Heritability estimation using relationship matrices

Jing Hua Zhao
University of Cambridge

August 18, 2014

1 Introduction

Twin and family studies have been the standard approach for heritability estimation, where differences between monozygotic and dizygotic twin pairs are attributed to genetics and familial relationships are linked with a polygenic effect. Usually the estimate from twin studies is higher than that from family studies. It is difficult to tease out influence of the common environment for both types of data.

There has been a lot of interest recently in use of genomic relationship matrices (GRMs) regardless their famiilial background so unrelated individuals can also be used (Yang et al. (2010)). The GRM associated with a polygenic component in a random effects or mixed model mirrors the role of a relationship matrix based on family structures. A dedicated computer program called GCTA is available (Yang et al. (2011)). Work has been done to show the utility of GRM in linkage studies (Day-Williams et al. (2011)) and heritability estimation (Klimentidis et al. (2013)).

Here we use a very simple family to illustrate heritability estimation. As GRMs typically involve large quantity of genomic data, we will use the relationship matrix derived from the family structure as if it was a GRM. We then provide examples to read/write GRMs either in text or binary format as required by GCTA. A version showing estimated GRM in the computer program PLINK is also provided.

2 Data

The data is on a single family from the computer program Morgan.

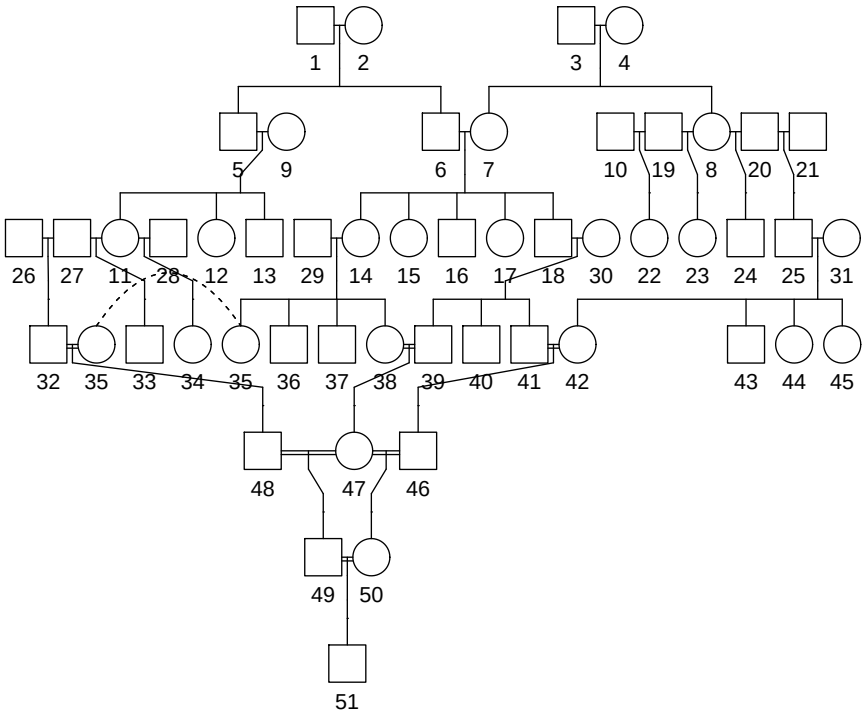
```
> library(gap)
> head(l51,10)

   id fid mid sex aff      qt
1   1   0   0   1   1 -0.9642
2   2   0   0   2   1  1.0865
3   3   0   0   1   1 -0.5363
4   4   0   0   2   1  0.4514
5   5   1   2   1   1  0.0538
```

```
6   6   1   2   1   1 -1.2667
7   7   3   4   2   1      NA
8   8   3   4   2   1  0.1743
9   9   0   0   2   1  0.2923
10 10   0   0   1   1      NA
```

```
> library(kinship2)
> ped <- with(l51,pedigree(id,fid,mid,sex))
> pdf("figures/l51.pdf")
> plot(ped)
> dev.off()
```

```
null device
          1
```

