BAYESIAN LEARNING UNDER NONNORMALITY

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

YILDIZ ELİF YILMAZ

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE
OF
MASTER OF SCIENCE
IN
COMPUTER ENGINEERING

DECEMBER 2004

Approval of the Graduate School of Natural and Applied Sciences

_____
Prof. Dr. Canan Özgen
Director

I certify that this thesis satisfies all the requirements as a thesis for the degree of Master of Science.

_____
Prof. Dr. Ayşe Kiper
Head of Department

This is to certify that we have read this thesis and that in our opinion it is fully adequate, in scope and quality, as a thesis for the degree of Master of Science.

_____	_____
Assoc. Prof. Dr. Ayşen Akkaya	Assoc. Prof. Dr. Ferda Nur Alpaslan
Co-Supervisor	Supervisor

Examining Committee Members

Prof. Dr. Mehmet R. Tolun	(Çankaya Univ., CENG)	_____
Assoc. Prof. Dr. Ferda Nur Alpaslan	(METU, CENG)	_____
Assoc. Prof. Dr. Ayşen Akkaya	(METU, STAT)	_____
Prof. Dr. Faruk Polat	(METU, CENG)	_____
Assist. Prof. Dr. Ayşenur Birtürk	(METU, CENG)	_____

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last name : Yıldız Elif Yılmaz

Signature :

ABSTRACT

BAYESIAN LEARNING UNDER NONNORMALITY

Yılmaz, Yıldız Elif

M.S., Department of Computer Engineering

Supervisor : Assoc. Prof. Dr. Ferda Nur Alpaslan

Co-Supervisor : Assoc. Prof. Dr. Ayşen (Dener) Akkaya

December 2004, 72 pages

Naive Bayes classifier and maximum likelihood hypotheses in Bayesian learning are considered when the errors have non-normal distribution. For location and scale parameters, efficient and robust estimators that are obtained by using the modified maximum likelihood estimation (MML) technique are used. In naive Bayes classifier, the error distributions from class to class and from feature to feature are assumed to be non-identical and Generalized Secant Hyperbolic (GSH) and Generalized Logistic (GL) distribution families have been used instead of normal distribution. It is shown that the non-normal naive Bayes classifier obtained in this way classifies the data more accurately than the one based on the normality assumption. Furthermore, the maximum likelihood (ML) hypotheses are obtained under the assumption of non-normality, which also produce better results compared to the conventional ML approach.

Keywords: Bayesian Learning, Non-normality, Generalized Secant Hyperbolic, Generalized Logistic, Robustness.

ÖZ

NORMAL DAĞILIMA SAHİP OLMAMA VARSAYIMI ALTINDA BAYES ÖĞRENMESİ

Yılmaz, Yıldız Elif

Yüksek Lisans, Bilgisayar Mühendisliği

Tez Yöneticisi : Doç. Dr. Ferda Nur Alpaslan

Ortak Tez Yöneticisi : Doç. Dr. Ayşen (Dener) Akkaya

Aralık 2004, 72 sayfa

Bayes öğrenmesinde naive Bayes sınıflandırıcısı ve en çok olabilirlik önsavları için hata terimlerinin normal olmayan dağılıma sahip olması durumu düşünülmüştür. Uyarlanmış en çok olabilirlik metodu ile yerleştirme ve ölçek parametreleri için etkin ve sağlam tahmin ediciler elde edilmiştir. Naive Bayes sınıflandırıcısında hata terimi dağılımlarının sınıftan sınıfa ve özellikten özelliğe özdeş olmadığı varsayılmıştır ve normal dağılımı yerine Genelleştirilmiş Sekant Hiperbolik (GSH) ve Genelleştirilmiş Lojistik (GL) dağılım aileleri kullanılmıştır. Bu yolla elde edilen normal olmayan naive Bayes sınıflandırıcısı, normallik varsayımına dayanana göre verileri daha doğru sınıflandırdığı gösterilmiştir. Ayrıca geleneksel en çok olabilirlik yaklaşımına göre daha iyi sonuçlar veren normal olmama varsayımı altında en çok olabilirlik önsavları elde edilmiştir.

Anahtar Kelimeler: Bayes öğrenmesi, Normal olmayan dağılımlar, Genelleştirilmiş Sekant Hiperbolik, Genelleştirilmiş Lojistik, Güçlülük.

To My Grandfather

ACKNOWLEDGMENTS

The author wish to express her deepest gratitude to her supervisor Assoc. Prof. Dr. Ferda Nur Alpaslan and co-supervisor Assoc. Prof. Dr. Ayşen (Dener) Akkaya for their invaluable guidances, advices, criticisms and encouragements and insight throughout the research.

The author also would like to thank the research assistant Candemir Çığşar for his help.

The scholarship provided by the Scientific and Technical Research Council of Turkey is gratefully acknowledged.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Bayesian learning is a statistical approach to the problem of pattern classification. Most Bayesian learning procedures are based on the assumption that the underlying distribution is normal. In practice, however, non-normal distributions occur so frequently. To quote Geary (1947): “Normality is a myth; there never was, and never will be, a normal distribution.” Hence to assume normality instead might lead to erroneous statistical inferences (Tiku et al., 1986). It is, therefore, very important to develop statistical procedures which are appropriate and efficient for non-normal distributions.

Naive Bayesian classification is the optimal method of supervised learning if the assumptions are satisfied. It basically assumes the features are independent given the class. Although this assumption is almost always violated in practice, Domingos and Pazzani (1996) showed that the naive Bayes classifier is remarkably robust to the failure of this assumption. However, in the literature, it is also assumed that the model distribution is normal without investigating the plausible distribution of the training set. For instance, Sebe et al. (2002) showed that in naive Bayes classifier the Cauchy distribution assumption provides better results than the normal distribution assumption for an emotion recognition problem. Hence, in this study, naive Bayesian classification technique is modified according to non-normal distribution assumption.

In machine learning determining the most probable hypothesis from some space, given the observed training data is an important task. Bayesian learning method provides a probabilistic approach to find the best hypothesis. In the literature, when the Bayesian learning method is used and the learning problem is a continuous-valued target function,

the target value of each training example is assumed to be corrupted by random noise drawn according to a normal distribution. In this study, the underlying distribution of noise is assumed to be non-normal and statistical procedures which are efficient and robust are developed.

The aim of this thesis is to introduce the Generalized Secant Hyperbolic and Generalized Logistic naive Bayes classifiers and to find the maximum likelihood hypothesis when the noise corrupting the target value is non-normally distributed.

The outline of this thesis is given as follows: Chapter 1 presents the naive Bayes classifier and the maximum likelihood hypotheses under the assumption of normality and the properties of the Generalized Secant Hyperbolic and Generalized Logistic families. In Chapter 2, the Generalized Secant Hyperbolic and Generalized Logistic naive Bayes classifiers are defined. In Chapter 3, maximum likelihood hypothesis is found when the noise is non-normally distributed. Finally, applications and conclusions are presented in Chapter 4.

1.1 Historical Perspective

1.1.1 Model Description and Test Procedures Under Normality

i) Naive Bayes Classifier

The naive Bayes classifier is a simple classifier which assumes the features are independent given the class. Although independence is a poor assumption, in practice the naive Bayes classifier often competes well with the more sophisticated classifiers. In some domains its performance has been shown to be comparable to that of neural network and decision tree learning (Mitchell, 1997). Domingos and Pazzani (1996) verified that the naive Bayes classifier performs quite well in practice even when strong

attribute dependences are present, and showed that this is at least in part due to the fact that the naive Bayes classifier does not depend on feature independence to be optimal.

Consider a classification problem in which w_j (j = 1, 2,..., c) denotes the state of nature with the prior probability $P(w_j)$ and $\underline{x} \in R^d$ denotes the feature vector. Suppose that a collection of samples according to class is separated, so that there are c data sets, D_1, ..., D_c, with the samples in D_j having been drawn independently according to the probability law $p(\underline{x}|w_j)$. In the literature, it is assumed that $p(x_i|w_j) \sim N(\mu_i, \sigma_i^2)$, i = 1, 2,..., d (Duda et al., 1973).

With the assumption that samples in D_j give no information about the parameters corresponding to w_k where $j \neq k$, there are c separate problems of the following form:

Since the naive Bayes classifier is based on the simplifying assumption that the feature values are conditionally independent given the target value, we have

$$p(\underline{x}|w_j) = \prod_{i=1}^{d} p(x_i|w_j) \qquad (1.1.1.1)$$

where $p(x_i|w_j) \sim N(\mu_i, \sigma_i)$, j = 1, 2,..., c.

The problem is to use the information provided by the training samples to obtain optimal estimators for the unknown parameters μ_i, σ_i $(1 \leq i \leq d)$. For the i^{th} feature in the j^{th} class $(1 \leq i \leq d, 1 \leq j \leq c)$, the Fisher likelihood function is

$$L = \prod_{k=1}^{n} p(x_{ik}|w_j). \qquad (1.1.1.2)$$

The likelihood equations for estimating μ_i and σ_i are

$$\frac{\partial \ln L}{\partial \mu_i} = -\frac{n}{\sigma_i^2}(\bar{x}_i - \mu_i) = 0 \qquad (1.1.1.3)$$

and

$$\frac{\partial \ln L}{\partial \sigma_i} = -\frac{n}{\sigma_i} + \frac{1}{\sigma_i^3}\sum_{k=1}^{n}(x_{ik} - \mu_i) = 0 \qquad (1.1.1.4)$$

where $\bar{x}_i = \frac{1}{n}\sum_{k=1}^{n} x_{ik}$.

The simultaneous solutions of (1.1.1.3) and (1.1.1.4) are the maximum likelihood estimators:

$$\tilde{\mu}_i = \bar{x}_i \qquad (1.1.1.5)$$

and

$$\tilde{\sigma}_i^2 = \frac{\sum_{k=1}^{n}(x_{ik} - \bar{x}_i)^2}{n-1} \qquad (1.1.1.6)$$

where $\tilde{\sigma}_i^2$ is an unbiased estimator of σ_i^2.

Hence, the following can be obtained:

$$p(\underline{x}\,|\,w_j, D) = \prod_{i=1}^{d} p(x_i\,|\,w_j, D)$$

$$= \prod_{i=1}^{d} \frac{1}{\sqrt{2\pi}\hat{\sigma}_i} \exp\left(-\frac{1}{2}\left(\frac{x_i - \hat{\mu}_i}{\hat{\sigma}_i}\right)^2\right). \qquad (1.1.1.7)$$

To make classification, the class which maximizes

$$p(w_j|\underline{x}, D) = \frac{p(\underline{x}|w_j, D)P(w_j|D)}{\sum_{j=1}^{c} p(\underline{x}|w_j, D)P(w_j|D)} \quad (j = 1, 2, ..., c) \qquad (1.1.1.8)$$

is selected. Here, maximizing (1.1.1.8) is equivalent to maximizing

$$p(\underline{x}|w_j, D)P(w_j|D) \qquad (1.1.1.9)$$

where $P(w_j|D)$ are obtained from a trivial calculation.

ii) Maximum Likelihood Hypotheses

Consider the problem of learning a continuous-valued target function (Mitchell, 1997):

Learner considers a hypothesis space H consisting of some class of real-valued functions defined over an instance space X (i.e., $\forall h \in H$ is a function of the form $h : X \to \Re$). The problem faced by the learner is to learn an unknown function $f : X \to \Re$ drawn from H. A set of n training examples is provided and the target value of each example is corrupted by random noise drawn according to normal distribution. More precisely, each training example is a pair of the form $\langle x_i, d_i \rangle$ where

$$d_i = f(x_i) + e_i \qquad (i = 1, 2, ..., n), \qquad (1.1.1.10)$$

$f(x_i)$ is the noise-free value of the target function and e_i is a random variable representing the noise. It is assumed that the values of the e_i are drawn independently and that they are distributed according to a normal distribution with zero mean and variance σ^2. The task of the learner is to output a maximum likelihood hypothesis given the observed data D by assuming all hypotheses are equally probable a priori. More precisely, h_{ML} is a maximum likelihood hypothesis provided

$$h_{ML} = \arg\max_{h \in H} \prod_{i=1}^{n} p(d_i | h)$$

$$= \arg\max_{h \in H} \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\left(\frac{d_i - \mu}{\sigma} \right)^2 \right) \qquad (1.1.1.11)$$

where $-\infty < d_i < \infty$ (i = 1, 2,..., n), $\mu \in \Re$, $\sigma > 0$.

By substituting $\mu = f(x_i) = h(x_i)$, we obtain

$$h_{ML} = \arg\max_{h \in H} \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\left(\frac{d_i - h(x_i)}{\sigma} \right)^2 \right)$$

$$= \arg\max_{h \in H} \left\{ -n \ln \sigma - \frac{1}{2} \sum_{i=1}^{n} \left(\frac{d_i - h(x_i)}{\sigma} \right)^2 \right\}. \qquad (1.1.1.12)$$

Now, to estimate σ, the likelihood equation is

$$\frac{\partial \ln L}{\partial \sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^3} \sum_{i=1}^{n} (d_i - h(x_i))^2 = 0. \qquad (1.1.1.13)$$

The solution of the equation (1.1.1.13) is the maximum likelihood estimator:

$$\tilde{\sigma}^2 = \frac{\sum_{i=1}^{n} (d_i - h(x_i))^2}{n}. \qquad (1.1.1.14)$$

Since maximum likelihood estimation has invariance property, by substituting the estimator of the variance (1.1.1.14) into the equation (1.1.1.12), the maximum likelihood hypothesis becomes

$$h_{ML} = \arg\min_{h \in H} \sum_{i=1}^{n} (d_i - h(x_i))^2 . \qquad (1.1.1.15)$$

Hence the maximum likelihood hypothesis is the one that minimizes the sum of the squared errors between the observed training values and the hypothesis predictions.

1.1.2 Robustness

Estimation is the art of inferring information about some unknown quantity on the basis of available data. The estimator is chosen to perform well under the conditions that are assumed to underly the data. Since these conditions are never known exactly, estimators must be chosen which are robust, which perform well under a variety of underlying conditions (Andrews et al., 1972). An estimator is called robust if it is fully efficient (or nearly so) for an assumed distribution but maintains high efficiency for plausible alternatives and a fully efficient estimator is one which is unbiased and its variance is equal to the Cramer-Rao minimum variance bound (Tiku and Akkaya, 2004).

Statistical methods are derived under certain assumptions. However, in practice, many of these assumptions do not hold. For example, the assumption of normality is an unrealistic one. Hence it is very important to obtain estimators which have certain optimal properties with respect to an assumed error distribution.

1.2 Theoretical Backround

1.2.1 Generalized Secant Hyperbolic (GSH) Distribution

The properties of a family of distributions generalizing the secant hyperbolic were developed by Vaughan (2002). This family consists of

- symmetric distributions, with kurtosis[1] ranging from 1.8 to infinity,
- the logistic as a special case,
- the uniform as a limiting case, and
- closely approximates the normal and Student t[2] distributions with corresponding kurtosis.

A significant difference between this family and Student t is that for any member of the Generalized Secant Hyperbolic family, all moments are finite. Thus, technical difficulties associated with evaluating moments of Student t are not present with this family. Moreover, the Student t distribution represents only long-tailed symmetric distributions, i.e. its kurtosis $\beta_2 = \mu_4 / \mu_2^2$ is greater than 3. However, short-tailed symmetric distributions with $\beta_2 < 3$ do occur in practice. To have a unified approach to symmetric non-normal distributions, we need a family of distributions which represents both short-tailed and long-tailed distributions. Such a family is represented by the Generalized Secant Hyperbolic (GSH) distribution.

Let the random variable X has a GSH distribution with the location parameter μ, scale parameter σ and shape parameter t.

$$\text{GSH}(\mu, \sigma; t):\ f(x) = \frac{c_1}{\sigma} \frac{\exp(c_2(x-\mu)/\sigma)}{\exp(2c_2(x-\mu)/\sigma) + 2a\exp(c_2(x-\mu)/\sigma) + 1} \quad (-\infty<x<\infty) \tag{1.2.1.1}$$

where for $-\pi < t \le 0$:

$$a = \cos(t),\ c_2 = \sqrt{(\pi^2 - t^2)/3} \text{ and } c_1 = \frac{\sin(t)}{t} c_2$$

and for $t > 0$:

$$a = \cosh(t),\ c_2 = \sqrt{(\pi^2 + t^2)/3} \text{ and } c_1 = \frac{\sinh(t)}{t} c_2 .$$

[1] Kurtosis is the degree of flatness of a density near its center.
[2] Student t distribution is symmetrical about the population mean, unimodal and extends to infinity in both directions.

For $t > \pi$, $t < \pi$ and $t = \pi$, GSH(μ, σ; t) represents short-tailed, long-tailed and approximately normal distributions, respectively.

