Bibliographic report

Florian Privé

September 21, 2017

Introduction

The main objective of my thesis is to make Polygenic Risk Scores (PRS) that can differentiate an healthy person (control) from a diseased person (case) to be used for precision medecine. These PRS consist in a combination of information on DNA mutations at multiple loci of the genome, typically from hundreds of thousands to dozens of millions. PRS have been used for other goals such as finding a common genetic contribution between two diseases (S. M. Purcell et al. 2009).

There are three main concerns when constructing a PRS. The first one is that the scores have to be constructed while taking care of confounding effects. The second one, which is partially related to the first one and which has been on particular interest recently, is that these PRS can be used on a global basis, i.e. not only for the population they were train on. Finally, the third concern that need to be overcome is the size of the datasets. These datasets can require several gigabytes of memory or even terabytes for the largest datasets, e.g. the UK Biobank (Bycroft et al. 2017).

Polygenic Risk Scores

Since 2007, genome-wide association studies (GWAS) have multiplied. The goal of these studies is to find loci which variation is associated with a trait of interest, e.g. a status of disease. In a GWAS, each locus is tested independently. Then, researchers tried to find a way to combine all the GWAS results, i.e. the size and significance of the effects of all loci, in a predictive score.

For computing PRS for human diseases, what is widely used is the Pruning + Thresholding (P+T) model (Chatterjee et al. 2013; Dudbridge 2013; Golan and Rosset 2014). Under the P+T model, a coefficient of regression is learned independently for each locus along with a corresponding p-value (the GWAS part).
The loci are first clumped (P) so that there remains only loci that are weakly correlated with each other. Thresholding (T) consists in removing loci that are under a certain level of significance (P-value threshold to be determined). A polygenic risk score is defined as the sum of allele counts of the remaining SNPs weighted by the corresponding regression coefficients. As the weights are learned independently, this model can be applied to very large dataset, which is also why this is widely used.

Nowadays, people usually don't do the GWAS themselves but recover public results from published GWAS and use these summary statistics to train predictive models (Vilhjálmsson et al. 2015). Few people have used some more complex models such as sparse machine learning models to train accurate, yet simple models (Abraham et al. 2012).

All steps of the P+T procedure

The steps required in the P+T procedure are described in figure 1. As described in the previous section, the P+T procedure has 3 required steps before computing the scores: the GWAS, the clumping and the thresholding. Yet, computing the GWAS is not straight-forward because it requires to compute the first Principal Components of the genotype matrix. Indeed, Principal Components is used in genetics to assess the population structure of the dataset (Patterson, Price, and Reich 2006). Population structure is used as covariables in the models because it can be a major confounding effect to association with a trait (Patterson, Price, and Reich 2006).

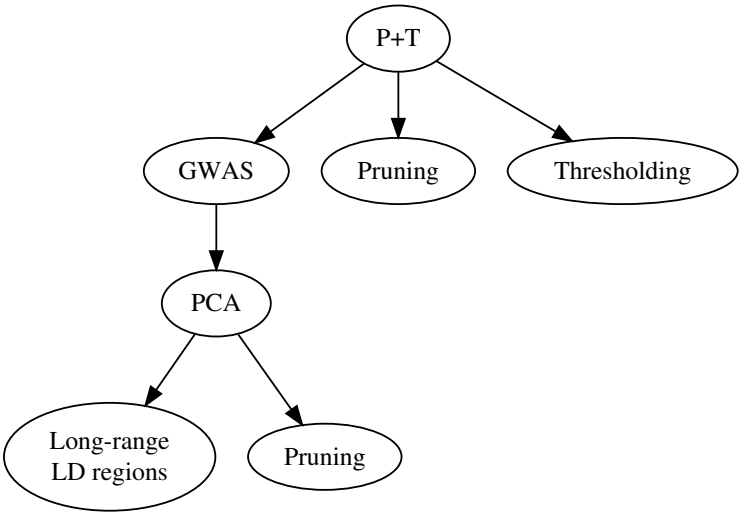

Figure 1: All steps required in the P+T procedure.

Moreover, computing PCs on the whole genotype matrix is not recommended. Indeed, because of recombination events or other biological events such as inversions, there is Linkage Disequilibrium (LD – correlation between nearby loci) in the data. To remove correlation alongside the genome, pruning is used and aims at better ascertaining population structure (Abdellaoui et al. 2013). Pruning is an algorithm that sequentially scan the genome for nearby loci in LD, performing pairwise thinning based on a given threshold of correlation. Then, there are also long-range of LD regions that will affect the PCA even after pruning. This is region will have a larger effect in loadings of PCA which can lead to false discoveries (Alkes L Price et al. 2008). So, even if the model is rather simple, because of population structure and patterns of LD, computing the model is not straight-forward.

Software for the computation of the P+T procedure

GWAS, pruning and thresholding can be done using the PLINK piece of software (Chang et al. 2015; Purcell et al. 2007). PLINK is often used because it is fast and memory-efficient tool written in C++ and used from the command line. It also provides some tools for conversion and quality control, which are mandatory processing steps.

