NON-GAUSSIAN MODELS FOR CONTEXT AWARE ANOMALY DETECTION

by

Gregory Randall Lakomski, B.S. M.B.A.

A thesis submitted to the Graduate Council of
Texas State University in partial fulfillment
of the requirements for the degree of
Master of Science
with a Major in Computer Science
May 2017

Committee Members:

Dan Tamir, Chair

Mina Guirguis

Tahir Ekin

FAIR USE AND AUTHOR'S PERMISSION STATEMENT

Fair Use

Duplication Permission

ACKNOWLEDGEMENTS

This thesis would not have been possible without my advisor Dr. Dan Tamir. He has been instrumental in giving guidance and direction for the work done in this thesis and has suffered through many semesters of my slow progress. I would like to especially thank him for his patience.

In addition, I would like to thank Dr. Mina Guirguis for his inspiring instruction and Dr. Tahir Ekin for his patient support and explanation of statistical concepts.

Lastly, I want to thank my wife Donna LaKomski for her patience and support. I cherish her love and encouragement.

TABLE OF CONTENTS

Page

LIST OF TABLES

Table **Page**

LIST OF FIGURES

Figure **Page**

1. INTRODUCTION

Problem Definition

The availability of large multivariate mixed data sets has created significant new opportunities to uncover previously hidden insights. Of particular interest are data instances or patterns with characteristics that set them apart from the body of data in some way. Called outliers if based solely on a specific measure of separation, or anomalies if determined to be strange in a quantitative or comparative sense, they are often averaged out or discarded due to the analytical complexities of addressing them. Traditional analytic processes, based on normative Gaussian statistics, are clearly useful in many cases. At the same time, these processes often fail to recognize the complexities of real data sets and inappropriately apply simplifying assumptions. Of special interest is the ability to deeply understand complex data that affect the welfare of the general public. In particular, identification, mitigation, and management of low probability, unexpected events (aka "Black Swans" [1]) such as disasters and security threats represent an important end goal.

The problem addressed in this thesis is that widely used methodologies for the detection of anomalies frequently ignore the context of the data [2]. Those methodologies that take context into account often treat the data as normally distributed and do not attempt to deal with complex probability distributions that would better represent the data and identify anomalies that are truly interesting. In particular, many highly important systems such as the stock market, weather, and economic phenomena exhibit power law probability density functions such as the exponential-logarithmic.

By overlooking the context of data, anomaly detection is less selective and points that "in context" would not be identified as anomalies become false positives and waste resources, such as time, that are needed to evaluate them. Furthermore, many interesting and important data sets are multivariate of order greater than two and are distributional mixtures. Due to the complexities of representing multivariate data, visual analytic techniques are of limited value in the evaluation of outliers in these real world data sets. One option is to model complex density functions as a mixture of several Gaussians. Nevertheless, this significantly increases the computational complexity of this approach.

The proposed solution is to extend an existing, novel, Gaussian density function based, context sensitive, anomaly detection scheme called Conditional Anomaly Detection Probability Density Function (FCAD) developed by Song to use non-Gaussian probability distributions to decrease computational complexity and improve detection [3]. This thesis achieved this modification by applying the skew-normal density function to the Gaussian Mixture Model - Conditional Anomaly Detection – Split (GMM-CAD-Split) version of the FCAD anomaly detection algorithm in order to improve the ability to detect meaningful anomalies in non-normal data sets while including normal data sets as well.

The hypothesis of this thesis is that current methodologies used for the context sensitive identification of anomalies can be extended to include the use of non-Gaussian probability distributions while including the effects of context, leading to improved identification of significant, interesting, low probability anomalies. These new distributions can be used to deal with both Gaussian and non-Gaussian data sets.

The contribution of this thesis is that it demonstrates non-Gaussian distributions can be successfully integrated into the FCAD algorithm and that it establishes that the use of an alternative probability distribution in association with FCAD can result in improvement of anomaly identification in non-normal data sets. To the best of our knowledge this is the first use of non-Gaussian distributions used with FCAD to identify anomalies in non-Gaussian and Gaussian data sets.

The remainder of this thesis is organized as follows. Section 2 covers relevant background where skew-normal and FCAD are described in detail. Section 3 describes pertinent related work. Section 4 presents the research methodology including solution and experimental setup. Section 5 presents the results of the experiments. Section 6 contains the global evaluation of these results, and finally Section 7 includes conclusions and proposals for further work.

2. BACKGROUND

This section covers several relevant definitions and concepts. In particular, the skew-normal distribution is introduced and defined and Expectation Maximization reviewed.

Variable Types

The bulk of research into anomaly detection has focused on data sets containing continuous variables. Continuous variables are referred to as analog variables or quantitative variables and are defined as being able to take on any real number value between its minimum value and its maximum value. Categorical variables are discrete variables that have two or more categories, and are differentiated by the presence of an intrinsic order. Nominal categorical variables have no intrinsic order. For example, real estate agents could classify their types of property into distinct categories such as houses, condos, co-ops or bungalows. So, "type of property" is a nominal variable with 4 categories called houses, condos, co-ops and bungalows. Additionally, the different categories of a nominal variable can be referred to as groups or the levels of the nominal variable. Ordinal categorical variables have a clear ordering. An example is economic status with values low, medium, and high. A random variable takes on a specific value based on the associated probability distribution. There are two kinds of random variables, discrete and continuous. A discrete random variable is associated with a mass function and a continuous random variable is associated with a density function.

Anomaly

Although the term anomaly is frequently used interchangeably with the term outlier they are conceptually different. An anomaly is a data point that has a low probability of occurring either in value or in relationship to other data points. An outlier is a data point that is, by some metric, separated from other observations. One implication is that a data point that would not be considered an outlier based on lack of separation from other data points could be an anomaly if lack of separation was not expected. Furthermore, a **context** sensitive anomaly is a data point that is probabilistically out of place only in the context in which it is occurring. An example would be a day with the temperature over 100 degrees is an anomaly if this is occurring in January in Alaska but not if it occurs in the Sahara Desert in July. This definition implies that there is no separation requirement and that discrete variables can be anomalies just as well.

Errors of Type 1 and Type 2

An error of type one occurs when the hypothesis being tested is really true and that it is concluded that it is false. An error of type 2 occurs when the hypothesis being tested is false and it is concluded that it is true.

Probability Density Function (PDF) [4]

A Probability Density Function describes the relative likelihood for a continuous random variable to take on a given value. The probability of the random variable falling

within a particular range of values is given by the integral of this variable's density over that range—that is, it is given by the area under the density function. An example of a PDF of a random variable x, where x $\epsilon\mathcal{R}$, is the normal distribution (Gaussian), which is parameterized in terms of the mean μ and varianceσ:

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \quad (1)$$

The Exponential Family [4]

The exponential family of probability distributions shares the PDF form:

$$f(x|\theta)=h(x)g(\theta)\exp\big(\eta(\theta)\cdot T(x)\big) \quad (2)$$

where h, g, T and η are all known functions of x ϵ $|R|$ and are the parameters of a probability density function.

A large number of common likelihood distributions are members of this family. Well known examples are normal, exponential-logarithmic, Gamma, Beta, Dirichlet, Bernoulli, and Poisson. The skew-normal family is an extension of the exponential family and will be covered in the next section.

The Skew-Normal Distribution [7]

The skew-normal distribution pioneered by is a powerful and flexible distribution that is being used in a rapidly growing body of work [5, 6]. The skew normal extends the normal distribution through the addition of a parameter λ that defines skewness or "heavy

tailness". The actual parameters of the normal distribution, mean μ and variance Σ are used. The effect of these parameters can be observed in figure 1 taken from the Boost toolkit description[1]. It can be observed that where λ is equal to zero, the distribution becomes the standard normal distribution. An increasing λ increases the skew in the distribution.

The skew-normal function has the PDF:

$$\mathrm{SN}(y|\mu,\Sigma,\lambda) = 2\phi(y|\mu,\Sigma)\Phi\left(\lambda^{T}\Sigma^{-1/2}(y-\mu)\right) \tag{4}$$

where ϕ stands for the density of the p-variate normal distribution and Φ stands for the distribution of the standard univariate normal distribution. It can be observed that when $\lambda = 0$, the result is a normal distribution.

[1]http://www.boost.org/doc/libs/1_50_0/libs/math/doc/sf_and_dist/html/math_toolkit/dist/dist_ref/dists/skew_normal_dist.html

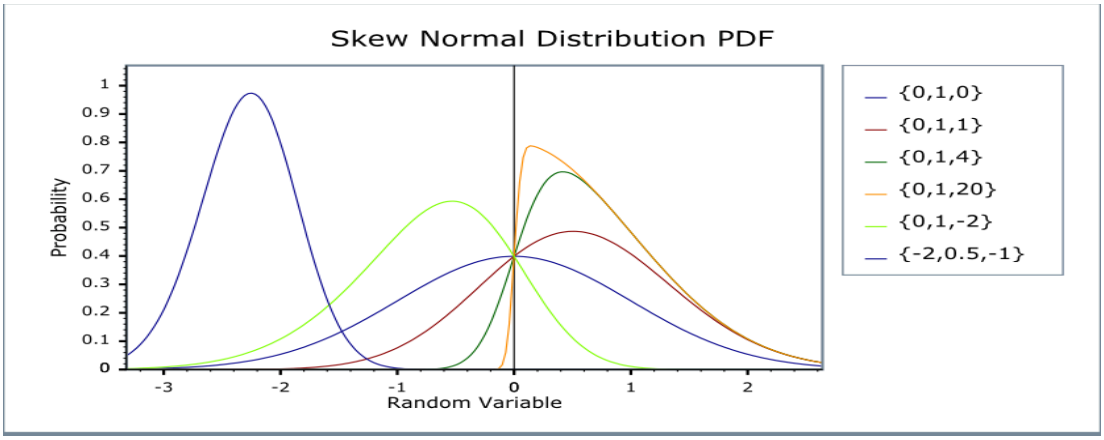

Figure 1 – Plots of skew-normal distribution

Maximum Likelihood Estimation (MLE)

Maximum Likelihood Estimation is a method of estimating the parameters of a statistical distribution or model. The likelihood function for the distribution is "maximized' by the parameter specific function derivative to zero and solving for the parameter of interest. Intuitively, this maximizes the "agreement" of the selected model with the observed data.

Based on [8], starting with a statistical model consisting of a set, X, of observed data, a set of latent data Z (hidden variables), and a set of unknown parameters θ, along with a likelihood function, the Maximum Likelihood Estimation (MLE) of the unknown parameters is determined by what's called the marginal likelihood of the observed data:

$$\mathrm{L}(\theta; \mathrm{X}, \mathrm{Z}) = \mathrm{p}(\mathrm{X}|\theta) = \sum_{\mathrm{Z}} \mathrm{p}(\mathrm{X}, \mathrm{Z}|\theta) \tag{5}$$

Maximizing the actual likelihood function for a particular distribution is often mathematically extremely difficult. The log-likelihood is often more convenient to use. Because the natural logarithm is a monotonically increasing function, the log of a function will have the same maximum as the original function. Taking the derivatives of the likelihood function frequently results in taking derivatives of products of terms and the log-likelihood allows taking derivatives of sums of terms instead. An example is the Gamma function [4]:

$$L(\alpha, \beta|x) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} \tag{6}$$

Taking the log makes differentiation possible. Where Γ is the Gamma function:

$$\log L(\alpha, \beta|x) = \alpha \log\beta - \log\Gamma(\alpha) + (\alpha - 1)\log x - \beta x \tag{7}$$

Expectation Maximization [10]

Expectation Maximization (EM) is an algorithm that can be used to estimate the parameters in a MLE when they cannot be calculated directly. The EM algorithm was explained and given its name in a classic 1977 paper by Dempster et al. [9]. EM introduced the concept of "hidden variables" (Z) that allows the model to be formulated in a simple way. For example, a mixture model can be described more simply by assuming that each observed data point has a corresponding unobserved data point, or latent variable, specifying the mixture component that each data point belongs to. The algorithm assumes a starting value for the parameters and in an iterative fashion refines

the parameters based on the quality of fit of the actual data to the implied probability distributions.

The algorithm seeks to find the MLE of the marginal likelihood by iteratively applying the following two steps:

Expectation Step: Calculate the expected value of the log likelihood function with respect to the conditional distribution of Z given X under the current estimate of the parameter $\theta^{(t)}$:

$$\mathrm{Q}\left(\theta|\theta^{(t)}\right) = \text{Expected Value}[\log \mathrm{L}(\theta; \mathrm{X}, \mathrm{Z})] \tag{8}$$

Maximization Step: Find the parameters that maximize this quantity using Maximum Likelihood Estimation:

$$\theta^{(t+1)} = \mathrm{argmax}\mathrm{Q}\left(\theta|\theta^{(t)}\right) \tag{9}$$

Mixture Models

If a data set can be described as a combination of different individual distributions, for example two normal distributions, this is called a mixture model. Mixture models have been widely applied in many applications. Given enough mixture components, they can approximate many complicated probability densities and accommodate skewness and heavy tails.

Gaussian Univariate Mixture Model

It would be easy to explicitly state a Gaussian univariate mixture model if the membership of every point was known. Since this is almost never the case, methods such as EM were developed to facilitate membership identification and the calculation of model parameters [11]. Starting with a probability function represented as a weighted sum of M Gaussian (normal) component densities:

For a data set $\{x_1 : x_k\}$

$$\mathrm{p}(\mathrm{x_i}|\mathrm{M}) = \sum_{\mathrm{k}=1}^{\mathrm{K}} \pi_\mathrm{k} \mathrm{p}(\mathrm{x_i}|\theta_\mathrm{k}) \tag{10}$$

where M is the set of mixtures and there are K mixtures with mixing weight π_k, mean μ_k, and standard deviation σ_k.

To apply EM to the univariate Gaussian Mixture model the log-likelihood function is

$$E_z[\log p(x|z)] = \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma(z_{nk}) \left(\log \pi_k + \log\left(p(x|\mu_k, {\sigma_k}^2)\right)\right) \tag{11}$$

E-Step: Maximize the log likelihood function to get the E-Step and M-Step functions.

Compute the expected values of the latent variable using the current parameter set.

$$\gamma(z_{nk}) = \frac{\pi_k^{old} p\left(x_n|\mu_k^{old}, \sigma_k^{old^2}\right)}{\sum_{j=1}^{K} \pi_j^{old} p\left(x_n|\mu_k^{old}, \left(\sigma_k^{old^2}\right)\right)} \tag{12}$$

M Step: Update $\pi_k^{old}, \mu_k^{old}, \sigma_k^{old}$

$$\pi_k^{new} = \frac{\sum_{n=1}^{N} \gamma(z_{nk})}{N} \tag{13}$$

$$\mu_k^{new} = \frac{\sum_{n=1}^{N} \gamma(z_{nk}) x_k}{\sum_{n=1}^{N} \gamma(z_{nk})} \tag{14}$$

$$\sigma_k^{new\,2} = \frac{\sum_{n=1}^{N} \gamma(z_{nk})(x_n - \mu_k^{new})^2}{\sum_{n=1}^{N} \gamma(z_{nk})} \tag{15}$$

Gaussian Multivariate Mixture Model

The multivariate Gaussian mixture model is similar to the univariate mixture model. The PDF for multivariate mixtures is of the form:

$$p_j(\boldsymbol{x}|\phi) = \frac{1}{(2\pi)^{d/2}(det\Sigma_i)} e^{-1/2(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i)} \tag{16}$$

where p_i is the density for an individual normal mixture. μ_i is the mean vector for the μ_i mixture, Σ_i is the d x d symmetric covariance matrix for the i^{th}mixture. The covariance matrix defines the relationship between the individual mixture components and indicates how a change in one variable affects the other variables.

