A High Dimensional Mixture Model: A Mixture Multivariate Probabilistic Model

Mian Arif Shams Adnan
Bowling Green State University, Bowling Green, OH 43403

Abstract
A Mixture Multivariate Probabilistic Models has been found to be appropriate for different characteristics of fish species in fish market sales. This is a convexly combined mixture multivariate probabilistic model which is also relatively optimum.

Key Words: Density Plot, Squared Distance, Goodness of Fit Test, Maximum likelihood, Mixture Distribution, p-value.

1. Introduction

Different hallmarks of fish species in fish market like weight, height, various lengths are important features for the agro-based industries especially food industries, market places, scientists for bio-diversity, fisheries, etc. However, to develop and implement strategies that minimize the adverse effects of consequences of biodiversity, a more complete knowledge of how the joint distribution of the multiple variables or features of fish species are related or. We should know at what and how much rate the one feature is getting changed by the other feature. This model will help us to predict how quickly the various steps such as the suitable alternatives of steps (like what amount of anti-pesticides will be mixed to each water reservoirs) will be introduced that can maintain proper growth of fish species.

Generally, the joint distribution of weight, height, length follow multivariate normal distribution. So, Multivariate Distribution can be a useful model. Various mixture models had been proposed for the entire distribution function, simultaneously capturing the bulk of the distribution (typically the main mode) with the flexibility of a probabilistic model for the upper/lower tails. These mixture models either explicitly include the threshold as a parameter to be estimated, or somewhat bypass this choice by the use of smooth transition functions between the bulk and tail components, thus overcoming the issues of threshold choice and uncertainty estimation. Frigessi et al. (2002), Mendes and Lopes (2004) present mixture models that combine parametric form for the bulk distribution (e.g. Gamma, Weibull or Normal). The drawback with all the aforementioned approaches is the prior specification of a parametric model for the bulk of the distribution (and associated weight function where appropriate).

Tancredi et al. (2006) proposed a semi-parametric mixture model, A. MacDonald (2011) *et al* proposed a flexible model which includes a non-parametric smooth kernel density estimator below some threshold accompanied with the PP model for the upper tail above the threshold. A mixture of hybrid-Pareto has been carried by Carreau and Bengio (2009). Ciarlini *et al* (2004) have introduced the use of a probabilistic tool, a mixture of probability

distributions, to represent the overall population in such a temperature comparison. This super-population is defined by combining the local populations in given proportions. The mixture density function identifies the total data variability, and the key comparison reference value has a natural definition as the expectation value of this probability density.

Adnan *et al* (2021) developed bagging and boosting based convexly combined mixture probabilistic models for extreme temperatures. Attempts have been made here to develop a mixture multivariate probability model which is suitable for the fish species.

Section 2 describes the available statistical methodologies used to find a probabilistic model to be suitable. The concern data analysis is explained in section 3. Section 4 reveals the proposed mixture model for the aforementioned region. Final section draws the conclusion.

2. Methods and Methodologies

Various statistical methodologies applied here are explained below chronologically.

2.1 Multivariate Normal Distribution

The Multivariate Normal Distribution is widely used for modeling multivariate normal features. Let $\boldsymbol{X} = [X_1, X_2, \dots, X_p]$ be a p-variate multivariate normal variate that has the p-dimensional normal density function $N_p(\mu, \Sigma)$ of the form

$$f(\boldsymbol{x}) = \frac{1}{(2\pi)^{\frac{p}{2}}|\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)}, -\infty < x_i < \infty, \quad i = 1,2, \dots p, \quad (1)$$

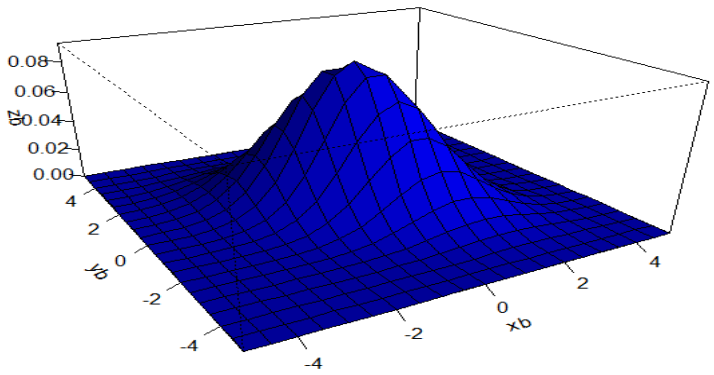

Figure-2.1 Graph of probability density function of s bivariate normal distribution.

