Project Proposal **Stat 495**

While this is framed as a project proposal, it's been laid out below in a framework to assist you in figuring out what you need to do to satisfy the project requirements. To get started, answer the six questions below. Then, you'll be taking your responses and using them to craft a proposal/synopsis (last page).

1. What is the new statistical topic you will be exploring for the project? If you have a large topic area in mind (like "clustering" or "imputation"), you should work (with me!) to narrow it down to one or several sub-topics (particular methods) within the area.

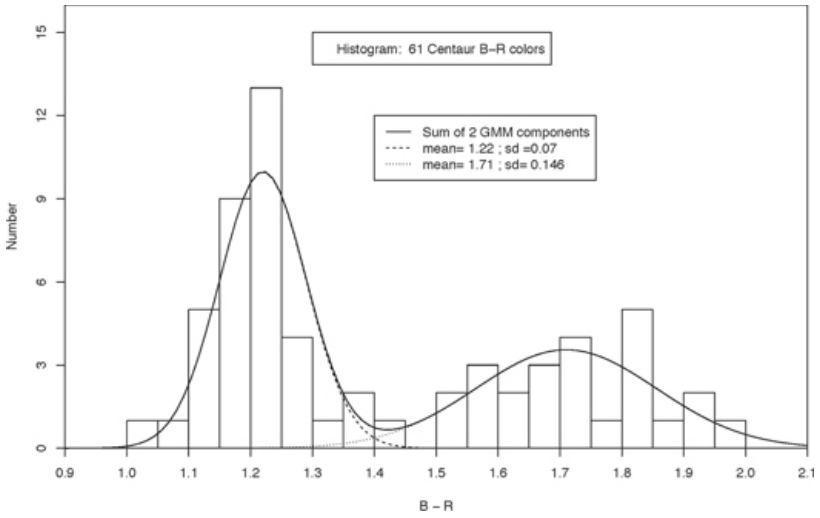
   For the project, the new statistical topic I will be exploring is Gaussian mixture models (GMM).

2. Consider where you have encountered this topic. The audience for the paper is a peer in the class, who won't necessarily have had the same electives/background experiences as you. Besides details about the new statistical topic, what other topics will your peers need information about in order to understand the topic? (For example, if you want to study a particular method of imputation, you are going to need to describe the types of missing data and imputation methods in general in order for it to be understandable to a classmate).

   I encountered this topic in a statistical paper that I found for a STAT-225 homework assignment, where the researchers used a GMM to decompose their Centaur (asteroid-like bodies in space that revolve around the Sun in the outer solar system) sample into two normal distributions. Since mixture models are a generalization of k-means, it would be helpful lay a solid foundation of clustering before working to understand how GMM's take the analysis method one step further. I would do this for the expectation-maximization (EM) algorithm as well. Also, just for this proposal, I've seen a visual or two of GMM's, and rather than confidence intervals, the technique generates confidence ellipsoids for multivariate models. So I would have to recall the confidence interval, but then expand to higher dimensional confidence metrics. I would give a review of the Gaussian distribution and maybe go a bit deeper into covariance, since we only lightly used the concept in our 300-level STAT classes.

3. Are you currently leaning more towards an application or simulations for the required second component? Provide a bit of your thinking about why this is your current thinking. You may or may not have very specific ideas here yet.

   I am currently leaning more towards an application for the second component. The paper in which I first came across GMM showed the figure below to illustrate how they applied the model to the comet data. The researchers did not go into the inner workings and implementation of GMM's at all, and I'd of course be doing those things, but I'd also love to find data on an interesting topic and specifically seek out data sets that exhibit multimodality. This would take some thought on what kinds of real studies might result in observations that follow multiple Normal distributions, and some wrangling to clean the data before I could draw any distributional patterns out.



4. If you have any other thoughts on planned applications or particular things you'd want to do in your simulations, feel free to list those here.

5. The first major project component is an expository review / exposition (an extremely solid Literature / Background / Introduction to the topic section, whatever name you give it) of the new topic. This is likely to include additional information about the general statistical topic area, and example applications in the literature. Thinking of your statistical and writing skills (and goals you have set for class), what do you expect will be the most challenging aspect of preparing this exposition for you? (Expositions may have a code component too, with toy examples.) What skills will this section help you demonstrate?

   Based on my statistical and writing experience, I would say that the reading part will be the most difficult for me, i.e., collecting enough outside material to form a complete picture in my mind, not just reading one relevant paper and calling it a day. I don't think I have too much problem understanding and implementing the math behind statistical material (and seeing connections to the conceptual/structural groundwork of other statistical concepts is cool too!). I have more issue staying focused on the important information surrounding the equations that give context and relationship to specific math. I will also be challenged by not purely regurgitating what I read, but really understanding it enough to use it myself and to make bridges between papers that only cover specific topics. This section will help me demonstrate my critical thinking, synthesis, and organizational skills.