E-Step: Maximize the log likelihood function to get the E-Step and M-Step functions. Compute the expected values of the latent variable using the current parameter set [11].

$$\gamma(z_{nk}) = \frac{\pi_k^{old}\, p\left(x_n|\mu_k^{old}, \sigma_k^{old\,2}\right)}{\sum_{j=1}^{K} \pi_j^{old}\, p\left(x_n|\mu_j^{old}, \sigma_j^{old\,2}\right)} \tag{17}$$

M Step: Update π_k^{old}, μ_k^{old}, $\sigma_k^{old^2}$

$$\pi_k^{new} = \frac{\sum_{n=1}^{N} \gamma(z_{nk})}{N} \tag{18}$$

$$\mu_k^{new} = \frac{\sum_{n=1}^{N} \gamma(z_{nk}) x_k}{\sum_{n=1}^{N} \gamma(z_{nk})} \tag{19}$$

$$\Sigma_k^{new^2} = \frac{\sum_{n=1}^{N} \gamma(z_{nk})(x_n - \mu_k^{new})(x_n - \mu_k^{new})^T}{\sum_{n=1}^{N} \gamma(z_{nk})} \tag{20}$$

Multivariate Skew-Normal Mixture Model

As cataloged in [12], the original form of the skew-normal distribution (FUSN) has many extensions. The extension used for this thesis is the class SMSN (Scale Mixture Skew Normal) described in [13]. This class of flexible distributions can accommodate skewness and discrepant observations. A family of scale mixtures of the skew-normal distribution that differ by their distribution in the mix was described in [16] and was used for coding the mixsmsn R package used in this thesis. Looking directly at the multivariate case as summarized in [14]:

The p-dimensional random vector Y belongs to the Scale Mixture Skew Normal (SMSN) family when $Y = \mu + U^{-1/2}Z$ where μ is a $px1$ location vector, $Z \sim SN(0, \Sigma, \lambda)$, and U is a positive random variable, independent of Z with a distribution

factor $H(.|\nu)$which is known as the mixing scale distribution indexed by the parameter ν.

From [13] and [14], the marginal skew-normal family density function is given by:

$$SMSN(y|\mu,\Sigma,\lambda,\nu) = 2\int_0^\infty \phi(y|\mu,\mathrm{u}^{-1},\Sigma)\Phi\left(u^{1/2}\lambda^T\Sigma^{-1/2}(y-\mu)\right)dH(u|\nu) \quad (21)$$

Where $\phi(\cdot)$ is the density of the p-variate normal distribution and $\Phi(\cdot)$ represents the distribution function of the standard univariate normal distribution [13][14]. The choice of $H(\cdot|\nu)$determines which member of the family you are using. For example:

- When U=1 and λ = 0 this is the multivariate normal distribution.
- When U=1 this is the multivariate skew-normal distribution.
- When $U \sim Gamma(\nu/2,\nu/2)$ with $\nu > 0$ and $Gamma$ (a, b) denotes the distribution with mean a/b this is the multivariate skew student-t distribution

Estimation of the parameters of the FMSN can be accomplished using a form of Expectation Maximization called Expectation-Conditional Maximization where the actual density function is maximized rather than the log likelihood. Complete details of the derivation of the parameter estimates and Maximum Likelihood function can be found at [14] and [15].

Conditional Anomaly Detection Probability Density Function (FCAD)

Given a data set with each tuple being an ordered set of variables $(x_1,\ldots\ldots,x_k,y_1,\ldots\ldots y_k)$, which can be represented as an ordered pair of two sets of

attribute values(x, y), f describes a probability distribution function that gives the likelihood that a single experiment with input x will give an output **y** where x is the set of context variables and y the set of indicator variables. For an individual data point its FCAD value describes the probability that the set of indicator values would be associated with the set of context variable values given the relationship between x and y values in the entire data set. In other words, how unusual the combination of context and indicator values is.

Formally, this PDF is described as $f(\mathrm{y}|\theta, \mathrm{x})$ where the PDF is conditioned on x and θ is the set of model parameters that generate the set of indicator values **y**. Θ is the complete set of parameters that describe the relationship between the context variables and the indicator variables. The following three sets of parameters which make up the overall parameter set Θ are used:

- A mixture model U that contains n_u mixture components each of dimensionality d_u. U models the data sets context variables and the ith mixture component is denoted U_i.
- A set V of n_v additional mixture components each of dimensionality d_v that model the indicator portion of the data space. V models the data sets indicator variables and the jth mixture component is denoted V_j.
- Probabilistic mappings function $p(V_j|U_i)$. This gives the probability that a given U_i maps to a particular V_j.

Figure 2 shows the probabilistic mapping model used by FCAD. It can be seen that both the context and the indicator variables form clusters of data points as in the above bivariate example. The context portion of the tuple will form clusters of points U that can be mapped to clusters of indicator value points V. As seen in the figure 2, $p(V|U)$ is the probability that a particular cluster U_i is mapped to a particular cluster V_j.

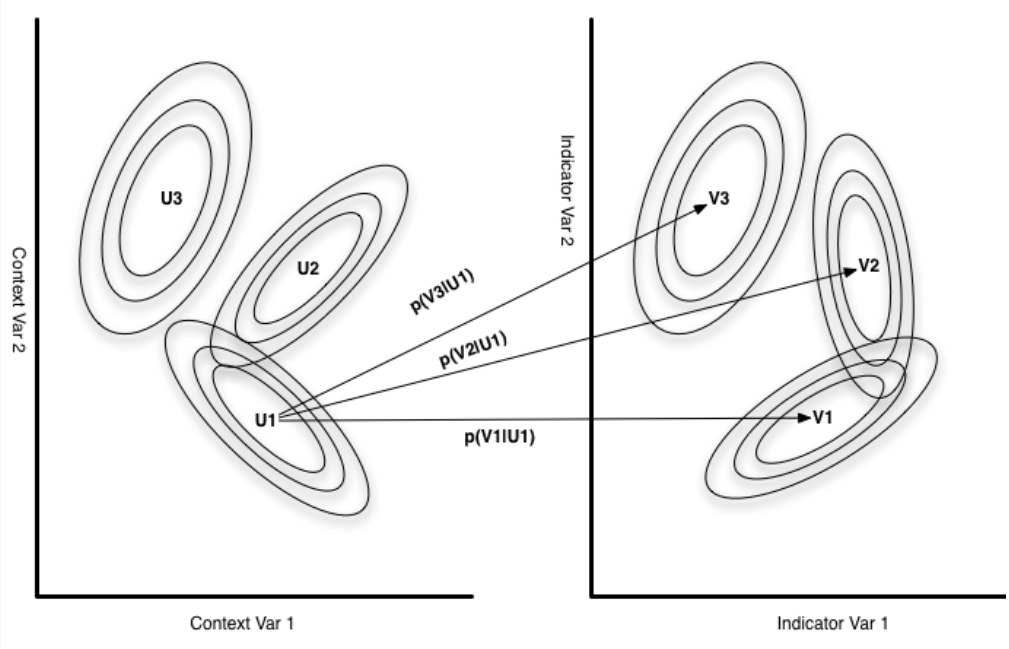

Figure 2 – FCAD Model

Therefore,

$$FCAD(y|\theta, x) = \sum_{i=1}^{n_u} p(x \in U_i) \sum_{j=1}^{n_v} f(y|V_j) p(V_j|U_i) \tag{22}$$

Where:

$$p(x \in U_i) = \frac{f(x|U_j)p(U_j)}{\sum_{k=1}^{n_U} f(x|U_k)p(U_k)} \tag{23}$$

Which is the Bayesian Probability that x was produced by the i^{th} mixture component in U.

$f(y|V_j)$ is the likelihood that the j^{th} mixture component in V would produce y.

$p(V_j|U_i)$ is the probability that the i^{th} mixture component from U maps to the j^{th} mixture component from V. This is directly given as a parameter in θ.

θ is chosen so as to maximize the log-likelihood of $f(y|\theta, x)$ for all possible values of θ using expectation-maximization.

Song et al. explored three different methodologies for θ. Two of the three involved jointly learning the parameters of U and V and mathematically did not lend themselves to incorporating a non-Gaussian distribution. The third, called **GMM-CAD-Split** algorithm learns the mixture models for U and V separately and then learns the mapping function between U and V and is the focus of this thesis. The algorithm becomes:

1. Learn parameters for U and V by doing separate Expectation-Maximization optimizations.

2. Compute joint probabilities for all k, i, j where k refers to number of data points, i refers to number of Gaussians in U, and j refers to number of Gaussians in V.

3. Compute updated $\overline{p(V_j|U_i)}$ which is the best guess:

$$\overline{p(V_j|U_i)} = \sum_{k=1}^{n} b_{kij} \Big/ \sum_{k=1}^{n}\sum_{h=1}^{n_v} b_{kih} \tag{24}$$

where:

$$b_{kij} = \frac{f(x_k|U_i)p(U_i)f(y_k|V_j)p(V_j|U_i,\theta)}{\sum_{t=1}^{n_u}\sum_{h=1}^{n_v}\{f(x_k|U_t)p(U_t)f(y_k|V_h)p(V_h|U_t,\theta)\}} \tag{25}$$

4. Set $p(V_j|U_i) = \overline{p(V_j|U_i)}$.

Although Song, et al used Gaussian distributions in their paper, $f(x|U)$ and $f(y|V)$ can be calculated for other distributions. In this thesis, the skew-normal distribution was substituted and explored.

Quantile – Quantile Plot (Q-Q Plot)

Q-Q plots are a way to compare whether a data set comes from a specific distribution or whether two data sets come from a common distribution. The data can be compared to a

random sample drawn from a specific distribution to determine if the data comes from that distribution. A perfect match plots as a 45-degree line. Deviations from this 45-degree line are an indication of how badly and where in the distribution there is deviation.

Kolmogorov-Smirnov Test

The Kolmogorov-Smirnov (KS) test is a method for comparing a data sample to a specific probability distribution or to another sample. The test evaluates the distance between the "empirical distribution function" of the sample and the cumulative distribution function of either a predetermined distribution or the empirical distribution function of another sample. The test generates a p statistic and having $p > \alpha$, where α is the significance level, accepts the null hypothesis. Consequently, big p-values establish a strong confidence in the hypothesis that the two samples came from the same distribution. In the case of this Thesis, the significance level is .05.

3. RELATED WORK

Contextual anomaly detection is a relatively new area of research with the bulk of historical research focused on point data and separation based outlier detection. [1]

This research is based on and extends the often cited paper: “Conditional Anomaly Detection” by Song et al. [3]. Song used Gaussian based Expectation Maximization and training sets of data to determine the parameters of the Gaussian distribution for each context and indicator variable. Then, they proceeded to determine probabilistic mapping functions between the context Gaussian distributions and indicator Gaussian distributions. Their algorithm, called FCAD, generates a ranking value for each data point, hereby referred to as an FCAD value. These ranking values indicate the degree to which a given data point is an anomaly based on how unusual a point’s indicator variable values are in relation to the point's context variable values. To the best of our knowledge, the work presented in [3] is the first work explicitly segregating variables into contextual and indicator variables and attempting find probabilistic relationships between these variables. Nevertheless, Song et al. did not explore the use of non-Gaussian probability distributions.

Babbar and Chawla have cited this paper ([3]) in their work and denote that it was the first to offer an approach to discover contextual anomalies and have shown no pertinent additional research [5], but to date, as far as we know, no research has extended the concepts in the proposed direction. This paper has been cited extensively by others. For example, in his Doctoral thesis, "Detecting Patterns of Anomalies," Das developed a set

of methods applicable to spatial scan data [6]. His method purports to treat the contextual and attribute variables as completely overlapping. His approach is interesting but clearly not an extension of [3].

Tang et al. developed a methodology for finding contextual outliers in categorical data using the concept of parent / child relationships which might be considered to be loosely based on the methodology of Song et al. but is otherwise dissimilar [15].

Other researchers in related fields of study have proposed data analysis methodologies that have similarities to the overall mapping methodology of [3] but are very different in execution. For example, Deodhar et al. propose a methodology using co-clustering that maps the joint relationships of dyadic data where the data variables are each in turn represented by a vector of characteristics [16]. Although finding probabilistic relationships between sets of variables, context is not an explicit goal.

To the best of our knowledge, and based on the extensive literature review we have performed, this proposal outlines the first research using contextual mapping assuming non-Gaussian distributions. Several papers have been written in recent years that apply the skew-normal distribution to situations historically reserved for the normal distribution. These papers applied the skew-normal distribution to modeling non Gaussian data but did neither use the skew-normal for modeling as an extension of FCAD or to improve the ability to detect anomalies of any sort.

4. METHODOLOGY

Experimental Setup

Experiments were performed on a Macbook Pro. The data sets were prepared using the R language environment RStudio. R is an open source statistical computing language. Both skew-normal and normal modeling of the data was done in R using the mixmsmn package [13]. The mixsmsn package is used to model the context and indicator data. Its role is to determine the best fit including the most appropriate number of mixture components, the parameters of the mixtures to be used in the PDF and the estimated probability values for each data point for each mixture cluster. FCAD can then be configured to use this information to calculate the FCAD values for each data tuple. This package provided Expectation-Maximization based modeling of the univariate and multivariate mixed model skew-normal family of distributions. In addition, it supports capabilities such as random variable generation based on model parameters. The FCAD algorithm was implemented in MATLAB using the R output as an input. Finally, the FCAD results were imported back to R for analysis and visualization.

Experimental Design

Five experiments were conducted. Table 1 lists the five experiments and their characteristics.

Table 1 – List of Experiments Run

1	Compare skew-normal distribution modeling of exponential and normal mixtures with normal distribution modeling of the same mixtures
2	Compare FCAD results when modeling normal context/normal indicator data using both skew-normal and Gaussian distribution modeling
3	Compare FCAD results when modeling normal context/exponential indicator data using both skew-normal and Gaussian distribution modeling
4	Compare FCAD results when modeling exponential context/exponential indicator data using both skew-normal and Gaussian distribution modeling
5	Compare FCAD results when modeling weather data using both skew-normal and Gaussian distributions

Experiments 1, 2, 3, and 4 all use data sets that were constructed from a base set of exponential and normal data. The base exponential data parameters are shown in Table 2.

Table 2 – Parameters of the original exponential mixture data

Exponential Mix Component	λ	**Offset**
Mix Component 1	.4	0
Mix Component 2	.4	15

The exponential data is a two component mixture model created with 1000 random variable samples from the exponential PDF $f(x|\lambda) = \lambda e^{\lambda x}$. Another 1000 samples were created by adding the offset 15 to the first set of samples. This data is shown in Figure 3.