2.2 Maximum Likelihood Estimation

Suppose $X_1, X_2, \dots, X_n$ denote n sample observations from a multivariate normal (mvn) distribution $N_p(\mu, \Sigma)$. The method of maximum likelihood is used to fit mvn distribution

(1) to these data. Assuming independence of the data, the likelihood is the product of the densities of extreme value distribution for the observations as

$$L(\mu, \Sigma) = \prod_{i=1}^{n} \frac{1}{(2\pi)^{\frac{p}{2}}|\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(x_i-\mu)'\Sigma^{-1}(x_i-\mu)}$$

The $\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$ and $\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})'$ are the mle estimators of μ and Σ.

2.3 Mixture Probabilistic Model

The infinite analogue (in which g is a density function) is $\int f(x;\ \theta)\ g(\theta)d\theta$; where θ is the parameter o the distributions, for the density function $f(x;\ \theta)$. The mixing distribution, say $g(\theta)$, is then a probability distribution on the parameter of the distribution $f(x;\ \theta)$. Several authors including Adnan (2009, 2010, 2021) worked on mixture distributions.

Let X be a multivariate random variable or vector taking values in sample space with the probability density function

$$g(\boldsymbol{x}) = \pi_1 f_1(\boldsymbol{x}) + \ldots + \pi_k f_k(\boldsymbol{x}), \quad x \sim N_p(\mu_i, \Sigma_i),$$

where $0 \leq \pi_i \leq 1,\ \ i = 1, \ldots, k,\ \ \pi_1 + \ldots + \pi_k = 1$.

Such a model can arise if one is sampling from a heterogeneous population that can be decomposed into k distinct homogeneous subpopulations, called *component populations*. If these components have been "*mixed*" together, and we measure only the variable X without determining the particular components, then this model holds. We say that X has a finite mixture distribution and that $g(\cdot)$ is a finite mixture density function. The parameters $\pi_1, \ldots, \pi_k$ are called *mixing weights* or *mixing proportions*, and each π_i represents the proportion of the total population in the i^{th} component. There is no requirement that the component densities should all belong to the same parametric family, but in this paper, we keep to the simplest case where $f_1(x), \ldots, f_k(x)$ have a common functional form but different parameters. Note that "mixed" distribution, which we defined in to be distributions with both continuous and discrete parts, are actually a special type of mixture. Indeed, if X is a mixed random variable,

$$f_X(x) = (1-k)f_C(x) + kf_D(x) \tag{2}$$

for some continuous random variable C, some discrete random variable D, and some number $k \in (0, 1)$. Hence, a mixed random variable is a discrete mixture of a continuous and a discrete random variable.

2.4 Goodness of Fit Test

The goodness of fit (GOF) tests measure the compatibility of a random sample with a theoretical probability distribution function. In other words, these tests show how well the distribution is selected fits to given data. That is, goodness of fit tests can be used to compare the fitted distributions, select one of the models, and determine how well it fits to data. In assessing whether a given distribution is suited to a data-set, the goodness of fit tests like Anderson Darling, Cramer-von Mises, Chi-square, Doornik-Hansen, Henze-Zirkler, Royston have been conducted. Wichitchan, S. et al (2021) has proposed a new goodness of fit test that observes several type I errors. Some state that the data came from a multivariate normal distribution and some decides the opposite. The following is the Multivariate Chi-square statistic (Voinov, V. et al, 2016) where the Dzhaparidze–Nikulin (DN) test statistic is mainly discussed.

$$D_p{}^{-1} = \frac{4}{1-2pr\sum d_i{}^2}\begin{pmatrix} \frac{(1-2(p-1)r\sum d_i{}^2)\sigma_{11}{}^2}{2} & r\sum d_i{}^2\sigma_{11}\sigma_{22} & \dots & r\sum d_i{}^2\sigma_{11}\sigma_{pp} \\ r\sum d_i{}^2\sigma_{22}\sigma_{11} & \dots & \dots & r\sum d_i{}^2\sigma_{pp}\sigma_{22} \\ \dots & \dots & \dots & \dots \\ r\sum d_i{}^2\sigma_{pp}\sigma_{11} & r\sum d_i{}^2\sigma_{pp}\sigma_{22} & \dots & \frac{(1-2(p-1)r\sum d_i{}^2)\sigma_{pp}{}^2}{2} \end{pmatrix}$$

$$\hat{B} = \frac{4}{1-2pr\sum d_i{}^2}\begin{pmatrix} \frac{d_1}{\sigma_{11}} & \frac{d_1}{\sigma_{22}} & \dots & \frac{d_1}{\sigma_{pp}} \\ \frac{d_2}{\sigma_{11}} & \dots & \dots & \frac{d_2}{\sigma_{11}} \\ \dots & & & \dots \\ \frac{d_r}{\sigma_{11}} & \frac{d_r}{\sigma_{22}} & \dots & \frac{d_r}{\sigma_{pp}} \end{pmatrix}$$

