Solution: GMM Clustering With Cross Validation

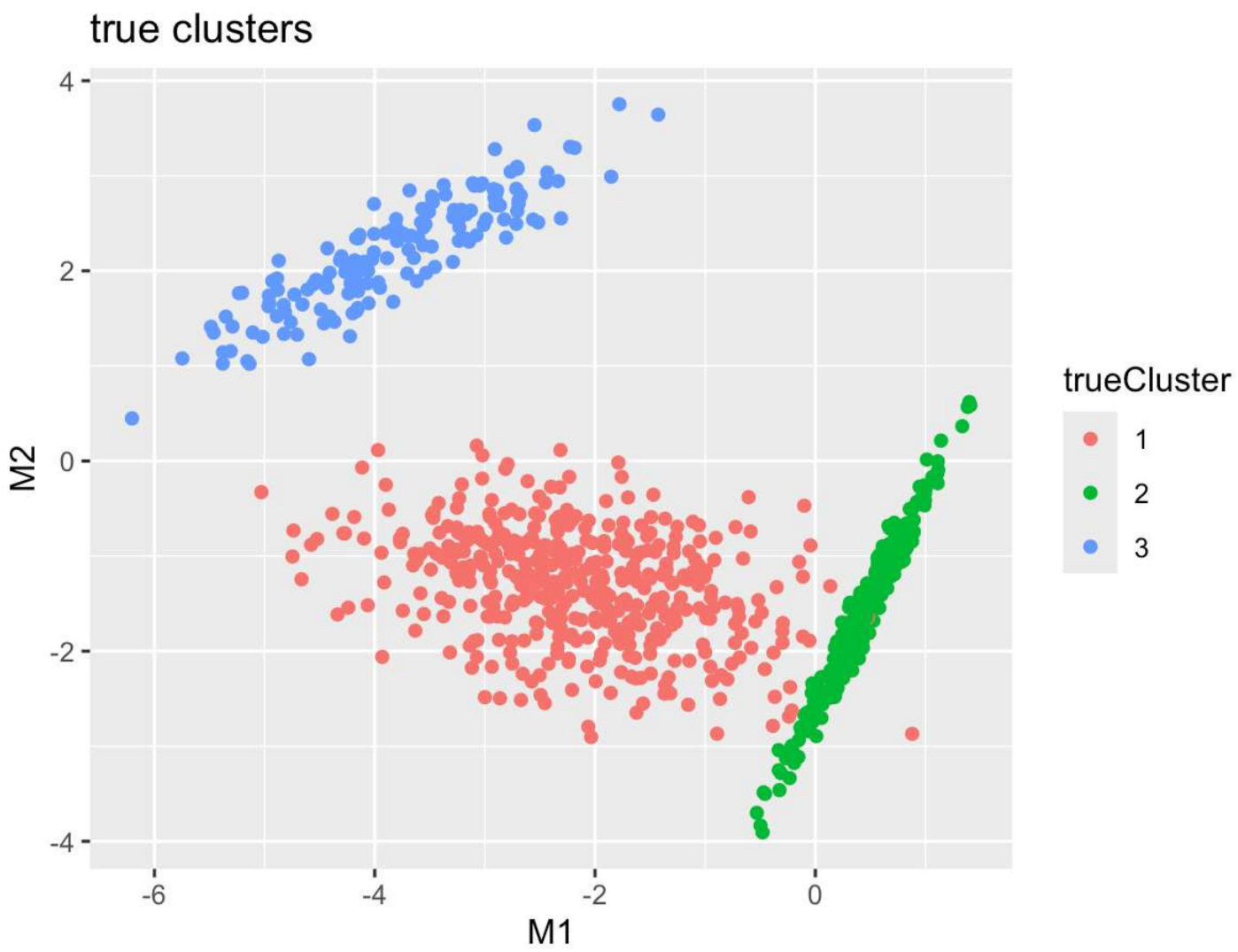
Directions: Please upload your solutions to the following questions before the deadline.

In this homework you will cluster numerical measurements using the Gaussian Mixture Model method. Specifically you will use cross-validation to select the best model and compare the selected model to the measurements.