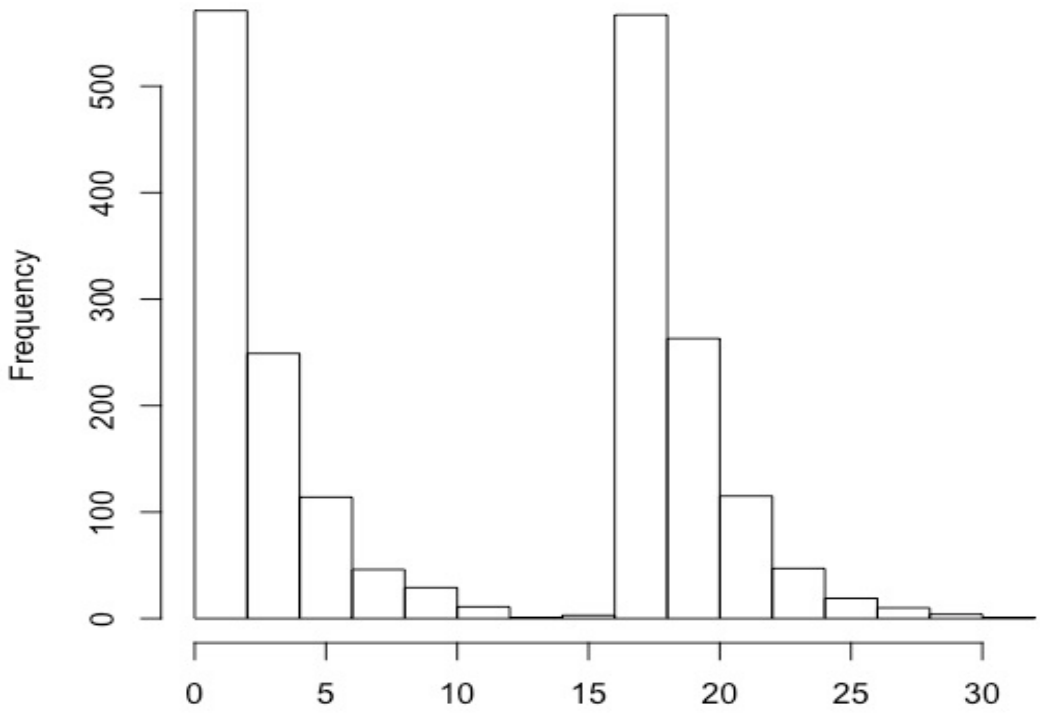

Figure 3 - Histogram of Original Exponential Data

The normal data is a two component mixture model created with 2000 random variable samples from the normal PDF $f(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{\frac{(x-\mu)^2}{2\sigma^2}}$ using the parameters shown in Table 3.

Table 3 – Parameters of original normal mixture data

Normal Mix Components	$\boldsymbol{\mu}$	$\boldsymbol{\sigma}$
Mix Component 1	4	.5
Mix Component 2	12	.5

The data is shown in Figure 4.

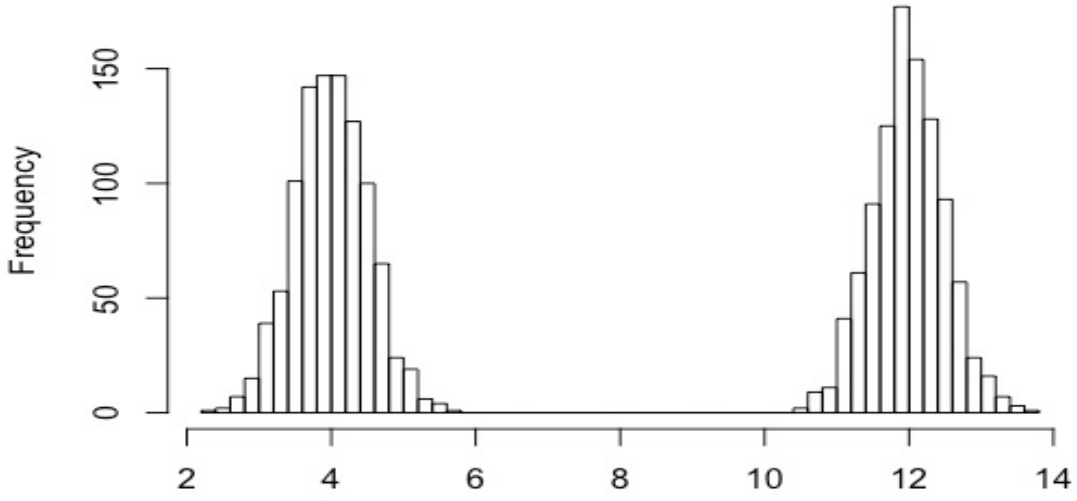

Figure 4 - Histogram of Original Normal Data

Experiment one compared the ability of the skew-normal to model the univariate mixture of exponential distributions from Figure 3 and the univariate mixture of normal distributions from Figure 4 to the ability of the normal distribution to model the same mixtures. To evaluate exponential modeling both the skew-normal and the normal distribution based modeler then used the random samples from Figure 3 to generate their own set of parameters describing the original exponential mixture model. 2000 random samples were generated from each parameter set and used to visualize and analyze how well each distribution modeled the original exponential mixture model. The number 2000 was arbitrarily chosen to be large enough adequately represent the distribution both visibly and computationally. To evaluate normal modeling both the skew-normal and the normal distribution based modeler then used the random samples from Figure 4 to

generate their own set of parameters describing the original normal mixture model. 2000 random samples were generated from each parameter set and used to visualize how well each distribution modeled the original exponential mixture model. Histograms, Q-Q plots, and the KS test were used for analysis.

The null hypothesis H_0 for the two sample KS test is that both samples are drawn from the same distribution. The alternate hypothesis, H_1, says that the samples came from different distributions. The KS test compares the original normal and exponential data sets to random samples generated from both models. Explicitly, the parameters generated modeling the normal sample data set using both normal and skew-normal distributions are used to generate 2000 point random samples that can be compared to the original 2000-point normal data. The specific hypotheses for normal data modeling are:

$H_{0-normal:normal}$: The sample from the model based on the normal distribution was drawn from the same distribution as the original normal data modeled.

$H_{1-normal:normal}$: The sample from the model based on the normal distribution was drawn from a different distribution than the original normal data modeled.

$H_{0-normal:skew}$: The sample from the model based on the skew-normal distribution was drawn from the same distribution as the original normal data modeled.

$H_{1-normal:skew}$: The sample from the model based on the skew-normal distribution was drawn from a different distribution than the original normal data modeled.

Additionally, the parameters generated modeling the exponential sample data set using both normal and skew-normal distributions are used to generate 2000 point random samples that can be compared to the original 2000-point exponential data. The specific hypotheses for exponential data modeling are:

$H_{0-exponential:normal}$: The sample from the model based on the normal distribution was drawn from the same distribution as the original exponential data modeled.

$H_{1-exponential:normal}$: The sample from the model based on the normal distribution was drawn from a different distribution than the original exponential data modeled.

$H_{0-exponential:skew}$: The sample from the model based on the skew-normal distribution was drawn from the same distribution as the original exponential data modeled.

$H_{1-exponential:skew}$: The sample from the model based on the skew-normal distribution was drawn from a different distribution than the original exponential data modeled.

In this Thesis, we have used a decision threshold of 0.05. Hence, if the p-value from the KS is greater than 0.05 we reject the alternative H_1 hypothesis that the samples came from different distributions and can conclude that they had a high likelihood of being generated by the same distribution.

Experiment two compared the results of using the normal and skew-normal distributions to model a synthetic data set as part of using FCAD for anomaly detection where both context and indicator variables are bivariate sets of normally distributed random variables. The context data set was created by generating 1000 random normal sample pairs for each mixture component using the parameters in Table 3 and then concatenating the rows for a total of 2000 data tuples. The 2000 was chosen arbitrarily to adequately represent the distribution both visibly and in computation. The distributions were not mixed. The indicator data is a copy of the context data generating a four

variable final data tuple for each point. Hence, there is no mixing between distributions, the context and indicator tuples are the same, and no cross mapping between context and indicator distributions. How well the modeler was able to fit the data was evaluated by comparing the number and distribution of the 10 smallest, most anomalous, log FCAD values where log FCAD is the log of the FCAD value for each tuple. A data model of experiment 2 is shown in Figure 5.

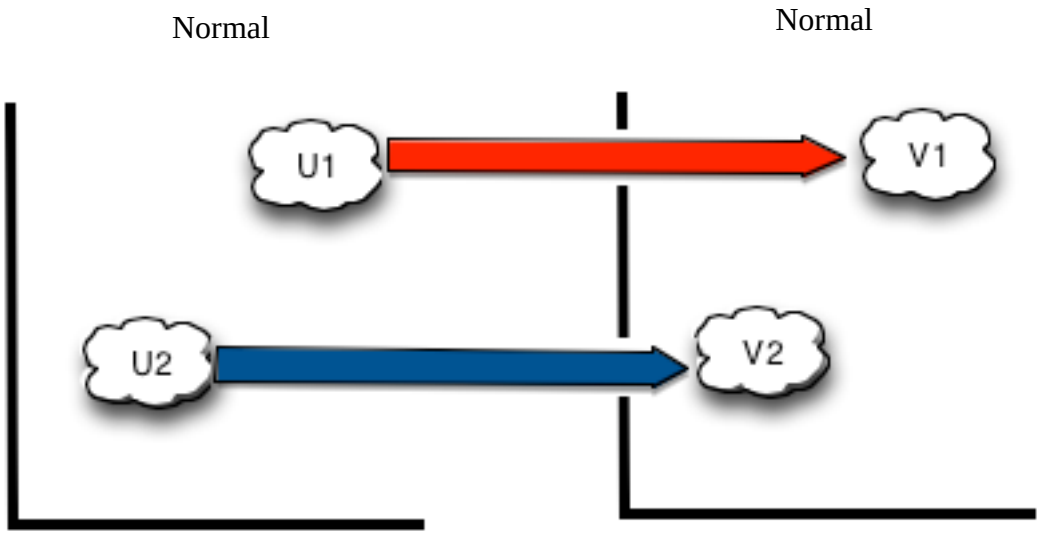

Figure 5 – Data model for Experiment 2

Experiment three compared the results of using the normal and skew-normal distributions to model a synthetic data set as a part of using FCAD for anomaly detection where the context variables are bivariate sets of exponentially distributed random variables and the indicator variables are bivariate normally distributed. The context data set was created by generating 1000 random exponential sample pairs for each mixture

component using the parameters in Table 2 and then concatenating the rows for a total of 2000 data tuples. The 2000 was chosen arbitrarily to adequately represent the distribution both visibly and in computation. The distributions were not mixed. The context data set was created by generating 1000 random normal sample pairs for each mixture component using the parameters in Table 3 and then concatenating the rows for a total of 2000 data tuples. The distributions were not mixed. Hence, there is no mixing between distributions and no cross mapping between context and indicator distributions. The context and indicator data sets were combined to form a 2000 point, four-variable, data set. How well the modeler was able to fit the data was evaluated by comparing the number and distribution of the 10 smallest, most anomalous, log FCAD values. A data model of experiment 3 is shown in Figure 6.

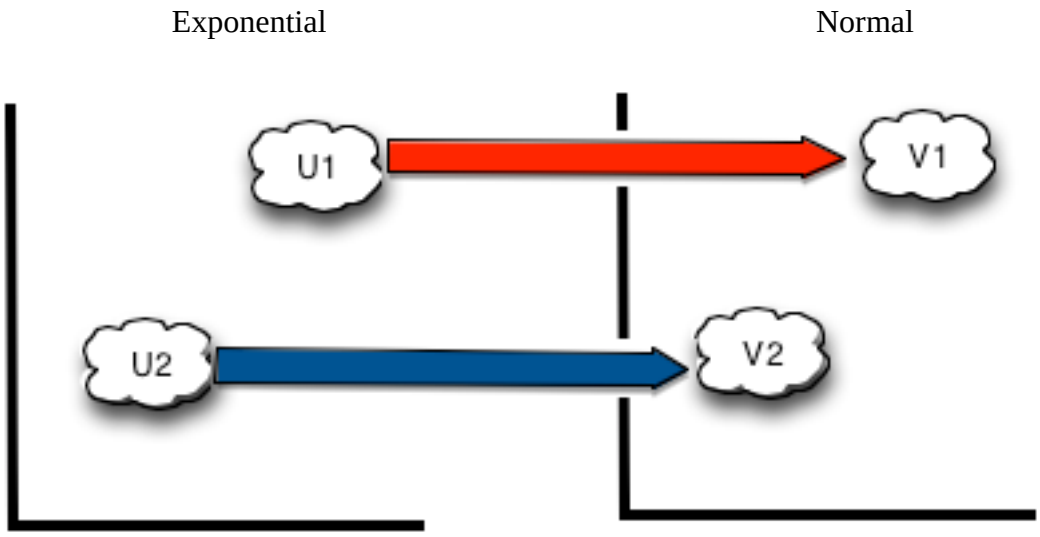

Figure 6 – Data model for Experiment 3

Experiment four compared the results of using the normal and skew-normal distributions to model a synthetic data set where both context and indicator variables are bivariate sets of exponentially distributed random variables. Data sets were created as above. Additionally, 40 points out of the 2000 were modified such that there was cross mapping between the lower context distribution and the upper indicator distribution. The behavior of the modelers with respect to these points was evaluated by considering the 50 most anomalous points identified by each modeler. A data model of experiment 4 is shown in Figure 7.

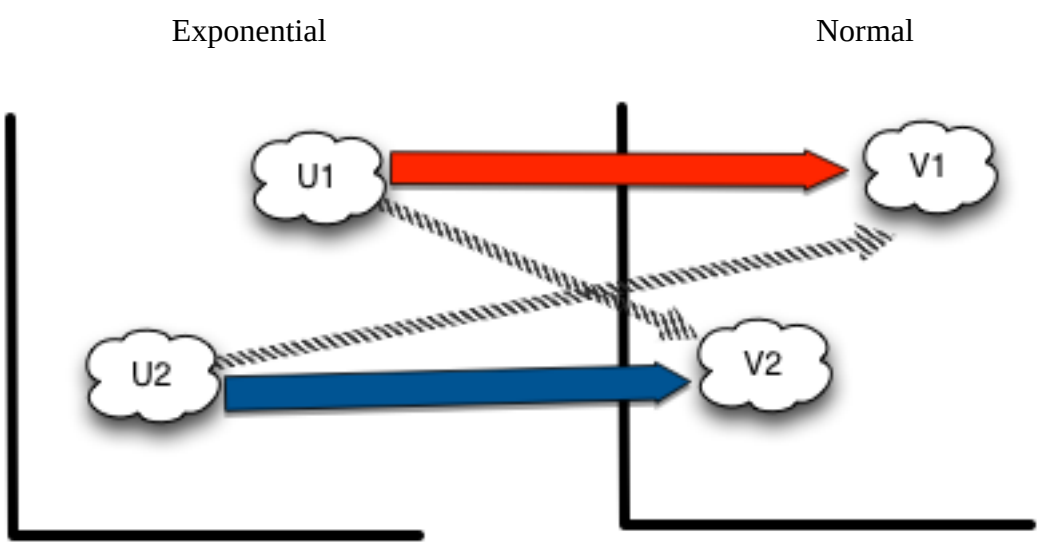

Figure 7 – Data model for Experiment 4

Experiment five compared the identification of anomalies when modeling a real world data set where both normally and exponentially distributed data is present. This experiment used two variables for the context and one variable for the indicator with both

exponentially and skewed normally distributed data and evaluated the generation of type 1 errors where the modelers differed from and expert selecting anomalies.

5. EXPERIMENTAL RESULTS

Experiment 1

Table 4 includes the parameters generated by the skew-normal model when modeling the target exponential data.

Table 4 - Skew-normal based model parameters for exponential target data

Skew-Normal Mix Component	μ	Σ	λ
1	-0.25	11.07	29.79
2	15.9	11.1	23.8

Figure 8 shows the histogram obtained from generating 2000 random samples on the using the skew-normal model parameters μ, Σ and λ that were obtained from the R **mixmsmn** skew-normal modeling package when modeling the exponential data described in Table 2 and shown in Figure 3.

