Algorithms in Bioinformatics
(Exercise 6 Report)
CMPUT 606
Winter 2020
ccid: afia

In the analysis of genome-wide association study (GWAS), genotype imputation is an important factor. Various reasons like a design flaw, unknown genomics information, the difference between the individuals, experimental errors, etc. can give rise to missing values in genotype data. As a result, it becomes important to rely on imputation techniques, which in turn helps investigators to test association at the genetic markers. Additionally, it helps to combine results across different studies that are built upon different genotyping platforms. So far, there have been different studies [1, 2, 3] that focus on applying various imputation techniques in order to impute missing genotypes in a dataset and thereby measure the efficiency of the imputation.

The main objective of this exercise is to impute values in datasets, given that they have various proportions of missing entries (0.1%, 1%, and 10%). For missing SNP genotype dataset and HAPI-UR genotype dataset, we utilize different imputation methods like row imputation (no to trivial), k nearest neighbor imputation (linear dependencies), iterative imputation and sequential imputation. For the provided HAPI-UR genotype dataset, additionally, we check the imputation capacity of phasing programs like HAPI-UR and FastPhase. The imputation accuracies from all the above-mentioned methods are reported in terms of their NRMSE scores.

The sample SNP genotype dataset consists of 305-row entries, depicting the population size, and 1040 SNP markers and an EBV value for cattle milk yield. The SNP markers are represented in terms of a genotype encoding where a 1, 0 and -1 represents a homozygous major, heterozygous and homozygous minor respectively. For this particular exercise, we keep the EBV values out of consideration.

On the other hand, the genotype dataset of the human genome, collected from the International HapMap Project consists of an individual file, a Single Nucleotide Polymorphism (SNP) marker file and a genotype file for 88 samples in the EIGENSTRAT format. We utilize only the supplied genotype file for this exercise. This file contains one line per SNP where each row contains a sequence of characters and each character encodes the genotype of one individual at the given SNP. Each of the column numbers corresponds to data for the individual at the same line number

in the individual file. The encodings in each row are between 0, 1, 2 or a 9, with 2 encoding the homozygous reference, 1 encoding the heterozygous, 0 encodings the homozygous alternate and 9 encodings the missing values. All these files contain information about the autosomes (22 pairs). For the span of the assignment, we will only focus on chromosome 1 data.

With the help of a python program, missing values are inserted at random positions in the above-mentioned datasets in 0.1%, 1% and 10% entries of the genotype values. Different imputation techniques from the scikit learn package [4] of python are used for imputing the values in the missing datasets. For the imputation accuracy calculation, the original SNP genotype data (genotype_dataset_input.csv) and the data inside .geno file provided by HAPI-UR are considered as the ground truth. An NRMSE score is obtained for all the above imputation techniques in order to evaluate the performance of the obtained imputation accuracy scores. The program to create missing values is repeated multiple times to obtain the average NRMSE scores for each method and each dataset.

In the following part, we discuss the terminologies RMSE and NRMSE and report the NRMSE scores obtained from various imputation methods.

Root Mean Square Error (RMSE) is the standard deviation of the residuals which is a measure of how far from the regression line our data points are. Since RMSE is measured in terms of squared error difference that is why it is useful where large errors are undesirable. In other words, RMSE gives relatively high weight to large errors. Lower the RMSE score, better is the performance of the model.

On the other hand, NRMSE is the normalization of RMSE. For simulation studies, the evaluation for NRMSE is done in the following way:

From a complete matrix, we select the missing entries and the true values which correspond to the missing entries. Let a_i denote the true value, a_i^* denote the imputed value, $\mu 2$ denote the mean squared errors (MSE) and $\sigma 2$ denotes the variance, the ratio of the root squared values of MSE and Variance is expressed by NRMSE (μ/σ) in the following way:

$$\mu^2 = \frac{1}{q}\sum_{i=1}^{q}(a_i - a_i^*)^2$$

$$\sigma^2 = \frac{1}{q}\sum_{i=1}^{q}(a_i - \bar{a})^2, \text{ where } \bar{a} = \frac{1}{q}\sum_{i=1}^{q} a_i$$

The NRMSE scores for the four imputation methods on the three missing proportioned genotype dataset and on the missing proportioned HAPI-UR genotype dataset are in the following table. It is to be noted that, mean of the row values were used for the 'No to trivial' imputation method and Iterative Imputer pipelined with a KNN regressor was used as a Sequential Imputer. This is because, scikit learn and most of the imputation packages that are available do not provide any built-in sequential imputer, however, sequential imputation algorithms that can be implemented with Iterative Imputer by passing in different regressors for predicting missing feature values as mentioned in section 6.4.3.1 [5]. Moreover, in KNN imputer, the number of nearest neighbor (k) is selected on the basis of research work mentioned in [6]. This paper worked with four values of k=3, 5, 10 and 15 and found out that 15 is the optimal number of k. The python codes to derive all these values are appended in the "python codes" directory.