and the pedigree diagram is as follows,



3 Model

We can obtain a linear mixed model for the quantitative trait (qt) in `l51` above.

```
> library(gap)
> k2 <- kin.morgan(l51)$kin.matrix*2
> k2[1:10,1:10]

     1   2   3   4   5   6   7   8 9 10
1  1.0 0.0 0.0 0.0 0.5 0.5 0.0 0.0 0  0
2  0.0 1.0 0.0 0.0 0.5 0.5 0.0 0.0 0  0
3  0.0 0.0 1.0 0.0 0.0 0.0 0.5 0.5 0  0
4  0.0 0.0 0.0 1.0 0.0 0.0 0.5 0.5 0  0
5  0.5 0.5 0.0 0.0 1.0 0.5 0.0 0.0 0  0
6  0.5 0.5 0.0 0.0 0.5 1.0 0.0 0.0 0  0
7  0.0 0.0 0.5 0.5 0.0 0.0 1.0 0.5 0  0
8  0.0 0.0 0.5 0.5 0.0 0.0 0.5 1.0 0  0
9  0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 1  0
10 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0  1

> library(regress)
> r <- regress(qt ~ 1, ~k2, data=l51)
> r$sigma

       k2        In
0.2817099 0.4444962

> r$sigma.cov

            k2          In
k2  0.07163300 -0.03991478
In -0.03991478  0.04042731
```

The function `kin.morgan` is readily used for the well-ordered pedigree. The relationship matrix is supplied to `regress` function for parameter estimation. We can also generate a binary trait (bt) and run through the regression model similarly,

```
> N <- dim(l51)[1]
> w <- with(l51,quantile(qt,probs=0.75,na.rm=TRUE))
> l51 <- within(l51, bt <- ifelse(qt<=w,0,1))
> with(l51,table(bt))

bt
 0  1
32 11

> d <- regress(bt ~ 1, ~k2, data=l51)
> d$sigma
```

```
       k2        In
0.0307703 0.1678370
```

```
> d$sigma.cov
```

```
             k2           In
k2  0.003615481 -0.002525622
In -0.002525622  0.003492826
```

4 Heritabilities

Once the mixed models are obtained, we can get the heritability estimates. Note that although we set a population prevalence (K) to be 0.25, there were 11 cases and 40 controls from the simulation, leading to a case/control proportion (P) of 11/51=0.2156863.

The heritability estimate is a ratio of polygenic and phenotypic variance and available from function `h2G` which also gives the associate variance estimate. Internally, this involves function `VR` for calculating variance of a ratio. We illustrate with the example given above,

```
> library(gap)
> # qt
> sigma <- c(0.2817099, 0.4444962)
> sigma.cov <- matrix(
+ c(0.07163300, -0.03991478,
+ -0.03991478, 0.04042731), 2, 2)
> h2G(sigma,sigma.cov)
```

```
Vp = 0.7262061 SE = 0.1795292
h2G = 0.38792 SE = 0.3136308
```

```
> # bt
> sigma <- c(0.0307703, 0.1678370)
> sigma.cov <- matrix(
+ c(0.003615481, -0.002525622,
+  -0.002525622,  0.003492826), 2, 2)
> h2G(sigma,sigma.cov)
```

```
Vp = 0.1986073 SE = 0.04535486
h2G = 0.1549304 SE = 0.2904298
```

As only a single family is involved in the analysis, it is not surprising to see large standard errors. For a case-control study, the heritability estimation is based on a liability threshold model and the connection is furnished through the function `h2l` taking into account the population prevalence and the proportion of cases in the sample (Lee et al. (2011)).

```
> h2l(K=0.25, P=11/51, h2=0.1549304, se=0.2904298)
```

```
K =  0.25 P =  0.2156863
h2 = 0.1549304 SE = 0.2904298 h2l = 0.3188476 SE = 0.597706
```

which yields a larger point estimate nevertheless with larger standard error. The relationship between population prevalence and heritability will be seen more clearly later.

It makes sense to illustrate with real data. Before doing that, we would like to indicate that when a model includes gene-environment interaction, (restricted) maximum likelihood estimateors would involve three variance components, heritabilities associated with both polygenic and interaction are obtained via function `h2GE`.

Below is an example from a real session of GCTA analysis but we only keep the variance components and their (lower-triangular) variance-covariance matrix as input to the relevant functions described above.

