Project 1 – Naïve Bayes

*Joseph Park - x172917 Date: 4/10/20*

Objective

The assignment is an illustration of classification based on the naïve Bayes model using particle collision data from a particle accelerator to identify particles.

Naïve Bayes Prediction Model

Step 1: Download the data set from Kaggle, look for missing data and look at summary, size and shape.

Step 2: Since the data set is too large to process, randomly sample. (Also, based on summary and histograms, as shown below, run the process with and without transformation and compare results.)

Step 3: Partition data into training vs testing and target vs classified data.

Step 4: Train data

Step 5: Classify data.

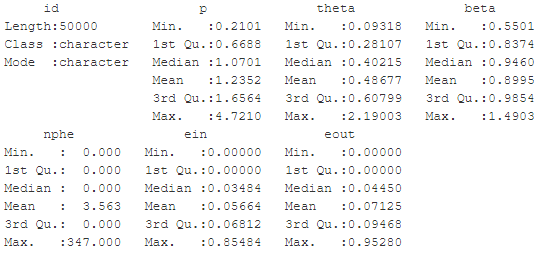
Step 6: Create a loop to optimize classification metric of choice, which is accuracy (for particles classification), with respect to Laplace estimator.

Step 7: Plot accuracy of the k’s and write confusion matrix for the k with max accuracy.

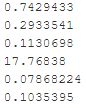
Step 7: Normalize the random sample and repeat steps 3-7 and compare results.

Step 8: Reset seed to what it was right before (pseudo)randomly sampling the 5000 particles (because I used same variable names), and repeat steps 2-7, this time with z-score standardization.