Table 1: NRMSE values for different imputation methods (genotype dataset)

Percentages	0.1%				1%				10%			
Methods	R1	R2	R3	Avg.	R1	R2	R3	Avg.	R1	R2	R3	Avg.
No to trivial	1.001	1.001	0.999	1.00	0.999	1.001	0.999	0.999	1.000	1.000	0.999	0.999
KNN imp.	0.858	0.808	0.806	0.824	0.785	0.801	0.802	0.796	0.807	0.809	0.808	0.808
Iterative imp.	0.782	0.752	0.793	**0.775**	0.725	0.732	0.713	**0.723**	0.727	0.712	0.732	**0.724**
Sequential imp.	0.857	0.865	0.889	0.870	0.867	0.865	0.861	0.864	0.875	0.869	0.862	0.868

For the hapi-ur dataset, the .geno file is taken to create missing entries with different proportions and then the values are imputed using the above mentioned four methods. The imputed files are then compared with the original .geno file in order to compute NRMSE scores in the following table:

Table 2: NRMSE values for different imputation methods on HAPI-UR dataset)

Methods	0.1%	1%	10%
No to trivial imp.	0.72924	0.73620	0.73245
KNN impute	**0.54884**	**0.55048**	**0.56756**
Iterative impute	0.71480	0.72367	0.72117
Sequential impute	0.86542	0.86204	0.87158

Additionally, we checked the imputation properties of HAPI-UR and FastPhase programs using the HAPI-UR dataset. For the hapi-ur dataset, different runs for

missing values are generated by running the HAPI-UR program three times and then a single phgeno file is produced by taking consensus with the help of the VotePhase program of HAPI-UR. The entries of this consensus file is compared with the original file at the missing sites of different missing files (0.1, 1 and 10%). The origchars files (denoting the reference and alternate alleles for a particular SNP site) and switch.out files from the FastPhase output are used together to bring back the genotype entries. Then this file is compared against the true genotype values to evaluate the NRMSE scores.

Table 3: NRMSE values to check the imp. capacity of HAPI-UR and FastPhase

Phasing programs	0.1%	1%	10%
HAPI-UR	1.804613	1.808338	1.2771
FastPhase	**0.82103**	**0.836035**	**0.88193**

Looking at Tables 1, 2 and 3, we can infer that Iterative Imputer outperforms other imputers when the provided genotype dataset (dataset with the EBV values) is used. On the other hand, KNN imputer, with k=15 (according to previous literature), outperforms all other imputers when the genotype data from HAPI-UR is utilized. We have checked the imputation capacities of HAPI-UR and FastPhase and looking at the result we can say that FastPhase outperforms HAPI-UR (Table 3) in terms of its imputation capability but cannot outperform when it is compared against the four imputation methods in Table 2. We also came across a work [6] which is very similar to our assignment. It works on a number of imputation methods and also checks the imputation capability of phasing programs. FastPhase is considered to perform the best (error in the range of 0.309–0.321) on their tuned parameters and settings.

References:

1. Isik F, Holland J, Maltecca C. Imputing Missing Genotypes. In Genetic Data Analysis for Plant and Animal Breeding 2017 (pp. 287-309). Springer, Cham.
2. Browning BL, Browning SR. Genotype imputation with millions of reference samples. The American Journal of Human Genetics. 2016 Jan 7;98(1):116-26.
3. Scheet P, Stephens M. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. The American Journal of Human Genetics. 2006 Apr 1;78(4):629-44.
4. https://scikit-learn.org/stable/index.html

5. https://scikit-learn.org/stable/modules/impute.html
6. Yu Z, Schaid DJ. Methods to impute missing genotypes for population data. Human genetics. 2007 Dec 1;122(5):495-504.