Problem 1

(a)

As indicated explictly in the paper, the purpose of the simulation study is to assess the accuracy of the proposed asymptotic approximation in finite samples and to examine the power of the EM test.

The key metrics that the authors consider in the simulation study are the test model order, null/alternative models, sample sizes and number of simulation repitions.

(b)

The first choice is the test model order, in the study of this paper, normal mixtures with order of 2 and 3 are chosen to demonstrate the EM test statistic. The next, and probabaly the most important, choices are the null/alternative model selections, i.e. the normal mixture parameter specifications (α, θ, σ and their different combinations). The final choices are the significance levels (i.e. 1% and 5%), sample sizes (i.e. 200 and 400)and number of simulation repititions (i.e. 1000 and 5000) for this simulation study.

As for the data generating mechanism, the key aspects that will likely affect the statistical power of the EM test should be the RNG of normal mixtures. Not many details are given in the paper, but the RNG of normal mixtures are not straightforward, which should be clarified in a better way. It could have siginificant impact on the final results if naive RNG yields a lot of overlaps and reductions.

(c)

There are some data-generating scenarios missing in their simulation study. First, for order-2 normal mixture models, the parameter specifications have missed out (θ_1, θ_2) with both positive or negative numbers, i.e. the current 3 levels of (θ_1, θ_2) are all pairs of positive/negative numbers. Similar scenarios exist with order-3 normal mixture models.

Second, the alternative models are designed intentionally to mimic those null models, which though maybe close to reality, are missing some non-ideal scenarios of alternative models with large discrepency to null models.

(d)

In order to test the hypothesis, e.g. $H_0 : m_0 = 2$ vs. $H_A : m_0 > 2$, the design space is quite large. We could use the principles of basic experiemental design to set up the simulation study, but it would be difficult to cover the whole space. As shown in the current setup of this paper, null models are just really a few samples from the huge design space, only 2 (α_1, α_2) levels, 3 (θ_1, θ_2) levels and 2 (σ_1, σ_2) levels are considered to compose a full factorial design of $2 \times 3 \times 2 = 12$ levels. With the continuous selections of α, θ, σ numbers, level choices that cover wider design space will yield a huge number of full-factorial design levels.

However, it is possible to increase the random sample size levels from 2 (i.e. 200,400) to more (e.g. 200, 400, 600). And this will give a much clearer trend of EM test power scaling with sample sizes.

(e)

The tables in the paper have done a fairly good job in representating EM test power under different models and sample sizes. However, there are inconsistencies between Table 4 and 6, as θ values split in Table 4 while α values split in Table 6. There is not much explanation in the text and cause some confusion on the reader's side.

As for the figures, the representation could be better if the box plots are composed together with different labeling and grouping, so that the readers can have more comparisons among different combinations.

At last, from the original text in the paper, there is no discussion on the issue of simulation uncertainty or standard errors. It is not so convincing to the readers that the authors have done enough simulation replications. The choice on the number of simulation repititions is rather arbitrary and lacks supporting evidence.

(f)

From the interpretation of the results shown in Tables 4 and 6, we can speculate that the EM test power is positively correlated with the random sample size, since all the numbers in the $n = 400$ column are larger than those in the $n = 200$ column. This shows that the data generating mechanism DOES have large impact on the final results and should be analyzed more in the discussion and representation of the results.

(g)

The JASA's guidelines on simulation studies: "Results Based on Computation - Papers reporting results based on computation should provide enough information so that readers can evaluate the quality of the results. Such information includes estimated accuracy of results, as well as descriptions of pseudorandom-number generators, numerical algorithms, computers, programming languages, and major software components that were used."

The authors fail to provide enough info suggested in the JASA's guidelines on the simulation study in the paper. We do not see descriptions of pseudorandom-number generators and computers performance specifications that were used.