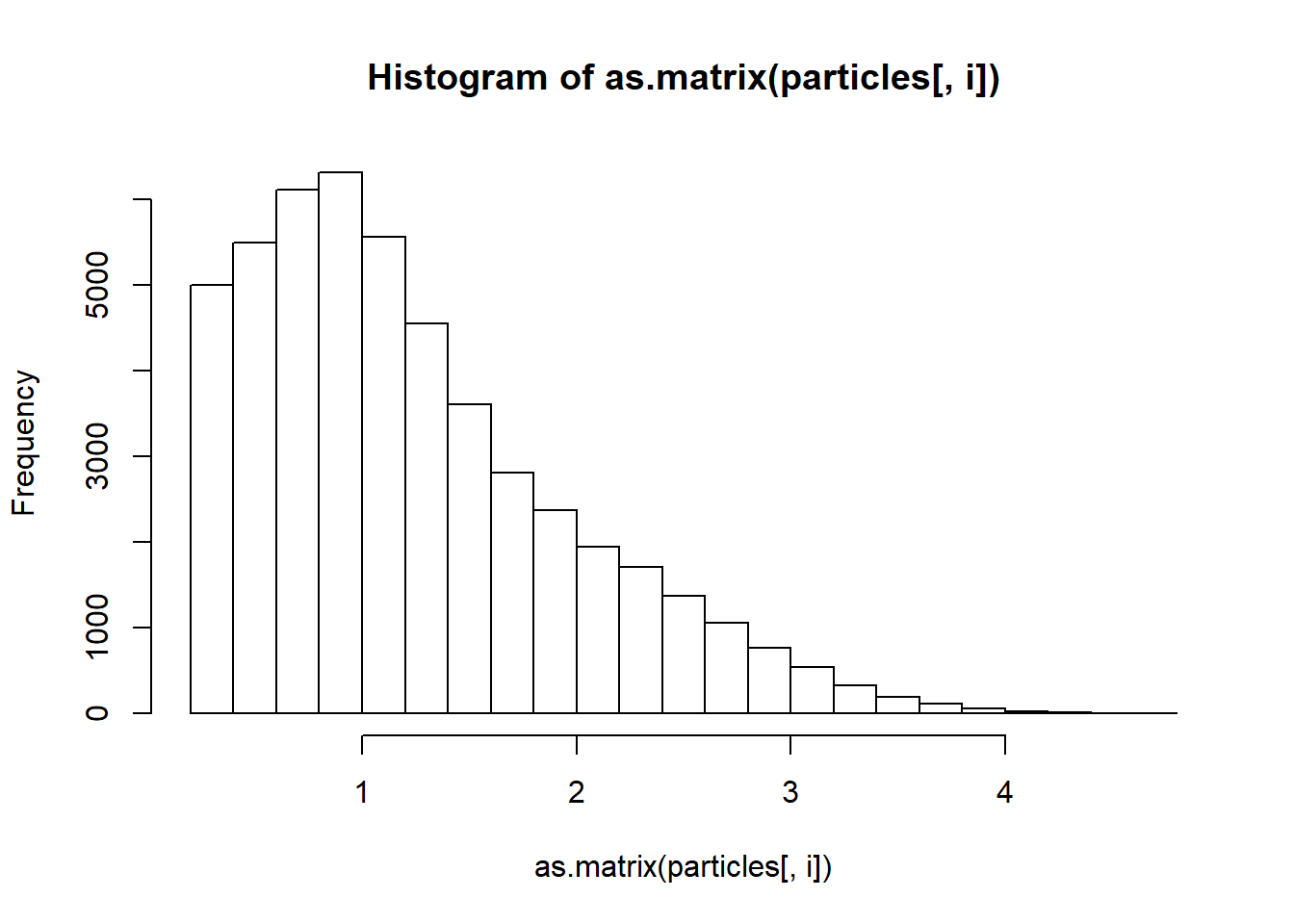
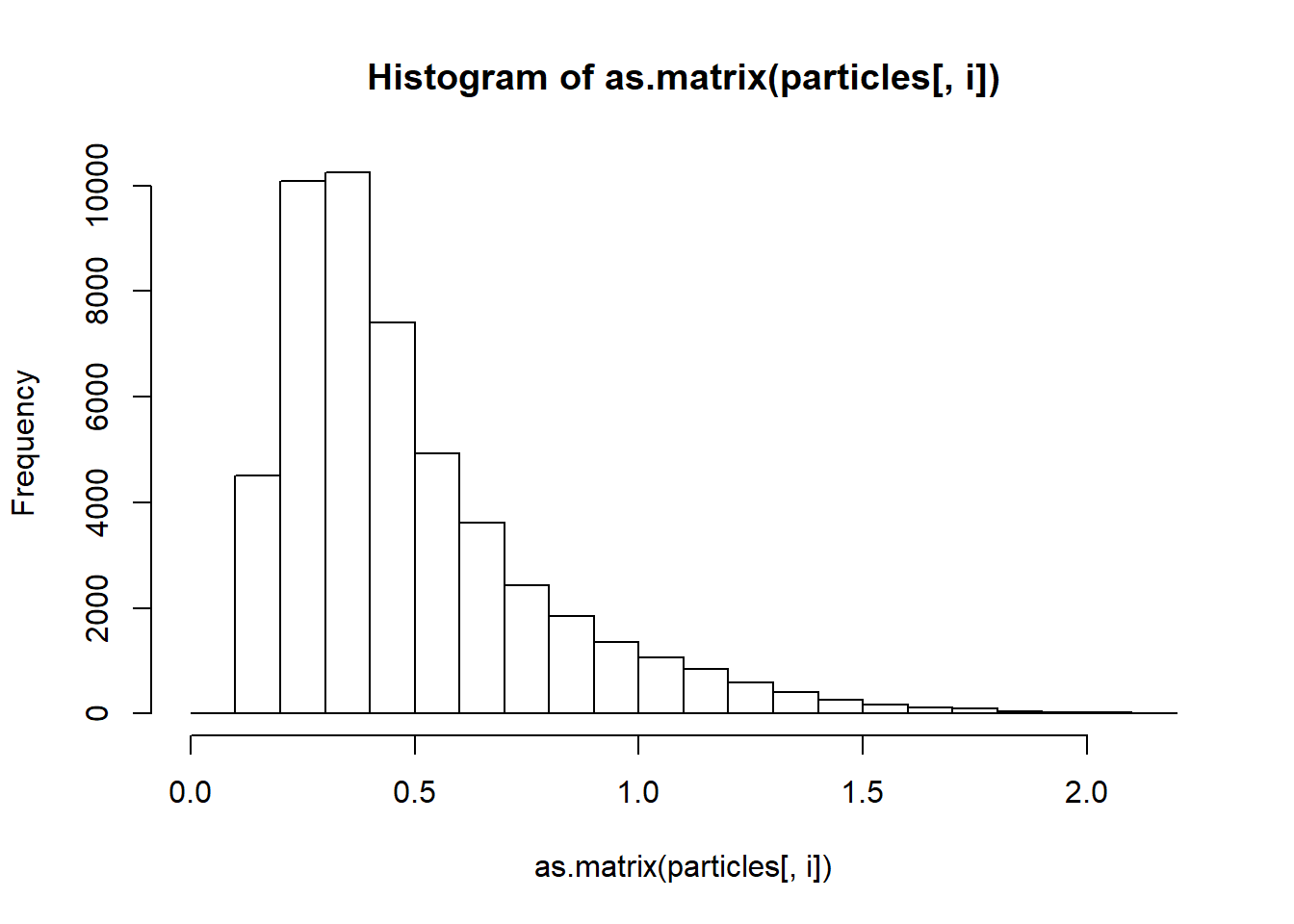