$$\hat{B}D_p{}^{-1}\hat{B}' = \frac{2pr}{1-2pr\sum d_i{}^2}\begin{pmatrix} d_1{}^2 & d_1d_2 & \dots & d_1d_p \\ d_2d_1 & \cdots & & d_2d_p \\ \cdots & \dots & & \dots \\ d_pd_1 & d_pd_2 & \dots & d_p{}^2 \end{pmatrix}$$

$$V'\hat{B}D_p{}^{-1}\hat{B}'\mathrm{V} = \frac{2pr(\sum V_i d_i)^2}{1-2pr\sum d_i{}^2}$$

$$U_n{}^2 = V'(\hat{\theta})\left[I - \hat{B}(\hat{B}'\hat{B})^{-1}\hat{B}'\right]\mathrm{V}(\hat{\theta})$$

where $U_n{}^2$ is the Dzhaparidze–Nikulin (DN) test statistic which is distributed as Chi-square test. However, for the simplicity, average squared distance between observed frequencies and expected frequencies has been found to measure the discrepancy between data based observed frequencies (nf) and theoretical frequencies (ne).

$$\text{Average Squared Discrepancy} = \frac{1}{n}(nf-ne)'(nf-ne).$$

3. Analysis of the Data

The information gathered from the features of the fish species is described in section 3.1. Section 3.2 describes about the determination of the multivariate normal distribution. QQ plot, Dzhaparidze–Nikulin (DN) Chi-square goodness of fit test of the different distributions are shown in section 3.3, 3.4 and 3.5 respectively. In section 3.6 and 3.7 represent the contour plot and density plot of the mixture multivariate normal distribution.

3.1 Information gathered from the fish species of sales data

The data of the fish species of sales has been collected from Kaggle (https://www.kaggle.com/aungpyaeap/fish-market). The data set has 159 observations and 7 variables. The variables' information and summary statistics are given as below.

```
'data.frame':	159 obs. of  7 variables:
 $ ï..Species: chr  "Bream" "Bream" "Bream" "Bream" ...
 $ Weight    : num  242 290 340 363 430 450 500 390 450 500 ...
 $ Length1   : num  23.2 24 23.9 26.3 26.5 26.8 26.8 27.6 27.6 28.5 ...
```

```
$ Length2   : num  25.4 26.3 26.5 29 29 29.7 29.7 30 30 30.7 ...
$ Length3   : num  30 31.2 31.1 33.5 34 34.7 34.5 35 35.1 36.2 ...
$ Height    : num  11.5 12.5 12.4 12.7 12.4 ...
$ Width     : num  4.02 4.31 4.7 4.46 5.13 ...
```

From the following figure, it is observed that there are trends among all variables.

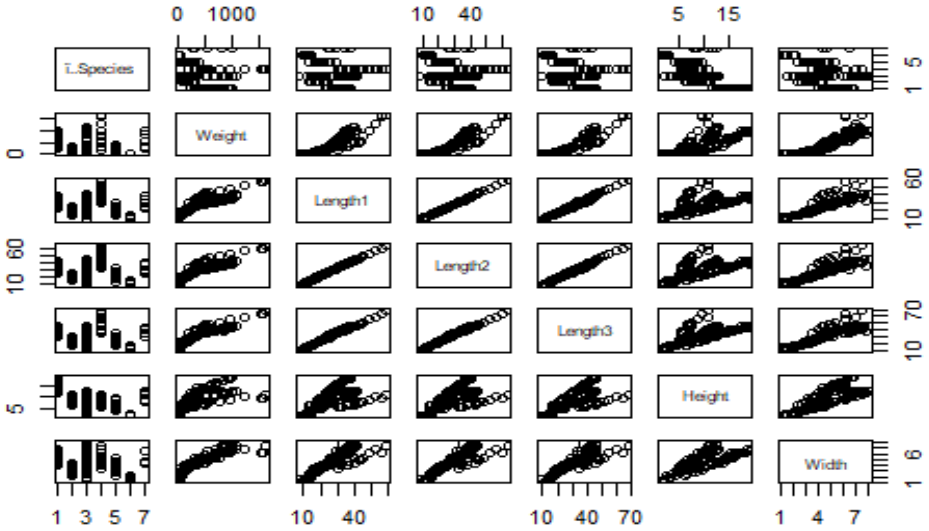

Figure 2: Scatterplot of the variables of the fish species sales.

