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 dataset.

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 with a power rating 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 nears 100% for larger sample sizes.

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.

Assuming that the twelve models represent what the researchers consider "typical" scenarios, there is still the underlying assumption of normality behind the mixture. The samples are drawn from a specified distribution, so there is no stochastic noise to the model. None of the real-life dataset are perfectly normal unless actually analyzed across thousands of iid samplings, so there is no metric for robustness towards deviations from normality. Furthermore there is no variance in the reported data on Figures 1+2 or Tables 4+6. Each model has one power rating or one Type I error rate, based off the 1000 and 5000 replications respectively, and there is no repeated test for the same model.

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. While there is no way to test the efficiency of this simulation given the lack of implementation details, the study is rather defficient on the other three principles as well. There is no reporting of uncertainty *in Type I error or power rates.* These two are themselves measurements of uncertainty, so the questions is a bit convoluted. We see variability across models, but no one model is tested twice, thus there is no data on how reproduceable or variables these measures of robustness really are. There is ample documentation on the *method* but little on the *simulation*; perhaps the raw R code has some additional notes and information to help recreate the simulation.

The variation in power across the twelve models suggests that they capture enough of the sample space to properly probe the algorithms. I would prefer to see some more models with extremes (overlapping or fully-separated distributions) and a larger selection of mixture orders. What the presentation is really missing, however, is the connection between the models and the Type I error performance. Rather than seeing the box plots, it would be nice to have the calculated error matched to the model that yielded it.

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

The main take-away from the figures is that the median error rate is comparable to the desired α rate, which could have been accomplished with a single point rather than a box plot. Since each 5000-iteration sampling can only yield a single value (the rejection rate), each box plot must only have twelve values corresponding to the twelve models. A figure that takes up half a page really should not encode a single value. There really is no talk about variability or reproduceablility since we don't know which model yielded which part of the box plot, and the models were not tested more than once.

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

Tables 4 and 6 are the power measures for different models under the *wrong* hypothesis that m=2. The higher the power, the more often the model was correctly rejected. The tables reveal that the more overlapped the distributions of the mixture, the harder it is for the EM test to reject the hypothesis. This is consistent with intuition, but rather surprising just *how poorly* the test performs at these extremes.

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

Having not read the Supplementary Information, there is very little detail in the paper itself regarding the specifics of the simulation. The authors provide the implementation of the EM test in R and document the overall testing procedure, but there is no information regarding the RNG seed or the specifics of the sampling procedure. As it stands, the *simulation* aspect of the paper is not up to JASA's guidelines.