The relation between the coefficient of kurtosis β_2 and the shape parameter t is in the following:

$$\beta_2 = \frac{21\pi^2 - 9t^2}{5\pi^2 - 5t^2}, \quad -\pi < t \le 0$$

$$= \frac{21\pi^2 + 9t^2}{5\pi^2 + 5t^2}, \quad t > 0. \qquad (1.2.1.2)$$

The coefficient of kurtosis is given below for a few representative values of the shape parameter, t:

$t =$	$-\pi\sqrt{2/3}$	$-\pi/2$	0	π	$\pi\sqrt{11}$	∞
Kurtosis $\mu_4/\mu_2^2 =$	9.0	5.0	4.2	3.0	2.0	1.8

Now, let Z has a standard GSH distribution, i.e. Z ~ GSH(0, 1, t). The cumulative distribution function is

$$F(z) = 1 + \frac{1}{t}\cot^{-1}(-(\exp(c_2 z) + \cos t)/\sin t), \quad -\pi < t < 0$$

$$= \exp(\pi z/\sqrt{3})/(1 + \exp(\pi z/\sqrt{3})), \quad t = 0$$

$$= 1 - \frac{1}{t}\coth^{-1}((\exp(c_2 z) + \cosh t)/\sinh t), \quad t > 0. \qquad (1.2.1.3)$$

From the cumulative distribution function, the percentage points are calculated as follows:

$$z = \frac{1}{c_2}\ln[\sin(tu)/\sin(t(1-u))], \quad -\pi < t < 0$$

$$= \frac{\sqrt{3}}{\pi} \ln(u/(1-u)), \qquad t = 0$$

$$= \frac{1}{c_2} \ln\left[\sinh(tu)/\sinh(t(1-u))\right], \qquad t > 0 \qquad (1.2.1.4)$$

where u = F(z).

1.2.2 Generalized Logistic (GL) Distribution

Let the random variable X has a Generalized Logistic (GL) distribution with the scale parameter σ and shape parameter b.

$$\text{GL}(\sigma, b): \; f(x) = \frac{b}{\sigma} \frac{\exp(-x/\sigma)}{\left[\exp(-x/\sigma)+1\right]^{b+1}} \qquad (-\infty < x < \infty) \qquad (1.2.2.1)$$

where b > 0.

For b < 1, b = 1 and b > 1, GL(b, σ) represents negatively skewed, symmetric and positively skewed distributions, respectively. In fact, when b = 1, it represents the logistic distribution as GSH(μ, σ, 0).

Now, let Z has a standard GL distribution, i.e. Z ~ GL(1, b). The moment generating function of Z is

$$M_t(z) = E(e^{tz}) = b \frac{\Gamma(1-t)\Gamma(b+t)}{\Gamma(b+1)} \qquad (1.2.2.2)$$

where $|t| < 1$. Hence the r^{th} moment of Z is

$$\mu'_r = \left\{ \frac{d^r}{dt^r} M_t(z) \right\}_{t=0}. \qquad (1.2.2.3)$$

In particular,

$$E(Z) = \psi(b) - \psi(1) \text{ and } V(Z) = \psi'(b) + \psi'(1) \qquad (1.2.2.4)$$

where $\psi(x) = \Gamma'(x)/\Gamma(x)$. The properties of $\psi(x)$ are given in Abramowitz and Stegun (1985).

For a few representative values of the shape parameter, b, the coefficients of kurtosis and skewness are given (Tiku and Akkaya, 2004):

b =	0.5	1.0	2.0	4.0	6.0
Skewness $\mu_3/\mu_2^{3/2}$ =	-0.855	0.0	0.577	0.868	0.961
Kurtosis μ_4/μ_2^2 =	5.400	4.2	4.333	4.758	4.951

The cumulative distribution function is

$$F(z) = (\exp(-z) + 1)^{-b}. \qquad (1.2.2.5)$$

From the cumulative distribution function, the percentage points are calculated as follows:

$$z = -\ln(u^{-1/b} - 1) \qquad (1.2.2.6)$$

where u = F(z).

CHAPTER 2

NAIVE BAYES CLASSIFIER UNDER NONNORMALITY

In this chapter, the Generalized Secant Hyperbolic (GSH) and Generalized Logistic (GL) naive Bayes classifiers using the GSH and GL distributions as the model distributions are introduced.

2.1 Generalized Secant Hyperbolic Naive Bayes Classifier

Suppose there are c data sets D_1, D_2,…, D_c with the samples in D_j having been drawn independently and identically according to probability law $p(\underline{x}|w_j)$ (j = 1, 2,…, c) and assume that the features are independently distributed according to Generalized Secant Hyperbolic (GSH) distribution:

$$p(x_i|w_j) = \frac{c_{1i}}{\sigma_i} \frac{\exp(c_{2i}(x_i - \mu_i)/\sigma_i)}{\exp(2c_{2i}(x_i - \mu_i)/\sigma_i) + 2a_{ij}\exp(c_{2i}(x_i - \mu_i)/\sigma_i) + 1} \qquad (2.1.1)$$

where $-\infty < x_i < \infty$, $-\infty < \mu_i < \infty$, $\sigma_i > 0$ (i = 1, 2, ..., d; j = 1, 2,..., c),

for $-\pi < t_i \le 0$:

$$a_i = \cos(t_i),\ c_{2i} = \sqrt{(\pi^2 - t_i^2)/3} \text{ and } c_{1i} = \frac{\sin(t_i)}{t_i} c_{2i}$$

and for $t_i > 0$:

$$a_i = \cosh(t_i),\ c_{2i} = \sqrt{(\pi^2 + t_i^2)/3} \text{ and } c_{1i} = \frac{\sinh(t_i)}{t_i} c_{2i}.$$

Since the naive Bayes classifier is based on the simplifying assumption that the feature values are conditionally independent given the target value, we have

$$p(\underline{x}\,|\,w_j) = \prod_{i=1}^{d} p(x_i\,|\,w_j) \tag{2.1.2}$$

where $p(x_i\,|\,w_j) \sim GSH(\mu_i, \sigma_i ; t_i)$, $j = 1, 2,..., c$.

2.1.1 Maximum Likelihood Estimation

In order to estimate μ_i and σ_i, first assume the shape parameter t_i is known. For the i^{th} feature in the j^{th} class $(1 \le i \le d, 1 \le j \le c)$, the Fisher likelihood function is

$$L = \prod_{k=1}^{n_i} p(x_{ik}\,|\,w_j)\,. \tag{2.1.1.1}$$

Note that the formulation is based on the fact that there may be missing attribute values. If there is no missing attribute values, take $n_1 = n_2 = ... = n_d = n$.

The likelihood equations for estimating μ_i and σ_i are

$$\frac{\partial \ln L}{\partial \mu_i} = -n_i \frac{c_{2i}}{\sigma_i} + 2\frac{c_{2i}}{\sigma_i}\sum_{k=1}^{n_i} g(z_{ik}) = 0 \tag{2.1.1.2}$$

and

$$\frac{\partial \ln L}{\partial \sigma_i} = -n_i \frac{1}{\sigma_i} - \frac{c_{2i}}{\sigma_i}\sum_{k=1}^{n_i} z_{ik} + 2\frac{c_{2i}}{\sigma_i}\sum_{k=1}^{n_i} z_{ik}\, g(z_{ik}) = 0 \tag{2.1.1.3}$$

where $z_{ik} = \dfrac{x_{ik} - \mu_i}{\sigma_i}$ and $g(z_i) = \dfrac{\exp(2c_{2i}z_i) + a_i \exp(c_{2i}z_i)}{\exp(2c_{2i}z_i) + 2a_i \exp(c_{2i}z_i) + 1}$.

The likelihood equations (2.1.1.2) and (2.1.1.3) do not admit explicit solutions since the terms involve the nonlinear function $g(z_i)$. An iterative process can be used to solve these equations, but without extensive simulations, the properties of the resulting maximum likelihood estimates are difficult to determine, especially for small samples. An alternative estimation procedure called the modified maximum likelihood solves the problems mentioned above. Therefore, Vaughan (2002) used modified maximum likelihood estimation technique in his analysis.

2.1.2 Modified Maximum Likelihood Estimation

Tiku and Suresh (1992) introduced modified maximum likelihood estimation in general location-scale models, with the following properties:

1. The estimates are explicit functions of sample observations and are easier to compute than the maximum likelihood estimates.
2. It is aymptotically equivalent to maximum likelihood when regularity conditions hold (Tiku et al.(1986), Vaughan and Tiku (2000), Bhattacharyya (1985)).
3. The estimates are almost fully efficient in terms of the Minimum Variance Bounds (MVBs) even for small samples.
4. The estimates have little bias or no bias.
5. The method is essentially self-censoring, since it assigns small weights to extremes.

Tiku's modified maximum likelihood methodology proceeds in three steps as follows:

1. Express the likelihood equations in terms of ordered variates $z_{i(k)} = \frac{x_{i(k)} - \mu_i}{\sigma_i}$ $(1 \le i \le d, 1 \le k \le n_i)$,
2. linearize the intractable terms in the likelihood equations by using the first two terms of the Taylor series expansion and
3. solve the resulting equations to get the modified maximum likelihood estimators.

Since complete sums are invariant to ordering, the likelihood equations can be written as follows:

$$\frac{\partial \ln L}{\partial \mu_i} = -n_i \frac{c_{2i}}{\sigma_i} + 2\frac{c_{2i}}{\sigma_i}\sum_{k=1}^{n_i} g(z_{i(k)}) = 0 \qquad (2.1.2.1)$$

and

$$\frac{\partial \ln L}{\partial \sigma_i} = -n_i \frac{1}{\sigma_i} - \frac{c_{2i}}{\sigma_i}\sum_{k=1}^{n_i} z_{i(k)} + 2\frac{c_{2i}}{\sigma_i}\sum_{k=1}^{n_i} z_{i(k)} g(z_{i(k)}) = 0 \qquad (2.1.2.2)$$

where $z_{i(k)}\ (1 \le i \le d)$ is the k^{th} order statistic of the values $z_{i1}, ..., z_{in}$.

Let $t_{i(k)} = E(z_{i(k)})$, $t_{i(k)}\ (1 \le i \le d, 1 \le k \le n_i)$ are the expected values of the standardized ordered variates. For large sample size, the values of $t_{i(k)}$ can be found as follows:

$$t_{i(k)} = \frac{1}{c_{2i}} \ln\left[\frac{\sin\left(t_i \frac{k}{n+1}\right)}{\sin\left(t_i\left(1-k/(n+1)\right)\right)}\right], \quad -\pi < t_i < 0$$

$$= \frac{\sqrt{3}}{\pi} \ln\left(\frac{k/(n+1)}{1-k/(n+1)}\right) \qquad , \quad t_i = 0$$

$$= \frac{1}{c_{2i}} \ln\left[\frac{\sinh\left(t_i \frac{k}{n+1}\right)}{\sinh\left(t_i\left(1-k/(n+1)\right)\right)}\right], \quad t_i > 0. \qquad (2.1.2.3)$$

Since $z_{i(k)}$ is located in the vicinity of $t_{i(k)}$, the nonlinear function $g(z_i)$ can be approximated by the Taylor series expansion as follows:

$$g(z_{i(k)}) \cong g(t_{i(k)}) + (z_{i(k)} - t_{i(k)}) g'(t_{i(k)})$$

$$= \alpha_{ik} + \beta_{ik} z_{i(k)} \qquad (2.1.2.4)$$

where

$$\alpha_{ik} = \frac{\exp(2c_{2i} t_{i(k)}) + a_i \exp(c_{2i} t_{i(k)})}{\exp(2c_{2i} t_{i(k)}) + 2a_i \exp(c_{2i} t_{i(k)}) + 1} - \beta_{ik} t_{i(k)}$$

and

$$\beta_{ik} = \frac{a_i c_{2i} \exp(3c_{2i} t_{i(k)}) + 2c_{2i} \exp(2c_{2i} t_{i(k)}) + a_i c_{2i} \exp(c_{2i} t_{i(k)})}{\left[\exp(2c_{2i} t_{i(k)}) + 2a_i \exp(c_{2i} t_{i(k)}) + 1\right]^2}.$$

Specifically, if there is no missing attribute values,

$$t_{1(k)} = t_{2(k)} = ... = t_{d(k)} = t_{(k)},$$

$$\alpha_{1k} = \alpha_{2k} = ... = \alpha_{dk} = \alpha_k$$

and

$$\beta_{1k} = \beta_{2k} = ... = \beta_{dk} = \beta_k. \qquad (2.1.2.5)$$

Incorporating (2.1.2.4) into (2.1.2.1) and (2.1.2.2), the following modified likelihood equations are obtained:

$$\frac{\partial \ln L}{\partial \mu_i} \cong \frac{\partial \ln L^*}{\partial \mu_i} = -n_i \frac{c_{2i}}{\sigma_i} + 2\frac{c_{2i}}{\sigma_i} \sum_{k=1}^{n_i} \left[\alpha_{ik} + \beta_{ik} z_{i(k)}\right] = 0 \qquad (2.1.2.6)$$

and

$$\frac{\partial \ln L}{\partial \sigma_i} \cong \frac{\partial \ln L^*}{\partial \sigma_i} = -n_i \frac{1}{\sigma_i} - \frac{c_{2i}}{\sigma_i} \sum_{k=1}^{n_i} z_{i(k)} + 2\frac{c_{2i}}{\sigma_i} \sum_{k=1}^{n_i} z_{i(k)} \left[\alpha_{ik} + \beta_{ik} z_{i(k)}\right] = 0. \qquad (2.1.2.7)$$

The simultaneous solutions of the equations (2.1.2.6) and (2.1.2.7) are the MML estimators:

$$\hat{\mu}_i = \frac{\sum_{k=1}^{n_i} \beta_{ik} x_{i(k)}}{m_i} \qquad (2.1.2.8)$$

and

$$\hat{\sigma}_i = \frac{-B_i + \sqrt{B_i^{\,2} + 4n_i C_i}}{2\sqrt{n_i(n_i - 1)}} \qquad (2.1.2.9)$$

where $m_i = \sum_{k=1}^{n_i} \beta_{ik}$, $B_i = n_i c_{2i}(\bar{x}_i - \bar{x}_{ia})$, $\bar{x}_i = \frac{1}{n_i}\sum_{k=1}^{n_i} x_{ik}$, $\bar{x}_{ia} = \frac{2}{n_i}\sum_{k=1}^{n_i} \alpha_{ik} x_{i(k)}$ and

$$C_i = 2c_{2i}\sum_{k=1}^{n_i} \beta_{ik}(x_{i(k)} - \hat{\mu}_i)^2 = 2c_{2i}\left[\sum_{k=1}^{n_i} \beta_{ik} x_{i(k)}^2 - m_i \hat{\mu}_i^2\right].$$

The divisor n_i in the original expression for $\hat{\sigma}_i$ is replaced by $\sqrt{n_i(n_i - 1)}$ to reduce the bias.

It may be noted that (2.1.2.4) are asymptotically strict equalities. Moreover, in the limit when n_i tends to infinity

$$\lim_{n_i \to \infty} \frac{1}{n_i}\frac{\partial \ln L}{\partial \mu_i} \equiv \lim_{n_i \to \infty} \frac{1}{n_i}\frac{\partial \ln L^*}{\partial \mu_i} \equiv 0$$

and

$$\lim_{n_i \to \infty} \frac{1}{n_i}\frac{\partial \ln L}{\partial \sigma_i} \equiv \lim_{n_i \to \infty} \frac{1}{n_i}\frac{\partial \ln L^*}{\partial \sigma_i} \equiv 0. \qquad (2.1.2.10)$$

Consequently, the MML estimators $\hat{\mu}_i$ and $\hat{\sigma}_i$ above are asymptotically equivalent to ML estimators and, thus, $\hat{\mu}_i$ and $\hat{\sigma}_i$ are asymptotically unbiased and efficient estimators, at least heuristically. Note, however, that $\hat{\mu}_i$ is unbiased for all n_i. That follows from symmetry.

Lemma 2.1.1: Asymptotically, the estimator $\hat{\mu}_i$ $(1 \le i \le d)$ is the MVB estimator of μ_i and is normally distributed with variance

$$V(\hat{\mu}_i) \cong \frac{\sigma_i^2}{2m_i c_{2i}}. \qquad (2.1.2.11)$$

Proof: Since $\partial \ln L^* / \partial \mu_i$ is asymptotically equivalent to $\partial \ln L / \partial \mu_i$ and assumes the form

$$\frac{\partial \ln L^*}{\partial \mu_i} = \frac{2m_i c_{2i}}{\sigma_i^2}(\hat{\mu}_i - \mu_i), \qquad (2.1.2.12)$$

(2.1.2.11) is obtained. By dividing both sides of (2.1.2.12) by n_i, we can apply central limit theorem and since $E\left[\frac{\partial^r \ln L^*}{\partial \mu_i^r}\right] = 0$ for all $r \geq 3$, $\hat{\mu}_i$ is asymptotically normally distributed.