3.2 Determining the distribution for the data

To search the appropriate probability distribution for fish-species data, the goodness of fit tests like Dzhaparidze–Nikulin (DN) Chi-square goodness of fit tests, Q-Q plot, contour plot, density plot, etc have been observed. For the sake of easy visualization, only 73 observations and two variables (weight and length1) have been considered. The scatter plot of these two variables is shown below.

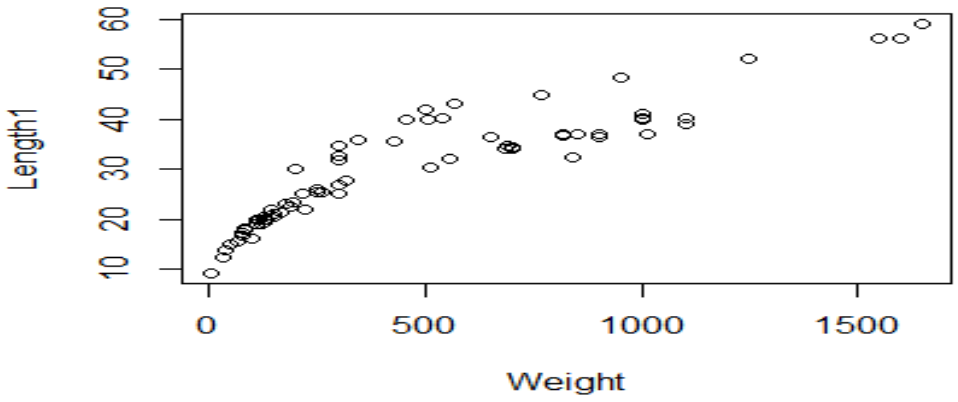

Figure 3: Scatterplot between Weight and Length1.

Maximum likelihood estimated parameters of multivariate normal distribution are given in the following table.

Table 1: Estimated parameters of Multivariate Data.

Parameters	*Estimates*
Mean Vector (μ)	$\begin{pmatrix}460.6055\\29.6589\end{pmatrix}$
Variance Covariance Matrix (S)	$\begin{pmatrix}167063 & 4151\\4151 & 124\end{pmatrix}$

The density function of Multivariate Normal fails to fit the data along with the following density according to the Chi-square test with p-value (0.055).

$$f(\boldsymbol{x}) = \frac{1}{2\pi\left|\begin{pmatrix}167063 & 4151\\4151 & 124\end{pmatrix}\right|^{\frac{1}{2}}} e^{-\frac{1}{2}\left(\begin{pmatrix}Weight\\Length1\end{pmatrix}-\begin{pmatrix}460.6055\\29.6589\end{pmatrix}\right)' \begin{pmatrix}167063 & 4151\\4151 & 124\end{pmatrix}^{-1}\left(\begin{pmatrix}Weight\\Length1\end{pmatrix}-\begin{pmatrix}460.6055\\29.6589\end{pmatrix}\right)},$$

$$i = 1,2,\dots p.\ 0 < Weight < 1650, 7.50 < Length1 < 59.$$

3.3 Quantile-Quantile (Q-Q) plot

Following Figure represent the Q-Q plot of the sorted values (in ascending order) of the variables weight and length1 is determined by $y_i = F^{-1}\left(\frac{i-0.5}{n+1}\right)$. Figure 3(a) checks whether the Q-Q plots follow univariate normal distributions.

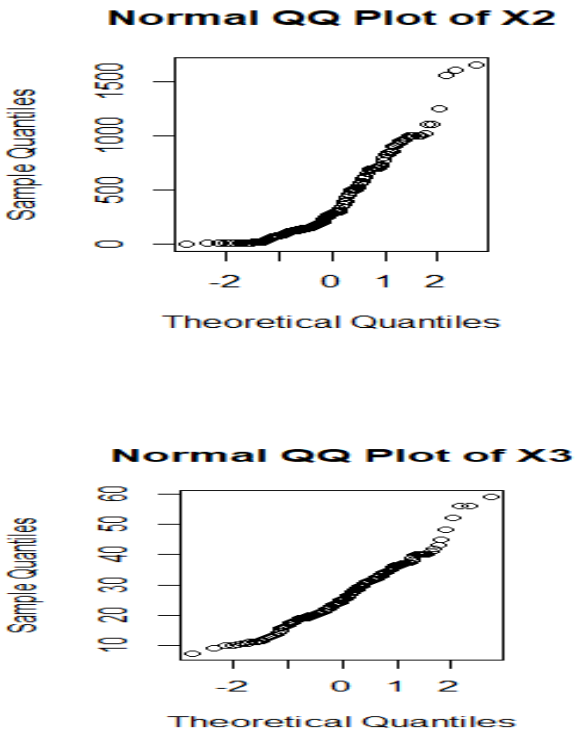

Figure 4: Q-Q plot for Normal distribution.

