Estimating the contribution of DNA to correlations between human traits

Ronald de Vlaming

October 12, 2021

Abstract. Almost without exception, behaviour, personality, and even socio-economic outcomes have been shown to have a heritable component (i.e., genes help explain a certain proportion of variation in such traits). As large-scale collection of genetic data becomes ever more affordable, genetics plays an increasingly important role in the social sciences. The widespread availability of raw genetic data creates exciting venues for development and application of new statistical methods and econometric techniques. Application of such methods help us better understand how differences and similarities between us are shaped. Here, I discuss the concepts of heritability and genetic correlation. These parameters help us gauge to what extent the genetic information that we can actually measure contributes to variation within traits and covariance between traits. More specifically, as an example at the intersection of genetics and econometrics, I will explain a method here, called MGREML (recently published in *Communications Biology*), that I developed together with colleagues here in the Netherlands and abroad. MGREML enables researchers to simultaneously estimate heritabilities and genetic correlations for many outcomes observed in a large sample. As you will see, such methods rely heavily on matrix algebra, statistics, and numerical methods—topics you all learn about in your training as econometrician!

Introduction

Imagine two individuals who have been exposed to highly similar environments in childhood. Intuitively, we expect these individuals to be more similar in terms of outcomes, such as educational attainment, than two individuals who have grown up in very dissimilar environments. Put differently, environmental similarity (to some degree) maps to similarity in a given outcome.

Such reasoning can easily be extended to the domain of genetic data, where one could argue that genetic similarity also maps (again to a certain degree) to

similarity in the outcome of interest. For example, we expect identical twins to be more similar for all kinds of outcomes (e.g., hair colour, longevity, educational attainment, etc.) than fraternal twins, siblings, nieces and nephews, and so on. This expectation gives rise to the concept of heritability (h^2): the proportion of variance in an outcome that can be 'explained' by genes. Decades of studies using data on twin pairs have revealed that basically any conceivable trait, characteristic, or other outcome at the individual level has a heritable component (Polderman *et al.*, 2015).

The concept of heritability can be further generalised to a so-called genetic correlation (r_G), which basically quantifies the degree to which genetic similarity between two individuals, i and j, maps to similarity between outcome X for individual i and outcome Y for individual j. More precisely, decomposing a trait as the sum of a genetic and environmental component, genetic correlation is defined as the correlation between G_X and G_Y, where these are the genetic components of outcome X and Y respectively. Genetic correlations, it turns out, are as omnipresent as correlations in general (e.g., see Bulik-Sullivan *et al.*, 2015).

In the olden days, geneticists would use expected genetic similarity (based on pedigree) to identify h^2 and r_G (Falconer and Mackay, 1996). For instance, identical twins have an expected genetic similarity of one, while fraternal twins have an expected genetic similarity of only $1/2$. These differences in expected genetic similarity can be used to glean h^2 and r_G (under a plethora of assumptions, of course).

In recent years, however, directly measuring the genetic data for many individuals has become quite affordable (e.g., see Loos, 2020), as illustrated by Figure 1. This direct availability of genetic data has paved the way for many exciting applications, including estimation of h^2 and r_G directly from such genetic data, even if we only have individuals who are (approximately) unrelated (Yang *et al.*, 2011; Lee *et al.*, 2012).

Here, I discuss (1) how genetic data is typically measured, (2) what kind of model is assumed to estimate h^2 and r_G. Next, I discuss (3) how we can use a form of maximum-likelihood estimation (MLE) in conjunction with (4) an appropriate multivariate model, matrix algebra, and numerical methods to make estimation of these parameters scalable for large sample sizes (N) and a large number of outcomes (T). This overall approach is called MGREML (De Vlaming *et al.*, 2021). Finally, I will (5) highlight results from an application of MGREML to brain data and related outcomes, such as educational attainment, using data from the UK Biobank.

