em: A Generic Function of the EM Algorithm for Finite Mixture Models in R

Dongjie Wu
Københavns Universitet

Abstract

Our *em* package follows R's feature of generic functions and the function `em()` can be implemented after a model fitting with one component using R's pre-existing functions and packages such as `glm()`, `lm()`, and so on.

Currently, it supports the following models: linear models (`lm()`), generalized linear models (`glm()`), generalized non-linear model (`gnm()`), survival models (mainly conditional logistic regression) (`survival::clogit()`), multinomial models(`nnet::multinom()`).

Keywords: R, the EM Algorithm, finite mixture models, model based clustering, latent class regression.

1. Introduction

Finite mixture models (FMMs) are widely applied in both natural science and social science to model unobserved heterogeneity[1]. This modelling technique usually assumes that data can be divided into unobserved groups known as latent classes, each following a distinct probability density or a regression model with its unique model parameter.

One reason for the extensive usage of finite mixture modelling is its flexibility. Its core assumption of unobserved heterogeneity can be examined on a variety of models and analysis including generalized linear regression models (GLMs), generalized non-linear regression models (GNMs), survival analysis, and etc. It can also be used on a model-based clustering technique and adopts different data structure, for example, categorical data, multidimensional data and hierarchical data(Vermunt 2008). In addition, FMMs can adopt parameterized probabilities for the latent class (i.e. a concomitant model(Wedel 2002)) or extend the model structure to a nested or hierarchical structure(Vermunt and Magidson 2005).

In general, FMMs can be estimated by the following four methods: Methods of moments (MoM), Maximum log-likelihood estimation (MLE), the Bayesian methods, and the unsupervised machine learning methods. MLE using *Expectation-maximization* (EM) algorithm is the mainstream method of estimating FMM models due to its performance and accuracy. However, it is relatively computationally intensive.

There are many packages or software available for FMM models: e.g. *fmm*(Deb 2007) and *gllamm*(Rabe-Hesketh, Skrondal, and Pickles 2004) in Stata, *flexmix*(Leisch 2004), *mixtools*(Benaglia, Chauveau, Hunter, and Young 2010) and *mclust*(Scrucca, Fop, Murphy, and

[1]See McLachlan, Lee, and Rathnayake (2019) for an introduction of finite mixture models

Raftery 2016) in R, etc. We make our own package *em* for estimating FMM using EM because packages mentioned above did not cover the case for our own research.

In addition, we adopt a framework based on generic functions in R, which integrates better with other functions and packages in R and makes implementing FMM models more flexible and straightforward.

In section 2, we specify FMMs and their extended variations. In section 3, we present EM algorithms and the approaches we use to fit FMMs. In section 4, we introduce the *em* package and the generic *em* function for fitting FMMs in R. Section 5 demonstrates some examples of using *em*. Section 6 provides a short summary of the paper.

2. Finite Mixture Models

2.1. The Finite Mixture Models

Finite mixture models can be described in the following equations given J components:

$$f(\mathbf{y}|\mathbf{x}, \phi) = \sum_{j=1}^{J} \pi_j (f_j(\mathbf{y}|\mathbf{x}, \theta_j)) \tag{1}$$

where $\sum_{j=1}^{J} \pi_j = 1$, $\mathbf{y}$ is a dependent variable with conditional density f, $\mathbf{x}$ is a vector of independent variables, π_j is the prior probability of component j, θ_j is the component specific coefficients for the density function f_j.

The model $f(\mathbf{y}|\mathbf{x}, \theta_j)$ can be one from a wide range of models: probability distributions, generalized linear models (GLM), generalized non-linear models (GNM), survival models, categorical models, etc.

2.2. The Concomitant Models

2.3. The Hierarchical Mixture Models

3. Fitting Finite Mixture Models

3.1. The EM Algorithm and its extensions

3.2. The Mix Likelihood Function and the Complete-Data Likelihood Function

3.3. Starting Values

4. Using the Generic em Function

5. Examples

6. Summary

Acknowledgments

This research was supported by ... (ERC) under grant ... ().

References

Benaglia T, Chauveau D, Hunter DR, Young DS (2010). "mixtools: an R package for analyzing mixture models." *Journal of statistical software*, **32**, 1–29.

Deb P (2007). "FMM: Stata module to estimate finite mixture models." Statistical Software Components, Boston College Department of Economics. URL https://ideas.repec.org/c/boc/bocode/s456895.html.

Leisch F (2004). "Flexmix: A general framework for finite mixture models and latent glass regression in R."

McLachlan GJ, Lee SX, Rathnayake SI (2019). "Finite mixture models." *Annual review of statistics and its application*, **6**, 355–378.

Rabe-Hesketh S, Skrondal A, Pickles A (2004). "GLLAMM manual."

Scrucca L, Fop M, Murphy TB, Raftery AE (2016). "mclust 5: clustering, classification and density estimation using Gaussian finite mixture models." *The R journal*, **8**(1), 289.

Vermunt JK (2008). "Latent class and finite mixture models for multilevel data sets." *Statistical methods in medical research*, **17**(1), 33–51.

Vermunt JK, Magidson J (2005). "Hierarchical mixture models for nested data structures." In *Classification—The ubiquitous challenge*, pp. 240–247. Springer.

Wedel M (2002). "Concomitant variables in finite mixture models." *Statistica Neerlandica*, **56**(3), 362–375.

Affiliation:

Dongjie Wu
Sociologisk Institut

Københavns Universitet
Øster Farimagsgade 5
1353 Copenhagen
E-mail: dongjie.wu@soc.ku.dk