IGSS 2020 Workshop in Statistical Genetics: Computer Exercise

September 17, 2020

Alexander Young
alextisyoung@gmail.com

1 Introduction

In this exercise, we introduce the relatedness disequilibrium regression (RDR) method for estimating heritability from SNP data[1]. We introduce the SNIPar software package, and show how to use it to impute missing parental genotypes from siblings and parent-offspring pairs, and to use imputed/observed parental genotypes to estimate direct and indirect genetic effects[2].

2 Installation

The exercise is performed within a customized distribution of the SNIPar package. It is necessary to install the SNIPar package to perform the exercise. You will need an installation of Python 3 to install the package. (Note that the exercise has been setup and tested for a 64 bit Linux environment.)

In the main SNIPar directory, use the following commands to install the package in a virtual environment:

```
python3 -m venv env
source env/bin/activate
python setup.py install
```

To compile the C modules in the package and to ensure everything is working properly (do not skip this), use the following command:

```
python setup.py pytest
```

This should complete without errors. Note that the tests may take some time.

3 Exercise data

The data we will use is contained within the IGSS subdirectory of the SNIPar package. This is also where additional scripts used for the exercise are contained. Please change to the IGSS subdirectory. In here, you will find genetic data from a simulated population and a simulated trait.

A sample of 4,500 independent families was simulated, where each family consists of a mother and a father, and their two full sibling offspring. Parents' genotypes were simulated independently, so that no assortative mating or population structure is considered here. The genotype data at 1,000 independent SNPs for this sample is contained in Plink format in sample.bed. To illustrate the imputation of missing parental data, we also provide genotype data for a subsample of the full sample where genotypes of some family members are missing in sample_reduced.bed. In the reduced sample, there are 1,500 families with both parents missing but both siblings genotyped, 1,500 families with one parent missing and between one and two siblings genotyped, and 1,500 families with both parents genotyped and between one and two siblings genotyped.

The simulated trait is in y.ped. The trait was simulated by giving each of the 1,000 SNPs direct, paternal, and maternal effects drawn from a normal distribution. Let δ_l, η_{pl}, and η_{ml} be the direct, indirect paternal, and indirect maternal effect of SNP l respectively; then the phenotype of individual j from family i is

$$Y_{ij} = \sum_{l=1}^{1000} \delta_l g_{ijl} + \sum_{l=1}^{1000} \eta_{ml} g_{m(i)l} + \sum_{l=1}^{1000} \eta_{pl} g_{p(i)l} + \epsilon_{ij}, \tag{1}$$

where g_{ijl} is the genotype of individual j from family i at SNP l, $g_{m(i)l}$ is the genotype of the mother in family i at SNP l, and $g_{p(i)l}$ is the genotype of the father in family i at SNP l. We drew effects from a multivariate normal distribution independently for each SNP:

$$\begin{bmatrix} \delta_l \\ \eta_{pl} \\ \eta_{ml} \end{bmatrix} \sim \mathcal{N}\left(0, \frac{v}{1000} \begin{bmatrix} 1 & 0.5 & 0.5 \\ 0.5 & 1 & 0.5 \\ 0.5 & 0.5 & 1 \end{bmatrix}\right), \tag{2}$$

where v was chosen so that the total variance explained by direct, paternal, maternal effects was 0.8. The residual variance was drawn from a normal distribution with variance 0.2 independently for each individual, giving a total trait variance of 1.

4 RDR and GREML

The first part of the exercise looks at estimating heritability using the RDR method, which controls for parental genotypes/relatedness, and compares this to the GREML[3], which does not.

If we consider the trait model above with genotypes normalised to have mean zero and variance 1, then we have that

$$\mathrm{Var}(Y_{ij}) = \sum_{l=1}^{1000} \delta_l^2 + \sum_{l=1}^{1000} \eta_{ml}^2 + \sum_{l=1}^{1000} \eta_{pl}^2 + \sum_{l=1}^{1000} \delta_l \eta_{pl} + \sum_{l=1}^{1000} \delta_l \eta_{pl} + \mathrm{Var}(\epsilon_{ij}), \tag{3}$$

where we have used the fact that the correlation between parent and offspring genotype is 1/2.