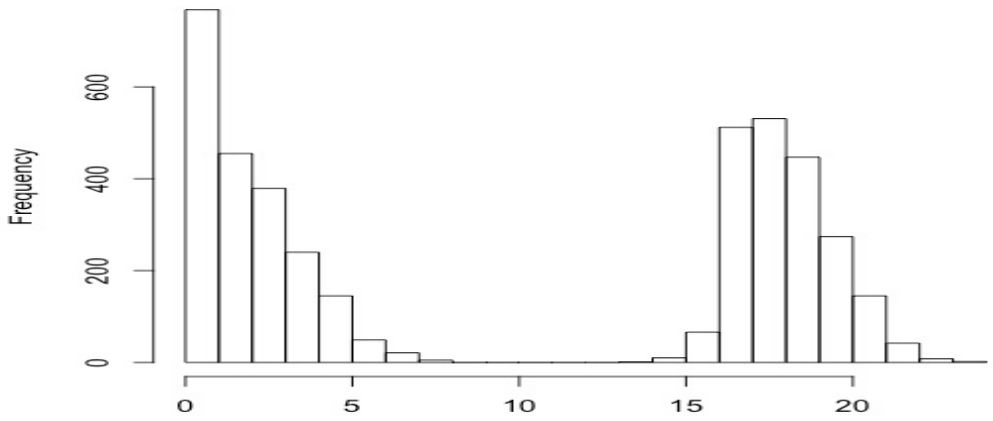

Figure 8 – Random samples from skew-normal based model of exponential data

It can be observed that the lower mixture was modeled closely by the skew-normal distribution and although there is some distortion at the beginning of the upper distribution, the bulk of the upper mixture is closely modeled by comparing them to the original distribution histogram in Figure 3. Table 5 includes the parameters generated by the normal modeler when modeling the target exponential data.

Table 5 – Normal based model parameters for exponential target data

Normal Mix Component	μ	σ
1	2.28	2.06
2	18.26	2.53

Figure 9 shows the histogram obtained from generating 2000 random points based on the using the normal model parameters μ and σ that were obtained from the R **mixmsmn** - normal modeling package when modeling the exponential data.

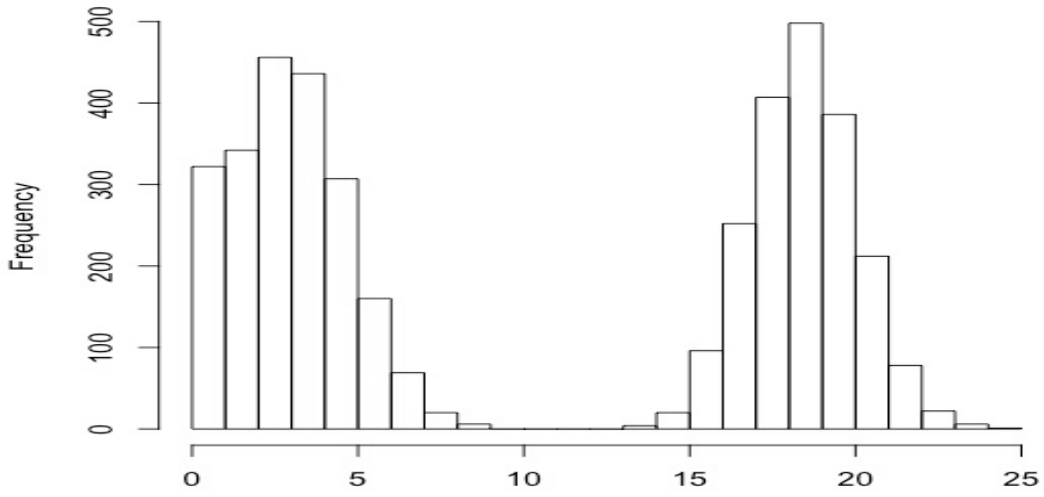

Figure 9 –Random samples from normal based model of exponential data

It can be observed that both the lower and upper exponential mixtures are poorly modeled by the normal distribution by comparing them to the original distribution histogram in Figure 3.

Table 6 includes the resulting parameters generated by the skew-normal model when modeling the normal data.

Table 6 - Skew-normal based model parameters for normal target data

Skew-Normal Mix Component	μ	σ	λ
1	3.66	.378	.941
2	11.66	.359	.934

Figure 10 shows the histogram obtained from generating 2000 random points based on using the skew-normal model parameters μ, σ and λ from Table 6 that were obtained from the R **mixmsmn** skew-normal modeling package.

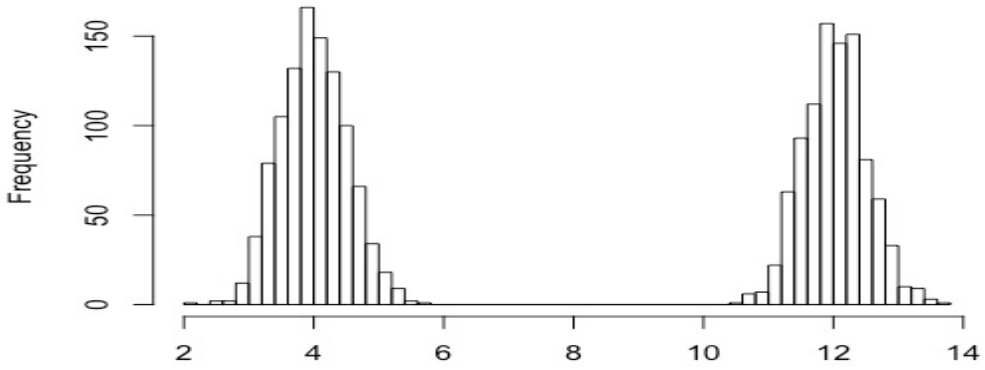

Figure 10 –Random samples from skew-normal based model of normal data

It can be observed that both mixture components were modeled closely using the skew-normal distribution by comparing them to the original distribution histogram in Figure 3. Table 7 includes the parameters generated by the normal modeler when modeling the normal data.

Table 7 - Normal based model parameters for normal target data

Normal Mix Component	μ	σ
1	3.99	.513
2	11.99	.501

Figure 11 shows the histogram obtained from generating 2000 random points based on the using the normal model parameters μ and σ that were obtained from the R mixmsmn normal modeling package and detailed above.

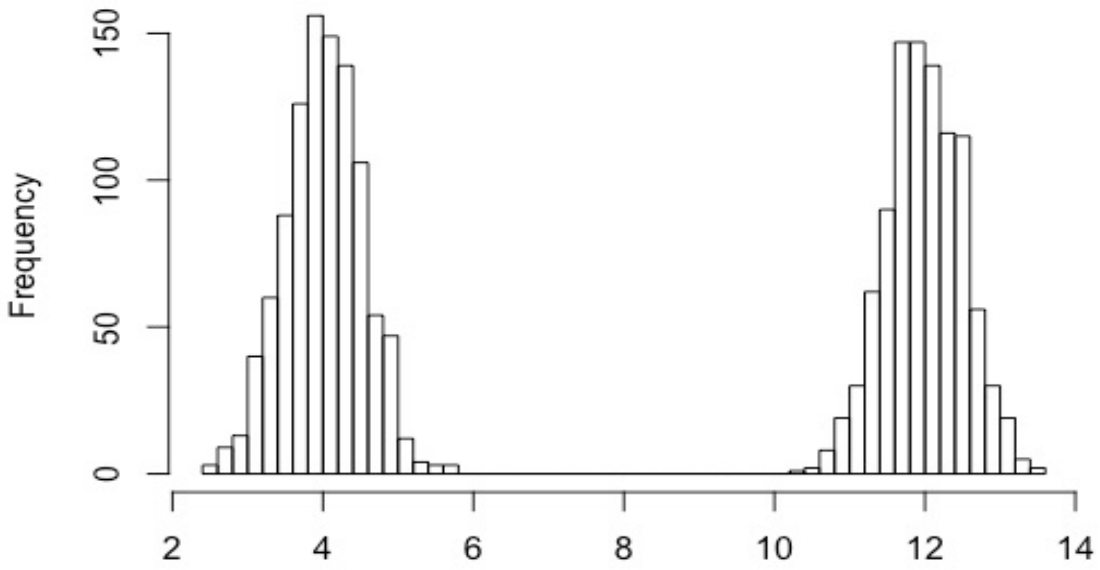

Figure 11 –Random samples from normal based model of normal data

It can be observed that both the lower and upper exponential mixtures are well modeled by the normal distribution by comparing them to the original distribution histogram in Figure 4.

Another way of expressing the quality of modeling is the Q-Q plot. The same random samples obtained from using the parameters of the two models were used for the plots. Since the Q-Q plot is best used with a single distribution, the random samples were separated into 2 groups with the break at 8.

Figure 12 shows a Q-Q plot which compares the lower distribution obtained from using the skew-normal distribution to model the exponential data to the exponential CDF. It can be observed that the majority of the points fall closely along the 45-degree line. This implies that the cumulative density functions for both the model and an exponential are very similar and that the generation models are similar.

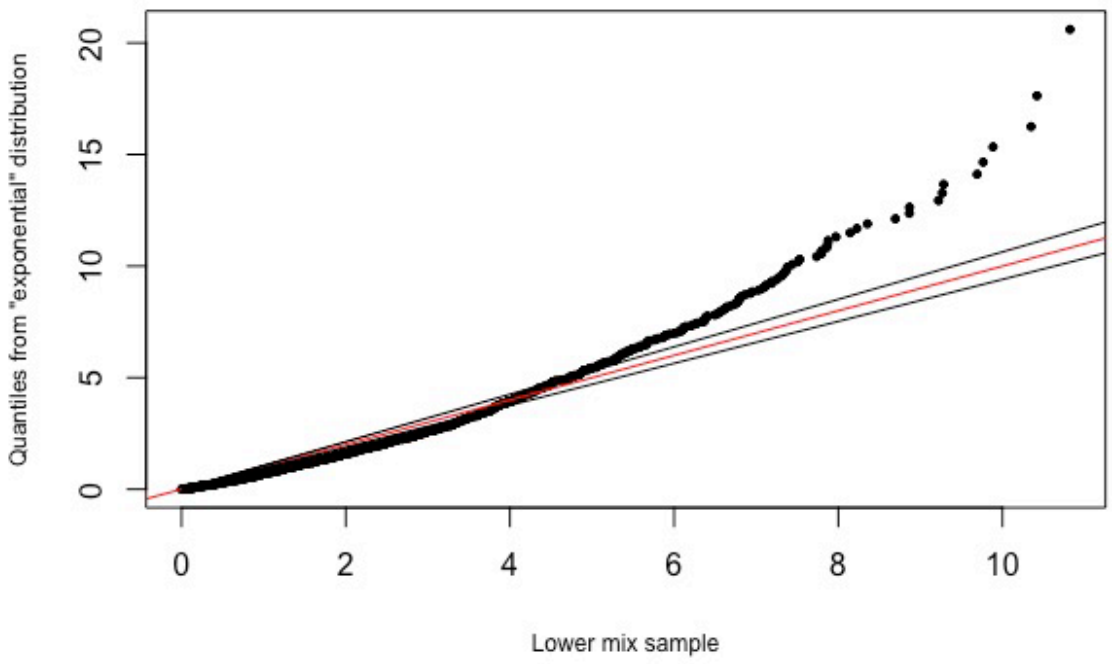

Figure 12 - Lower distribution of skew-normal based model compared to exponential CDF

Figure 13 shows a Q-Q plot which compares the lower distribution obtained from using the normal distribution to model the exponential data to the exponential CDF.

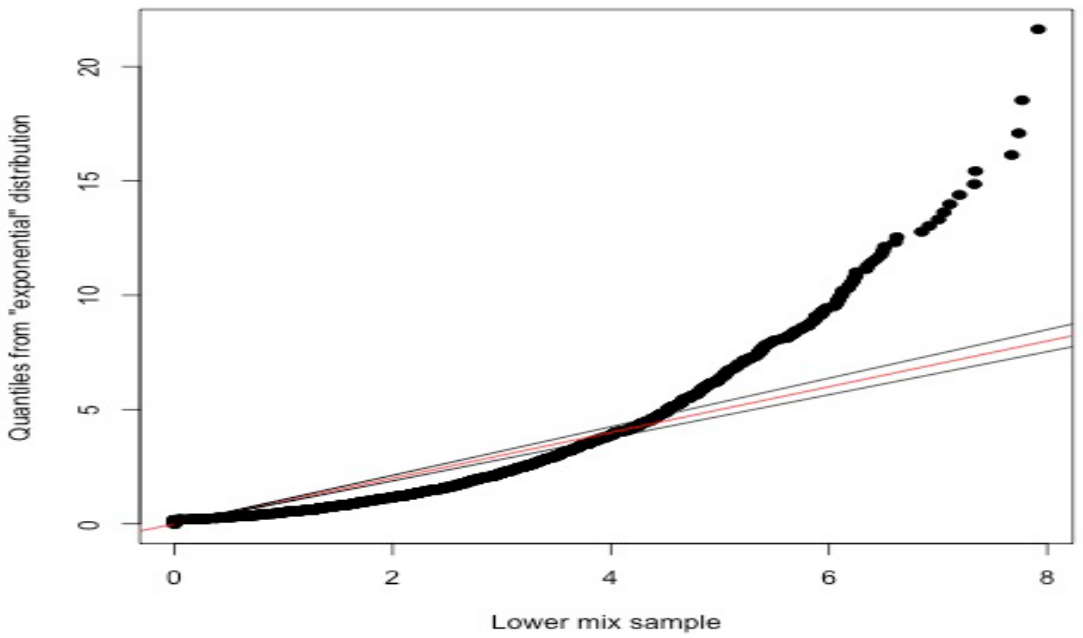

Figure 13 - Lower distribution of normal based model compared to exponential CDF

It can be observed that the majority of the points do not fall closely along the 45-degree line which implies that the cumulative density functions for the model and an exponential function are not similar and that the normal distribution does not model an exponential function accurately.

Figure 14 shows a Q-Q plot which compares the upper distribution obtained from using the skew-normal distribution to model the exponential data to the exponential function CDF.

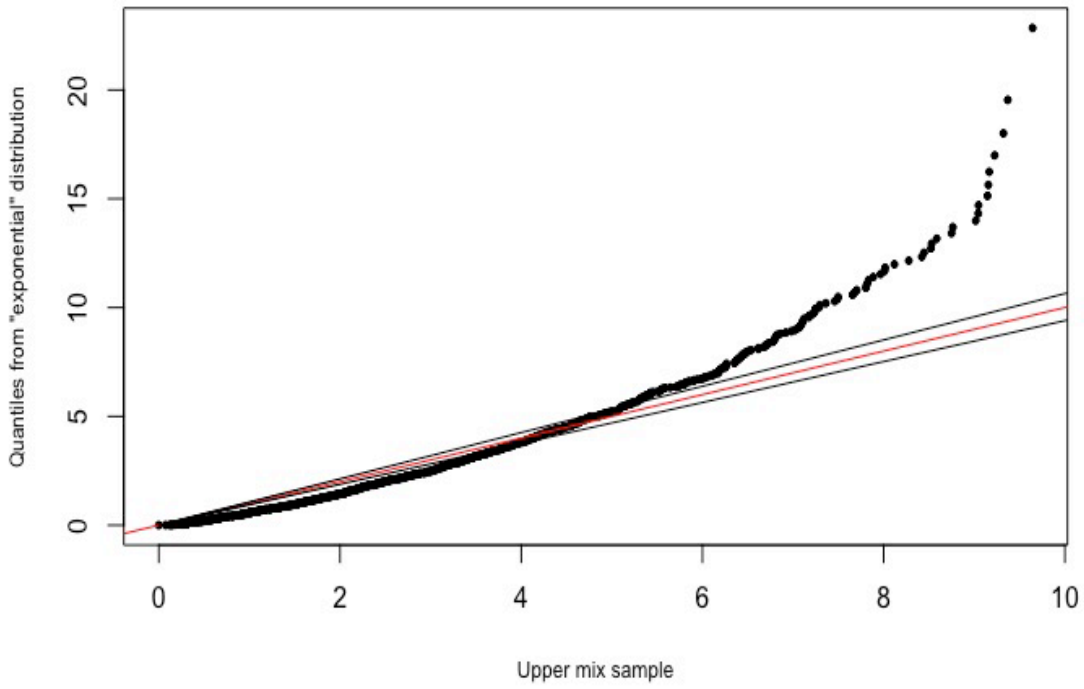

Figure 14 - Upper distribution of skew-normal based model compared to exponential CDF

It can be observed that points fall closely along the 45-degree line for the first half of the plot and then diverge. This means that the cumulative density functions for both the model and an exponential function are similar. This is in agreement with the histogram which shows modeling issues in the beginning of the plot.