This plot shows that the points fall approximately along with 45^0 reference line. We therefore conclude that the individual variable of these data follows univariate normal distribution.

3.4 Chi-square Test

Let the null and alternative hypothesis are H_0: Data follows multivariate normal distribution and H_A: Data does not follow multivariate normal distribution. The calculation of Chi-square statistic on the basis of 73 observations under the postulated models is performed. So, we use the estimated parameter $\hat{\mu} = \begin{pmatrix} 460.6055 \\ 29.6589 \end{pmatrix}$, and $S = \begin{pmatrix} 167063 & 4151 \\ 4151 & 124 \end{pmatrix}$ to perform this test. The estimated value of chi-square statistic on the basis of 51 years of maximum temperature data under the postulated models are shown in the following Table 6.

Table 2: Calculation for Chi-square goodness of fit test

Distribution	*Statistic* (χ^2)	*p-value*
Extreme Value distribution	7.6	0.055

So, the null hypothesis is not rejected at 5% level of significance. Therefore, the data follows Multivariate Normal distribution.

3.6 Probability Plot

The Probability plot of the sorted values (in ascending order) of the observed data has been found in the following figure. We observe that these data have roughly normal shape distribution.

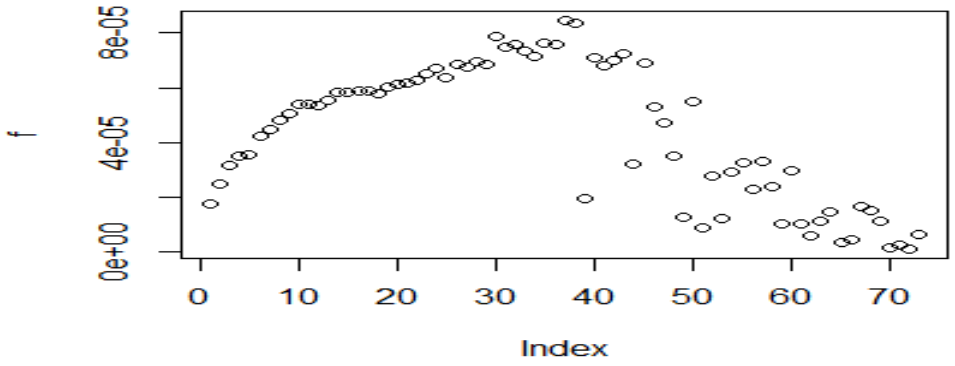

Figure 5: Empirical Probability plot for data.

3.7 Bivariate plots

In Figure 5, density plot also shows that data plots are approximately close to the original density. We therefore conclude that data follows bivariate normal distribution.

