GWAS 10: Genotype imputation

Matti Pirinen, University of Helsinki

27-Feb-2019

The slide set referred to in this document is "GWAS 10".

10.1 Genotype imputation

Since long haplotype segments are shared between individuals, the reference panels of haplotypes in different human populations are useful, e.g., to give information for phasing or for choosing tag-SNPs to include on the genotyping arrays. The HapMap project, the 1000 Genomes project, and, more recently, the Haplotype Reference Consortium have collected such data. One of the main use of the haplotype structure in GWAS context is the genotype imputation.

Genotype imputation (slides 2-9) is the process of predicting or imputing genotypes that are not directly genotyped in a sample of individuals. Most often imputation is done using a reference panel of haplotypes at a dense set of SNPs to impute into a study sample of individuals that have been genotyped at a subset of the SNPs of the reference panel. A review on imputation by Howie & Marchini (2010). Imputation software: IMPUTE4, Beagle and Minimac4.

As an example, the UK Biobank has released genotype data together with thousands of phenotypes on 500,000 British volunteers for research purposes. The samples were genotyped using a genotyping chip with a bit less than 1 million SNPs, but then imputed to 80 million variants using the existing haplotype reference data. The whole imputed data set is 2 terabytes. The GWAS results on many traits are available for browsing in Global Biobank Engine or GWAS atlas and also available for download through Ben Neale's lab.

Imputation is important to harmonize cohorts to have the same set of variants before meta-analysis. Imputation is also needed to have as a comprehensive set of variants as possible to be used in fine-mapping.

Imputation produces probabilistic genotype predictions, i.e., we don't know with certainty whether each individual has genotype 0, 1 or 2, but instead a probability model gives a probability distribution $(p_{il0}, p_{il1}, p_{il2})$ that describes what are the probabilities that individual i has genotype 0, 1 or 2 at locus l. From this distribution, we can compute the expected value of the number of copies of allele 1, so called allele 1 **dosage** as : $\overline{x}_{il} = 0 \cdot p_{il0} + 1 \cdot p_{il1} + 2 \cdot p_{il2} = p_{il1} + 2 \cdot p_{il2}$. This dosage is then used in the GWAS analysis as if it was an observed genotype. Regression models show their flexibility through their ability to include the imputed dosages in the analysis without any changes to the analysis pipeline of directly observed genotypes. (ANOVA-type group comparisons are not similarly flexible.)

Example 10.1. Let's consider three SNPs 1,2 and 3, which are not on the genotyping chip that we have used to genotype individual i, but which are present in our imputation reference panel, with allele 1 frequencies 0.32, 0.42 an 0.01, respectively. Suppose our imputation results for individual i are

SNP	p_0	p_1	p_2	AF1
1	0.8	0.19	0.01	0.32
2	0.01	0.99	0.0	0.42
3	0.98	0.02	0.001	0.01

These tell us that we are quite confident that i has genotype 1 at SNP 2, but we have a considerable uncertainty about the genotype at SNP 1. We come back to SNP 3 later.

We say that the imputed genotypes are **correctly calibrated**, if, among all individual-locus pairs for which the algorithm estimates genotype g with a probability of p, a proportion of p indeed have the genotype g. For example, if we collect all the individual-locus pairs, such as SNP 1 for individual i in Example 10.1 above, where the imputation algorithm has proposed that the individual-locus pair corresponds to genotype 0 with probability in (0.79,0.81), a proportion close to 0.80 do indeed have the genotype 0. Of course we cannot know that the algorithm is correctly calibrated for variants we have not genotyped, but we can at least test the algorithm by masking some of the known genotypes and imputing them and comparing those results to the truth.

10.1.1 Imputation quality metrics

Assuming that the probabilistically imputed genotypes are correctly calibrated, we define a measure of the available information given by imputation compared to the perfect information we would have, had we genotyped the locus of interest.