Figure 15 shows a Q-Q plot which compares the upper distribution obtained from the using the normal distribution to model the exponential data to the exponential function CDF.

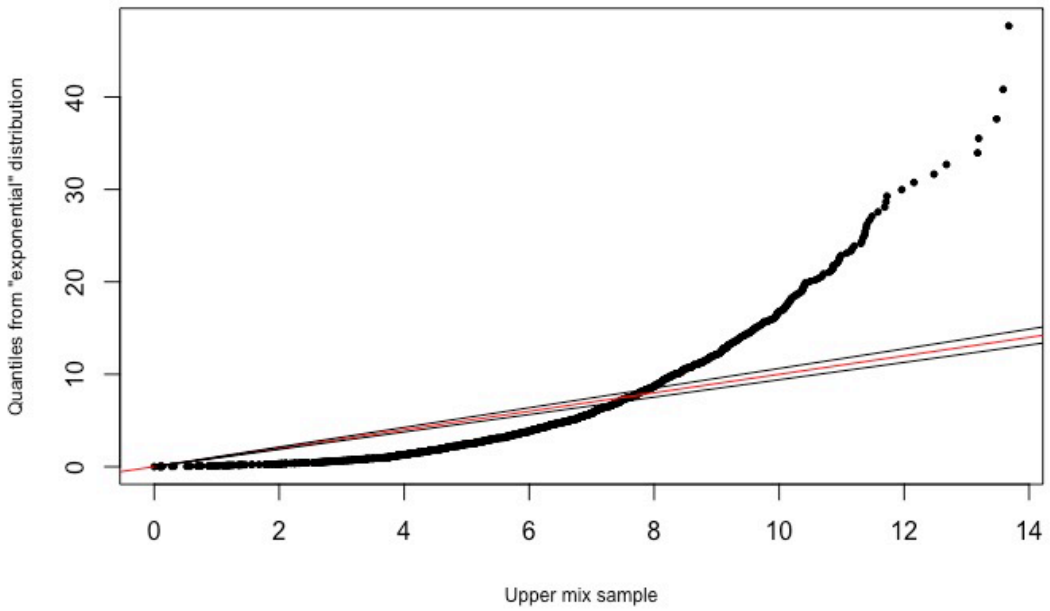

Figure 15 - Upper distribution of normal based model compared to exponential CDF

It can be observed that the points poorly track the 45-degree line. This means that the cumulative density functions for both the model and an exponential function are very different and using the normal distribution does not model the exponential data well.

Figure 16 shows a Q-Q plot which compares the lower distribution obtained from using the skew-normal distribution to model the normal data to the normal CDF.

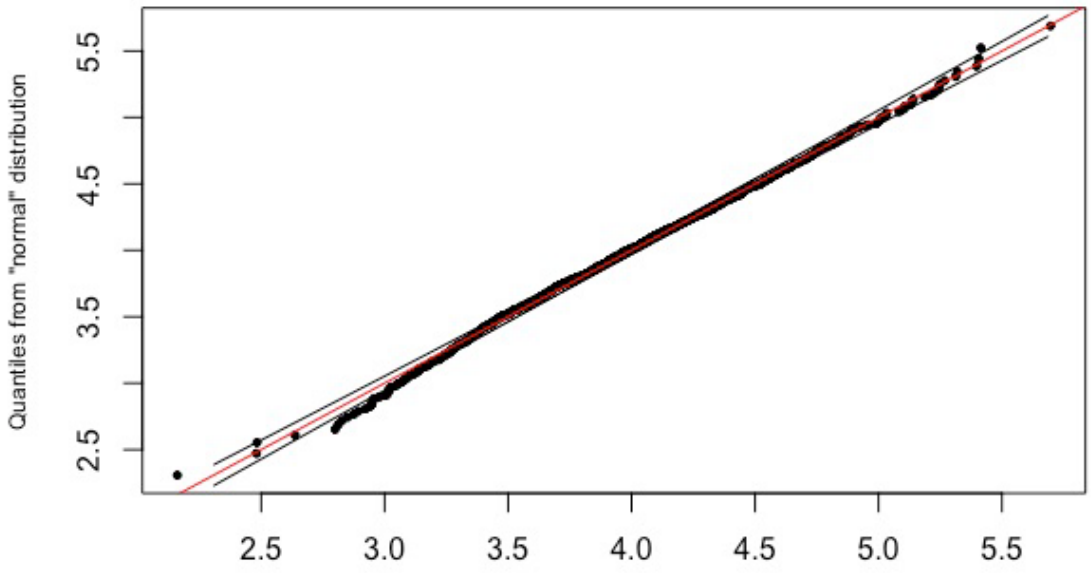

Figure 16 - Lower distribution of skew-normal based model compared to normal CDF

It can be observed that the majority of the points fall closely along the 45-degree line. This implies that the cumulative density functions for both the model and normal are very similar and that the generation models are similar.

Figure17 shows a Q-Q plot which compares the lower distribution obtained from using the normal distribution to model the normal data to the normal CDF.

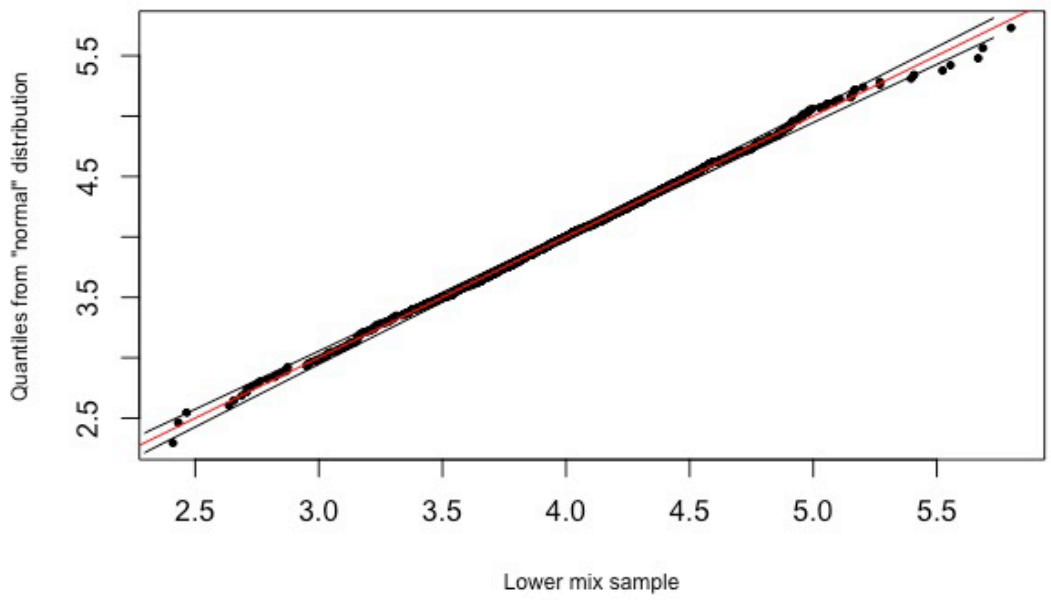

Figure 17 - Lower distribution of normal based model compared to normal CDF

It can be observed that the majority of the points fall closely along the 45-degree line. This implies that the cumulative density functions for both the model and normal are very similar and that the generation models are similar.

Figure 18 shows a Q-Q plot which compares the upper distribution obtained from using the skew-normal distribution to model the normal data to the normal CDF.

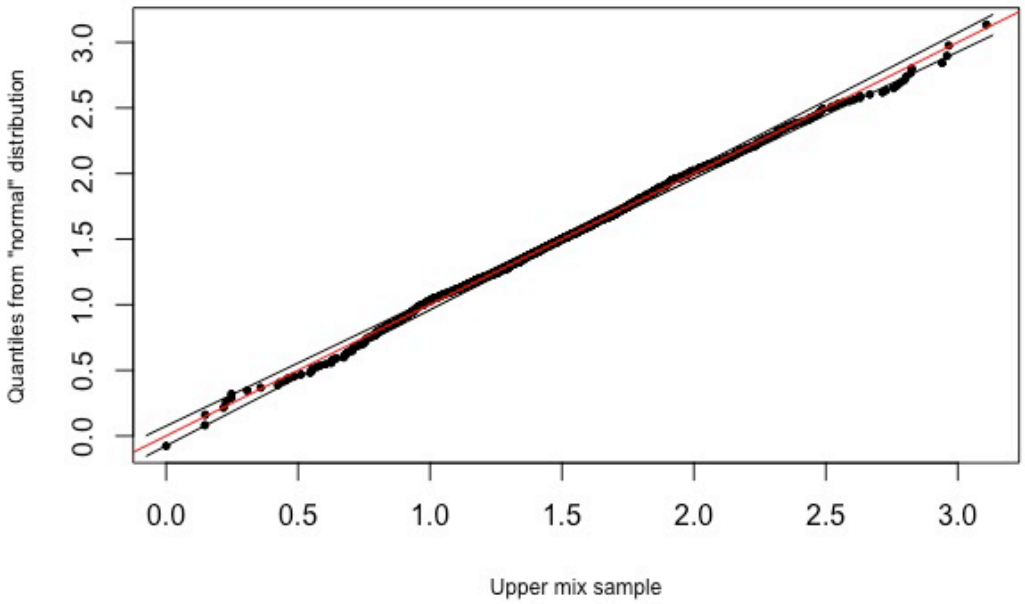

Figure 18 - Upper distribution of skew-normal based model compared to normal CDF

It can be observed that the majority of the points fall closely along the 45-degree line. This implies that the cumulative density functions for both the model and normal are very similar and that the generation models are similar.

Figure 19 shows a Q-Q plot which compares the upper distribution obtained from using the normal distribution to model the normal data to the normal CDF.

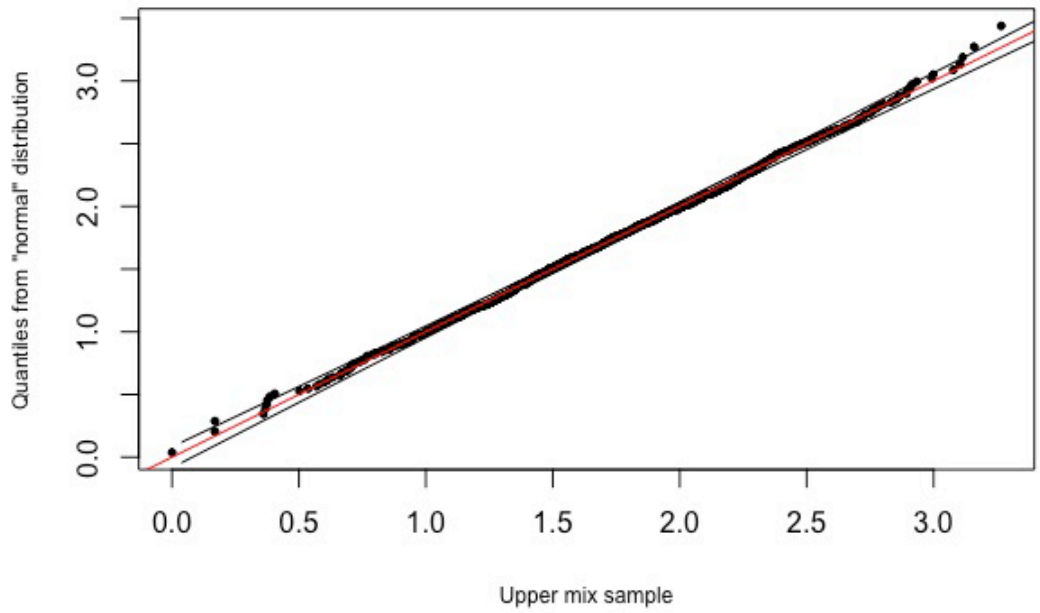

Figure 19 - Upper distribution of normal based model compared to normal CDF

It can be observed that the majority of the points fall closely along the 45-degree line. This implies that the cumulative density functions for both the model and normal are very similar and that the generation models are similar.

An additional way of evaluating the quality of modeling is to use the Kolmogorov-Smirnov (KS) test. The benefit of using this test is that the entire data set can be evaluated at once since the KS test compares the samples generated by modeled parameters to the original mixture model data. Table 8 shows the p-values from the KS test comparing the modeling of exponential and normal data by the two model types.

Table 8 – Comparison of Kolmogorov-Smirnov p values

Model	Exponential Mixture	Normal Mixture
Skew-Normal	p = .002804	p = .2414
Normal	p = 1.25 e-06	p = .8186

The two hypotheses $H_{1-normal:normal}$ and $H_{1-normal:skew}$, described in section chapter 4, were tested with the KS test to determine whether random samples drawn from the models of the normal data set based on the normal and skew-normal distributions were drawn from a different distribution than the original normal data set modeled. From Table 8, the p-values for modeling a normal data set are both greater than .05. This indicates that both $H_{1-normal:normal}$ and $H_{1-exponential:skew}$ can be rejected and that samples drawn from the model and the original data came from the same distribution. We can conclude that both models did an acceptable job modeling the normal data set but using a normal distribution to model normal data provided a better fit.

Modeling the exponential data is slightly more complicated. $H_{1-exponential:normal}$ and $H_{1-exponential:skew}$ were tested using the KS test to determine whether random samples drawn from the models of the exponential data set based on the normal and skew-normal distributions were drawn from a different distribution than the original exponential data set modeled. From table 8 it can be observed that both p-values for exponential modeling are less than 0.05 indicating that the alternate hypotheses $H_{1-exponential:normal}$ and $H_{1-exponential:skew}$ should be accepted, suggesting that the samples came from different distributions. At the same time, although grounds for

accepting the alternative hypotheses, the skew-normal p-value of .0028 is much larger than the normal p-value of .00000125. This implies that the skew-normal based model, although not superb, is significantly better than the normal distribution based model when modeling exponential data.

Experiment 2

Figure 20 shows a histogram of 2000 log FCAD values when the context and indicator variable distributions, in this case both normal data were modeled using the normal distribution.

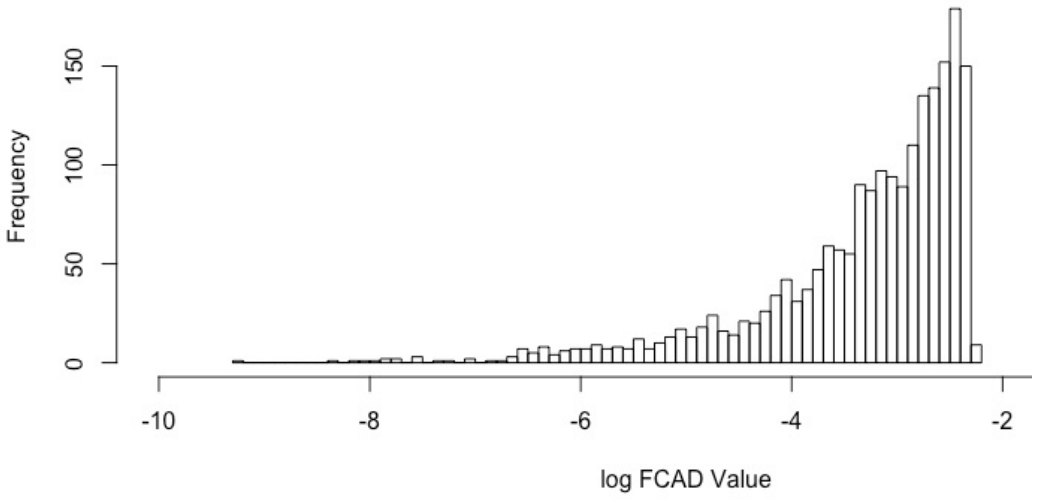

Figure 20 – Log FCAD data from normal based modeling of normal data

Figure 21 shows a histogram of 2000 log FCAD values when the context and indicator variable distributions, in this case both normal data were modeled using the skew-normal distribution.