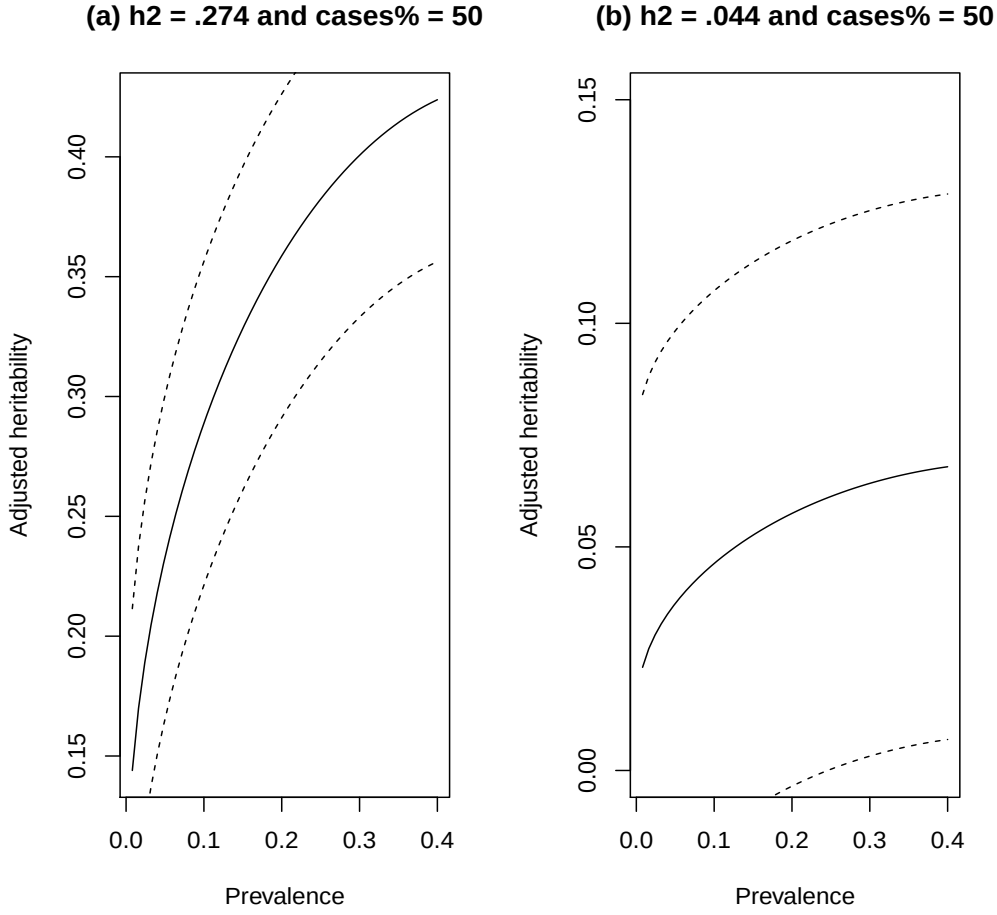
```
> library(gap)
> V <- c(0.017974, 0.002451, 0.198894)
> VCOV <- matrix(0,3,3)
> diag(VCOV) <- c(0.003988, 0.005247, 0.005764)^2
> VCOV[2,1] <- -7.93348e-06
> VCOV[3,1] <- -5.54006e-06
> VCOV[3,2] <- -1.95297e-05
> z <- h2GE(V,VCOV)

Vp = 0.219319 SE = 0.003263797
h2G = 0.08195368 SE = 0.01799574 h2GE = 0.0111755 SE = 0.02392398
```

Here is an example for case-control data,

```
> library(gap)
> h2 <- 0.274553
> se <- 0.067531
> P <- 0.496404
> z <- h2l(P=P,h2=h2,se=se)

K =  0.05 P =  0.496404
h2 = 0.274553 SE = 0.067531 h2l = 0.2329584 SE = 0.05730009

> R <- 50
> kk <- h2all <- seall <- rep(0,R)
> for(i in 1:R)
+ {
+   kk[i] <- 0.4*i/R
+   z <- h2l(kk[i],P=P,h2=h2,se=se,verbose=FALSE)
+   h2all[i] <- z$h2l
+   seall[i] <- z$se
+ }
> h2 <- 0.044
> se <- 0.061
> z <- h2l(P=P,h2=h2,se=se)

K =  0.05 P =  0.496404
h2 = 0.044 SE = 0.061 h2l = 0.03733402 SE = 0.05175853
```

```
> h2alls <- sealls <- rep(0,R)
> for(i in 1:R)
+ {
+   z <- h2l(kk[i],P=P,h2=h2,se=se,verbose=FALSE)
+   h2alls[i] <- z$h2l
+   sealls[i] <- z$se
+ }
> pdf("figures/h2l.pdf")
> par(mfrow=c(1,2))
> plot(kk,h2all,type="l",ylab="Adjusted heritability",xlab="Prevalence")
> lines(kk,h2all-seall,lty="dashed")
> lines(kk,h2all+seall,lty="dashed")
> title("(a) h2 = .274 and cases% = 50")
> plot(kk,h2alls,type="l",ylab="Adjusted heritability",xlab="Prevalence",ylim=c(0,0.15))
> lines(kk,h2alls-sealls,lty="dashed")
> lines(kk,h2alls+sealls,lty="dashed")
> title("(b) h2 = .044 and cases% = 50")
> dev.off()

null device
          1
```

