The EM Algorithm for Latent Class Analysis

Jeroen K. Vermunt
Department of Methodology and Statistics, Tilburg University
www.jeroenvermunt.nl

Introduction

- The Expectation Maximum (EM) algorithm is a very popular algorithm for maximum likelihood estimation with missing data (or latent variables).

- The article from Dempster, Laird, and Rubin from 1977 is one of the most cited articles in statistics (63407 cites in Google Scholar).

- EM is ideal for LC analysis. It is very simple to implement and, moreover, very stable (it is guaranteed to increase the LL value every iteration).

- In this video, I will explain how EM works in LC analysis.

The EM algorithm for a simple LC model

- E-step:
  - Compute the expected value of the missing data given the observed data and the "current" parameter values.
  - For a LC model this involves computing the $P(X = c|y_1, \dots, y_J)$ using the "current" parameter estimates.

- M-step:
  - Re-estimate the model probabilities $P(X = c)$ and $P(y_j|X = c)$ treating class membership as if were observed and using the $P(X = c|y_1, \dots, y_J)$ as "weights".
  - This involves obtaining the *J* 2-way class-indicator "observed" frequency tables and transforming these to probabilities.

- Multiple iterations of the E-step and M-step are performed till convergence, or till ready to switch to Newton-Raphson.

Two illustrations

- I created a spreadsheet which allows me to perform EM iterations “manually” with the antireli.dat data set.

- I will run Latent GOLD a single iteration at a time with the GSS82.sav data (3 class model), and show how the Profile changes and the LL value increases at every iteration step.