A Likelihood Based Approach to the Assessment of Large Sample Convergence and Model Based Clustering.

By

Milan Bimali

Submitted to the graduate degree program in Biostatistics and the Graduate Faculty of the University of Kansas in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

_____

Chairperson, Michael Brimacombe, PhD

_____

Brooke L. Fridley, PhD

_____

Francisco J. Diaz, PhD

_____

Jo A. Wick, PhD

_____

Udayan Apte, PhD

Date Defended: December 07, 2015

The Dissertation Committee for Milan Bimali

certifies that this is the approved version of the following dissertation:

A Likelihood Based Approach to the Assessment of Large Sample Convergence and Model Based Clustering.

_____

Chairperson, Michael Brimacombe.

Date approved: December 07, 2015

Abstract

The likelihood is a function of model parameter(s) and data using a pre-defined probability density function (pdf). Thus, the likelihood can be viewed as model-data combination that can be utilized to address questions of interest. The relative likelihood function is the likelihood function scaled by its mode so as to have its maximum at one. Unlike likelihood functions, relative likelihood functions have attracted little attention and use by statisticians. The proposed dissertation work explores the properties and applications of relative likelihood functions in examining the large-sample convergence properties of maximum likelihood estimator (MLE) and in relation to clustering.

The dissertation consists of three chapters. The first chapter presents a simulation based approach to examine the relationship between sample size and the asymptotic behavior of the MLE. The convergence of the observed relative likelihood function (RLF) to the asymptotic relative likelihood function (RLF) is assessed for different sample sizes using two measures of convergence; difference in areas and dissimilarity in shape. The proposed approach has been applied to data from the literature as well as to data simulated from different exponential family distributions.

The second chapter proposes a novel clustering approach based on the observed RLFs. Observations in the dataset are assumed to follow a known distribution and observed RLFs are obtained. The observed RLFs are further scaled by the inverse of the asymptotic variation (Fisher Information) evaluated at the mode of the likelihood functions. The weighted RLFs reflect information based similarity among observations in the data. A data matrix is then developed by evaluating the weighted RLFs at different values in the parameter space. The data matrix allows for direct application of standard clustering algorithms such as k-means algorithm. This

clustering approach was applied to simulated dataset based on real data and to datasets simulated from known distributions.

The third chapter examines the proposed RLF based clustering approach to a publicly available gene expression dataset consisting of 70 gene expression profiles used to classify patients into prognostic groups. The agreement between the RLF clustering results and previous classification is also presented. The clusters obtained are also examined in relation to differences in two clinical features – time to overall survival; and time to metastases**.**

Acknowledgements

I would like to thank my academic advisor as well as dissertation supervisor, Dr. Michael Brimacombe, whose rare combination of serenity, depth of statistical knowledge, coupled with patience and encouragement throughout these years made the dissertation possible. I would also like to thank my committee members, Dr. Brooke Fridley, Dr. Francisco J. Diaz, and Dr. Jo A. Wick for their unparalleled guidance, and support throughout my graduate student years. My work on Dr. Fridley's projects as a graduate research assistant enhanced my computing skills and also introduced me to novel areas and methods within biostatistics. Dr. Diaz's feedback in my work has always been rewarding. Dr. Wick was important in multiple ways. She was the go-to person for any question regarding the graduate school. I would also like to thank Dr. Udayan Apte for his time and effort in serving in the committee as an external committee member. I would also like to acknowledge department chair, Dr. Matthew S. Mayo for his generous support in helping us attend and present at regional as well as national conferences. Most importantly, I want to express my deepest gratitude and love for my parents, Bishnu Maya Bimali and Tika Prasad Bimali, for their love, care, and support.

Table of Contents

List of Tables

List of Figures

Introduction

"Likelihood" is arguably the most used term in the statistical realm and was defined and popularized by the eminent geneticist and statistician R.A. Fisher. Likelihood function, since its inception in the early 1900s has emerged as an indispensable and fundamental tool in inferential statistics. The likelihood function is a function of model parameter(s), data and a pre-defined probability density function (pdf). Thus the likelihood function reflects the observed model-data combination and can be used to investigate research questions of interest.

Let $f(\boldsymbol{x}|\theta)$ be the joint pdf of $\boldsymbol{X} = (X_1, \dots, X_n)$. For observed values of $\boldsymbol{X}$, $\boldsymbol{x} = (x_1, \dots, x_n)$, the likelihood function (under iid assumption) is formally defined as:

$$L(\theta|\boldsymbol{x}) = f(\boldsymbol{x}|\theta) = \prod_{i=1}^{n} f(x_i|\theta)$$

Likelihood functions can be used to examine the plausibility of data for different values of the parameter θ under the given model $(f(x|\theta))$. Log transformation allows us to work with the sum of log-transformed pdfs instead of product of the pdfs. Owing to computational simplicity and mathematical convenience; log-likelihood functions in lieu of likelihood functions are more commonly applied.

The value of the parameter $(\hat{\theta})$ which makes the data most "likely" under the given model is called the MLE of $\theta, i.e. \hat{\theta}$ is the value of the parameter that maximizes the likelihood function $L(\theta|\boldsymbol{x})$. MLEs have been of particular interest to statisticians due to several optimal properties such as asymptotic normality, invariance of parameterization, and consistency. In particular, the asymptotic normality of the MLEs has been used in drawing inference about the parameter.

Besides the MLE, the curvature of the log-likelihood function is also of particular interest. The observed Fisher Information evaluated at the value of the MLE, reflects the local

curvature of a log-likelihood function and summarizes the estimation related accuracy present in an individual likelihood. The observed Fisher Information evaluated at the MLE has also been used as an approximation to the average (expected) Fisher information.

Relative likelihood functions (RLFs) are the likelihood functions re-scaled by their mode. Thus if $\hat{\theta}$ is the MLE of θ, then the RLF of θ is defined as:

$$R(\theta) = \frac{L(\theta|\boldsymbol{x})}{L(\hat{\theta}|\boldsymbol{x})}$$

It is obvious that unlike likelihood functions, RLFs are bounded above by one. The initial observed likelihood and its relative version have the same *MLE* and Fisher Information values. RLFs evaluated across the values of θs measure the plausibility of the value of θ relative to the mle $(\hat{\theta})$. In situations where we are interested in a group of subjects measured on a set of variables, using a relative scale makes a set of likelihood functions more comparable as they all have a maximum value of one.

The proposed dissertation work explores the properties and applications of RLFs. The dissertation consists of three chapters. The first chapter proposes a bootstrap/simulation based approach in examining the relationship between sample size and asymptotic behavior of the MLEs by assessing the convergence of the observed RLF (based on data) to the asymptotic RLF. The second chapter proposes a novel clustering approach based on the weighted observed RLF of each observation in the dataset weights being Fisher Information evaluated at the mode of the likelihood functions. The third chapter applies the proposed likelihood based clustering approach to a publicly available gene expression data that contains patients classified into two prognostic groups based on 70 gene expression profiles. The agreement between the proposed clustering approach and authors' classification has been examined. The clusters obtained are also examined

in relation to differences in two clinical features – time to overall survival; and time to metastases**.**

1.1 Relative Likelihood Differences to Examine the Convergence of Asymptotic: A Simulation Study.

MLEs and their large sample properties are extensively used in descriptive as well as inferential statistics. In the framework of large sample distribution of *MLE*, the relationship between the sample size and asymptotic convergence of *MLE* is important i.e. for what sample size does the *MLE* behave satisfactorily attaining asymptotic normality. Sprott (1969) has discussed the undesirable impacts of using large sample approximations of the *MLE*s when such approximations do not hold. It has been argued that RLFs must be examined before making inferences based on *MLE* (Sprott and Kalbfleisch 1969). Sprott (1969) proposed the comparison of the observed RLF (based on the data) with the normal RLF (based on the asymptotic normality of mle) as a diagnostic measure for assessing the satisfactory behaviour of large sample normality of *MLE*.

The two RLFs can be assessed graphically and a large difference would make the large sample normality assumption for *MLE* questionable. It was also demonstrated that transformation of *MLE* can help achieve asymptotic normality with smaller sample sizes. In other words, he proposed examining the RLFs based on a transformation of mle while holding the sample and hence the sample size fixed. This naturally begets an interesting question – if we examine the RLFs for different (increased) sample sizes, what is the impact on convergence of the observed RLF to the asymptotic RLF? There has been little work regarding the appropriate sample size that would allow the *MLE* achieve asymptotic normality from relative likelihood perspective directly.

Our work proposes simulation based approach in examining the relationship between sample size and asymptotic behaviours of *MLE*. We propose two measures of the convergence of observed RLF to the asymptotic RLFs namely: differences in areas and dissimilarity in shape between the two RLFs. We argue that, for a given sample size, if the difference in the area under the two RLFs and the dissimilarity index between them is both close to zero (below a pre-specified threshold), the asymptotic approximation of *MLE* is satisfactorily achieved.

To study the properties of these measures and related likelihood convergence, we use the simulation methods to generate samples of varying size based on initial samples from examples in literature. Our results show that the sample size requirements for the large sample normality of mle vary remarkably depending on the distribution assumed to have generated the sample. The sample size requirement can range from surprisingly small (as in the case of Poisson distribution) to large (e.g. exponential distribution).

1.2 Likelihood Transformation and Information Based Approach to Clustering.

Clustering is a grouping procedure focused on identifying subgroups within a dataset. Traditional non-parametric clustering methods such as hierarchical clustering and k-means clustering are commonly used and do not assume any density function on the data. Parametric clustering approaches such as model-based clustering, however assume that data is generated from a mixture of distribution. Despite the differences in assumptions and approaches, the objective of most clustering algorithms is to classify subjects or observations into one of a finite set of disjoint clusters while ensuring that subjects within a cluster are more similar than subjects across clusters.

The use of likelihood function for clustering has remained largely unexplored. Spatial clustering algorithm proposed by Kerby et al (2009) makes use of pair-wise defined likelihood functions together with a grouping algorithmic hierarchical approach (Kerby, Marx et al. 2007). Likelihood based hierarchical clustering has also been pursued by Castro et al (2003) (Castro and Nowak 2003). However to the best of our knowledge no clustering algorithm has been proposed based on similarities or dissimilarities between the information based on a set of likelihood functions.

In the second chapter we propose a more basic clustering approach based on likelihood function that takes into account the structure of data, as well as the similarity in Fisher Information across the subjects in the data. Each observation in the dataset is assumed to follow a known distribution and thus a set of likelihood functions can be constructed which are further scaled by their value at mode to obtain RLF. The relative scale makes a set of likelihood functions more comparable as they all have maximum of one. The RLFs are further scaled by their Fisher information evaluated at the MLE to obtain weighted RLFs. The weighted RLFs are

then evaluated at different values within the parameter space to obtain a data matrix which can be subsequently examined for patterns or subgroups in the data using the traditional non parametric clustering approach such as k means. The proposed approach was applied to simulated data based on real data set as well as to data simulated from known distributions.

1.3 Information Based Clustering of Gene Expression Signatures in Primary Breast Carcinoma Patients.

In the context of gene expression data, clustering techniques have been employed to identify sub-groups of patients at the molecular level, to understand gene function and regulation. It has been applied successfully to group similarly expressed genes across a set of subjects, as well as, in grouping subjects that share similar gene expression profiles (Jiang, Tang et al. 2004). In the context of clustering gene expression data, hierarchical clustering and k means clustering are more commonly used (D'Haeseleer 2005). Other approaches such as fuzzy c means clustering, self-organizing maps and model-based clustering have also been employed. This chapter contains an application of the proposed RLF based clustering approach based on the properties of the observed likelihood and Fisher Information in the subjects across a set of gene expressions profiles.

In the context of gene expression, the proposed method assumes the gene expression profile for each subject follows a known distribution and thus a set of RLFs can be constructed. The likelihood functions can be viewed as a transformation of the original gene expression profiles. These RLFs are further weighted by the Fisher Information to obtain the weighted RLF. This is then evaluated at different values of the parameter to obtain a data matrix which can be subjected to the clustering algorithms. The proposed clustering approach takes into account the variation in

mean expression levels as well as the variation in the observed Fisher Information across the patients. We make use of the publicly available dataset by Van De Vijer et al (2012) in clustering primary breast carcinomas patients based on a previously recommended set of 70 gene expression profiles. The agreement between the proposed clustering approach and authors' classification has been examined. The clusters obtained are also examined in relation to two clinical features – time to overall survival; and time to metastases**.**

Chapter 1: Relative Likelihood Differences to Examine Asymptotic Convergence: A Simulation Study[1].

1.1 Introduction[1]

"Likelihood" is arguably the most pronounced terminology in the statistical realm and was defined and popularized by the eminent geneticist and statistician R.A. Fisher (1922) (Fisher 1932, Fisher 1934, Fisher 1934, Fisher 1941). The likelihood function is a function of model parameter(s) based on a given set of data and a pre-defined probability density function (pdf). The Likelihood function is formally defined as follows:

Let $f(\boldsymbol{x}|\theta)$ be the joint pdf of $\boldsymbol{X} = (X_1, \dots, X_n)$. For observed values of $\boldsymbol{X}$, $\boldsymbol{x} = (x_1, \dots, x_n)$, the likelihood function (under iid assumption) is formally defined as:

$$L(\theta|\boldsymbol{x}) = f(\boldsymbol{x}|\theta) = \prod_{i=1}^{n} f(x_i|\theta)$$

A key point often reiterated in textbooks is that the likelihood function is a function of θ and not to be viewed as a probability density itself (Casella 2001). However, the shape of the likelihood function relative to its mode is often of interest in estimating θ. Likelihood functions can be mathematically constructed for most statistical distributions; however MLEs may not always have closed form (Altman, Gill et al. 2003). Nevertheless most of the distributions commonly used allow the computation of MLEs either analytically, numerically or graphically. Several properties of MLEs such as asymptotic normality, invariance, and ease of computation have made MLEs popular (Efron 1982). In this paper we assume θ is a scalar throughout.

[1] This chapter has been published at Journal of Biometric and Biostatistics (Bimali and Brimacombe, J Biom Biostat 2015, 6:1 http://dx.doi.org/10.4172/2155-6180.1000220).

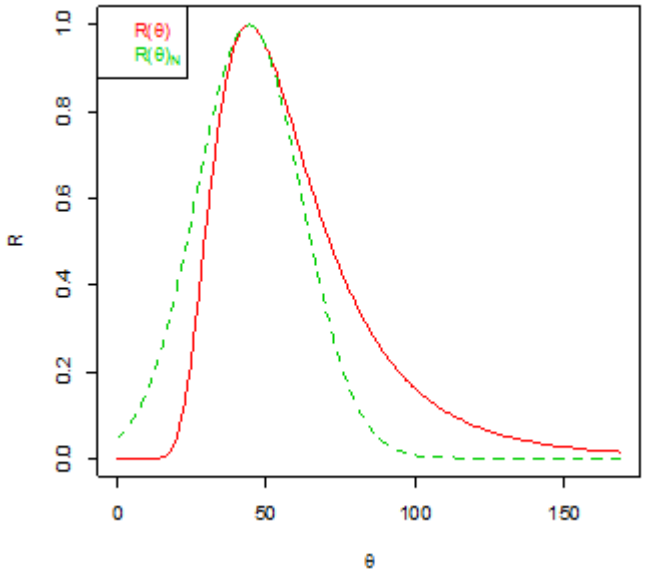
The large sample distribution of the MLEs is often used for inferential purposes. If $\hat{\theta}$ is the *MLE* of θ, then for large sample, $\hat{\theta} \sim N(\theta, I^{-1}(\hat{\theta}))$ where $I(\hat{\theta}) = E\left(\left(\frac{\partial}{\partial\theta} logf(\boldsymbol{X}|\theta)\right)^2\right)$ is the Fisher Information evaluated at $\hat{\theta}$. In situations where the computation of the expected Fisher Information is analytically intractable, $J(\hat{\theta}) = \left(\left(\frac{\partial}{\partial\theta} logf(\boldsymbol{X}|\theta)\right)^2\right)\Big|_{\theta=\hat{\theta}}$, the observed Fisher Information, has been used as an approximation in computation of $I(\hat{\theta})$ (Efron and Hinkley 1978).

A common question that often arises in statistics is in regard to sample size. In the framework of the large sample distribution of the *MLE*, we are interested in knowing for what sample size the *MLE* behaves satisfactorily, attaining the asymptotic normal distribution. Put a different way, does the existing sample size allow us to use the large sample properties of the *MLE* with confidence? If not, what would be an ideal sample size?

Sprott et al (1977) have elicited some of the undesirable impacts of using large sample approximation of the *MLE* when such approximations do not seem to hold (Sprott and Kalbfleisch 1969). They argue in favor of examining likelihood functions before making inferences about *MLE*. They demonstrate via an example from Bartholomew (Bartholomew 1957) that drawing inferences from the *MLE* without first examining the likelihood functions can be misleading. Fig 1.1 gives the plot of the observed RLF (likelihood function based on observed data, assumed pdf scaled by their mode) as obtained from Bartholomew's data and normal RLF (likelihood function based on large sample normality of MLE scaled by their mode). The plot shows that for a pre-specified value of relative likelihood, the range of θs can be in complete disagreement between the two likelihood functions. E.g. for relative likelihood of 10% or higher, the ranges θ are roughly $(20,110)$ and $(7,81)$ for the observed RLF and normal RLF (Sprott and

Kalbfleisch 1969), approximately 17% drop in coverage. Sprott et al also demonstrated that transformation of the parameter can help achieve the asymptotic normality with smaller sample sizes. However little has been explored regarding the appropriate sample size that would allow the *MLE* to achieve asymptotic normality from a relative likelihood perspective directly.

Figure 1.1: Bartholomew's Data - Observed and Normal Relative Likelihood Functions



Our work proposes a simulation based approach to the above question via the behavior and properties of RLF. In particular we measure the proximity of the observed likelihood function based on the actual sample to the likelihood function based on large sample properties, both of which are scaled here by their modes to have a maximum at one. The two convergence measures proposed by the authors are (i) difference in area under the two RLFs and (ii) dissimilarity in the shape of the two RLFs (dissimilarity index). We propose that, for a given sample size, if the difference in the area under the two RLFs and the dissimilarity index between them are both close to 0, the asymptotic approximation of *MLE* is satisfactorily achieved. To study the properties of these measures and related likelihood convergence, we use the bootstrap to generate samples of varying size based on initial samples for examples in literature.

The paper is laid out as follows. Section 1.2 provides a review of bootstrap method and some of the proposed measures of distance between distributions. In section 1.3, we provide the mathematical details of the two measures of convergence. In section 1.4 we provide examples by simulating data from exponential families of distributions and apply our method to some of the data available in literature and textbook.

1.2 Review of Bootstrap and Measuring Distances between Distributions

1.2.1 Bootstrap

The Bootstrap is a resampling technique introduced by Efron (1979) with a related long history (Hall 1994) and has attracted immense attention in the past three decades primarily due to its conceptual simplicity and due to the computational empowerment of statisticians due to advances in computer science (Efron 1979). The past three decades have witnessed numerous works dedicated to developing bootstrap methods (Hall 1986, Singh 1986, Hall 1988, Hall 1988, Hall 1989, Hall and Wilson 1992, Leger, Politis et al. 1992).

Bootstrap at its core, is a resampling technique that treats the data at hand as a "surrogate population" and allows resampling with replacement with a goal of re-computing the statistic of interest many times. This allows us to examine its distribution. Efron has demonstrated that bootstrap method outperforms other resampling methods such as jackknifing and cross-validation (Efron 1979). The distribution of the computed statistics is referred to as the bootstrap distribution. Despite the mathematical modesty of bootstrap algorithm, the large sample properties of bootstrapping distributions are surprisingly elegant. Singh (1981), for example has demonstrated that the sampling distribution of $(\hat{\theta} - \theta)$, where $\hat{\theta}$ is an estimate of θ, is approximated well enough by its bootstrap distribution(Singh 1981). Bickel and Freedman have also made substantial contributions in developing bootstrap theory (Bickel and Freedman 1981, Bickel and Freedman 1984, Freedman 1986). The most common applications of the bootstrap in its basic form involve approximating standard error of sample estimate, correcting the bias in the sample estimate, and in constructing confidence intervals. However in situations involving bootstrapping dependent data, modified bootstrap approaches such as moving-block bootstrap

are recommended (Carlstein, Do et al. 1998). Romano (1992) has discussed extensively the applications of bootstrap (Leger, Politis et al. 1992).

1.2.2 Distance between distributions

Kullback-Leibler distance is a commonly used measure of difference between two statistical distributions (Kullback and Leibler 1951). If $p(x)$ and $q(x)$ are two continuous distributions the KL distance between them is defined as follows:

$$KL(p\|q) = \int_{x \in \Omega} p(x) \log \frac{p(x)}{q(x)} dx = \int_{x \in \Omega} p(x) \log p(x)\, dx - \int_{x \in \Omega} p(x) \log q(x)\, dx.$$

Kullback-Leibler distance has been applied in areas such as functional linear model, Markovian process, model selection, and classification analysis (Gersch, Martinelli et al. 1980, Rodrigues 1992, Kon'kov, Morozov et al. 2007, Kubokawa and Tsukuma 2007). It should be noted that the Kullback-Leibler distance is not symmetric, $KL(p||q) \neq KL(q||p)$, but can be expressed in a symmetric form (Johnson and Sinanovic 2001).

Bhattacharya distance is another popular measure of difference between two distributions (Bhattacharyya 1943). If $p(x)$ and $q(x)$ are two continuous distributions the Bhattacharya distance between $p(x)$ and $q(x)$ is defined as follows:

$$B(p, q) = \int_{x \in \Omega} \sqrt{p(x)q(x)}\, dx$$

The measure for discrete distribution is identical with integral replaced by summation. Bhattacharya distance has also found extensive applications in several fields (Schweppe 1967, Jain 1976, Chaudhuri, Borwankar et al. 1991, Chen, Li et al. 2014). Bhattacharya distance assumes the product $p(x)q(x)$ to be non-negative.

RLFs typically yield non-negative values. Thus in lieu of the above two distance measures, we could simply use $\int_{\theta\in\Omega}(f_1(\theta) - f_2(\theta))d\theta$ as a measure of proximity between the two functions $f_1(\theta)$ and $f_2(\theta)$. Geometrically this measure is the difference in the area under the two curves generated by $f_1(\theta)$ and $f_2(\theta)$.

In this paper we make use of the bootstrap approach to resample from the actual sample (or simulate data from known distributions) to obtain a "bootstrap sample". The size of the resampled "bootstrap sample" is taken to exceed the size of actual sample. For each "bootstrap sample", the observed RLF and corresponding normal (asymptotic) RLF are constructed and the area under the two RLFs computed. As the size of "bootstrap sample" increases we measure the convergence of the observed RLF to the asymptotic RLF. The convergence is also measured by a form of "Kullback-Leibler" distance and a dot product based measure of curve similarity. We note that simulated data is not a real world data and the sample sizes determined here are obtained in an ideal situation.

1.3 Method

1.3.1 Background

Let $X_1, \dots, X_n$ be iid random variables from a specified distribution $f(X|\theta)$ with observed values $x_1, \dots, x_n$. The relative observed likelihood function of θ i.e. $R(\theta)$ is defined as follows:

$$R(\theta) = \frac{L(\theta|\boldsymbol{x})}{L(\hat{\theta}|\boldsymbol{x})}.$$

For large samples and under regularity conditions, $\hat{\theta} \sim N(\theta, I^{-1}(\hat{\theta}))$, the relative large sample normal likelihood of θ can be defined as follows:

$$R_N(\theta) = \frac{L_N(\theta)}{L_N(\hat{\theta})}.$$

For exponential families, the density function can be expressed in the following form:

$$f(x|\theta) = exp(h(x)c(\theta) - a(\theta) + b(x)),$$

and the likelihood function can be expressed as:

$$L(\theta|\boldsymbol{X}) = exp(c(\theta)\sum_i h(x_i) - na(\theta) + \sum_i b(x_i))$$

If $\hat{\theta}$ is the *MLE* of θ, then the likelihood function evaluated at $\hat{\theta}$ is:

$$L(\hat{\theta}|\boldsymbol{X}) = exp\big(c(\hat{\theta})\sum_i h(x_i) - na(\hat{\theta}) + \sum_i b(x_i)\big)$$

Thus the RLF $R(\theta)$ is:

$$R(\theta) = \frac{L(\theta|\boldsymbol{X})}{L(\hat{\theta}|\boldsymbol{X})} = \frac{exp(c(\theta)\sum_i h(x_i) - na(\theta) + \sum_i b(x_i))}{exp(c(\hat{\theta})\sum_i h(x_i) - na(\hat{\theta}) + \sum_i b(x_i))} = \frac{exp\big(\sum_i h(x_i)(c(\theta) - c(\hat{\theta}))\big)}{exp\big(n(a(\theta) - a(\hat{\theta}))\big)}$$

The asymptotic RLF for $\hat{\theta}$ assumes the following form:

$$R_N(\theta|\hat{\theta}) = exp\left(\frac{-1}{2}\left(\frac{\hat{\theta}-\theta}{std.err(\hat{\theta})}\right)^2\right) = exp\left(\frac{-1}{2} \times I(\hat{\theta}) \times (\theta - \hat{\theta})^2\right)$$

Where,

$std.err(\hat{\theta}) = \sqrt{I^{-1}(\hat{\theta})}$ and

$I(\hat{\theta}) = -E\left(\frac{\partial^2}{\partial\theta^2}\log(L(\theta|X))\right)\Big|_{\theta=\hat{\theta}} = -c''(\hat{\theta})\sum_i E(h(x_i)) + na''(\hat{\theta}).$

In situations where computation of expectations are not analytically tractable, $I(\hat{\theta})$ will be estimated by $J(\hat{\theta}) = -\left(\frac{\partial^2}{\partial\theta^2}\log(L(\theta|X))\right)\Big|_{\theta=\hat{\theta}}$.

Here both $R(\theta)$ and $R_N(\theta)$ are positive since both are exponential functions.

1.3.2 Measure of distance between $R(\theta)$ and $R_N(\theta)$.

If $R(\theta)$ and $R_N(\theta)$ are defined over the interval (θ_L, θ_U), the difference in area under the two likelihood curves will serve as the measure of discrepancy between $R(\theta)$ and $R_N(\theta)$ and can be computed analytically as follows:

$$\Delta R = \int_{\theta_L}^{\theta_U} R(\theta)d\theta - \int_{\theta_L}^{\theta_U} R_N(\theta)d\theta = \int_{\theta_L}^{\theta_U}\left(R(\theta) - R_N(\theta)\right)d\theta \qquad (1).$$

If the expression does not have a closed form solution, numerical methods such as Simpson's rule (Atkinson 1989) can be applied:

$\Delta R = \sum_{i=1}^{k}\left(R(\theta_{i-1}) - R(\theta_i)\right)\delta\theta - \left(R_N(\theta_{i-1}) - R_N(\theta_i)\right)\delta\theta$

$= \sum_{i=1}^{k}\left[\left(R(\theta_{i-1}) - R_N(\theta_{i-1})\right) - \left(R(\theta_i) - R_N(\theta_i)\right)\right]\delta\theta.$

Where k is the number of intervals.

For similar curves we would expect ΔR to be very small. "How small is small?" – the examples in the next section demonstrate that different distributions have different thresholds. This is primarily related to the fact that the domain of the parameters varies for different distributions. For example, in binomial distribution– $Bin(n,p)$, $p \in (0,1)$ whereas in the exponential distribution $exp(\beta), \beta \in (0,\infty)$. It is thus recommended that the measure of proximity should be considered on case by case basis. A tolerance level may also bet set: $\left|A_{n_1} - A_{n_2}\right| < \epsilon; n_1 < n_2$. Typically $\epsilon = 0.01$ will be acceptable.

1.3.3 Property of ΔR:

1. On a log scale, $R(\theta|X)$ can be approximated by $R_N(\theta|X)$ up to a second term.

Proof:

The general expression for Taylor expansion of a function $f(x)$ around "a" is as follows:

$$f(x) = \sum_{k=0}^{\infty} \frac{(x-a)^k}{k!} f^{(k)}(a) \text{ where } f^{(k)}(x) = \frac{\partial^k}{\partial x^k} f(x)$$

Using Taylor expansion on $\log(R(\theta|X))$ around $\hat{\theta}$ we have:

$$\log\big(R(\theta|X)\big) = \log\big(R(\theta)\big)\Big|_{\theta=\hat{\theta}} + \big(\theta - \hat{\theta}\big)\frac{\partial}{\partial\theta}\log\big(R(\theta)\big)\Big|_{\theta=\hat{\theta}} + \frac{\big(\theta-\hat{\theta}\big)^2}{2}\frac{\partial^2}{\partial\theta^2}\log\big(R(\theta)\big)\Big|_{\theta=\hat{\theta}} +$$

$$higher\ order\ powers\ of\ k$$

Now:

$$\log\big(R(\theta)\big)\Big|_{\theta=\hat{\theta}} = log\left(\frac{L\big(\hat{\theta}|X\big)}{L\big(\hat{\theta}|X\big)}\right)\Bigg|_{\theta=\hat{\theta}} = log\left(\frac{L\big(\hat{\theta}|X\big)}{L\big(\hat{\theta}|X\big)}\right) = \log(1) = 0$$

$\frac{\partial}{\partial\theta}\log\big(R(\theta)\big)\Big|_{\theta=\widehat{\theta}} = \frac{\partial}{\partial\theta}\Big(log\big(L(\theta)\big) - log\left(L\big(\widehat{\theta}\big)\right)\Big)\Big|_{\theta=\widehat{\theta}} = \frac{\partial}{\partial\theta}\Big(log\big(L(\theta)\big)\Big)\Big|_{\theta=\widehat{\theta}} = 0$. This is the score function evaluated at *MLE*.