1.) (0 points) Please follow the linked file **EM.R (https://northeastern.instructure.com/courses/193161/files/30835292?wrap=1)** **(https://northeastern.instructure.com/courses/193161/files/30835292/download?download_frd=1)** for an R script to guide you in this assignment. You may use this script as a basis for porting to your favorite IDE, or may use it directly in Rstudio. Also, follow the links to **measurementPoint.csv (https://northeastern.instructure.com/courses/193161/files/30835307?wrap=1)** **(https://northeastern.instructure.com/courses/193161/files/30835307/download?download_frd=1)** and **trueClass.csv (https://northeastern.instructure.com/courses/193161/files/30835311?wrap=1)** **(https://northeastern.instructure.com/courses/193161/files/30835311/download?download_frd=1)** to download the data for this assignment.

2.) (20 points) Provide a scatter plot of the 2D feature vectors, and indicate their true cluster assignment by color. Provide a legend for this coloring, label the horizontal axis "M1", the vertical axis "M2", and title the plot "true clusters". Provide answers to the following questions. Provide all the reasons why this data set might be challenging for any clustering algorithm. Be as specific as you can.

The scatter plot follows.



There are three reasons why this data set might prove challenging to some clustering algorithms. First, the clusters are not spherical, but rather elongated. Second, the clusters overlap and are not linearly separable. Third, each cluster differs in density, size, shape and rotation.

3.) (10 points) You will now create 10 "folds" of the feature vectors by indexing each measurement with an integer from 1 to 10, inclusive. Make this indexing random and uniformly distributed on these integers. Each vector should now belong to exactly one fold. Retain this fold index for the remainder of the assignment. Please answer the following question. Specify the number of vectors in each of your folds.

I expect a random count for each fold, an iid set of binomial random variable with success probability 1/10 and 990 trials. So, the average fold count should be 99 and the variance should be 89 (standard deviation = 9ish). Of course, the sample averages will be close to these numbers.

4.). (0 points) We will now consider a standard method to estimate the number of folds. This will be done by repeating a calculation 10 times, each time we will use a different fold as the test or holdout fold. Here is the calculation.

a.) Perform Kmeans clustering on nine folds. Keep the number of iterations and the number of restarts low, as we use this only to initialize the EM algorithm.

b.) Pass the output of a.) into the M step of the EM algorithm. This will get us started with the EM algorithm by creating a so-called "model" and initializing the cluster shapes and centers.

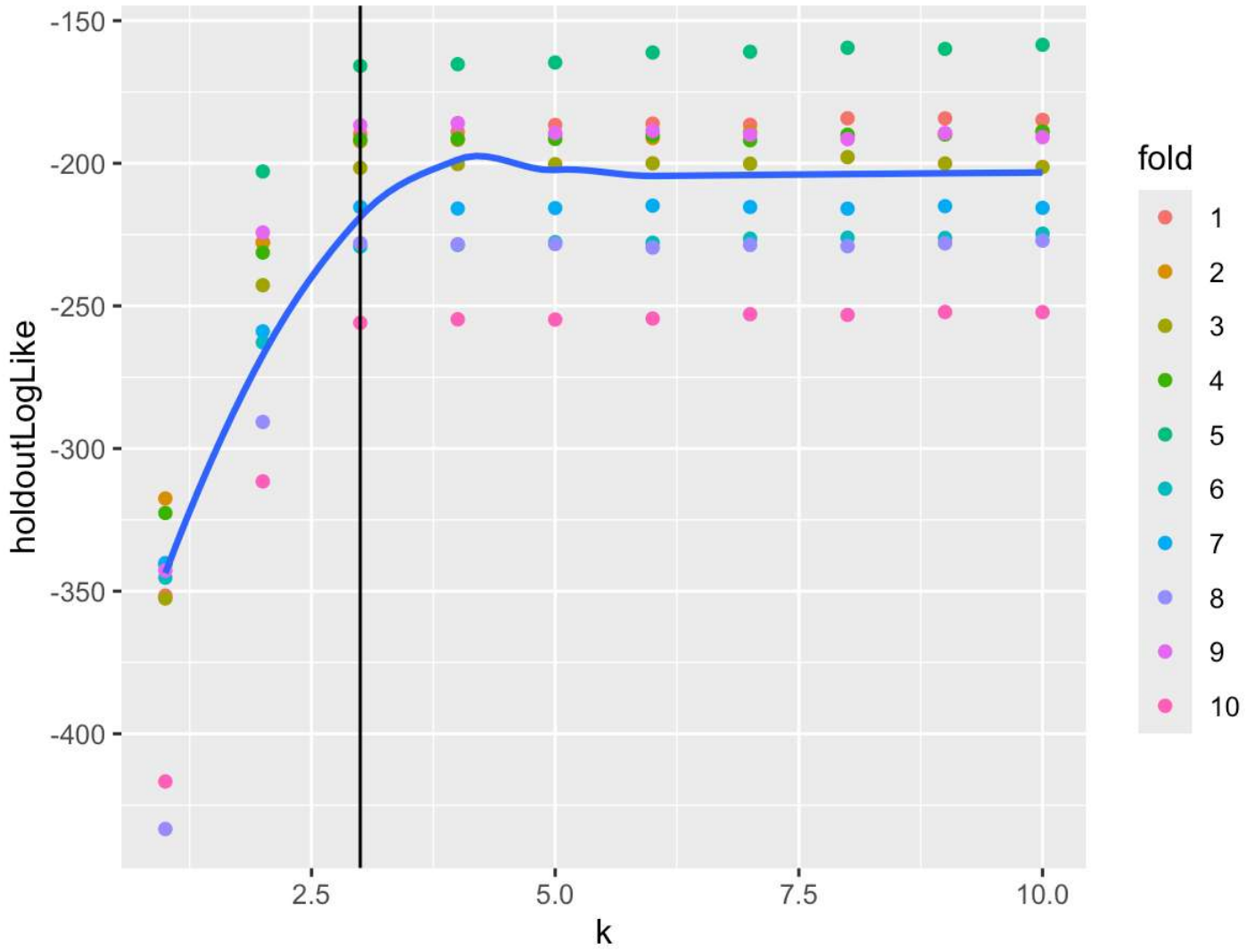
c.) Run the EM algorithm onto a convergence, using only the nine folds. Save the output as a variable of your choosing.