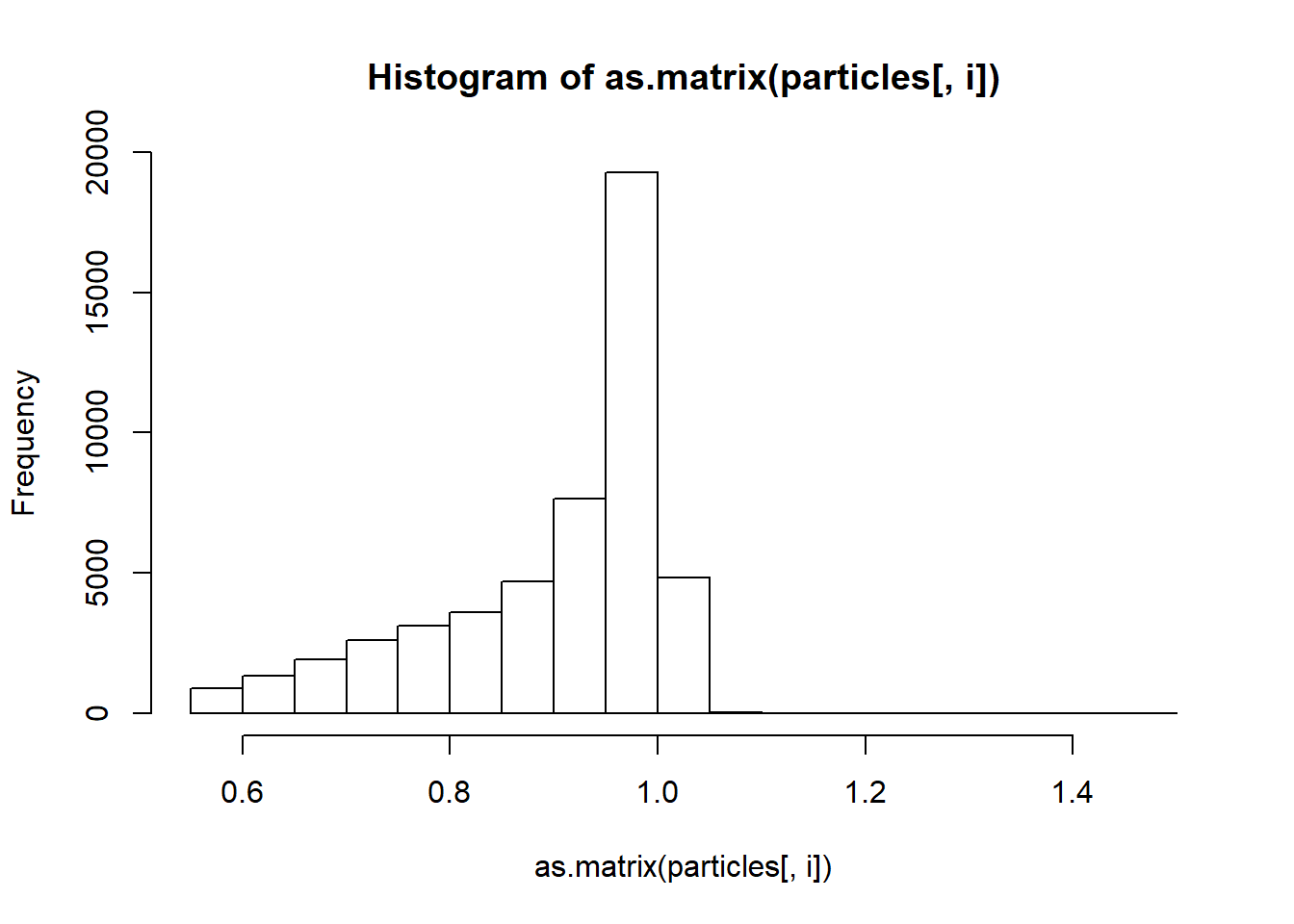
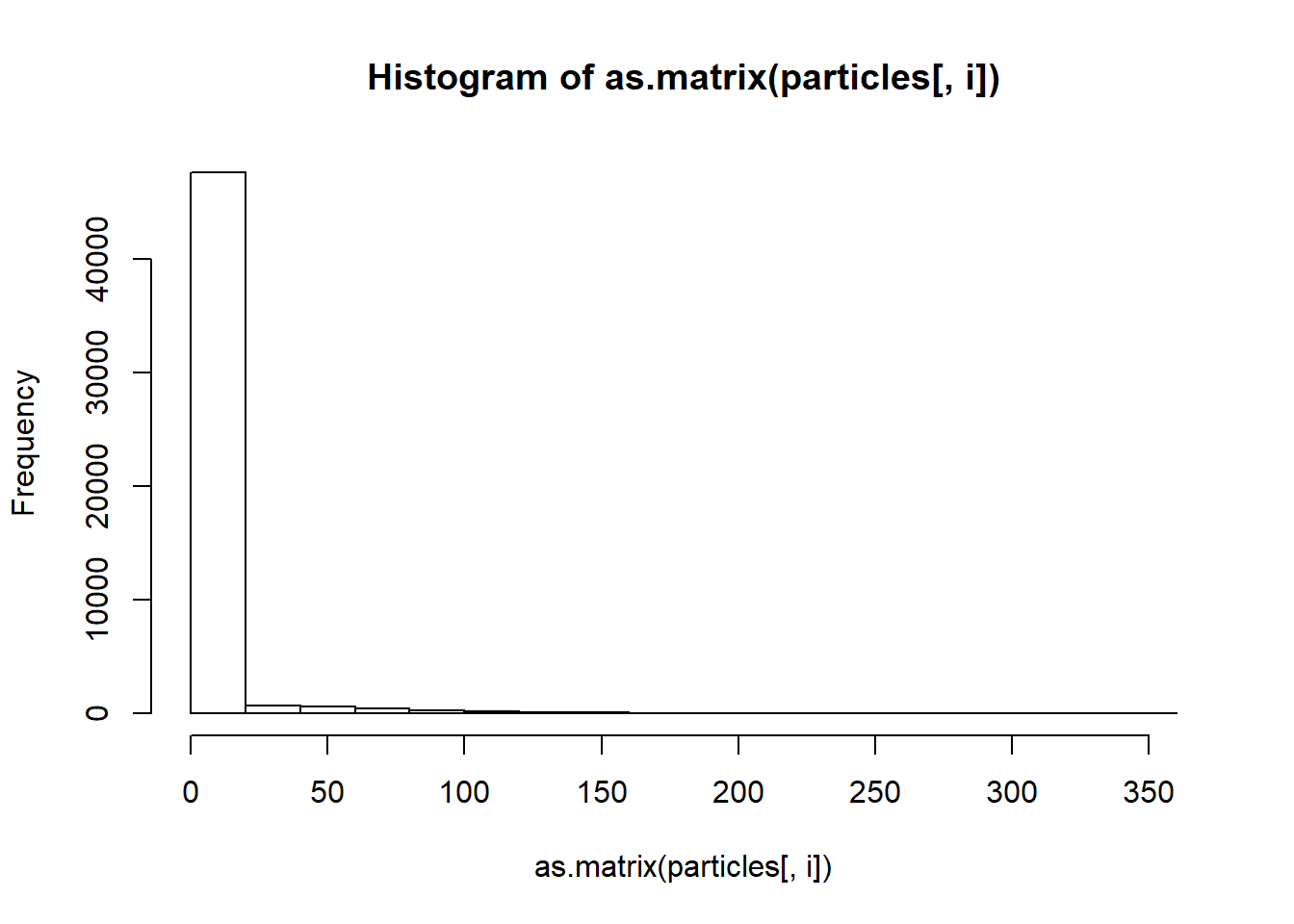
