Compressing HapMap Data using Sufficient Subpopulation Markov Chain Statistics

Kevin Cummiskey, Caleb Lareau, Matthew Ploenzke*

Harvard School of Public Health
ploenzke@g.harvard.edu

I. Introduction

Recent advances in sequencing technology have led to an explosion in the amount of genetic data available for researchers to analyze. In light of this, cost and time efficient means of storing these large amounts of data are necessary. In what follows, we present a novel approach for reducing the size of sequencing data based on the principle of sufficient statistics [1]. Sufficiency provides a means of reducing data to a minimal set of parameters which adequately summarize the distribution. Any inference made on the full dataset then would be equivalent to inference made with a randomly-generated dataset based on the sufficient statistics.

There are two key advantages to this approach: gains in storage and the ability to adequately summarize individual level data. This latter benefit is of special importance; adequate summarization of individual level data would allow researchers to share sufficient statistics only and discontinue the need for obtaining special data access or privileges. This would have far-reaching implications on the reproducability of research as any genetic study could be reproduced by the researcher simulating a representative dataset from the sufficient statistics.

II. Theory

The key issue in undergoing the aforementioned approach is in finding the sufficient statistic(s). We base our method on the theory of linkage disequilibrium, defined as the "nonrandom association of alleles at different loci" [5]. The motivation arises from the idea the individuals with tagged variants can have their haplotypes inferred from a reference panel. If we assume that allele associations along different stretches of the chromosome are different from what they would be under an independent association (linkage disequilibrium), then we can make use of a Markov chain to summarize the probability of the next allele being a certain type. Each variant is a state and the probability of transitioning to a new state is based on the average number of transitions per pre-defined subgroup (e.g. ethnicity) at that marker.

We hypothesize that a matrix of transition probabilities provides a sufficient statistic for reconstructing population substructure. We begin by discussing the proposed algorithm and how to determine subpopulations. We then present runtime efficiency results and close by addressing the question of sufficiency.

III. Methods

Consider a matrix of size $n \times 2p$ where n is the number of markers and p is the number of individuals, each with ethnicity i. There are twice as many columns as individuals since each individual has two alleles per marker. The use of the term *marker* and *variant* will be exchangable throughout.Then compressing the data via the Sufficient Markov Chain Transition algorithm (SMCT) consists of two main steps:

- Calculate the mean haplotype for each ethnicity i for the first marker $j = 1$
- Calculate the mean probability of transitioning to a new state for ethnicity i for

*BIO234 Final Paper Submission for Matthew Ploenzke

markers $j \in \{2, n\}$

The resultant file will be a matrix of size $n \times I$, where I is the number of ethnicities reported in the input data. The compression will be by a factor of $I/2p$ and it is clear to see that the compression gains are a direct consequence of the number of ethnicities and initial sample size. Thus we may consider combining similar ethnicities if we have *a priori* knowledge of their similarity. Similarly, we may consider thinning the markers, say collapsing to every third marker and averaging the transition probability. These ideas will be discussed in detail later.

Any benefits of reducing the data to the sufficient statistics are lost if we are unable to reproduce a randomly-simulated sample from the statistics with features representative of the original data. Consider the output from the SMCT algorithm as input, i.e. a $n \times I$ matrix. Then the data regeneration process may be summarized as follows:

- Randomly assign p individuals to I ethnicity groups
- Randomly generate the first haplotype for each individual based on mean haplotype for ethnicity i
- Randomly transition to new states based on the sufficient markov chain transition probabilities

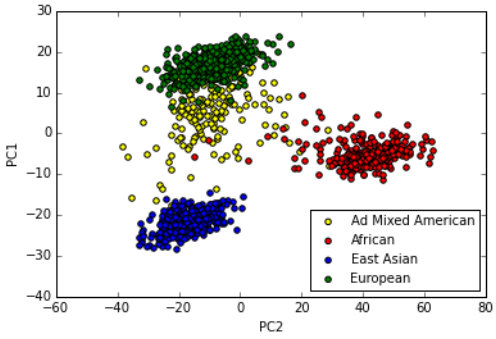
The algorithm presented above has expected complexity $O(p)$, given n, since each individual must have their two alleles simulated across the entire chromosome. We see that the runtime of this process is heavily dependent on the size of the sample one wishes generate as well as the number of markers and ethnicities. Before we can address the sufficiency of the markov chain transition statistic, we must first define the sort of population substructure we wish to reproduce.

IV. Results

I. Defining Substructure

There are many popular methods for determining similarities between observations. One such technique is Principal Components Analysis (PCA) and is widely used in the field of genetics [4, 6]. Glossing over much of the mathematics, PCA can be understood intuitively as a procedure which uses orthogonal components from a singular value matrix decomposition to "explain" as much of the variance in the data as possible. These components are ranked such that first principal component accounts for the most variance in the data while each subsequent orthogonal component accounts for less. In terms of data reduction, if the first few components explain a high proportion of the overall variance then the subsequent components may be discarded without losing information regarding the variance.