The components of the variance can be related to the RDR variance decomposition. Due to the fact that offspring genotype is conditionally independent of environment given parental genotype, the variance of a trait can be decomposed as:

$$\mathrm{Var}(Y) = v_g + v_{e\sim g} + c_{e,g} + \mathrm{Var}(\epsilon), \tag{4}$$

where v_g is the variance due to direct effects, which is equal to the heritability when divided by the phenotypic variance; $v_{e\sim g}$ is the variance of the environmental component of the trait that is correlated with parental genotype; $c_{e,g}$ is the variance due to covariance between direct and environmental effects; and $\mathrm{Var}(\epsilon)$ is the variance of the component of the trait that is uncorrelated with parent and offspring genotype.

For this simulation, we have that

$$v_g = \sum_{l=1}^{1000} \delta_l^2 = v;\ v_{e\sim g} = \sum_{l=1}^{1000} \eta_{ml}^2 + \sum_{l=1}^{1000} \eta_{pl}^2 = 2v;\ \text{and } c_{e,g} = \sum_{l=1}^{1000} \delta_l \eta_{pl} + \sum_{l=1}^{1000} \delta_l \eta_{pl} = v; \tag{5}$$

where we have used the fact that direct and parental effects are correlated to derive $c_{e,g}$. This implies that the total variance explained by direct and indirect effects is $4v = 0.8$, so that $h^2 = v_g = v = 0.2$.

The RDR covariance model is

$$\mathrm{Cov}(\mathbf{Y}) = v_g R + v_{e\sim g} R_{\mathrm{par}} + c_{e,g} R_{\mathrm{o,par}} + \sigma_\epsilon^2 \mathrm{I}, \tag{6}$$

where $\mathbf{Y}$ is the vector of phenotype observations, R is the relatedness matrix, R_{par} is the parental relatedness matrix, and $R_{\mathrm{o,par}}$ is the parent-offspring relatedness matrix. These matrices can be computed from SNP data by using the following formulae:

$$R = \frac{XX^T}{L};\ \ R_{\mathrm{par}} = \frac{X_{\mathrm{par}} X_{\mathrm{par}}^T}{2L};\ \ R_{\mathrm{o,par}} = \frac{XX_{\mathrm{par}}^T + X_{\mathrm{par}} X^T}{2L}, \tag{7}$$

where X is the matrix of normalised genotypes, X_{par} is the matrix of parental genotypes (sum of normalised maternal and paternal genotypes), and L is the number of SNPs in X. To compute these matrices, you can use the provided script:

```
python make_rdr_grms.py sample sample --ped sample_fams.ped
```

This takes the genotypes from sample.ped and the pedigree in sample_fams.ped to compute the matrices, which are output as binary lower-triangular matrices in sample_R.grm.bin, sample_R_par.grm.bin, sample_R_o_par.grm.bin.

To estimate the variance components using RDR, we use *GCTA* to perform restricted maximum likelihood inference of the variance components. To do this, use the command:

../gcta_1.93.2beta/gcta64 --reml --reml-no-lrt --mgrm RDR_GRMs.txt --pheno y.ped --out y --thread-num 4

This uses the provided *GCTA* executable for a Linux environment and takes advantage of multi-threading with the --thread-num 4 option. If you are running Windows/Mac, you can find an appropriate binary at https://cnsgenomics.com/software/gcta/Download. You should find that this gives the heritability estimate, $V(G1)/Vp$, close to 0.2, indicating unbiased estimation of heritability.

The GREML approach does not control for parental/genotypes and relatedness, fitting the following covariance model:

$$\mathrm{Cov}(\mathbf{Y}) = v_g R + \sigma^2_\epsilon \mathrm{I}. \tag{8}$$

This is expected to give a biased estimate of v_g since the offspring relatedness matrix is highly correlated with the parental and parent-offspring relatedness matrices. To see what GREML gives for this trait, run

../gcta_1.93.2beta/gcta64 --reml --reml-no-lrt --grm sample_R --pheno y.ped --out y --thread-num 4

5 Imputing parental genotypes

The RDR method uses the fully observed genotypes of all family members. However, in real data, parental genotypes are often missing. Common data types include sibling pairs, and parent-offspring pairs. To simulate a mix of different missing data patterns, the sample_reduced.bed file contains genotype data for this sample with one parent removed from 1,500 families, and both parents removed from 1,500 other families. By using the imputation methods outlined in[2], missing parental genotypes are imputed from sibling genotypes and identity-by-descent sharing data on the siblings for the families without any genotyped parents, and from single parents and sibling offspring of that parent for families with one parent genotyped.