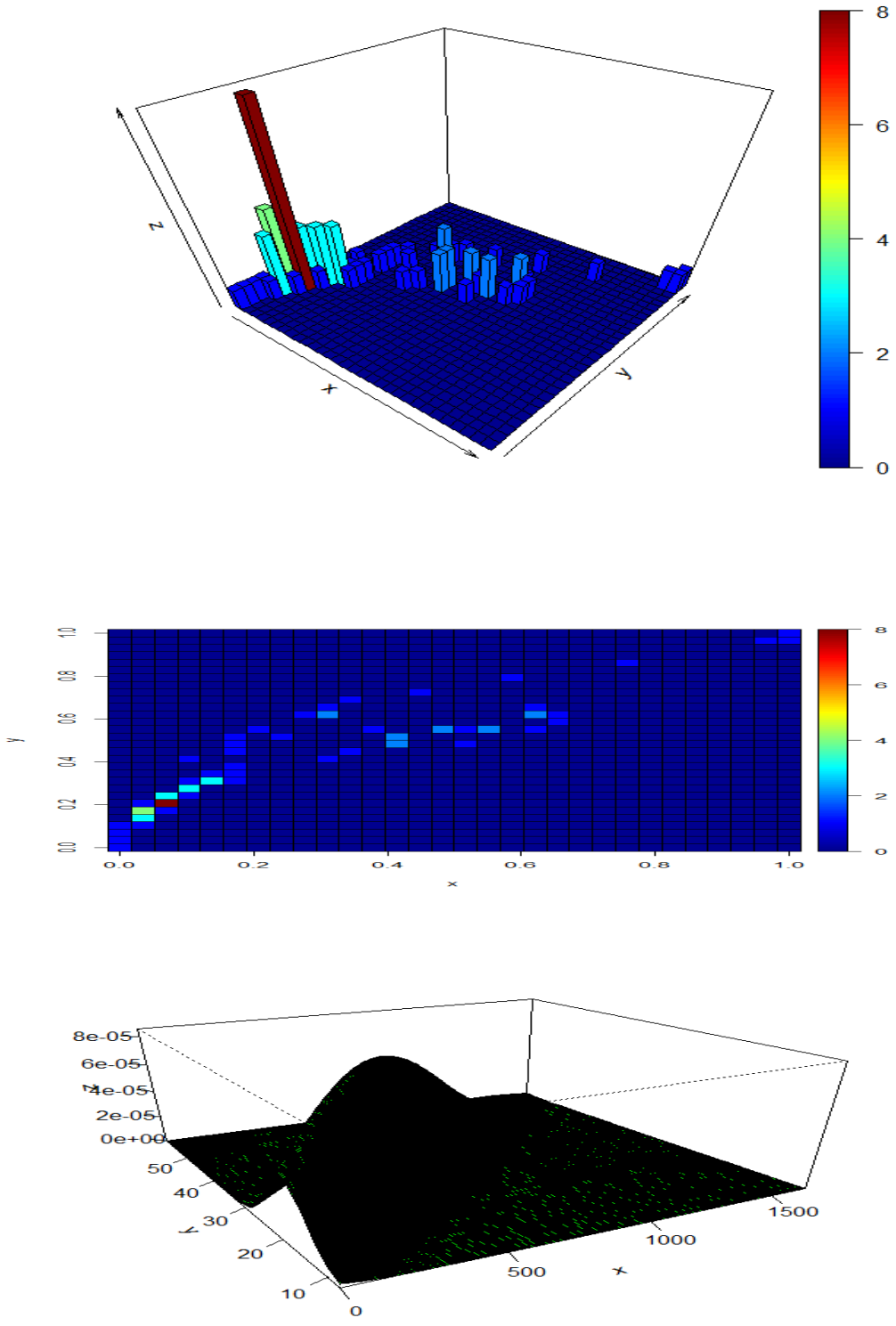

Figure 6: Bivariate histogram, countour points plot and Density plot for the data.

By observing the bivariate plots, it seems that there is more than one mean vector (one near $\binom{460.6055}{29.6589}$ and another near $\binom{5100.6055}{29.6589}$) in the observed bivariate normal distribution for the data. So, bimodality of the fish species indicates us to impose the mixture of two multivariate normal distribution in order to identify the better mixture multivariate normal distribution for the observed data.

4. Proposed mixture extreme value distribution

The distribution of the bivariate normal distribution of weight and leanth1, $f(x)$, is estimated by a mixture of two distributions having same probability distribution but different parameters. Mixture bivariate normal distribution can be formed with weights $(1-p)$ and p (where, p refers p-value). If we get the higher p-value for the goodness of fit test in case of the mixture bivariate normal distribution, and that p-value is greater than those of the other distributions, then we can say that the mixture bivariate normal distribution is the best probabilistic model for the observed data. So, the mixture model of the bivariate normal distribution with weights $(1-p)$ and p is given (Adnan *et al*,2011) as of the following form

$$f(x) = (1-p) * f_1(x) + p * f_{2'}(x)$$

where, $f_1(x)$ is the density function of the one bivariate normal distribution with the estimated value of mean vector parameter and $f_2(x)$ is the density function of another bivariate normal distribution with the estimated value of different mean vector parameter.

Table 3: Calculation for Chi-square goodness of fit test.

Distribution	*Squared Distance Between Observed and Expected Frequency*
Mixture multivariate normal distribution	8.700009e-07

Therefore, probability density function of mixtured extreme value distribution is given by

$$f(x) = (1-0.0551237) * \frac{1}{2\pi \left|\begin{pmatrix} 167063 & 4151 \\ 4151 & 124 \end{pmatrix}\right|^{\frac{1}{2}}} e^{-\frac{1}{2}\left(\binom{Weight}{Length1}-\binom{460.6055}{29.6589}\right)' \begin{pmatrix} 167063 & 4151 \\ 4151 & 124 \end{pmatrix}^{-1} \left(\binom{Weight}{Length1}-\binom{460.6055}{29.6589}\right)} + 0.0551237 * \frac{1}{2\pi \left|\begin{pmatrix} 167063 & 4151 \\ 4151 & 124 \end{pmatrix}\right|^{\frac{1}{2}}} e^{-\frac{1}{2}\left(\binom{Weight}{Length1}-\binom{460.6055}{29.6589}\right)' \begin{pmatrix} 2*167063 & 2*4151 \\ 2*4151 & 2*124 \end{pmatrix}^{-1} \left(\binom{Weight}{Length1}-\binom{5100.6055}{29.6589}\right)}.$$

Now, let us consider now the null and alternative hypothesis will be H_0: Data follows mixture bvn distribution against H_A: H_0 is not true.