$\frac{\partial^2}{\partial\theta^2}\log\big(R(\theta)\big)\Big|_{\theta=\widehat{\theta}} = \frac{\partial^2}{\partial\theta^2}\log\big(L(\theta)\big)\Big|_{\theta=\widehat{\theta}} = -J_{\widehat{\theta}}$.

Thus the above expression can be written as:

$\log\big(R(\theta|X)\big) \approx \frac{\left(\theta-\widehat{\theta}\right)^2}{2}(-J_{\widehat{\theta}}) \approx \log(R_N(\theta|X))$

The $k!$ in the higher order terms of the Taylor expansion shrinks them to 0.

Using the approximation $\log\big(R(\theta)\big) \approx \log(\mathrm{R_N}(\theta))$ for exponential families we have:

$\log(R(\theta)) = log\left(\frac{exp\left(\sum_i h(x_i)\left(c(\theta)-c(\widehat{\theta})\right)\right)}{exp\left(n\left(a(\theta)-a(\widehat{\theta})\right)\right)}\right) = \sum_i h(x_i)\big(c(\theta) - c(\widehat{\theta})\big) - n\big(a(\theta) - a(\widehat{\theta})\big)$, and

$\log\big(R_N(\theta)\big) = log\left(exp\left(-\frac{1}{2}\big(\theta - \widehat{\theta}\big)^2\big(-c''\big(\widehat{\theta}\big)\sum_i E\big(h(x_i)\big) + na''(\widehat{\theta})\big)\right)\right) = -\frac{1}{2}\big(\theta -$

$\widehat{\theta}\big)^2\big(-c''\big(\widehat{\theta}\big)\sum_i E\big(h(x_i)\big) + na''(\widehat{\theta})\big)$

So,

$\sum_i h(x_i)\big(c(\theta) - c(\widehat{\theta})\big) - n\big(a(\theta) - a(\widehat{\theta})\big) \approx -\frac{1}{2}\big(\theta - \widehat{\theta}\big)^2\big(-c''\big(\widehat{\theta}\big)\sum_i E\big(h(x_i)\big) + na''(\widehat{\theta})\big)$

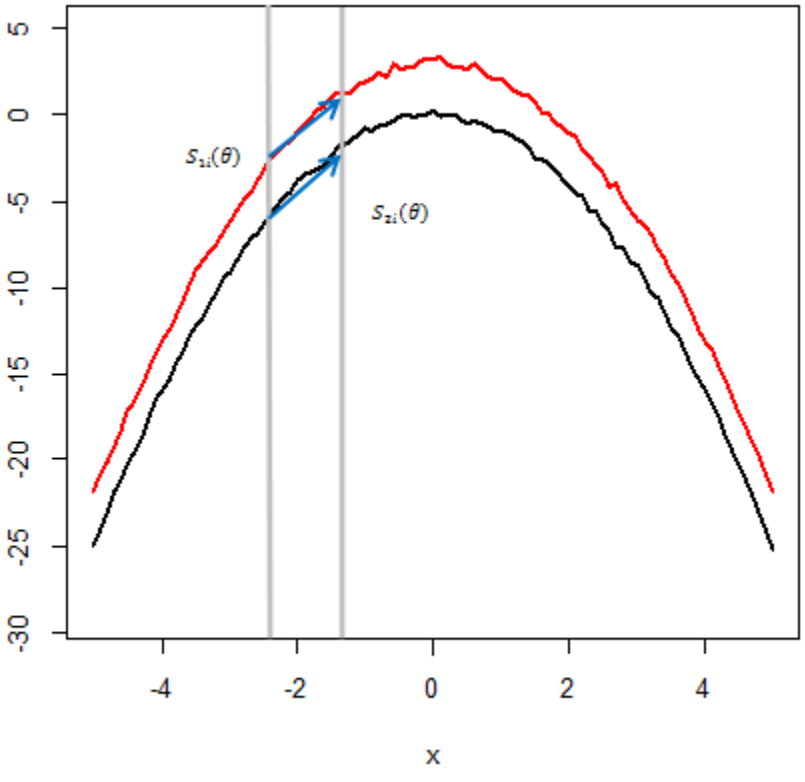
This implies that higher order terms in the Taylor expansion are converging to zero. Our method here graphically demonstrates this as a function of n.

1.3.4 Curve Dissimilarity Index

Let $L_1(\theta)$ and $L_2(\theta)$ be two different functions of θ with the same domain Ω. Graphically $L_1(\theta)$ and $L_2(\theta)$ can be visualized as two curves constructed on the same domain. The two curves need not necessarily have closed functional form. Here we propose a simple and computationally efficient algorithm that uses the dot product to measure the similarity of the two curves in terms of their curvature.

The idea is to divide the domain of the two curves into sufficiently small segments so that each of them can be approximated by a line segment (Figure 1.2). Each of these segments is equivalent to a vector in two dimensions and hence we can compute the dot product for the two vectors in each of these segments. If in general the two vectors are parallel in each of these segments, this would imply that the two curves have similar local curvature and hence the curves are locally similar. In other words, for similar curves, the dot product between the two vectors is equal to the product of their individual L_2 norms over each segment.

Figure 1.2: Dissimilarity Index.



Let $i = 1, \dots, n+1$, be the number of points over which the two curves are segmented i.e. there are n segments of the two curves in total. $S_{1i}(\theta)$ and $S_{2i}(\theta)$ be the two vectors that approximate the segments of the two curves. Let

$$d_i(S_{1i}, S_{2i}) = \frac{S_{1i}.S_{2i}}{\|S_{1i}\|_2 \|S_{2i}\|_2} \qquad (2)$$

The following are properties of $\boldsymbol{d_i}$:

1. $|d_i| \leq 1$.
2. $d_i = 1$ if S_{1i} and S_{2i} are parallel. This is the case for perfect similarity.
3. $d_i = -1$ if S_{1i} and S_{2i} are in opposite direction. This is the case for perfect dissimilarity.

Ideally if two curves were exactly same, we would expect:

$\sum_{i=1}^{n} d_i = n$ (3)

1.3.5 Dissimilarity Index

Equation (3) can be used to express disagreement between the two curves (here referred to as dissimilarity index). If D is the dissimilarity index between the two curves then,

$D = \frac{n - \sum_i d_i}{n}$

Note that: $0 \leq D \leq 1$.

1.3.6 Bootstrap Algorithm

The proposed bootstrap algorithm can be summarized in the following steps.

1. For a given sample $(X_1, \dots, X_n)$, compute $R(\theta)$ and $R_N(\theta)$.
2. Choose tolerance level for ΔR and D.
3. Compute ΔR and D for the given sample.
4. If ΔR and D are not sufficiently close to 0, bootstrap from the original sample and compute ΔR and D again for the bootstrapped sample (increased sample size).
5. Repeat step 3 until satisfactory convergence is achieved i.e. $D \; and \; \Delta R$ is less than chosen tolerance level.

The next section contains several simulated examples to demonstrate the application of the above method.

1.4 Results

In this section, we examine the convergence of likelihood functions for some of the common distributions, using simulated data as well as for data obtained from the literature. Expression for $R(\theta)$ and $R_N(\theta)$ for some common distributions are tabulated in Table 1.1. We would like to reiterate that $R(\theta)$ and $R_N(\theta)$ are the observed and asymptotic likelihood functions scaled by their modes.

Table 1.1: Relative and relative Normal likelihood functions for some exponential families

Distribution	$R(\theta)$	$R_N(\theta)$	Note
Poisson	$exp(n(\bar{x}-\lambda))\left(\frac{\lambda}{\bar{x}}\right)^{n\bar{x}}$	$exp\left(-\frac{n}{2\bar{x}}(\lambda-\bar{x})^2\right)$	$\hat{\lambda}=\bar{x}$
Binomial	$\left(\frac{p}{\hat{p}}\right)^{\sum x_i}\left(\frac{1-p}{1-\hat{p}}\right)^{\sum(n-x_i)}$	$exp\left(-\frac{n(p-\hat{p})^2}{2\times\hat{p}\times(1-\hat{p})}\right)$	$\hat{p}=\frac{\sum_i x_i}{n}$
Exponential	$\left(\frac{\hat{\beta}}{\beta}\right)^n exp\left(\sum_i x_i\left(\frac{1}{\hat{\beta}}-\frac{1}{\beta}\right)\right)$	$exp\left(\frac{-n}{2\hat{\beta}^2}\left(\beta-\hat{\beta}\right)^2\right)$	$\hat{\beta}=\bar{x}$
Weibull (Shape parameter fixed)	$\left(\frac{\hat{\beta}}{\beta}\right)^n exp\left(\left(\frac{1}{\hat{\beta}}-\frac{1}{\beta}\right)\left(\sum_{i=1}^n x_i^{\gamma^*}\right)\right)$	$exp\left(-\frac{1}{2}I(\hat{\beta})\left(\beta-\hat{\beta}\right)^2\right)$	γ^*: a constant $\hat{\beta}=\frac{\sum_{i=1}^n x_i^{\gamma^*}}{n}$ $I(\hat{\beta})=\frac{2}{\hat{\beta}^3}\left(\sum_{i=1}^n x_i^{\gamma^*}\right)-\frac{n}{\hat{\beta}^2}$

1.4.1 Simulations

The convergence of the observed RLF to asymptotic RLF was first examined using simulated dataset. For different families of exponential distributions, data were simulated for a given sample size. For the given data, the two convergence measures namely ΔR and D were

computed. This process was repeated for different sample sizes and the values of ΔR and D thus obtained were recorded. The following are some of the examples of the above distribution (table 1.1) and the required sample size that makes the large sample approximation of the *MLE* reasonable.

Example 1: Poisson distribution$(\lambda = 10)$

Table 1.2: Poisson Distribution - Convergence measures.

n	ΔR: Difference in Area	D: Dissimilarity Index
1	0.0691	0.01968
3	0.01318	0.01183
5	0.00604	0.00767
7	0.00359	0.00575
10	0.00201	0.00374
15	0.00114	0.00303

Figure 1.3: Poisson distribution – Change in values of Difference in Area and Dissimilarity Index

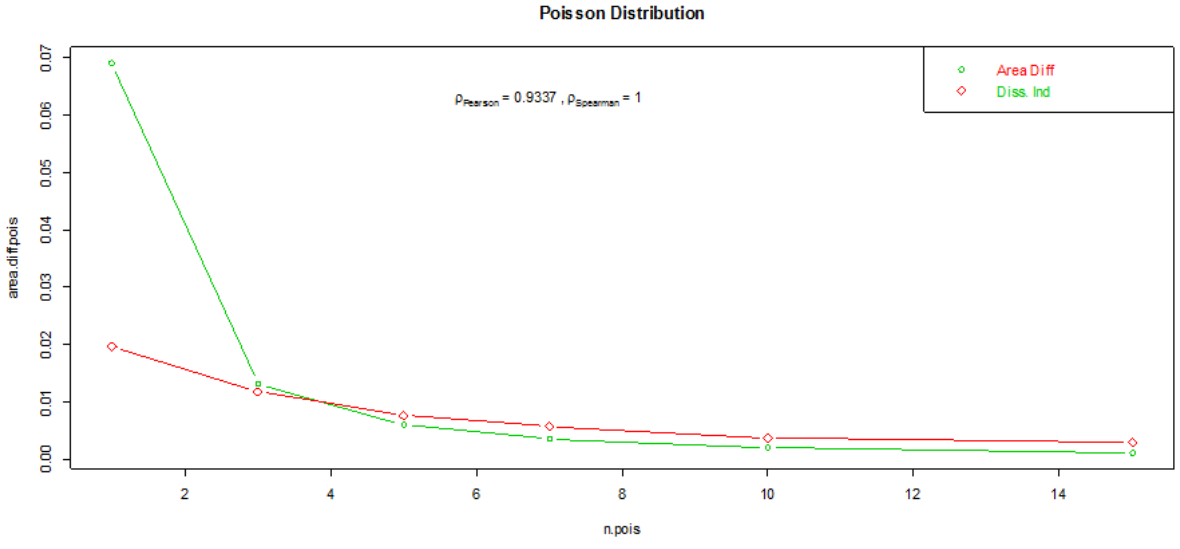
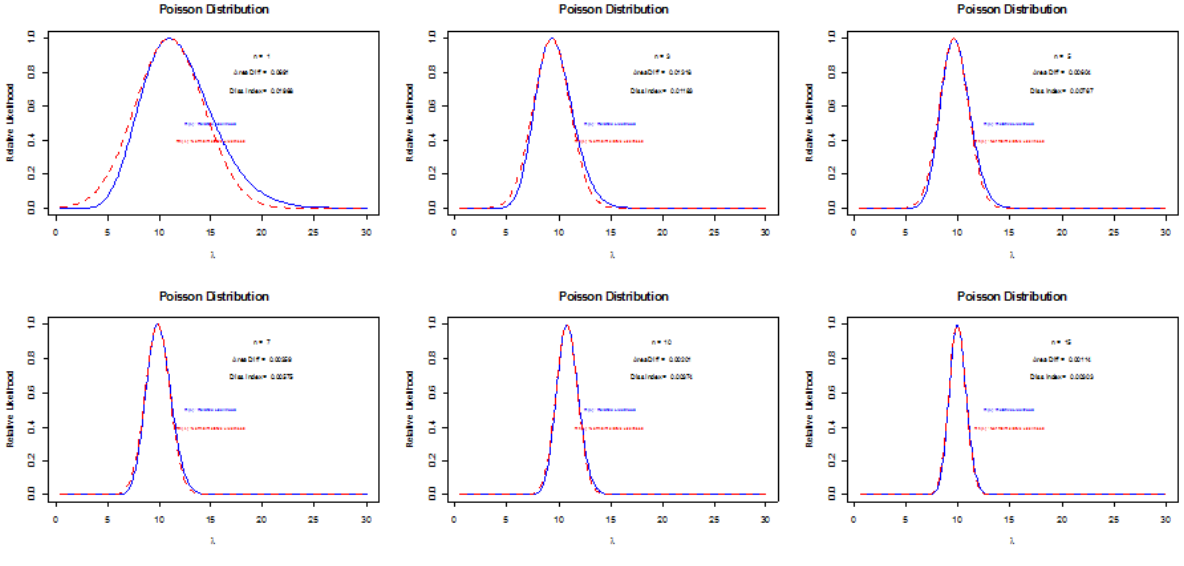


Figure 1.4: Poisson distribution – Observed and normal relative likelihood functions



Example 2: Weibull Distribution $(\gamma = 2, \beta = 6)$

The value of β was held fixed at 6 and thus γ is the parameter of interest.

Table 1.3: Weibull Distribution – Convergence measures

n	ΔR: Difference in Area	D: Dissimilarity Index
10	0.35706	0.34744
15	0.20736	0.23947
30	0.08872	0.16687
50	0.04498	0.10846
75	0.02406	0.07855
100	0.0158	0.0615

Figure 1.5: Weibull distribution – Change in values of Difference in Area and Dissimilarity Index

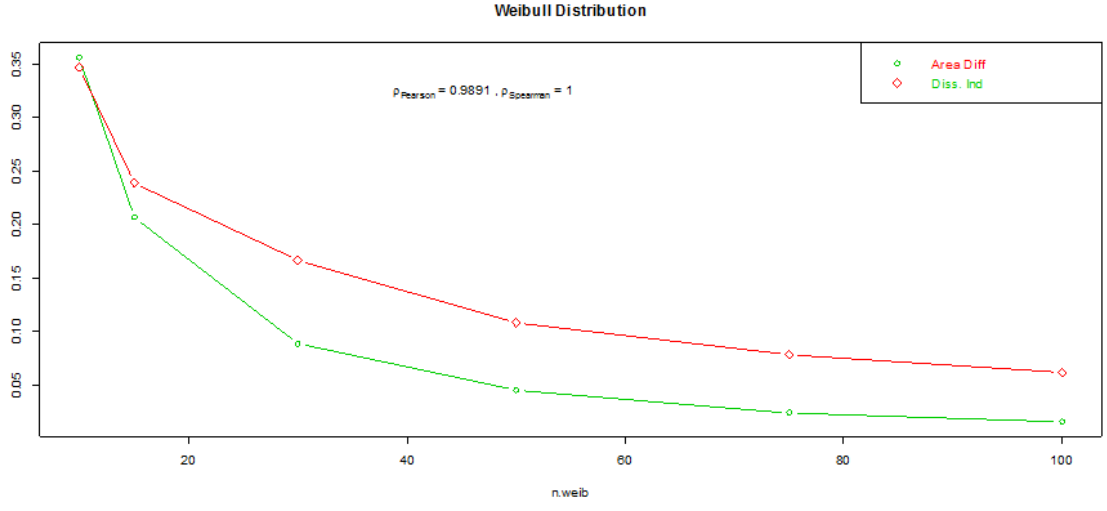
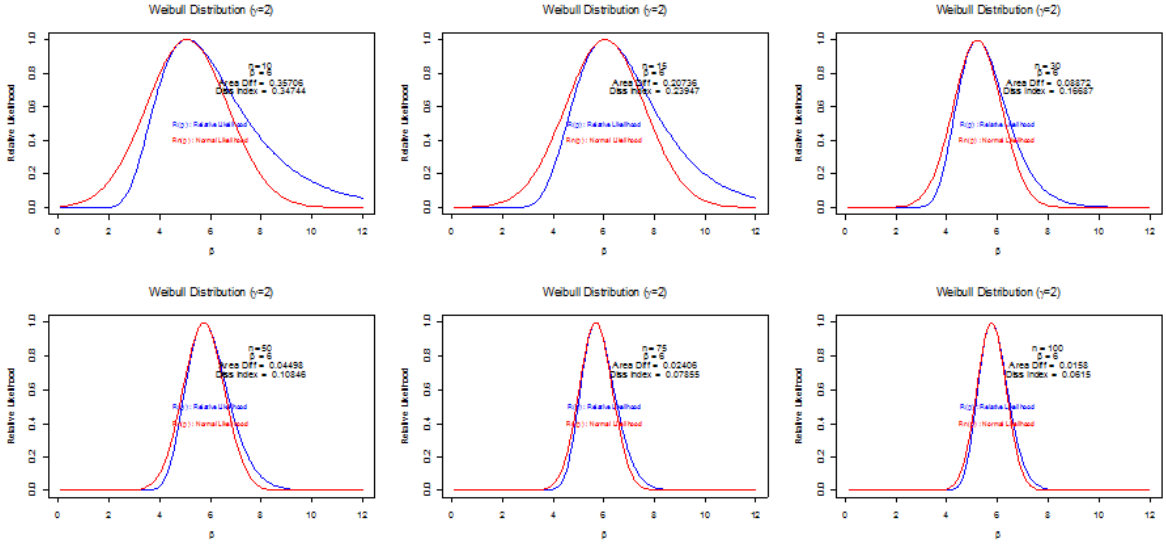


Figure 1.6: Weibull distribution – Observed and normal relative likelihood functions



Example 3: Exponential Distribution ($\beta = 6$)

Table 1.4: Exponential Distribution – Convergence measures.

n	ΔR: Difference in Area	D: Dissimilarity Index
20	0.16544	0.21671
30	0.09338	0.16252
50	0.04439	0.10903
70	0.02781	0.08202
85	0.02178	0.06883
100	0.01772	0.05941

Figure 1.7: Exponential distribution – Change in values of Difference in Area and Dissimilarity Index

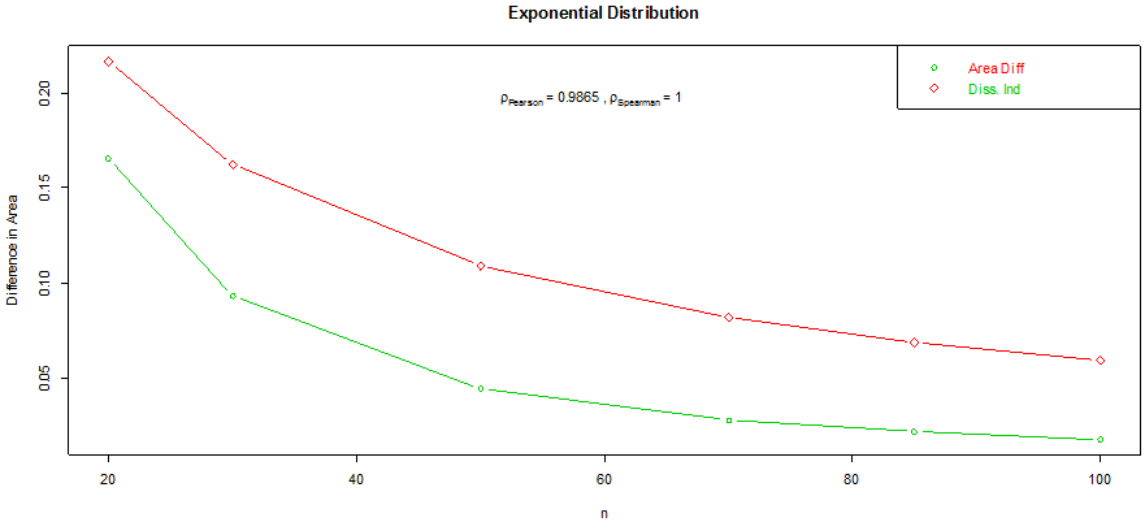
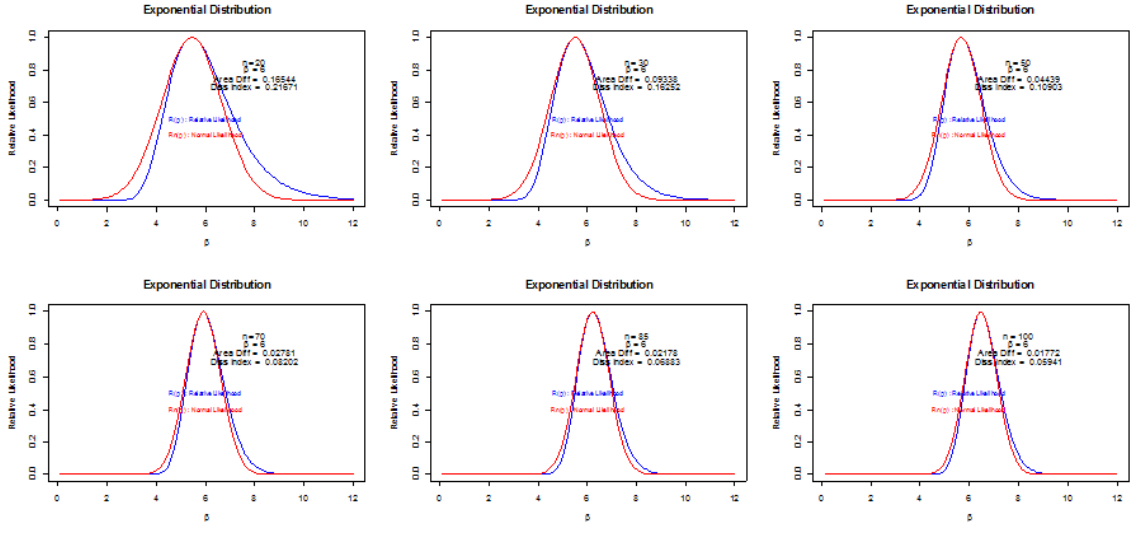


Figure 1.8: Exponential distribution – Observed and normal relative likelihood functions



1.4.2 Example Involving Real Data:

a) Data from Gibbons et al's book "Nonparametric Statistical Inference" (Gibbons and Chakraborti 2011).

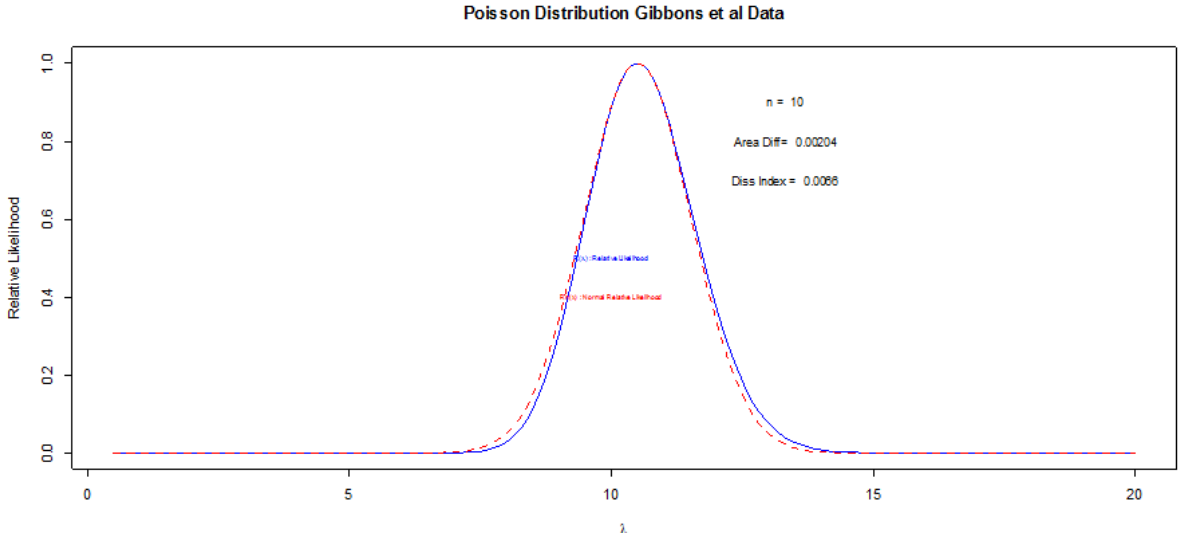
A group of 20 mice are allocated to individual cages randomly. The cages are then assigned randomly to two treatments namely control A and drug B. All animals were infected with tuberculosis. The number of days until the mice die is recorded (Table 1.5).

Table 1.5: Data from Gibbons et al's book

	Number of days until death	Mean	Variance
Control	5,6,7,7,8,8,8,9,12	7.778	3.944
Drug	7,8,8,8,9,9,12,13,14,17	10.5	10.944

For mice assigned to drug the mean and variance are roughly equal and the data is count data. So a Poisson model for it is a reasonable choice. Based on the proposed methods, the values of difference in area under curves ΔA and dissimilarity index were found to be D: 0.00204 and 0.0066 respectively. It indicates that the asymptotic normality approximation of the *MLE*s holds for the data (Drug) above.

Figure 1.9: Gibbons et al – Observed and normal relative likelihood functions.



b) Data from Breslow (1984)

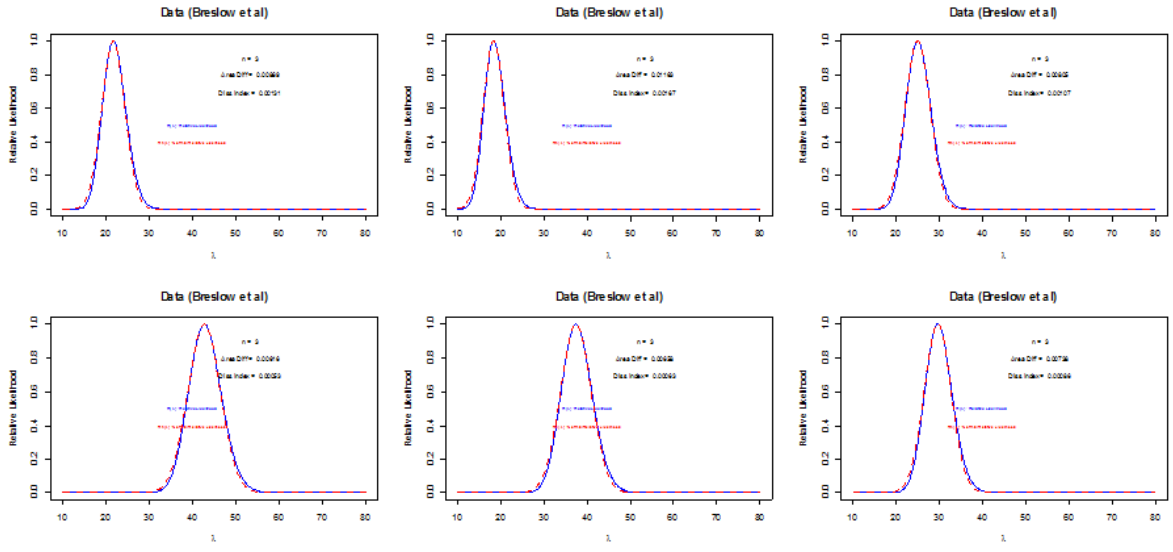
"Breslow (1984) analyses some mutagenicity assay data (Table 1.6) on salmonella in which three plates have been processed at each dose *i* of quinoline and the number of revertant colonies of TA98 Salmonella measured. A certain dose-response curve is suggested by theory." – Winbugs Example 1 volume (Breslow 1984).