d.) Now take the output of c.), which includes the centroids and cluster shapes, and assign clustering membership to the hold out fold. This is done by passing both of the model from c.) and the holdout fold into the E-step one last time.

e.) Recall that maximizing a log likelihood function is a good thing for classification. Specifically, we choose the class whose log likelihood function is sufficiently high. It turns out that the same applies to clustering. That is, for a fixed set of feature vectors, we estimate the GMM parameters by maximizing a log likelihood function. Fortunately, the E-step provides an estimate of the log likelihood function, evaluated on the holdout fold and the model parameters. You may find this log likelihood in the output from d.). For each execution, retain this value, along with the holdout fold index and the chosen value of K.

5.) (30 points) Apply the process in 4.) on the $1 <= K <= 10$. Produce a plot summarizing the results from e.) as follows. Provide a scatterplot of the log likelihood values (vertical axis) for all values of K (horizontal axis) and holdout fold index (coloring). Color code these points by the holdout fold index and provide a legend explaining the coloring. Label the horizontal axis "k", and the vertical axis "holdoutLogLike". For each value of K, compute the average holdout log likelihood across folds. Include in your plot a line connecting these averages, either piece linear, or by smoothing. Find the smallest value of K which reaches the asymptotic (high-K) value of log likelihood within 15%. Draw a vertical line at this value of K. This will be the estimated cluster count, bestK, for the remainder of this work. Using this value of K, and **both** test and training points together, reevaluate the GMM model as you did earlier. This is your final GMM model.