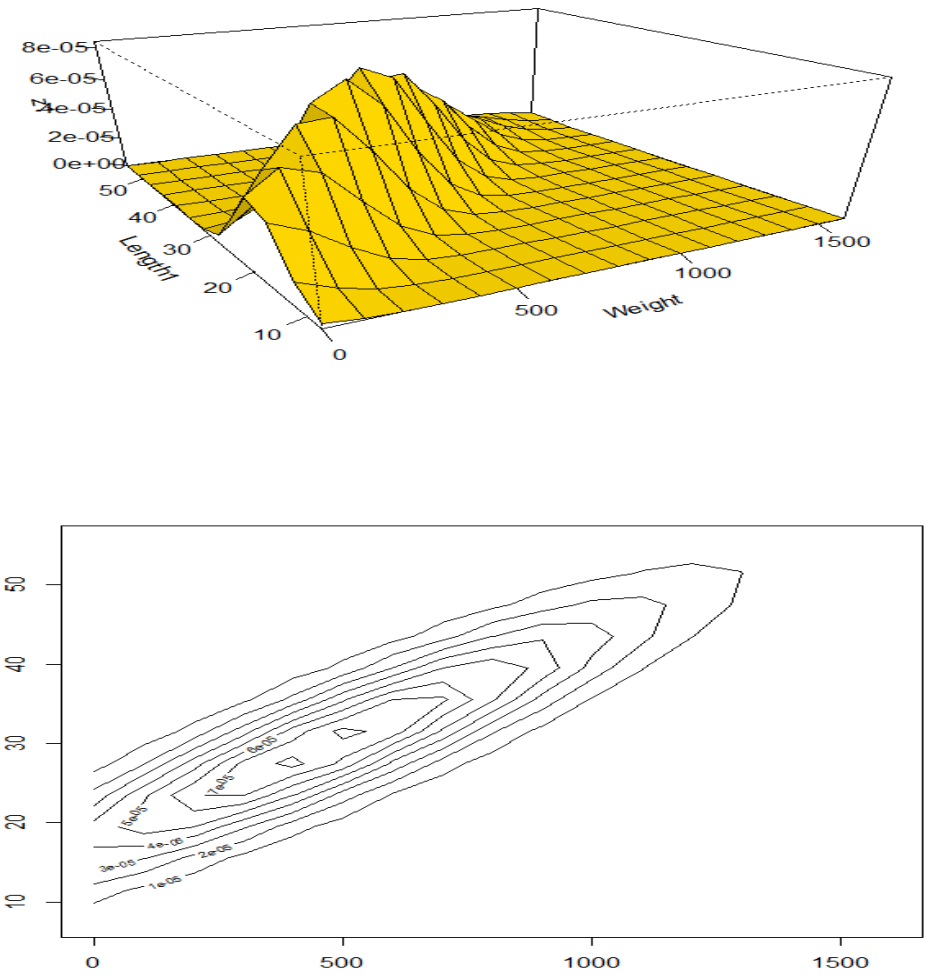

Figure 7: Mixture Bivariate density plot, countour plot for the data.

Therefore, the final mixture model for temperatures of the aforementioned region is as follows:

$$f(x) = (1 - 0.0551237) * \frac{1}{2\pi \left|\begin{pmatrix} 167063 & 4151 \\ 4151 & 124 \end{pmatrix}\right|^{\frac{1}{2}}} e^{-\frac{1}{2}\left(\begin{pmatrix} Weight \\ Length1 \end{pmatrix} - \begin{pmatrix} 460.6055 \\ 29.6589 \end{pmatrix}\right)' \begin{pmatrix} 167063 & 4151 \\ 4151 & 124 \end{pmatrix}^{-1} \left(\begin{pmatrix} Weight \\ Length1 \end{pmatrix} - \begin{pmatrix} 460.6055 \\ 29.6589 \end{pmatrix}\right)}$$

$$+ 0.0551237 * \frac{1}{2\pi \left|\begin{pmatrix} 167063 & 4151 \\ 4151 & 124 \end{pmatrix}\right|^{\frac{1}{2}}} e^{-\frac{1}{2}\left(\begin{pmatrix} Weight \\ Length1 \end{pmatrix} - \begin{pmatrix} 460.6055 \\ 29.6589 \end{pmatrix}\right)' \begin{pmatrix} 2*167063 & 2*4151 \\ 2*4151 & 2*124 \end{pmatrix}^{-1} \left(\begin{pmatrix} Weight \\ Length1 \end{pmatrix} - \begin{pmatrix} 5100.6055 \\ 29.6589 \end{pmatrix}\right)}.$$

; 0 ≤ weight ≤ 1650, 7.5 ≤ length1 ≤ 59.