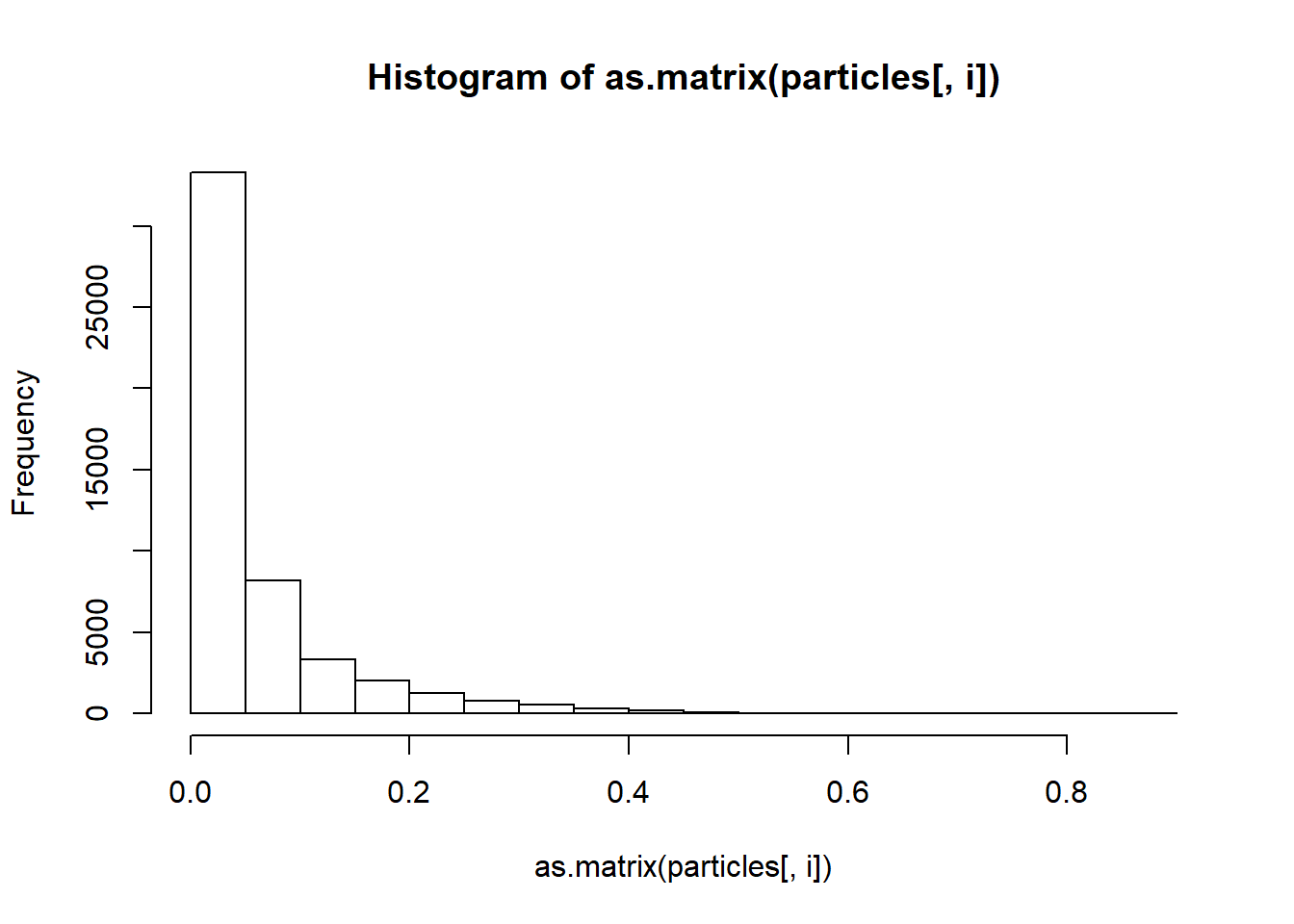
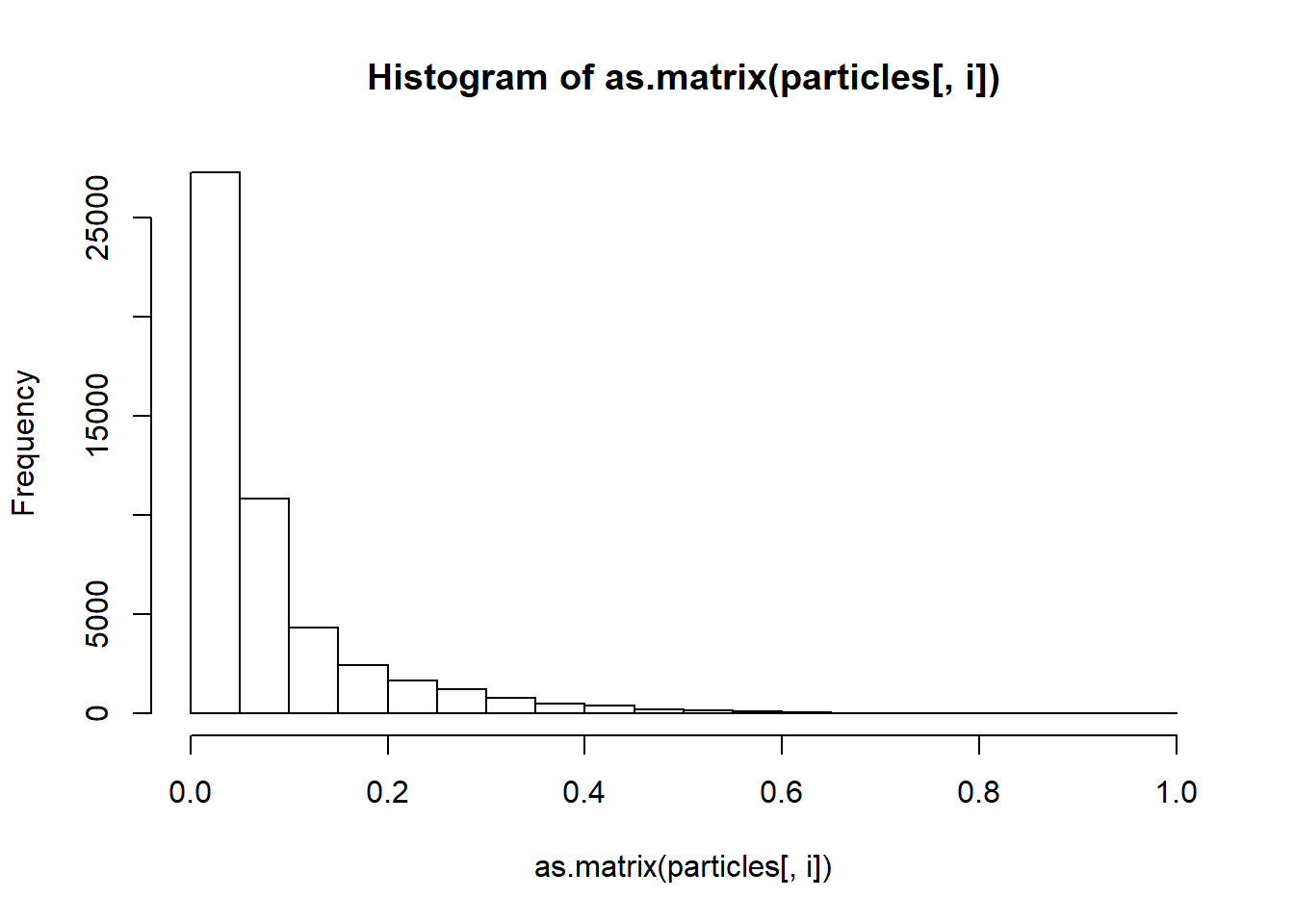
Step 9: Reset seed to what it was right before (pseudo)randomly sampling the 5000 particles (because I used same variable names), and repeat steps 2-7, this time optimizing with respect to variable removal.