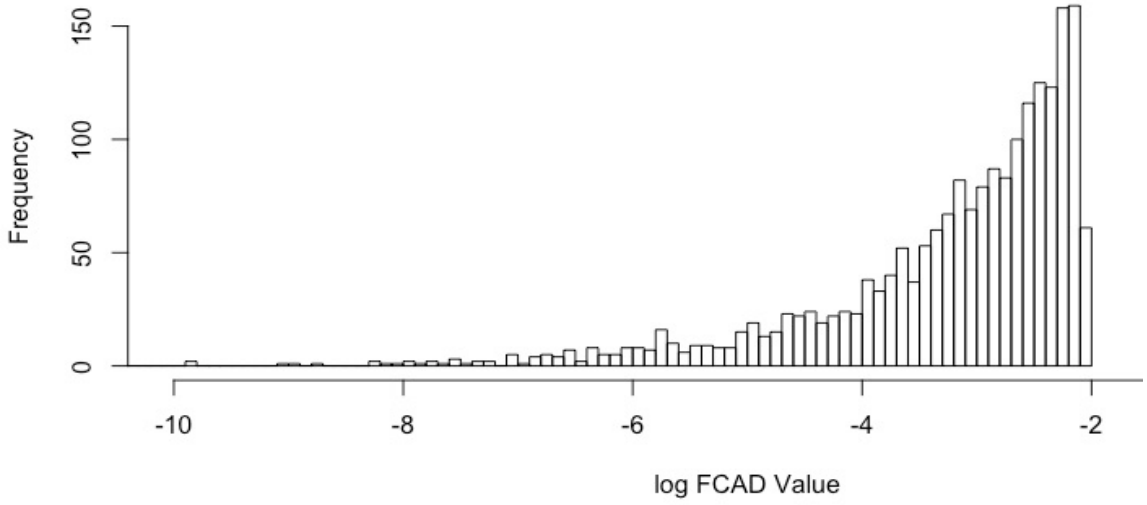

Figure 21 – Log FCAD data from skew-normal based modeling of normal data

Based on visual comparison, both FCAD data plots look similar. The mean log FCAD values, -3.331 for norm and -3.2442 are very close with standard deviations of 1.003 and 1.179 respectively. This implies that both modeling distributions yield roughly the same results at the overall level. The FCAD data from using a normal distribution has a narrower range of values than the FCAD data from using a skew-normal distribution. This smaller spread might imply less ability to discriminate between adjacent FCAD values for normal modeling. Figure 22 shows the context variables plotted and the top 10 anomalies generated using normal modeling indicated in red.

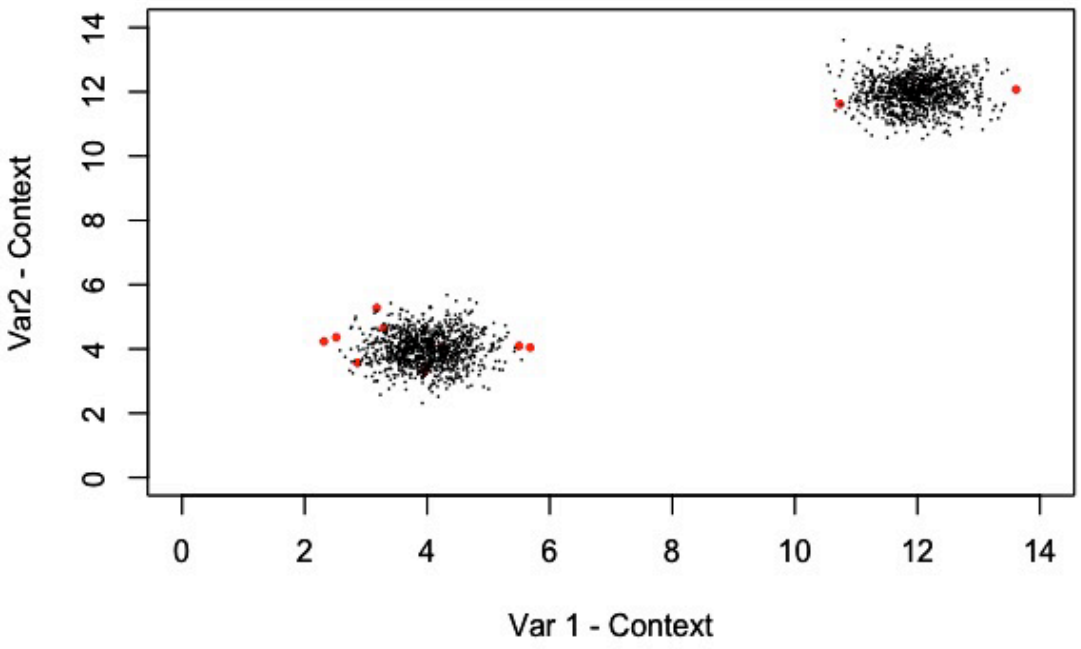

Figure 22 – FCAD anomalies for context variables from normal based modeling

Figure 23 shows the context variables plotted and the top 10 anomalies generated using skew-normal modeling indicated in red.

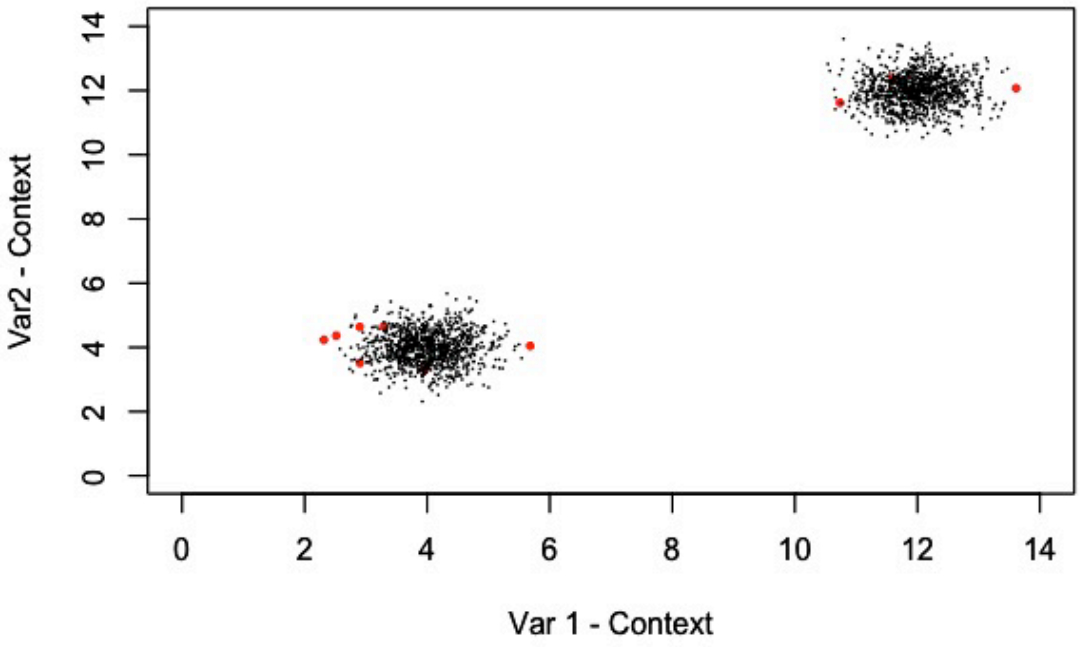

Figure 23 – FCAD anomalies for context variables from skew-normal based modeling

It can be observed that several of the points identified as anomalous appear in both plots.

Experiment 3

Figure 24 shows a histogram of FCAD values for 2000 data tuples with exponential context variables and normally distributed indicator variables when modeled using the normal distribution.

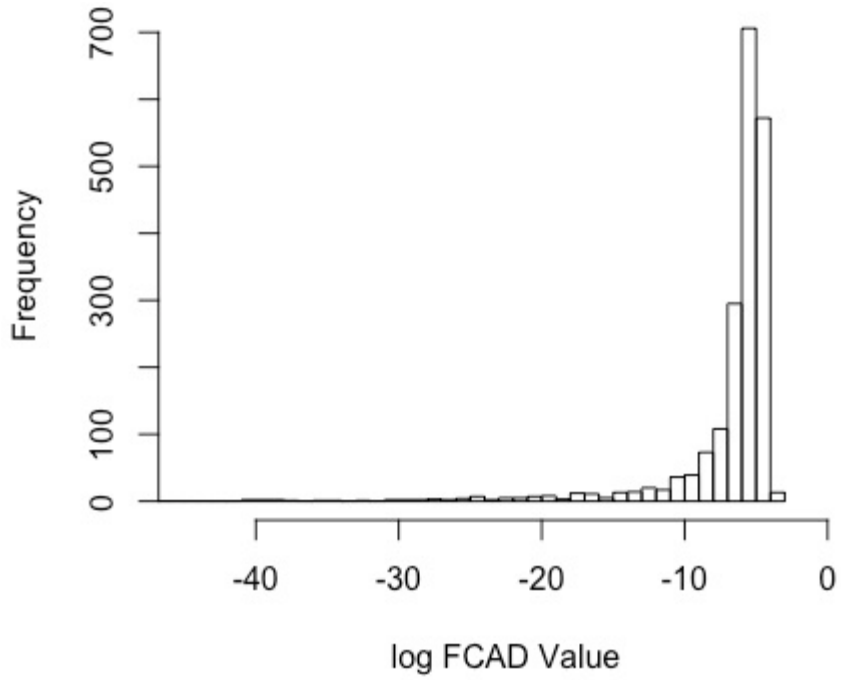

Figure 24 – Log FCAD data from normal based modeling of exponential context / normal indicator data

Figure 25 shows a histogram of FCAD values for 2000 data tuples with exponential context variables and normally distributed indicator variables when modeled using the skew-normal distribution.

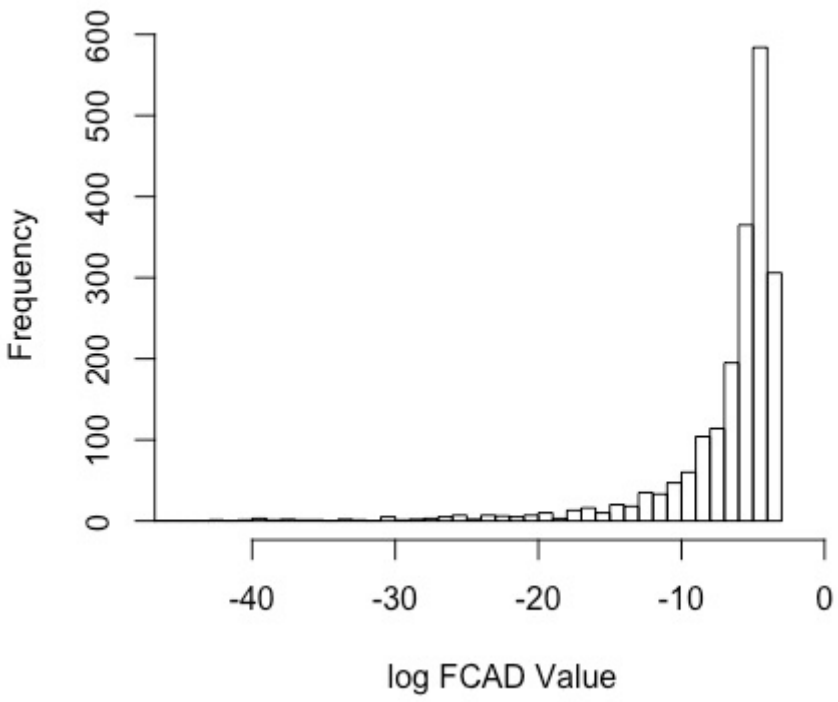

Figure 25 –Log FCAD skew-normal based modeling of exponential context / normal indicator data

It can be observed that the highest frequency values of log FCAD in the normal histogram have shifted toward the left or toward more anomalous. Additionally, it can be observed that the tail when using the normal distribution for modeling is shorter that when using the skew-normal distribution. This implies that using the skew-normal distribution allows for a more informed modeling and a more refined discrimination between individual tuples. At the same time the mean of the 200 most anomalous points (smallest log FCAD values) when using the skew-normal distribution is -19.68 versus -17.78 when using the normal distribution with standard deviations of 5.5 and 4.9 respectively. This implies that the skew-normal distribution is capable of making a "stronger" decision that a point is considered anomalous.

To further highlight the differences, Figure 26 shows a scatter plot of the context data with the 10 smallest log FCAD points when using the normal distribution for modeling highlighted in red.

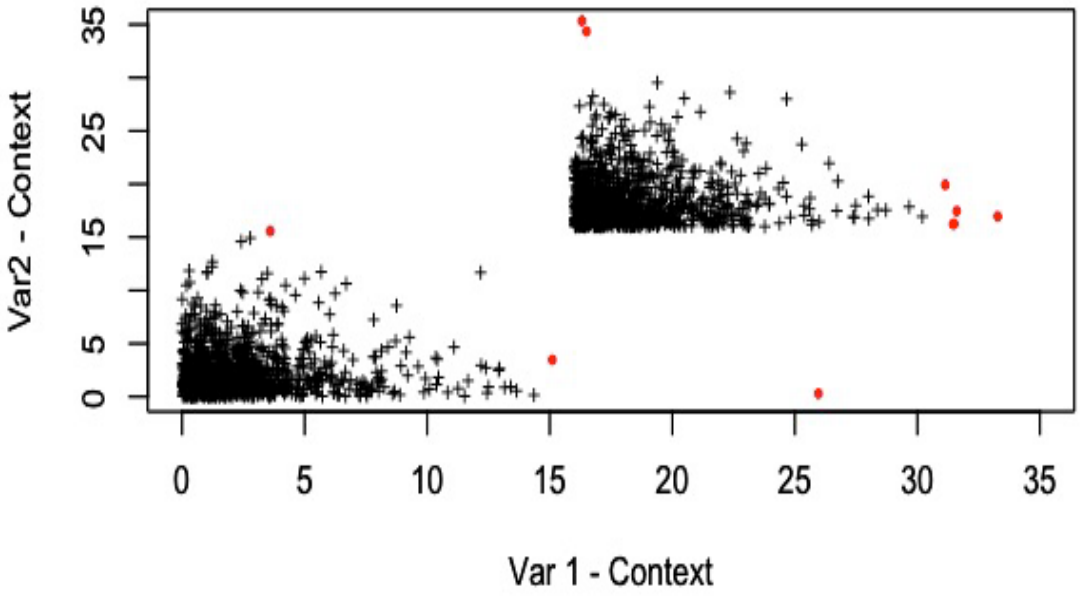

Figure 26 – Context data with small log FCAD values based on normal modeling

Figure 27 shows a scatter plot of the context data with the 10 smallest log FCAD points derived from using the skew-normal distribution for modeling highlighted in red.