The dataset is as follows:

Table 1.6: Data from Breslow.

Dose of quinoline (μg per plate)					
0	**10**	**33**	**100**	**333**	**1000**
15	16	16	27	33	20
21	18	26	41	38	27
29	21	33	69	41	42

Figure 1.10: Breslow et al – Observed and normal relative likelihood functions.



The two convergence measures suggest that the data at each dose level is large enough for the *MLE* to satisfy asymptotic normality (Figure 1.10).

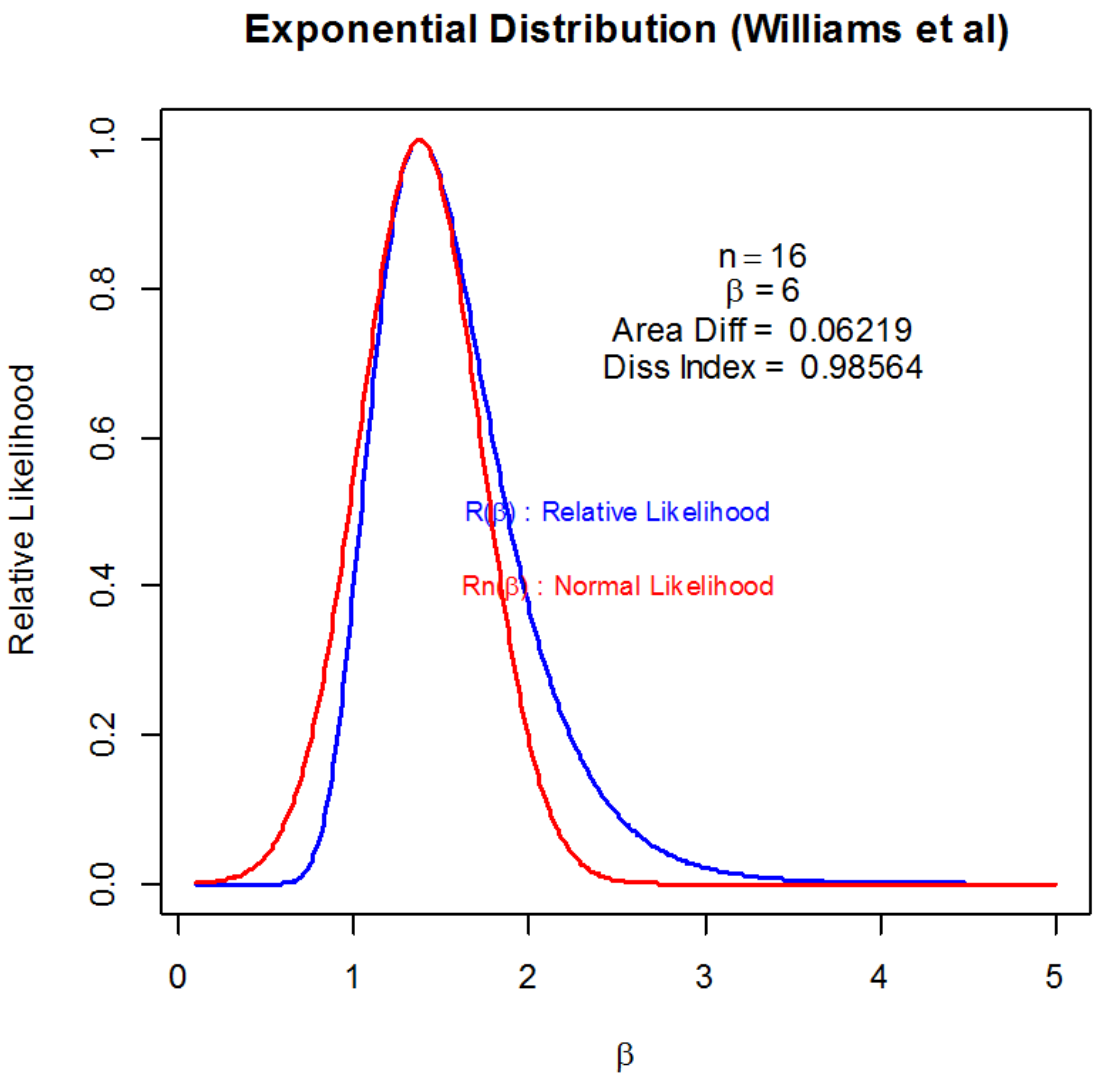
c) Data from Williams et al.

The following data was obtained from Williams et al (Williams, Bradshaw et al. 1995). The data is the weight (in grams) of dry seed in the stomach of each spinifex pigeon captured in desert.

0.457,3.751,0.238,2.967,2.509,1.384,1.454,0.818,0.335,1.436,1.603,1.309,0.201,0.530,2.144,0.834.

The plot of observed and normal RLFs together with the values of ΔA and D is in Figure 1.11.

Figure 1.11: Williams et al - Observed and normal relative likelihood functions



While the difference in area is small enough, the value of dissimilarity index seems fairly high. Table 1.7 shows that with larger samples (bootstrap) the dissimilarity index and difference in area both decrease.

Table 1.7: Data from Williams et al.

n	ΔR: Difference in Area	D: Dissimilarity Index
16	0.06219	0.98564
20	0.0508	0.80232
30	0.0269	0.58368
50	0.01266	0.38409
70	0.00702	0.29476
85	0.00508	0.25017

Figure 1.12: Williams et al – Change in values of Difference in Area and Dissimilarity Index

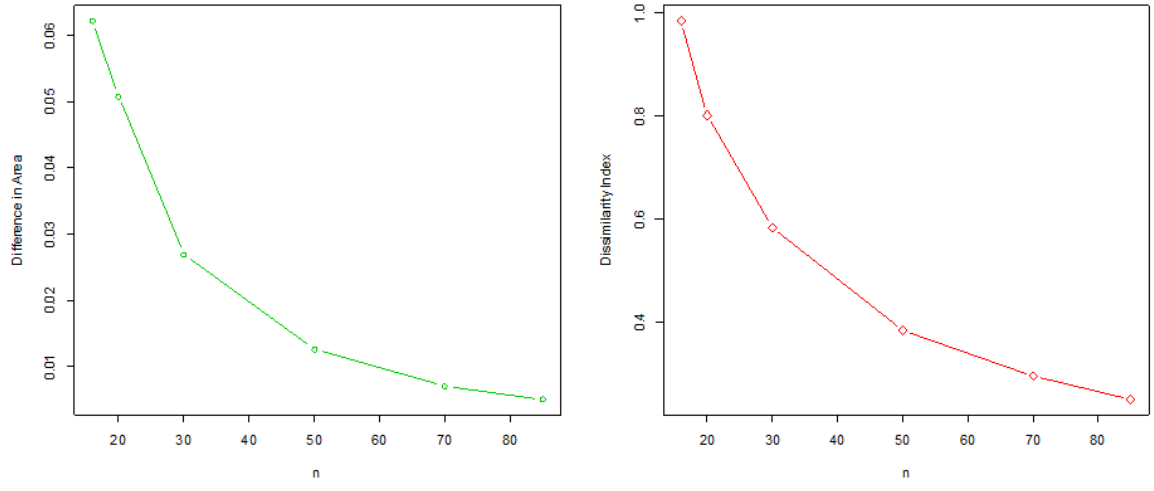
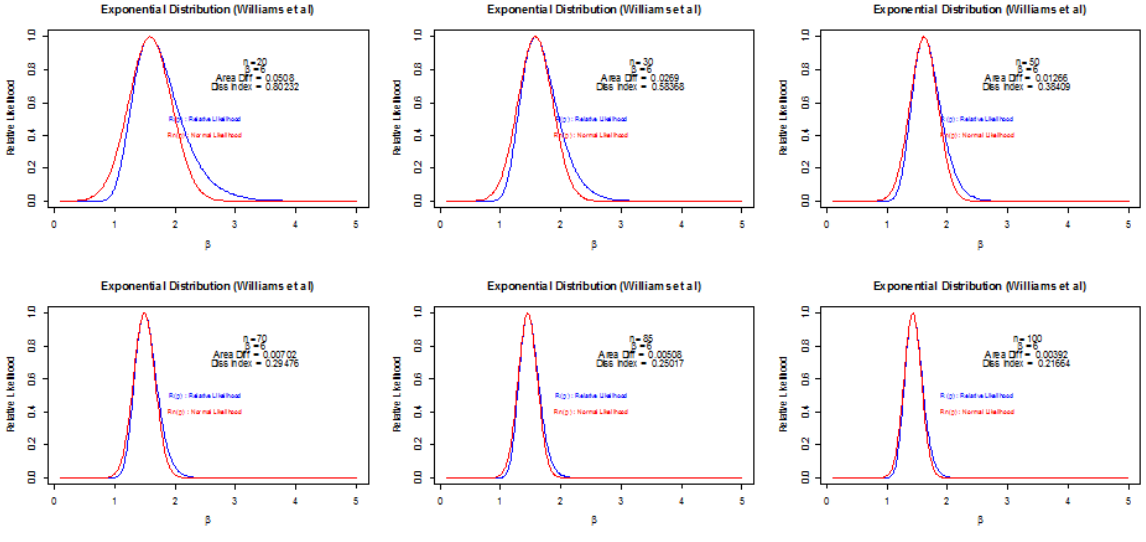


Figure 1.13: Williams et al – Observed and normal relative likelihood functions (bootstrapped samples)



1.5 Discussion

Our work discusses the issue of appropriateness of sample size required for asymptotic normality of *MLE*s to hold true. We essentially proposed two different diagnostic measures for this purpose viz. $\Delta R -$ difference in the area under the relative observed likelihood and relative asymptotic likelihood curves and $D -$ dissimilarity index which measures the shape of the curves. The simulated results show that different distributions have different threshold of ΔR and D. It gives an informal measure of convergence in real world. For example if we believe that the data at hand follows $Poisson(\lambda = 10)$ distribution we could compute ΔR and D and compare it with the tabulated values in table 1.2. If the $\Delta R_{computed}$ and $D_{computed}$ are close to the tabulated values for the given sample size, assumption of asymptotic normality of mle is reasonable.

The two measures of convergence were also applied to data from literature and bootstrap samples were used in assessing the convergence of RLFs. As seen from the simulated examples as well as the example from literature, the myth of "sample size of 30" can be far more than what is actually needed and the sample size requirements for satisfactory asymptotic convergence differ for different distributions. For example with Poisson $(\lambda = 10)$ distribution, it was seen that samples of sizes less than 10 show convincing convergence. Our future work is directed at generalizing these diagnostic measures to distributions taking into account parameters with in more than one dimension.

References

1. Fisher, R.A., *Inverse probability and the use of likelihood.* Proceedings of the Cambridge Philosophical Society, 1932. **28**: p. 257-261.
2. Fisher, R.A., *Probability likelihood and quantity of information in the logic of uncertain inference.* Proceedings of the Royal Society of London Series a-Containing Papers of a Mathematical and Physical Character, 1934. **146**(A856): p. 0001-0008.
3. Fisher, R.A., *Two new properties of mathematical likelihood.* Proceedings of the Royal Society of London Series a-Containing Papers of a Mathematical and Physical Character, 1934. **144**(A852): p. 0285-0307.
4. Fisher, R.A., *The likelihood solution of a problem in compounded probabilities.* Annals of Eugenics, 1941. **11**: p. 306-307.
5. Casella, G., Berger Roger L, *Statistical Inference*. Second ed. 2001, Belmont, California: Duxbury Press.
6. Altman, M., J. Gill, and M. McDonald, *Convergence Problems in Logistic Regression*, in *Numerical Issues in Statistical Computing for the Social Scientist*. 2003, Wiley. p. 219-233.
7. Efron, B., *Maximum Likelihood and Decision Theory.* The Annals of Statistics, 1982. **10**(2): p. 340-356.
8. Efron, B. and D. Hinkley, *Assessing the Accuracy of the Maximum Likelihood Estimator: Observed Versus Expected Fisher Information.* Biometrika, 1978. **65**(3): p. 457-482.
9. Sprott, D.A. and Kalbflei.Jd, *Examples of Likelihoods and Comparison with Point Estimates and Large Sample Approximations.* Journal of the American Statistical Association, 1969. **64**(326): p. 468-&.
10. Bartholomew, D.J., *A problem in life testing.* Journal of Americal Statistical Association, 1957. **52**: p. 350-355.

11. Hall, P. *A short history of the bootstrap*. in *Acoustics, Speech, and Signal Processing, IEEE International Conference on*. 1994. IEEE.

12. Efron, B., *1977 Rietz Lecture - Bootstrap Methods - Another Look at the Jackknife.* Annals of Statistics, 1979. **7**(1): p. 1-26.

13. Hall, P., *On the Bootstrap and Confidence-Intervals.* Annals of Statistics, 1986. **14**(4): p. 1431-1452.

14. Hall, P., *Rate of Convergence in Bootstrap Approximations.* Annals of Probability, 1988. **16**(4): p. 1665-1684.

15. Hall, P., *On Symmetric Bootstrap Confidence-Intervals.* Journal of the Royal Statistical Society Series B-Methodological, 1988. **50**(1): p. 35-45.

16. Hall, P., *On Efficient Bootstrap Simulation.* Biometrika, 1989. **76**(3): p. 613-617.

17. Hall, P. and S.R. Wilson, *Bootstrap Hypothesis-Testing - Reply.* Biometrics, 1992. **48**(3): p. 970-970.

18. Leger, C., D.N. Politis, and J.P. Romano, *Bootstrap Technology and Applications.* Technometrics, 1992. **34**(4): p. 378-398.

19. Singh, K., *Jackknife, Bootstrap and Other Resampling Methods in Regression-Analysis - Discussion.* Annals of Statistics, 1986. **14**(4): p. 1328-1330.

20. Singh, K., *On the Asymptotic Accuracy of Efrons Bootstrap.* Annals of Statistics, 1981. **9**(6): p. 1187-1195.

21. Bickel, P.J. and D.A. Freedman, *Some Asymptotic Theory for the Bootstrap.* Annals of Statistics, 1981. **9**(6): p. 1196-1217.

22. Bickel, P.J. and D.A. Freedman, *Asymptotic Normality and the Bootstrap in Stratified Sampling.* Annals of Statistics, 1984. **12**(2): p. 470-482.

23. Freedman, D.A., *Jackknife, Bootstrap and Other Resampling Methods in Regression-Analysis - Discussion.* Annals of Statistics, 1986. **14**(4): p. 1305-1308.

24. Carlstein, E., et al., *Matched-block bootstrap for dependent data.* Bernoulli, 1998. **4**(3): p. 305-328.

25. Kullback, S. and R.A. Leibler, *On Information and Sufficiency.* Annals of Mathematical Statistics, 1951. **22**(1): p. 79-86.

26. Gersch, W., et al., *A Kullback Leibler Nearest Neighbor Rule Classification of Eegs - the Eeg Population Screening Problem, an Anesthesia Level Eeg Classification Application.* Computers and Biomedical Research, 1980. **13**(3): p. 283-296.

27. Kon'kov, E.A., et al., *Application of the Kullback-Leibler measure for estimating the instants of a change in the statistical properties of a binary Markovian process.* Journal of Communications Technology and Electronics, 2007. **52**(12): p. 1350-1354.

28. Kubokawa, T. and H. Tsukuma, *Estimation in a linear regression model under the Kullback-Leibler loss and its application to model selection.* Journal of Statistical Planning and Inference, 2007. **137**(7): p. 2487-2508.

29. Rodrigues, J., *The Kullback-Leibler Approximation of the Marginal Posterior Density - an Application to the Linear Functional-Model.* Communications in Statistics-Theory and Methods, 1992. **21**(10): p. 2861-2868.

30. Johnson, D.H. and S. Sinanovic, *Symmetrizing the kullback-leibler distance.* IEEE Transactions on Information Theory, 2001. **1**(1): p. 1-10.

31. Bhattacharyya, A., *On a measure of divergence between two statistical populations defined by their probability distribution.* Calcutta Mathematical Society, 1943. **35**: p. 99-110.

32. Chaudhuri, G., J.D. Borwankar, and P.R.K. Rao, *Bhattacharyya Distance Based Linear Discriminant Function for Stationary Time-Series.* Communications in Statistics-Theory and Methods, 1991. **20**(7): p. 2195-2205.

33. Chen, K., et al., *Vessel attachment nodule segmentation using integrated active contour model based on fuzzy speed function and shape-intensity joint Bhattacharya distance.* Signal Processing, 2014. **103**: p. 273-284.

34. Jain, A.K., *Estimate of Bhattacharyya Distance.* Ieee Transactions on Systems Man and Cybernetics, 1976. **6**(11): p. 763-766.

35. Schweppe, F.C., *On Bhattacharyya Distance and Divergence between Gaussian Processes.* Information and Control, 1967. **11**(4): p. 373-&.

36. Atkinson, K.E., *An Introduction to Numerical Analysis*. 2nd ed. 1989: John Wiley & Sons.

37. Sprott, D.A., *Normal Likelihoods and Their Relation to Large Sample Theory of Estimation.* Biometrika, 1973. **60**(3): p. 457-465.

38. Gibbons, J.D. and S. Chakraborti, *Nonparametric Statistical Inference*, ed. N. Balakrishnan and S.W. R. 2011: CRC Press, Taylor and Francis Group.

39. Breslow, N.E., *Extra-Poisson variation in log-linear models.* Applied Statistics, 1984. **33**: p. 38-44.

40. Williams, J.B., D. Bradshaw, and L. Schmidt, *Field metabolism and water requirements of spinifex pigeons (Geophaps plumifera) in Western Australia.* Australian Journal of Zoology, 1995. **43**(1): p. 1-15.

Chapter 2: Likelihood Transformation and Information Based Approach to Clustering

2.1 Introduction

For a sample of iid observations $X_i's, \boldsymbol{X} = (X_1, \dots, X_n)$ with pdf $f(\boldsymbol{x}|\theta)$, the likelihood function, introduced and established by Fisher (Fisher 1932, Fisher 1934, Fisher 1934, Fisher 1941, Casella 2001) is formally defined as:

$$L(\theta|\boldsymbol{x}) = f(\boldsymbol{x}|\theta) = \prod_{i=1}^{n} f(x_i|\theta).$$

The likelihood function reflects the model-data combination being used to investigate the research question of interest. The value of the parameter that maximizes the likelihood function is the MLE of the parameter. Thus if $\hat{\theta}$ maximizes the above likelihood function, then $\hat{\theta}$ is the mle of θ. For most well-known families of distributions, *MLEs* can be computed analytically, numerically or graphically.

RLFs are the likelihood functions re-scaled by their mode. Thus if $\hat{\theta}$ is the mle of θ, then the RLF of θ is defined as:

$$R(\theta) = \frac{L(\theta|\boldsymbol{x})}{L(\hat{\theta}|\boldsymbol{x})}$$

It is obvious that unlike likelihood functions, RLFs are bounded above by one. Note that the initial observed likelihood and its relative version have the same *MLE* and Fisher Information values. Using a relative scale makes a set of likelihood functions more comparable as they all have a maximum value of one.

While likelihood functions are virtually omnipresent in statistics, RLFs have attracted little application. Sprott et al (1969) proposed the application of RLFs to examine the asymptotic behavior of the mle and suggested examining RLFs based on the observed data as well as the

large sample distribution of the *MLE* before making inferences based on the asymptotic distribution of *MLE* (Sprott and Kalbfleisch 1969).

Recently bootstrap and simulation approaches were used to estimate the sample sizes that would allow the distribution of the *MLE* to attain asymptotic normality (Bimali and Brimacombe 2015). The proximity of the observed RLF to its asymptotic form was examined using two criteria – (i) difference in area under the two RLFs and (ii) local dissimilarity in the shape of the two RLFs. It was argued that for a given sample size, if the difference in the area under the two RLFs and the dissimilarity index between them are both close to 0, the asymptotic approximation of *MLE* is satisfactorily achieved (Bimali and Brimacombe 2015).

The application of the observed likelihood function in the context of multivariate empirical data analysis has been fairly limited. Recent spatial clustering algorithm proposed by Kerby et al makes use of pairwise defined likelihood functions together with a grouping algorithmic hierarchical approach (Kerby, Marx et al. 2007). The idea is to start with n clusters (each observation forming its own cluster as in hierarchical clustering) followed by computation of pair-wise likelihoods for all possible pairs of two clusters and eventually merging the pairs with largest likelihood thereby producing $n - 1$ clusters. The process is repeated until all the observations are grouped together into a single cluster. Likelihood based hierarchical clustering has also been pursued by Castro et al (Castro and Nowak 2003). However to the best of our knowledge no clustering algorithm has been proposed based on similarities or dissimilarities between the information based properties of a set of likelihood functions.

The approach here uses a set of likelihood functions evaluated across a range of θ_i values to obtain a data matrix whose columns are the evaluated likelihood functions on a relative scale. This allows us to directly apply standard multivariate data-analytic clustering techniques such as

k-means, centroid to investigate information based similarities in the likelihood transformed data.

The proposed approach is justified as follows: for each observation in the dataset, a set of (relative) likelihood functions are constructed, viewed as a transformation of the original data. Noting that the Fisher Information reflects the local curvature of a log-likelihood function and summarizes the estimation related accuracy present in an individual likelihood, this can be used as a weight for each respective likelihood function. The weighted RLF is then evaluated at different values of the parameter to obtain a data matrix. The data matrix thus constructed can be examined using multivariate data analytic clustering algorithms such as k-means, centroid or PCA (Johnson and Wichern 2002, Rencher 2002). Initially we assume the set of likelihoods are independent. This is a new scale upon which to cluster studies in a direct and interpretable manner. Depending on the structure of the data, and the appropriateness of the density function chosen, the method developed here can be applied to datasets with several measured variables on each observation.

The paper is organized as follows. In section 2.2 we discuss the likelihood based distance matrix and its properties. We also define and discuss the Fisher Information and its use as a weighting element in various definitions of multivariate distance. Section 2.3 and 2.4 present results of the method applied to simulated datasets. The analysis was done using R (version 3.2.0) (Team 2015).

2.2 Method

Let us consider data matrix $\boldsymbol{X} = (\boldsymbol{x_1} \quad \dots \quad \boldsymbol{x_n})'$ where $\boldsymbol{x_i} = (x_{i1} \quad \cdots \quad x_{iT_i})$ $x_{i1}, \dots, x_{iT_i}$ are iid observations with pdf $f_i(\boldsymbol{x_i}|\theta_i); j = 1, \dots, T_i$ and $\boldsymbol{\theta} = (\theta_1 \quad \dots \quad \theta_n)$

We assume that $\theta_i's$ share the same support. Thus for each set of θ_i values we can construct likelihood functions reflecting assumed pdf which give rise to n likelihood functions collected as columns in the data matrix $\boldsymbol{X}$.

$$\boldsymbol{L_X}(\boldsymbol{\theta}) = (L_{\boldsymbol{x_1}}(\theta_1) \quad \dots \quad L_{\boldsymbol{x_n}}(\theta_n))'$$

The RLF can then be constructed as:

$$\boldsymbol{R_X}(\boldsymbol{\theta}) = (R_{\boldsymbol{x_1}}(\theta_1) \quad \dots \quad R_{\boldsymbol{x_n}}(\theta_n))'$$

with $R_{\boldsymbol{x_i}}(\theta_i) = \frac{L_{\boldsymbol{x_i}}(\theta_i)}{L_{\boldsymbol{x_i}}(\hat{\theta}_i)}$. Note that the Fisher information, by definition is the same for both the initial and relatively re-weighted likelihood function. The relative aspect however is useful in graphical summaries related to the clustering and is maintained here.

To improve the assessment of similarity across the set of evaluated likelihoods, the Fisher information matrix for each observation vector can be used as a weight and we have;

$$w_{\boldsymbol{x_i}}(\hat{\theta}) = I(\hat{\theta}) \text{ where } I(\theta) = E\left(\left(\frac{\partial}{\partial\theta} log(L(\theta|\boldsymbol{x_i}))\right)^2\right)$$

For exponential families with iid observations, note that the Fisher Information matrix can be simplified to $I(\theta) = -E\left(\frac{\partial^2}{\partial\theta^2}\log l(\theta|\boldsymbol{x}_i)\right)$ and we can construct a matrix $\boldsymbol{P_X}$ with rows containing the weighted RLFs evaluated at different values of θ_i .

$$\boldsymbol{P_X} = \begin{pmatrix} w_{\boldsymbol{x_1}}(\hat{\theta}) & \cdots & w_{\boldsymbol{x_1}}(\hat{\theta}) \\ \vdots & \ddots & \vdots \\ w_{\boldsymbol{x_n}}(\hat{\theta}) & \cdots & w_{\boldsymbol{x_n}}(\hat{\theta}) \end{pmatrix} \circ \begin{pmatrix} R_{\boldsymbol{x_1}}(\theta_1) & \cdots & R_{\boldsymbol{x_1}}(\theta_k) \\ \vdots & \ddots & \vdots \\ R_{\boldsymbol{x_n}}(\theta_1) & \cdots & R_{\boldsymbol{x_n}}(\theta_k) \end{pmatrix}$$

where, $w_{x_i}(\hat{\theta})$ is the Fisher Information evaluated at the *MLE*. $R_{x_i}(\theta_j)$ is the value of the RLF for x_i evaluated at θ_j and $\circ$ is the Hadamard product operator between the two matrices. The

matrix $\boldsymbol{P_X}$ can be subjected to various standard clustering algorithms to explore for patterns and clusters in the data matrix $\boldsymbol{X}$. In situations where the expected Fisher information is analytically intractable, the observed Fisher Information can be used as an estimate of expected Fisher Information (Efron and Hinkley 1978).

As the Fisher Information reflects the local curvature of the log likelihood or relative log likelihood function, the approach here clusters the set of likelihoods based on the level of accuracy in their likelihood functions in regard to estimating the common parameter θ. This approach jointly reflects the observed data values and sample size.

2.3 Application: Simulation Based on Rates of Salmonellosis in Kansas

The dataset used in the analysis is simulated based on the rates of Salmonellosis− a common form of food poisoning, in Kansas in 2012 (Data.Cms.gov 2014). Salmonellosis is an illness characterized by sudden onset of fever, headache, diarrhoea, abdominal pain, nausea, and sometimes vomiting. The rate of Salmonellosis across different counties in Kansas was reported in "Reportable Infectious Diseases in Kansas 2012 Summary" published by Kansas Department of Health and Environment, Bureau of Epidemiology and Public Health Informatics. The rates (per 100,000) across different counties were used to simulate 30 counts from Poisson distribution (Table 2.1). We included only those counties where at least 3 counts of the disease were reported. The analysis will be carried out using the simulated data given the low counts and issues of subject identifiability in the respective counties.

If $\boldsymbol{x_i} \sim Poi(\lambda_i), i = 1, \dots, n$, then the RLFs and observed Fisher Information can be shown to be:

$$R(\lambda|\boldsymbol{x}) = e^{n(\hat{\lambda}-\lambda)}\left(\frac{\lambda}{\hat{\lambda}}\right)^{n\hat{\lambda}}$$

$$I(\hat{\lambda}) = \frac{n}{\hat{\lambda}}$$

Thus the weighted RLF for each county based on the simulated data is:

$$w(\lambda|\boldsymbol{x}) = I(\hat{\lambda}) \times R(\lambda|x) = \frac{n}{\hat{\lambda}} \times e^{n(\hat{\lambda}-\lambda)}\left(\frac{\lambda}{\hat{\lambda}}\right)^{n\hat{\lambda}}$$

The related likelihood based data matrix $\boldsymbol{P_X}$ is constructed as described in Section 2.

Table 2.1: Simulated Data from Poisson distribution using observed rate of Salmonellosis for different counties in Kansas.

