Imputation

Steven Snyder
CS 124 - Spring 2008

Introduction - What is Imputation?

Imputation is a method of dealing with missing data points by filling in values.

This can be as simple (and naive) as choosing a random value from the domain of the data, or as complicated as analyzing all existing values to determine the most likely value for missing data points.

Imputation of SNPs

What is imputation in genetics?

- In genetics, imputation usually refers to the substitution of missing SNP values

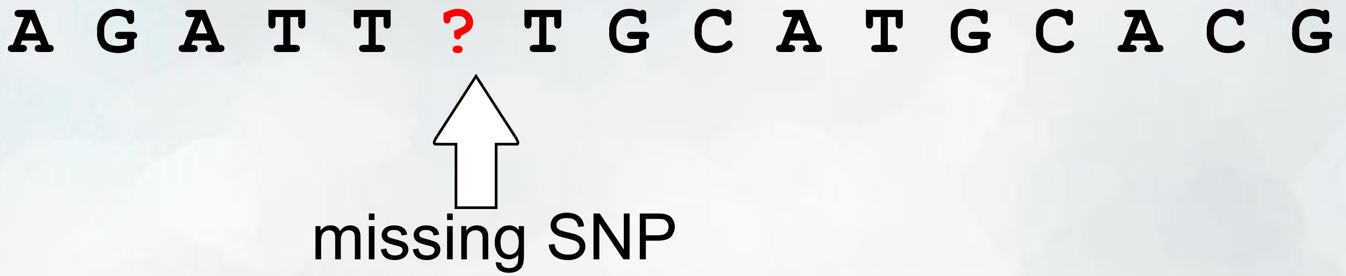
Why should we use imputation?

- Missing SNP data is fairly common in association studies, sometimes with rates as high as 5-10% [J. Dai, et al 2006].
- Re-genotyping is usually not possible due to financial constraints.
- Individuals with missing SNP data are usually thrown out, decreasing the effective sample size.
- Recovery of SNP values can keep costs down and restore some of the power lost by errors in data

How do we use SNP imputation?

Example:
We measured 16 SNPs from an individual, but the value of one SNP was missing due to lab equipment problems.

- Suppose we measured and phased the following sequence of SNPs for one of the individual's haplotypes:



A G A T T ? T G C A T G C A C G

missing SNP

- If we could impute the value of the missing SNP, we wouldn't have to re-sequence the individual.

How do we impute? - Naive Method

Example (continued):

Naive imputation: insert the major allele!

A G A T T ? T G C A T G C A C G

⇓

A G A T T A T G C A T G C A C G

Now we have the value of the missing SNP.

What if the individual's haplotype actually had the minor allele? Clearly this isn't the best way to do it!

How do we impute? - LD Method

Example (continued):

Linkage disequilibrium (LD) method:

A G A T T ? T G C A T G C A C G

Suppose we have LD data from the population that this individual comes from.

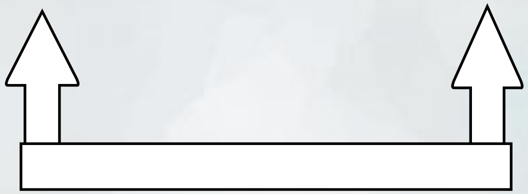
If any of the SNPs we measured are in LD with the missing SNP, we can use this information to predict the value of the missing SNP.

- This is the method I chose to work with.

How do we impute? - Correlation

Linkage disequilibrium (LD) method:

A G A T T ? T G C A T G C A C G



LD between these two SNPs

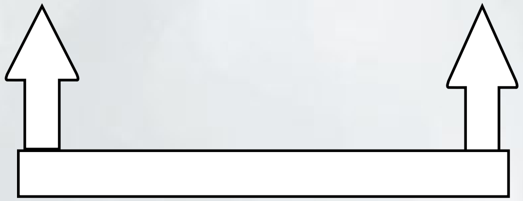
From the HapMap project we can get a measure of the LD between the missing SNP and SNPd that we have measured successfully. This is generally expressed as the r^2 value.
The r^2 value tells us the strength of the association between pairs of SNPs.

How do we Impute? - Correlation

Back to the example:

Let's suppose that the r^2 correlation value between our missing SNP and the SNP shown in blue is 0.85.

