., and Watkins, W. (2002). Is human height bimodal? *The American Statistician*, 56(3):223–229. 113

Serfling, R. J. (1980). *Approximation theorems of mathematical statistics*. John Wiley & Sons. 25, 76

Simar, L. (1976). Maximum likelihood estimation of a compound Poisson process. *The Annals of Statistics*, 4:1200–1209. x, 18, 19

Thyrion, P. (1960). Contribution *à* l'*é*tude du bonus pour non sinsitre en assurance automobile. *Astin Bulletin*. x, 18

Tibshirani, R., Hastie, T., Narasimhan, B., Eisen, M., Sherlock, G., Brown, P., and Botstein, D. (2002). Exploratory screening of genes and clusters from microarray experiments. *Statistica Sinica*, 12(1):47–59. 129

Tibshirani, R., Walther, G., and Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society, Series B, Methodological*, 63(2):411–423. 129

Titterington, D. M., Smith, A. F. M., and Makov, U. E. (1985). *Statistical analysis of finite mixture distributions*. John Wiley & Sons. 2, 10

Vlassis, N. and Likas, A. (1999). A kurtosis-based dynamic approach to gaussian mixture modeling. *IEEE Trans. Systems, Man, and Cybernetics, Part A*, 29:393–399. 19

Vlassis, N. A., Papakonstantinou, G., and Tsanakas, P. (1999). Mixture density estimation based on maximum likelihood and sequential test statistics. *Neural Processing Letters*, 9(1):63–76. 19

von Mises, R. (1947). On the assymptotic theory of differentiable statistical functions. *Annals of Mathematical Statistics*, pages 309–348. 24, 76

Wasserman, L. (2000). Asymptotic inference for mixture models using data-dependent priors. *Journal of the Royal Statistical Society, Series B, Methodological*, 62(1):159–180. 17

Wilson, S. R. (1982). Sound and exploratory data analysis. In *COMPSTAT 1982, Proceedings in Computational Statistics*, pages 447–450. 55

Yoo, D.-S. and Stark, W. E. (2003). Stochastic degrees of freedom in a wide-sense stationary uncorrelated scattering channel. Technical report, The University of Michigan. 60

Yosida, K. (1980). *Functional Analysis*. Springer-Verlag. 32

Vita

Surajit Ray

- **Education**
  - Ph. D. Statistics, The Pennsylvania State University, December 2003 (anticipated).
  - Master of Statistics, Indian Statistical Institute, India, May 1999
  - B.Sc. in Statistics, Presidency College, Calcutta, India, July 1997
- **Professional Experience**
  - 5/2002-Present: *Research Assistant*, Dept. of Statistics, The Pennsylvania State University.
  - 8/2001-12/2001: *Instructor (Stat 401),*The Pennsylvania State University.
  - 7/2000-7/2001: *Research Assistant, Add Health Project* , Pennsylvania State University.
  - 8/1999-5/2000: *Teaching Assistant (Stat 200)*, Dept. of Statistics, The Pennsylvania State University.
- **Publication**
  - Ayanendranath Basu, Surajit Ray, Chanseok Park and Srabashi Basu, Improved Power in Multinomial Goodness-of-fit Tests, *Journal of the Royal Statistical Society Series D*, September 2002, vol. 51, no. 3, pp. 381-393.