County	Count	Rate	Simulated Counts
Allen	5	37.5	33,43,31,41,37,40,38,50,41,29,38,39,34,41,40,35,45,33,41,37,39,32,51,50,29,42,33,37,33,44
Anderson	4	50.5	47,42,42,50,53,44,60,60,43,54,41,54,56,39,48,66,50,54,56,56,52,46,39,49,50,46,50,38,44,50
Barton	7	25.4	29,18,34,26,29,26,29,32,21,22,27,31,27,32,19,25,26,24,30,28,19,20,24,21,31,26,23,24,26,21
Butler	5	7.6	7,6,7,15,5,4,6,11,8,6,9,2,7,5,6,9,5,4,6,5,7,9,7,8,9,3,8,6,6,11
Cherokee	5	23.6	25,20,26,22,24,27,21,26,25,23,26,24,16,28,26,22,28,19,20,18,24,23,24,20,22,26,24,23,20,31
Clay	4	46.9	49,39,50,38,50,38,51,39,54,42,47,48,61,40,49,46,54,46,59,49,50,47,43,51,53,48,43,44,48,35
Cloud	4	42.6	35,35,43,43,36,37,46,42,43,37,42,38,51,50,49,41,44,40,40,44,45,42,37,38,54,44,40,40,51,40
Cowley	12	33.1	44,48,30,34,21,21,27,43,33,28,33,26,28,29,33,36,27,26,38,33,28,33,22,30,38,33,34,32,35,37
Crawford	7	17.8	19,11,13,17,10,19,8,11,24,16,25,15,23,14,14,25,20,13,16,18,17,14,18,26,18,14,14,18,18,13
Dickinson	6	30.4	26,26,34,38,32,34,27,23,29,40,31,35,35,29,29,42,33,43,31,31,29,20,21,30,41,35,24,27,24,31
Douglas	23	20.4	23,21,22,18,17,22,19,15,24,19,17,14,24,13,23,26,18,19,23,18,28,23,25,19,23,15,18,19,34,18
Edwards	3	100.7	97,93,89,105,106,105,91,118,113,97,100,104,104,109,89,94,97,90,88,88,103,83,88,110,123,95,88,108,100,99
Ellis	7	24.1	29,25,18,27,25,31,25,28,32,24,21,26,29,32,20,22,29,26,26,19,35,35,32,19,24,35,19,22,24,27
Ellsworth	3	46.2	45,48,40,54,47,39,50,55,48,38,51,52,59,39,49,50,49,36,59,44,38,63,51,33,53,44,40,49,49,58
Finney	4	10.2	8,8,11,9,8,10,12,7,8,15,8,13,11,17,11,15,14,8,11,8,11,6,12,16,12,9,15,10,14,15
Ford	5	14.4	14,9,14,11,12,17,13,12,12,12,17,18,22,11,16,17,12,23,17,18,20,10,17,16,20,20,15,21,11,18
Franklin	8	30.9	32,28,29,29,23,23,36,29,23,28,29,30,39,36,30,22,27,38,38,31,39,40,29,31,33,24,33,18,35,36
Geary	9	23.7	15,28,29,28,29,22,27,22,26,20,21,19,17,26,26,19,21,30,27,25,18,23,22,32,14,22,21,26,21,21
Graham	10	387.9	375,389,386,358,404,364,402,384,376,390,400,353,415,407,381,338,392,393,373,396,351,366,406,373,384,356,378,373,404,407
Harvey	9	25.8	20,31,24,18,30,24,25,31,28,22,22,24,19,25,24,22,31,28,24,21,19,28,29,24,19,24,21,31,27,21
Jackson	5	37.2	41,44,40,44,32,40,41,34,31,38,37,42,34,47,38,34,41,29,39,38,41,35,28,44,39,43,42,31,38,37
Jefferson	5	26.4	30,19,18,33,23,28,21,22,27,22,35,29,33,28,21,26,22,29,34,27,26,28,20,29,17,16,26,30,28,19
Johnson	69	12.3	14,16,19,14,8,14,17,10,10,14,12,8,11,14,17,17,9,13,11,6,17,16,10,8,13,9,15,13,9,12
Labette	3	14.1	10,15,20,16,22,14,12,14,14,12,17,21,8,22,17,10,12,13,13,12,10,13,15,14,8,12,8,15,15,15
Leavenworth	16	20.6	21,22,19,21,20,11,18,23,22,19,19,18,14,16,24,17,28,21,31,17,24,26,26,20,25,18,22,23,21,15
Lincoln	9	283.6	265,303,268,259,304,250,327,281,291,273,297,267,310,291,290,295,277,289,290,291,282,293,266,273,274,286,291,303,292,278
Linn	3	31.8	34,32,30,30,36,26,38,38,32,36,27,27,32,32,27,26,25,37,35,31,34,29,43,33,37,32,35,33,30,26
Lyon	3	8.9	13,7,7,9,10,9,4,8,15,16,10,10,9,9,7,8,13,16,19,10,14,14,11,13,10,10,10,10,4,8
Marion	3	24.3	28,28,30,33,17,19,22,15,22,24,19,25,23,25,27,19,23,21,24,22,23,26,25,27,33,29,27,33,20,20
McPherson	5	17	12,22,17,18,28,25,16,20,16,22,20,17,13,12,14,17,16,14,16,10,19,13,19,19,14,18,21,15,20,19
Miami	5	15.3	18,12,17,20,10,11,17,15,22,15,13,14,14,14,9,17,22,19,14,11,13,9,13,15,19,14,25,14,11,10
Mitchell	4	62.9	54,84,61,69,64,58,55,49,78,54,61,58,56,71,66,67,62,76,67,52,46,65,62,70,51,60,75,66,63,63
Montgomery	4	11.6	9,10,14,9,13,11,12,9,9,12,10,11,15,12,14,8,15,14,10,17,12,12,16,9,13,14,10,8,10,15
Nemaha	8	79	80,80,63,71,78,62,86,76,73,84,64,102,91,76,73,92,72,70,79,73,84,88,60,96,89,77,93,72,78,78
Osage	3	18.6	18,17,26,24,12,19,14,18,21,22,23,12,12,26,16,12,19,14,22,14,17,25,20,22,25,24,20,13,14,23
Pawnee	4	57.7	62,51,65,60,46,52,50,61,53,67,64,59,60,62,68,65,58,55,63,56,81,57,67,50,58,59,70,45,53,56
Phillips	3	54.4	52,50,61,59,44,50,48,50,53,62,59,56,49,50,68,54,38,50,53,44,49,44,59,60,57,55,46,44,60,49
Pottawatomie	4	17.9	18,19,18,24,21,11,19,20,18,26,18,18,9,21,17,23,13,13,17,17,12,12,15,21,13,14,14,15,16,22
Reno	10	15.5	11,13,19,16,21,21,8,17,15,21,20,15,16,17,16,11,14,16,20,22,15,12,13,22,18,20,12,12,27,19
Rice	3	30	38,39,24,21,32,43,28,27,33,33,30,31,33,39,18,28,30,32,28,25,28,29,37,34,23,37,30,28,30,25
Riley	15	19.9	21,20,19,22,23,23,18,25,23,15,21,11,18,21,20,24,17,20,28,19,26,19,13,21,17,18,25,20,21,26
Saline	10	17.9	11,17,15,14,15,17,16,20,28,20,23,18,21,15,25,24,24,33,18,12,20,19,15,16,19,14,15,17,12,16
Sedgwick	49	9.7	10,11,12,8,8,10,7,13,9,9,16,7,7,10,12,7,12,10,15,9,13,9,9,10,7,10,9,7,11,8
Shawnee	32	17.9	21,14,25,17,14,12,17,17,12,13,12,22,14,19,16,13,26,19,15,29,19,22,19,21,19,23,16,17,16,15
Sumner	4	16.9	22,17,14,16,22,21,15,20,17,20,14,13,14,20,17,17,17,22,16,26,16,13,19,19,20,10,19,17,11,18
Washington	4	69.5	68,65,84,62,72,63,77,78,68,79,75,76,69,69,55,79,52,59,67,50,79,82,51,64,72,69,61,67,68,79
Wichita	3	133	132,128,137,130,133,128,151,150,109,137,125,149,137,133,118,136,118,115,152,136,122,119,138,141,113,131,144,142,124,130
Wyandotte	24	15.1	19,15,15,17,14,24,16,18,11,9,19,15,19,16,21,12,8,13,9,15,19,11,15,14,15,9,14,14,9,9

The plot of the weighted RLF is provided in Figure 2.1.

The observed rate for Salmonellosis was found to range from 7.6 to 387.9. The range for λ was thus chosen to be $(0, 400)$ so as to include the observed range of rates. Note the rate parameter for Poisson distribution is non-negative; thus we have 0 as the default lower bound. For the construction of $\boldsymbol{P_X}$, the weighted likelihood functions is evaluated between 0 and 400 in the increment of 0.1. This gave adequate smoothness to the likelihood functions evaluated.

Figure 2.1: Weighted relative likelihood function based on simulated counts of Salmonellosis for different counties in Kansas. The horizontal line represents a height of 1 - height of relative likelihood function, and thus the vertical deviation represents weight.

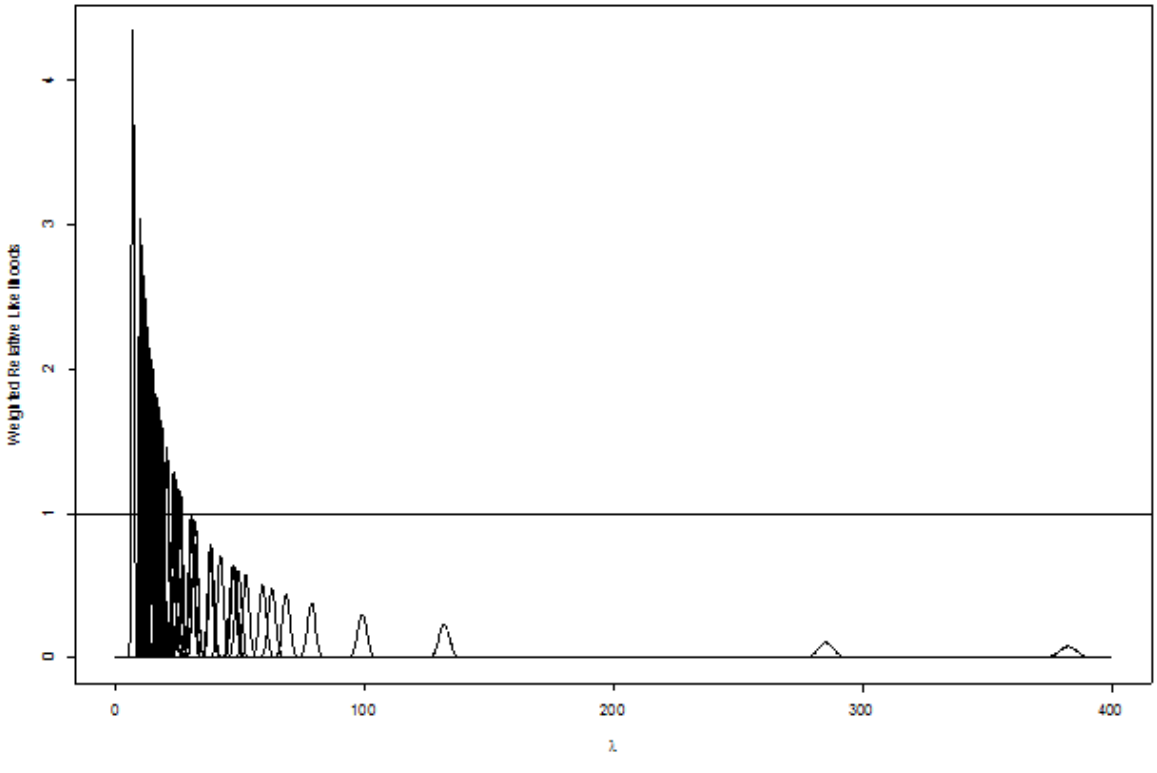
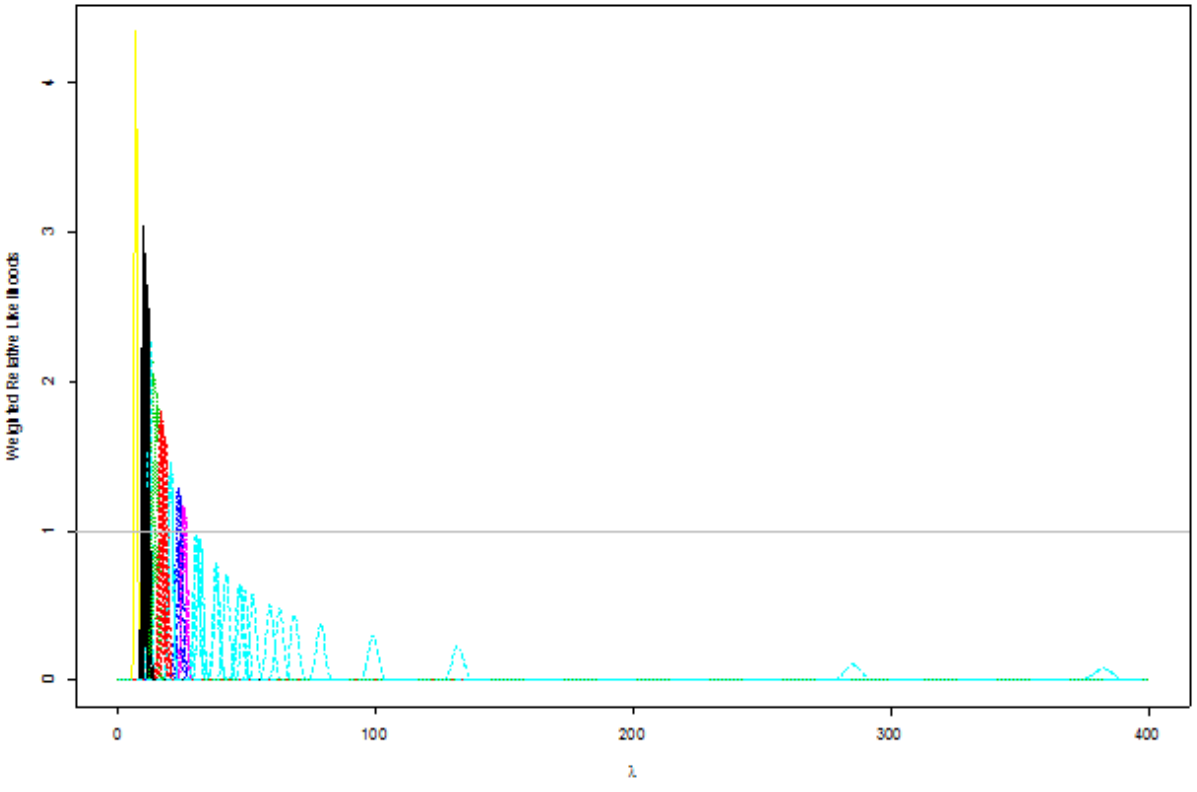


Table 2.2: Counties clustered based on their similarity across weighted relative likelihood functions.

Cluster	Counties
1	Finney, Lyon, Montgomery, Sedgwick
2	Crawford, McPherson, Osage, Pottawatomie, Reno, Saline, Shawnee, Sumner
3	Ford, Labette, Miami, Wyandotte
4	Cherokee, Geary, Harvey, Marion
5	Allen, Anderson, Clay, Cloud, Cowley, Dickinson, Douglas, Edwards, Ellsworth, Franklin, Graham, Jackson, Johnson, Leavenworth, Lincoln, Linn, Mitchell, Nemaha, Pawnee, Phillips, Rice, Riley, Washington, Wichita
6	Barton, Ellis, Jefferson
7	Butler

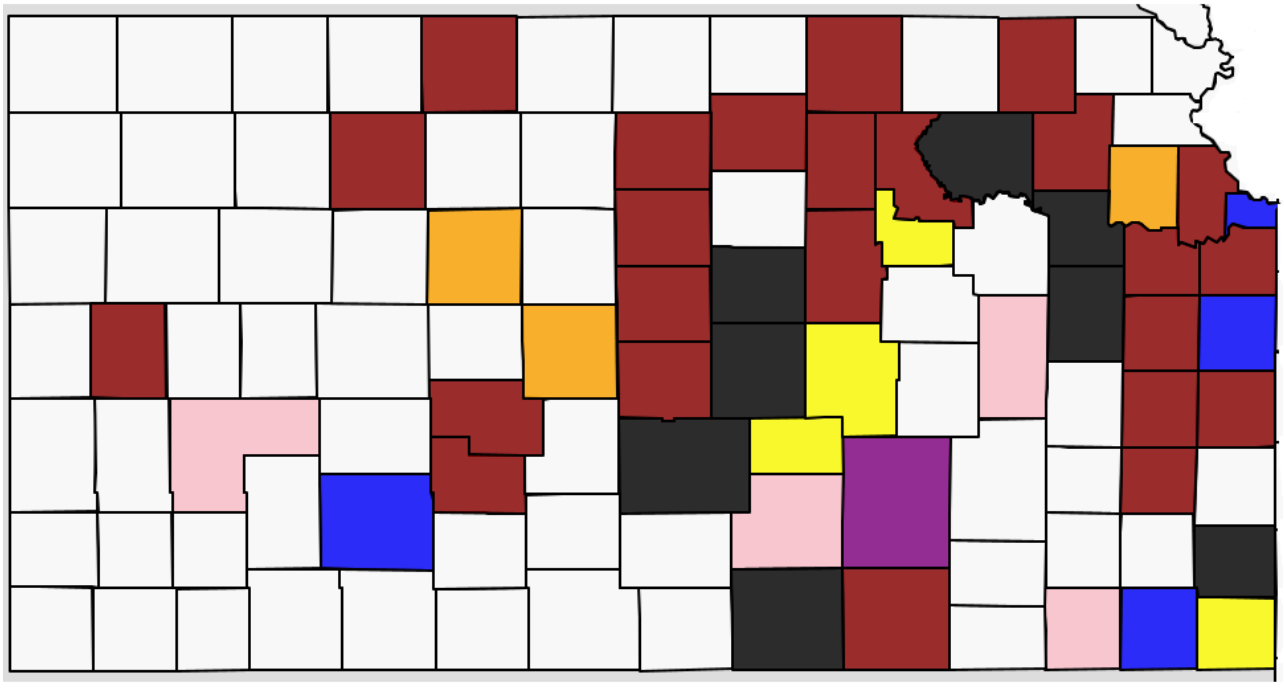
The k-means clustering algorithm was also applied to the matrix $\boldsymbol{P_X}$. The numbers of clusters were chosen based on the within sum of square plot. The elbow was observed at 7 clusters. Thus the k-means clustering algorithm was carried for 7 clusters. Table 2.2 provides the clustering of counties. Figure 2.2 provides the plot of weighted RLFs colored by the clusters.

Figure 2.2: Weighted relative likelihood function based on simulated counts of Salmonellosis for different counties in Kansas.



The plot of weighted RLF for the counties reflects the variation in *MLEs* (mean count) as well as the variation in the observed Fisher Information across the counties. The variation in *MLE* is reflected along the x-axis and the variation in the height and curvatures of the weighted likelihood functions across the y-axis reflect the variation in the observed Fisher Information. The clustering of counties takes into account the similarity of Fisher Information as well as the proximity of the *MLEs*. As a whole they provide insight into the clustering patterns in the overall data where we transform the observed data onto a relative likelihood scale. Table 2.2 shows that observations with similar rates as well as information are grouped together. Cluster 5 contains counties with high rates of Salmonellosis; while cluster 7 contains Butler County which has the lowest rate. The plot of counties colored by their cluster assignment does not show obvious spatial distribution (Figure 2.3).

Figure 2.3: Map of Kansas with counties colored by their cluster assignment (White color represents counties not included in study).



2.4 Further Simulation Studies

The use of the weighted RLF based approach in clustering observations can further be assessed by simulating data from a set of known distributions, constructing weighted relative likelihoods and clustering the observations based on the methods described in section 2.2. The clusters thus formed can be compared to those underlying the simulation and used to assess the efficiency of the weighted RLF based approach.

The data were simulated from the Cauchy, Poisson, and t distributions. Two distributions (2 shift parameters in case of t-distribution) were used in generating each dataset. 30 observations were generated from each of the aforementioned distributions with different parameters (e.g. Poisson distribution with two rate parameters). Each observation consists of 15 values. Weighted RLFs were constructed for each observation and a matrix based on the resulting likelihood functions evaluated at different values of the parameter was constructed.

The k-means algorithm with 2 clusters was applied to the matrix and a misclassification rate was computed as the proportion of observations from one distribution clustered with observations from another distribution. The process was repeated 30 times and the average misclassification rate was computed. In addition k-means algorithm was also applied to the simulated data and the average misclassification rate was computed.

2.4.1 Simulation from Cauchy distribution $\left(f(x|\theta) = \frac{1}{\pi}\frac{1}{1+(x-\theta)^2}\right)$

The likelihood function is given by:

$$L(\theta|\boldsymbol{x}) = \frac{1}{\pi^n}\prod_i (1 + (x_i - \theta)^2)^{-1}$$

The *MLE* is computed numerically using the *optimize* function in R. The weight is computed as follows:

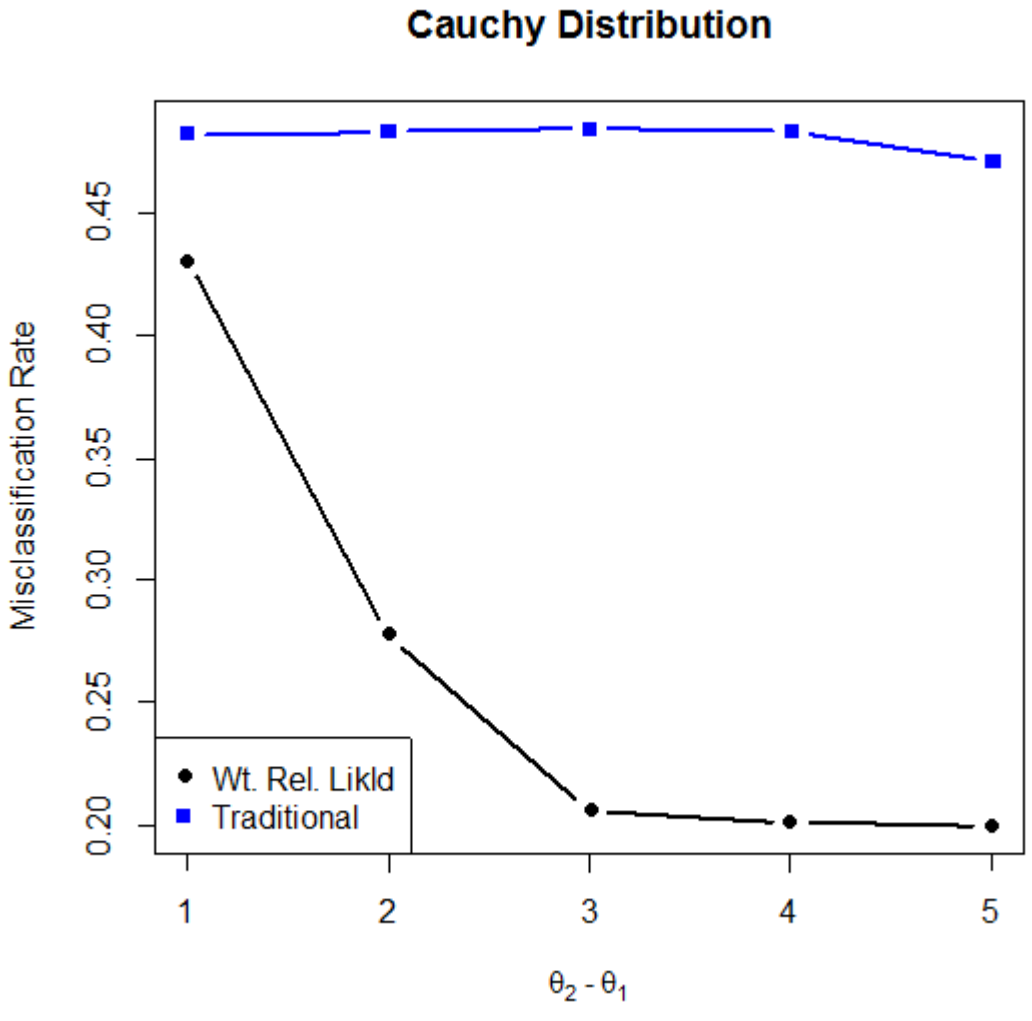
$$w(\theta) = E\left(\left(\frac{\partial}{\partial\theta} logL(\theta|\boldsymbol{x})\right)^2\right) = E\left(\left(2\sum_i \frac{(x_i-\theta)}{1+(x_i-\theta)^2}\right)^2\right) \approx 4\left(\sum_i \frac{(x_i-\hat{\theta})}{1+\left(x_i-\hat{\theta}\right)^2}\right)^2$$

The misclassification rates are provided in Table 2.3 and Figure 2.4.

Table 2.3: Cauchy distribution - Misclassification rates based on the k-means clustering applied to the matrix of values with evaluated weighted relative likelihood function and simulated data.

Cauchy Distribution	Wt. Relative Likld.	Simulated Data
$\theta_1 = 5, \theta_2 = 6$	0.431	0.481
$\theta_1 = 5, \theta_2 = 7$	0.278	0.483
$\theta_1 = 5, \theta_2 = 8$	0.206	0.484
$\theta_1 = 5, \theta_2 = 9$	0.201	0.483
$\theta_1 = 5, \theta_2 = 10$	0.200	0.471

Figure 2.4: Misclassification rates versus difference in θ parameters for data simulated from Cauchy distributions



2.4.2 Simulation from t distribution

The density function of a central t-distribution with shift parameter μ and degrees of freedom ν is given by:

$$f(x|\mu,\nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} \frac{1}{\sqrt{\pi\nu}} \frac{1}{\left(1+\frac{(x-\mu)^2}{\nu}\right)^{\frac{\nu+1}{2}}}$$

For a given degrees of freedom, the likelihood function of μ can be constructed as follows:

$$L(\mu|\boldsymbol{x},\nu) = \prod_i \left(k_1 \left(1 + \frac{(x-\mu)^2}{\nu} \right)^{k_2} \right)$$

Where,

$$k_1 = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} \frac{1}{\sqrt{\pi\nu}} \text{ and } k_2 = -\frac{\nu+1}{2}.$$

The mle was computed numerically using the $optimize$ function in R. The weight function was computed as:

$$w(\mu) = E\left(\left(\frac{\partial}{\partial \mu} logL(\mu|\boldsymbol{x},\nu) \right)^2 \right) = E\left(\left(2k_2 \sum_i \frac{x_i - \mu}{\nu + (x_i - \mu)^2} \right)^2 \right)$$

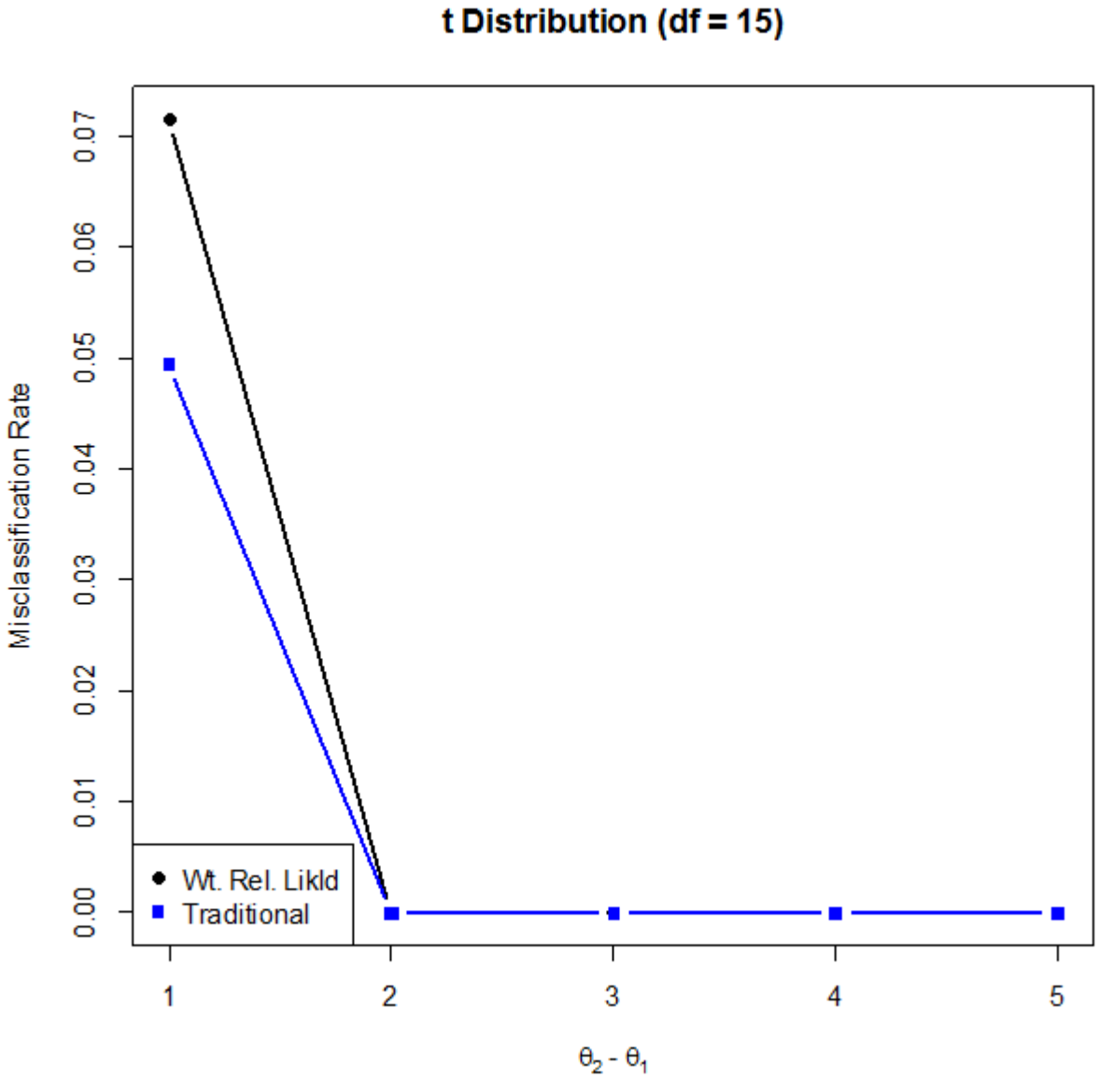
$$= 4k_2^2 E\left(\left(\sum_i \frac{x_i - \mu}{\nu + (x_i - \mu)^2} \right)^2 \right) \approx 4k_2^2 \left(\sum_i \frac{x_i - \hat{\mu}}{\nu + (x_i - \hat{\mu})^2} \right)^2$$

The misclassification rates are provided in Table 2.4 and Figure 2.5.

