Methods to Improve Expectation–Maximization Algorithm Estimation

Joshua Head

November 2022

1 Introduction

The Expectation–Maximization (EM) algorithm is frequently used to estimate parameters from mixture models, and can produce accurate estimates for a given data set, but a draw back of using this method is the requirement of providing initial estimates of the given parameters. This means that an individual would need to know the number of sub populations as well as the estimates for each population. In this paper, a solution to the optimal number of sub-population will not be discussed. However, it will instead focus on providing better initial estimates by using hard clustering methods, which will increase the accuracy of the fitted values as well as reducing the number of iterations required for fitting the EM algorithm.

2 Algorithms

The EM algorithm was introduced in 1977 by Arthur Dempster, Nan Laird, and Donald Rubin (Dempster et al. 1-38). The algorithm itself involves alternates between two steps: the expectation step and the maximization step. The expectation step calculates the expected value of the likelihood function of an assumed distribution given a set of fitted parameters (the assumed distribution in this research will be a multivariate normal distribution with multiple sub population model). The maximization step updates the parameters using its previous expectation step. After reaching convergence or a predetermined amount of iterations, the algorithm is stopped.

In addition to the EM algorithm, the research will be using 2 different hard clustering methods, which are k-means and k-medians. The k-means clustering algorithm first randomly selects k data points, which are set as the first centroid, and then each data point from a data set is assigned to a group by the closest centroid. After that, in the next step the centroids are recomputed as the mean distance of each subgroup before repeating until convergence. The k-median applies the same process, but instead of using the mean as the centroid, the

median is used. Other clustering methods were attempted, but failed during the EM fitting phase.

3 Data

The data set selected for experimentation of this method is the National Basketball Association(NBA) Player data provided on Kaggle ("NBA Players stats since 1950"). It contains 15042 rows and 53 columns. However, the data set is subsetted to the season 2016-2017 to reduce the possible influence time could have on the players' performance. To reduce computation time, only 9 columns are kept: Player, Pos, 3P%, 2P%, TRB%, AST%, STL%, BLK%, and TOV%. There are only two categorical columns in this data set, which are Player and Pos. Pos is the position of each player. C is Center, PF is Power Forward, SF is Short Forward, PG is Point Guard, and finally SG stands for Shooting Guard. The Pos column will be used at the sub populations in this project. All the numeric value represents the average game performance of each NBA player.

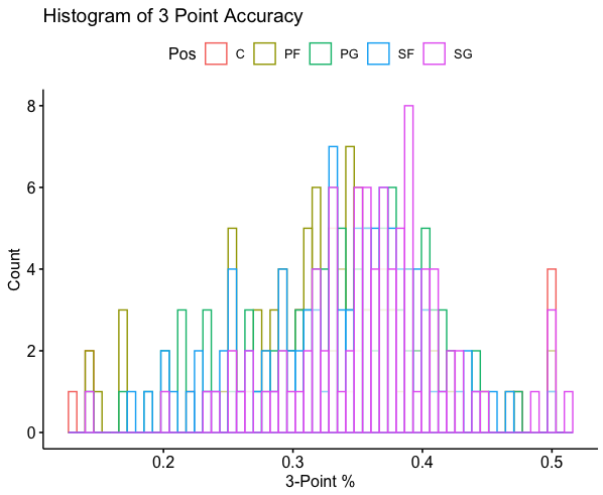


Figure 1: 3-Point Accuracy %

The above plot shows a relatively normal distributed of the 3P metric. Missing value handling was used to produce better estimations. After this process, the data set number of rows was reduced to 416. The table below is the mean statistics by position.

Pos	3P%	2P%	TRB	AST	STL	BLK	TOV
C	0.34	0.53	15.37	11.50	1.46	3.52	12.97
PF	0.32	0.50	12.13	9.64	1.27	1.98	12.36
PG	0.34	0.45	6.06	25.78	1.80	0.62	15.28
SF	0.33	0.50	8.64	9.89	1.59	1.22	11.53
SG	0.36	0.46	6.37	12.21	1.49	0.79	11.32

Figure 2: Mean Values by Position

4 Methodology

Usually, when fitting an EM algorithm, random values are calculated and given to the algorithm as initial values. However, this could lead to slow computation as well as poor fitting of the parameters since local minimums could influence the fitting. Thus "priming" the algorithm with better initial values is critical to producing more accurate result with less iterations. A possible method to prime the EM algorithm is using parameters of a data set, which has been clustered into sub populations. And the two cluster options used in this research is k-means and k-median(hierarchical and DB scanning was attempted, but was unsuccessful).

So once an cluster algorithm is fit, the estimated parameters of each cluster is fed into the EM algorithm, and then tests are done to see the effectiveness of each method. These two cluster methods will be compared to the general random initial values way.

