Imputation using the linkage disequilibrium method

Steven Snyder - <stsnyder@ucla.edu>
University of California – Los Angeles
Thursday, June 5, 2008

Abstract

Imputation is a method of dealing with missing data points by filling in values. In genetics, imputation generally refers to the substitution of missing SNP values. Missing SNP values commonly cause data to be thrown out, as re-genotyping is limited by financial constraints. Recovery of SNP values can keep costs down and restore power lost due to missing data. The linkage disequilibrium (LD) method imputes the value of missing SNPs based on LD correlation data between missing SNPs and SNPs which have been measured successfully. This approach is easy to implement, is generalizable, has decent accuracy, and is fairly fast. On the other hand, it does not have optimal accuracy, and makes decisions without an explicit statistical confidence value.

1. Introduction

Imputation is a method of dealing with missing data points by filling in values. This can be as simple (and naïve) as choosing a random value from the domain of the data, or as complicated as analyzing all measured values to determine the most likely value for missing data points. The latter approach often takes advantage of existing data from other experiments (the reference data) to improve the accuracy of the imputation.

In genetics, imputation usually refers to the substitution of missing SNP values. Missing SNP data is fairly common in association studies, sometimes with rates as high as 5-10% [1]. Individuals with missing SNP data are usually thrown out, decreasing the effective sample size and thus the power of the study. Since re-genotyping is usually not possible due to financial constraints, recovery of SNP values using imputation can keep costs down and restore some of the power lost from error-causing processes.

2. Methods of Imputation

2.1. Naïve Imputation of SNPs

The most simple method for imputing the value of missing SNPs is the major allele method. With this method of imputation, the major allele (the allele occurring most often in the population at this SNP location) is substituted for the missing value. Naïve SNP imputation usually results in a high error rate, equal to 1-P where P is the proportion of the major allele in the population.

2.2. EM Algorithm / Haplotype-Based Imputation

Haplotype-based imputation uses the conditional probability of the value of a SNP based on various covariates including case/control group placement of the individual, known disease factors, and other properties not necessarily within the haplotype itself. It also takes into account internal covariates such as linkage disequilibrium.
The EM algorithm finds the maximum likelihood SNP values that conform to the unphased haplotype data.

2. EM Algorithm / Haplotype-based Imputation

Linkage disequilibrium (LD) refers to the correlation of SNP values caused by their tendency to travel together during recombination. This generally occurs because SNPs are close together, with crossing over and other recombination events unlikely in the few base pairs between them. The HapMap project studied this phenomenon in humans, and provided us with an understanding of LD in the human genome that LD can help us with a successful approach to imputation.

Using the *r^2* values from the HapMap project, we can find measured SNPs correlated with missing SNPs. We can determine the most likely value of the missing SNP based on the values of the correlated SNPs. Multiple correlated SNPs can be used for imputation by averaging their *r^2* values, signed according to whether or not they imply the major (+) or minor (-) allele. If the resulting value is positive, the likelihood that the missing SNP is of the major allele is higher. If the value is negative, the minor allele has a higher likelihood.

3. LD Imputation Implementation

I implemented the LD method using the C++ programming language. There are a few reasons why the LD method was chosen. The two most important reasons why the LD method was chosen are because it is fairly accurate (within 2% error rate of other methods) and easy to implement. Additionally, it is generalizable in that it doesn't depend on associations besides LD data, which we gather directly from the reference haplotypes (sometimes referred to as the training data set).

In my implementation, each step of the imputation process has been abstracted from the rest. This allows functional components to be replaced with different implementations that retain the same interface. For example, precomputed correlation data could be stored on a web server, as it is on the HapMap project. In my current implementation, correlation data is computed locally, but by changing a few functions in the Correlation class, it could retrieve the web-based data and imputation could proceed as usual. I implemented the software using this approach to also make it easy to distribute as a library of imputation-related functions.