Lemma 2.1.2: Asymptotically, $\frac{n_i \hat{\sigma}_i^2(\mu_i)}{\sigma_i^2}$ $(1 \leq i \leq d)$ is conditionally distributed as chi-square with n_i degrees of freedom.

Proof: For large n_i, $\frac{B_i}{\sqrt{n_i C_{1i}}} \cong 0$ where $C_{1i} = 2c_{2i}\sum_{k=1}^{n_i}\beta_{ik}(x_{i(k)} - \mu_i)^2$. Therefore, it can be shown that

$$\frac{\partial \ln L^*}{\partial \sigma_i} \cong \frac{n_i}{\sigma_i^3}\left(\frac{C_{1i}}{n_i} - \sigma_i^2\right). \qquad (2.1.2.13)$$

Asymptotically, $\frac{C_{1i}}{n_i}$ is the MVB estimator of σ_i^2. Evaluation of the cumulants of $\frac{\partial \ln L^*}{\partial \sigma_i}$ in terms of the expected values of the derivatives of $\frac{\partial \ln L^*}{\partial \sigma_i}$ immediately leads

to the result that $\frac{n_i \hat{\sigma}_i^2(\mu_i)}{\sigma_i^2}$ is distributed as chi-square with n_i degrees of freedom (Bartlett, 1953).

Corollary 2.1.1: Asymptotically, $\frac{n_i \hat{\sigma}_i^2}{\sigma_i^2}$ $(1 \le i \le d)$ is distributed as chi-square with n_i-1 degrees of freedom.

At the beginning, it was assummed that the shape parameters t_i $(1 \le i \le d)$ are known. Now, plausible values for these parameters can be found by the following method (Tiku and Akkaya, 2004):

To locate the most plausible value of t_i, the MML estimators of μ_i and σ_i are calculated from (2.1.2.8) and (2.1.2.9) for different values of t_i. Then, the values of

$$\ln \hat{L} = n_i \ln c_{1i} - n_i \ln \hat{\sigma}_i + c_{2i} \sum_{k=1}^{n_i} \hat{z}_{ik} - \sum_{k=1}^{n_i} \ln\left[\exp(2c_{2i}\hat{z}_{ik}) + 2a_i \exp(c_{2i}\hat{z}_{ik}) + 1\right] \quad (2.1.2.14)$$

where $\hat{z}_{ik} = \frac{x_{ik} - \hat{\mu}_i}{\hat{\sigma}_i}$ $(1 \le i \le d,\ 1 \le k \le n_i)$,

are calculated. The value of t_i that maximizes $\ln \hat{L}$ is the most appropriate choice. As a result, the estimates of μ_i, σ_i and t_i $(1 \le i \le d)$ are obtained for all classes. Since modified maximum likelihood estimation has invariance property, the following is obtained:

$$p(\underline{x} \mid w_j, D) = \prod_{i=1}^{d} p(x_i \mid w_j, D)$$

$$= \prod_{i=1}^{d} \frac{c_{1i}}{\hat{\sigma}_i} \frac{\exp(c_{2i}(x_i - \hat{\mu}_i)/\hat{\sigma}_i)}{\exp(2c_{2i}(x_i - \hat{\mu}_i)/\hat{\sigma}_i) + 2a_i \exp(c_{2i}(x_i - \hat{\mu}_i)/\hat{\sigma}_i) + 1} \quad (2.1.2.15)$$

and to make classification, the class which maximizes (1.1.1.9) is selected.

2.1.3 Efficiency Properties

The estimator $\hat{\mu}$ is unbiased, in fact, it is asymptotically MVB estimator of μ, and is normally distributed. Therefore, $\hat{\mu}$ is best asymptotically normal (BAN) estimator. The MVB for estimating μ is as follows:

for $-\pi < t \le 0$

$$\mathrm{MVB}(\mu) = \frac{2\sigma^2 t \sin^2 t}{n c_2^2 (t - \sin t \cos t)}, \tag{2.1.3.1}$$

for $t \ge 0$

$$\mathrm{MVB}(\mu) = \frac{2\sigma^2 t \sinh^2 t}{n c_2^2 (\sinh t \cosh t - t)}. \tag{2.1.3.2}$$

The estimator $\hat{\sigma}^2$ is asymptotically the MVB estimator of σ^2 and is distributed as a multiple of chi-square; see Lemma 2.1.2. The MVB for estimating σ^2 is as follows:

for $-\pi < t \le 0$

$$\mathrm{MVB}(\sigma) = \left(\frac{6\sigma^2}{n}\right) \Bigg/ \left(\frac{\pi^2 - t^2}{\sin^2 t} - \frac{(\pi^2 - 3t^2)\cos t}{t \sin t}\right) \tag{2.1.3.3}$$

for $t \ge 0$

$$\mathrm{MVB}(\sigma) = \left(\frac{6\sigma^2}{n}\right) \Bigg/ \left(\frac{(\pi^2 + 3t^2)\cosh t}{t \sinh t} - \frac{\pi^2 + t^2}{\sinh^2 t}\right) \tag{2.1.3.4}$$

Given in Table 2.1.1 are the simulated values (based on N=100,000/n Monte Carlo runs) of the variances of the MML and LS estimators of μ_i, relative efficiency (RE) of the LS estimator $\tilde{\mu} = \bar{x}$, the MVB of μ and the efficiency (E) of $\hat{\mu}$.

Table 2.1.1 Variances of the MML and LS estimators of μ

(1) $V(\tilde{\mu})/\sigma^2$ (2) $V(\hat{\mu})/\sigma^2$ (3) $RE(\tilde{\mu}) = [V(\hat{\mu})/V(\tilde{\mu})]*100$

(4) $MVB(\mu)/\sigma^2$ (5) $E(\hat{\mu}) = [MVB(\mu)/V(\hat{\mu})]*100$

		$\beta_2 =$	2.0	3.0	4.2	5.0	9.0
n = 10	(1)		0.099	0.100	0.100	0.101	0.980
	(2)		0.070	0.101	0.094	0.086	0.055
	(3)		70.60	100.44	94.46	85.66	56.59
	(4)		0.053	0.097	0.091	0.081	0.046
	(5)		75.54	96.68	96.77	93.69	82.63
n = 15	(1)		0.066	0.067	0.066	0.065	0.064
	(2)		0.043	0.068	0.061	0.055	0.033
	(3)		64.65	100.44	92.54	83.67	52.23
	(4)		0.035	0.065	0.061	0.054	0.031
	(5)		82.31	96.09	98.97	98.70	91.34
n = 20	(1)		0.049	0.050	0.050	0.050	0.049
	(2)		0.030	0.049	0.046	0.041	0.025
	(3)		61.14	99.13	92.31	82.43	50.96
	(4)		0.026	0.049	0.046	0.041	0.023
	(5)		87.29	98.97	98.85	98.93	91.76
n = 50	(1)		0.020	0.021	0.019	0.021	0.020
	(2)		0.011	0.020	0.017	0.017	0.009
	(3)		55.99	97.66	92.14	81.38	47.14
	(4)		0.011	0.019	0.018	0.016	0.009
	(5)		93.73	95.79	104.94	96.77	96.84

It can be seen that $\hat{\mu}$ is considerably more efficient than $\tilde{\mu}$ even for small sample sizes other than approximately normal distribution ($\beta_2 = 3.0$). Actually, for approximately normal distribution $\hat{\mu}$ is as efficient as $\tilde{\mu}$. A disconcerting feature of $\tilde{\mu}$ is that its relative efficiency decreases as sample size n increases. Realize that both $\hat{\mu}$ and $\tilde{\mu}$ are unbiased estimators of μ.

The MML estimator $\hat{\sigma}$ can sometimes have larger bias than $\tilde{\sigma}$ for small sample sizes. Thus, deficiency of MML and LS estimators are calculated through simulations.

Given in Table 2.1.2 are the simulated values of the deficiencies (Def) of the least square estimators $\tilde{\mu} = \bar{x}$ and $\tilde{\sigma}^2 = \sum_{k=1}^{n} (x_k - \bar{x})^2 / (n-1)$ and the MML estimators $\hat{\mu}$ and $\hat{\sigma}^2$. Note that since $\hat{\mu}$ and $\hat{\sigma}^2$ are uncorrelated with one another and so are the LS estimators $\tilde{\mu}$ and $\tilde{\sigma}^2$ (this follows from symmetry), the joint deficiencies can be calculated as follows:

$$\mathrm{Def}(\tilde{\mu}, \tilde{\sigma}) = \mathrm{MSE}(\tilde{\mu}) + \mathrm{MSE}(\tilde{\sigma})$$

and

$$\mathrm{Def}(\hat{\mu}, \hat{\sigma}) = \mathrm{MSE}(\hat{\mu}) + \mathrm{MSE}(\hat{\sigma}). \qquad (2.1.3.5)$$

Table 2.1.2 Deficiencies of the MML and LS estimators of μ and σ
(1) Def($\tilde{\mu}, \tilde{\sigma}$) (2) Def($\hat{\mu}, \hat{\sigma}$)

	β_2 =	2.0	2.5	3.0	4.2	5.0	7.0	9.0
n = 10	(1)	0.132	0.142	0.154	0.175	0.189	0.223	0.256
	(2)	0.105	0.137	0.158	0.176	0.180	0.213	0.262
n = 15	(1)	0.085	0.095	0.103	0.120	0.125	0.157	0.181
	(2)	0.061	0.091	0.105	0.115	0.115	0.134	0.160
n = 20	(1)	0.065	0.070	0.077	0.087	0.099	0.117	0.134
	(2)	0.044	0.066	0.078	0.083	0.087	0.094	0.109
n = 50	(1)	0.025	0.026	0.031	0.036	0.039	0.049	0.060
	(2)	0.015	0.024	0.031	0.033	0.032	0.035	0.038

Deficiency of MML estimators are considerably smaller than the defficiency of LS estimators even for sample size n = 10 other than approximately normal ($\beta_2 = 3.0$), near normal (logistic, $\beta_2 = 4.2$) and very long-tailed ($\beta_2 = 9.0$) distributions. However, for n $\geq$ 11 defficiency of MML estimators becomes smaller than that of LS estimators for near normal and long-tailed distributions.

2.1.4 Robustness of Estimators

It is very important to obtain estimators which have certain optimal properties with respect to an assumed distribution. In spite of our best efforts to identify the underlying distribution through graphical techniques (Q-Q plots, for example) or goodness-of-fit tests, in practice, the shape parameters might be misspecified or the data might contain outliers (inliers) or be contaminated. Thus deviations from an assumed distribution occur. That brings the issue of robustness in focus. An estimator is called robust if it is fully efficient (or nearly so) for an assumed distribution but maintains high efficiency for plausible alternatives (Tiku et al., 1986).

To show the robustness of both MML estimators, we consider, for illustration, the following plausible alternatives (1)-(4) to the assumed distribution GSH in (2.1.1) with $t = -\pi / 2$:

(1) Misspecification of the distribution: GSH(μ, σ, $-3\pi / 4$)

(2) Dixon's outlier model: (n-1) observations come from GSH(μ, σ, $-\pi / 2$) but one observation (we do not know which one) comes from GSH(μ, 4σ, $-\pi / 2$)

(3) Mixture model: 0.90 GSH(μ, σ, $-\pi / 2$) + 0.10 GSH(μ, 4σ, $-\pi / 2$)

(4) Contamination model: 0.90 GSH(μ, σ, $-\pi / 2$) + 0.10 Uniform(-1/2, 1/2)

The simulated variances of $\tilde{\mu}$ and $\hat{\mu}$, the simulated means of $\tilde{\sigma}$ and $\hat{\sigma}$ are given in Table 2.1.3. Also given are the values of the relative efficiency of the LS estimators of μ and σ. It is obvious that the MML estimators $\hat{\mu}$ and $\hat{\sigma}$ are remarkably efficient and robust.

Table 2.1.3 Means, variances and relative efficiencies; n = 10, σ = 1

	Variance		Mean		RE	
Model	$\tilde{\mu}$	$\hat{\mu}$	$\tilde{\sigma}$	$\hat{\sigma}$	$\tilde{\mu}$	$\tilde{\sigma}$
(1)	0.057	0.041	0.711	0.754	70.96	82.94
(2)	0.257	0.116	1.417	1.417	45.23	71.67
(3)	0.254	0.126	1.389	1.415	49.45	81.38
(4)	0.093	0.073	0.914	0.988	78.16	97.31

2.2 Generalized Logistic Naive Bayes Classifier

Assume that the features are independently distributed according to Generalized Logistic (GL) distribution:

$$p(x_i \mid w_j) = \frac{b_i}{\sigma_i} \frac{\exp(-(x_i - \mu_i)/\sigma_i)}{\left[1 + \exp(-(x_i - \mu_i)/\sigma_i)\right]^{b_i+1}} \tag{2.2.1}$$

where $-\infty < x_i < \infty$, $-\infty < \mu_i < \infty$, $\sigma_i > 0$,

2.2.1 Maximum Likelihood Estimation

In order to estimate μ_i and σ_i, first assume the shape parameter b_i is known. For i^{th} feature in the j^{th} class, the Fisher likelihood function is

$$L = \prod_{k=1}^{n_i} p(x_{ik} \mid w_j) . \tag{2.2.1.1}$$

The likelihood equations for estimating μ_i and σ_i are

$$\frac{\partial \ln L}{\partial \mu_i} = n_i \frac{1}{\sigma_i} - \frac{b_i + 1}{\sigma_i} \sum_{k=1}^{n_i} g(z_{ik}) = 0 \tag{2.2.1.2}$$

and

$$\frac{\partial \ln L}{\partial \sigma_i} = -n_i \frac{1}{\sigma_i} + \frac{1}{\sigma_i} \sum_{k=1}^{n_i} z_{ik} - \frac{b_i + 1}{\sigma_i} \sum_{k=1}^{n_i} z_{ik} g(z_{ik}) = 0 \qquad (2.2.1.3)$$

where $z_{ik} = \frac{x_{ik} - \mu_i}{\sigma_i}$ and $g(z_i) = \frac{1}{1 + \exp(z_i)}$.

Şenoğlu (2000) used modified maximum likelihood estimation technique in his analysis.

2.2.2 Modified Maximum Likelihood Estimation

Since complete sums are invariant to ordering, the likelihood equations can be written as follows:

$$\frac{\partial \ln L}{\partial \mu_i} = n_i \frac{1}{\sigma_i} - \frac{b_i + 1}{\sigma_i} \sum_{k=1}^{n_i} g(z_{i(k)}) = 0 \qquad (2.2.2.1)$$

and

$$\frac{\partial \ln L}{\partial \sigma_i} = -n_i \frac{1}{\sigma_i} + \frac{1}{\sigma_i} \sum_{k=1}^{n_i} z_{i(k)} - \frac{b_i + 1}{\sigma_i} \sum_{k=1}^{n_i} z_{i(k)} g(z_{i(k)}) = 0 . \qquad (2.2.2.2)$$

Since $z_{i(k)}$ is located in the vicinity of $t_{i(k)}$, it is approximated by the Taylor series expansion

$$\begin{aligned} g(z_{i(k)}) &\cong g(t_{i(k)}) + (z_{i(k)} - t_{i(k)}) g'(t_{i(k)}) \\ &= \alpha_{ik} - \beta_{ik} z_{i(k)} \end{aligned} \qquad (2.2.2.3)$$

where

$$\alpha_{ik} = \frac{t_{i(k)} \exp(t_{i(k)}) + \exp(t_{i(k)}) + 1}{(\exp(t_{i(k)}) + 1)^2},$$

$$\beta_{ik} = \frac{\exp(t_{i(k)})}{(\exp(t_{i(k)}) + 1)^2}$$

and for large sample size,

$$t_{i(k)} = -\ln\left[\left(\frac{k}{n_i+1}\right)^{-1/b_i} - 1\right].$$

Incorporating (2.2.2.3) into (2.2.2.1) and (2.2.2.2) gives the modified likelihood equations as follows:

$$\frac{\partial \ln L}{\partial \mu_i} \cong \frac{\partial \ln L^*}{\partial \mu_i} = n_i \frac{1}{\sigma_i} - \frac{b_i+1}{\sigma_i}\sum_{k=1}^{n_i}\left[\alpha_{i(k)} - \beta_{i(k)} z_{i(k)}\right] = 0 \qquad (2.2.2.4)$$

and

$$\frac{\partial \ln L}{\partial \sigma_i} \cong \frac{\partial \ln L^*}{\partial \sigma_i} = -n_i \frac{1}{\sigma_i} + \frac{1}{\sigma_i}\sum_{k=1}^{n_i} z_{i(k)} - \frac{b_i+1}{\sigma_i}\sum_{k=1}^{n_i} z_{i(k)}\left[\alpha_{i(k)} - \beta_{i(k)} z_{i(k)}\right] = 0. \quad (2.2.2.5)$$

The simultaneous solutions of the equations (2.2.2.4) and (2.2.2.5) are the MML estimators:

$$\hat{\mu}_i = \hat{\mu}_{i0} + \frac{\Delta_i}{m_i}\hat{\sigma}_i \qquad (2.2.2.6)$$

and

$$\hat{\sigma}_i = \frac{B_i + \sqrt{B_i^2 + 4n_i C_i}}{2\sqrt{n_i(n_i-1)}} \qquad (2.2.2.7)$$

where

$$\hat{\mu}_{i0} = \frac{1}{m_i}\sum_{k=1}^{n_i}\beta_{ik} x_{i(k)}, \quad m_i = \sum_{k=1}^{n_i}\beta_{ik}, \quad \Delta_{ik} = (b_i+1)^{-1} - \alpha_{ik}, \quad \Delta_i = \sum_{k=1}^{n_i}\Delta_{ik},$$

$$B_i = (b_i+1)\sum_{k=1}^{n_i}\Delta_{ik}(x_{i(k)} - \hat{\mu}_{i0}) \quad \text{and} \quad C_i = (b_i+1)\sum_{k=1}^{n_i}\beta_{ik}(x_{i(k)} - \hat{\mu}_{i0})^2.$$

Lemma 2.2.1: Asymptotically, the estimator $\hat{\mu}_i(\sigma_i) = \hat{\mu}_{i0} + \frac{\Delta_i}{m_i}\sigma_i$ is conditionally (σ_i known) the MVB estimator of μ_i $(1 \le i \le d)$ and is normally distributed with variance

$$V(\hat{\mu}_i(\sigma_i)) \cong \frac{\sigma_i^2}{m_i(b_i+1)}. \qquad (2.2.2.8)$$

Lemma 2.2.2: Asymptotically, $\frac{n_i\hat{\sigma}_i^2(\mu_i)}{\sigma_i^2}$ $(1 \le i \le d)$ is conditionally distributed as chi-square with n_i degrees of freedom.