Table 2.4: t-distribution - Misclassification rates based on the k-means clustering applied to the matrix of values with evaluated weighted relative likelihood function and simulated data.

t Distribution	Wt. Relative Likld.	Simulated Data
$\theta_1 = 5, \theta_2 = 6$	0.072	0.049
$\theta_1 = 5, \theta_2 = 7$	0	0
$\theta_1 = 5, \theta_2 = 8$	0	0
$\theta_1 = 5, \theta_2 = 9$	0	0
$\theta_1 = 5, \theta_2 = 10$	0	0
Note: θ_1 *and* θ_2 are the shift parameters. Degrees of freedom set at 15.		

Figure 2.5: Misclassification rates versus difference in shift parameters for data simulated from in t distributions



The simulation results above for different degrees of freedom indicate that as the distinction between the distributions from which data are sampled increases, the misclassification error rates based on the proposed clustering algorithm decreases. This is not surprising. However, with distributions that are closer to each other, the error rate is smaller in case of likelihood based clustering.

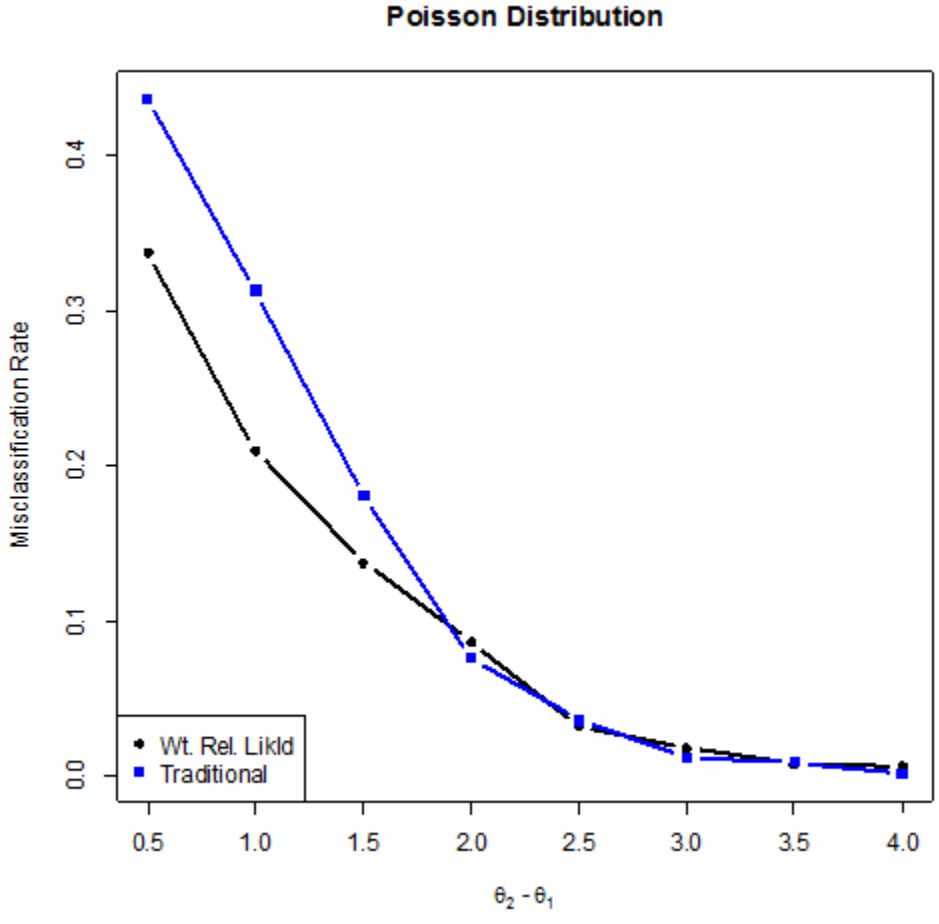
2.4.3 Simulation from Poisson distribution

The misclassification rates are provided in Table 2.5 and Figure 2.6.

Table 2.5: Poisson distribution - Misclassification rates based on the k-means clustering applied to the matrix of values with evaluated weighted relative likelihood function and simulated data.

	Misclassification rate	
Poisson Distribution	**Wt. Relative Likld.**	**Simulated Data**
$\lambda_1 = 5, \lambda_2 = 5.5$	0.337	0.437
$\lambda_1 = 5, \lambda_2 = 6$	0.209	0.313
$\lambda_1 = 5, \lambda_2 = 6.5$	0.138	0.181
$\lambda_1 = 5, \lambda_2 = 7$	0.087	0.076
$\lambda_1 = 5, \lambda_2 = 7.5$	0.032	0.036
$\lambda_1 = 5, \lambda_2 = 8$	0.018	0.012
$\lambda_1 = 5, \lambda_2 = 8.5$	0.008	0.009
$\lambda_1 = 5, \lambda_2 = 9$	0.007	0.002

Figure 2.6: Misclassification rates versus difference in rate parameters for data simulated from Poisson distributions.



2.5 Discussion

Our work develops and applies a novel approach for clustering based on RLFs. Our proposed method takes into account the structure of data via the density functions involved, as well as the sufficient statistics, in constructing the RLFs. The weighted RLF includes the data, parameter space, as well as the Fisher information. The dissimilarity in weighted RLFs (over different values in parameter space) across different observations was subjected to multivariate data analytic clustering techniques.

The proposed method was applied to simulated datasets from known distributions. Simulated datasets from Cauchy distribution had lower misclassification rates when the k means clustering was applied to the matrix of weighted relative likelihoods. Simulated data from t distribution had marginally higher misclassification rate when the shift parameter differed by 1 and the error rate dropped to zero for differences greater than 1. Simulated data from Poisson distribution had lower misclassification rates when clustering was applied to the matrix of weighted relative likelihood for close distributions (difference in rate parameter less than 2) and as the difference increased the error rate converged to zero. Simulated datasets based on observed rates of Salmonellosis across different counties in Kansas produced intuitive clusters of counties i.e. counties with similar rates were clustered together.

Since the proposed approach is based on likelihood functions it can be applied to datasets having different numbers of observations per subject i.e. the multiplicity of observations could vary across the subjects. We believe that the proposed methods can be extended to multivariate data. Future efforts are aimed at generalizing the proposed methods to multivariate data and associated multidimensional likelihood and profile likelihood functions. In addition we are also interested in deriving the optimal number of different parameters that need to be evaluated.

References

1. Casella, G., Berger Roger L, *Statistical Inference*. Second ed. 2001, Belmont, California: Duxbury Press.
2. Fisher, R.A., *Inverse probability and the use of likelihood.* Proceedings of the Cambridge Philosophical Society, 1932. **28**: p. 257-261.
3. Fisher, R.A., *Probability likelihood and quantity of information in the logic of uncertain inference.* Proceedings of the Royal Society of London Series a-Containing Papers of a Mathematical and Physical Character, 1934. **146**(A856): p. 0001-0008.
4. Fisher, R.A., *Two new properties of mathematical likelihood.* Proceedings of the Royal Society of London Series a-Containing Papers of a Mathematical and Physical Character, 1934. **144**(A852): p. 0285-0307.
5. Fisher, R.A., *The likelihood solution of a problem in compounded probabilities.* Annals of Eugenics, 1941. **11**: p. 306-307.
6. Sprott, D.A. and J.D. Kalbfleisch, *Examples of Likelihoods and Comparison with Point Estimates and Large Sample Approximations.* Journal of American Statistical Association, 1969. **64**(326): p. 468-484.
7. Bimali, M. and M. Brimacombe, *Relative Likelihood Differences to Examine Asymptotic Convergence: A Bootstrap Simulation Approach.* J. Biomet Biostat, 2015. **6**(220).
8. Kerby, A., et al., *Spatial Clustering Using the Likelihood Function.* Proceedings of the Sevents IEEE International Conference on Data Mining Workshops, 2007: p. 637-642.
9. Castro, R. and R. Nowak, *Likelihood based hierarchical clustering and network topology identification.* Proceedings of the 4th International Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition 2003. **2683**: p. 113-129.

10. Rencher, A.C., *Methods of multivariate analysis*. 2nd ed. Wiley series in probability and mathematical statistics. 2002, New York: J. Wiley. xxii, 708 p.

11. Johnson, R.A. and D.W. Wichern, *Applied multivariate statistical analysis*. 5th ed. 2002, Upper Saddle River, N.J.: Prentice Hall. xviii, 767 p.

12. Team, R.C., *R: A language and environment for statistical computing*. 2015, R Foundation for Statistical Computing: Vienna, Austria.

13. Efron, B. and D. Hinkley, *Assessing the Accuracy of the Maximum Likelihood Estimator: Observed Versus Expected Fisher Information.* Biometrika, 1978. **65**(3): p. 457-482.

14. Data.Cms.gov. *Inpatient Prospective Payment System (IPPS) Provider Summary for the Top 100 Diagnosis*. 2014; Available from: https://data.cms.gov/Medicare/Inpatient-Prospective-Payment-System-IPPS-Provider/97k6-zzx3.

Chapter 3: Information Based Clustering of Gene Expression Signatures in Primary Breast Carcinoma Patients

3.1 Introduction

Clustering is a grouping procedure focused on identifying subgroups within a dataset (Rencher 2002). While traditional non-parametric clustering methods such as hierarchical clustering and k-means clustering are commonly used clustering algorithms (D'Haeseleer 2005), there has been work developing parametric clustering approaches such as model-based clustering (Bouveyron and Brunet-Saumard 2014). Despite the differences in assumptions and approaches, the objective of most clustering algorithms is to classify subjects or observations into one of a finite set of disjoint clusters while ensuring that subjects within a cluster are more similar than subjects across clusters.

In the context of gene expression data, clustering techniques have been employed to identify sub-groups of patients at the molecular level, to understand gene function and regulation. It has been applied successfully to group similarly expressed genes across a set of subjects as well as in grouping subjects that share similar gene expression profiles (Jiang, Tang et al. 2004). In the context of clustering gene expression data, hierarchical clustering and k means clustering are more commonly used (D'Haeseleer 2005). Other approaches such as fuzzy c means clustering, self-organizing maps, and model-based clustering have also been employed (Toronen, Kolehmainen et al. 1999, Yeung, Fraley et al. 2001, Gasch and Eisen 2002, Nikkila, Toronen et al. 2002, Covell, Wallqvist et al. 2003, Huang, Wei et al. 2006, Arima, Hakamada et al. 2008, Zhang, Adamu et al. 2011, Shahdoust, Hajizadeh et al. 2013, Zhang and Shen 2014).

Our recent work has proposed a clustering approach based on the properties of the observed likelihood and Fisher Information for each observation in the dataset (Bimali and

Brimacombe 2015). Unlike the traditional non-parametric and model based approach, the proposed method takes into account the structure of data in relation to the distributional assumption as well as information based similarity among observations in the data. In the context of gene expression, the proposed method assumes that the gene expression profile for each subject is assumed to follow a known distribution and thus a set of RLFs can be constructed. The likelihood functions can be viewed as a transformation of the original gene expression profiles. These RLFs are further weighted by the Fisher Information to obtain the weighted RLFs. This is then evaluated at different values of the parameter to obtain a data matrix which can be subjected to the clustering algorithms. The proposed clustering approach takes into account the variation in mean expression levels as well as the variation in the observed Fisher Information across the patients.

Here we apply the proposed clustering approach to the publicly available dataset by Van De Vijer et al (2002) in clustering primary breast carcinomas patients based on a previously recommended set of 70 gene expression profile (van de Vijver, He et al. 2002). The agreement between the proposed clustering approach and authors' classification has been examined. The clusters obtained are also examined in relation to two clinical features – time to overall survival; and time to metastases**.**

3.2 Data

The dataset has been made available by Van De Vijver et al at http://ccb.nki.nl/data/. Personal communication with the corresponding author clarified the availability of the dataset for analysis. The authors describe the study subject group as patients having either I or II breast cancer and younger than 53 years. The authors have made available expression profiles for 24496 genes, of which 70 genes formed a subset. Clinical covariates such as time to overall survival, time to distant metastases, death status, and the number of positive nodes were also provided. Van De Vijer et al used 70 gene expression profiles that were identified by Veer et al, to classify 295 patients with primary breast carcinomas into two groups – poor prognosis groups and good prognosis group (van 't Veer, Dai et al. 2002). Among the 295 patients, 180 were classified into poor prognosis groups while 115 were classified into good prognosis groups.

The prognostic classification was based on correlation of these 70 genes with the average profile of these 70 genes in tumors from patients with a good prognosis. The threshold of 0.4, used for correlation coefficients, was determined based on a previous study of 78 tumors which resulted in a false negative rate of 10 percent. The two groups differed significantly with respect to the overall 10-year survival time as well as with respect to time to distant metastases. The authors mentioned that the classification system based on 70 genes outperformed all clinical variables in predicting the likelihood of distant metastases within 5 years. The dataset provided used by Van De Vijer has been made available publicly. We restrict our attention to the 70 gene expression profiles and examine the subsequent clusters of 295 patients formed on based on these 70 gene expressions.

3.3 Method

Genes in each subject are assumed to follow a known distribution and thus likelihood functions can be constructed. The likelihood functions are further scaled by their maxima to transform them into RLFs. A data matrix is then developed by evaluating the weighted RLFs at different values in the parameter space, the weights being the Fisher Information matrix evaluated at the mode of the likelihood functions. The proposed approach thus takes into account the structure of data via the distributional assumption as well as information similarity between observations in the data. We assume that the genes for each subject follow a normal distribution.

Let us consider data matrix $\boldsymbol{X} = (\boldsymbol{x_1} \quad \dots \quad \boldsymbol{x_n})'$ where $\boldsymbol{x_i} = (x_{i1} \quad \dots \quad x_{ik})$

$x_{i1}, \dots, x_{ik}$ are iid observations with pdf $f_i(\boldsymbol{x_i}; \theta_i); j = 1, \dots, T_i$ and $\boldsymbol{\theta} = (\theta_1 \quad \dots \quad \theta_n)$

We assume that $\theta_i's$ share the same support. Thus for each θ_i we can construct likelihood functions reflecting assumed pdf giving rise to n likelihood functions based the data matrix $\boldsymbol{X}$.

$$\boldsymbol{L_X}(\boldsymbol{\theta}) = (L_{\boldsymbol{x_1}}(\theta_1) \quad \dots \quad L_{\boldsymbol{x_n}}(\theta_n))'$$

where $L_{\boldsymbol{x_i}}(\theta_i) = \prod_{j=1}^{k} f(x_{ij}|\theta_i)$. Let $\hat{\theta}_i$ be the mle of θ_i. Then the RLF for each θ_i can be constructed as follows:

$$\boldsymbol{R_X}(\boldsymbol{\theta}) = (R_{\boldsymbol{x_1}}(\theta_1) \quad \dots \quad R_{\boldsymbol{x_n}}(\theta_n))'$$

with $R_{\boldsymbol{x_i}}(\theta_i) = \frac{L_{\boldsymbol{x_i}}(\theta_i)}{L_{\boldsymbol{x_i}}(\hat{\theta}_i)}$. Note that the Fisher information, by definition is the same for both the initial and relatively re-weighted likelihood function. The relative aspect however is useful in graphical summaries related to the clustering and is maintained here.

To improve the assessment of similarity across the set of evaluated likelihoods, the Fisher information matrix for each observation can be used as a weight and we have;

$w_{x_i}(\hat{\theta}) = I(\hat{\theta})$ where $I(\theta) = E\left(\left(\frac{\partial}{\partial\theta} log(L(\theta|\boldsymbol{x_i}))\right)^2\right)$

For exponential families with iid observations, note that the Fisher Information matrix can be simplified to $I(\theta) = -E\left(\frac{\partial^2}{\partial\theta^2}\log l(\theta|\boldsymbol{x}_i)\right)$

The value of the likelihood functions can be evaluated at different values of the $\theta_i's$. For each observation $\boldsymbol{x_i}$, we can compute the value of likelihood functions at k different θ_i values. Thus we can construct a matrix $\boldsymbol{P_X}$ with rows containing the weighted RLFs evaluated at different values of θ_i .

$$\boldsymbol{P_X} = \begin{pmatrix} w_{\boldsymbol{x_1}}(\hat{\theta}) & \cdots & w_{\boldsymbol{x_1}}(\hat{\theta}) \\ \vdots & \ddots & \vdots \\ w_{\boldsymbol{x_n}}(\hat{\theta}) & \cdots & w_{\boldsymbol{x_n}}(\hat{\theta}) \end{pmatrix} \circ \begin{pmatrix} R_{\boldsymbol{x_1}}(\theta_1) & \cdots & R_{\boldsymbol{x_1}}(\theta_k) \\ \vdots & \ddots & \vdots \\ R_{\boldsymbol{x_n}}(\theta_1) & \cdots & R_{\boldsymbol{x_n}}(\theta_k) \end{pmatrix}$$

where, $w_{x_i}(\hat{\theta})$ is the Fisher Information evaluated at the *MLE*. $R_{x_i}(\theta_j)$ is the value of the RLF for x_i evaluated at θ_j and $\circ$ is the Hadamard product operator between the two matrices. The matrix $\boldsymbol{P_X}$ can be subjected to various standard clustering algorithms to explore for patterns and clusters in the data matrix $\boldsymbol{X}$.

Under the assumption of normality of genes for each subject, the weighted RLF for each subject can be shown to be as follows:

$w_{\boldsymbol{x_i}}(\theta) = \frac{n}{\hat{\sigma}^2} \times \exp\left(-0.5 \times \frac{n}{\hat{\sigma}^2}\left(\theta - \hat{\theta}\right)^2\right)$

The above weighted RLF can be evaluated across different values of θ for each subject to obtain a matrix of weighted RLFs.

3.4 Analysis

The assumption of normality of genes for each subject was tested using the Shapiro-Wilk's test of non-normality. Among the 295 subjects, 88 subjects showed significant deviation from the normality assumption based on α −level of 0.01, and were thus excluded from the analysis. Table 3.1 provides summary statistics on the survival time, time to distant metastases, for good and poor prognosis subjects. The pair-wise correlation of genes across the subjects was examined. The correlations of the gene expression profiles across 207 patients were examined and genes that were moderately to highly correlate with other genes were excluded to be consistent with the iid assumption. The absolute correlation threshold was set at 0.8,0.7, and 0.6 respectively. Thus the data matrix that was analyzed consisted of 207 patients with gene expression profiles whose correlation (absolute value) was below the specified threshold.

For each of the 207 subjects, a weighted RLF was constructed. The weighted RLF was then evaluated at 1000 equi-spaced intervals within $(-0.4,0.3)$. The range was chosen so as to cover the variation across the MLEs as well as the support of the observed likelihood functions where the evaluated likelihood is greater than 0. The matrix of evaluated weighted RLF was then subjected to k means clustering with 2 clusters. Choosing two clusters allows us to examine the agreement between the authors classification of poor and good prognosis as the cluster formed based on proposed approach.

Table 3.1: Summary Statistics on overall survival time and time to metastases on 207 patients.

Prognosis	Count	Median Overall Survival	Median Time to Metastases
Good	92	8.91 (4)	4.98 (12)
Poor	115	6.93 (40)	2.94 (50)
Note: Value in bracket indicate number of subjects experiencing the event			

3.4.1 Correlation threshold set at 0.8

The number of gene expression profiles dropped from 70 to 64 $i.e.$ there were 6 genes that were highly correlated (correlation $\geq$ 0.8) with other genes and were dropped from analysis. The plot of weighted relative observed likelihood function is provided in Figure 3.1. The data matrix obtained by evaluating the weighted relative observed likelihood function was subjected to k means cluster with 2 clusters. Figure 3.2 provides a plot of the weighted relative observed likelihood functions colored by their cluster assignment.

Figure 3.1: Weighted relative likelihood functions for the 207 subjects

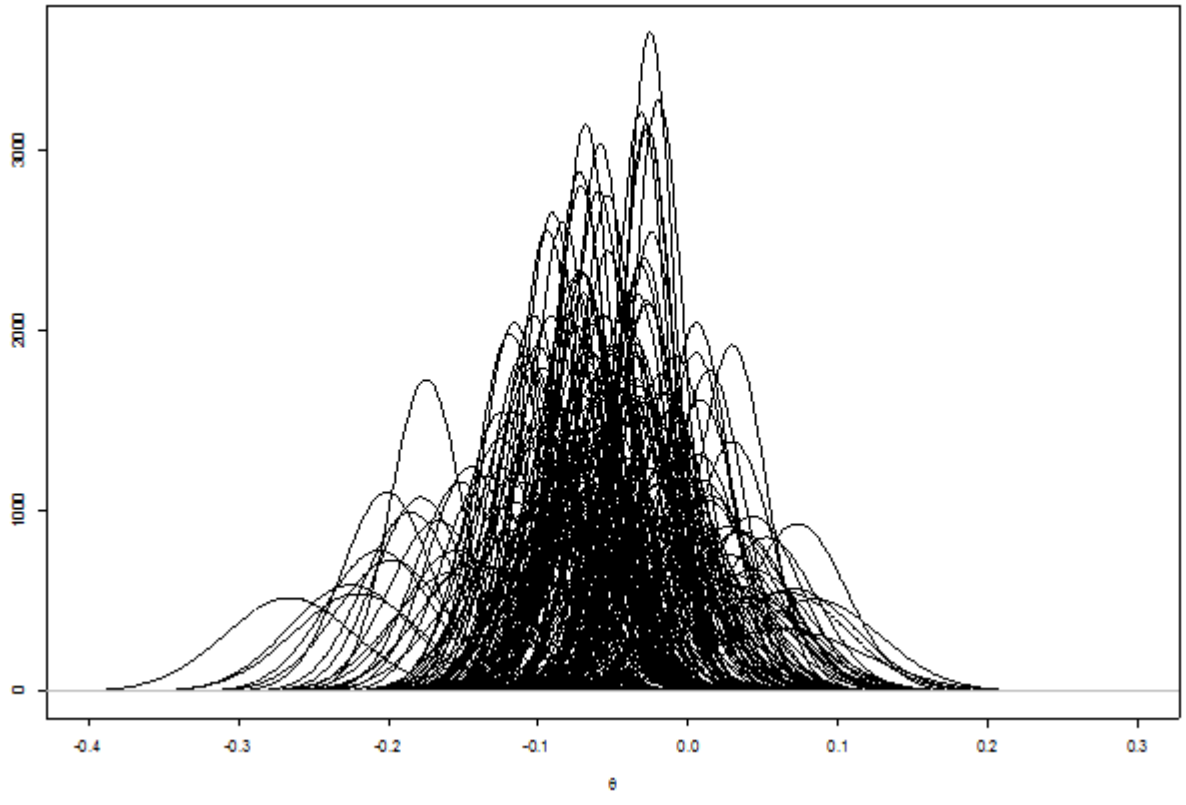


Figure 3.2: Weighted relative likelihood functions colored by their cluster assignment.

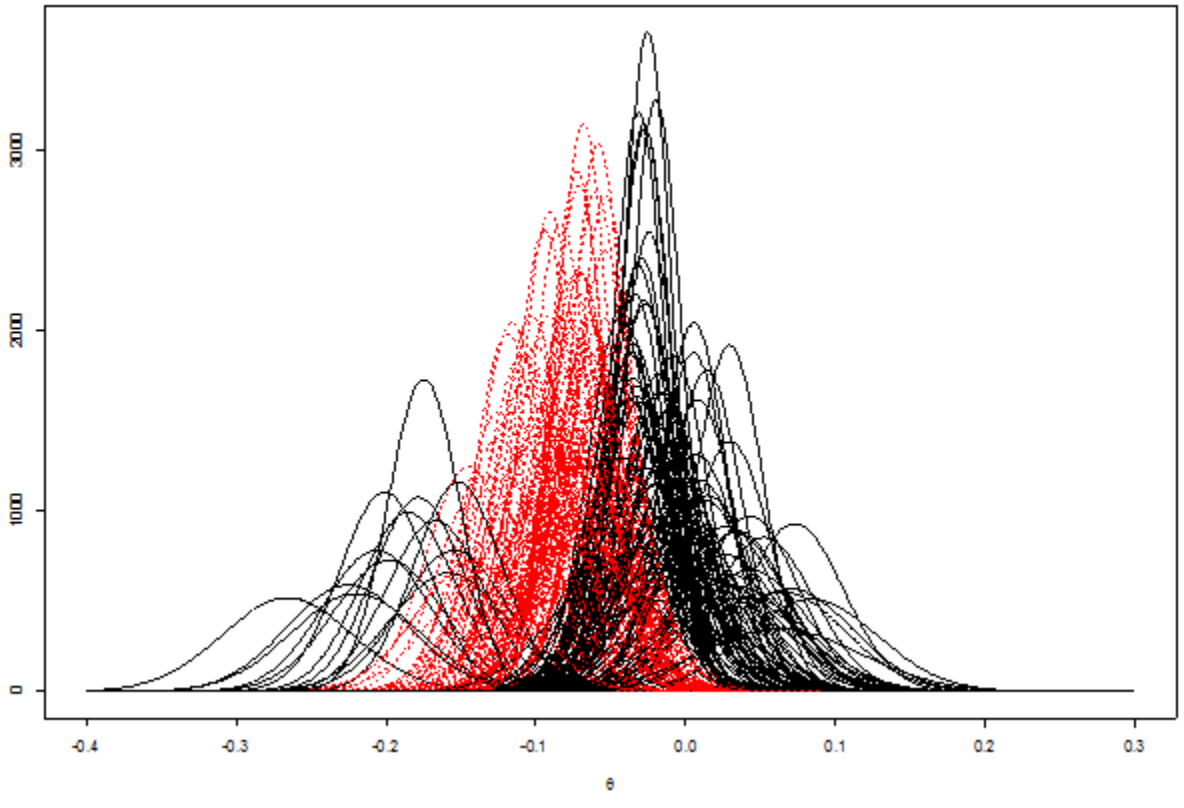


Table 3.2 provides a summary of the agreement between the authors' classification and the clustering based on weighted RLF.

Table 3.2: Bivariate classification of cluster assignment by prognosis classification.

	Cluster 1	Cluster 2
Good Prognosis	22	70
Poor Prognosis	82	33

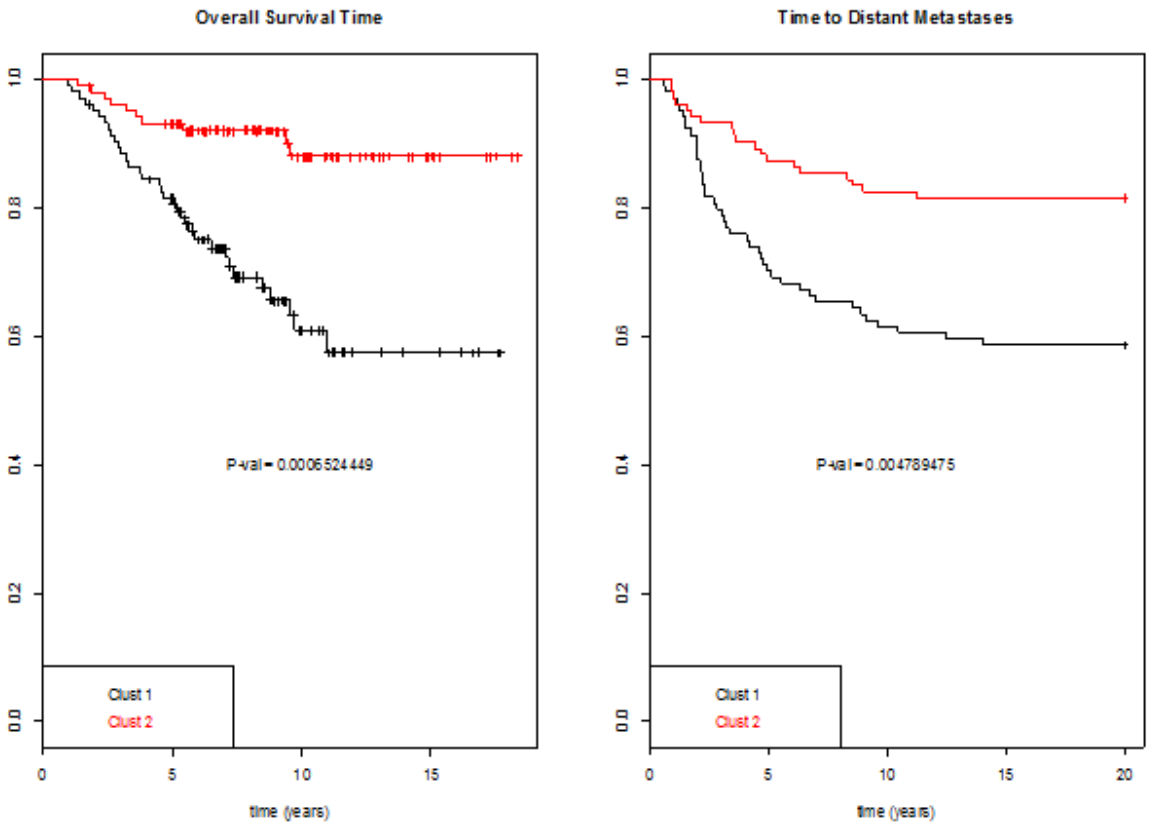
The summary statistics of the two clusters is provided in Table 3.3.

