Kevin Cummiskey
PhD Candidate
Department of Biostatistics
May 3, 2016

Compressing HapMap Data using Sufficient Subpopulation Markov Chain Statistics

with Caleb Lareau and Matthew Ploenzke

In genetic recombination, gene crossover results in offspring inheriting sections of DNA from both parents. The regions involved in recombination are not selected randomly leading to an association between alleles at different loci, which is commonly referred to as linkage disequilibrium. Our hypothesis is the transition probabilities between each pair of variants in a subpopulation are sufficient statistics for the data. In other words, we believed that given only the transition probabilities for each ethnicity and a random number generator, it would be possible to reproduce a data set that would be indistinguishable from the original data in terms of its large sample properties. Furthermore, we assumed the probability of transitioning only depended on the current state, resulting in a Markov Chain. We refer to our algorithm as the Sufficient Markov Chain Transition (SMCT) algorithm.

SMCT consists of two main steps. First, the transition probabilities are calculated for each ethnicity compressing the data into a matrix with number of rows equal to the number of variants and number of columns equal to the number of ethnicities. Second, we hoped to be able to reproduce a data set equivalent to the original data set by using the transition probabilities and a random number generator. More formally, the algorithm is:

1. Calculate sufficient statistics (transition probabilities).
   - Calculate the mean haplotype for each population for the first marker.
   - For $i = 1, \ldots, k$ where k is the number of ethnicities in the dataset and j is the position index of the variant, calculate the probability, $\pi_{i,j}$, of transitioning to a new haplotype from position $j - 1$ to j in ethnicity i.
2. Import sufficient states and randomly generate states.
   - Randomly assign individuals to ethnicity groups
   - Randomly assign states based on transition probabilities.

Table 1 shows storage and runtime comparisons for SMCT. When the transition probabilities were thinned by 50%, we were able to achieve comparable size reductions as a zipped file. We were not able to get close to obtaining the compress time speed of a zipped file as it took SMCT about 10 times more time than zip to compress the data . SMCT requires one pass through the data to calculate the transition probabilities, so it is an $O(n)$ complexity algorithm where n is the number of variants.

File Type	Size (MB)	Rel. Size	Compress Time (s)	Load Time (s)
Plain Text Document	3,609.6	1.0	NA	519.74
Zipped File	72.4	0.020	60.34	543.96
SMCT File (No Thinning)	146.6	0.041	658.23	326.35
SMCT File (Thinning=2)	87.0	0.024	905.08	180.49
SMCT File (Thinning=5)	40.5	0.011	761.23	70.25
SMCT File (Thinning=10)	21.0	0.006	705.93	32.66
SMCT File (Thinning=100)	2.1	<0.001	663.23	3.33

Table 1: Storage and Runtime Comparisons.

In order to determine the subpopulations present in the data, we used Principal Component Analysis (PCA). PCA is a widely used technique to map high-dimensional data into new coordinates, called components. Each component can be thought of as the transformation that maximizes the variation in the data subject to the constraints that the norm of the transformation is 1 (or else you could just make the variance as large as you want) and the direction of the transformation is orthogonal to all previous components. PCA is generally implemented on a computer using Singular Value Decomposition (SVD). Most SVD algorithms are $O(n^3)$ complexity.

Figure 1(a) shows a scatterplot of a random subset ($\approx 1\%$) of the original data set projected onto the first two principal components. The first two principal components together explain about 75% of the variation in the data set. The 1000 genomes project classifies populations into one of five super-populations. For example, the population CHS (Southern Han Chinese) is classified into the super population EAS (East Asian). Applying this classification system to the original data, we see the first two components are sufficient to completely separate Africans, Europeans, and East Asians into distinct groups.

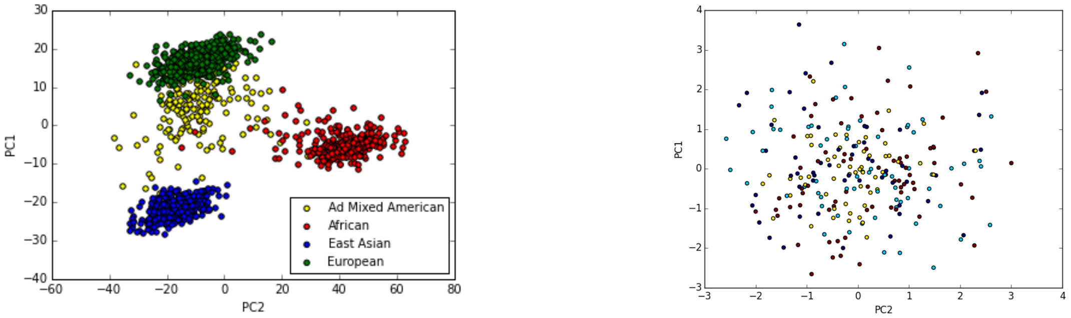

Figure 1: PCA plots of the original data (left panel) and the SMCT simulated data (right panel) by superpopulation.

Figure 1(b) shows the PCA of the simulated individual level haplotype data from SMCT. Unfortunately, SMCT was not able to adequately reproduce the data as we had expected, as the first two components of the PCA explained only 20% of the variation as compared to 75% in the original dataset and there is no apparent clustering by superpopulation. One potential improvement to the current implementation would be to calculate the probability of transitioning from state i to j conditional on the current state. Our current algorithm calculates a marginal probability of transitioning that is the same over all the possible current states.

While this current method was not successful, a procedure like this would be an important contribution to the field. Providing sufficient statistics for genetic data would delink the analysis from the individual level data, thus reducing privacy concerns, reducing costs of getting access to data, and increasing the reproducibility of research.

Works Cited

- Casella G, Berger RL. Statistical inference. Pacific Grove, CA: Duxbury; 2002.

- Golub, Gene H., and Charles F. Van Loan. Matrix computations. Vol. 3. JHU Press, 2012.

- Jolliffe, Ian. Principal component analysis. John Wiley & Sons, Ltd, 2002.

- Marchini J, Howie B, Myers S, McVean G, Donnelly P. A new multipoint method for genome-wide association studies by imputation of genotypes. Nature genetics. 2007 Jul 1;39(7):906-13.

- Ostrer H. A genetic profile of contemporary Jewish populations. Nature Reviews Genetics. 2001 Nov 1;2(11):891-8.