Corollary 2.2.1: Asymptotically, $\frac{n_i\hat{\sigma}_i^2}{\sigma_i^2}$ $(1 \le i \le d)$ is distributed as chi-square with n_i-1 degrees of freedom.

To locate the most plausible value of the shape parameters b_i $(1 \le i \le d)$, the MML estimators of μ_i and σ_i are calculated from (2.2.2.6) and (2.2.2.7) for different values of b_i. Then, the values of

$$\ln \hat{L} = n_i \ln b_i - n_i \ln \hat{\sigma}_i - \sum_{k=1}^{n_i} \hat{z}_{ik} - (b_i+1)\sum_{k=1}^{n_i} \ln[\exp(-\hat{z}_{ik})+1] \qquad (2.2.2.9)$$

where $\hat{z}_{ik} = \frac{x_{ik} - \hat{\mu}_i}{\hat{\sigma}_i}$ $(1 \le i \le d,\ 1 \le k \le n_i)$

are calculated. The value that maximizes $\ln \hat{L}$ is the most appropriate choice (Tiku and Akkaya, 2004). As a result, the estimates of μ_i, σ_i and b_i $(1 \le i \le d)$ are obtained for all classes. Since modified maximum likelihood estimation has invariance property, the following is obtained:

$$p(\underline{x}|w_j, D) = \prod_{i=1}^{d} p(x_i|w_j, D)$$

$$= \prod_{i=1}^{d} \frac{b_i}{\hat{\sigma}_i} \frac{\exp(-(x_i - \hat{\mu}_i)/\hat{\sigma}_i)}{[\exp(-(x_i - \hat{\mu}_i)/\hat{\sigma}_i)+1]^{b_i+1}} \qquad (2.2.2.10)$$

and to make classification, the class which maximizes (1.1.1.9) is selected.

2.2.3 Efficiency Properties

The estimator $\hat{\mu}$ is asymptotically MVB estimator of μ, and is normally distributed. Therefore, $\hat{\mu}$ is BAN estimator. The estimator $\hat{\sigma}^2$ is asymptotically the MVB estimator of σ^2 and is distributed as a multiple of chi-square; see Lemma 2.2.2. Given in Table 2.2.1 are the simulated values (based on N=100,000/n Monte Carlo runs) of the relative efficiencies (RE) of the bias corrected LS estimators $\tilde{\mu} = \bar{x} - (\psi(b) - \psi(1))\tilde{\sigma}$ and $\tilde{\sigma}^2 = \sum_{k=1}^{n} (x_k - \bar{x})^2 / \{(n-1)(\psi'(b) + \psi'(1))\}$.

Table 2.2.1 Relative efficiencies of the LS estimators of μ and σ
(1) $RE(\tilde{\mu}) = [MSE(\hat{\mu}) / MSE(\tilde{\mu})] * 100$
(2) $RE(\tilde{\sigma}) = [MSE(\hat{\sigma}) / MSE(\tilde{\sigma})] * 100$

		b =	0.5	1.0	2.0	4.0	8.0
n = 10	(1)		95.11	93.95	99.30	88.82	76.92
	(2)		93.90	109.37	97.85	86.30	80.05
n = 15	(1)		92.91	93.53	98.68	85.15	73.39
	(2)		88.66	101.82	93.05	78.77	75.74
n = 20	(1)		93.19	92.13	97.99	83.06	70.17
	(2)		84.83	99.66	90.08	76.21	72.68
n = 50	(1)		90.70	90.96	96.95	79.19	62.74
	(2)		74.24	91.02	83.30	70.62	64.02

It can be seen that $\hat{\mu}$ and $\hat{\sigma}$ are considerably more efficient than $\tilde{\mu}$ and $\tilde{\sigma}$ even for small sample sizes. A disconcerting feature of $\tilde{\mu}$ and $\tilde{\sigma}$ is that their relative efficiencies decrease as sample size n increases.

2.2.4 Robustness of Estimators

To show the robustness of both MML estimators, we consider, for illustration, the following plausible alternatives (1)-(4) to the assumed distribution GL in (2.2.1) with b = 4.0:

(1) Misspecification of the distribution: GL(3.5, σ)

(2) Misspecification of the distribution: GL(4.5, σ)

(3) Dixon's outlier model: (n-1) observations come from GL(4.0, σ) but one observation (we do not know which one) comes from GL(4.0, 4σ)

(4) Mixture model: 0.90 GL(4.0, σ) + 0.10 GL(4.0, 4σ)

(5) Contamination model: 0.90 GL(4.0, σ) + 0.10 Uniform(-1/2, 1/2)

The simulated relative efficiencies of the LS estimators of μ and σ are given in Table 2.2.2. It is obvious that the MML estimators $\hat{\mu}$ and $\hat{\sigma}$ are remarkably efficient and robust.

Table 2.2.2 Relative efficiencies of the LS estimators of μ and σ; n = 10, σ = 1

	Models				
Model	(1)	(2)	(3)	(4)	(5)
RE($\tilde{\mu}$)	85.14	84.19	28.47	34.76	84.20
RE($\tilde{\sigma}$)	80.71	95.70	44.31	53.93	84.01

CHAPTER 3

MAXIMUM LIKELIHOOD HYPOTHESES UNDER NONNORMALITY

In this chapter, the problem faced by the learner is to learn an unknown function $f : X \to \Re$ drawn from a hypothesis space H consisting of some class of real-valued functions defined over an instance space X (i.e., $\forall h \in H$ is a function of the form $h : X \to \Re$). In the first section, noise is assumed to have a distribution from Generalized Secant Hyperbolic (GSH) family and in the second section, it is assumed to have a distribution from Generalized Logistic (GL) family.

3.1 Maximum Likelihood Hypotheses with Symmetric Non-normally Distributed Noise

Assume the target value of each example is corrupted by random noise drawn according to a GSH distribution with mean zero, variance σ^2 and shape parameter t in (1.1.1.10). Therefore, the probability density function is

$$p(d_i \mid h) = \frac{c_1}{\sigma} \frac{\exp(c_2(d_i - \mu)/\sigma)}{\exp(2c_2(d_i - \mu)/\sigma) + 2a\exp(c_2(d_i - \mu)/\sigma) + 1} \qquad (3.1.1)$$

where $-\infty < d_i < \infty$ (i = 1, 2,..., n), $\mu \in \Re$, $\sigma > 0$ and

for $-\pi < t \le 0$: $a = \cos(t)$, $c_2 = \sqrt{(\pi^2 - t^2)/3}$ and $c_1 = \frac{\sin(t)}{t} c_2$,

for $t > 0$: $a = \cosh(t)$, $c_2 = \sqrt{(\pi^2 + t^2)/3}$ and $c_1 = \frac{\sinh(t)}{t} c_2$.

The maximum likelihood hypothesis is

$$h_{ML} = \arg\max_{h \in H} \prod_{i=1}^{n} p(d_i | h)$$

$$= \arg\max_{h \in H} \prod_{i=1}^{n} \frac{c_1}{\sigma} \frac{\exp(c_2 (d_i - \mu)/\sigma)}{\exp(2c_2 (d_i - \mu)/\sigma) + 2a \exp(c_2 (d_i - \mu)/\sigma) + 1}. \qquad (3.1.2)$$

By substituting $\mu = f(x_i) = h(x_i)$, we obtain

$$h_{ML} = \arg\max_{h \in H} \left\{ \frac{c_1^n}{\sigma^n} \frac{\exp(c_2 \sum_{i=1}^{n} (d_i - h(x_i))/\sigma)}{\prod_{i=1}^{n} \left[\exp(2c_2 (d_i - h(x_i))/\sigma) + 2a \exp(c_2 (d_i - h(x_i))/\sigma) + 1\right]} \right\}$$

$$= \arg\max_{h \in H} -n \ln \sigma + \frac{c_2}{\sigma} \sum_{i=1}^{n} (d_i - h(x_i)) - \sum_{i=1}^{n} \ln\left[\exp(2c_2 (d_i - h(x_i))/\sigma) + 2a \exp(c_2 (d_i - h(x_i))/\sigma) + 1\right]. \qquad (3.1.3)$$

i) Maximum Likelihood Estimation

Now, to estimate σ, the likelihood equation is

$$\frac{\partial \ln L}{\partial \sigma} = -n \frac{1}{\sigma} - \frac{c_2}{\sigma} \sum_{i=1}^{n} z_i + 2 \frac{c_2}{\sigma} \sum_{i=1}^{n} z_i g(z_i) = 0 \qquad (3.1.4)$$

where

$$g(z_i) = \frac{\exp(2c_2 z_i) + a \exp(c_2 z_i)}{\exp(2c_2 z_i) + 2a \exp(c_2 z_i) + 1} \quad \text{and} \quad z_i = \frac{d_i - h(x_i)}{\sigma}.$$

Equation (3.1.4) has no explicit solution since the terms involve the nonlinear function $g(z_i)$. Therefore, modified maximum likelihood estimation technique is used (Vaughan, 2002).

ii) Modified Maximum Likelihood Estimation

Since complete sums are invariant to ordering, the likelihood equations can be written as follows:

$$\frac{\partial \ln L}{\partial \sigma} = -n\frac{1}{\sigma} - \frac{c_2}{\sigma}\sum_{i=1}^{n} z_{(i)} + 2\frac{c_2}{\sigma}\sum_{i=1}^{n} z_{(i)} g(z_{(i)}) = 0 \qquad (3.1.5)$$

where $z_{(i)} = \dfrac{d_{(i)} - h(x_{[i]})}{\sigma}$ $(1 \le i \le n)$.

Let $t_{(i)} = E(z_{(i)})$, $t_{(i)}$ $(1 \le i \le n)$ are the expected values of the standardized ordered variates. For large n, the values of $t_{(i)}$ can be found as follows:

$$t_{(i)} = \frac{1}{c_2}\ln\left[\frac{\sin\left(t\dfrac{i}{n+1}\right)}{\sin\left(t\left(1-i/(n+1)\right)\right)}\right], \quad -\pi < t < 0$$

$$= \frac{\sqrt{3}}{\pi}\ln\left(\frac{i/(n+1)}{1-i/(n+1)}\right) \qquad , \quad t = 0$$

$$= \frac{1}{c_2}\ln\left[\frac{\sinh\left(t\dfrac{i}{n+1}\right)}{\sinh\left(t\left(1-i/(n+1)\right)\right)}\right], \quad t > 0. \qquad (3.1.6)$$

For small sample size n, the values of $t_{(i)}$ can be found by using the formula of $E(d_{(i)})$ given by Vaughan (2002).

Since $z_{(i)}$ is located in the vicinity of $t_{(i)}$, it is approximated by the Taylor series expansion

$$\begin{aligned} g(z_{(i)}) &\cong g(t_{(i)}) + (z_{(i)} - t_{(i)})g'(t_{(i)}) \\ &= \alpha_i + \beta_i z_{(i)} \end{aligned} \tag{3.1.7}$$

where

$$\alpha_i = \frac{\exp(2c_2 t_{(i)}) + a\exp(c_2 t_{(i)})}{\exp(2c_2 t_{(i)}) + 2a\exp(c_2 t_{(i)}) + 1} - \beta_i t_{(i)}$$

and

$$\beta_i = \frac{ac_2\exp(3c_2 t_{(i)}) + 2c_2\exp(2c_2 t_{(i)}) + ac_2\exp(c_2 t_{(i)})}{\left[\exp(2c_2 t_{(i)}) + 2a\exp(c_2 t_{(i)}) + 1\right]^2}.$$

Since $t_{(i)} = -t_{(n-i+1)}$ from symmetry, $\sum_{i=1}^{n}\alpha_i = \frac{n}{2}$ and $\sum_{i=1}^{n}\beta_i t_{(i)} = 0$.

Incorporating (3.1.7) in (3.1.5) gives the modified likelihood equation as follows:

$$\frac{\partial \ln L}{\partial \sigma} \cong \frac{\partial \ln L^*}{\partial \sigma} = -n\frac{1}{\sigma} - \frac{c_2}{\sigma}\sum_{i=1}^{n} z_{(i)} + 2\frac{c_2}{\sigma}\sum_{i=1}^{n}\left[\alpha_i + \beta_i z_{(i)}\right] = 0. \tag{3.1.8}$$

The solution of the equation (3.1.8) is the modified maximum likelihood estimator:

$$\hat{\sigma} = \frac{-B + \sqrt{B^2 + 4nC}}{2\sqrt{n(n-1)}} \tag{3.1.9}$$

where $B = nc_2(\bar{d} - \bar{d}_a)$, $\bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_{(i)}$, $\bar{d}_a = \frac{2}{n}\sum_{i=1}^{n}\alpha_i d_{(i)}$,

$$C = 2c_2 \sum_{i=1}^{n} \beta_i (d_{(i)} - \hat{\mu})^2 = 2c_2 \left[\sum_{i=1}^{n} \beta_i d_{(i)}^2 - m\hat{\mu}^2 \right],$$

$$\hat{\mu} = \frac{1}{m} \sum_{i=1}^{n} \beta_i d_{(i)} \text{ and } m = \sum_{i=1}^{n} \beta_i .$$

Now, we can find the maximum likelihood hypothesis by substituting the estimator of the variance (3.1.9) into the equation (3.1.3). Hence the maximum likelihood hypothesis is

$$h_{ML} = \arg\max - n \ln \hat{\sigma} + \frac{c_2}{\hat{\sigma}} \sum_{i=1}^{n} (d_i - h(x_i)) - \sum_{i=1}^{n} \ln\left[\exp(2c_2 (d_i - h(x_i))/\hat{\sigma}) + 2a \exp(c_2 (d_i - h(x_i))/\hat{\sigma}) + 1\right] \quad (3.1.10)$$
$$h \in H$$

3.2 Maximum Likelihood Hypotheses with Skewed Non-normally Distributed Noise

Now, assume that the noise e_i is coming from a GL distribution with zero mean, variance σ^2 and shape parameter b. Therefore, the probability density function is

$$p(d_i \mid h) = \frac{b}{\sigma} \frac{\exp(-(d_i - \mu)/\sigma)}{\left[1 + \exp(-(d_i - \mu)/\sigma)\right]^{b+1}} \quad (3.2.1)$$

where $-\infty < d_i < \infty$ (i = 1, 2,..., n), $\mu \in \Re$, $\sigma > 0$.

