Seminar Report - Fitting Mixture of Distributions with R package mixdist

Nik Počuča - 1322223

January 19, 2019

Chapter 1

Introduction

1.1 Peter Macdonald

Dr. Peter Macdonald, professor emeritus from McMaster Unviersity dives into his lifetime research and life experiences working with finite mixture models in fishery-length frequency analysis. Dr. Macdonald served on numerous United States Environmental Protection Agency FIFRA Scientific Advisory Panels, reviewing methodology proposed by the EPA for pesticide risk assessment. Dr. Macdnolad presents the pike measurements taken in 1965 in that of Heming lake and demonstrates how mixture modelling can be used to track length of pike of fish. This report consists of a methodological introduction in the subsequent sections, an estimation methodology detailing the outline of parameter estimation. An application section also considers the use of the mixdist package.

1.2 Finite Mixture Model

A finite mixture model is a model in which multiple distributions play a role on the sample sample space. These models often arise when sampling from heterogeneous populations with a specific probability density function on each partition, For example in frequency distributions of animal populations with various age-groups. A finite mixture has a finite number of groups contributing to the density that lives on sample space Ω.

Definition 1.2.1. Finite mixture model Suppose a random variable $\boldsymbol{X}$ takes on values in a sample space Ω in which it's distribution can be represented by a probability density function (continuous case) of the form

$$p(x;\boldsymbol{\theta}) = \sum_{k=1}^{G} \pi_k f_k(x;\boldsymbol{\theta_k}).$$

Here $\boldsymbol{x}$ are realizations of $\boldsymbol{X}$, π_k is the mixing proportion of group k, where $0 \leq \pi_k \leq 1, k = 1, \ldots, G$ and $\sum_{k=1}^{G} \pi_k = 1$. $f_k(\boldsymbol{x};\boldsymbol{\theta_k})$ is the probability density function of partition k and $\boldsymbol{\theta_k}$ as a parameter vector for that group k. Subsequently f_k is usually taken to be the same for all mixtures thus $f_k = f \forall k$. $p(x;\boldsymbol{\theta})$ can be thought of as a convex combination of densities with mixture portions summing to one. For extensions and specifics pertaining to identifiability see Everitt (1985), Smith and Makov (1995). Other implementations and formulations of finite mixture models including parsimonious models can be found in McNicholas (2015).

1.3 Gaussian Mixture Modelling

For the purposes of modelling size of fish, samples were taken at random where measurements of fish length showed symmetry. Provided a Gaussian distribution as the assumption for the density. The function for $f_k(x;\boldsymbol{\theta}_k)$ is defined as :

$$f_k(x;\boldsymbol{\theta}_k) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(- \frac{(x-\mu_k)^2}{2\sigma_k^2} \right)$$

where $\boldsymbol{\theta}_k = (\mu_k, \sigma_k)$. $\mu_k \in \mathbb{R}$ is the mean size of the fish and $0 < \sigma \in \mathbb{R}^+$ is the standard deviation of fish lengths. During sampling it was noted that the small sized fish were considered to be under-sampled. Hence, for the smallest of the groups, the assumption is that of a truncation. Gaussian mixture modelling is assumed to be symmetric in univariate space. Within this symmetric assumption there are various tests for symmetry involving likelihood ratios but for now only the implementation of the model will be considered (Garel, 2001).

Chapter 2

Estimation Methodology

2.1 Maximum Likelihood Estimation

There are several ways to estimate the coefficients of a finite mixture model. The best methods are using the maximum likelihood estimation (MLE). The incomplete data Likelihood is defined as

$$L(\boldsymbol{\theta}; \boldsymbol{x}) = \prod_{i=1}^{n} p(x_i; \theta) = \prod_{i=1}^{n} \sum_{k=1}^{G} \pi_k f_k(x; \theta_k).$$

Within the incomplete data likelihood, membership of observations are unknown, thus estimating the correct mixing proportions and the parameters of the density is difficult.

Bibliography

Garel, B., 2001. Likelihood ratio test for univariate gaussian mixture. Journal of Statistical Planning and Inference 96 (2), 325 – 350.

URL http://www.sciencedirect.com/science/article/pii/S0378375800002160

McNicholas, P. D., 2015. Mixture Model-Based Classification. Chapman and Hall.