Sarah Kostick
Crop_Sci545: Homework 2

Part (1) The goal of part 1 was to examine the relationship between missing rate and stochastic imputation accuracy. It was hypothesized that as the missing rate increased, the stochastic imputation accuracy would decrease. In other words, it was expected that when more data was missing, the imputation accuracy correlation will be lower. The methods that were followed are outlined below.

Methods:

1. Missing rate was set 5%, 10%, 25%, 50%, and 75%. Imputation was repeated for each of these missing rates.
2. Missing values imputed with stochastic method.
3. Accuracy correlation coefficient calculated.
4. Process repeated 30 times for each missing rate.
5. Average correlation coefficient and standard deviation were calculated for each missing rate.

Results:

Table 1: Missing rate (%), number of replications, average accuracy correlation coefficient, and standard deviation for imputation accuracy of stochastic imputation.

Missing Rate (%)	# Reps	Avg. Cor.	Sd.
5%	30	0.387156	0.0043269
10%	30	0.387210	0.0028279
25%	30	0.386927	0.0018864
50%	30	0.387222	0.0013473
75%	30	0.387293	0.0009043

The average accuracy correlation coefficient remained about the same (~0.38) regardless of the missing rate (Table 1). This result differed from what was originally expected. However, if one really thinks about it, the result makes sense. The stochastic imputation method used is based upon allele frequency. For example, if you have inbreds with two possible alleles (A and B), the frequency of allele A is f(A). A missing allele N can be imputed as $N = \begin{cases} A\ if\ \ x < f(X) \\ B\ if\ otherwise \end{cases}$ if x is uniformly distributed (U(0,1)). Based on how stochastic imputation works, if the missing values were set at random, the frequency of the alleles should remain relatively constant regardless of missing rate. As a result, this could explain why regardless of the missing rate, the average accuracy correlation remained relatively constant. What can be concluded from this exercise is that for stochastic imputation, the accuracy will remain relatively constant regardless of how many data points are missing.

Part (2) The purpose of part 2 was to examine the relationship between sample size and stochastic imputation accuracy. It was hypothesized that larger sample sizes would result in higher imputation accuracy. The methods that were followed in order to test this hypothesis are outlined below.

Methods:

1. Missing rate was fixed at 25%.
2. 75%, 50%, and 25% of individuals were sampled and stochastic imputation was performed for missing values of these samples.
3. Accuracy correlation coefficients were calculated for each imputation.
4. Process was repeated 30 times for each sample size (i.e. 75%, 50%, and 25%)
5. Average correlation coefficient and standard deviation were calculated for each sample size.

Results:

Table 2: Sample size (%), number of replications, average correlation coefficient, and standard deviation for imputation accuracy of stochastic imputation.

Sample size (%)	# Reps	Avg. Cor.	Sd.
25%	30	0.388466	0.0045340
50%	30	0.387403	0.0033587
75%	30	0.386853	0.0017348

The average accuracy correlation coefficient remained about the same (~0.38) regardless of the sample size (Table 2). This result differed from what was expected with the average accuracy correlation coefficient remaining relatively constant regardless of sample size. The reason for this result may be similar to the reason for the result in part 1. Since stochastic imputation is based upon allele frequency, it would make sense that sample size would not affect imputation accuracy if the missing values were set at random. If the missing values were set relatively at random, a change in sample size should not result in a large change in the allele frequency. As a result, a change in sample size should not cause a change in imputation accuracy if the missing values were set at random.

Part (3) The purpose of part 3 was to examine the relationship between the number of nearest neighbors (K) and imputation accuracy. It was hypothesized that as the number of nearest neighbors (K) increased, the imputation accuracy would also increase. The methods followed to test this are listed below.

Methods:

1. Fixed the missing rate at 25%.
2. Imputed with KNN method with K = 2, 5, 10, and 20.
    a. It is important to note that KNN imputation was performed on the transpose of the matrix because the impute() function considers columns as individuals. Since the

data matrix has individuals as rows and columns as SNPs, the transpose of the matrix had to be used for imputation.

3. Calculated accuracy correlation coefficient.
4. Repeated with each K value, 20 times.
5. Calculated average accuracy correlation coefficient and standard deviation.

Results:

Table 3: number of nearest neighbors (K), number of replications, average correlation coefficient, and standard deviation for imputation accuracy of KNN method.

# Nearest Neighbors (K)	# Reps.	Avg. Cor.	Sd.
2	20	0.6015026	0.0003211
5	20	0.6391945	0.0005408
10	20	0.6486037	0.0007212
20	20	0.6506768	0.0002188

As expected, the average accuracy correlation coefficient increased as the number of nearest neighbors (K) increased (Table 3). What these results indicate is that the larger the number of nearest neighbors (K), the higher the imputation accuracy. This result intuitively makes sense. As you increase the number of nearest neighbors, the more likely you will be able to get an accurate picture of what the missing value should be. However, at a certain optimal point increasing K will not give you any added benefit. We can see this even in this dataset. For example, when we increased K from 2 to 5 we saw a relatively large increase (0.602 to 0.639) in the correlation coefficient (Table 3). In contrast, when we increased K from 10 to 20 we saw a much smaller increase (0.649 to 0.651) in the correlation coefficient (Table 3). What this indicates is that there is an optimal K value and that optimum is dependent upon the specific dataset.

Part (4) The goal of part 4 was to examine the relationship between imputation method (i.e. stochastic, KNN, and BEAGLE) and imputation accuracy. It was hypothesized that the methods like KNN and BEAGLE would have higher imputation accuracy than the stochastic method. The methods followed to test this hypothesis are listed below.