The maximum likelihood hypothesis is

$$h_{ML} = \arg\max_{h \in H} \prod_{i=1}^{n} \frac{b}{\sigma} \frac{\exp(-(d_i - \mu)/\sigma)}{\left[1 + \exp(-(d_i - \mu)/\sigma)\right]^{b+1}} . \quad (3.2.2)$$

By substituting $\mu = f(x_i) = h(x_i)$, we obtain

$$h_{ML} = \arg\max_{h \in H} \left\{ \frac{b^n}{\sigma^n} \frac{\exp(-\sum_{i=1}^{n} (d_i - h(x_i))/\sigma)}{\prod_{i=1}^{n} [1 + \exp(-(d_i - h(x_i))/\sigma]^{b+1}} \right\}. \quad (3.2.3)$$

By taking the natural logarithm of (3.2.3), the following maximum likelihood hypothesis is obtained:

$$h_{ML} = \arg\max_{h \in H} \left\{ -n \ln \sigma - \frac{1}{\sigma} \sum_{i=1}^{n} (d_i - h(x_i)) - (b+1) \sum_{i=1}^{n} \ln[1 + \exp(-(d_i - h(x_i))/\sigma)] \right\}. \quad (3.2.4)$$

i) Maximum Likelihood Estimation

Now, to estimate σ, the likelihood equation is

$$\frac{\partial \ln L}{\partial \sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma} \sum_{i=1}^{n} z_i - \frac{b+1}{\sigma} \sum_{i=1}^{n} z_i g(z_i) = 0 \quad (3.2.5)$$

where $g(z_i) = \frac{1}{1 + \exp(z_i)}$ and $z_i = \frac{d_i - h(x_i)}{\sigma}$.

Since equation (3.2.5) does not admit explicit solution, Tiku's modified maximum likelihood estimation technique can be used (Şenoğlu, 2000).

ii) Modified Maximum Likelihood Estimation

Since complete sums are invariant to ordering, the likelihood equations can be written as follows:

$$\frac{\partial \ln L}{\partial \sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma} \sum_{i=1}^{n} z_{(i)} - \frac{b+1}{\sigma} \sum_{i=1}^{n} z_{(i)} g(z_{(i)}) = 0. \quad (3.2.6)$$

Since $z_{(i)}$ is located in the vicinity of $t_{(i)}$, it is approximated by the Taylor series expansion

$$\begin{aligned} g(z_{(i)}) &\cong g(t_{(i)}) + (z_{(i)} - t_{(i)})g'(t_{(i)}) \\ &= \alpha_i - \beta_i z_{(i)} \end{aligned} \tag{3.2.7}$$

where

$$\alpha_i = \frac{t_{(i)} \exp(t_{(i)}) + \exp(t_{(i)}) + 1}{(\exp(t_{(i)}) + 1)^2},$$

$$\beta_i = \frac{\exp(t_{(i)})}{(\exp(t_{(i)}) + 1)^2}$$

and for large n

$$t_{(i)} = -\ln\left[\left(\frac{i}{n+1}\right)^{-1/b} - 1\right].$$

Incorporating (3.2.7) in (3.2.6) gives the modified likelihood equation as follows:

$$\frac{\partial \ln L}{\partial \sigma} \cong \frac{\partial \ln L^*}{\partial \sigma} = -n\frac{1}{\sigma} + \frac{1}{\sigma}\sum_{i=1}^{n} z_{(i)} - \frac{b+1}{\sigma}\sum_{i=1}^{n} z_{(i)}\left[\alpha_i - \beta_i z_{(i)}\right] = 0. \tag{3.2.8}$$

The solution of the equation (3.2.8) is the modified maximum lilelihood estimator:

$$\hat{\sigma} = \frac{B + \sqrt{B^2 + 4nC}}{2\sqrt{n(n-1)}} \tag{3.2.9}$$

where $B = (b+1)\sum_{i=1}^{n} \Delta_i (d_{(i)} - \hat{\mu})$, $\Delta_i = (b+1)^{-1} - \alpha_i$, $\hat{\mu} = \frac{1}{m}\sum_{i=1}^{n} \beta_i d_{(i)}$ and

$$C = (b+1)\sum_{i=1}^{n}\beta_i (d_{(i)} - \hat{\mu})^2 = (b+1)\left[\sum_{i=1}^{n}\beta_i d_{(i)}^2 - m\hat{\mu}^2\right],\ m = \sum_{i=1}^{n}\beta_i .$$

Now, we can find the maximum likelihood hypothesis by substituting the estimator of the variance (3.2.9) into the equation (3.2.4). Hence the maximum likelihood hypothesis is

$$h_{ML} = \arg\max_{h\in H}\left\{-n\ln\hat{\sigma} - \frac{1}{\hat{\sigma}}\sum_{i=1}^{n}(d_i - h(x_i)) - (b+1)\sum_{i=1}^{n}\ln\left[1+\exp(-(d_i - h(x_i))/\hat{\sigma})\right]\right\}. \quad (3.2.10)$$

CHAPTER 4

APPLICATIONS AND CONCLUSIONS

4.1 Applications

Example 1: Lindsey et al. (1987) give the measurements of ten insects having three attributes for each of three species of a type of insect, Chaetocnema. This data is reproduced in Hand et al. (1994, p.190) and given in Table 4.1.1.

Table 4.1.1 Insect data

Species I			Species II			Species III		
x_1	x_2	x_3	x_1	x_2	x_3	x_1	x_2	x_3
191	131	53	186	107	49	158	141	58
185	134	50	211	122	49	146	119	51
200	137	52	201	144	47	151	130	51
173	127	50	242	131	54	122	113	45
171	128	49	184	108	43	138	121	53
160	118	47	211	118	51	132	115	49
188	134	54	217	122	49	131	127	51
186	129	51	223	127	51	135	123	50
174	131	52	208	125	50	125	119	51
163	115	47	199	124	46	130	120	48

The variable x_1 (microns) is the width of the first joint of the first tarsus, x_2 (microns) is the width of the first joint of the second tarsus and x_3 (microns) is the maximal width of the aedegus. Since in a real situation it is not known which insect belongs to which species, the object is to classify new insects with high accuracy. The measurements of six new insects given by Lindsey (1987) are given in Table 4.1.2.

Table 4.1.2 Insect data: new insects

x_1	x_2	x_3
190	143	52
174	131	50
211	129	49
128	126	49
130	131	51
138	127	52

Since the aim of this study is to determine which species new insects belong to, training set contains all of the measurements of ten insects. To locate the plausible distribution of the i^{th} (i = 1, 2, 3) feature in the j^{th} (j = 1, 2, 3) class, the Q-Q plot is used. Hamilton (1992, p.16) has very useful Q-Q plots constructed from random samples which identify a variety of distributions, e.g., long-tailed, short-tailed, negatively skewed, positively skewed, etc. The Q-Q plots of data generally indicate symmetric distributions. Therefore, GSH distribution assumption is used for the model distribution. For the j^{th} class, to find the most appropriate value of the shape parameter t_i, the MML estimators of μ_i and σ_i are calculated from (2.1.2.8) and (2.1.2.9) for a series of values of t_i and the values of (2.1.2.14) are calculated. The kurtosis values corresponding to the shape parameters t_i (i = 1, 2, 3) which maximize (2.1.2.14) for each class are as follows:

		Feature		
		x_1	x_2	x_3
Class	Species I	2.0	2.0	2.0
	Species II	2.0	6.4	5.3
	Species III	2.0	6.0	9.1

The LS and MML estimates of the parameters μ_i and σ_i for each class are given in Table 4.1.3.

Table 4.1.3 Insect data: LS and MML estimates of the parameters μ and σ

Class	Feature	$\tilde{\mu}$	$\tilde{\sigma}$	$\hat{\mu}$	$\hat{\sigma}$
Species I	x_1	179.1000	12.87935	179.5563	12.22341
	x_2	128.4000	6.995237	126.1130	6.743501
	x_3	50.50000	2.368778	50.42988	2.180407
Species II	x_1	208.2000	17.22272	211.6443	17.72048
	x_2	122.8000	10.71655	122.9357	11.69911
	x_3	48.90000	3.034981	49.10543	3.227953
Species III	x_1	136.8000	11.55470	139.6468	11.02152
	x_2	122.8000	8.148619	121.2471	8.685871
	x_3	50.70000	3.368152	50.68508	3.835470

To compare the accuracy of naive Bayes classifiers under normal and GSH distribution assumptions, the training set is used again and to make classification in the training set, the class which maximizes (1.1.1.9) is selected by using the class conditional probability density in (1.1.1.7) or (2.1.2.15) under the Normal and GSH distributions assumptions, respectively. As a result, the classification rates of naive Bayes classifier under normality and GSH naive Bayes classifier are obtained as in Table 4.1.4.

Table 4.1.4 Insect data: classification rates

	Normal	GSH
Classification rate	96.67%	100.00%

Since the MML estimators in GSH naive Bayes classifier are robust, the method can easily adapts itself to the data and it classifies the data correctly. Hence, when determining which species new insects belong to, GSH naive Bayes classifier is expected to classify the data more accurately. Table 4.1.5 shows the species that are matched to the new insects when the classifiers are applied.

Table 4.1.5 Insect data: classification of new insects

Insect	x_1	x_2	x_3	Class obtained by naive Bayes classifier	Class obtained by GSH naive Bayes classifier
1	190	143	52	1	2
2	174	131	50	1	1
3	211	129	49	2	2
4	218	126	49	2	2
5	130	131	51	3	3
6	138	127	52	3	3

First insect is classified differently by the two classification methods. As it can be seen from that application, different distribution assumptions cause different results. Therefore, the underlying distribution should be fitted by using Q-Q plots and goodness-of-fit tests or determining the value of a shape paramter by maximizing lnL (Tiku and Akkaya, 2004).

Example 2: The data in Appendix C, taken from Fisher (1936), are the measurements of the sepal length, sepal width, petal length and petal width in centimeters of fifty plants for each of the three types of iris: Iris setosa, Iris versicolor and Iris virginica. Clearly the main problem is classification. The data is split in two nonoverlapping sets: the training set and the test set. Training set contains 60% of the observations of each class. The estimation of the parameters is done using only the training set. The classification is performed using only the test set. By using Q-Q plots of training set given in Appendix C, it can be said that for all of the features of the first class, GSH distribution; for all of the features of the second class, GL distribution and for other than the third feature of the third class, GSH distribution assumptions are appropriate. Furthermore, if the i^{th} (i = 1, 2, 3, 4) feature of the j^{th} (j = 1, 2, 3) class has GSH distribution, by the method that is explained in the previous example for the determination of shape parameters, the following kurtosis values corresponding to the shape parameters t_i which maximize (2.1.2.14) for each class are obtained:

		Feature			
		Sepal length	Sepal width	Petal length	Petal width
Class	Iris setosa	1.9	3.4	4.0	7.2
	Iris virginica	4.4	3.8		2.3

If the i^{th} (i = 1, 2, 3, 4) feature of the j^{th} (j = 1, 2, 3) class has GL distribution, to find the most appropriate value of the shape parameter b_i, the MML estimators of μ_i and σ_i are calculated from (2.2.2.6) and (2.2.2.7) for a series of values of b_i and the values of (2.2.2.9) are calculated. The shape parameters b_i which maximize (2.2.2.9) for each class are as follows:

		Feature			
		Sepal length	Sepal width	Petal length	Petal width
Class	Iris versicolor	2.3	0.3	0.2	0.4
	Iris virginica			9.5	

The MML and LS estimates of the parameters μ_i and σ_i for each class are obtained as in Table 4.1.6.

Table 4.1.6 Iris data: LS and MML estimates of the parameters μ and σ

Class	Feature	$\tilde{\mu}$	$\tilde{\sigma}$	$\hat{\mu}$	$\hat{\sigma}$
Iris setosa	Sepal length	5.053333	0.3936858	5.048231	0.4292984
	Sepal width	3.480000	0.4397492	3.482978	0.4436810
	Petal length	1.436666	0.1629117	1.433125	0.1647454
	Petal width	0.2466667	0.1105888	0.2185031	0.1234213
Iris versicolor	Sepal length	5.863333	0.5555261	5.405012	0.3983227
	Sepal width	2.750000	0.3319223	3.038163	0.1012718
	Petal length	4.180000	0.4693356	4.643937	0.1000344
	Petal width	1.293333	0.1799106	1.426778	0.068472303
Iris virginica	Sepal length	6.523333	0.6317372	6.519863	0.6507546
	Sepal width	2.956666	0.3500575	2.952808	0.3581837
	Petal length	2.086667	0.2674701	1.965022	0.1858948
	Petal width	2.086667	0.2674701	2.084731	0.2626103

Note that LS estimates of the parameters of GL distribution are not bias corrected.

To make classification in the test set, the class which maximizes (1.1.1.9) is selected by using the class conditional probability density in (1.1.1.7) and

$$p(\underline{x}|w_j, D) = \prod_{i=1}^{d} \left[\frac{c_{1i}}{\hat{\sigma}_i} \frac{\exp(c_{2i}(x_i - \hat{\mu}_i)/\hat{\sigma}_i)}{\exp(2c_{2i}(x_i - \hat{\mu}_i)/\hat{\sigma}_i) + 2a_i \exp(c_{2i}(x_i - \hat{\mu}_i)/\hat{\sigma}_i) + 1} \right]^{q} \times \left[\frac{b_i}{\hat{\sigma}_i} \frac{\exp(-(x_i - \hat{\mu}_i)/\hat{\sigma}_i)}{[\exp(-(x_i - \hat{\mu}_i)/\hat{\sigma}_i) + 1]^{b_i+1}} \right]^{1-q} \quad (4.1.1)$$

where q = 0 if the distribution of i^{th} feature of the j^{th} class is GL and
1 if the distribution of i^{th} feature of the j^{th} class is GSH

under the normal and non-normal distributions assumptions, respectively. As a result, the classification rates of naive Bayes classifiers under normality and non-normality are given in Table 4.1.7.

Table 4.1.7 Iris data: classification rates

	Normal	Non-normal
Classification rate	93.33%	95.00%

In conclusion, non-normal naive Bayes classifier improves the classification rate.

Example 3: Atkinson and Riani (2000) give the data in Table 4.1.8 taken from Carr (1960). The observations are from an experiment on the catalytic isomerization of n-pentane to iso-pentane in the presence of hydrogen. There are 24 observations and the variables are rate of disappearance of n-pentane (d), partial pressure of hydrogen (x_1), partial pressure of n-pentane (x_2) and partial pressure of iso-pentane (x_3).

Table 4.1.8 Reaction rate for the catalytic isomerization of n-pentane to isopentane

Partial Pressures (psia)			Rate
x_1	x_2	x_3	d
205.8	90.9	37.1	3.541
404.8	92.9	36.3	2.397
209.7	174.9	49.4	6.694
401.6	187.2	44.9	4.722
224.9	92.7	116.3	0.593
402.6	102.2	128.9	0.268
212.7	186.9	134.4	2.797
406.2	192.6	134.9	2.451
133.3	140.8	87.6	3.196
470.9	144.2	86.9	2.021
300.0	68.3	81.7	0.896
301.6	214.6	101.7	5.084
297.3	142.2	10.5	5.686
314.0	146.7	157.1	1.193
305.7	142.0	86.0	2.648
300.1	143.7	90.2	3.303
305.4	141.1	87.4	3.054
305.2	141.5	87.0	3.302
300.1	83.0	66.4	1.271
106.6	209.6	33.0	11.648
417.2	83.9	32.9	2.002
251.0	294.4	41.5	9.604
250.3	148.0	14.7	7.754
145.1	291.0	50.2	11.590

Islam and Tiku (2004) found that multiple linear regression model is reasonable for this data set. The data is split in two nonoverlapping sets: the training set and the test set. Training set contains 70% of the observations. The hypotheses are constructed by using only the training set. If the distribution of the error terms is assumed to be normal, the following model with the least square estimates is obtained:

$$h_1(x) = 2.832 - 0.006x_1 + 0.033x_2 - 0.033x_3.$$

However, Islam and Tiku (2004) showed that Generalized Logistic distribution is appropriate for the error distribution. A plausible value of the shape parameter is

identified as 0.5 for the training set. Therefore, the model with the modified maximum likelihood estimates is in the following:

$$h_2(x) = 3.275 - 0.007x_1 + 0.034x_2 - 0.032x_3 .$$

Note that the hypothesis $h_2(x)$ is obtained by using the modified maximum likelihood estimators of the parameters in multiple linear regression model (Islam and Tiku, 2004; Tiku and Akkaya, 2004).