Here is the requested plot



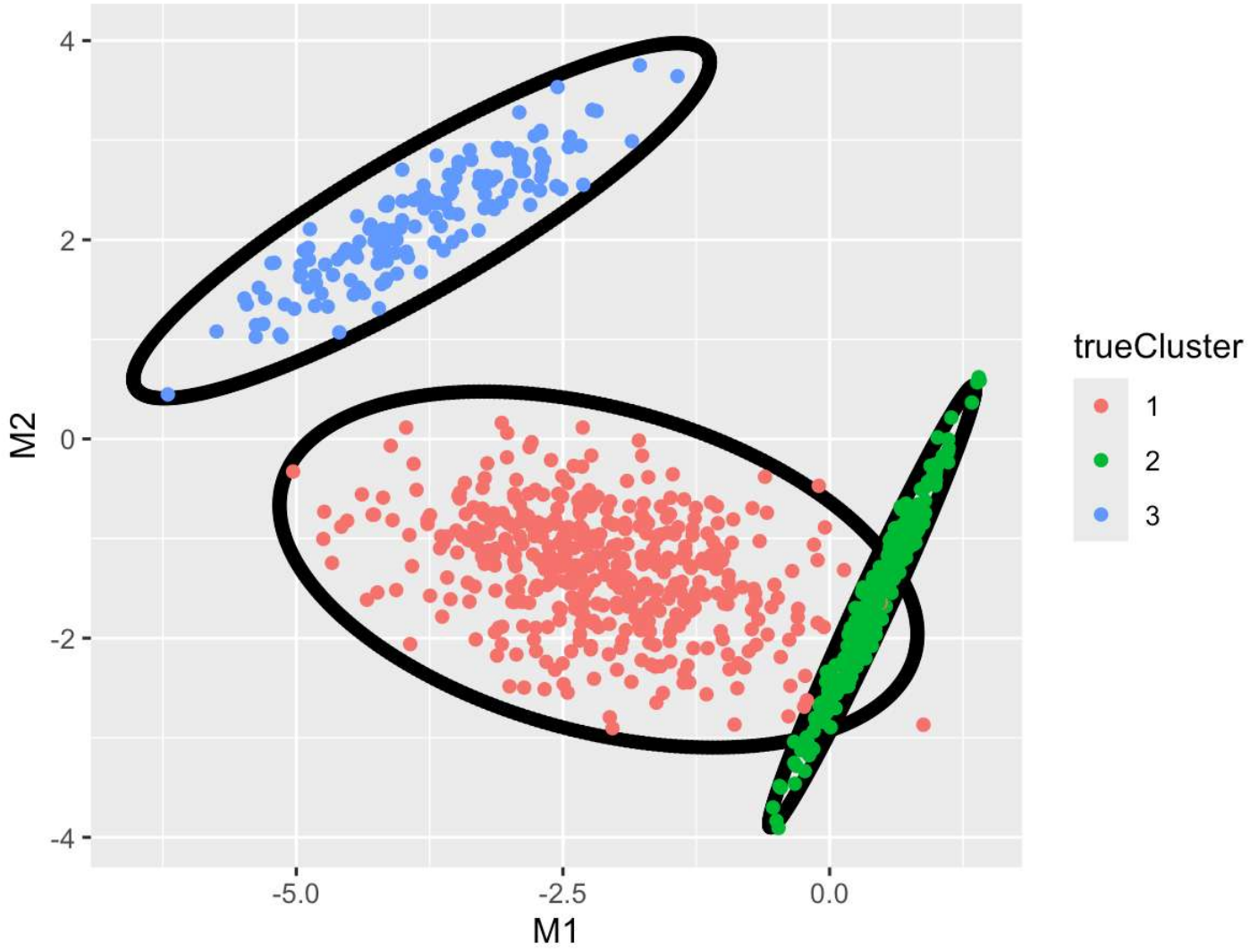
6.) (20 points) Please answer the following questions about the plot in 5.). For a given horizontal value, why is there variation in the altitude? Interpret both the highest and lowest altitude for a value of K...what does that mean? If two adjacent values of K produce the same (fold-averaged) altitude, why do we choose the smaller K? What, if any, is the risk of choosing the smaller value?

You will now find shape contours for each of the clusters. This is easily done as follows. Look at the contents of the final GMM model. In that model, there will be estimates of the covariance matrices, one for each cluster. If you are lucky, there will also be the "square root factor" or Cholesky factor of this covariance matrix. For each Cholesky matrix, multiply by the vectors [3cos x, 3sin x], where x is a fine grid on [0,2 pi]. The output will be the coordinates of a vector on the shape contour, with a centroid at the origin. Add to this the vector the corresponding cluster centroid, also found in the final GMM model. This method will allow you to calculate an ellipse of points corresponding to a particular cluster.

For a given value of K, the variation of the altitude is due to a relative mismatch between the model parameters and the holdout fold. The highest altitude refers to the best match between the model and the test data, while the smallest value indicates the poorest match. When two models yield similar results, we choose the simpler model (with fewer parameters). This is because there would

be more tuning vectors per parameter for the simpler model. The risk of doing this is the unintended merger of two clusters.

7.) (20 points) Produce a graph comparing the true cluster assignments to the model ellipses as follows. Begin by repeating step 2.) above in new plot. Then, include scatter plots of the vectors found in 6.) above. Please provide a response to the following question. What is your interpretation of the model fit to the measurements in this case? Why was the numeral "3" chosen to construct the ellipse contours above? Provide an accurate interpretation.



The contours indicate an excellent fit with the measurement vectors. The scalar "3" was chosen so that each ellipse included 3 standard deviations from the centroid on each size (6 total) for each ellipse axis. If the feature vector are Gaussian (they are), then the ellipse axes contain approximately 99% of the realized points.