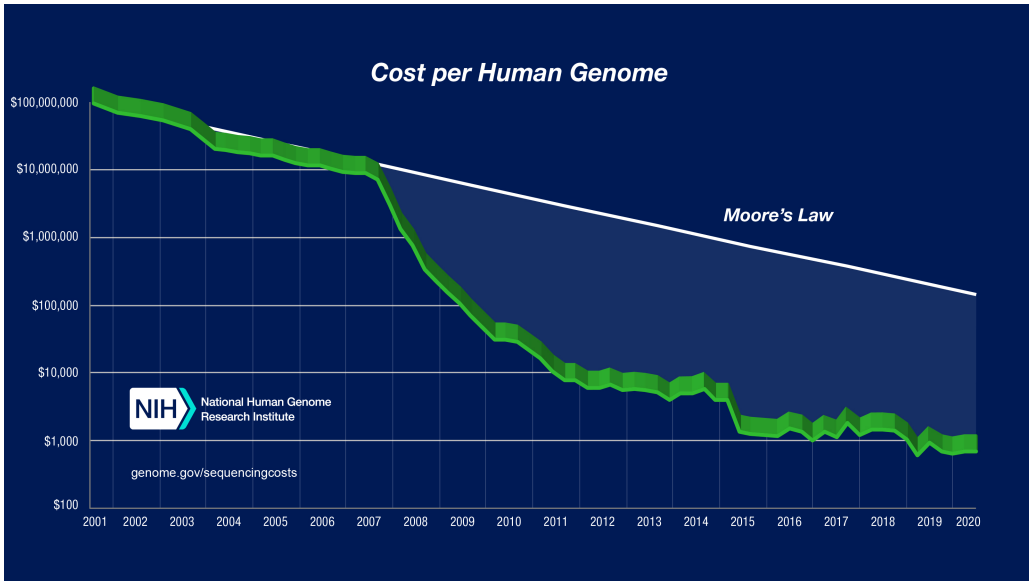

Figure 1: The costs of genotyping over time. Source: *National Human Genome Research Institute*, `https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost`.

1 Measuring genetic data

The human genome comprises roughly three billion so-called basepairs, many of which contain identical information for all mankind: these 'constant' parts of the DNA are what makes us human. The easiest way to conceptualise these basepairs is using four letters from the alphabet: 'A', 'C', 'G', and 'T'. Now, using this 'genetic alphabet', the easiest way to think about the human genome is by considering it as a humongous string which is a whopping three billion characters long, comprising substrings such as "[...]ACGTCA[...]" as can been seen in Figure 2.

Notice, based on that figure, that DNA is a two-stranded molecule. One strand, however, can safely be 'ignored' when reading the genetic data, as the two strands are complementary to each other. That is, if at a given position in the DNA there is an 'A' on the one strand then there must be a 'T' on the other strand (and vice versa), and if there is a 'C' on the one strand there must be a 'G' on the other (and vice versa). Thus, the string of letters "[...]ACGTCA[...]" on the one strand corresponds to "[...]TGCAGT[...]" on the other, as can also be seen in Figure 2. In other words: if you know the string that describes a given strand you also know the string that describes the complementary strand.

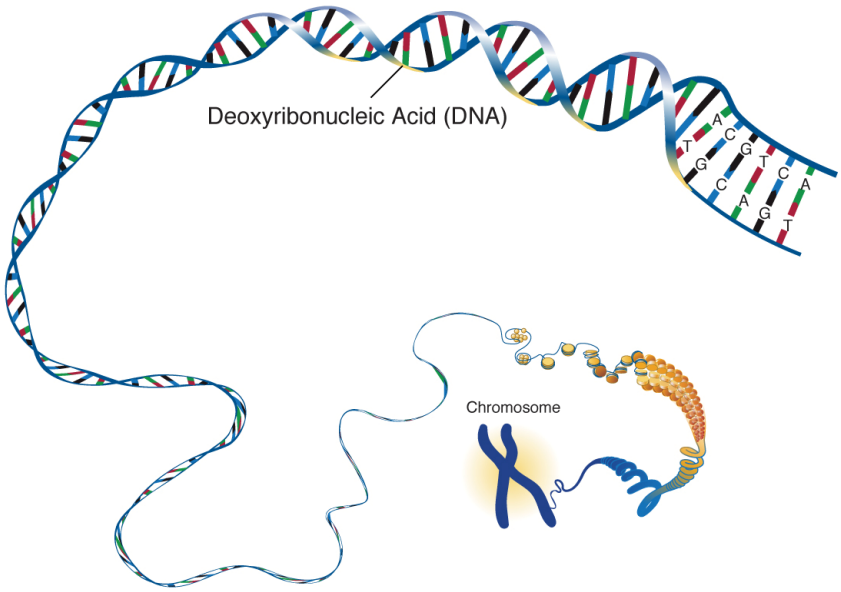

Figure 2: A short portion of DNA from a given chromosome. Source: *National Human Genome Research Institute*, `http://www.genome.gov/glossary/index.cfm?id=48`.