Here are the standard deviations of the variables in order of above left to right then top down.

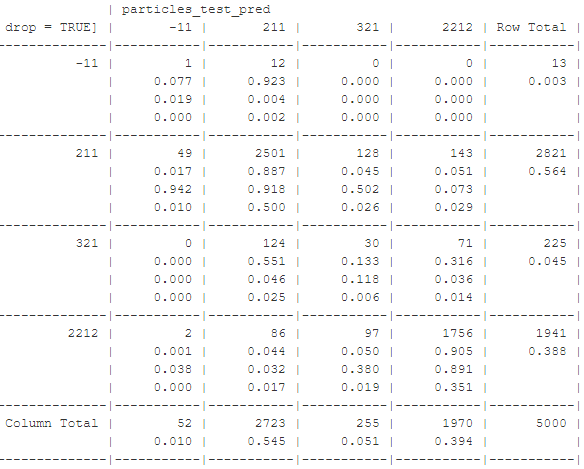


Here the corresponding histograms and the aforementioned order.

Results

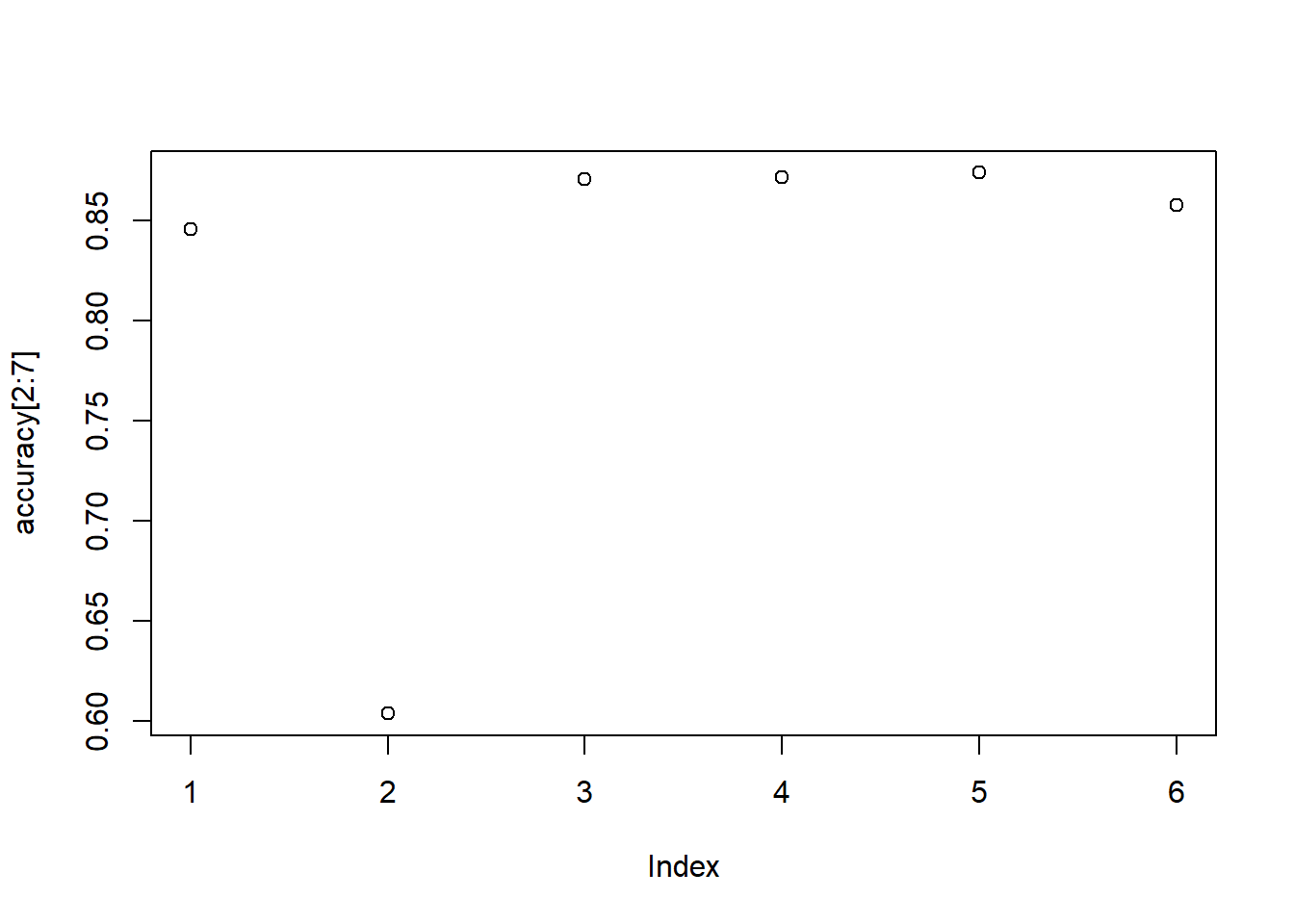
Results are same for all with and without normalization and z-score standardization (as anticipated) as well as Laplace estimator of zero and one per each.