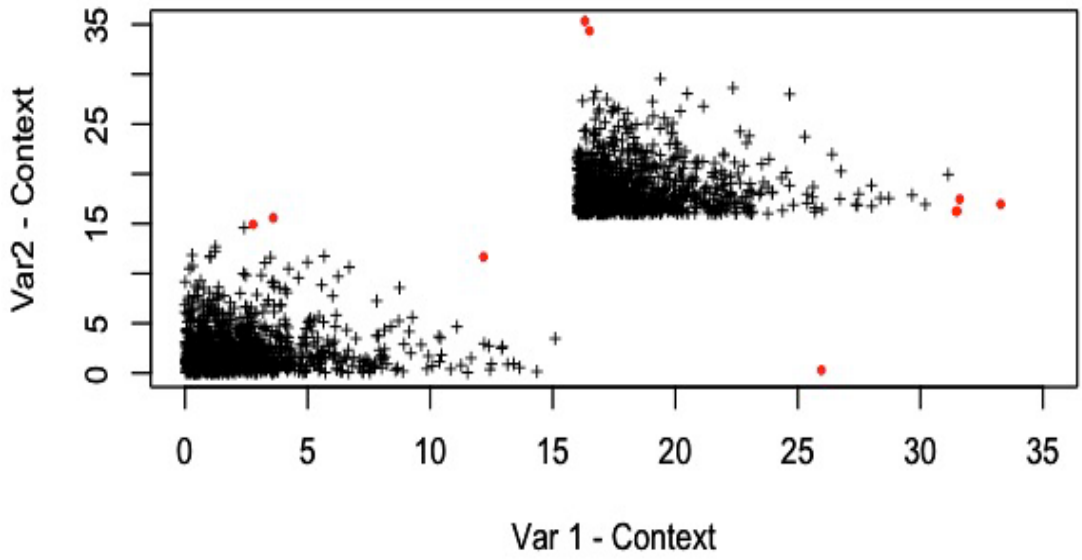

Figure 27 – Context data with small log FCAD values based on skew-normal modeling

. It can be observed that although there are slight differences between the two plots, in the main both modeling methods identify roughly the same points as anomalies.

Experiment 4

Figure 28 shows a histogram of log FCAD values for 2000 data tuples with exponential context variables and exponential indicator variables with 40 points cross mapped when modeled using the normal distribution.

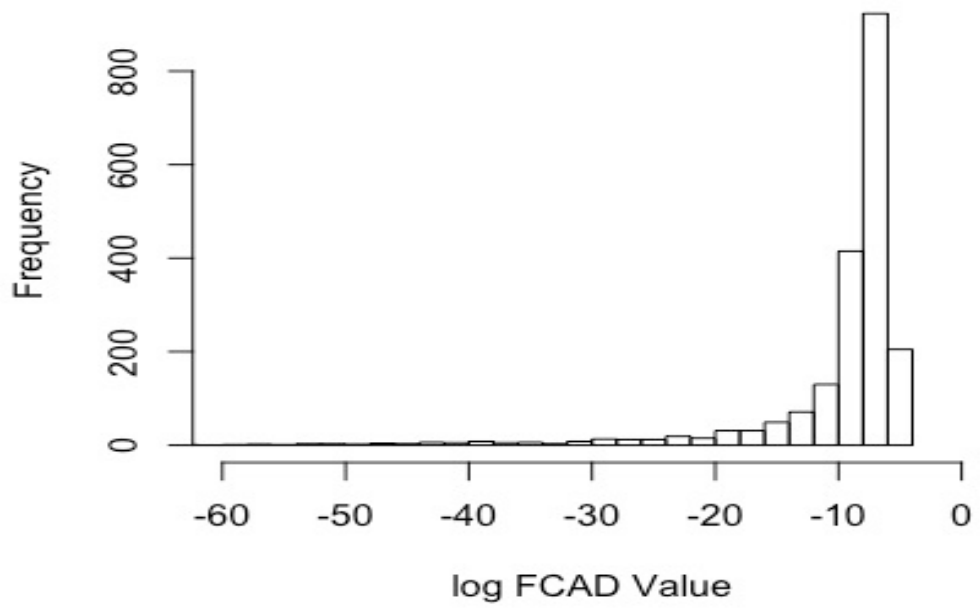

Figure 28 –Log FCAD distribution when using normal distribution in modeling

Figure 29 shows a histogram of FCAD values for 2000 data tuples with exponential context variables and exponential indicator variables with 40 points cross mapped when modeled using the skew-normal distribution.

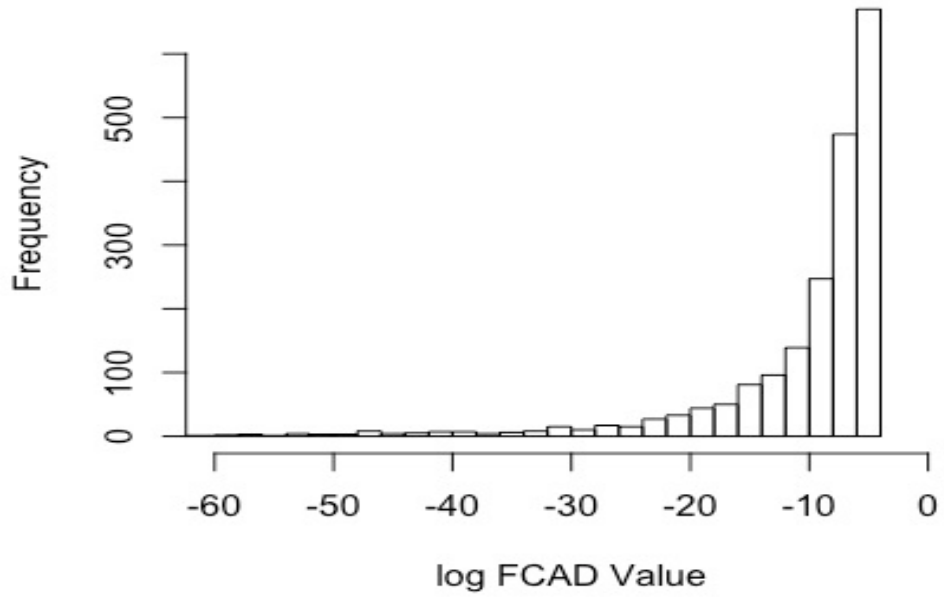

Figure 29 – Log FCAD distribution when using skew-normal distribution in modeling

It can be observed that the highest frequency values of FCAD in the normal histogram have shifted toward the left or more anomalous. Additionally, it can be observed that the spread of the skew-normal log FCAD values continues to be larger than the spread of the log FCAD values obtained from using a normal distribution for modeling. Furthermore, the mean of the 200 most anomalous points (lowest log FCAD values) when using the skew-normal distribution for modeling is -35.84 versus -32.05 when using the normal distribution for modeling with standard deviations of 10.78 and 9.60 respectively. This implies that using the skew-normal distribution is capable of making a "stronger" decision that a point is considered anomalous. Figure 30 shows a scatter plot of the context data with the smallest 50 log FCAD points derived from using the normal distribution for modeling highlighted in red.

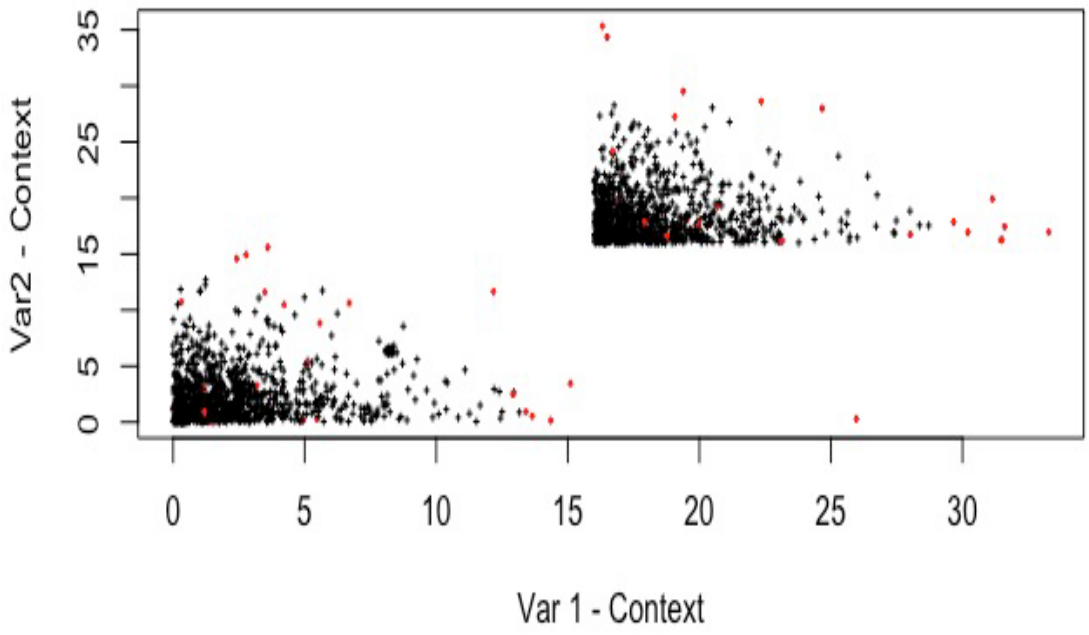

Figure 30 – Normal based modeling anomalies displayed in context data

Figure 31 shows a scatter plot of the context data with smallest 50 log FCAD points derived from using the skew-normal distribution for modeling highlighted in red.

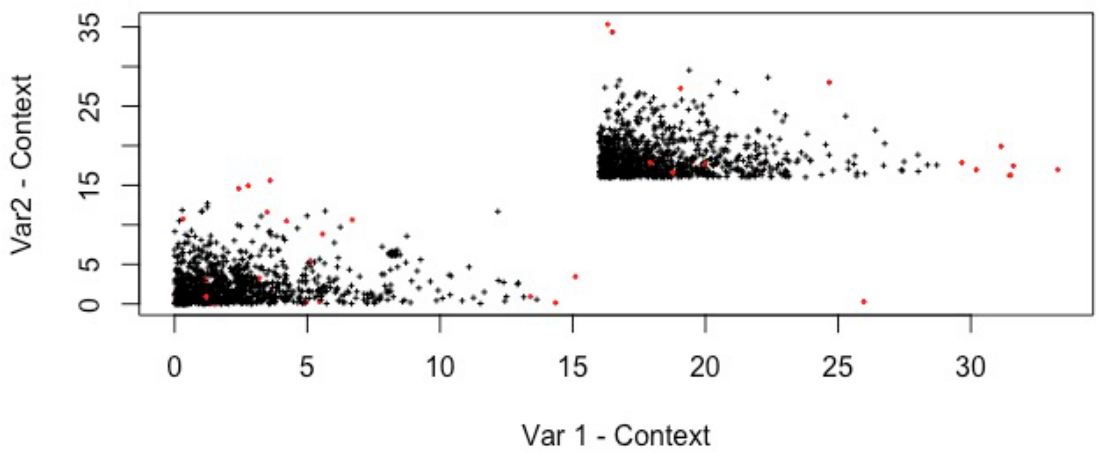

Figure 31 – Skew-normal based modeling anomalies displayed in context data

It can be observed that both modeling methods identified roughly the same points as anomalies. There was a trend exhibited when using the skew-normal distribution toward less distance based anomalies but it was subtle.

Finally, looking at the 40 injected points, none of the injected points were included in the top 50 anomalies by either method.

Experiment 5

Figure 32 shows a histogram of the maximum Central Texas temperatures from 1940 to present.

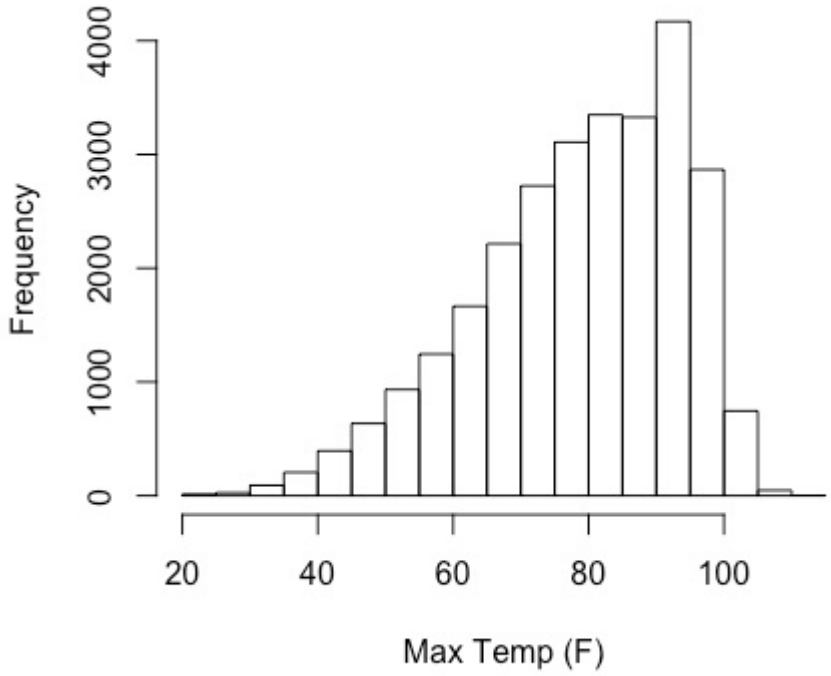

Figure 32 - Histogram of Central Texas Max Temperatures from 1940 to Present

It can be observed that the data is skewed substantially to the right. Figure 33 shows a histogram of precipitation over the same period.

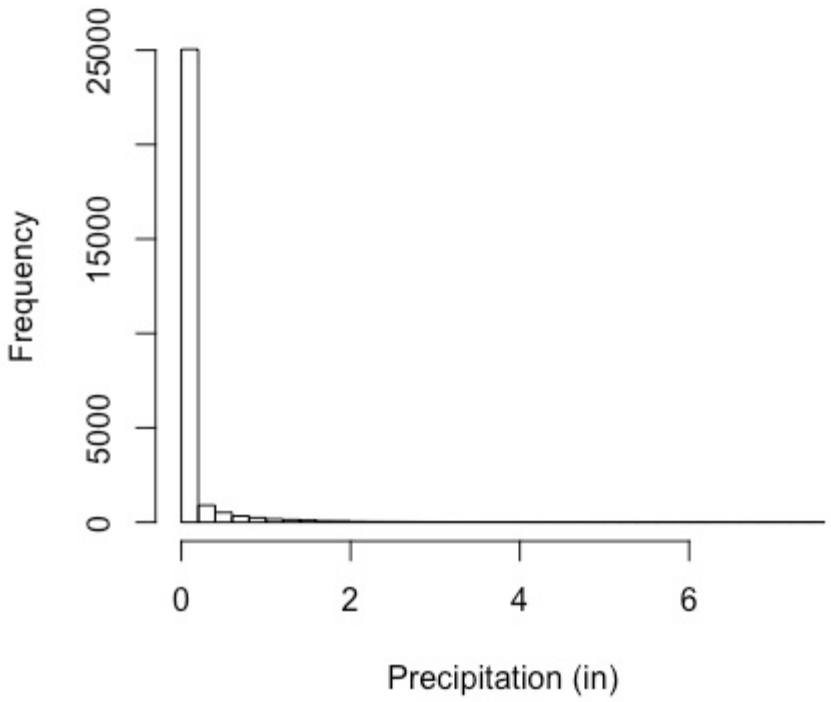

Figure 33 - Histogram of Central Texas Precipitation from 1940 to Present

It can be observed that the data is highly exponential. Figure 34 shows a scatter plot of daily rainfall and maximum temperatures for Central Texas covering 1940 to the present day. In reality, the precipitation data is exponentially distributed with the majority of the values being zero or no rainfall.