These 'genetic strings' are located on chromosomes, 22 of which are regular chromosomes and one of which is the so-called sex chromosome. For each chromosome, you have two versions: one inherited from your mother and one from your father. From hereon out, for simplicity, we ignore the sex chromosomes from further discussion. The fact that you have inherited two versions of each chromosome effectively doubles the number of basepairs in your DNA, to about six billion basepairs, as you have each unique basepair twice, *viz.*, once on each of the two versions of the given chromosome that you have inherited.

Importantly, once every so many basepairs, a bit of variation may occur across the population. A form of genetic variation that is often studied is a so-called single-nucleotide polymorphism (SNP; pronounced as *snip*). A SNP is a single basepair in the genome where different 'letters' of the genetic alphabet are observed across the population. For example, perhaps 75% of the population has a 'G' at the first basepair on the first chromosome and 25% has an 'A' there. We then say this position in the genome is a SNP with 'G' and 'A' as its alleles, where 'A' is the minor allele (i.e., occurring least frequently). Although SNPs with three or even four alleles can exist, such SNPs are far less common than SNPs with just two alleles. Thus, much research focusses exclusively on these so-called biallelic SNPs.

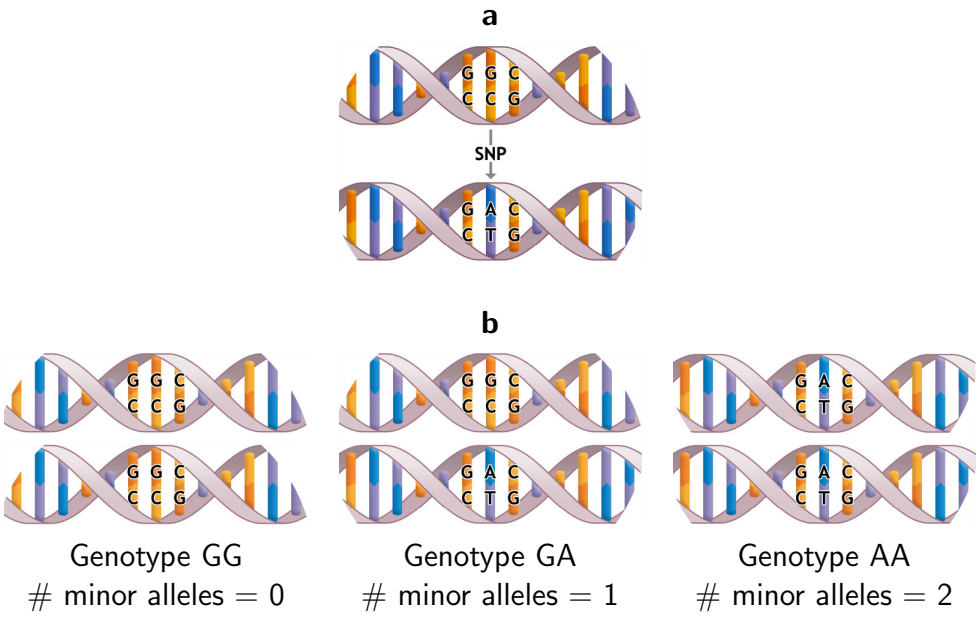

Figure 3: Panel a: example of a SNP with two alleles, *viz.*, 'G' (major allele) and 'A' (minor allele). Panel b: example of the three different genotypes that a given individual in the population can have for this SNP, *viz.*, 'GG', 'GA' (or equivalently 'AG'), and 'AA'. Source: *Integrated Biobank of Luxembourg*, `https://www.ibbl.lu/wp-content/uploads/2012/07/SNPs.jpg`. Edited by Ronald de Vlaming.

Panel a of Figure 3 shows the SNP in the example: at this location in the DNA, two variants are seen in the upper strand across the population, *viz.*, 'G' and 'A'. The fact that you have two copies of each basepair (one on each of the two chromosomes that form a pair) implies that you three unique so-called genotypes for the SNP in the example: 'GG' (i.e., the 'G' allele on both chromosomes for the given base pair), 'GA' or 'AG' (i.e., the 'G' allele on one chromosome and the 'A' allele on the other, for the given basepair), and finally 'AA'. These three possible genotypes are shown in Panel b of Figure 3.