Methods:

1. Fixed the missing rate at 25%.
2. Impute with stochastic, KNN (with K=10), and BEAGLE imputation methods.
    a. For the KNN method, the reason K=10 was used was because 10 is the default of the impute() function in R.
    b. The R functions for stochastic and KNN methods that developed in parts 1 and 3 were used in this part of the assignment.
3. Calculated accuracy correlation coefficient for each method.
4. Repeated imputation with each method 10 times.
5. Calculated average accuracy correlation coefficient and standard deviation for each method.

Results:

Table 4: Imputation method, number of replications, average accuracy correlation coefficient, and standard deviation of correlation coefficient.

Imputation Method	# Reps	Avg. Cor	Sd.
Stochastic	10	0.38617	0.00258
KNN	10	0.64875	0.00073
BEAGLE	10	0.57956	0.00103

The average accuracy correlation was lowest for the stochastic method of imputation and the highest for KNN method (Table 4). The correlation coefficient for BEAGLE was greater than the stochastic method but was slightly lower than the KNN method (Table 4). Based upon these results, if one wanted to use the imputation method with the highest accuracy, one should use KNN. These results explain why KNN and BEAGLE imputation methods are used much more often in studies than the stochastic method.

Part (5) The goal of part 5 was to examine how switching neighbors and attributes in the KNN method would affect imputation accuracy. It was hypothesized that switching neighbors and attributes would not affect imputation accuracy. The methods that were followed are listed below.

Methods:

1. Missing rate was fixed at 25%.
2. Missing data was imputed using the KNN method with K=2, 5, 10, and 20.
    a. It is important to note that unlike in part 3, KNN imputation was carried out on the matrix not the transpose of the matrix. The reason for this is that the impute() function in R considers columns as individuals and rows as markers. Since we wanted to switch neighbors and attributes, I ran imputation on the matrix and not the transpose of the matrix.
3. Calculated accuracy correlation coefficient for each K value.
4. Repeated imputation 20 times for each K value.
5. Calculated average accuracy correlation coefficient and standard deviation for each K value.

Results:

Table 5: Number of nearest neighbors (or attributes in this case), number of reps, average accuracy correlation coefficient, and standard deviation for KNN method of imputation.

# Nearest Neighbors (K)	# Reps.	Avg. Cor.	Sd.
2	20	0.6798696	0.000000
5	20	0.6798696	0.000000
10	20	0.6798696	0.000000
20	20	0.6798696	0.000000

In part 3, by increasing K, the average correlation coefficient increased indicating that increasing the number of nearest neighbors one could increase the imputation accuracy (Table 3). The highest average correlation coefficient in part 3 was 0.6506768 when K=20 (Table 3). In part 5, the average correlation was higher than in part 3 for any K value (Table 5). What was most interesting in part 5 was that the average correlation coefficient remained the same regardless of the value of K (Table 5). Also, there was no among replication variation for a given K value as can be observed by the standard deviations of 0 (Table 5). This lack of change in correlation coefficient must have something to do with how missing values are imputed with the KNN methods. Specifically, it must be related to the Euclidean distance calculation.

Extra credit

Part (5) I did not answer this question.

Part (6) The purpose of part 6 (extra credit) was to find an imputation method that was more accurate than the KNN and the BEAGLE methods. As we know from part 4 that KNN and BEAGLE had average correlation coefficients of 0.64875 and 0.57956, respectively. Any method that was chosen in this section must have had an accuracy correlation coefficient of greater than 0.64875. The method that was chosen for imputation in this section is random forest imputation with the R package missForest(). It was hypothesized that Random Forest imputation would have a higher imputation accuracy correlation than both KNN and BEAGLE. This hypothesis was based upon previous evidence presented by Xavier et al. (2016). Xavier et al. (2016) examined how different imputation methods (i.e. multivariate mixed model (MMM), Hidden Markov Models (HMM), a logic alogorithm, KNN, Single Value Decomposition (SVD), and Random Forest) on SNP data in soybeans. What they found was that RF (Random Forest) had higher imputation accuracy than both KNN and BEAGLE. This information was what formed the basis of my hypothesis that Random Forest imputation should have a higher correlation coefficient than KNN and BEAGLE. The methods that were followed to test this hypothesis are outlined below.

Methods:

1. Missing rate was set to 25%.
2. missForest package was loaded in R.

3. The doParallel and foreach packages were loaded so that missForest could be run in parallel.
4. A backend was registered so that missForest could be run in parallel.
5. Missing data was imputed using missForest() command.
6. Imputation was only carried out once since the imputation process took about 2 hours to run in R.
7. Accuracy correlation coefficient was calculated.

Results:

Imputation with RF resulted in an accuracy correlation coefficient of 0.72225, which was higher than both the coefficients for KNN and BEAGLE calculated in part 4. As a result, RF imputation appears to be a more accurate method of imputation. The question becomes why is RF have higher accuracy than other methods used in this assignment. The answer probably has to do with how RF works. RF is a non-parametric method of imputation which means that it doesn't make any assumptions about data distribution (Analytics Vidhya Content Team, 2016). According to Analytics Vidhya Content Team (2016), RF uses a model based upon the observed to predict the missing values. As a result, this method of imputation can be highly accurate.

Citations:

Analytics Vidhya Content Team. 2016. Tutorial on 5 Powerful R Packages for imputing missing values. https://www.analyticsvidhya.com/blog/2016/03/tutorial-powerful-packages-imputing-missing-values/ (Accessed 2/14/2017).

Xavier, A., W. M. Muir, and K. M. Rainey. 2016. Impact of imputation methods on the amount of genetic variation captured by a single-nucleotide polymorphism panel in soybeans. BCM Bioinformatics 17: 2-9. https://www.ncbi.nlm.nih.gov/pubmed/26830693 (Accessed 2/14/2017).