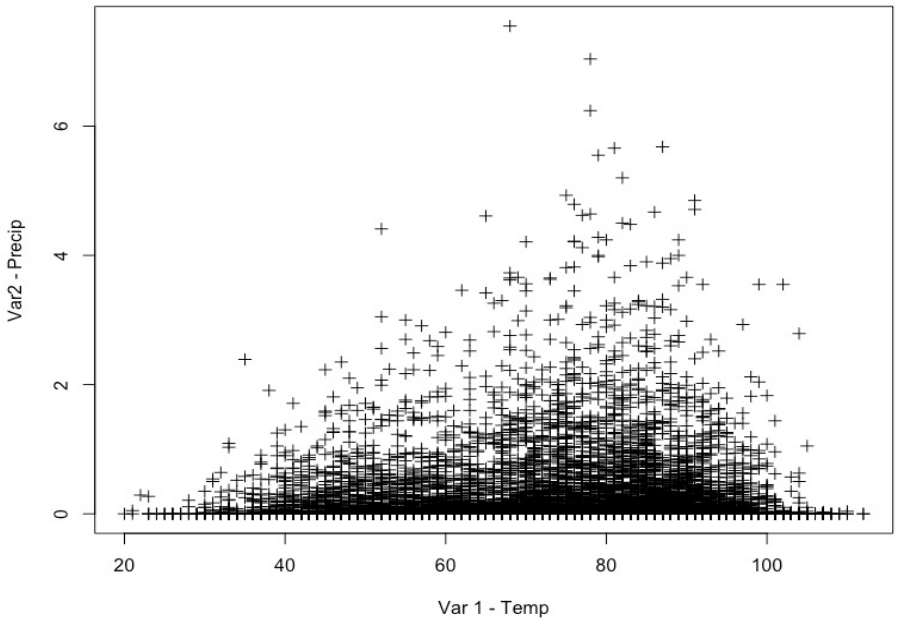

Figure 34- Max Temperature and Rainfall data for Central Texas 1940-Present

It can be observed that to the casual analyst, the data appears to be normally distributed with some extreme precipitation values that might be considered anomalies.

Figure 35 shows the same scatter plot with the points manually considered to be anomalies colored blue.

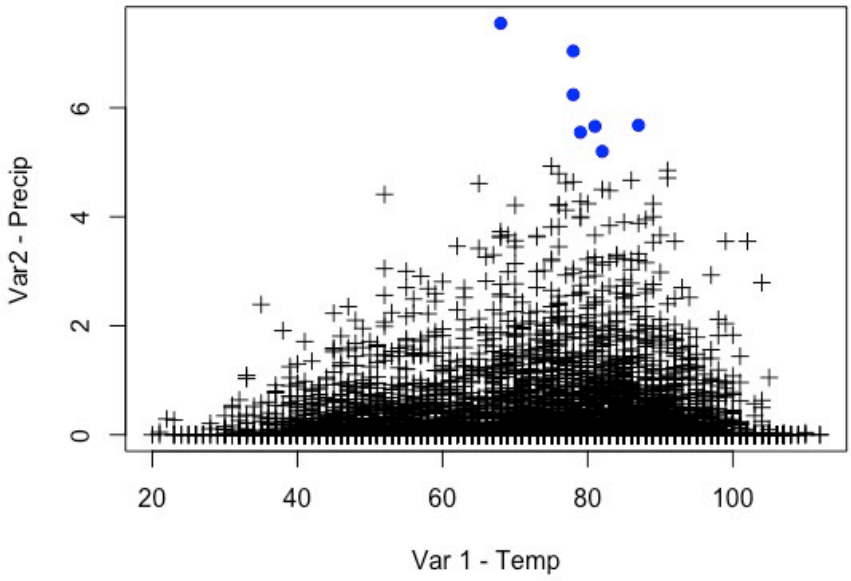

Figure 35 – Weather data with anomalies manually selected

It can be observed that these data points represent extremes of precipitation and that there are no temperature related points that stand out as extremes.

Figure 36 shows the same scatter plot with the smallest 12 FCAD points when the normal distribution is used for FCAD modeling highlighted in red.

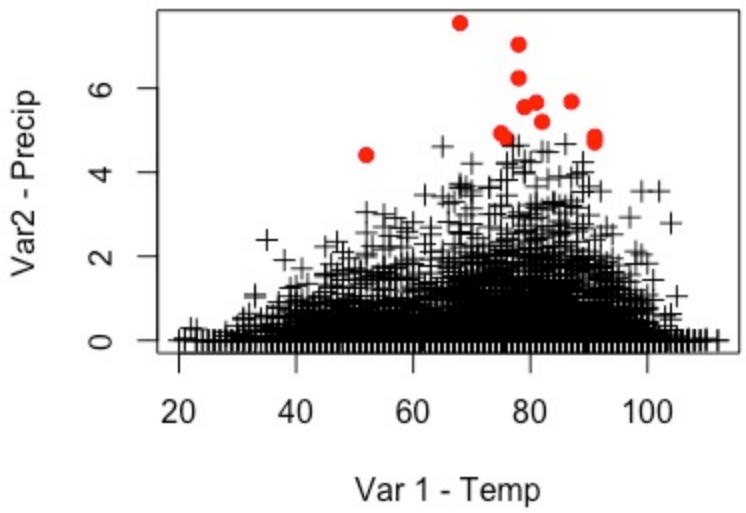

Figure 36 – FCAD anomalies identified using normal based modeling of weather data

It can be observed that all of the points that were manually selected as anomalies are selected by FCAD when using the normal distribution for modeling as well as additional distance based precipitation related points.

Figure 37 shows the same scatter plot with the smallest 12 FCAD points when the skew-normal distribution is used for FCAD modeling highlighted in red.

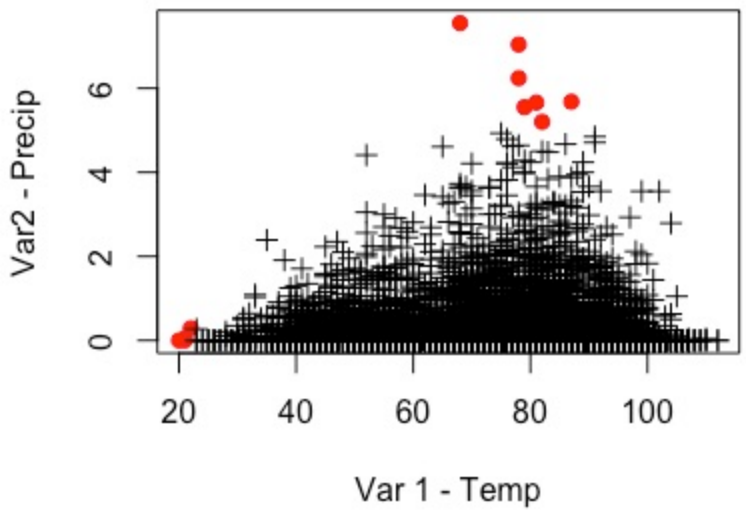

Figure 37 – FCAD anomalies identified using skew-normal modeling of weather data

It can be observed that additional points have been identified as anomalies when using the skew-normal distribution for modeling and it can be further observed that these points are all low temperature based points and are not distance based points. Normally, these points would not be selected by a human observer. All points that were manually selected as anomalies are identified when using the skew-normal distribution for modeling.

Table 9 shows all data points identified as anomaly points using the three methods of anomaly determination: manual, using the normal distribution for modeling and using the skew-normal distribution for modeling. Type 1 errors, points identified as anomalies not visually identified, are detailed.

Table 9- Identified Anomaly Points

Month	Max Temp	Precipitation	Manual Baseline	Normal	Skew Normal	Point Number
11	68	7.55	Yes	Yes	Yes	22600
9	78	7.04	Yes	Yes	Yes	25818
10	78	6.24	Yes	Yes	Yes	21475
8	87	5.68	Yes	Yes	Yes	19945
6	81	5.66	Yes	Yes	Yes	15138
5	79	5.55	Yes	Yes	Yes	14386
1	20	0	No	No	T1	8046
5	82	5.2	Yes	Yes	Yes	27539
1	21	0	No	No	T1	4048
1	21	0.05	No	No	T1	4049
1	22	0.29	No	No	T1	3318
1	52	4.41	No	T1	T1	18637
10	75	4.93	No	T1	No	27697
7	91	4.85	No	T1	No	14445
10	76	4.79	No	T1	No	27691
9	91	4.71	No	T1	No	12323

It can be observed that there are differences between the points identified as anomalies. No type two errors are present indicating that the FCAD modeling did as least as well as a human observer in anomaly identification.

Using both the normal and skew-normal distributions for modeling identified anomalies that were not identified manually. This can be explained based on the difference between a human visually observing a plotted data set in two dimensions and assuming that the data is normally distributed. Anomalies are visually identified based on being distance based outliers. The observer selected points are all exceptional in that they represent very large daily precipitation values. Both modeling methods calculated the total probability of a tuples context variables resulting in membership in a particular of indicator variable values. Manual inspection the plot identified seven of the points. Using both the normal distribution for modeling and skew-normal distribution for modeling identified five additional anomalies each. There was agreement on only one of these points. The remaining four points identified by using the normal distribution for modeling were all precipitation events. The four additional anomalies identified by using the skew-normal distribution represent temperature based anomalies where none are outliers. This implies that the ability to correctly incorporate the exponential nature of rainfall data results in a more complex selection of anomalies.

6. RESULT EVALUATION

This thesis demonstrates that an innovative context sensitive anomaly detection algorithm can be extended to use non-Gaussian probability distributions in the calculation of PDF values and enhance the ability to discover anomalies. Under the right circumstances, the skew-normal distribution was capable of providing enhanced detection performance when used with non-Gaussian data and had acceptable performance when used with normal data.

Modeling with the skew-normal distribution results in the majority of points in a data set being treated as less anomalous than when modeling with the normal distribution. Table 10 shows mean log FCAD values for Experiments 2, 3 and 4 considering the 1000 points out of 2000 with the largest log FCAD values and the 1400 points out of 2000 with the largest log FCAD values for both normal and skew-normal distribution based models. The mean log FCAD value for skew-normal is always larger than the mean value for normal distribution based modeling.

Table 10 – Mean log FCAD values for least anomalous points

Experiment	Mean FCAD Upper 50% Points	Mean FCAD Upper 70% Points
Experiment 2 - Normal	-2.62	-2.80
Experiment 2- Skew-Normal	-2.42	-2.63
Experiment 3 - Normal	-4.87	-5.15
Experiment 3 – Skew-Normal	-4.71	-4.90
Experiment 4 - Normal	-6.69	-7.14
Experiment 4 – Skew-Normal	-6.15	-6.74

Table 11 shows the means of log FCAD values for the 200 most anomalous points identified by each model.

Table 11 – Mean log FCAD values for most anomalous points

Experiment	Mean FCAD Lowest 200 Points
Experiment 2 - Normal	-5.6
Experiment 2- Skew-Normal	-6.0
Experiment 3 - Normal	-17.78
Experiment 3 – Skew-Normal	-19.98
Experiment 4 - Normal	-32.05
Experiment 4 – Skew-Normal	-35.84

In Table 11, it can be observed from the 200 most anomalous points from each experiment, that the mean log FCAD value when using the skew-normal distribution is smaller than the mean when using normal distribution based modeling. This suggests that skew-normal based modeling more positively identifies those points that are truly lower probability.

Both skew-normal and normal distribution based modeling appear to be effective at anomaly detection of clearly outlier anomalies and capture what would likely be identified by a human being. Finally, given all of the above, Experiment 5 strongly suggests that using skew–normal distribution based modeling provides a means of identifying interesting, important anomalies that previously would have gone unnoticed. Because of better data modeling afforded by the skew–normal distribution, and the application of context based probabilistic analysis, the unusual nature of the decline in temperature extrema became apparent.

7. CONCLUSIONS AND FUTURE WORK

The usefulness of FCAD in identifying anomalies in complex context driven data sets depends on the ability to correctly calculate the PDF of the multivariate mixture model representing the context and the indicator variables. Furthermore, in many areas of interest, exponentially biased data is relatively common.

In this thesis, it was shown that, when data that is normally distributed is modeled using the skew-normal distribution, there is little difference in the anomalies identified. In the case of data sets containing skewed data, the skew-normal distribution is able to take into account the skew and provide a PDF that more accurately reflected the data and therefore, under the right circumstances, allowed FCAD to more accurately identify low probability tuples.

It can be concluded that the context sensitive anomaly detection algorithm FCAD can be extended to utilize non-Gaussian probability distributions to better detect non-obvious anomalies in non-Gaussian data sets. In particular, the skew-normal distribution can be used to model both Gaussian and non-Gaussian data and provide greater flexibility. We have shown that the use of an alternative distribution to model non-Gaussian data leads to better anomaly detection when used with FCAD and has little negative impact when used with Gaussian data sets.

There are several avenues for future work including: 1) the skew-normal is part of a family of skew density functions that includes the skew-student-t, skew-slash, and skew-contaminated. Each of these distributions provides the potential for an even better fit to

specific data sets and could be explored in the same fashion. 2) Rewrite the extension of FCAD to analyze discrete data sets.

REFERENCES

1. N. Talleb, *The Black Swan*. New York: Random House, 2007.

2. V. Chandola, V. Banerjee and A. Kumar, "Anomaly Detection, A Survey," *ACM Computing Surveys*, vol. 41, no. 1, article 15, July 2009.

3. X. Song, M. Wu, C. Jermaine and S. Ranka, "Conditional Anomaly Detection," *IEEE Transactions on Knowledge and Data Engineering*, vol.19, no.5, 2007, pp.631–645.

4. Evans, Merran, N. Hastings, B. Peacock, *Statistical Distributions*., 3rd ed. New York: John Wiley & Sons, 2000

5. A. Azzalini, "A class of distributions which includes the normal ones,". *Scandinavian Journal of Statistics*, vol.12, pp.171–178, 1985.

6. K. Das. "Detecting Patterns of Anomalies," Ph.D. Dissertation, Carnegie Mellon University, Computer Science, Pittsburgh, PA, USA, 2009.

7. A. Azzalini and A Dalla Valle," The multivariate skew-normal distribution," *Biometrika*, vol. 83(4), pp.715–726, 1996.

8. R. Millar, *Maximum Likelihood Estimation and Inference: with Examples in R, SAS, and ADMB.* New York: John Wiley & Sons, 2011.

9. A. P. Dempster, N. M. Laird and D. B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society*, series 39(1), 1977, pp.1–38.

10. G. McLachlan, T. Krishnan, *The EM Algorithm and Extensions.*, 2nd ed. New York: John Wiley & Sons, 2008.

11. G. McLachlan, D. Peel, *Finite Mixture Models.* New York: John Wiley & Sons, 2000.

12.S. Lee, and G.J. McLachlan, “Finite mixtures of multivariate skew t-distributions: some recent and new results,” *Stat Comput.* vol.24, 2014

13. V.H. Lachos, P. Ghosh and R.B. Arellano-Valle, “Likelihood based inference for skew normal independent linear mixed models,” *Stat. Sin.* vol.20, pp. 303–322, 2010

14. M.O. Prates, C.R.B. Cabral, and V.C. Lachos, "mixsmsn: Fitting Finite Mixture of Scale Mixture of Skew-Normal Distributions," *Journal of Statistical Software*, vol.54, no.12, 2013.

15. C.R.B. Cabral, V.H. Lachos and M.O. Prates, “Multivariate Mixture Modeling Using Skew- normal Independent Distributions,” *Computational Statistics & Data Analysis*, vol.56, pp.126 – 142., 2012.

16.R. Vernic, “On the multivariate Skew-Normal distribution and its scale mixtures,” *An. St. Univ Ovidius Constanta*, vol. 13(2), pp.83-96, 2005.