Hence, there are two hypotheses in the hypotheses space H and the object is finding the maximum likelihood hypothesis. The maximum likelihood hypothesis is found from the training set and the test set is used for showing the validty of the hypothesis. If the distribution of noise is assumed to obey normal distribution, the following sum of squared errors are obtained for the training and test sets:

	Hypotheses:	$h_1(x)$	$h_2(x)$
Training set:	$\sum_{i=1}^{17}(d_i - h(x_i))^2$	3.149	3.936
Test set:	$\sum_{i=1}^{7}(d_i - h(x_i))^2$	20.645	14.050

Under the normality assumption, since the maximum likelihood hypothesis is the one that minimizes the sum of the squared errors between the observed training values and the hypothesis predictions, the training set indicates that $h_1(x)$ is the maximum likelihood hypothesis. However, the Euclidean distance between the observed test values and the hypothesis predictions points out that $h_2(x)$ gives more accurate results that $h_1(x)$. Now, assume noise is distributed according to GL distribution. Since the values of

$$q = -n\ln\hat{\sigma} - \frac{1}{\hat{\sigma}}\sum_{i=1}^{n}(d_i - h(x_i)) - (b+1)\sum_{i=1}^{n}\ln\left[1 + \exp(-(d_i - h(x_i))/\hat{\sigma})\right]$$

in equation (3.2.10) are as in the following,

Hypotheses:	$h_1(x)$	$h_2(x)$
q	3.149	3.936

the maximum likelihood hypothesis is $h_2(x)$. Hence it chooses the true hypothesis under the assumption that noise is distributed according to GL distribution.

Example 4: Atkinson and Riani (2000) give the stack loss data taken from Brownlee (1965). The data is reproduced in Table 4.1.9.

Table 4.1.9 Stack loss data on the oxidation of ammonia

Air Flow	Cooling Water Inlet Temperature	Acid Concentration	Stack Loss
x_1	x_2	x_3	d
80	27	89	42
80	27	88	37
75	25	90	37
62	24	87	28
62	22	87	18
62	23	87	18
62	24	93	19
62	24	93	20
58	23	87	15
58	18	80	14
58	18	89	14
58	17	88	13
58	18	82	11
58	19	93	12
50	18	89	8
50	18	86	7
50	19	72	8
50	19	79	8
50	20	80	9
56	20	82	15
70	20	91	15

There are observations from 21 days of operation of a plant for the oxidation of ammonia as a stage in the production of nitric acid. The air flow (x_1) measures the rate of operation of the plant. The nitric oxides produced are absorbed in a countercurrent absorption tower; x_2 is the inlet temperature of cooling water circulating through coils in this tower, x_3 (=10*(acid concentration-50)) is proportional to the concentration of acid in the tower and d represents stack loss that is 10 times the percentage of ingoing ammonia escaping unconverted.

As indicated in Islam and Tiku (2004), a multiple linear regression model is appropriate and the observation (x_1 = 70, x_2 = 20, x_3 = 91 and d = 15) has an abnormally large residual when a standard least squares regression is fitted as it can be understood from the normal Q-Q plot given in Andrews (1974). Therefore, the training set which contains 70% of the observations is chosen without including that anomalous observation. Although it is mentioning that the observations (x_1 = 80, x_2 = 27, x_3 = 89 and d = 42), (x_1 = 75, x_2 = 25, x_3 = 90 and d = 37) and (x_1 = 62, x_2 = 24, x_3 = 87 and d = 28) are anomalous, they are included in the training set and the following hypothesis is constructed when the least square estimates are used:

$$h_1(x) = -60.734 + 0.598x_1 + 1.688x_2 + 0.090x_3 .$$

Furthermore, since Islam and Tiku (2004) showed that the distribution of residuals is long-tailed symmetric, the following hypothesis is constructed by using the modified maximum likelihood estimators which were derived in Islam and Tiku (2004):

$$h_2(x) = -60.683 + 0.608x_1 + 1.676x_2 + 0.085x_3 .$$

First, let the test set do not contain the observation (x_1 = 70, x_2 = 20, x_3 = 91 and d = 15). If the distribution of noise is assumed to obey normal distribution, the following sum of squared errors are obtained for the training and test sets:

Hypotheses:		$h_1(x)$	$h_2(x)$
Training set:	$\sum_{i=1}^{14}(d_i - h(x_i))^2$	58.426	58.468
Test set:	$\sum_{i=1}^{6}(d_i - h(x_i))^2$	131.217	128.091

The training set indicates that $h_1(x)$ is the maximum likelihood hypothesis. However, the Euclidean distance between the observed test values and the hypothesis predictions points out that $h_2(x)$ gives more accurate results that $h_1(x)$. Now, assume noise is distributed according to GSH distribution. A plausible value of the shape parameter is identified as 9.6 for the training set. Since the values of

$$q = -n\ln\hat{\sigma} + \frac{c_2}{\hat{\sigma}}\sum_{i=1}^{n}(d_i - h(x_i)) - \sum_{i=1}^{n}\ln\left[\exp\left(2c_2\frac{d_i - h(x_i)}{\hat{\sigma}}\right) + 2a\exp\left(c_2\frac{d_i - h(x_i)}{\hat{\sigma}}\right) + 1\right]$$

in equation (3.1.10) are as in the following,

Hypotheses:	$h_1(x)$	$h_2(x)$
q	-22.38572	-22.38540

the maximum likelihood hypothesis is $h_2(x)$. Hence it chooses the true hypothesis under the assumption that noise is distributed according to GSH distribution.

Now, consider the test set containing the anomalous observation. Since square of the Euclidean distances between the observed test values and the hypothesis predictions are obtained as in the following,

Hypotheses:		$h_1(x)$	$h_2(x)$
Test set:	$\sum_{i=1}^{7}(d_i - h(x_i))^2$	196.4384	194.2203

it is again obvious that $h_2(x)$ predicts the stack loss more accurately. As a result, the maximum likelihood hypothesis found under GSH distribution assumption gives better results than the one which is found under normality assumption.

4.2 Conclusions

In machine learning, Bayesian learning is a statistical approach which depends on normality assumption. However, assuming normal as the underlying distribution is unrealistic and it might cause erroneous statistical inferences (Tiku and Akkaya, 2004). Hence a plausible underlying distribution should be identified. In this study, two families of distributions are considered: Generalized Secant Hyperbolic (GSH) and Generalized Logistic (GL). These families contain five types of distributions:

i) long-tailed symmetric,
ii) short-tailed symmetric,
iii) negatively skewed,
iv) positively skewed, and
v) approximately normal.

Generalized Secant Hyperbolic and Generalized Logistic naive Bayes classifiers using GSH and GL distribution families as the model distribution, respectively are introduced. The difficulty of these models is in estimating the parameters of GSH and GL distributions. Since the maximum likelihood equations do not have explicit solutions, they have to be solved by iteration which can be problematic for reasons of

i) multiple roots,
ii) slow convergence, or
iii) convergence to wrong values.

Furthermore, since it is not possible to identify the underlying distribution exactly from a sample, a method of estimation which yields robust estimators is needed. Modified

maximum likelihood estimators are explicit functions of the sample observations and easy to compute. Besides, these estimators are fully efficient asymptotically, highly efficient for small sample size and robust to plausible deviations, therefore, it is sufficient to locate a distribution in reasonable proximity to the true distribution. This can easily be accomplished by constructing Q-Q plots followed by a formal goodness-of-fit test and a viable alternative to a goodness-of-fit test is to determine the value of the shape parameter by maximizing the natural logarithm of the likelihood function (Tiku and Akkaya, 2004). By real life applications of the procedures, it is shown that GSH and GL naive Bayes classifiers improve the true classification rates even for small sample sizes.

In addition, the maximum likelihood hypotheses are obtained under the assumption of non-normality. When the real life applications are carried out, it is shown that GSH and GL distribution assumptions provide better results than normality assumption.

REFERENCES

Abramowitz, M. and Stegun, I. A. (1985). Handbook of Mathematical Functions. Dover: New York.

Andrews D. F., Bickel P. J., Hampel F. R., Huber P. J., Rogers W. H. and Tukey J. W. (1972). Robust Estimates of Location. Princeton University Press: Princeton, N. J.

Atkinson, A. and Riani, M. (2000). Robust Diagnostic Regression Analysis. Springer: New York

Bartlett, M.S. (1953). Approximate confidence intervals. Biometrika 40, 12-19.

Bhattacharyya, G. K. (1985). The asymptotics of maximum likelihood and related estimators based on type II censored data. J. Amer. Statist. Assoc. 80, 398-404.

Brownlee, K. A. (1965). Statistical Theory and Methodology in Science and Engineering. Wiley: New York.

Carr, N. L. (1960). Kinetics of catalytic isomerisation of n-pentane. Industrial and Engineering Chemistry 52, May, 391-396.

Domingos, P. and Pazzani, M. (1996). Beyond independence: Conditions for the optimality of the simple Bayesian classifier. In Proceedings of the Thirteenth International Conference on Machine Learning, 105-112.

Duda, R.O. and Hart, P.E. (1973). Pattern Classification and Scene Analysis. John Wiley & Sons.

Fisher, R.A. (1936). The use of multiple measurements in taxonomic problems. Annals of Eugenics, 7, 179-184.

Geary, R.C. (1947). Testing for normality. Biometrika 34, 209-242.

Hamilton, L.C. (1992). Regression with Graphics. California: Brooks/Cole Publishing Company: California.

Hand, D. J., Daly, F., Lunn, A. D., McConway, K. J. and Ostrowski, E. (1994). Small Data Sets. Chapman & Hall: New York.

Islam, M.Q., Tiku, M.L. (2004). Multiple linear regression model under non-normality. Commun. Stat.- Theory Meth., 33, 2443-2467.

Lindsey, J.C., Herzberg, A.M. and Watts, D.G. (1987). A method for cluster analysis based on projections and quantile-quantile plots. Biometrics, 43, 327-341.

Mitchell, T. M. (1997). Machine Learning. Mc Graw-Hill.

Sebe, N., Lew, M. S., Cohen, I., Garg, A. and Huang, T. S. (2002). Emotion recognition using a Cauchy naive Bayes classifier. In Proceedings of the Sixteenth International Conference on Pattern Recognition, 1, 17-20.

Şenoğlu, B. (2000). Experimental design under non-normality: Skew distributions. Ph.D. Thesis, Middle East Technical University, Ankara.

Tiku, M. L., Tan, W. Y. and Balakrishnan, N. (1986). Robust Inference. Marcel Dekker: New York.

Tiku, M. L. and Suresh, R. P. (1992). A new method of estimation for location and scale parameters. J. Stat. Plann. Inf., 30, 281-292.

Tiku, M. L. And Akkaya, A. (2004). Robust Estimation and Hypothesis Testing. New Delhi: New Age International Lim. Pub.

Vaughan, D. C. and Tiku, M. L. (2000). Estimation and hypothesis testing for a non-normal bivariate distribution with applications. J. Mathematical and Computer Modeling, 32, 53-67.

Vaughan, D. C. (2002). The Generalized Secant Hyperbolic distribution and its properties. Commun. Stat.-Theory Meth., 31, 219-238.

APPENDIX A

COVARIANCE MATRIX

A.1 GSH Distribution

The Fisher information matrix is

$$I = \begin{bmatrix} -E\left(\frac{\partial^2 \ln L}{\partial \mu_i^2}\right) & -E\left(\frac{\partial^2 \ln L}{\partial \mu_i \partial \sigma_i}\right) \\ -E\left(\frac{\partial^2 \ln L}{\partial \sigma_i \partial \mu_i}\right) & -E\left(\frac{\partial^2 \ln L}{\partial \sigma_i^2}\right) \end{bmatrix} \qquad \text{(A.1.1)}$$

where $E\left(\frac{\partial^2 \ln L}{\partial \mu_i \partial \sigma_i}\right) = 0$,

for $-\pi < t_i < 0$, $\quad E\left(\frac{\partial^2 \ln L}{\partial \mu_i^2}\right) = -\frac{c_{2i}^2 n(t_i - \sin t_i \cos t_i)}{2\sigma_i^2 t_i \sin^2 t_i}$ and

$$E\left(\frac{\partial^2 \ln L}{\partial \sigma_i^2}\right) = -\frac{n}{6\sigma_i^2}\left(\frac{\pi^2 - t_i^2}{\sin^2 t_i} - \frac{(\pi^2 - 3t_i^2)\cos t_i}{t_i \sin t_i}\right)$$

for $t_i \geq 0$, $\quad E\left(\frac{\partial^2 \ln L}{\partial \mu_i^2}\right) = -\frac{c_{2i}^2 n(\sinh t_i \cosh t_i - t_i)}{2\sigma_i^2 t_i \sinh^2 t_i}$ and

$$E\left(\frac{\partial^2 \ln L}{\partial \sigma_i^2}\right) = -\frac{n}{6\sigma_i^2}\left(\frac{(\pi^2 + 3t_i^2)\cosh t_i}{t_i \sinh t_i} - \frac{\pi^2 + t_i^2}{\sinh^2 t_i}\right).$$

The asymptotic variance-covarince matrix is $\mathbf{V} = (V_{ij})$, where

$$V_{11} = \frac{-1}{E\left(\frac{\partial^2 \ln L}{\partial \mu_i^2}\right)}, \quad V_{12} = V_{21} = 0, \quad V_{22} = \frac{-1}{E\left(\frac{\partial^2 \ln L}{\partial \sigma_i^2}\right)}. \tag{A.1.2}$$

A.2 GL Distribution

The Fisher information matrix is

$$I(\mu, \sigma) = \frac{n}{\sigma^2}\begin{bmatrix} \frac{b}{b+2} & \frac{b\{\psi(b+1) - \psi(2)\}}{b+2} \\ \frac{b[\psi(b+1) - \psi(2)]}{b+2} & 1 + \frac{b\{[\psi'(b+1) + \psi'(2)] + [\psi(b+1) - \psi(2)]^2\}}{b+2} \end{bmatrix} \tag{A.2.1}$$

where $\psi(x) = \Gamma'(x)/\Gamma(x)$ is the psi-function. The asymptotic variance-covarince matrix of $\hat{\mu}$ and $\hat{\sigma}$ is $\mathbf{V} = I^{-1}(\mu, \sigma)$.