3.1. Algorithm Overview

For each missing SNP in the input haplotype, a stack of correlated SNPs (the “implicators”) is generated by calculating the *r* correlation value between the missing SNP and measured SNPs[1], using the first-allele[2] probabilities rather than the minor allele probabilities, which are more commonly use for calculating the *r* value. The major/minor allele status is abstracted from the process by using the first-allele value for all calculations where normally the minor allele would be used. This does not affect the value of the result, but can make the sign vary from that calculated using the other method. However, the sign is appropriate as we understand that the direction of positive association is from the first allele to the first allele, rather than from minor allele to minor allele between the SNPs.
Implicators associated below a threshold of *r*^2 = 0.35 are ignored to prevent poorly-associated SNPs from decreasing the accuracy of the imputation.

For each implicator, the signed *r* value is added or subtracted to a running sum that starts at zero for each new missing SNP. If the implicator's measured SNP is of the first-allele, the *r* value is added. Otherwise it is subtracted.

When all implicators have had their correlation values added to the sum, the value of the SNP can be imputed. If the sum is positive or equal to zero, the first-allele of the missing SNP is the imputed value. Otherwise the other allele is assigned to the missing SNP.

A confidence value is assigned by dividing the sum by the number of implicators. While this does keep the range between -1.0 and 1.0, the value has no statistical significance, being only a rough indication of the confidence in the imputation value.

4. Results

This method produces a 6% error rate when used on an input haplotype of 198,000 SNPs with 5% of the SNPs randomly deleted. The reference data contained 120 haplotypes from the HapMap phased data.

This algorithm is fast even on such a large data set. The total calculation took 45 minutes on a 2.4Ghz single-core Pentium 4 machine and used 500-megabytes of RAM. A Pentium M machine running at 3 Ghz performed the imputation in 39 minutes.

4.1. Drawbacks

Unfortunately, the 6% error rate of the LD imputation method is not optimal among all methods of imputation. Other methods produce an error rate of 4-5% [1].

The LD imputation method doesn't produce a statistically usable confidence value. The confidence value is dependent on the number of implicators, and may even decrease as the number of implicators increases. It is not easy to interpret unless at one of the extremes (-1 or 1) where it indicates that all correlated SNPs were perfectly correlated and indicated the same value.

Due to the fact that my implementation of the LD imputation method works on a per-haplotype basis, its results are dependent on haplotype phasing, which may have errors.

4.2. Future Improvements

There are a few possible future improvements for this implementation of the LD imputation method. One is to use precomputed *r*^2 values from HapMap project or another source of correlation data. The *r*^2 values are the most computationally expensive parts of the imputation, so improving performance here would greatly reduce the time required to impute a haplotype. On the other hand, relatively few *r*^2 values are

[1] Note that SNPs with a distance greater than 5,000,000 bases from the missing SNP are not considered for the calculation as they are almost invariably uncorrelated.

[2] The *first-allele* is the allele assigned to the number '0' in the haplotype legend file. This is arbitrary in the scope of an imputation, but must be used consistently throughout the calculation of *r* values and the imputation.

required, and according to HapMap data [2], correlation is specific to some populations, so it may be desirable to calculate the r^2 values directly and (only as needed) from local reference data.

A statistically useful confidence value for each imputed SNP would greatly increase the utility of the LD method. The confidence value would allow imputed SNPs to be easily used in association studies without adding an unknown factor to noncentrality parameters.

4.3. Conclusion

Overall, the performance of my implementation of the LD imputation method is good, both in error rate and processing speed.

With a few further optimizations to improve the error rate, and a way of producing a statistically significant confidence value for imputations, the implementation could find use in high-throughput applications where the addition one or two percentage points of accuracy from the other more computationally intensive methods are not required.

4. References

[1] J. Dai et al., Imputation Methods to Improve Inference in SNP Association Studies, Genetic Epidemiology 30 (2007) 690-702
[2] The International HapMap Consortium, A haplotype map of the human genome, Nature Vol 437 doi: 10.1038/nature04226 (2005) 1299-1320