The highest possible information corresponds to the case where no individual has any uncertainty in their imputed genotype. Technically, this happens when the genotype distribution of every individual has one of the probabilities p_0, p_1 and p_2 equal to 1 and the other two equal to 0. The **variance** v_{il} of individual i's imputed genotype distribution at locus l:

$$v_{il} = \mathrm{E}(x_{il}^2|\boldsymbol{p}_i) - \mathrm{E}(x_{il}|\boldsymbol{p}_i)^2 = 4 \cdot p_{il2} + 1 \cdot p_{il1} + 0 \cdot p_{il0} - \overline{x}_{il}^2 = 2\,p_{il2} + \overline{x}_{il} - \overline{x}_{il}^2,$$

is 0 exactly when there is no uncertainty in the genotype. On the other hand, if LD from imputation reference panel has not provided any new information about the genotype of individual i over what we already had by knowing the allele 1 frequency f_l in the reference panel, then the variance of the imputed genotype for individual i corresponds to the HW variance $w_l = 2f_l(1-f_l)$. Thus, if the observed variance v_{il} equals to the HW variance w_l, then we have not gained any new information from imputation. Based on these two extreme cases of either perfect information or no information, we define the imputation INFO measure for locus l as

$$\mathrm{INFO}_l = 1 - \frac{1}{n}\sum_{i=1}^{n}\frac{v_{il}}{w_l},$$

where the sum is over all imputed individuals. When INFO is 1, there is no uncertainty after imputation, and if INFO is 0, then we have not learned anything new from the imputation over what we could had already guessed based on the population allele frequencies. (Technically, INFO could also go negative, which is interpreted as INFO=0.) This measure is used by versions of the IMPUTE software.

Minimac uses a very similar information measure except that the observed variance is not computed as mean over individuals' variances but instead mean of the allele dosages are used to derive the observed variance. For more details about the info measures see Supplementary 2 from Marchini & Howie 2010.

Example 10.2. Consider the imputation probabilities given above in Example 10.1. The variances of the imputed genotype distributions for individual i are

```
p = matrix(c(0.8, 0.19, 0.01,
             0.01, 0.99, 0.0,
             0.98, 0.02, 0.001), byrow = T, ncol = 3)
f = c(0.32, 0.42, 0.01)
cbind(p,f) #check that these are correct
```

```
##                         f
## [1,] 0.80 0.19 0.010 0.32
## [2,] 0.01 0.99 0.000 0.42
## [3,] 0.98 0.02 0.001 0.01
```

```
x = p %*% 0:2 #allele dosages
v = 2*p[,3] + x - x^2 #variance of imputed genotype distribution
w = 2*f*(1-f) #HW genotype variance
info = 1-v/w #info based on one individual
info
```

```
##            [,1]
## [1,]  0.5728401
## [2,]  0.9796798
## [3,] -0.1876768
```

We see that even if loci 2 and 3 have similar looking distributions (highest value 0.99 or 0.98), still they have very different info. This is because the distribution at SNP 3 corresponds to the population frequencies under HWE and thus there is no information from imputation, whereas at SNP 2, the imputation has clearly provided new information.

10.1.2 Summary statistics imputation

Suppose that we have access to only the summary association statistics at set T of SNPs genotyped at a GWAS. We would like to impute the summary statistics to a set U of untyped SNPs that we have in our imputation reference panel. The individual level genotype imputation cannot be applied because we don't have access to any individual level genotypes. However, we can impute the summary statistics based on the LD between SNP sets T and U estimated from the reference panel and the observed summary statistics at T. We use a standard assumption that the z-scores of the SNPs follow the multivariate Normal distribution. This approach was introduced by Pasaniuc et al. 2014 to their software Imp-G.