The pdf of the following part is given as below.

$$0.0551237 * \frac{1}{2\pi\left|\begin{pmatrix}167063 & 4151\\ 4151 & 124\end{pmatrix}\right|^{\frac{1}{2}}} e^{-\frac{1}{2}\left(\begin{pmatrix}Weight\\ Length1\end{pmatrix}-\begin{pmatrix}460.6055\\ 29.6589\end{pmatrix}\right)' \begin{pmatrix}2*167063 & 2*4151\\ 2*4151 & 2*124\end{pmatrix}^{-1}\left(\begin{pmatrix}Weight\\ Length1\end{pmatrix}-\begin{pmatrix}5100.6055\\ 29.6589\end{pmatrix}\right)}.$$

; 0 ≤ weight ≤ 1650, 7.5 ≤ length1 ≤ 59.

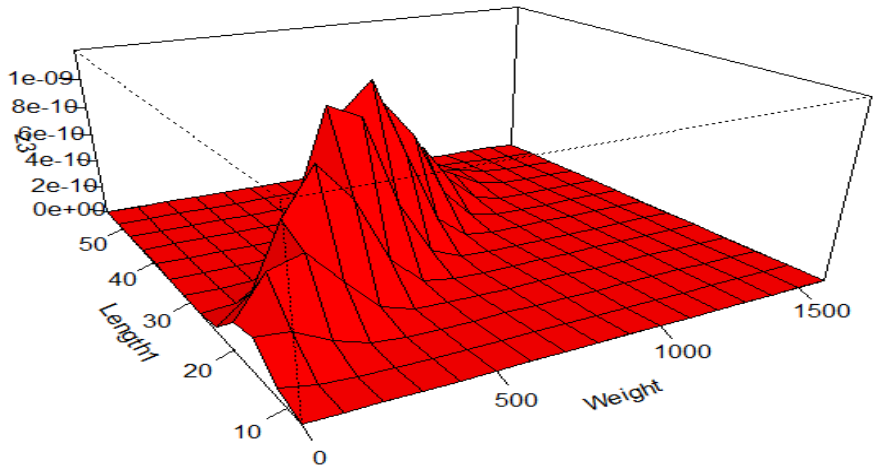

Figure 8: Bivariate density plot, countour plot for the second mixing compnent for the data.

Conclusion

Optimum Mixture Probabilistic Model based several convexly combined multivariate distributions can be found.

References

Adnan, M. A. S. and Mahmud, H. M. M. (2021). A Bagging and Boosting Based Convexly Combined Optimum Mixture Probabilistic Model. arXiv:2106.05840.

Adnan, M.A.S. and Mahmud, M., (2011). A Mixture Probabilistic Model for Extreme Temperatures. In *JSM Proceedings*, Statistics and the Environment Section. Alexandria, VA: American Statistical Association. 2728-2739.

Carreau, J. and Bengio, Y. (2009). A hybrid pareto mixture for conditional asymmetric fat-tailed distributions. IEEE Trans Neutral Netw. 20(7):1087-101. Epub: doi: 10.1109/TNN.2009.2016339.

Ciarlini, P., Cox, M., Pavese, F. and Regoliosi, G. (2004). The Use of a Mixture of Probability Distributions in Temperature Interlaboratory Comparisons. Metrologia. 41 93). 116 – 121.

Frigessi, A., Haug, O., Rue, H., (2002). A dynamic mixture model for unsupervised tail

estimation without threshold selection. Extremes. **5 (3)**, 219–235.

MacDonald, A., Scarrott, C.J., Lee, D., Darlow, B., Reale, M. and Russell, G. (2011) A Flexible Extreme Value Mixture Model. Computational Statistics and Data Analysis. 55 (2137 – 2157).

Mendes, B. V. M. and Lopes, H. F. (2004). Data driven estimates for mixtures. Computational Statistics and Data Analysis. 47 (583 - 598). doi:10.1016/j.csda.2003.12.006.

Tancredi, A., Anderson, C., O'Hagan, A., (2006). Accounting for threshold uncertainty in extreme value estimation. Extremes. **9 (2)**, 87–106.

Voinov, V. *et al* (2016). New invariant and consistent chi-squared type goodness-of-fit tests for multivariate normality and a related comparative simulation study. Communications in Statistics - Theory and Methods, 45:11, 3249-3263, DOI: 10.1080/03610926.2014.901370.

Wichitchan, S. *et al* (2021). A new class of multivariate goodness of fit tests for multivariate normal mixtures. Communications in Statistics - Simulation and Computation, DOI: 10.1080/03610918.2020.180868.