A G A T T ? T G C A T G C A C G



r^2 = 0.85

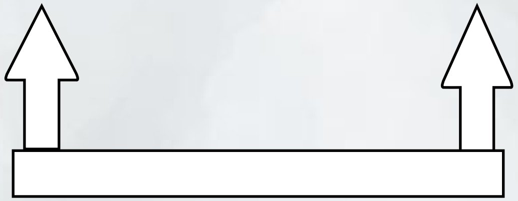
This means it is highly likely that the major allele of our missing SNP occurs when we see the major allele of the blue SNP, and vice versa. In practice, this is not always true, but we assume it is for now. (More on this later.)

- In this case suppose that T is the minor allele of the blue SNP.

How do we Impute?

Let's assume that the r^2 correlation value between our missing SNP and the SNP shown in blue is 0.85.

A G A T T ? T G C A T G C A C G



r^2 = 0.85

Since the correlated SNP has the minor allele in this individual, we impute the missing SNP's value to be its minor allele.

A G A T T C T G C A T G C A C G

Multiple correlated SNPs

What if there are multiple correlated SNPs?

- We can use multiple correlated SNPs by averaging their r^2 values, signed according to whether or not they imply the major (+) or minor (-) allele.
- If the resulting average is positive, we choose the major allele. If the resulting value is negative (less than zero), we choose the minor allele.

Imputation of SNPs - Methods

There are other, more complex approaches to SNP imputation.

EM algorithm / Haplotype-Based Imputation

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

Imputation of SNPs - My Method

Why my approach? (Linkage Disequilibrium method)

- Easy to implement
- Generalizable -- doesn't depend on associations besides LD data (which we gather directly from the reference haplotypes)
- Good accuracy - 6% error rate when ~5% of SNPs are missing from a ~195,000 SNP haplotype
- Can be used to impute values on individual haplotypes, rather than the genome.
    - Framework is easily modified to impute SNPs in genotypes.
- Fast (~8000 missing SNPs in 200,000 SNP chromosome imputed in 45 minutes!)

My Method (cont.)

Drawbacks of my method

- Doesn't have optimal accuracy
  - Other methods have as low as 4 or 5% error rate [J. Dai et al. 2007]
- Doesn't produce a statistically usable confidence value
  - Confidence value is dependent on sample size (and can decrease due to increased sample size)
  - Not easy to interpret unless it is -1 or 1
- Results are dependent on haplotype phasing, which may have errors

Optimizations and Fixes

There are a few problems I had to deal with in my algorithm that were not expressed in the example.

Problem:

In practice, we can't use the r^2 value directly for per-haplotype SNP imputation, due to it being ambiguous for some cases.
For example, an r^2 of 1.0 indicates that two SNPs are perfectly correlated. However, if the SNP has two alleles of similar frequency, the r^2 value cannot tell us which allele from the first SNP is correlated with which allele from the second SNP.

Optimizations and Fixes

My fix:

- Calculate the *signed* pairwise r^2 values between SNPs.
- This is done by calculating the r value directly, then squaring it and assigning the r^2 value the sign of r.
  - If allele 0 of SNP A is correlated with allele 1 of SNP B, the r^2 value is signed positive.
  - If the allele 0 of SNP A is correlated with the allele 1 of SNP B, the r^2 value is signed negative.
- This abstracts the concept of major and minor alleles.
- Using the minor/major allele association idea resulted in 10% error rate... this fix reduced that to 6%.

Optimizations

Optimizations:

- Only use SNPs for imputation if they are well-correlated with the missing SNP, otherwise they may bring the accuracy down.
  - I conducted trials with r^2 thresholds from 0.10 to 0.80 in increments of 0.05. Accuracy ranged from 6 to 13%. The optimal value was 0.35.
- Only check r^2 values if SNPs are within a certain number of base pairs. SNPs that are far away from each other are less likely to be associated, so its a waste of time to calculate the r^2 values between them.

Demonstration

Demonstration of imputation software.

Future Improvements

- Use pre-computed r^2 values from the HapMap project.
    - This would greatly improve performance because the r^2 calculations are one of the most computationally expensive parts of the imputation.
- Develop statistically usable confidence value.
- Add more processing options:
    - Partial chromosome imputation
    - User-modifiable r^2 and distance thresholds
    - Automatically download HapMap reference data