Given (1) we consider only biallelic SNPs and (2) by setting one of its two alleles as the so-called coded allele, we can thus reduce SNP data to simple counts of the coded allele, with counts being equal to zero, one, or two. In the example, by setting minor allele 'A' as the coded allele, the count is zero for the 'GG' genotype, one for the 'GA' and 'AG' genotypes, and two for the 'AA' genotype. Now, for each individual for whom we have collected data on biallelic SNPs, for each of those SNPs, we effectively have this count.

In short, we have a clearly defined measure of genetic data on the molecular level with a clear interpretation and a numerical value (zero, one, or two), making this data highly suitable for various statistical analyses. And, perhaps most importantly, this measure can be inferred with great accuracy using affordable genotyping platforms. What a marvel this is! We can measure the genotype of many individuals with great precision, and the resulting data lend themselves extremely well for statistical analyses.

2 Statistical model

Once the initial excitement has subsided a bit, you may find there is also a more sobering aspect to working with genetic data: its sheer size! Just considering SNPs with two alleles, there are already many millions of genetic variants, while the largest samples are of the order of hundreds of thousands of individuals. Thus, from a regression perspective, we have M potential regressors (SNPs) and N individuals for whom we observe these SNPs, where $M \gg N$. That is, we are solidly in the domain of multicollinearity and overfitting (e.g., see Friedman *et al.*, 2009). In other words, if we would use ordinary least squares to estimate the effect of all SNPs jointly on a given outcome Y, it would fail spectacularly.

However, there are three important things to keep in mind here. First, genetic variants (such as SNPs) that are very close to each other within the genome tend to be strongly correlated. Thus, we only need to consider a broad subset of

SNPs (called common variants) to 'tag' most of the variation in a trait that can be explained by genes. Second, there is much evidence that for many outcomes, ranging from diseases, to intelligence, personality traits, behaviour, and even socio-economic outcomes, there are thousands of genetic variants that affect these traits, all with small effects. This pattern is referred to as high polygenicity. Third, there are many techniques to deal with $M \gg N$, two elegant approaches being the imposition of (1) a reasonable prior on the distribution of the effects of the regressors and (2) a penalty on the effects of the regressors in the loss function of interest (e.g., in the definition of the sum of squared regression residuals). In fact, maximum *a posteriori* estimates obtained under certain priors can be shown to be mathematically equivalent to simply imposing an appropriate penalty in the loss function (e.g., see De Vlaming and Groenen, 2015).

An important prior is the so-called infinitesimal model, where each SNP is assumed to have a very small effect, with mean zero, where some effects are slightly negative and others slightly positive. This prior, in fact, aligns rather neatly with the high polygenicity of outcomes, a conceptual model for which there is much evidence (Visscher *et al.*, 2017). More specifically, we typically assume the following model holds for a normally distributed continuous outcome Y:

$$\mathbf{y} = \mathbf{G}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{1}$$

$$\boldsymbol{\beta} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{I}_M \sigma^2_\beta\right) \tag{2}$$

$$\boldsymbol{\varepsilon} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{I}_N \sigma^2_E\right). \tag{3}$$

In this model, $\mathbf{y}$ denotes the $N \times 1$ outcome vector and $\mathbf{I}_N$ denotes the $N \times N$ identity matrix. The main parameters here are σ^2_β and σ^2_E. The distribution of $\mathbf{y}$ can now be written quite compactly as follows:

$$\mathbf{y} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{G}\mathbf{G}^\mathsf{T}\sigma^2_\beta + \mathbf{I}_N \sigma^2_E\right). \tag{4}$$

Now, defining $\mathbf{A} = M^{-1}\mathbf{G}\mathbf{G}^\mathsf{T}$ and $\sigma^2_G = M\sigma^2_\beta$, we can rewrite the model for $\mathbf{y}$ as follows:

$$\mathbf{y} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{A}\sigma^2_G + \mathbf{I}_N \sigma^2_E\right). \tag{5}$$

Matrix $\mathbf{A}$ is often referred to as the genomic-relatedness matrix (GRM). Importantly, as SNPs are standardised, the diagonal element of $\mathbf{A}$ are approximately equal to one. Notice that the off-diagonal elements of $\mathbf{A}$ basically quantify how similar two given individuals are in terms of their SNP data.