Since each untyped SNP will be imputed on its own, we can derive the formulas for imputing a single untyped SNP u. We assume that, after reorganising the SNPs so that the typed SNPs T come first and u is the last SNP, the (prior) distribution of the z-scores is

$$\boldsymbol{z} = \begin{bmatrix} \boldsymbol{z}_T \\ z_u \end{bmatrix} \sim \mathcal{N}_L \left(\boldsymbol{0}, \left[\begin{array}{c|c} \boldsymbol{R}_{TT} & \boldsymbol{R}_{Tu} \\ \hline \boldsymbol{R}_{uT} & \boldsymbol{R}_{uu} \end{array} \right] \right)$$

where we have divided the LD matrix $\boldsymbol{R}$ into 4 parts according to SNP sets T and u. Given the obsereved z-scores $\boldsymbol{z}_T$ at the typed SNPs, we then use the conditional distribution for the untyped SNPs, that, by the properties of the multivartiate Normal distribution, is

$$(z_u \,|\, \boldsymbol{z}_T) = \boldsymbol{R}_{uT} \boldsymbol{R}_{TT}^{-1} \boldsymbol{z}_T.$$

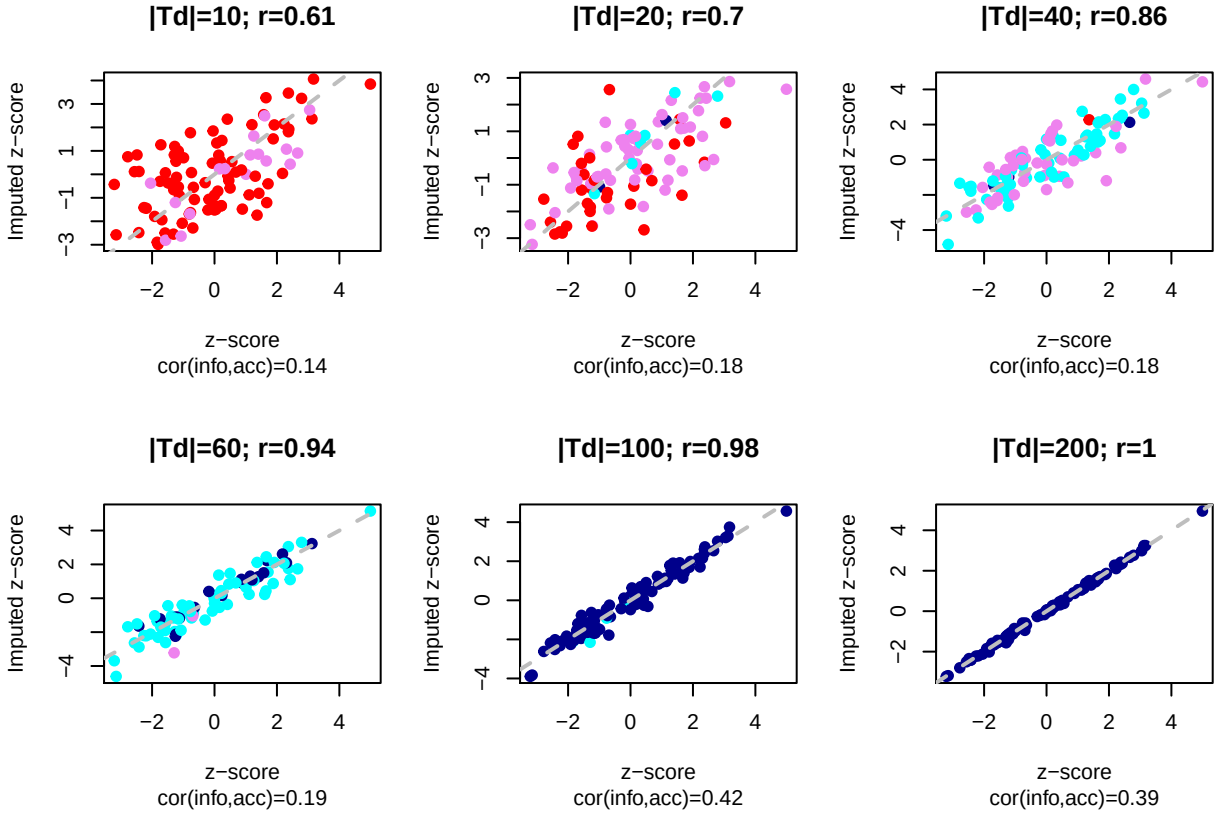
Under the null, this z-score has variance of $v_{u|T} = \boldsymbol{R}_{uT} \boldsymbol{R}_{TT}^{-1} \boldsymbol{R}_{Tu}$ and we use this variance as an information measure of the imputed z-score. If this info value is close to 1 (the true variance of the null z-scores), then the LD structure is very informative about u given T, whereas if there is little correlation between u and T, then this info value will be close to 0.