5 Fitting and performance

The fitting of the EM algorithms through each priming method produced different results. The random method did performed worse than both k-means and k-median. The initial values of the hierarchical and DB scanning clusters methods could not be used to fit an EM algorithm because of the size of the smallest cluster. The likely reason for the inability of the fitting is the normality assumption that the EM algorithm assumed, and all clusters smaller than 40 observations could not be fitted(a worthy note is that the required threshold of 40 is relatively close to the rule of thumb of 30 observations or more to assume normality). Because of the inability to fit EM algorithms, the two cluster methods were removed from this research.

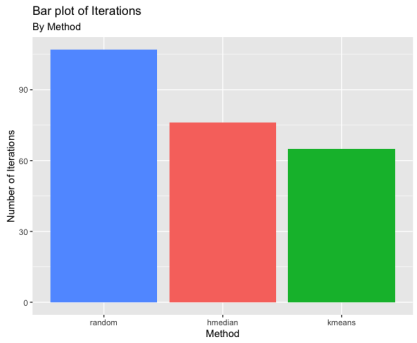
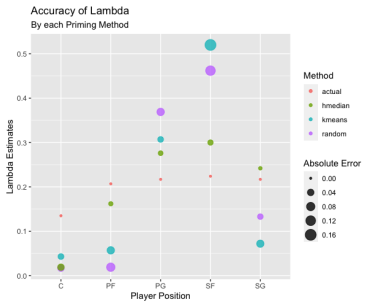
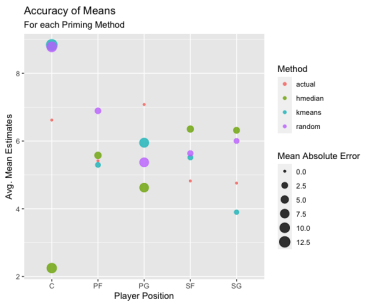


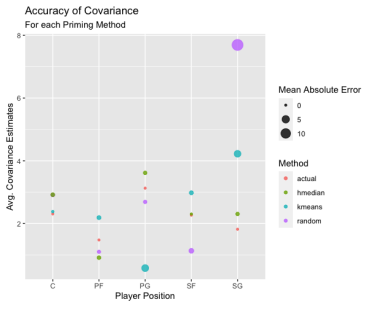
Figure 3: Number of Iterations by Method

As for the speed of fitting each EM algorithm, the number of iterations was used to see performance. k-means produced the best in regards to iterations, followed by k-medians, and finally the random primer.

To compare the estimated values of the proportion of number of observations per sub populations, mean, and covariance. The absolute value of the error for each corresponding estimate was used. The following plots show each sub population and its estimates with the corresponding true value. The size of the points represent the square of the error to show the performance of each method.







As the plots display, the worst method is randomly choosing an initial value,

the next is the k-means, and finally k-medians. There are two possible reasons the median clustering produced better estimates over the mean. The first is that the sub populations are not normality distributed, meaning that the k-means is pulled by the skewness of the data. The second possible reason is that outliers exist in the data set, which also end up pulling on the mean. k-medians would be the best option since most actual data sets are not normally distributed, and the ability to account for extreme values by relying on the median would be the optimal choice.

Additionally, the plots show that the player position Center(C) is poorly estimated by all methods, and this is likely because the proportions are wrong, which will naturally lead to worse fitting of the means and covariance matrices. The metrics of the player position Center is relatively close to Point Guard(PG) and Small Forward(SF) data metrics. In the game of basketball, this makes sense because the Center and Small Forward play similar roles, but not for the Point Guard.

6 Conclusion

After fitting, the k-median clustering method shows its relevance in the real world where the normality assumption cannot be guaranteed. And priming the EM model with any clustering method will lead to better performance than non-priming methods. Because the cluster estimates are already relatively close to the actual estimates, the EM algorithm is able to avoid local minimums during fitting, which leads to the model reaching a true global minimum. The results shows the power of the priming method, but better estimates could be obtained with a deeper look into the data set, and more consideration given towards preparation of the data before being fed into the process.

If this research were to be repeated, missing imputation should be investigated, which could lead to increased performance of the entire process. Since this was not utilized, the fitting of the initial clusters were poorly fit and it also led to some clustering method to be completely skipped over because of the normality assumption. Additionally, there was no outlier handling in this research as well. Because of that, it likely lead to some cluster methods, which are good at finding outliers, to create a sub population that consisted of only outlier, meaning the EM algorithm could not be used. Thus proper missing data and outlier handling would have lead to much better result.

7 Citations

Dempster, Authur P, et al. “Maximum Likelihood from Incomplete Data via the EM Algorithm.” Journal of the Royal Statistical Society, Wiley, 8 Dec. 1977, https://www.jstor.org/stable/2984875.

Author Goldstein, Ormi. ”NBA Players stats since 1950” Kaggle, 2017. ”https://www.kaggle.com/datasets/drgilermo/nba-players-stats?select=Seasons_Stats.csv”. 16 Nov. 2022.