The fact that the diagonal elements are (approximately) equal to one implies that, under this model, the variance in Y for each individual $i = 1, \ldots, N$ can

be decomposed as $\sigma^2_G + \sigma^2_E$, where σ^2_G is the so-called additive genetic variance explained by SNPs and σ^2_E the residual or environmental variance. Now, SNP-based heritability is defined as

$$h^2_{\text{SNP}} = \frac{\sigma^2_G}{\sigma^2_G + \sigma^2_E}. \tag{6}$$

Given a GRM $\mathbf{A}$, derived from SNP data, and an outcome vector $\mathbf{y}$, one could use MLE to estimate σ^2_G and σ^2_E. Here, MLE aims to find values for the parameters such that the following function is maximised:

$$\ell(\sigma^2_G, \sigma^2_E) = -\frac{1}{2}\left(N \log(2\pi) + \log\left|\mathbf{A}\sigma^2_G + \mathbf{I}_N\sigma^2_E\right| + \mathbf{y}^\mathsf{T}\left(\mathbf{A}\sigma^2_G + \mathbf{I}_N\sigma^2_E\right)^{-1}\mathbf{y}\right) \tag{7}$$

3 Restricted maximum likelihood estimation

For a given outcome, there are often certain factors that we want to control for, as our results could otherwise suffer from an omitted-variable bias. If we have an $N \times K$ matrix of covariates, $\mathbf{X}$, that we want to take into account, the most straightforward way to incorporate this in our model is by changing Equation 1, by assigning so-called fixed effects, $\boldsymbol{\gamma}$, to $\mathbf{X}$ as follows:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\gamma} + \mathbf{G}\boldsymbol{\beta} + \boldsymbol{\varepsilon}. \tag{8}$$

In this model, we thus a have a mix of linear fixed effects, $\boldsymbol{\gamma}$, and linear random effects, $\boldsymbol{\beta}$. Therefore, this type of model is often referred to as a linear mixed model (LMM). Please keep in mind that the 'fixed' and 'random' effects as mentioned here and elsewhere in the literature on LMMs are conceptually different from random and fixed effects as often seen in the literature on panel data!

By using properties of the multivariate normal distribution, we can show that the outcome of interest is distributed as follows under this model:

$$\mathbf{y} \sim \mathcal{N}\left(\mathbf{X}\boldsymbol{\gamma}, \mathbf{A}\sigma^2_G + \mathbf{I}_N\sigma^2_E\right). \tag{9}$$

Now, one could formulate the log-likelihood function in terms of σ^2_G, σ^2_E, and $\boldsymbol{\gamma}$ and apply MLE. However, there is a downside there: when estimating these so-called variance components σ^2_G and σ^2_E (i.e., parameters that shape the variance matrix), these tend to get underestimated by MLE, as MLE fails to take the degrees of freedom lost by controlling for $\mathbf{X}$ into account.

However, there exists an alternative approach that takes the lost degrees of freedom into account. This approach is called restricted maximum likelihood

(REML) estimation. Effectively, REML applies maximum likelihood estimation to $\mathbf{P}^\mathsf{T}\mathbf{y}$ instead of $\mathbf{y}$, where $\mathbf{P}$ is an $N \times (N-k)$ matrix, such that $\mathbf{P}^\mathsf{T}\mathbf{X} = \mathbf{0}$, where $k = \text{rank}(\mathbf{X})$ and $\text{rank}(\mathbf{P}) = N - k$. Consequently, $\mathbf{P}$ accounts for the lost degrees of freedom by changing the 'effective' sample size and by 'partialing out' the fixed effects, as the columns of $\mathbf{P}$ all lie in the left null space of $\mathbf{X}$.

Importantly, REML does not make this transformation explicit. Rather, the transformation is subsumed in the formulation of the corresponding log-likelihood function which is defined as follows (up to a constant):

$$\ell = -\frac{1}{2}\left(\log|\mathbf{V}| + \log\left|\mathbf{X}^\mathsf{T}\mathbf{V}^{-1}\mathbf{X}\right| + \mathbf{y}^\mathsf{T}\mathbf{M}\mathbf{y}\right), \text{ where} \tag{10}$$

$$\mathbf{M} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}\left(\mathbf{X}^\mathsf{T}\mathbf{V}^{-1}\mathbf{X}\right)\mathbf{X}^\mathsf{T}\mathbf{V}^{-1} \text{ and} \tag{11}$$

$$\mathbf{V} = \mathbf{A}\sigma^2_G + \mathbf{I}_N\sigma^2_E. \tag{12}$$

Now, numerical methods (such as Newton's method) can be used to find values for σ^2_G and σ^2_E that maximise the REML function. This approach of estimating the variance accounted for by all available SNPs is called genomic-relatedness restricted maximum likelihood (GREML) estimation.