Note that since some information is lost in imputation compared to the directly observed z-scores, the imputed z-scores will be deflated under the null. This is a consequence of *regression dilution bias*, that causes the regression coefficients to shrink towards zero when predictors in regression model are measured with uncertainty. To correct for this dilution in imputed z-scores, we need to scale them by $1/\sqrt{v_{u|T}}$. This scaling option is the default z-score in Imp-G and we will also use it in the example below. Thus, our imputed z-score is

$$(z_u \,|\, \boldsymbol{z}_T) = \frac{\boldsymbol{R}_{uT} \boldsymbol{R}_{TT}^{-1} \boldsymbol{z}_T}{\sqrt{v_{u|T}}} = \frac{\boldsymbol{R}_{uT} \boldsymbol{R}_{TT}^{-1} \boldsymbol{z}_T}{\sqrt{\boldsymbol{R}_{uT} \boldsymbol{R}_{TT}^{-1} \boldsymbol{R}_{Tu}}}.$$

Example 10.3. Let's do a GWAS using data from Exercise 6.5 and choose randomly 100 SNPs to be imputed using 20 to 200 other SNPs and compare the imputation results to the true z-scores. Let's color the SNPs according to the estimated imputation info from the summary statistics imputation.

```r
set.seed(118)
n = 1000
p = 2000
X = matrix(scan("https://www.mv.helsinki.fi/home/mjxpirin/GWAS_course/2019/material/ex5.2.txt"), byrow =
R = cor(X) #LD matrix
y = rep(c(1,0),c(n/2,n/2)) #1st 500 cases rest controls
gwas = t(apply(X, 2, function(x){ summary( glm( y ~ x, family = "binomial") )$coeff[2,] }))
Ud = sample(1:p, size = 100) #indexes of the "untyped" SNPs to be imputed
par(mfrow=c(2,3))
par(mar=c(5.5,4.5,4.5,1))
for(n.Td in c(10,20,40,60,100,200)){ #
  Td = sample(setdiff(1:p, Ud), size = n.Td)
  R.TdUd = R[Td,Ud] #LD of typed - untyped pairs
  inv.R.Td = solve( R[Td,Td] + 0.001*diag(n.Td) ) #inverse of LD of typed SNPs
  W = R[Ud, Td] %*% inv.R.Td #these are the weights that turn typed to imputed
  infos = as.numeric(rowSums(W * R[Ud, Td])) #info measures per each untyped
  z.imp = (W %*% gwas[Td, 3])/sqrt(infos) #use scaling 1/sqrt(infos) to get var=1 for z-scores
  cols = rep("darkblue", length(Td))
  cols[infos < 0.75] = "cyan"
  cols[infos < 0.5] = "violet"
  cols[infos < 0.25] = "red"
  plot(gwas[Ud,3],z.imp, col = cols, pch = 19,
       main = paste0("|Td|=",n.Td,"; r=", signif(cor(gwas[Ud,3],z.imp),2)),
       xlab = "z-score", ylab = "Imputed z-score",
       sub = paste0("cor(info,acc)=", signif( cor(-(gwas[Ud,3]-z.imp)^2,infos, method="spearman"),2) ))
  abline(0,1, lty=2, col="gray", lwd=2)
}
```



We can see how the accuracy of the imputed z-scores gets better as we include more SNPs to be used in the imputation. The info measure is also able to tell when the LD-structure is more informative (dark blue) and when it carries little info (red/violet). The correlation value below plots show how the accuracy of the imputed z-scores is correlated with info measure. When info is well calibrated, we expect a positive correlation where increasing info goes with higher imputation accuracy.