Seminar Report - Fitting Mixture of Distributions with R package mixdist

Nik Počuča - 1322223

February 5, 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 $0 < \sigma_k \in \mathbb{R}^+$ 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). Gaussian mixture modelling assumes that tail probabilities are very low and does not account for extreme values. This appropriate for the fish setting as rare do we see extremely large sized fish in nature.

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 as we do not know parameters π_k for all k. Often we derive a latent indicator variable Z_{ik} where $Z_{ik} = 1$ when observation i belongs to group k and $Z_{ik} = 0$ otherwise. The likelihood can be rewritten as the complete data likelihood as

$$L(\boldsymbol{\theta}; \boldsymbol{x}) = \prod_{i=1}^{n} \prod_{k=1}^{G} (f(x; \theta_k))^{z_{ik}}.$$

Here z_{ik} is a realization of Z_{ik}. From this complete data likelihood representation maximization is difficult. Maximization is usually performed on the log-likelihood due to nice asymptotic properties defined as follows:

$$l(\boldsymbol{\theta}; \boldsymbol{x}) = \log(\prod_{i=1}^{n} \prod_{k=1}^{G} (\pi_k f(x; \theta_k))^{z_{ik}}) = \sum_{i=1}^{n} \sum_{k=1}^{G} z_{ik}[log(\pi_k) + log(f(x; \theta_k))]$$

To find maximum likelihood estimators the expectation-maximization (EM) algorithm provides convergence very quickly (Dempster et al., 1977).

2.2 Expectation Maximization Algorithm

The EM algorithm is an iterative technique which alternates between the expectation of the likelihood (E-step) and the maximization step (M-step) of the parameters in the likelihood. Convergence is discussed in Dempster et al. (1977). The E-step is expressed by taking the expectation of the latent variable Z_{ik} within the likelihood. While the M-step is expressed by taking derivatives with respect to each parameter μ_k and σ_g.

2.2.1 E-Step

The expectation step is performed by taking the expectation of the latent variable within the complete data log-likelihood. Since z_{ik} is a binomial variable it's expectation is a posteriori calculation with probability $f(x;\theta_k)$ of being selected. Thus the E-step is calculated as

$$\tau_{ik} = \mathbb{E}[z_{ik}|l(\boldsymbol{\theta};\boldsymbol{x_i})] = \frac{\pi_k f(x_i;\theta_k)}{\sum_{g=1}^{G}\pi_g f(x_i;\theta_g)}$$

Where τ_{ik} is the posteriori probability that observation i belongs to group k. Now given the posteriori we wish to maximize parameters μ_k, and σ_k accordingly .

Bibliography

Dempster, A. P., Laird, N. M., Rubin, D. B., 1977. Maximum likelihood from incomplete data via the EM-algorithm. Journal of the Royal Statistical Society B 39, 1–38.

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.