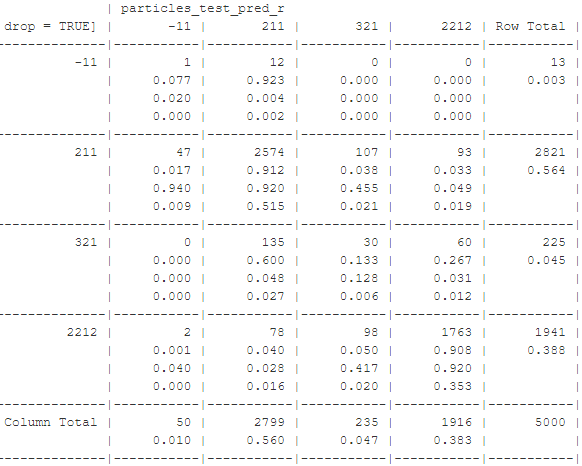
^Columns are predicted values and rows are actual values

"The accuracy is 0.8576"

Now with variable removals:



"The best accuracy is 0.8736"



Interpretation of the Results

Note 1: If particle -11 was of particular interest, a supercomputer (and thus ability to process a sample size significantly more 5000) would be of use, as 13 data points is not reliable.

Note 2: The accuracy is good for naïve Bayes and all attempts to improve performance.

Note 3: Whereas for kNN, when the order(s) of magnitude difference in actual numbers of particles appear to have the classification overpowered by the numbers of two particles (211 and 2212) that are more prevalent, here particles -11 and 321 where not under-classified (even over-classified). But, rather the majority of the particles classified as -11 and 321 were actually other particles. And worse, the number of particles that were correctly classified as -11 and 321 was nowhere the number of actual -11 and 321 particles.

Note 4: Just like with kNN, particle 321 is greatly misclassified as these two and disproportionately more to the one that is more prevalent: 100/28 > 2821/1941. Particle -11 is even more misclassified and with similar behavior of disproportion.

Note 5: This may be explained by high correlations between particles -11 and 211, and between particles 321 and 211 and particles 321 and 2212 (the faulty independence assumption of conditional classes).

Note 6: The other faulty assumption (of equally important variable) may be indicated as a playing a role by the graph of accuracy vs index of removal (particularly by the second index). This suggests a weighted model may be a better choice (with significantly more weight on the second index than the others).

Note 7: This naïve Bayes classification performed more poorly for -11 and 321 than did kNN and more poorly overall.

Note 8: If we cared in particular about particle -11 or particle 321, we could optimize with respect to metric of choice for whichever particle(s).

Note 9: Since one of the features has significantly higher variance (though probably not enough), this suggest we may want to transform, as this (among other things) can cause lower accuracy that can be remedied by transformation. Doing the two transformations taught so far (as suspected) made no difference, nor did the difference in Laplace estimator.

Note 10: Although I was right about removal of the nphe variable giving better accuracy, the improvement in accuracy may not be enough to justify losing the information from that variable and possible robustness to classify when more emphasis is on one of the less prevalent particles. But, we check investigate this further by cross validation, bootstrapping, using greater computational power (e.g. supercomputer) for a significantly bigger sample (though naïve Bayes is supposed to already work well with little data), and/or choosing metrics specific to the less prevalent particles.

Note 11: The histograms look Poisson, multinomial, or gamma (probably not beta

because of bounded domain). There are not Poisson or multinomial (since their domains are not subsets of the set of integers). So, we can design a naïve Bayes classifier to use gamma distribution pdf. Or, we can homogeneously discretize (by histogram binning) the distribution and scale appropriately to make the integers make sense for the available (R) naïve Bayes classifiers that handle Poisson or multinomial distribution pdf’s.

Note 12: Given that so many of the values are close to zero, designing a naïve Bayesian classifier that uses log likelihood could be another option to improve accuracy (though I’m not sure that they are overall close enough to zero to make a significant difference).