4 Multivariate GREML

Let $\mathbf{y}_t$ denote the $N \times 1$ outcome vector for trait $t = 1, \ldots, T$ and let $\mathbf{X}_t$ denote the corresponding $N \times K_t$ matrix of covariates with fixed effects in $K_t \times 1$ vector γ_t. Now, the univariate model can be generalised to a multivariate model on T outcomes observed in the same set of N individuals fairly straightforwardly as follows:

$$\text{vec}(\mathbf{Y}) = \begin{pmatrix} \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_T \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mathbf{X}_1 & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \mathbf{X}_T \end{pmatrix}\begin{pmatrix} \gamma_1 \\ \vdots \\ \gamma_T \end{pmatrix}, \mathbf{V}_G \otimes \mathbf{A} + \mathbf{V}_E \otimes \mathbf{I}_N\right), \tag{13}$$

where $\mathbf{Y} = (\mathbf{y}_1 \quad \ldots \quad \mathbf{y}_T)$ is the $N \times T$ matrix of outcomes and

$$\mathbf{V}_G = \begin{pmatrix} \sigma_{G_{11}} & \ldots & \sigma_{G_{1T}} \\ \vdots & \ddots & \vdots \\ \sigma_{G_{1T}} & \ldots & \sigma_{G_{TT}} \end{pmatrix}, \text{ and } \mathbf{V}_E = \begin{pmatrix} \sigma_{E_{11}} & \ldots & \sigma_{E_{1T}} \\ \vdots & \ddots & \vdots \\ \sigma_{E_{1T}} & \ldots & \sigma_{E_{TT}} \end{pmatrix}. \tag{14}$$

Here, $\text{vec}(\cdot)$ denotes the vectorisation operator and '$\otimes$' the Kronecker product. In this model, $\sigma_{G_{ts}}$ is the genetic covariance between traits t and s, and $\sigma_{E_{ts}}$ the environmental covariance for the corresponding two traits.

We can now define SNP-based heritability of trait t and the genetic correlation between traits t and s as follows under this model:

$$h^2_{\text{SNP}}(t) = \frac{\sigma_{G_{tt}}}{\sigma_{G_{tt}} + \sigma_{E_{tt}}} \text{ and} \tag{15}$$

$$r_G(t,s) = \frac{\sigma_{G_{ts}}}{\sqrt{\sigma_{G_{ts}}\sigma_{G_{ts}}}}. \tag{16}$$

Notice that the grand variance matrix, $\mathbf{V}_G \otimes \mathbf{A} + \mathbf{V}_E \otimes \mathbf{I}_N$, is a full $NT \times NT$ matrix. Thus, a naïve application of REML (which involves the log-determinant and inverse of that matrix) is computationally prohibitive, as it would require $O(N^3T^3)$ time per iteration, just to calculate the log-likelihood.

Importantly, however, the GRM is a symmetric positive (semi)-definite matrix, and as such as has eigenvalue decomposition given by $\mathbf{A} = \mathbf{Q}\mathbf{D}\mathbf{Q}^\mathsf{T}$, where $\mathbf{D}$ is a diagonal matrix with non-negative diagonal entries and $\mathbf{Q}$ is an orthonormal matrix (i.e., such that $\mathbf{Q}\mathbf{Q}^\mathsf{T} = \mathbf{Q}^\mathsf{T}\mathbf{Q} = \mathbf{I}_N$). As a consequence, when we consider $\text{vec}\left(\mathbf{Q}^\mathsf{T}\mathbf{Y}\right)$ as grand outcome vector, instead of $\text{vec}\left(\mathbf{Y}\right)$, the following model should then hold:

$$\text{vec}\left(\mathbf{Q}^\mathsf{T}\mathbf{Y}\right) \sim \mathcal{N}\left(\begin{pmatrix} \mathbf{Q}^\mathsf{T}\mathbf{X}_1 & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \mathbf{Q}^\mathsf{T}\mathbf{X}_T \end{pmatrix} \begin{pmatrix} \gamma_1 \\ \vdots \\ \gamma_T \end{pmatrix}, \mathbf{V}_G \otimes \mathbf{D} + \mathbf{V}_E \otimes \mathbf{I}_N\right). \tag{17}$$