Table 3.3: Summary Statistics on overall survival time and time to metastases on 207 patients with gene expression profiles whose correlation was below 0.8 based on cluster assignment.

Cluster	Count	Median Overall Survival	Median Time to Metastases
1	104	6.88 (34)	3.02 (43)
2	103	8.81 (10)	3.66 (19)
Note: Value in bracket indicate number of subjects experiencing the event			

The two clusters were tested for difference in overall survival time as well as time to distant metastases using log-rank test (Figure 3.3). The two clusters differed significantly in terms of overall survival times (p value $= 6.5 \times 10^{-4}$) as well as time to distant metastases (p-value = 4.88×10^{-3}).

Figure 3.3: Kaplan-Meier Survival Plot for difference in overall survival time (left) as well as time to distant metastases between the two clusters.



3.4.2 Correlation threshold set at 0.7

The number of gene expression profiles dropped from 70 to 58 i.e. there were 12 genes with moderately high correlation (correlation $\geq$ 0.7) with other genes and were dropped from analysis. The plot of weighted relative observed likelihood function is provided in Figure 3.4. The data matrix obtained by evaluating the weighted relative observed likelihood function was subjected to k means cluster with 2 clusters. Figure 3.5 provides a plot of the weighted relative observed RLFs colored by their cluster assignment.

Figure 3.4: Weighted relative likelihood functions for the 207 subjects and 58 genes.

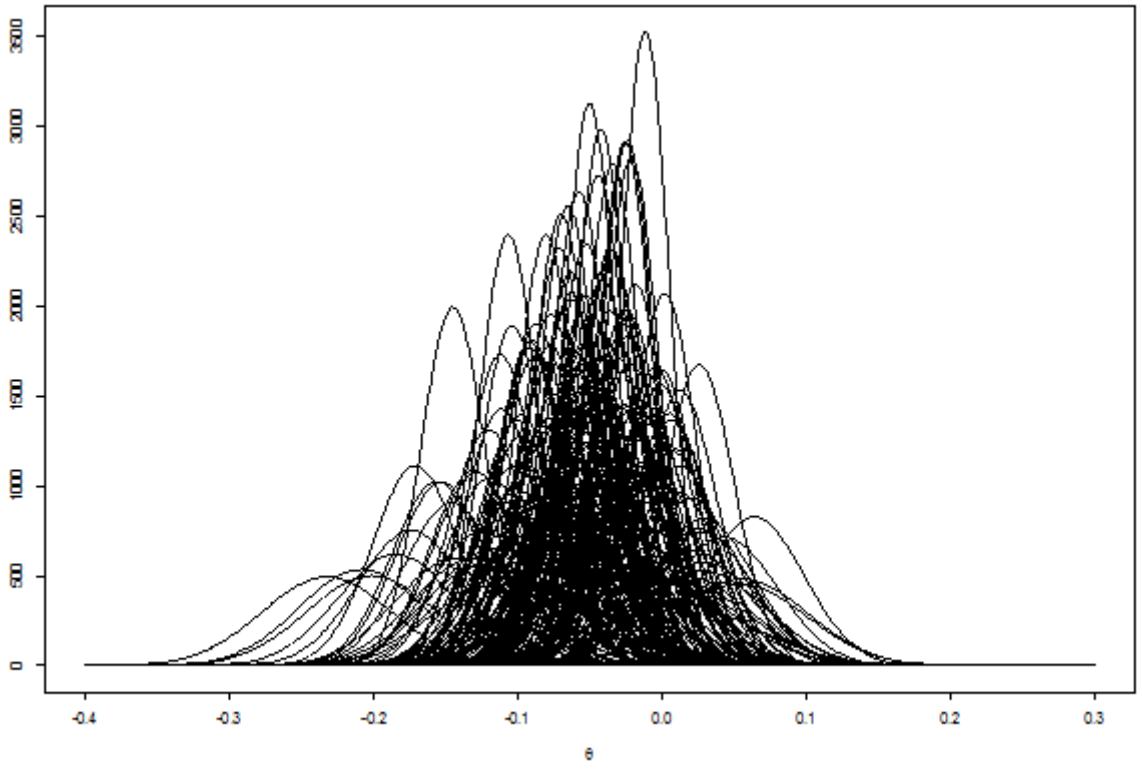


Figure 3.5: Weighted relative likelihood functions (207 subjects and 58 genes) colored by their cluster assignment.

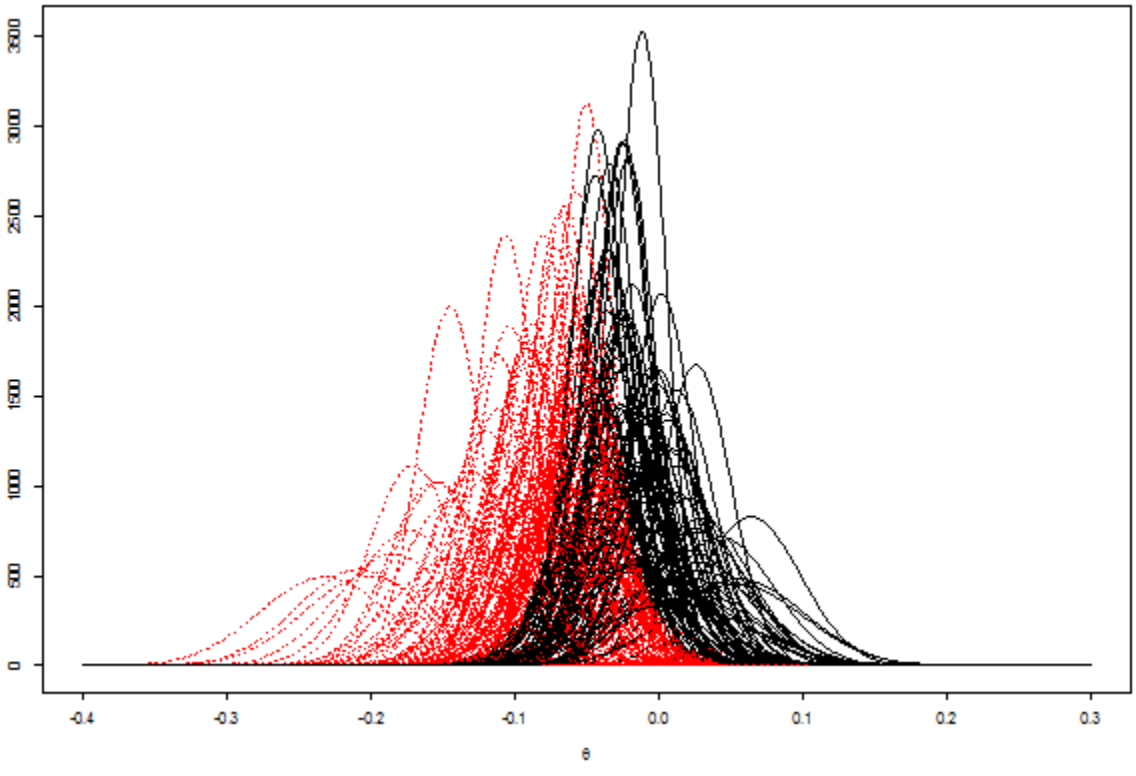


Table 3.4 provides a summary of the agreement between the authors' classification and the clustering based on weighted RLF.

Table 3.4: Bivariate classification of cluster assignment by prognosis classification.

	Cluster 1	Cluster 2
Good Prognosis	19	73
Poor Prognosis	72	43

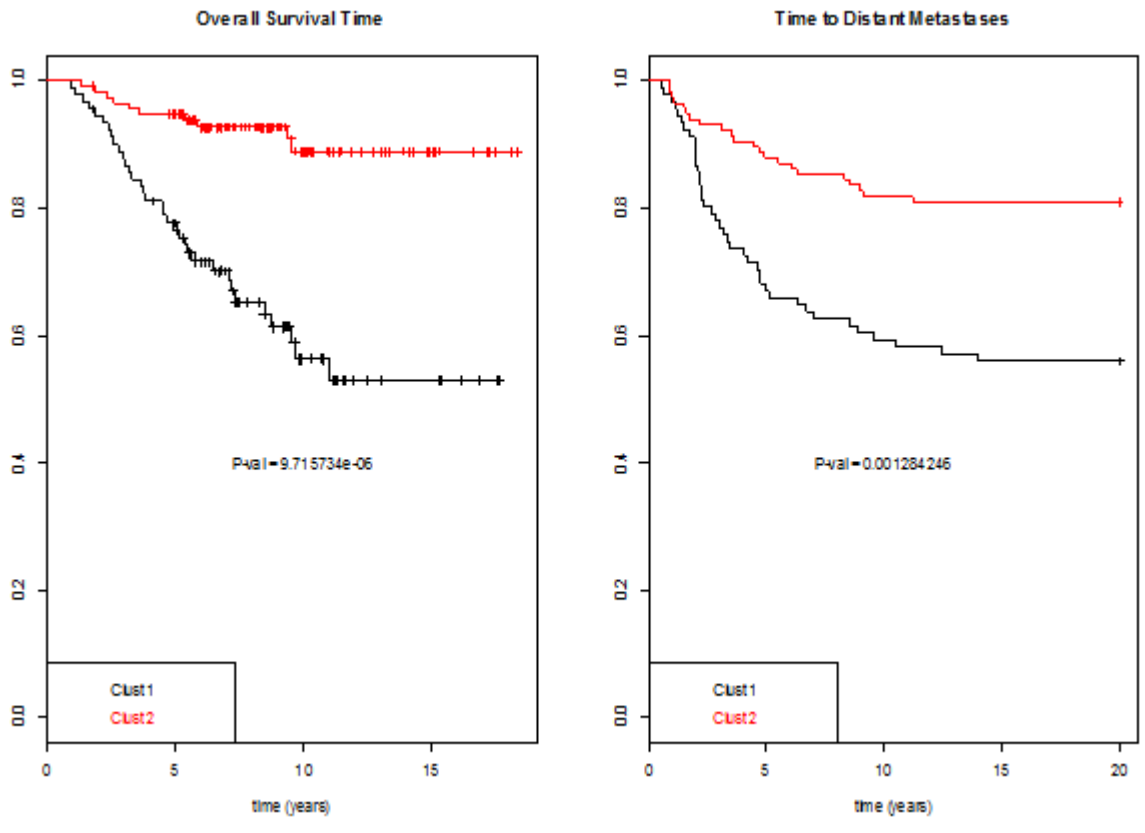
The summary statistics of the two clusters is provided in Table 3.5.

Table 3.5: Summary Statistics on overall survival time and time to metastases on 207 patients with gene expression profiles whose correlation was below 0.8 based on cluster assignment.

Cluster	Count	Median Overall Survival	Median Time to Metastases
1	116	6.93 (34)	4.14 (40)
2	91	8.77 (10)	4.05 (22)
Note: Value in bracket indicate number of subjects experiencing the event			

The two clusters were tested for difference in overall survival time as well as time to distant metastases using log-rank test (Figure 3.6). The two clusters differed significantly in terms of overall survival times (p value $= 9.72 \times 10^{-6}$) as well as time to distant metastases (p-value = 1.28×10^{-3}).

Figure 3.6: Kaplan-Meier Survival Plot for difference in overall survival time (left) as well as time to distant metastases between the two clusters.



3.4.3 Correlation threshold set at 0.6

The number of gene expression profiles dropped from 70 to 52 i.e. there were 18 genes with moderately high correlation (correlation $\geq$ 0.7) with other genes and were dropped from analysis. The plot of weighted relative observed likelihood function is provided in Figure 3.7. The data matrix obtained by evaluating the weighted relative observed likelihood function was subjected to k means cluster with 2 clusters. Figure 3.8 provides a plot of the weighted relative observed RLFs colored by their cluster assignment.

Figure 3.7: Weighted relative likelihood functions for the 207 subjects and 52 genes.

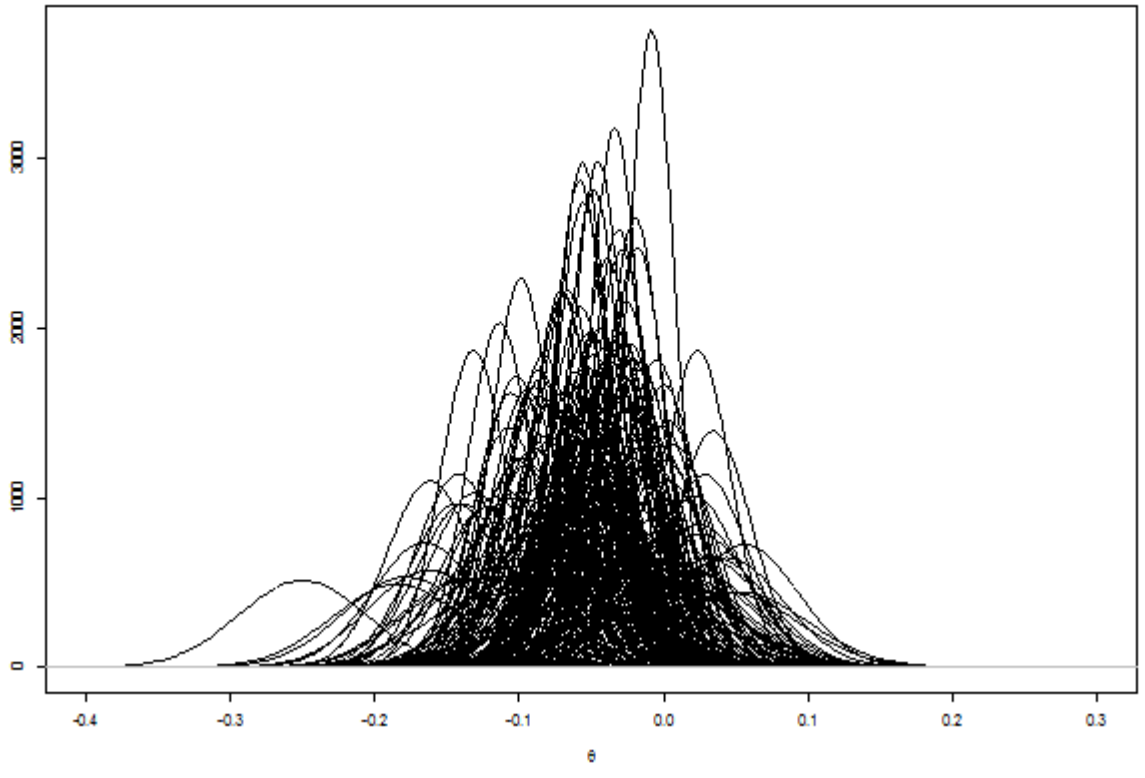


Table 3.6 provides a summary of the agreement between the authors' classification and the clustering based on weighted RLF. The summary statistics of the two clusters is provided in Table 3.7.

Figure 3.8: Weighted relative likelihood functions (207 subjects and 58 genes) colored by their cluster assignment.

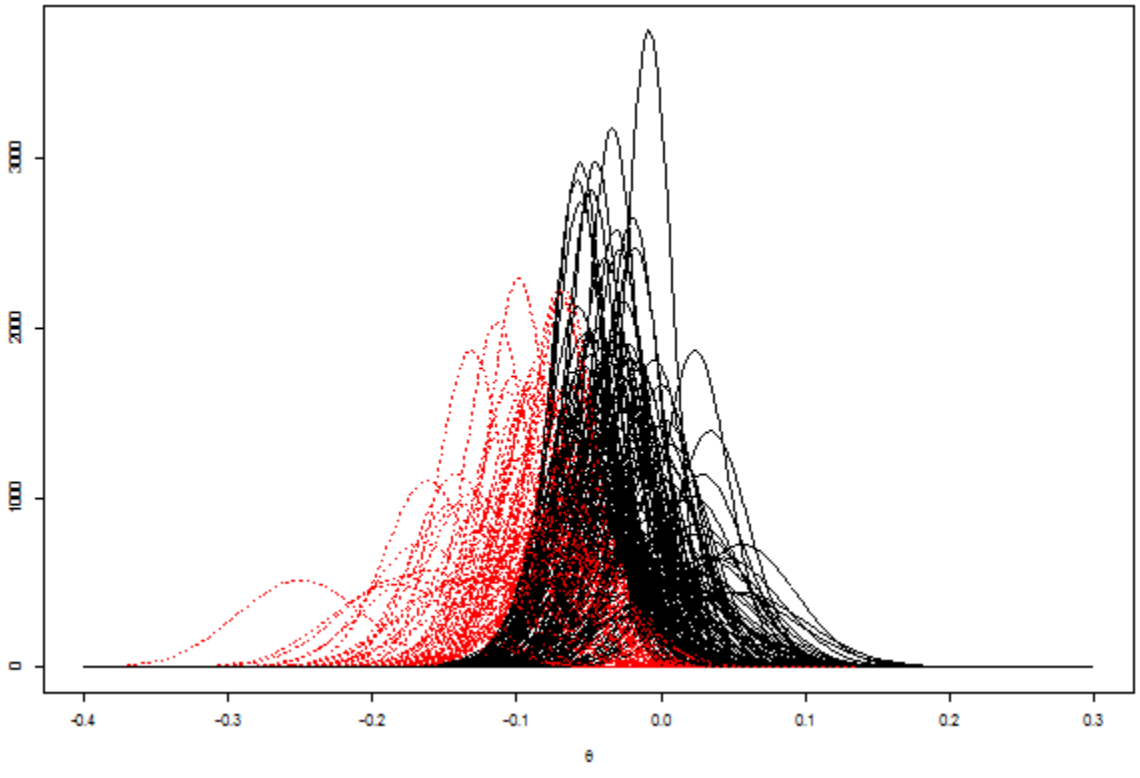


Table 3.6: Bivariate classification of cluster assignment by prognosis classification.

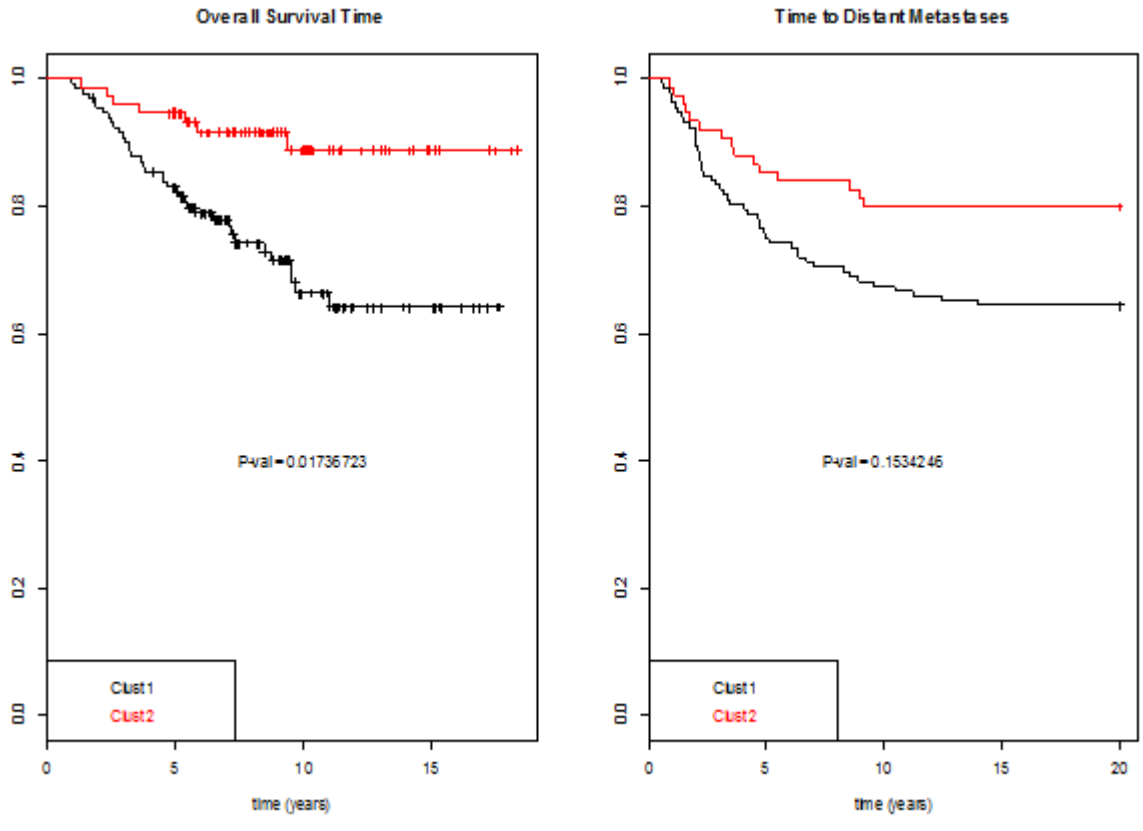
	Cluster 1	Cluster 2
Good Prognosis	46	46
Poor Prognosis	86	29

Table 3.7: Summary Statistics on overall survival time and time to metastases on 207 patients with gene expression profiles whose correlation was below 0.6 based on cluster assignment.

Cluster	Count	Median Overall Survival	Median Time to Metastases
1	132	7.89 (37)	3.22 (47)
2	75	9.11 (7)	3.58 (15)
Note: Value in bracket indicate number of subjects experiencing the event			

The two clusters were tested for difference in overall survival time as well as time to distant metastases using log-rank test (Figure 3.9). The two clusters differed significantly in terms of overall survival times (p value = 0.017) as well as time to distant metastases (p-value = 0.153).

Figure 3.9: Kaplan-Meier Survival Plot for difference in overall survival time (left) as well as time to distant metastases between the two clusters.



3.5 Discussion

The clusters of patients obtained takes into consideration the variation across the mean expression level of the genes as well as variation across observed Fisher Information. The correlation threshold was set at 0.8,0.7, and 0.6. It isn't surprising that as the correlation threshold was relaxed, the number of gene expression profiles decreased gradually from 64 to 58 to 52. The two clusters differ significantly with respect to overall survival time for each of the three correlation thresholds as well as time to distant metastases for the correlation thresholds of 0.8 and 0.7. Unlike the author's classification, our clustering algorithm uses fewer number of gene expression profiles to be consistent with the assumption in the proposed methodology. Our clustering results show clusters of patients that differed significantly with respect to overall survival time as well as time to distant metastases based on subset of the 70 gene expression profiles and thus provide further support to the authors' argument of the 70 gene expression profiles being strong predictors of the risk of metastases.

References

Arima, C., K. Hakamada, M. Okamoto and T. Hanai (2008). "Modified fuzzy gap statistic for estimating preferable number of clusters in fuzzy k-means clustering." J Biosci Bioeng **105**(3): 273-281.

Bimali, M. and M. Brimacombe (2015). "Likelihood Transformation and Information Based Approach to Clustering." Submitted.

Bouveyron, C. and C. Brunet-Saumard (2014). "Model-based clustering of high-dimensional data: A review." Computational Statistics and Data Analysis **72**: 52-78.

Covell, D. G., A. Wallqvist, A. A. Rabow and N. Thanki (2003). "Molecular classification of cancer: unsupervised self-organizing map analysis of gene expression microarray data." Mol Cancer Ther **2**(3): 317-332.

D'Haeseleer, P. (2005). "How does gene expression clustering work?" Nat Biotechnol **23**(12): 1499-1501.

Gasch, A. P. and M. B. Eisen (2002). "Exploring the conditional coregulation of yeast gene expression through fuzzy k-means clustering." Genome Biol **3**(11): RESEARCH0059.

Huang, D., P. Wei and W. Pan (2006). "Combining gene annotations and gene expression data in model-based clustering: weighted method." OMICS **10**(1): 28-39.

Jiang, D., C. Tang and A. Zhang (2004). "Cluster Analysis for Gene Expression Data: A Survey." IEEE Transactions on Knowledge and Data Engineering **16**(11): 1370-1386.

Nikkila, J., P. Toronen, S. Kaski, J. Venna, E. Castren and G. Wong (2002). "Analysis and visualization of gene expression data using self-organizing maps." Neural Netw **15**(8-9): 953-966.

Rencher, A. C. (2002). Methods of multivariate analysis. New York, J. Wiley.

Shahdoust, M., E. Hajizadeh, H. Mozdarani and A. Chehrei (2013). "Finding genes discriminating smokers from non-smokers by applying a growing self-organizing clustering method to large airway epithelium cell microarray data." Asian Pac J Cancer Prev **14**(1): 111-116.

Toronen, P., M. Kolehmainen, G. Wong and E. Castren (1999). "Analysis of gene expression data using self-organizing maps." FEBS Lett **451**(2): 142-146.

van 't Veer, L. J., H. Dai, M. J. van de Vijver, Y. D. He, A. A. Hart, M. Mao, H. L. Peterse, K. van der Kooy, M. J. Marton, A. T. Witteveen, G. J. Schreiber, R. M. Kerkhoven, C. Roberts, P. S. Linsley, R. Bernards and S. H. Friend (2002). "Gene expression profiling predicts clinical outcome of breast cancer." Nature **415**(6871): 530-536.

van de Vijver, M. J., Y. D. He, L. J. van't Veer, H. Dai, A. A. Hart, D. W. Voskuil, G. J. Schreiber, J. L. Peterse, C. Roberts, M. J. Marton, M. Parrish, D. Atsma, A. Witteveen, A. Glas, L. Delahaye, T. van der Velde, H. Bartelink, S. Rodenhuis, E. T. Rutgers, S. H. Friend and R. Bernards (2002). "A gene-expression signature as a predictor of survival in breast cancer." N Engl J Med **347**(25): 1999-2009.

Yeung, K. Y., C. Fraley, A. Murua, A. E. Raftery and W. L. Ruzzo (2001). "Model-based clustering and data transformations for gene expression data." Bioinformatics **17**(10): 977-987.

Zhang, J. and L. Shen (2014). "An improved fuzzy c-means clustering algorithm based on shadowed sets and PSO." Comput Intell Neurosci **2014**: 368628.

Zhang, M., B. Adamu, C. C. Lin and P. Yang (2011). "Gene expression analysis with integrated fuzzy C-means and pathway analysis." Conf Proc IEEE Eng Med Biol Soc **2011**: 936-939.

Summary

This dissertation work examines and extends likelihood functions, in general and RLFs, in particular, to assess the relationship between the sample size and large sample behavior of MLE; as well as in the context of clustering. Chapter 1 discusses the issue of appropriateness of sample size required for asymptotic normality of *MLE*s to hold true. We accept the argument by Sprott et al of examining the RLFs based on observed data and the large sample distribution of MLEs to gauge whether the data at hand is large enough for large sample normality of MLEs to hold satisfactorily. In order to examine the proximity between the two relative likelihoods we proposed two different diagnostic measures for this purpose viz. $\Delta R -$ difference in the area under the relative observed likelihood and relative asymptotic likelihood curves and $D -$ dissimilarity index which measures the shape of the curves. It was argued that for a given sample size if ΔR and D are both small, the sample size is large enough for asymptotic normality of MLE. The proposed measure was applied to data from literature as well as to simulated datasets.

Chapter 2 develops and applies a novel approach for clustering based on RLFs. For each observation in the dataset, a set of (relative) likelihood functions are constructed, viewed as a transformation of the original data. The Fisher Information reflects the local curvature of a log-likelihood function and summarizes the estimation related accuracy present in an individual likelihood and thus has been used as a weight for each respective likelihood function. The weighted RLF is then evaluated at different values of the parameter to obtain a data matrix. The data matrix is then subjected to multivariate data analytic clustering algorithms to explore for possible subgroups in the dataset. Our proposed method takes into account the structure of data via the density functions involved, as well as the sufficient statistics, in constructing the RLFs. The proposed method was applied to simulated datasets from known distributions as well as to simulated dataset based on data from literature. Since the proposed approach is based on

likelihood functions it can be applied to datasets having different numbers of observations per subject i.e. the multiplicity of observations could vary across the subjects. The misclassification rates based on proposed approach as well as traditional approach were examined. Chapter 3 is an application of the method proposed in chapter 2 to gene expression dataset. Here we examine a set of patients previously classified into two diagnostic groups. We cluster the patients and examine the agreement between our clusters and the previous classification. In addition, the differences between the clusters based on clinical covariates were also examined.

Bibliography

Altman, M., J. Gill and M. McDonald (2003). Convergence Problems in Logistic Regression. Numerical Issues in Statistical Computing for the Social Scientist, Wiley**:** 219-233.

