Stat 243 Section: Simulation Study

Eugene Yedvabny

November 10, 2014

What are the goals of the study and the metrics to assess the method?

The study by Chen *et al* is proposing a new likelihood test for testing the null hypothesis regarding the *order* of a normal mixture model. In English, the study claims to provide a rigorous test for whether a **proposed** order a normal mixture (the number of Normal distribution components in the mixture) is statistically significant given a particular mixture.

The proposed method is a variation of Expectation-Maximization with an imposed penalty for variance that is set by a tunable parameter. The quality of the tuning and the overall EM test is vetted by a simulation study. The primary metric is Type I error, e.g. the false positive rejection of the null hypothesis. The error rate is measured at 1% and 5% α significance across 5000 replications, 2 different sizes (n=200 and n=400), 12 unique models, and two different mixture orders (m=2 and m=3). The tests a further verified on 8 alternative models with 1000 repetitions for each model. The EM test is considered robust and successful because its Type I error is close to target and the power grows with sample size.

What choices did the authors make in the design of the study? What are the key aspects of the data generating mechanism?

There is a bit of a disconnect between the claims of the method's prowess and the testing limitations. The simulation only considers orders of 2 and 3, which is a far cry from the *any order* claimed by the algorithm. There is also no reason given for the choices of the number of models, the sample size, or number of repetitions. **Table 3** provides the parameters for the models but again, no justification for their selection is given. The simulation design involves drawing 5000 samples of size 200 or 400 from 24 models (12 x m=2 & m=3) to assess the Type I rate. The alternative models (**Table 4**) are sampled 1000 times each, which in itself is another arbitrary number.

The patterns of the 12-model parameters indicate that they were systematically chosen to represent different ratios of means, variances, and proportions for mixture components. There is no randomization of the parameters, which could potentially lead to undiscovered algorithm-breaking configurations. Futhermore, this is selection is certainly not exhaustive, but no metrics are given as to why only 12 models were chosen and how, if at all, additional parameters affect the results.

Possible design alternatives? How would you leverage the principles of basic experimental design?

The four key aspects of good experimental design are: Efficiency, Reporting of Uncertainty, Reproducibility, Documentation.

Do the figures present the results well? Do the authors address the issues of uncertainty and reproducibility?

Interpret Tables 4 & 6. Do the results make sense?

Does the study follow JASA's guidelines on simulation studies?