APPENDIX B

LISTING OF SIMULATION PROGRAMS

B.1 Simulation for GSH Distribution

```
PROGRAM GSH
USE NUMERICAL_LIBRARIES
REAL T,B2,Y(100),G(100),PI,C1,C2,A,SIGMA,MLMU,MLSIGMA,BET(100)
REAL ALFA(100),T1(100),M,MLMUMEAN,MLSIGMAMEAN,MLMUVAR,
REAL MLSIGMAVAR,X(100),XBAR,XABAR,C,B,LSMU,LSSIGMA
REAL LSMUMEAN,LSMUVAR,LSSIGMAMEAN,LSSIGMAVAR
REAL LSSIGMAMSE,MLSIGMAMSE,RELSSIGMA,RELSMU,LSMUMSE
REAL MLMUMSE,LSDEF,MLDEF,MVBMU,MVBSIGMA,I22,MLMUE
REAL MLSIGMAE

OPEN (unit=1,file='c:\Concon\Documents\sonuc.txt')
PI=22.0/7.0

PRINT *,'ENTER THE KURTOSIS'
READ *,B2
IF (B2.GT.4.2) THEN
  T=-PI*SQRT((5.0*B2-21.0)/(5.0*B2-9.0))
ELSE IF (B2.EQ.4.2) THEN
  T=0.0
ELSE IF (B2.LT.4.2.AND.B2.GT.1.8) THEN
  T=PI*SQRT((21.0-5.0*B2)/(5.0*B2-9.0))
ENDIF
WRITE(1,*) 'SHAPE PARAMETER=',T
PRINT*,'ENTER THE SAMPLE SIZE'
READ*,N

SIGMA=1.0

   IF (T.GT.(-PI).AND.T.LE.0.0) THEN
  A=COS(T)
  C2=SQRT((PI*PI-T*T)/3.0)
  C1=(SIN(T)/T)*C2
ELSE IF (T.GT.0.0) THEN
  A=COSH(T)
  C2=SQRT((PI*PI+T*T)/3.0)
```

```
      C1=(SINH(T)/T)*C2
      ENDIF

      IF (T.GT.(-PI).AND.T.LT.0.0) THEN
       DO 16 I=1,N
        V2=I
        V1=V2/(N*1.0+1.0)
        T1(I)=LOG(SIN(T*V1)/SIN(T*(1.0-V1)))/C2
16     CONTINUE
      ELSE IF (T.EQ.0.0) THEN
       DO 17 I=1,N
        V2=I
        V1=V2/(N*1.0+1.0)
        W=(V1)/(1.0-V1)
        T1(I)=(SQRT(3.0)/PI)*LOG(W)
17     CONTINUE
      ELSE IF (T.GT.0.0) THEN
       DO 18 I=1,N
         V2=I
         V1=V2/(N*1.0+1.0)
         T1(I)=LOG(SINH(T*V1)/SINH(T*(1.0-V1)))/C2
18     CONTINUE
      ENDIF

      DO 20 I=1,N
       BET(I)=A*C2*EXP(3.0*C2*T1(I))+2.0*C2*EXP(2.0*C2*T1(I))
       BET(I)=BET(I)+A*C2*EXP(C2*T1(I))
       BET(I)=BET(I)/(EXP(2.0*C2*T1(I))+2.0*A*EXP(C2*T1(I))+1.0)**2
       ALFA(I)=A*EXP(C2*T1(I))+EXP(2.0*C2*T1(I))
       ALFA(I)=ALFA(I)/(EXP(2.0*C2*T1(I))+2.0*A*EXP(C2*T1(I))+1.0)
       ALFA(I)=ALFA(I)-BET(I)*T1(I)
20    CONTINUE

      DO 23 I=1,N
       IF (BET(I).LT.0.0) THEN
        BET(I)=0.0
        ALFA(I)=A*EXP(C2*T1(I))+EXP(2.0*C2*T1(I))
        ALFA(I)=ALFA(I)/(EXP(2.0*C2*T1(I))+2.0*A*EXP(C2*T1(I))+1.0)
       ENDIF
23    CONTINUE

      M=0.0
      DO 24 I=1,N
       M=M+BET(I)
24    CONTINUE
```

```
      MLMUMEAN=0.0
      MLSIGMAMEAN=0.0
      MLMUVAR=0.0
      MLSIGMAVAR=0.0
      LSMUMEAN=0.0
      LSSIGMAMEAN=0.0
      LSMUVAR=0.0
      LSSIGMAVAR=0.0

      NN=100000/N

      DO 100 L=1,NN
      CALL RNUN(N,G)
      IF (T.GT.(-PI).AND.T.LT.0.0) THEN
      DO 5 I=1,N
        Y(I)=LOG(SIN(T*G(I))/SIN(T*(1.0-G(I))))/C2
 5    CONTINUE
      ELSE IF (T.EQ.0.0) THEN
      DO 6 I=1,N
        Y(I)=(SQRT(3.0)/PI)*LOG(G(I)/(1.0-G(I)))
 6    CONTINUE
      ELSE IF (T.GT.0.0) THEN
      DO 7 I=1,N
        Y(I)=LOG(SINH(T*G(I))/SINH(T*(1.0-G(I))))/C2
 7    CONTINUE
      ENDIF

C FINDING MMLE
      CALL SVRGN(N,Y,X)

      MLMU=0.0
      DO 25 I=1,N
        MLMU=MLMU+BET(I)*X(I)
 25   CONTINUE
      MLMU=MLMU/M
      MLMUMEAN=MLMUMEAN+MLMU
      MLMUVAR=MLMUVAR+MLMU**2

      XBAR=0.0
      XABAR=0.0
      DO 30 I=1,N
        XBAR=XBAR+X(I)
        XABAR=XABAR+ALFA(I)*X(I)
 30   CONTINUE
      XBAR=XBAR/(N*1.0)
      XABAR=XABAR*2.0/(N*1.0)
```

```
      B=N*C2*(XBAR-XABAR)

      C=0.0
      DO 31 I=1,N
        C=C+BET(I)*(X(I)-MLMU)**2
 31   CONTINUE
      C=2.0*C2*C

      MLSIGMA=(-B+SQRT(B**2+4.0*N*C))/(2.0*SQRT(1.0*N*(N-1)))
      MLSIGMAMEAN=MLSIGMAMEAN+MLSIGMA
      MLSIGMAVAR=MLSIGMAVAR+MLSIGMA**2

C FINDING LSE
      LSMU=XBAR
      LSMUMEAN=LSMUMEAN+LSMU
      LSMUVAR=LSMUVAR+LSMU**2

      LSSIGMA=0.0
      DO 35 I=1,N
        LSSIGMA=LSSIGMA+(X(I)-LSMU)**2
 35   CONTINUE
      LSSIGMA=LSSIGMA/(N*1.0-1.0)
      LSSIGMA=SQRT(LSSIGMA)
      LSSIGMAMEAN=LSSIGMAMEAN+LSSIGMA
      LSSIGMAVAR=LSSIGMAVAR+LSSIGMA**2

 100  CONTINUE

      MLMUMEAN=MLMUMEAN/(NN*1.0)
      MLMUVAR=MLMUVAR/(NN*1.0)-MLMUMEAN**2
      MLSIGMAMEAN=MLSIGMAMEAN/(NN*1.0)
      MLSIGMAVAR=MLSIGMAVAR/(1.0*NN)-MLSIGMAMEAN**2

      WRITE(1,*) '  '
      WRITE(1,*) 'MMLE OF M=',MLMUMEAN
      WRITE(1,*) 'MMLE OF SIGMA=',MLSIGMAMEAN
      WRITE(1,*) 'VAR OF MMLE OF M=',MLMUVAR
      WRITE(1,*) 'VAR OF MMLE OF SIGMA=',MLSIGMAVAR

      LSMUMEAN=LSMUMEAN/(NN*1.0)
      LSMUVAR=LSMUVAR/(NN*1.0)-LSMUMEAN**2
      LSSIGMAMEAN=LSSIGMAMEAN/(NN*1.0)
      LSSIGMAVAR=LSSIGMAVAR/(1.0*NN)-LSSIGMAMEAN**2

      WRITE(1,*) '  '
      WRITE(1,*) 'LSE OF M=',LSMUMEAN
```

```
      WRITE(1,*) 'LSE OF SIGMA=',LSSIGMAMEAN
      WRITE(1,*) 'VAR OF LSE OF M=',LSMUVAR
      WRITE(1,*) 'VAR OF LSE OF SIGMA=',LSSIGMAVAR

      MLMUMSE=MLMUVAR+MLMUMEAN**2
      LSMUMSE=LSMUVAR+LSMUMEAN**2

      WRITE(1,*) '  '
      WRITE(1,*) 'MSE OF MMLE OF M=',MLMUMSE
      WRITE(1,*) 'MSE OF LSE OF M=',LSMUMSE

      MLSIGMAMSE=(SIGMA-MLSIGMAMEAN)**2+MLSIGMAVAR
      LSSIGMAMSE=(SIGMA-LSSIGMAMEAN)**2+LSSIGMAVAR

      WRITE(1,*) 'MSE OF MMLE OF SIGMA=',MLSIGMAMSE
      WRITE(1,*) 'MSE OF LSE OF SIGMA=',LSSIGMAMSE

      MLDEF=MLMUMSE+MLSIGMAMSE
      LSDEF=LSMUMSE+LSSIGMAMSE

      WRITE(1,*) '  '
      WRITE(1,*) 'DEF OF MMLE=',MLDEF
      WRITE(1,*) 'DEF OF LSE=',LSDEF

      RELSMU=(MLMUVAR/LSMUVAR)*100.0
      RELSSIGMA=(MLSIGMAMSE/LSSIGMAMSE)*100.0

      WRITE(1,*) '  '
      WRITE(1,*) 'RE OF LSE OF M=',RELSMU
      WRITE(1,*) 'RE OF LSE OF SIGMA=',RELSSIGMA
      IF (T.GT.(-1.0*PI).AND.T.LT.0.0) THEN
      MVBMU=(2.0*(SIGMA**2)*T*SIN(T)**2)/(N*(C2**2)*(T-SIN(T)*COS(T)))
      I22=(PI**2-T**2)/(SIN(T)**2)
      I22=I22-((PI**2-3.0*T**2)*COS(T)/(T*SIN(T)))
      I22=-N*I22/(6.0*SIGMA**2)
      ELSE IF (T.GT.0.0) THEN
      MVBMU=(2.0*SIGMA**2*T*SINH(T)**2)/(N*(C2**2)*(SINH(T)*COSH(T)
-T))
      I22=(PI**2+3.0*T**2)*COSH(T)/(T*SINH(T))
      I22=I22-((PI**2+T**2)/(SINH(T)**2))
      I22=-N*I22/(6.0*SIGMA**2)
      ELSE IF (T.EQ.0.0) THEN
      MVBMU=(3.0*SIGMA**2)/(N*(C2**2))
      I22=-N*(PI**2+3.0)/(9.0*SIGMA**2)
      ENDIF
      MVBSIGMA=-1.0/I22
```

```
      WRITE(1,*)' '
      WRITE(1,*)'MVB OF M=',MVBMU
      WRITE(1,*)'MVB OF SIGMA=',MVBSIGMA

      MLMUE=(MVBMU/MLMUVAR)*100.0
      MLSIGMAE=(MVBSIGMA/MLSIGMAVAR)*100.0
      WRITE(1,*)' '
      WRITE(1,*)'EFF OF M=',MLMUE
      WRITE(1,*)'EFF OF SIGMA=',MLSIGMAE
      END
```

B.2 Simulation for GL Distribution

```
      PROGRAM GL
      USE NUMERICAL_LIBRARIES
      REAL B,Y(100),G(100),SIGMA,MLMU0,MLMU,MLSIGMA,T(100)
      REAL BET(100),ALFA(100),DEL(100),DELTA,M,MLMUMEAN
      REAL MLSIGMAMEAN,MLMUVAR,MLSIGMAVAR,X(100),XBAR,C
      REAL BB,LSMU,LSSIGMA,LSMUMEAN, LSMUVAR,LSSIGMAMEAN
      REAL LSSIGMAVAR,LSSIGMAMSE,MLSIGMAMSE
      REAL RELSSIGMA,RELSMU,LSMUMSE,MLMUMSE

      OPEN (unit=1,file='c:\Concon\output.txt')

      B=0.5

      PRINT*,'ENTER THE SAMPLE SIZE'
      READ*,N

      SIGMA=1.0

      DO 10 I=1,N
        V2=I
        V1=V2/(N*1.0+1.0)
        T(I)=-LOG(V1**(-1.0/B)-1.0)
10    CONTINUE

      DO 20 I=1,N
        BET(I)=EXP(T(I))/(EXP(T(I))+1.0)**2
        ALFA(I)=(T(I)*EXP(T(I))+EXP(T(I))+1.0)
        ALFA(I)=ALFA(I)/(EXP(T(I))+1.0)**2
20    CONTINUE

      DELTA=0.0
```

```
      DO 22 I=1,N
       DEL(I)=1.0/(B+1.0)-ALFA(I)
       DELTA=DELTA+DEL(I)
22    CONTINUE

      M=0.0
      DO 24 I=1,N
       M=M+BET(I)
24    CONTINUE

      MLMUMEAN=0.0
      MLSIGMAMEAN=0.0
      MLMUVAR=0.0
      MLSIGMAVAR=0.0
      LSMUMEAN=0.0
      LSSIGMAMEAN=0.0
      LSMUVAR=0.0
      LSSIGMAVAR=0.0

      NN=100000/N

      DO 100 L=1,NN
      CALL RNUN(N,G)
      DO 5 I=1,N
       Y(I)=-LOG(G(I)**(-1.0/B)-1.0)
 5    CONTINUE

C FINDING MMLE

      CALL SVRGN(N,Y,X)

      MLMU0=0.0
      DO 25 I=1,N
       MLMU0=MLMU0+BET(I)*X(I)
25    CONTINUE
      MLMU0=MLMU0/M

      BB=0.0
      DO 30 I=1,N
       BB=BB+DEL(I)*(X(I)-MLMU0)
30    CONTINUE
      BB=(B+1.0)*BB

      C=0.0
      DO 31 I=1,N
       C=C+BET(I)*(X(I)-MLMU0)**2
```

```
 31    CONTINUE
       C=(B+1.0)*C

       MLSIGMA=(BB+SQRT(BB**2+4.0*N*C))/(2.0*SQRT(1.0*N*(N-1.0)))
       MLSIGMAMEAN=MLSIGMAMEAN+MLSIGMA
       MLSIGMAVAR=MLSIGMAVAR+MLSIGMA**2

       MLMU=MLMU0+DELTA*MLSIGMA/M
       MLMUMEAN=MLMUMEAN+MLMU
       MLMUVAR=MLMUVAR+MLMU**2

C FINDING LSE

       XBAR=0.0
       DO 32 I=1,N
        XBAR=XBAR+X(I)
 32    CONTINUE
       XBAR=XBAR/(N*1.0)

       LSSIGMA=0.0
       DO 35 I=1,N
        LSSIGMA=LSSIGMA+(X(I)-XBAR)**2
 35    CONTINUE
       LSSIGMA=(LSSIGMA/(N*1.0-1.0))/(4.9348+1.6449)
       LSSIGMA=SQRT(LSSIGMA)
       LSSIGMAMEAN=LSSIGMAMEAN+LSSIGMA
       LSSIGMAVAR=LSSIGMAVAR+LSSIGMA**2

       LSMU=XBAR-(-1.9635+0.5772)*LSSIGMA
       LSMUMEAN=LSMUMEAN+LSMU
       LSMUVAR=LSMUVAR+LSMU**2

100    CONTINUE

       MLMUMEAN=MLMUMEAN/(NN*1.0)
       MLMUVAR=MLMUVAR/(NN*1.0)-MLMUMEAN**2
       MLSIGMAMEAN=MLSIGMAMEAN/(NN*1.0)
       MLSIGMAVAR=MLSIGMAVAR/(NN*1.0)-MLSIGMAMEAN**2

       WRITE(1,*) '  '
       WRITE(1,*) 'MMLE OF M=',MLMUMEAN
       WRITE(1,*) 'MMLE OF SIGMA=',MLSIGMAMEAN
       WRITE(1,*) 'VAR OF MMLE OF M=',MLMUVAR
       WRITE(1,*) 'VAR OF MMLE OF SIGMA=',MLSIGMAVAR

       LSMUMEAN=LSMUMEAN/(NN*1.0)
```

```
LSMUVAR=LSMUVAR/(NN*1.0)-LSMUMEAN**2
LSSIGMAMEAN=LSSIGMAMEAN/(NN*1.0)
LSSIGMAVAR=LSSIGMAVAR/(NN*1.0)-LSSIGMAMEAN**2


WRITE(1,*) '  '
WRITE(1,*) 'LSE OF M=',LSMUMEAN
WRITE(1,*) 'LSE OF SIGMA=',LSSIGMAMEAN
WRITE(1,*) 'VAR OF LSE OF M=',LSMUVAR
WRITE(1,*) 'VAR OF LSE OF SIGMA=',LSSIGMAVAR


MLMUMSE=MLMUVAR+MLMUMEAN**2
LSMUMSE=LSMUVAR+LSMUMEAN**2

WRITE(1,*) '  '
WRITE(1,*) 'MSE OF MMLE OF M=',MLMUMSE
WRITE(1,*) 'MSE OF LSE OF M=',LSMUMSE

MLSIGMAMSE=(SIGMA-MLSIGMAMEAN)**2+MLSIGMAVAR
LSSIGMAMSE=(SIGMA-LSSIGMAMEAN)**2+LSSIGMAVAR

WRITE(1,*) 'MSE OF MMLE OF SIGMA=',MLSIGMAMSE
WRITE(1,*) 'MSE OF LSE OF SIGMA=',LSSIGMAMSE

RELSMU=(MLMUMSE/LSMUMSE)*100.0
RELSSIGMA=(MLSIGMAMSE/LSSIGMAMSE)*100.0

WRITE(1,*) '  '
WRITE(1,*) 'RE OF LSE OF M=',RELSMU
WRITE(1,*) 'RE OF LSE OF SIGMA=',RELSSIGMA
END
```

APPENDIX C

IRISES AND Q-Q PLOTS

C.1 Irises

Fisher (1936) gives the data, collected by E. Anderson, in Table C.1.1.