Arima, C., K. Hakamada, M. Okamoto and T. Hanai (2008). "Modified fuzzy gap statistic for estimating preferable number of clusters in fuzzy k-means clustering." J Biosci Bioeng **105**(3): 273-281.

Atkinson, K. E. (1989). An Introduction to Numerical Analysis, John Wiley & Sons.

Bartholomew, D. J. (1957). "A problem in life testing." Journal of Americal Statistical Association **52**: 350-355.

Bhattacharyya, A. (1943). "On a measure of divergence between two statistical populations defined by their probability distribution." Calcutta Mathematical Society **35**: 99-110.

Bickel, P. J. and D. A. Freedman (1981). "Some Asymptotic Theory for the Bootstrap." Annals of Statistics **9**(6): 1196-1217.

Bickel, P. J. and D. A. Freedman (1984). "Asymptotic Normality and the Bootstrap in Stratified Sampling." Annals of Statistics **12**(2): 470-482.

Bimali, M. and M. Brimacombe (2015). "Likelihood Transformation and Information Based Approach to Clustering." Submitted.

Bimali, M. and M. Brimacombe (2015). "Relative Likelihood Differences to Examine Asymptotic Convergence: A Bootstrap Simulation Approach." J. Biomet Biostat **6**(220).

Bouveyron, C. and C. Brunet-Saumard (2014). "Model-based clustering of high-dimensional data: A review." Computational Statistics and Data Analysis **72**: 52-78.

Breslow, N. E. (1984). "Extra-Poisson variation in log-linear models." Applied Statistics **33**: 38-44.

Carlstein, E., K. A. Do, P. Hall, T. Hesterberg and H. R. Kunsch (1998). "Matched-block bootstrap for dependent data." Bernoulli **4**(3): 305-328.

Casella, G., Berger Roger L (2001). Statistical Inference. Belmont, California, Duxbury Press.

Castro, R. and R. Nowak (2003). "Likelihood based hierarchical clustering and network topology identification." Proceedings of the 4th International Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition **2683**: 113-129.

Chaudhuri, G., J. D. Borwankar and P. R. K. Rao (1991). "Bhattacharyya Distance Based Linear Discriminant Function for Stationary Time-Series." Communications in Statistics-Theory and Methods **20**(7): 2195-2205.

Chen, K., B. Li, L. F. Tian, W. B. Zhu and Y. H. Bao (2014). "Vessel attachment nodule segmentation using integrated active contour model based on fuzzy speed function and shape-intensity joint Bhattacharya distance." Signal Processing **103**: 273-284.

Covell, D. G., A. Wallqvist, A. A. Rabow and N. Thanki (2003). "Molecular classification of cancer: unsupervised self-organizing map analysis of gene expression microarray data." Mol Cancer Ther **2**(3): 317-332.

D'Haeseleer, P. (2005). "How does gene expression clustering work?" Nat Biotechnol **23**(12): 1499-1501.

Data.Cms.gov. (2014). "Inpatient Prospective Payment System (IPPS) Provider Summary for the Top 100 Diagnosis." from https://data.cms.gov/Medicare/Inpatient-Prospective-Payment-System-IPPS-Provider/97k6-zzx3.

Efron, B. (1979). "1977 Rietz Lecture - Bootstrap Methods - Another Look at the Jackknife." Annals of Statistics **7**(1): 1-26.

Efron, B. (1982). "Maximum Likelihood and Decision Theory." The Annals of Statistics **10**(2): 340-356.

Efron, B. and D. Hinkley (1978). "Assessing the Accuracy of the Maximum Likelihood Estimator: Observed Versus Expected Fisher Information." Biometrika **65**(3): 457-482.

Fisher, R. A. (1932). "Inverse probability and the use of likelihood." Proceedings of the Cambridge Philosophical Society **28**: 257-261.

Fisher, R. A. (1934). "Probability likelihood and quantity of information in the logic of uncertain inference." Proceedings of the Royal Society of London Series a-Containing Papers of a Mathematical and Physical Character **146**(A856): 0001-0008.

Fisher, R. A. (1934). "Two new properties of mathematical likelihood." Proceedings of the Royal Society of London Series a-Containing Papers of a Mathematical and Physical Character **144**(A852): 0285-0307.

Fisher, R. A. (1941). "The likelihood solution of a problem in compounded probabilities." Annals of Eugenics **11**: 306-307.

Freedman, D. A. (1986). "Jackknife, Bootstrap and Other Resampling Methods in Regression-Analysis - Discussion." Annals of Statistics **14**(4): 1305-1308.

Gasch, A. P. and M. B. Eisen (2002). "Exploring the conditional coregulation of yeast gene expression through fuzzy k-means clustering." Genome Biol **3**(11): RESEARCH0059.

Gersch, W., F. Martinelli, J. Yonemoto, M. D. Low and J. A. Mcewen (1980). "A Kullback Leibler Nearest Neighbor Rule Classification of Eegs - the Eeg Population Screening Problem, an Anesthesia Level Eeg Classification Application." Computers and Biomedical Research **13**(3): 283-296.

Gibbons, J. D. and S. Chakraborti (2011). Nonparametric Statistical Inference, CRC Press, Taylor and Francis Group.

Hall, P. (1986). "On the Bootstrap and Confidence-Intervals." Annals of Statistics **14**(4): 1431-1452.

Hall, P. (1988). "On Symmetric Bootstrap Confidence-Intervals." Journal of the Royal Statistical Society Series B-Methodological **50**(1): 35-45.

Hall, P. (1988). "Rate of Convergence in Bootstrap Approximations." Annals of Probability **16**(4): 1665-1684.

Hall, P. (1989). "On Efficient Bootstrap Simulation." Biometrika **76**(3): 613-617.

Hall, P. (1994). A short history of the bootstrap. Acoustics, Speech, and Signal Processing, IEEE International Conference on, IEEE.

Hall, P. and S. R. Wilson (1992). "Bootstrap Hypothesis-Testing - Reply." Biometrics **48**(3): 970-970.

Huang, D., P. Wei and W. Pan (2006). "Combining gene annotations and gene expression data in model-based clustering: weighted method." OMICS **10**(1): 28-39.

Jain, A. K. (1976). "Estimate of Bhattacharyya Distance." Ieee Transactions on Systems Man and Cybernetics **6**(11): 763-766.

Jiang, D., C. Tang and A. Zhang (2004). "Cluster Analysis for Gene Expression Data: A Survey." IEEE Transactions on Knowledge and Data Engineering **16**(11): 1370-1386.

Johnson, D. H. and S. Sinanovic (2001). "Symmetrizing the kullback-leibler distance." IEEE Transactions on Information Theory **1**(1): 1-10.

Johnson, R. A. and D. W. Wichern (2002). Applied multivariate statistical analysis. Upper Saddle River, N.J., Prentice Hall.

Kerby, A., D. Marx, A. Samal and V. Adamchuck (2007). "Spatial Clustering Using the Likelihood Function." Proceedings of the Sevents IEEE International Conference on Data Mining Workshops: 637-642.

Kon'kov, E. A., O. A. Morozov, E. A. Soldatov and V. R. Fidel'man (2007). "Application of the Kullback-Leibler measure for estimating the instants of a change in the statistical properties of a binary Markovian process." Journal of Communications Technology and Electronics **52**(12): 1350-1354.

Kubokawa, T. and H. Tsukuma (2007). "Estimation in a linear regression model under the Kullback-Leibler loss and its application to model selection." Journal of Statistical Planning and Inference **137**(7): 2487-2508.

Kullback, S. and R. A. Leibler (1951). "On Information and Sufficiency." Annals of Mathematical Statistics **22**(1): 79-86.

Leger, C., D. N. Politis and J. P. Romano (1992). "Bootstrap Technology and Applications." Technometrics **34**(4): 378-398.

Nikkila, J., P. Toronen, S. Kaski, J. Venna, E. Castren and G. Wong (2002). "Analysis and visualization of gene expression data using self-organizing maps." Neural Netw **15**(8-9): 953-966.

Rencher, A. C. (2002). Methods of multivariate analysis. New York, J. Wiley.

Rodrigues, J. (1992). "The Kullback-Leibler Approximation of the Marginal Posterior Density - an Application to the Linear Functional-Model." Communications in Statistics-Theory and Methods **21**(10): 2861-2868.

Schweppe, F. C. (1967). "On Bhattacharyya Distance and Divergence between Gaussian Processes." Information and Control **11**(4): 373-&.

Shahdoust, M., E. Hajizadeh, H. Mozdarani and A. Chehrei (2013). "Finding genes discriminating smokers from non-smokers by applying a growing self-organizing clustering method to large airway epithelium cell microarray data." Asian Pac J Cancer Prev **14**(1): 111-116.

Singh, K. (1981). "On the Asymptotic Accuracy of Efrons Bootstrap." Annals of Statistics **9**(6): 1187-1195.

Singh, K. (1986). "Jackknife, Bootstrap and Other Resampling Methods in Regression-Analysis - Discussion." Annals of Statistics **14**(4): 1328-1330.

Sprott, D. A. (1973). "Normal Likelihoods and Their Relation to Large Sample Theory of Estimation." Biometrika **60**(3): 457-465.

Sprott, D. A. and Kalbflei.Jd (1969). "Examples of Likelihoods and Comparison with Point Estimates and Large Sample Approximations." Journal of the American Statistical Association **64**(326): 468-&.

Sprott, D. A. and J. D. Kalbfleisch (1969). "Examples of Likelihoods and Comparison with Point Estimates and Large Sample Approximations." Journal of American Statistical Association **64**(326): 468-484.

Team, R. C. (2015). R: A language and environment for statistical computing. Vienna, Austria, R Foundation for Statistical Computing.

Toronen, P., M. Kolehmainen, G. Wong and E. Castren (1999). "Analysis of gene expression data using self-organizing maps." FEBS Lett **451**(2): 142-146.

van 't Veer, L. J., H. Dai, M. J. van de Vijver, Y. D. He, A. A. Hart, M. Mao, H. L. Peterse, K. van der Kooy, M. J. Marton, A. T. Witteveen, G. J. Schreiber, R. M. Kerkhoven, C. Roberts, P. S. Linsley, R. Bernards and

S. H. Friend (2002). "Gene expression profiling predicts clinical outcome of breast cancer." Nature **415**(6871): 530-536.

van de Vijver, M. J., Y. D. He, L. J. van't Veer, H. Dai, A. A. Hart, D. W. Voskuil, G. J. Schreiber, J. L. Peterse, C. Roberts, M. J. Marton, M. Parrish, D. Atsma, A. Witteveen, A. Glas, L. Delahaye, T. van der Velde, H. Bartelink, S. Rodenhuis, E. T. Rutgers, S. H. Friend and R. Bernards (2002). "A gene-expression signature as a predictor of survival in breast cancer." N Engl J Med **347**(25): 1999-2009.

Williams, J. B., D. Bradshaw and L. Schmidt (1995). "Field metabolism and water requirements of spinifex pigeons (Geophaps plumifera) in Western Australia." Australian Journal of Zoology **43**(1): 1-15.

Yeung, K. Y., C. Fraley, A. Murua, A. E. Raftery and W. L. Ruzzo (2001). "Model-based clustering and data transformations for gene expression data." Bioinformatics **17**(10): 977-987.

Zhang, J. and L. Shen (2014). "An improved fuzzy c-means clustering algorithm based on shadowed sets and PSO." Comput Intell Neurosci **2014**: 368628.

Zhang, M., B. Adamu, C. C. Lin and P. Yang (2011). "Gene expression analysis with integrated fuzzy C-means and pathway analysis." Conf Proc IEEE Eng Med Biol Soc **2011**: 936-939.

Appendix

Chapter 1 Codes

```
#-------------------------
# Part 1
#-------------------------

life_days <- c(2,NA,51,NA,33,27,14,24,4,NA)
test_per <- c(81,72,70,60,41,31,31,30,29,21)

data <- data.frame(life_days,test_per)

n <- length(life_days)
k <- n - sum(is.na(life_days))

data_na <- subset(data, is.na(data$life_days))
T <- c(na.omit(life_days),data_na[,2])
S <- sum(T)

theta <- seq(0.1,170,1)

# Expression for R_theta and Rn_theta
std_err <- (S/k) / sqrt(sum(1 - exp(-data$test_per/(S/k))))

R_theta1 <- (((S/k)/theta)^7)*exp(k - S/theta)
Rn_theta <- exp(-0.5*((theta - S/k)/std_err)^2)

# Area between the two curves
area_diff <- trapz(theta,R_theta1) - trapz(theta,Rn_theta)
area_diff

# Plotting the two rel. likld plots.
plot1 <- plot(theta,R_theta1,ylim = c(0,1),type = "l",lwd = 2,lty=1,col = 2
,xlab = "theta", ylab = "R")
plot2 <- lines(theta,Rn_theta,ylim = c(0,1),type = "l",lwd = 2,lty=2,col = 3)

#-------------------------
# Part 2
#-------------------------

require(pracma);library(MASS)

set.seed(132)
n.sim <- 15; lambda.sim <- 10
x <- rpois(n.sim,lambda.sim)
```

```
lambda <- seq(0.5,30,0.1)

rel.likld.pois <- exp(n.sim*(mean(x) - lambda))*(lambda/mean(x))^(n.sim*mean(x))
summary(rel.likld.pois)

norm.likld.pois <- exp( (-n.sim /(2*mean(x)))*((lambda - mean(x))^2) )
summary(norm.likld.pois)

# Area between the two curves
area_diff <- trapz(lambda,rel.likld.pois) - trapz(lambda,norm.likld.pois)
area_diff

# Computing the Dissimilarity index between two curves
sim.ind <- function(x,y)
{
        x <- as.vector(x)
        y <- as.vector(y)
        si <- sum(x*y) /(sqrt(sum(x*x))*sqrt(sum(y*y)))
        return(si)
}

y1 <- function(theta)
{
exp(n.sim*(mean(x) - theta))*(theta/mean(x))^(n.sim*mean(x))} #Rel likld

y2 <- function(theta)
{ exp( (-n.sim /(2*mean(x)))*((theta - mean(x))^2) ) } # Rel Normal likld

n <- 100
si <- NULL
for (i in 1:n-1){
seg <- seq(min(lambda),max(lambda),length = n)

v.x <- c(seg[i],seg[i+1])
v.y1 <- c(y1(seg[i]),y1(seg[i+1]))
v.y2 <- c(y2(seg[i]),y2(seg[i+1]))

l1 <- c(v.x[2] - v.x[1],v.y1[2] - v.y1[1])
l2 <- c(v.x[2] - v.x[1],v.y2[2] - v.y2[1])
si[i] <- sim.ind(l1,l2)
}
print(si)

diss.ind <- ((n-1) - (sum(si)))/(n-1)
diss <- diss.ind*100
diss

# Plotting the two relative likelihoods:
```

```
xlab <- expression(lambda);ylab <- "Relative Likelihood"

plot1 <- plot(lambda,rel.likld.pois,ylim = c(0,1),type = "l",lwd = 2,lty=1,
col = 4, xlab = xlab, ylab = ylab)
txt <- expression(paste("R(",lambda,") : Relative Likelihood"))
text(max(lambda/2),0.5,txt,col = 4, cex = 0.5)

plot2 <- lines(lambda,norm.likld.pois,ylim = c(0,1),type = "l",lwd = 2,lty=2,
col = 2)
txt <- expression(paste("Rn(",lambda,") : Normal Relative Likelihood"))
text(max(lambda/2),0.4,txt,col = 2, cex = 0.5)

txt <- paste("n = ", n.sim); text(max(lambda/1.5),0.9,txt,cex=0.8)
txt <- paste("Area Diff = ",round(area_diff,5)); text(max(lambda/1.5),0.8,txt,cex=0.8)
txt <- paste("Diss Index = ",round(diss,5)); text(max(lambda/1.5),0.7,txt,cex=0.8)
title("Poisson Distribution")

#-------------------------
# Part 3
#-------------------------

rm(list = ls())
set.seed(132)
n.sim <- 100; beta.sim <- 3; gamma = 2
x.sim <- rweibull(n.sim, shape = gamma, scale = exp(log(beta.sim)/gamma))
# Scale re-evaluated to match with casell and berger

rel.likld.weib <- function(beta)
{
        x.sim <- x.sim; gamma <- gamma
        val <- (beta.hat/beta)^(n.sim) * exp((1/beta.hat - 1/beta)*sum(x.sim^gamma))
        return(val)
}

rel.likld.weib.val <- NULL

for(i in 1:length(beta))
{
        rel.likld.weib.val[i] <- rel.likld.weib(beta[i])
}

norm.likld.weib <- function(beta)
{
        x.sim <- x.sim; gamma <- gamma
        I.theta <- (2/beta.hat^3)*sum(x.sim^gamma) - n.sim/beta.hat^2
        val <- exp(-0.5*I.theta*((beta-beta.hat)^2))
        return(val)
}
```

```
norm.likld.weib.val <- NULL

for(i in 1:length(beta))
{
        norm.likld.weib.val[i] <- norm.likld.weib(beta[i])
}
summary(norm.likld.weib.val)


# Area between the two curves
area_diff <- trapz(beta,rel.likld.weib.val) - trapz(beta,norm.likld.weib.val)
area_diff

# Computing the Dissimilarity index between two curves
sim.ind <- function(x,y)
{
        x <- as.vector(x)
        y <- as.vector(y)
        si <- sum(x*y) /(sqrt(sum(x*x))*sqrt(sum(y*y)))
        return(si)
}

y1 <- rel.likld.weib
y2 <- norm.likld.weib

n <- 100
si <- NULL
for (i in 1:n-1){
seg <- seq(min(beta),max(beta),length = n)

v.x <- c(seg[i],seg[i+1])
v.y1 <- c(y1(seg[i]),y1(seg[i+1]))
v.y2 <- c(y2(seg[i]),y2(seg[i+1]))

l1 <- c(v.x[2] - v.x[1],v.y1[2] - v.y1[1])
l2 <- c(v.x[2] - v.x[1],v.y2[2] - v.y2[1])
si[i] <- sim.ind(l1,l2)
}
print(si)

diss.ind <- ((n-1) - (sum(si)))/(n-1)
diss <- diss.ind*100
diss


# Plotting the two relative likelihoods:
xlab <- expression(beta)
```

```
ylab <- "Relative Likelihood"

plot1 <- plot(beta,rel.likld.weib.val,ylim = c(0,1),type = "l",lwd = 2,lty=1,
col = 4, xlab = xlab, ylab = ylab)
txt <- expression(paste("R(",beta,") : Relative Likelihood"))
text(max(beta/2),0.5,txt,col = 4, cex = 0.8)

plot2 <- lines(beta,norm.likld.weib.val,ylim = c(0,1),type = "l",lwd = 2,lty=1,
col = 2, xlab = xlab, ylab = ylab)
txt <- expression(paste("Rn(",beta,") : Normal Likelihood"))
text(max(beta/2),0.4,txt,col = 2, cex = 0.8)

txt <- bquote(n == .(n.sim)); text(max(beta/1.5),0.85,txt)
txt <- expression(paste(beta," = 6")); text(max(beta/1.5),0.8,txt)
txt <- paste("Area Diff = ",round(area_diff,5)); text(max(beta/1.5),0.75,txt)
txt <- paste("Diss Index = ",round(diss,5)); text(max(beta/1.5),0.70,txt)
title <- expression(paste("Weibull Distribution (",gamma, "=2")); title(title)

#-------------------------
# Part 4
#-------------------------

rm(list = ls())
set.seed(132)

n.sim <- 100;beta.sim <- 6

x <- rgamma(n.sim, shape = 1, scale = beta.sim)

beta.hat <- mean(x)
beta <- seq(0.1,12,0.1)

rel.likld.expo <- ((beta.hat/beta)^n.sim) * exp(sum(x)*(1/beta.hat - 1/beta))
norm.likld.expo <- exp( (-n.sim/(2*beta.hat^2)) * (beta - beta.hat)^2 )

# Area between the two curves
area_diff <- trapz(beta,rel.likld.expo) - trapz(beta,norm.likld.expo)
area_diff

# Computing the Dissimilarity index between two curves
sim.ind <- function(x,y)
{
        x <- as.vector(x)
        y <- as.vector(y)
        si <- sum(x*y) /(sqrt(sum(x*x))*sqrt(sum(y*y)))
        return(si)
}
```

```
y1 <- function(beta)
{((beta.hat/beta)^n.sim) * exp(sum(x)*(1/beta.hat - 1/beta)) } #Rel likld

y2 <- function(beta)
{ exp( (-n.sim/(2*beta.hat^2)) * (beta - beta.hat)^2 ) } # Rel Normal likld

n <- 100; si <- NULL
for (i in 1:n-1){
seg <- seq(min(beta),max(beta),length = n)

v.x <- c(seg[i],seg[i+1])
v.y1 <- c(y1(seg[i]),y1(seg[i+1]))
v.y2 <- c(y2(seg[i]),y2(seg[i+1]))

l1 <- c(v.x[2] - v.x[1],v.y1[2] - v.y1[1])
l2 <- c(v.x[2] - v.x[1],v.y2[2] - v.y2[1])
si[i] <- sim.ind(l1,l2)
}
print(si)

diss.ind <- ((n-1) - (sum(si)))/(n-1)
diss <- diss.ind*100
diss

# Plotting the two relative likelihoods:
xlab <- expression(beta)
ylab <- "Relative Likelihood"

plot1 <- plot(beta,rel.likld.expo,ylim = c(0,1),type = "l",lwd = 2,lty=1,
col = 4, xlab = xlab, ylab = ylab)
txt <- expression(paste("R(",beta,") : Relative Likelihood"))
text(max(beta/2),0.5,txt,col = 4, cex = 0.8)

plot2 <- lines(beta,norm.likld.expo,ylim = c(0,1),type = "l",lwd = 2,lty=1,
col = 2, xlab = xlab, ylab = ylab)
txt <- expression(paste("Rn(",beta,") : Normal Likelihood"))
text(max(beta/2),0.4,txt,col = 2, cex = 0.8)

txt <- bquote(n == .(n.sim));text(max(beta/1.5),0.85,txt)
txt <- expression(paste(beta," = 6"));text(max(beta/1.5),0.8,txt)
txt <- paste("Area Diff = ",round(area_diff,5));text(max(beta/1.5),0.75,txt)
txt <- paste("Diss Index = ",round(diss,5));text(max(beta/1.5),0.70,txt)
title("Exponential Distribution")

#-------------------------
# Part 5
#-------------------------
rm(list = ls())
```

```
set.seed(132)

# x <- c(5,6,7,7,8,8,8,9,12)
x <- c(7,8,8,8,9,9,12,13,14,17)
n.sim <- length(x); lambda.sim <- mean(x)

lambda <- seq(0.5,20,0.1)

rel.likld.pois <- exp(n.sim*(mean(x) - lambda))*(lambda/mean(x))^(n.sim*mean(x))
norm.likld.pois <- exp( (-n.sim /(2*mean(x)))*((lambda - mean(x))^2) )

# Area between the two curves
area_diff <- trapz(lambda,rel.likld.pois) - trapz(lambda,norm.likld.pois)

# Computing the Dissimilarity index between two curves

sim.ind <- function(x,y)
{
        x <- as.vector(x)
        y <- as.vector(y)
        si <- sum(x*y) /(sqrt(sum(x*x))*sqrt(sum(y*y)))
        return(si)
}

y1 <- function(theta)
{ exp(n.sim*(mean(x) - theta))*(theta/mean(x))^(n.sim*mean(x))} #Rel likld
y1(lambda)

y2 <- function(theta)
{ exp( (-n.sim /(2*mean(x)))*((theta - mean(x))^2) ) } # Rel Normal likld
y2(lambda)

n <- 30
si <- NULL
for (i in 1:n-1){
seg <- seq(min(lambda),max(lambda),length = n)

v.x <- c(seg[i],seg[i+1])
v.y1 <- c(y1(seg[i]),y1(seg[i+1]))
v.y2 <- c(y2(seg[i]),y2(seg[i+1]))

l1 <- c(v.x[2] - v.x[1],v.y1[2] - v.y1[1])
l2 <- c(v.x[2] - v.x[1],v.y2[2] - v.y2[1])
si[i] <- sim.ind(l1,l2)
}
print(si)

diss.ind <- ((n-1) - (sum(si)))/(n-1)
```

```
diss <- diss.ind*100;diss

# Plotting the two relative likelihoods:
xlab <- expression(lambda);ylab <- "Relative Likelihood"
plot1 <- plot(lambda,rel.likld.pois,ylim = c(0,1),type = "l",lwd = 2,lty=1,
col = 4, xlab = xlab, ylab = ylab)
plot2 <- lines(lambda,norm.likld.pois,ylim = c(0,1),type = "l",lwd = 2,lty=2,
col = 2)

txt <- expression(paste("Rn(",lambda,") : Normal Relative
Likelihood"));text(max(lambda/2),0.4,txt,col = 2, cex = 0.5)
txt <- expression(paste("R(",lambda,") : Relative Likelihood"));text(max(lambda/2),0.5,txt,col =
4, cex = 0.5)
txt <- paste("n = ", n.sim);text(max(lambda/1.5),0.9,txt,cex=0.8)
txt <- paste("Area Diff = ",round(area_diff,5));text(max(lambda/1.5),0.8,txt,cex=0.8)
txt <- paste("Diss Index = ",round(diss,5));text(max(lambda/1.5),0.7,txt,cex=0.8)
title("Poisson Distribution Gibbons et al Data")

#-------------------------
# Part 6
#-------------------------

rm(list = ls())
set.seed(132)
data <- matrix(c(15,21,29,16,18,21,16,26,33,27,41,60,33,38,41,20,27,42),ncol = 6)
colnames(data) <- c("Dose 0","Dose 10","Dose 33","Dose 100","Dose 333","Dose 1000")
m <- 1
x <- data[,m]
n.sim <- length(x)

lambda <- seq(10,80,0.1)

rel.likld.pois <- exp(n.sim*(mean(x) - lambda))*(lambda/mean(x))^(n.sim*mean(x))
norm.likld.pois <- exp( (-n.sim /(2*mean(x)))*((lambda - mean(x))^2) )

# Area between the two curves
area_diff <- trapz(lambda,rel.likld.pois) - trapz(lambda,norm.likld.pois)

# Computing the Dissimilarity index between two curves
sim.ind <- function(x,y)
{
        x <- as.vector(x)
        y <- as.vector(y)
        si <- sum(x*y) /(sqrt(sum(x*x))*sqrt(sum(y*y)))
        return(si)
}

y1 <- function(theta)
```

```
{ exp(n.sim*(mean(x) - theta))*(theta/mean(x))^(n.sim*mean(x))} #Rel likld
y1(lambda)

y2 <- function(theta)
{ exp( (-n.sim /(2*mean(x)))*((theta - mean(x))^2) ) } # Rel Normal likld
y2(lambda)

n <- 30
si <- NULL
for (i in 1:n-1){
seg <- seq(min(lambda),max(lambda),length = n)

v.x <- c(seg[i],seg[i+1])
v.y1 <- c(y1(seg[i]),y1(seg[i+1]))
v.y2 <- c(y2(seg[i]),y2(seg[i+1]))

l1 <- c(v.x[2] - v.x[1],v.y1[2] - v.y1[1])
l2 <- c(v.x[2] - v.x[1],v.y2[2] - v.y2[1])
si[i] <- sim.ind(l1,l2)
}
print(si)

diss.ind <- ((n-1) - (sum(si)))/(n-1)
diss <- diss.ind*100
diss

# Plotting the two relative likelihoods:
xlab <- expression(lambda);ylab <- "Relative Likelihood"

plot1 <- plot(lambda,rel.likld.pois,ylim = c(0,1),type = "l",lwd = 2,lty=1,
col = 4, xlab = xlab, ylab = ylab)
txt <- expression(paste("R(",lambda,") : Relative Likelihood"))
text(max(lambda/2),0.5,txt,col = 4, cex = 0.5)

plot2 <- lines(lambda,norm.likld.pois,ylim = c(0,1),type = "l",lwd = 2,lty=2,
col = 2)
txt <- expression(paste("Rn(",lambda,") : Normal Relative Likelihood"))
text(max(lambda/2),0.4,txt,col = 2, cex = 0.5)

txt <- paste("n = ", n.sim)
text(max(lambda/1.5),0.9,txt,cex=0.8)
txt <- paste("Area Diff = ",round(area_diff,5))
text(max(lambda/1.5),0.8,txt,cex=0.8)
txt <- paste("Diss Index = ",round(diss,5))
text(max(lambda/1.5),0.7,txt,cex=0.8)
title("Data (Breslow et al)")
```

```
#------------------------
# Part 7
#------------------------

rm(list = ls())
set.seed(132)

# Data Source: Williams, Bradshaw, Schmidt: Australian Journal of Zoology
x <-
c(0.457,3.751,0.238,2.967,2.509,1.384,1.454,0.818,0.335,1.436,1.603,1.309,0.201,0.530,2.144,0
.834)
x <- sample(x,100,replace = TRUE)

n.sim <- length(x)
beta.hat <- mean(x)
beta <- seq(0.1,5,0.01)

rel.likld.expo <- ((beta.hat/beta)^n.sim) * exp(sum(x)*(1/beta.hat - 1/beta))
norm.likld.expo <- exp( (-n.sim/(2*beta.hat^2)) * (beta - beta.hat)^2 )

# Area between the two curves
area_diff <- trapz(beta,rel.likld.expo) - trapz(beta,norm.likld.expo)

# Computing the Dissimilarity index between two curves

sim.ind <- function(x,y)
{
        x <- as.vector(x)
        y <- as.vector(y)
        si <- sum(x*y) /(sqrt(sum(x*x))*sqrt(sum(y*y)))
        return(si)
}

y1 <- function(beta)
{ ((beta.hat/beta)^n.sim) * exp(sum(x)*(1/beta.hat - 1/beta)) } #Rel likld
y1(beta)

y2 <- function(beta)
{ exp( (-n.sim/(2*beta.hat^2)) * (beta - beta.hat)^2 ) } # Rel Normal likld
y2(beta)

n <- 100
si <- NULL
for (i in 1:n-1){
seg <- seq(min(beta),max(beta),length = n)

v.x <- c(seg[i],seg[i+1])
v.y1 <- c(y1(seg[i]),y1(seg[i+1]))
```

```
v.y2 <- c(y2(seg[i]),y2(seg[i+1]))

l1 <- c(v.x[2] - v.x[1],v.y1[2] - v.y1[1])
l2 <- c(v.x[2] - v.x[1],v.y2[2] - v.y2[1])
si[i] <- sim.ind(l1,l2)
}
print(si)

diss.ind <- ((n-1) - (sum(si)))/(n-1)
diss <- diss.ind*100;diss

# Plotting the two relative likelihoods:
xlab <- expression(beta);ylab <- "Relative Likelihood"

plot1 <- plot(beta,rel.likld.expo,ylim = c(0,1),type = "l",lwd = 2,lty=1,
col = 4, xlab = xlab, ylab = ylab)
txt <- expression(paste("R(",beta,") : Relative Likelihood"))
text(max(beta/2),0.5,txt,col = 4, cex = 0.8)

plot2 <- lines(beta,norm.likld.expo,ylim = c(0,1),type = "l",lwd = 2,lty=1,
col = 2, xlab = xlab, ylab = ylab)
txt <- expression(paste("Rn(",beta,") : Normal Likelihood"))
text(max(beta/2),0.4,txt,col = 2, cex = 0.8)

txt <- bquote(n == .(n.sim));text(max(beta/1.5),0.85,txt)
txt <- paste("Area Diff = ",round(area_diff,5));text(max(beta/1.5),0.75,txt)
txt <- paste("Diss Index = ",round(diss,5));text(max(beta/1.5),0.70,txt)
title("Bootstrapped Distribution (Williams et al)")
```