Later we also examine the impact of disease prevalence, using a grid over 50 prevalences, on heritbaility estimation as shown in the following figure,



This suggests a nonlinear relationship between the observed and adjusted estimtes and high prevalence puts more weight on the estimator. However, the effect of prevalence and ascertainment adjustment is less pronounced when the heritability is low.

5 Exchange of GRMs between software

We can read or write the GRMs used by GCTA for the example above with the following code,

```
> p <- matrix(0,N,4)
> for(i in 1:N) p[i,] <- with(l51[i,],c(i,i,qt,bt))
> write(t(p),file="51.txt",4,sep="\t")
> NN <- rep(51, N * (N + 1)/2)
> WriteGRM(51,p[,1:2],NN,k2)
> one <- ReadGRM(51)
> grm <- one$grm
```

```
> WriteGRMBin(51,grm,NN,p[,1:2])
> two <- ReadGRMBin(51,TRUE)
> sum(one$GRM-two$GRM)
```

As well as illustrating how to manipulate GRMs in two formats, we also generate a phenotypic file called `51.txt`. Note the function `kin.morgan` result has an elemenet called `kin` which is similar to the vector `grm` above.

GRM from PLINK, i.e., the .genome file, can be read via a function called `ReadGRMPLINK`. Another function is called `WriteGRMSAS` can be used to output an `ldata` as required by type=LIN(1) in SAS PROC MIXED and PROC GLIMMIX. As for phenotypic data, we again turn to our pedigree `l51` and issue commands,

```
> library(foreign)
> write.dta(l51, "l51.dta")
```

to save the data as an external file in Stata format so that software system such as SAS can read it directly. Together with relationship matrix we can take a whole range of facilities available from there. Of course with this particular example, one could use PROC INBREED to generate a relationship matrix.

Morgan actually provides the relevant result for this pedigree as well. It is possible to work on kinship matrix generated from SOLAR, Earlier we discussed how to do this kind of analysis using SAS in Zhao and Luan (2012).

References

A.G. Day-Williams, J. Blangero, T.D. Dyer, K. Lange, and E.M. Sobel. Linkage analysis without defined pedigrees. *Genet Epidemiol.*, 35(5):360–370, July 2011. URL `PM:21465549`.

Y.C. Klimentidis, A.I. Vazquez, Campos G. de Los, D.B. Allison, M.T. Dransfield, and V.J. Thannickal. Heritability of pulmonary function estimated from pedigree and whole-genome markers. *Front Genet*, 4(174):1–5, 2013. URL `PM:24058366`.

S.H. Lee, N.R. Wray, M.E. Goddard, and P.M. Visscher. Estimating missing heritability for disease from genome-wide association studies. *Am.J.Hum.Genet*, 88(3):294–305, March 2011. URL `PM:21376301`.

J. Yang, B. Benyamin, B.P. McEvoy, S. Gordon, A.K. Henders, D.R. Nyholt, P.A. Madden, A.C. Heath, N.G. Martin, G.W. Montgomery, M.E. Goddard, and P.M. Visscher. Common snps explain a large proportion of the heritability for human height. *Nat Genet*, 42(7):565–569, July 2010. URL `PM:20562875`.

J. Yang, S.H. Lee, M.E. Goddard, and P.M. Visscher. GCTA: A tool for genome-wide complex trait analysis. *Am.J.Hum.Genet*, 88(1):76–82, January 2011. URL `PM:21167468`.

J.H. Zhao and J.A. Luan. Mixed modeling with whole genome data. *J Prob Stat*, doi 10.1155/2012.485174:1–16, 2012. URL `http://www.hindawi.com/journals/jps/2012/485174/`.