Table C.1.1 Irises data

Iris setosa				Iris versicolor				Iris virginica			
Sepal length	Sepal width	Petal length	Petal width	Sepal length	Sepal width	Petal length	Petal width	Sepal length	Sepal Width	Petal length	Petal width
5.1	3.5	1.4	0.2	7.0	3.2	4.7	1.4	6.3	3.3	6.0	2.5
4.9	3.0	1.4	0.2	6.4	3.2	4.5	1.5	5.8	2.7	5.1	1.9
4.7	3.2	1.3	0.2	6.9	3.1	4.9	1.5	7.1	3.0	5.9	2.1
4.6	3.1	1.5	0.2	5.5	2.3	4.0	1.3	6.3	2.9	5.6	1.8
5.0	3.6	1.4	0.2	6.5	2.8	4.6	1.5	6.5	3.0	5.8	2.2
5.4	3.9	1.7	0.4	5.7	2.8	4.5	1.3	7.6	3.0	6.6	2.1
4.6	3.4	1.4	0.3	6.3	3.3	4.7	1.6	4.9	2.5	4.5	1.7
5.0	3.4	1.5	0.2	4.9	2.4	3.3	1.0	7.3	2.9	6.3	1.8
4.4	2.9	1.4	0.2	6.6	2.9	4.6	1.3	6.7	2.5	5.8	1.8
4.9	3.1	1.5	0.1	5.2	2.7	3.9	1.4	7.2	3.6	6.1	2.5
5.4	3.7	1.5	0.2	5.0	2.0	3.5	1.0	6.5	3.2	5.1	2.0
4.8	3.4	1.6	0.2	5.9	3.0	4.2	1.5	6.4	2.7	5.3	1.9
4.8	3.0	1.4	0.1	6.0	2.2	4.0	1.0	6.8	3.0	5.5	2.1
4.3	3.0	1.1	0.1	6.1	2.9	4.7	1.4	5.7	2.5	5.0	2.0
5.8	4.0	1.2	0.2	5.6	2.9	3.6	1.3	5.8	2.8	5.1	2.4
5.7	4.4	1.5	0.4	6.7	3.1	4.4	1.4	6.4	3.2	5.3	2.3
5.4	3.9	1.3	0.4	5.6	3.0	4.5	1.5	6.5	3.0	5.5	1.8
5.1	3.5	1.4	0.3	5.8	2.7	4.1	1.0	7.7	3.8	6.7	2.2
5.7	3.8	1.7	0.3	6.2	2.2	4.5	1.5	7.7	2.6	6.9	2.3
5.1	3.8	1.5	0.3	5.6	2.5	3.9	1.1	6.0	2.2	5.0	1.5
5.4	3.4	1.7	0.2	5.9	3.2	4.8	1.8	6.9	3.2	5.7	2.3
5.1	3.7	1.5	0.4	6.1	2.8	4.0	1.3	5.6	2.8	4.9	2.0
4.6	3.6	1.0	0.2	6.3	2.5	4.9	1.5	7.7	2.8	6.7	2.0
5.1	3.3	1.7	0.5	6.1	2.8	4.7	1.2	6.3	2.7	4.9	1.8
4.8	3.4	1.9	0.2	6.4	2.9	4.3	1.3	6.7	3.3	5.7	2.1
5.0	3.0	1.6	0.2	6.6	3.0	4.4	1.4	7.2	3.2	6.0	1.8
5.0	3.4	1.6	0.4	6.8	2.8	4.8	1.4	6.2	2.8	4.8	1.8
5.2	3.5	1.5	0.2	6.7	3.0	5.0	1.7	6.1	3.0	4.9	1.8
5.2	3.4	1.4	0.2	6.0	2.9	4.5	1.5	6.4	2.8	5.6	2.1
4.7	3.2	1.6	0.2	5.7	2.6	3.5	1.0	7.2	3.0	5.8	1.6

4.8	3.1	1.6	0.2	5.5	2.4	3.8	1.1	7.4	2.8	6.1	1.9
5.4	3.4	1.5	0.4	5.5	2.4	3.7	1.0	7.9	3.8	6.4	2.0
5.2	4.1	1.5	0.1	5.8	2.7	3.9	1.2	6.4	2.8	5.6	2.2
5.5	4.2	1.4	0.2	6.0	2.7	5.1	1.6	6.3	2.8	5.1	1.5
4.9	3.1	1.5	0.2	5.4	3.0	4.5	1.5	6.1	2.6	5.6	1.4
5.0	3.2	1.2	0.2	6.0	3.4	4.5	1.6	7.7	3.0	6.1	2.3
5.5	3.5	1.3	0.2	6.7	3.1	4.7	1.5	6.3	3.4	5.6	2.4
4.9	3.6	1.4	0.1	6.3	2.3	4.4	1.3	6.4	3.1	5.5	1.8
4.4	3.0	1.3	0.2	5.6	3.0	4.1	1.3	6.0	3.0	4.8	1.8
5.1	3.4	1.5	0.2	5.5	2.5	4.0	1.3	6.9	3.1	5.4	2.1
5.0	3.5	1.3	0.3	5.5	2.6	4.4	1.2	6.7	3.1	5.6	2.4
4.5	2.3	1.3	0.3	6.1	3.0	4.6	1.4	6.9	3.1	5.1	2.3
4.4	3.2	1.3	0.2	5.8	2.6	4.0	1.2	5.8	2.7	5.1	1.9
5.0	3.5	1.6	0.6	5.0	2.3	3.3	1.0	6.8	3.2	5.9	2.3
5.1	3.8	1.9	0.4	5.6	2.7	4.2	1.3	6.7	3.3	5.7	2.5
4.8	3.0	1.4	0.3	5.7	3.0	4.2	1.2	6.7	3.0	5.2	2.3
5.1	3.8	1.6	0.2	5.7	2.9	4.2	1.3	6.3	2.5	5.0	1.9
4.6	3.2	1.4	0.2	6.2	2.9	4.3	1.3	6.5	3.0	5.2	2.0
5.3	3.7	1.5	0.2	5.1	2.5	3.0	1.1	6.2	3.4	5.4	2.3
5.0	3.3	1.4	0.2	5.7	2.8	4.1	1.3	5.9	3.0	5.1	1.8

C.2 Q-Q Plots

Normal Q-Q plots of the training set for the i^{th} (i = 1, 2, 3, 4) feature of the j^{th} (j = 1, 2, 3) class are as follows:

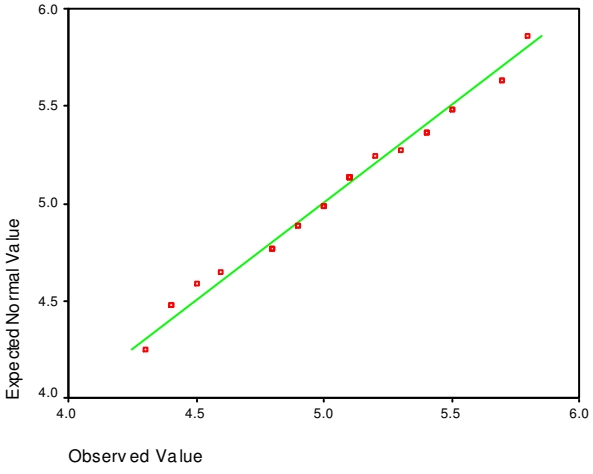

Figure C.1 Normal Q-Q plot of the training set for the feature sepal length of the class iris setosa

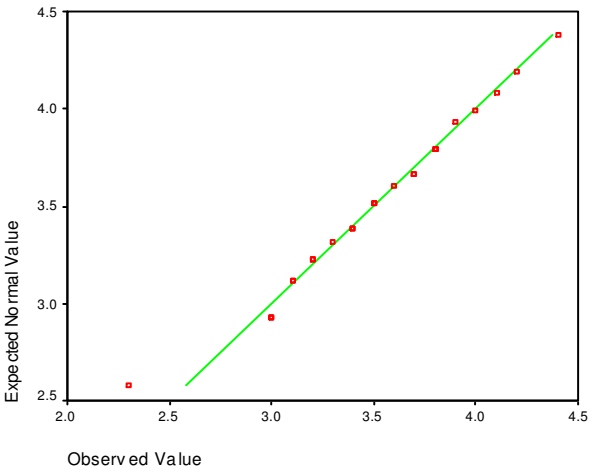

Figure C.2 Normal Q-Q plot of the training set for the feature sepal width of the class iris setosa

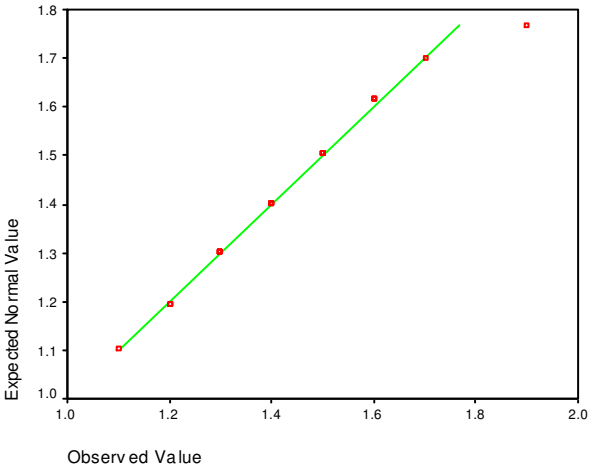

Figure C.3 Normal Q-Q plot of the training set for the feature petal length of the class iris setosa

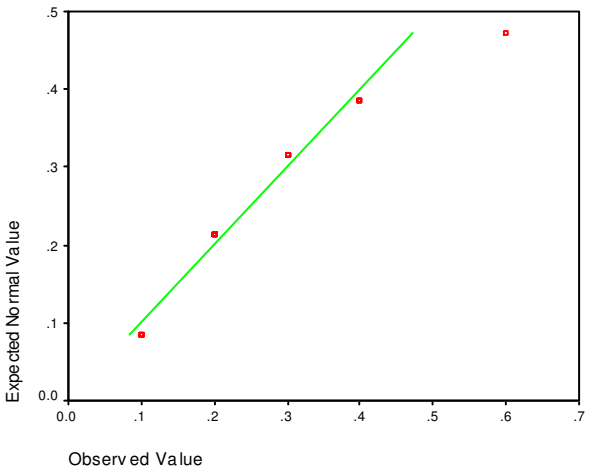

Figure C.4 Normal Q-Q plot of the training set for the feature petal width of the class iris setosa

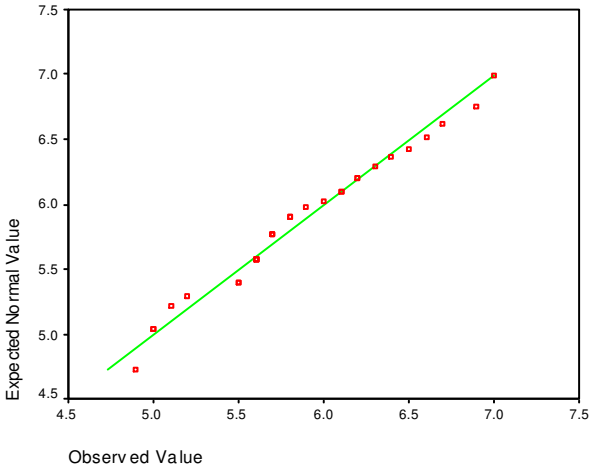

Figure C.5 Normal Q-Q plot of the training set for the feature sepal length of the class iris versicolor

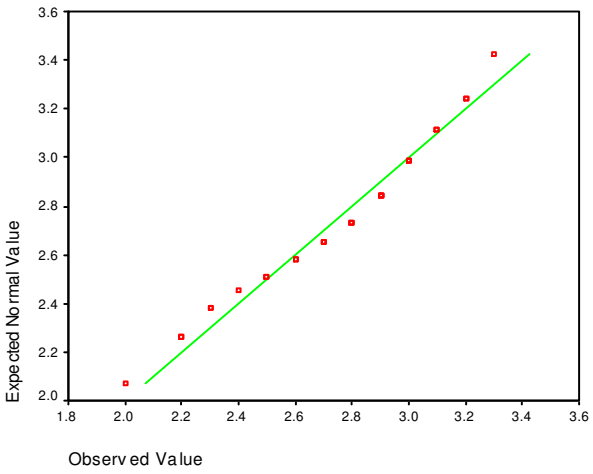

Figure C.6 Normal Q-Q plot of the training set for the feature sepal width of the class iris versicolor

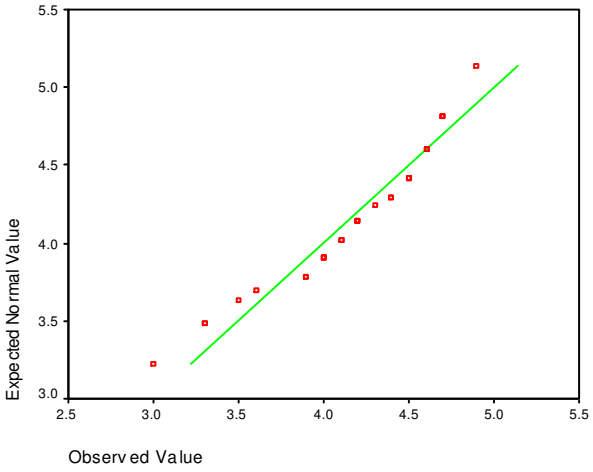

Figure C.7 Normal Q-Q plot of the training set for the feature petal length of the class iris versicolor

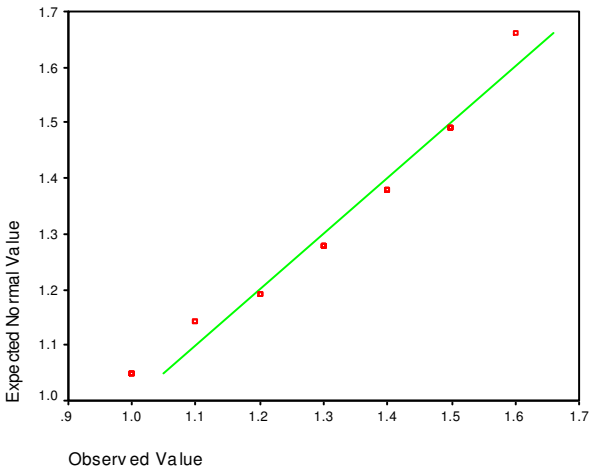

Figure C.8 Normal Q-Q plot of the training set for the feature petal width of the class iris versicolor

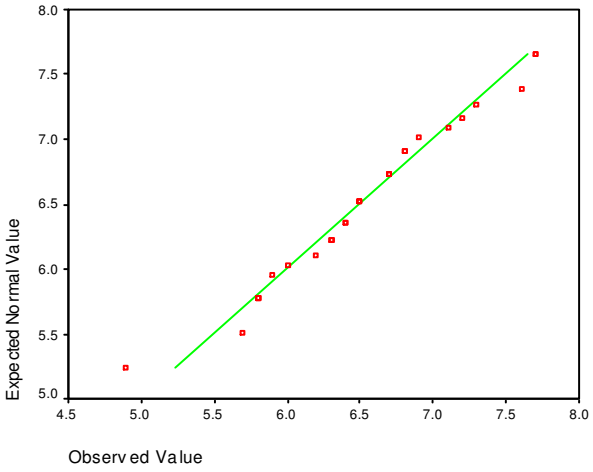

Figure C.9 Normal Q-Q plot of the training set for the feature sepal length of the class iris virginica

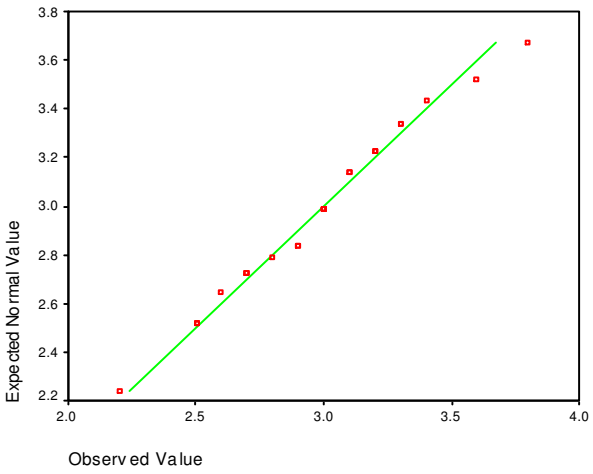

Figure C.10 Normal Q-Q plot of the training set for the feature sepal width of the class iris virginica

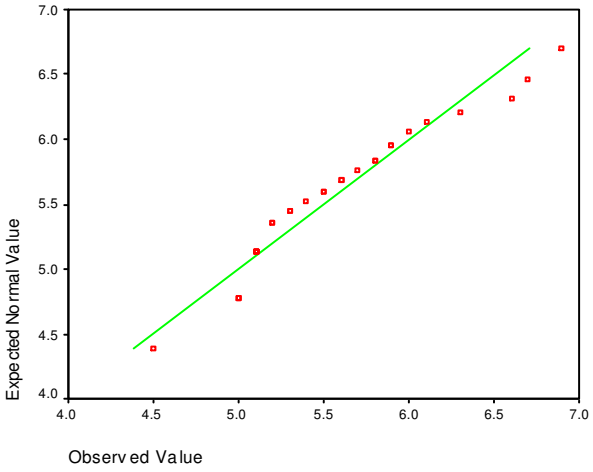

Figure C.11 Normal Q-Q plot of the training set for the feature petal length of the class iris virginica

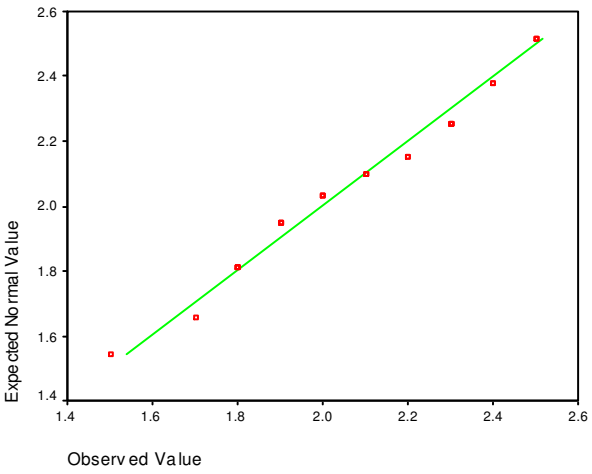

Figure C.12 Normal Q-Q plot of the training set for the feature petal width of the class iris virginica