Chapter 2 Codes

```
gc();rm(list = ls())
library(xlsx);library(clues)

# reading in simulated dataset
dat.sim = read.xlsx("S:\\Biostats\\BIO-STAT\\Brimacombe\\Dissertation\\Paper 2 Simulated
Study\\Simulated Data Salmonollesis 2012.xlsx",1)
rownames(dat.sim) <- dat.sim[,2]

# Including counties with counts of at least 3.
dat.use <- dat.sim[dat.sim[,3] >= 3,-c(1,2)]
dat.use[1:5,1:8]

#---------------------------------------------------------------
# Constructing the log relative likelihood function for each state:
#---------------------------------------------------------------
wt.rel.pois <- function(lambda)
{
        x <- x;n <- length(x);lambda.hat <- mean(x)
        wt.rel.log <- n*(lambda.hat - lambda) + log(n/lambda.hat) +
n*lambda.hat*log(lambda/lambda.hat)
        wt.rel <- exp(wt.rel.log)
        return(wt.rel)
}

#---------------------------------------------------------------
# Plot of Weighted relative likld function for each county.
#---------------------------------------------------------------
for (j in 1:dim(dat.use)[1])
{
        x <- as.numeric(dat.use[j,-c(1,2)])
        lambda <- seq(0,400,by = 0.1)
        rel.likld <- (wt.rel.pois(lambda))
        xlab <- expression(lambda)
        ylab <- "Weighted Relative Likelihoods"
        if (j == 1){ plot(lambda,rel.likld,type = "l",xlab = xlab,ylab=ylab,ylim = c(0,5))}
        else { lines(lambda,rel.likld,xlab = xlab) }
}
abline(h=1,col="gray",lwd=2)
title("Wt. Rel. Likld. Functions of Counties in KS")

#---------------------------------------------------------------
# Analysis with Weighted relative likld function for each county.
#---------------------------------------------------------------
lambda <- seq(0,400,by = 0.1)
mat.pca <- matrix(NA,nrow = dim(dat.use)[1],ncol = length(lambda))
```

```
rownames(mat.pca) <- rownames(dat.use)
colnames(mat.pca) <- 1:length(lambda)

# Log of relative likld function for each state.
for (j in 1:dim(dat.use)[1])
{
        x <- as.numeric(dat.use[j,-c(1,2)])
        mat.pca[j,]<- (wt.rel.pois(lambda))
}
mat.pca[1:5,1:5]
matplot(lambda,t(mat.pca),type="l",col=1,lty=1,xlab=xlab,ylab=ylab)
abline(h=1)

set.seed(132)
fit.kmeans <- kmeans(mat.pca,7)
groups <- fit.kmeans$cluster
sort(groups);table(groups)

# Plot with cluster color
xlab <- expression(lambda);ylab <- "Weighted Relative Likelihoods"
matplot(lambda,t(mat.pca),type="l",col=groups,lty=groups,xlab=xlab,ylab=ylab)
abline(h=1,col="gray",lwd=2)

dat.ord <- cbind(dat.use[,c(1,2)],groups)
dat.ord[order(dat.ord[,3]),]

# k means clustering
wssplot <- function(data, nc=15, seed=1234)
{
        wss <- (nrow(data)-1)*sum(apply(data,2,var))
                for (i in 2:nc)
                {
                set.seed(seed)
                wss[i] <- sum(kmeans(data, centers=i)$withinss)
                }
        plot(1:nc, wss, type="b", xlab="Number of Clusters",
        ylab="Within groups sum of squares",lwd=2)
}
wssplot(mat.pca,30)

##############################################
# Analysis with simulated Poisson Data
##############################################

rm(list=ls())
# Function to simulate Data from Poisson distribution
sim.pois <- function(lambda,n.obs,n.val)
{
```

```
	dat.sim <- vector("list",length(lambda))
	for (i in 1:length(lambda))
	{
		#set.seed(sample(.Random.seed,1))
		X <- matrix(NA,nrow = n.obs, ncol = n.val)
		for (j in 1:n.obs)
		{
			X[j,] <- rpois(n.val,lambda[i])
		}
	dat.sim[[i]] <- X
	}

dat.sim <- do.call(rbind,dat.sim)
rownames(dat.sim) <- 1:dim(dat.sim)[1]
return(dat.sim)
}

#-----------------------------------------------------------------
# Constructing the log relative likelihood function
#-----------------------------------------------------------------
wt.rel.pois <- function(lambda)
{
	x <- x;n <- length(x);lambda.hat <- mean(x)
	wt.rel.log <- n*(lambda.hat - lambda) + log(n/lambda.hat) +
n*lambda.hat*log(lambda/lambda.hat)
	wt.rel <- exp(wt.rel.log)
	return(wt.rel)
}

#----------------------------
lam2 <- seq(0.5,5,by = 0.3)
err.wt <- matrix(NA,nrow=length(lam2),ncol=30)
err.km <- matrix(NA,nrow=length(lam2),ncol=30)

for(q in 1:length(lam2)){
l1 <- 5; l2 <- l1 + lam2[q]

temp1 <- NULL; temp2 <- NULL
for (u in 1:30)
{
	lambda.sim <- c(l1,l2); n.obs <- 30; n.val <- 15
	dat.sim <- sim.pois(lambda.sim,n.obs,n.val)
	dim(dat.sim);dat.sim[1:5,1:5]

	lambda <- seq(2.1,15,len=100)
	mat.sim <- matrix(NA,ncol = length(lambda),nrow = dim(dat.sim)[1])
	colnames(mat.sim) <- 1:length(lambda)
	rownames(mat.sim) <-  1:(2*n.obs)
```

```
        for (i in 1:dim(dat.sim)[1])
        {
                x <- dat.sim[i,]
                mat.sim[i,] <- wt.rel.pois(lambda)
        }

        fit.kmeans1 <- kmeans(mat.sim,2)
        clust1 <- as.numeric(names(fit.kmeans1$cluster[fit.kmeans1$cluster == 1]))
        clust2 <- as.numeric(names(fit.kmeans1$cluster[fit.kmeans1$cluster == 2]))
        a1 <- setdiff(clust1,1:30);a2 <- setdiff(clust1,31:60)
        b1 <- setdiff(clust2,1:30);b2 <- setdiff(clust2,31:60)
        temp1[u] <- sum(min(length(a1),length(a2)),min(length(b1),length(b2)))

        #matplot(lambda,t(mat.sim),type="l",col = fit.kmeans1$cluster)

        fit.kmeans2 <- kmeans(dat.sim,2)
        clust1 <- as.numeric(names(fit.kmeans2$cluster[fit.kmeans2$cluster == 1]))
        clust2 <- as.numeric(names(fit.kmeans2$cluster[fit.kmeans2$cluster == 2]))
        a1 <- setdiff(clust1,1:30);a2 <- setdiff(clust1,31:60)
        b1 <- setdiff(clust2,1:30);b2 <- setdiff(clust2,31:60)
        temp2[u] <- sum(min(length(a1),length(a2)),min(length(b1),length(b2)))

}

err.wt[q,] <- temp1
err.km[q,] <- temp2
}

#--------------------------------------
# Plot of misclassification error rate.
#--------------------------------------
misclass.wt <- round(err.wt/60,3)
misclass.km <- round(err.km/60,3)
mean.wt <- apply(misclass.wt,1,mean)
mean.km <- apply(misclass.km,1,mean)
xlab <- expression(paste(theta[2]," - ",theta[1]))
ylim = c(min(c(mean.wt,mean.km)),max(c(mean.wt,mean.km)))
plot(lam2,mean.wt,pch=16,col=1,ylim=ylim,ylab="Misclassification
Rate",xlab=xlab,type="b",lwd=2)
lines(lam2,mean.km,pch=15,col=4,ylab="Misclassification Rate",xlab=xlab,type="b",lwd=2)
txt <- c("Wt. Rel. Likld","Traditional")
legend("bottomleft",txt,pch=c(16,15),col=c(1,4))
title("Poisson Distribution")



##########################################
```

```
# Analysis with simulated Cauchy Data
##########################################
rm(list=ls());gc()

# Function to simulate data from Cauchy distribution
sim.cauchy <- function(theta,n.obs,n.val)
{
        dat.sim <- vector("list",length(theta))
        for (i in 1:length(theta))
        {
                #set.seed(sample(.Random.seed,1))
                X <- matrix(NA,nrow = n.obs, ncol = n.val)
                for (j in 1:n.obs)
                {
                        X[j,] <- rcauchy(n.val,theta[i])
                }
        dat.sim[[i]] <- X
        }

dat.sim <- do.call(rbind,dat.sim)
rownames(dat.sim) <- 1:dim(dat.sim)[1]
return(dat.sim)
}

#-----------------------------------------------------------------
# Constructing the wt. relative likelihood function for each state:
#-----------------------------------------------------------------
rel.likld.cauchy <- function(theta,x)
{
        x <- y
        likld.cauchy <- function(theta,x)
        {
        x <- y; n <- length(x);
        log.lik <- -n*log(pi) - sum(log(1 + (x-theta)^2))
        return(exp(log.lik))
        }
        likld.cauchy <- Vectorize(likld.cauchy,"theta","x")

mle.theta <- optimize(likld.cauchy,c(-20,20),maximum=TRUE)$maximum

a <- (x-mle.theta)
wt <- 4*((sum(a/(1-a^2)))^2)
out <- wt*(likld.cauchy(theta)/likld.cauchy(mle.theta))
return(log(out))
}

#-----------------------------------------------------------------
lam2 <- seq(1,7,by=1)
```

```
err.wt <- matrix(NA,nrow=length(lam2),ncol=30)
err.km <- matrix(NA,nrow=length(lam2),ncol=30)

for(q in 1:length(lam2)) {
l1 <- 5; l2 <- l1 + lam2[q]

temp1 <- NULL; temp2 <- NULL
for (u in 1:30)
{
	theta.sim <- c(l1,l2); n.obs <- 30; n.val <- 15
	dat.sim <- sim.cauchy(theta.sim,n.obs,n.val)
	dim(dat.sim);dat.sim[1:5,1:5]

	theta <- seq(0,20,0.1)
	mat.sim <- matrix(NA,ncol = length(theta),nrow = dim(dat.sim)[1])
	colnames(mat.sim) <- 1:length(theta)
	rownames(mat.sim) <-  1:(2*n.obs)

	for(i in 1:dim(dat.sim)[1])
	{
		y <- dat.sim[i,]
		mat.sim[i,] <- rel.likld.cauchy(theta,x)
	}

	dim(mat.sim)

	fit.kmeans1 <- kmeans(mat.sim,2)
	clust1 <- as.numeric(names(fit.kmeans1$cluster[fit.kmeans1$cluster == 1]))
	clust2 <- as.numeric(names(fit.kmeans1$cluster[fit.kmeans1$cluster == 2]))
	a1 <- setdiff(clust1,1:30);a2 <- setdiff(clust1,31:60)
	b1 <- setdiff(clust2,1:30);b2 <- setdiff(clust2,31:60)
	temp1[u] <- sum(min(length(a1),length(a2)),min(length(b1),length(b2)))

	fit.kmeans2 <- kmeans(dat.sim,2)
	clust1 <- as.numeric(names(fit.kmeans2$cluster[fit.kmeans2$cluster == 1]))
	clust2 <- as.numeric(names(fit.kmeans2$cluster[fit.kmeans2$cluster == 2]))
	a1 <- setdiff(clust1,1:30);a2 <- setdiff(clust1,31:60)
	b1 <- setdiff(clust2,1:30);b2 <- setdiff(clust2,31:60)
	temp2[u] <- sum(min(length(a1),length(a2)),min(length(b1),length(b2)))

}

err.wt[q,] <- temp1
err.km[q,] <- temp2
}

#-------------------------------------
```

```
# Plot of misclassification error rate.
#-------------------------------------
misclass.wt <- round(err.wt/60,3)
misclass.km <- round(err.km/60,3)

mean.wt <- apply(misclass.wt,1,mean)
mean.km <- apply(misclass.km,1,mean)

diff <- 1:length(lam2)
xlab <- expression(paste(theta[2]," - ",theta[1]))
ylim = c(min(c(mean.wt,mean.km)),max(c(mean.wt,mean.km)))
plot(diff,mean.wt,pch=16,col=1,ylab="Misclassification
Rate",xlab=xlab,type="b",lwd=2,ylim=ylim)
lines(diff,mean.km,pch=15,col=4,ylab="Misclassification Rate",xlab=xlab,type="b",lwd=2)
txt <- c("Wt. Rel. Likld","Traditional")
legend("bottomleft",txt,pch=c(16,15),col=c(1,4))
title("Cauchy Distribution")

##########################################
# Analysis with simulated t Data
##########################################

options(expressions=5e5)
rm(list = ls()); gc()

sim.t <- function(shift,n.obs,n.val)
{
        dat.sim <- vector("list",length(shift))
        for (i in 1:length(shift))
        {
                set.seed(sample(.Random.seed,1))
                X <- matrix(NA,nrow = n.obs, ncol = n.val)
                for (j in 1:n.obs)
                {
                        X[j,] <- rt(n.val,df) + shift[i]
                }
        dat.sim[[i]] <- X
        }

dat.sim <- do.call(rbind,dat.sim)
rownames(dat.sim) <- 1:dim(dat.sim)[1]
dat.sim <- dat.sim[sample(rownames(dat.sim),replace=FALSE),]
return(dat.sim)
}

rel.likld.t <- function(shift,x)
{
        x <- y; df <- df
```

```
	likld.t <- function(shift,x)
	{
		x <- y; df <- df
		n <- length(x);
		k1 <- exp(lgamma((df+1)/2))/(exp(lgamma(df/2))*sqrt(pi*df))
		k2 <- (df+1)/2
		log.likld <- n*log(k1) - k2*sum(log(1+(x-shift)^2/df))
		return(exp(log.likld))
	}
	likld.t <- Vectorize(likld.t,"shift","x")
	mle.shift <- optimize(likld.t,c(-20,20),maximum=TRUE)$maximum

	k2 <- (df+1)/2
	wt <- 4*(k2^2)*sum((x-shift)/(df + (x-shift)^2))^2
	rel.likld <- wt*(likld.t(shift,x)/likld.t(mle.shift))
	return(rel.likld)
}

#---------------------------------------------------------------
#---------------------------------------------------------------
err.wt <- matrix(NA,nrow=5,ncol=30)
rownames(err.wt) <- 1:5

err.km <- matrix(NA,nrow=5,ncol=30)
rownames(err.km) <- 1:5

df <- 4
for (iter.out in 1:2){

l1 <- 5; l2 <- l1 + iter.out
store1 <- rep(NA,2); store2 <- rep(NA,2)
for (iter in 1:30)
{
	gc()
	shift.sim <- c(l1,l2); n.obs <- 30; n.val <- 15
	dat.sim <- sim.t(shift.sim,n.obs,n.val)

	shift <- seq(-20,20,len=50)
	mat.sim <- matrix(NA,nrow = dim(dat.sim)[1],ncol = length(shift))
	colnames(mat.sim) <- 1:length(shift)
	rownames(mat.sim) <-  1:(2*n.obs)

	# Constructing the Distance matrix

	for (i in 1:dim(dat.sim)[1])
	{
		y <- dat.sim[i,]
		rel.likld.t <- Vectorize(rel.likld.t,"shift","x")
```

```
                mat.sim[i,] <- rel.likld.t(shift)
        }

        dim(mat.sim)

        fit.kmeans <- kmeans(mat.sim,2)
        clust.save <- cbind(as.factor(rownames(dat.sim)),fit.kmeans$cluster)
        clust.ord <- clust.save[order(clust.save[,1],clust.save[,2]),]
        a1 <- setdiff(subset(clust.ord,clust.ord[,2]==1)[,1],31:60);a2 <-
setdiff(subset(clust.ord,clust.ord[,2]==2)[,1],1:30)
        b1 <- setdiff(subset(clust.ord,clust.ord[,2]==2)[,1],31:60);b2 <-
setdiff(subset(clust.ord,clust.ord[,2]==1)[,1],1:30)
        store1[iter] <- min(sum(length(a1),length(a2)),sum(length(b1),length(b2)))

        fit.kmeans <- kmeans(dat.sim,2)
        clust.save <- cbind(as.factor(rownames(dat.sim)),fit.kmeans$cluster)
        clust.ord <- clust.save[order(clust.save[,1],clust.save[,2]),]
        a1 <- setdiff(subset(clust.ord,clust.ord[,2]==1)[,1],31:60);a2 <-
setdiff(subset(clust.ord,clust.ord[,2]==2)[,1],1:30)
        b1 <- setdiff(subset(clust.ord,clust.ord[,2]==2)[,1],31:60);b2 <-
setdiff(subset(clust.ord,clust.ord[,2]==1)[,1],1:30)
        store2[iter] <- min(sum(length(a1),length(a2)),sum(length(b1),length(b2)))

}

err.wt[iter.out,] <- store1
err.km[iter.out,] <- store2

}

loc <- "S:/Biostats/BIO-STAT/Brimacombe/Dissertation/Paper2/R rpgm/Simulation K means"
write.table(err.wt,paste(loc,"/df=",df,"Wt.txt",sep=""))
write.table(err.km,paste(loc,"/df=",df,"km.txt",sep=""))


#---------------------------------------
# Plot of misclassification error rate.
#---------------------------------------
misclass.wt <- round(err.wt/60,3)
mean.wt <- apply(misclass.wt,1,mean)
misclass.km <- round(err.km/60,3)
mean.km <- apply(misclass.km,1,mean)

diff <- sort(rep(1:5,30))
xlab <- expression(paste(theta[2]," - ",theta[1]))
ylim <- c(min(c(mean.wt,mean.km)),max(c(mean.wt,mean.km)))
plot(1:2,mean.wt,pch="*",col=6,cex=3,type="b",lwd=2,ylab="Misclassification
Rate",xlab=xlab,ylim = ylim)
```

```
points(1:2,mean.km,pch="*",col=4,cex=3,type="b",lwd=2,ylab="Misclassification
Rate",xlab=xlab)
title("t Distribution")
```

Chapter 3 Codes

```
library(xlsx);
library(clues);library(survival);library(pROC);library(irr);library(caret);library(nortest)

# Reading in data filed
loc <- "S:\\Biostats\\BIO-STAT\\Brimacombe\\Dissertation\\Paper 3 Genomics
Application\\VanDeVijer Data Analysis\\Dataset\\Expression Data\\Log ratio.txt"
dat.allgene <- read.table(loc,sep="\t")
dat.allgene[1:5,1:5]

dat.clin <- read.xlsx("S:\\Biostats\\BIO-STAT\\Brimacombe\\Dissertation\\Paper 3 Genomics
Application\\VanDeVijer Data Analysis\\Dataset\\Clinical Data.xlsx",1)
dat.clin[1:5,1:7]

dat.70gene <- read.xlsx("S:\\Biostats\\BIO-STAT\\Brimacombe\\Dissertation\\Paper 3 Genomics
Application\\VanDeVijer Data Analysis\\Dataset\\70 gene.xlsx",1)
dat.70gene[1:5,]

rm(list = ls()[!(ls() %in% c("dat.allgene","dat.clin","dat.70gene"))])
gc();ls()

dat.subset <- dat.allgene[ which(as.factor(dat.allgene[,1]) %in% as.factor(dat.70gene[,1])), ]
class(dat.subset); dim(dat.subset); dat.subset[1:5,1:7]

dat.use <- data.matrix(t(dat.subset[,-c(1,2)])); dim(dat.use);
dat.use[1:5,1:5]

#----------------------------
# Test of normality for each sample

# Across all genes
# dat.use <- t(na.omit((dat.allgene[,-c(1,2)])))
dat.use[1,]
sample.normal <- matrix(NA,ncol = 2, nrow = dim(dat.use)[1])
sample.normal[,1] <- dat.clin[,2]
rownames(sample.normal) <- rownames(dat.use)
for (i in 1:dim(dat.use)[1])
{
        sample.normal[i,2] <- shapiro.test(dat.use[i,])$p.value
}
sample.normal <- sample.normal[sample.normal[,2] > 0.01,]
summary(sample.normal[,2]);dim(sample.normal)
dat.use <- dat.use[rownames(dat.use) %in% rownames(sample.normal),]
dat.use[1:5,1:5];dim(dat.use)

#----------------------------
```

```
# Correlation across genes
df1 <- dat.use; df2 <- cor(df1); dim(df2)
hc = findCorrelation(df2, cutoff=0.6) # putt any value as a "cutoff"
hc = sort(hc)
dat.use = df1[,-c(hc)]
dim (dat.use)
cor(dat.use)[1:5,1:5]

# Summary Statistics
samp.id <- matrix(unlist(strsplit(rownames(dat.use),"Log.Ratio.S.")),ncol=2,byrow=TRUE)[,2]
dat.clin.use <- dat.clin[dat.clin$SampleID %in% as.vector(samp.id),]
dat.clin.use[1:5,1:5]

table(dat.clin.use$ind.prog)
summary(dat.clin.use[dat.clin.use$ind.prog == 1,]$TIMEsurvival)
table(dat.clin.use[dat.clin.use$ind.prog == 1,]$EVENTdeath)
summary(dat.clin.use[dat.clin.use$ind.prog == 0,]$TIMEsurvival)
table(dat.clin.use[dat.clin.use$ind.prog == 0,]$EVENTdeath)

summary(dat.clin.use[dat.clin.use$ind.prog == 1,]$TIMEmeta)
table(dat.clin.use[dat.clin.use$ind.prog == 1,]$EVENTmeta)
summary(dat.clin.use[dat.clin.use$ind.prog == 0,]$TIMEmeta)
table(dat.clin.use[dat.clin.use$ind.prog == 0,]$EVENTmeta)

#-----------------------------------
wt.rel.norm <- function(x,mu)
{
        n <- length(x); sig.hat <- var(x); mu.hat <- mean(x)
        rel <- exp( -0.5*(n/sig.hat)*((mu - mu.hat)^2) )
        wt <- n/sig.hat
        wt.rel <- wt*rel
        return(wt.rel)
}

#-----------------------------------
len <- 1000
mat.wt <- matrix(NA,nrow = dim(dat.use)[1],ncol = len)
rownames(mat.wt) <- rownames(dat.use)
for (i in 1:dim(dat.use)[1])
{
        x <- dat.use[i,]
        mu <- seq(-0.4,0.3,len=len)
        y <- wt.rel.norm(x,mu)
        mat.wt[i,] <- y
}
mat.wt[1:5,1:5]
xlab <- expression(theta); ylab <- ""
matplot(mu,t(mat.wt),type="l",col = 1, lty=1,xlab=xlab,ylab=ylab)
```

```
abline(h=1,lwd=2,col="gray")

set.seed(132)
fit.kmeans <- kmeans(mat.wt,2)
table(fit.kmeans$cluster)
xlab <- expression(theta); ylab <- ""
matplot(mu,t(mat.wt),type="l",col = fit.kmeans$cluster,lty =
fit.kmeans$cluster,xlab=xlab,ylab=ylab)

Cluster <- as.factor(fit.kmeans$cluster)
dat.mem <- cbind(dat.clin.use,Cluster)
dat.mem[1:5,c(1:3,16:18)]
table(dat.mem$ind.prog,Cluster)

# Summary Statistics by clusters
summary(dat.mem[dat.mem$Cluster == 1,]$TIMEsurvival)
table(dat.mem[dat.mem$Cluster == 1,]$EVENTdeath)
summary(dat.mem[dat.mem$Cluster == 2,]$TIMEsurvival)
table(dat.mem[dat.mem$Cluster == 2,]$EVENTdeath)

summary(dat.mem[dat.mem$Cluster == 1,]$TIMEmeta)
table(dat.mem[dat.mem$Cluster == 1,]$EVENTmeta)
summary(dat.mem[dat.mem$Cluster == 2,]$TIMEmeta)
table(dat.mem[dat.mem$Cluster == 2,]$EVENTmeta)

par(mfrow = c(1,2))
# K-M survival plots  and logrank test with kmeans using overall survival time
fit.km <- survfit(Surv(TIMEsurvival,EVENTdeath == 1) ~ Cluster,data = dat.mem)
plot(fit.km,col= 1:max(fit.kmeans$cluster),xlab = "time (years)",main = "Overall Survival
Time")
test <- Surv(dat.mem$TIMEsurvival,dat.mem$EVENTdeath  == 1)
log.rank.stage <- pchisq(survdiff(test~Cluster,data=dat.mem)$chisq,3,lower.tail=F)
txt <- substitute(paste("P-val = ",a),list(a=log.rank.stage))
text(10,0.4,txt)
legend("bottomleft",c("Clust 1","Clust 2"),text.col = 1:max(fit.kmeans$cluster))

# K-M survival plots  and logrank test with kmeans using metastates
dat.mem$TIMEmeta[is.na(dat.mem$TIMEmeta)] <- 20
fit.km <- survfit(Surv(TIMEmeta,EVENTmeta == 1) ~ Cluster,data = dat.mem)
plot(fit.km,col= 1:max(fit.kmeans$cluster),xlab = "time (years)",main = "Time to Distant
Metastases")
test <- Surv(dat.mem$TIMEmeta,dat.mem$EVENTmeta  == 1)
log.rank.stage <- pchisq(survdiff(test~Cluster,data=dat.mem)$chisq,3,lower.tail=F)
txt <- substitute(paste("P-val = ",a),list(a=log.rank.stage))
text(10,0.4,txt)
legend("bottomleft",c("Clust 1","Clust 2"),text.col = 1:max(fit.kmeans$cluster))
```

Comments

After Discussion by the dissertation Committee the following comments are added in this appendix regarding Paper 1.

- A faculty member wished to emphasize the difference between simulation and drawing samples with replacement from the original data, which is termed bootstrap sampling here. The faculty member wanted to note that both can be seen as simulation.
- The sample sizes reported are to be viewed as minimums as they reflect non real world conditions and the particular sampling approach involved.
- The approach did not work well for very small values of the lambda parameter in the Poisson case.
- The proofs in the paper both derive from the central limit theorem for the maximum likelihood estimate in the exponential family. They result from considering the behavior of the likelihood function as sample size increases. These are well known results restated on a likelihood scale here.
- Inference may be undertaken in small sample sizes with parametric bootstrap methods or nonparametric statistics. The results in the paper reflect a diagnostic which assesses the accuracy of the normal approximation.