Now, observe that the grand variance matrix, $\mathbf{V}_G \otimes \mathbf{D} + \mathbf{V}_E \otimes \mathbf{I}_N$, is highly sparse. Yet, this sparsity is spread over many rows and columns. However, by using a so-called commutation matrix (which effectively re-orders data by individual rather than by trait), denoted by $\mathbf{C}$, which is such that $\mathbf{C}\text{vec}\left(\mathbf{Q}^\mathsf{T}\mathbf{Y}\right) = \text{vec}\left(\mathbf{Y}^\mathsf{T}\mathbf{Q}\right)$ and $\mathbf{C}(\mathbf{B} \otimes \mathbf{E})\mathbf{C}^\mathsf{T} = \mathbf{E} \otimes \mathbf{B}$ for $\mathbf{B}$ and $\mathbf{E}$ of appropriate sizes, we obtain the following model:

$$\text{vec}\left(\mathbf{Y}^\mathsf{T}\mathbf{Q}\right) \sim \mathcal{N}\left(\mathbf{C}\begin{pmatrix} \mathbf{Q}^\mathsf{T}\mathbf{X}_1 & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \mathbf{Q}^\mathsf{T}\mathbf{X}_T \end{pmatrix} \begin{pmatrix} \gamma_1 \\ \vdots \\ \gamma_T \end{pmatrix}, \mathbf{D} \otimes \mathbf{V}_G + \mathbf{I}_N \otimes \mathbf{V}_E\right). \tag{18}$$

Although the matrix of fixed-effect covariates now looks a bit clunky, crucially, the variance matrix has a greatly reduced complexity: it is now completely block diagonal. That is, there are N diagonal blocks, where each block is a full $T \times T$

matrix, with block i given by $d_i\mathbf{V}_G + \mathbf{V}_E$, where d_i is the i-th diagonal element from $\mathbf{D}$ for $i = 1, \ldots, N$.

The reason why a block-diagonal variance matrix is so useful is that computing the determinant and inverse of this grand $NT \times NT$ variance matrix is now only as difficult as computing the determinant and inverse of each of the N non-overlapping blocks. Thus, effectively, these calculations now have a time complexity that goes up linearly with sample size, N, instead of going up at an N^3 rate.

From this point on, further matrix-algebraic tricks can be used to compute the complete log-likelihood in $O(NT^2)$ time. Similarly, in the publication in *Communications Biology*, I show that even the gradient can be calculated in $O(NT^2)$ time (De Vlaming *et al.*, 2021).

Importantly, to find the parameter estimates that maximise the REML function, we need a suitable numerical method, as there exists no analytical solution. Unfortunately, Newton's method works poorly here for two reasons: (1) the Hessian becomes computationally prohibitively expensive when T is large, as the number of parameters scales quadratically with the number of the traits T and, thus, the Hessian has $O(T^4)$ unique elements that all need to be calculated in each iteration, and (2) there may be many points in the parameter space where the Hessian is (nearly) rank deficient, yielding numerically unstable updates.

We can resolve both issues by applying a Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm (e.g., see Fletcher, 1987), which is a so-called quasi-Newton approach. This widely used algorithm iteratively constructs a numerically stable approximation of the inverse of the Hessian using little more than the gradient from several subsequent iterations. To ensure each step guarantees an increase in the REML function, we combine the BFGS algorithm with a simple line-search method that is applied in each iteration.

The overall approach of estimating the genetic and environmental variance matrix, $\mathbf{V}_G$ and $\mathbf{V}_E$, for a set of T traits observed in a set of N individuals, using these numerical techniques, we refer to as MGREML. This technique is available as command-line tool here on GitHub.

Naturally, there are many more details that I would like to describe about this method. But these are best left to the manuscript itself, available as open-access article in *Communications Biology*. Detailed derivations of the method are available as "Supplementary Information" (De Vlaming *et al.*, 2021).

5 An application: brain morphology

In the study highlighted here, I applied MGREML, implemented as command-line tool using Python 3.x, to data from the UK Biobank (UKB). In the UKB, genotypes were collected for hundreds of thousands of UK residents. Amongst many other traits, the UKB data comprises brain-scan data. These data can be used to calculate so-called grey-matter volume in various regions of the brain.