Considering again the problem at hand, we define subpopulations by computing the variance/covariance matrix of the haplotype matrix, performing a PCA, and plotting the first two components. These first two components were found to account for nearly 75% of the variation in the data and when we color the observations by their reported Super-Population mapped ethnicity we see four distinct clusters (figure below)[1]. A brief overview of the SMCT algorithm efficiency will be now given before closing with the simulation results.



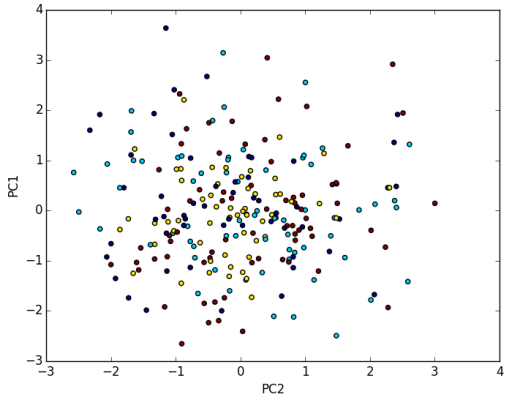
[1] Super-Population mapping is available on the 1000 Genomes Project website

II. Efficiency Analysis

The SMCT algorithm was implemented in both julia-0.3.11 and R-3.2.2. The primary computational overhead is in loading and storing the expression matrix and we therefore report results from the much quicker julia runtimes. The simulation analysis is based on a local machine with a 1.7 GHz Intel Dual-Core i7 Processor and 8 GB RAM. The raw test file was just over 3.6GB unzipped and included variants along chromosome 22. The zipped file could be compressed to 2% (72.4MB) the size of the raw data. In the absense of thinning (collapsing and averaging the transition probabilities), we were able to achieve a filesize twice that of the zipped file (146.6MB) when we collapsed to 14 ethnicities. We could of course zip the resulting file with no loss and achieve a much lower size (11.6MB). We introduced the idea of thinning earlier and thinning to every other transition probability halved the filesize as expected, at the cost of a 140% increase in runtime. The *increase* in runtime for the thinning process is a result of needing to calculate the full transition matrix first and subsequently collapsing. Overall, runtime was around a 10-fold increase to that of standard zip compression.
Runtime could be greatly shortened through code parallelization however our interest lies in determining whether the aforementioned population clusters are reproducible via the sufficient statistics. To address this, datasets were simulated from the transition probabilities for both the 14 reported ethnicities as well as the Super-Population mapping. We set the total number of simulated individuals to be 500 (1000 simulated alleles) and note the simulation times were comparable to that of the unzip-and-load time for the zip-compressed file. We now addres the primary question of interest, i.e. are ethnicity-level transition probabilities sufficient to reproduce the clustering among the first two principal components?

III. Simulation Analysis

As was previously mentioned, datasets were simulated for both the 14 individual ethnicities and the 4 Super-Population groups. No thinning, thinning to every second, fifth, tenth, and hundredth marker was also included, resulting in 10 simulated datasets. In all cases we were unable to capture the population clusters seen previously, as the plot below clearly indicates no discernable clustering. In fact, the first two principal components only accounted for around 20% of the total variance whereas in the original data they captured 75%. Given the theory mentioned in section I, we are very disappointed by these results.



V. Conclusion

We hypothesized that a matrix of transition probabilities would provide a sufficient statistic for reconstructing population substructure as defined by the clustering of the first two principal components. While the proposed algorithm had some attractive features and was believed to be founded upon theory, we were unable to reproduce meaningful substructure. We do believe this approach has merit and there are alternative statistics which may prove sufficient. The current algorithm uses unconditional transitions, however an alternative would be to use conditional transitions, calculated and stored at the expense of compression and efficiency gains. Another explanation for the negative results is that we simply did not simulate large enough samples to capture the true ethnicity variance structure. Indeed more work should be done on this topic before a conclusion is reached but we believe the high level idea of

sufficiency warrants further consideration.

VI. Notes

- Code is located at `https://github.com/caleblareau/BIO234FinalProject`.
- My primary involvement in the project was the algorithm development and implementation. As such, my paper focuses more heavily on that aspect and brushes over the theory and PCA.
- The transition matrix is located in the github repo as well.

References

[1] Casella G, Berger RL. Statistical inference. Pacific Grove, CA: Duxbury; 2002.

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

[3] Jolliffe, Ian. Principal component analysis. John Wiley & Sons, Ltd, 2002.

[4] 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.

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

[6] Yeung, Ka Yee, and Walter L. Ruzzo. Principal component analysis for clustering gene expression data. *Bioinformatics* 17.9 (2001): 763-774.