6. The application / simulation (or both, etc.) that you include will mean some coding (in R or another software). What do you expect will be the most challenging aspect of preparing this section and reporting the findings from your work? (This includes coding, statistical, and writing skills). What skills will this section help you demonstrate?

For the application component, I expect that the most challenging aspect will be wrangling the data both to clean it for more effective analysis and to get it into a format usable for GMM. I have not looked into any GMM packages in R yet, but if the command is one simple line, I may feel that I am doing insufficient work. I hope to find a very engaging data set that will push me to in terms of wrangling, analysis implementation, and interpretation. This section will help me demonstrate coding and communication skills, and I would love to thoroughly understand all of the components in whichever GMM package I use.

Now, let's synthesize all of this into a complete proposal via making it a project synopsis – e.g. you lay out what you plan to do. Here is an example assuming that random forests were the topic (this is not "new" for any of you though!), so I can demonstrate how all of these responses (from the previous pages) go together. See if you can see how each sentence relates to one of the questions above. Space for you to write your own synopsis (which is what I'll focus on reading in this submission) is below the example.

Example Proposed Project Title: Using Random Forests to Predict Company Bankruptcy

Example Project Synopsis:

For my project, I plan to explore the use of random forests in supervised learning (classification) problems. My paper will introduce my peers to the classification setting, review the concept of decision trees in order to motivate forests, and then delve into the details of random forests, as well as touch on the pre-cursor of bagging. I will demonstrate how random forests work with a toy example such as predicting Species in iris in the exposition to help make points clear for my peers. I will also include several examples of applications of random forests in the literature to demonstrate recent use of the method. For my application, I intend to apply random forests to a larger data set (potentially sourced from Kaggle) with a binary response that indicates whether a company is bankrupt or not. I intend to use a new R package for fitting the random forest, called ranger. Wrangling the data set will let me demonstrate appropriate wrangling skills and reporting my results will show my ability to communicate my data analysis process. I will include a comparison to logistic regression to showcase my modeling and variable selection skills in a regression setting, before comparing the models based on performance.

Note: In the synopsis, I tried to write about intentions (if some things were uncertain) or write about what I "will" do. This helps provide a road map for you – you know you need to do those things. Your synopsis may not have some of these details, but I am hoping writing it this way helps you envision what the final product is and how you can work to get there. Sharing intention is useful! You can write that you intend to find a data set that allows you to demonstrate skills X, Y, and Z, or that you intend to write a simulation to demonstrate A and B, etc.

Your turn

Your name: Cassandra Jin

Proposed Project Title: Using Gaussian Mixture Models to Deconstruct Data With Unusual Distributions

Project Synopsis:

For the project, the new statistical topic I will be exploring is Gaussian mixture models (GMM). I encountered this topic in a statistical paper that I found for a STAT-225 homework assignment, where the researchers used a GMM to decompose their Centaur (asteroid-like bodies in space that revolve around the Sun in the outer solar system) sample into two normal distributions. Since mixture models are a generalization of k-means, I will lay a foundation of clustering before working to understand how GMM's take the analysis method one step further. I will do this for the expectation-maximization (EM) algorithm as well. Also, I will recall the confidence interval and then expand to higher dimensional confidence metrics. Lastly, I will review the Gaussian distribution and go a bit deeper into covariance, since we only lightly used the concept in our 300-level STAT classes. I will also include applications of the GMM, as the paper in which I first came across GMM illustrated how they applied the model to the comet data. I will find data on an interesting topic and specifically seek out data sets that exhibit multimodality. This will take some thought on what kinds of real studies might result in observations that follow multiple Normal distributions, and some wrangling to clean the data before I can draw any distributional patterns out. Based on my statistical and writing experience, I expect that the reading component will be the most difficult for me, i.e., collecting enough outside material to form a complete picture in my mind, not just reading one relevant paper and calling it a day. I will need to stay focused on the important information surrounding the equations that give context and relationship to specific math. I will also be challenged by not purely regurgitating what I read, but understanding it enough to use it myself and to make bridges between papers that only cover specific topics. For the application component, I expect that the most challenging aspect will be wrangling the data both to clean it for more effective analysis and to get it into a format usable for GMM. I have not investigated any GMM packages in R yet, but if the command is one simple line, I may feel that I am doing insufficient work. I hope to find a very engaging data set that will push me to in terms of wrangling, analysis implementation, and interpretation, and through the process, I would love to thoroughly understand all of the components in whichever GMM package I use. These two sections will help me demonstrate my critical thinking, synthesis, organizational, wrangling, coding, and communication skills.