After stringent quality control (i.e., dropping individuals, trais, and SNPs that were problematic or incomplete), we were left with a set of $T = 86$ traits, of which 76 were directly related to brain morphology, observed in $N = 20,190$ individuals. MGREML estimation required 324 BFGS iterations to converge. In total, estimation of the full set of 86 SNP-based heritabilities and $86 \times 85 \times 1/2 = 3,655$ unique genetic correlations took about one hour.

To illustrate some of the results, Figure 4 shows the average SNP-based heritability for various sets of brain regions (Panel a), as well as the full genetic correlation matrix (shown as a heatmap in Panel b) for the brain-related traits. Results align strongly with standard anatomical subdivisions in the brain. A detailed description of these empirical results can also be found in the manuscript in *Communications Biology*.

In addition, simulation results show that MGREML currently is the most efficient tool to simultaneously estimate many SNP-based heritabilities and genetic correlations. Moreover, these simulations corroborate the consistency of its estimates.

Wrapping up, the take-home message of all this (at least to me, at a personal level) is the following: in the age of big data, there are many fields where you, as econometrician, can apply your skills. These fields may even include really unexpected areas, like quantitative genetics. Such serendipitous applications of skills and knowledge are, of course, what makes interdisciplinary research so exciting!

References

Bulik-Sullivan,B.K. *et al*. (2015) An atlas of genetic correlations across human diseases and traits. *Nat Genet*, **47**, 1236–1241.

De Vlaming,R. and Groenen,P.J.F. (2015) The current and future use of ridge regression for prediction in quantitative genetics. *BioMed Res Int*, **2015**, 1–18.

De Vlaming,R. *et al*. (2021) Multivariate analysis reveals shared genetic architecture of brain morphology and human behavior. *Commun Biol*, **4**, 1180.

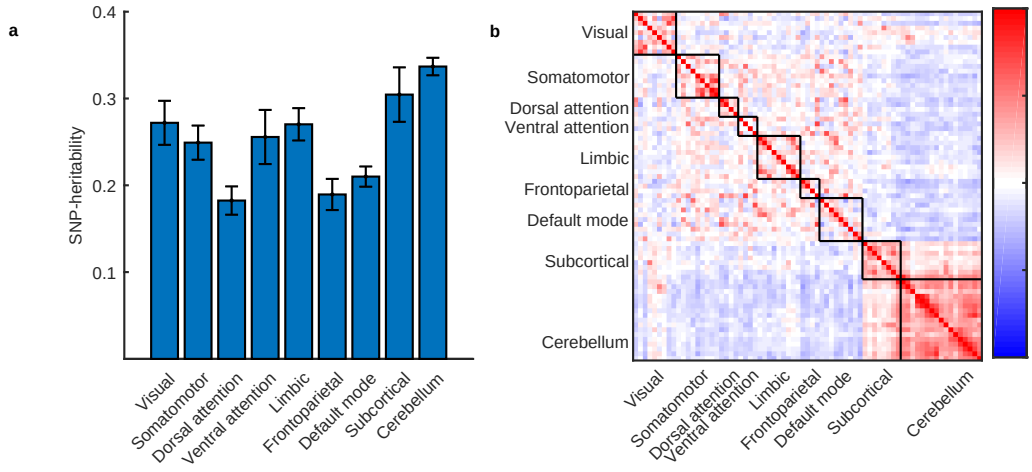

Figure 4: Results from MGREML estimation using data from the UK Biobank (from work by De Vlaming *et al.*, 2021).

Falconer,D.S. and Mackay,T.F.C. (1996) Introduction to quantitative genetics, fourth edition. *Pearson, Prentice Hall*.

Fletcher,R. (2013) Practical methods of optimization, second edition. *John Wiley & Sons*.

Friedman,J. *et al*. (2009) The elements of statistical learning, second edition. *Springer*.

Lee,S.H. *et al*. (2012) Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood. *Bioinformatics*, **28**, 2540–2542.

Loos,R.J.F. (2020) 15 years of genome-wide association studies and no signs of slowing down. *Nat Commun*, **11**, 1–3.

Polderman,T.J.C. *et al*. (2015) Meta-analysis of the heritability of human traits based on fifty years of twin studies. *Nat Genet*, **47**, 702–709.

Visscher,P.M. *et al*. (2017) 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am J Hum Genet*, **101**, 5–22.

Yang,J. *et al*. (2011) GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet*, **88**, 76–82.