To perform the imputation, use the impute_runner.py in the main SNIPar directory:

python ../impute_runner.py sample.segments.gz sample_reduced --agesex sample.agesex --king sample_reduced.king.kin0 --output_address sample --threads 4

This uses the recorded parent-offspring and sibling relations in sample_reduced.king.kin0, which is the same format as output by KING with the '--related --degree 1' options, and the age and

sex information in sample.agesex to construct a pedigree. Given the constructed pedigree, the script then uses the observed sibling and parental genotypes in sample.bed and IBD segments shared between siblings in sample.segments.gz to impute missing parental genotypes. The IBD segments are in the same format as output by KING with the '--ibdseg --degree 1' option. The script takes advantage of multi-threading using the --threads 4 option. It outputs the imputed parental genotypes in a HDF5 file sample.hdf5.

6 Estimating direct and indirect effects

The fGWAS.py script in the main SNIPar directory fits the following model for each SNP:

$$Y_{ij} = \delta_l g_{ijl} + \eta_{pl} g_{p(i)l} + \eta_{ml} g_{m(i)l} + u_i + \epsilon_{ij}, \tag{9}$$

where $u_i \sim \mathcal{N}(0, \sigma_u^2)$ is a random effect that models differences in phenotypic mean between families, and ϵ_{ij} is independent normal error with variance σ_ϵ^2. When parental genotypes are not observed, they are replaced with their imputed versions. When parental genotypes are imputed from siblings alone, the imputed paternal and maternal genotypes are identical. It is therefore necessary to have some observed parental genotypes to fit the above model, otherwise there is no information to separate paternal and maternal effects. If parental genotypes are imputed from siblings alone, a model that infers the average of the paternal and maternal effects can be used instead:

$$Y_{ij} = \delta_l g_{ijl} + \eta_l (\hat{g}_{p(i)l} + \hat{g}_{m(i)l}) + u_i + \epsilon_{ij}, \tag{10}$$

where $(\hat{g}_{p(i)l} + \hat{g}_{m(i)l})$ is the imputed sum of maternal and paternal genotypes. This model can be used by providing the '--parsum' option to the fGWAS.py script.

To use the imputed (and remaining observed) parental genotypes to estimate direct and indirect effects for the 1,000 SNPs, use the command:

```
python ../fGWAS.py sample sample.hdf5 y.ped y
```

This outputs estimates of direct, paternal, and maternal effects, along with their sampling variance-covariance matrices, to y.hdf5. The properties of the imputation we perform ensure that, even though we have not observed all of the variables, the estimates remain unbiased. To test this, run the following command:

```
python estimate_sim_effects.py
```

This regresses the estimates from y.hdf5 onto the true SNP effects in y.effects.txt, and checks for bias. You should see that the estimated bias term is within a couple of standard errors from zero. It also checks for the bias you would get if you used estimates from standard GWAS, which regresses phenotype onto genotype without control for parental genotypes. The effects estimated by standard GWAS, which we refer to as population effects, are approximately equal to the direct effect plus the average of the maternal and paternal effects. This bias term should be more than a couple of standard deviations from zero, indicating that population effects from standard GWAS give biased estimates of direct effects for this trait.

References

[1] Young, A. I., et al. Relatedness disequilibrium regression estimates heritability without environmental bias. *Nature Genetics*, **50**(9):1304–1310 2018. ISSN 1546-1718. doi:10.1038/s41588-018-0178-9.

[2] Young, A. I., et al. Mendelian imputation of parental genotypes for genome-wide estimation of direct and indirect genetic effects. *BioRxiv* 2020. doi:10.1101/xxxxxx.

[3] Yang, J., et al. Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics*, **42**(7):565–9 2010. ISSN 1546-1718. doi:10.1038/ng.608.