PCA is commonly performed using EIGENSOFT, written in Java and executed from the command line or a Perl script (Galinsky et al. 2016; Patterson, Price, and Reich 2006). Yet, the first implementation is not fast enough for current sizes and the second one, based on random projections, is not accurate enough so that other implementations have been proposed (Abraham, Qiu, and Inouye 2016; Privé, Aschard, and Blum 2017).

Other species

For animal and vegetal species, knowing a "breeding value" is often of interest, which is a continous trait. Moreover, in these studies, the number of samples usually don't exceed a few thousands, which are manageable datasets. The preferred model for predicting a "breeding value" is the genome Best Linear Unbiased Prediction (gBLUP) (Meuwissen, Hayes, and Goddard 2001). This is basically a linear mixed model. This type of algorithm is quadratic with the number of samples which makes it impractible for large human datasets. Moreover, it is not directly suited for binary outcomes such as disease status.

Genome-wide data analysis

More generally, genome-wide datasets produced for association studies have dramatically increased in size over the past years. A range of software and data formats have been developed to perform essential pre-processing steps and data analysis, often optimizing each of these steps within a dedicated implementation. This diverse and extremely rich software environment has been of tremendous benefit for the genetic community. However, it has two limitations: analysis pipelines are becoming very complex and researchers have limited access to diverse analysis tools due to growing data sizes.

Consider first the basic tools necessary to perform a standard genome-wide analysis. Conversions between standard file formats has become a field by itself with several tools such as VCFtools, BCFtools and PLINK, available either independently or incorporated within large framework (Danecek et al. 2011; Li 2011; Purcell et al. 2007). Similarly, quality control software for genome-wide analysis have been developed such as PLINK and the Bioconductor package GWASTools (Gogarten et al. 2012). There are also several software for the computation of principal components (PCs) of genotypes, commonly performed to account for population stratification in association studies, including EIGENSOFT (SmartPCA and FastPCA) and FlashPCA (Abraham and Inouye 2014; Abraham, Qiu, and Inouye 2016; Galinsky et al. 2016; A L Price et al. 2006). Then, implementation of GWAS analyses also depends on the data format and model analyzed. For example, the software ProbABEL (Aulchenko, Struchalin, and Duijn 2010) or SNPTEST (Marchini and Howie 2010) can handle dosage data (genotype likekihoods from imputation), while PLINK version 1 is limited to best guess genotypes because of its input file format. Finally, there exists a range of tools for Polygenic Risk Scores (PRS) such as LDpred (Vilhjálmsson et al. 2015) and PRSice (Euesden, Lewis, and O'Reilly 2015), which provide prediction for quantitative traits or disease risks based on multiple genetic variants. As a result, one has to make extensive bash/perl/R/python scripts to link these software together and convert between multiple file formats, involving many file manipulations and conversions. Overall, this might be a brake on data exploration. Overall, this means that researchers are usually restricted on how they can manipulate and analyse the data they have access to.

Secondly, increasing size of genetic datasets is the source of major computational challenges and many analytical tools would be restricted by the amount of memory (RAM) available on computers. This is particularly a burden for commonly used analysis languages such as R, Python and Perl. Solving the memory issues for these languages would give access to a broad range of tools for data analysis that have been already implemented. Hopefully, strategies have been developed to avoid loading large datasets in RAM. For storing and accessing matrices, memory-mapping is very attractive because it is seamless and usually much faster to use than direct read or write operations. Storing large matrices on disk and accessing them via memory-mapping has been available for several years in R through "big.matrix" objects implemented in the R package bigmemory (Kane, Emerson, and Weston 2013). We provide a similar format as filebacked "big.matrix" objects that we called "Filebacked Big Matrices (FBMs)". Thanks to this matrix-like format, algorithms in R/C++ can be developed or adapted for large genotype data.

We developed two R packages, bigstatsr and bigsnpr, that integrate the most efficient algorithms for the pre-processing and analysis of large-scale genomic data while using memory-mapping (Privé, Aschard, and Blum 2017). Package bigstatsr implements many statistical tools for several types of FBMs (unsigned char, unsigned short, integer and double). This includes implementation of multivariate sparse linear models, Principal Component Analysis, matrix operations, and numerical summaries. The statistical tools developed in bigstatsr can be used for other types of data as long as they can be represented as matrices. Package bigsnpr depends on bigstatsr, using a special type of FBM object to store the genotypes, called "FBM.code256". Package bigsnpr implements algorithms which are specific to the analysis of SNP arrays, such as calls to external software for processing steps, I/O (Input/Output) operations from binary PLINK files, and data analysis operations on SNP data (thinning, testing, plotting). We used both a real case-control genomic dataset for Celiac disease and large-scale simulated data to illustrate application of the two R packages, including association study and computation of Polygenic Risk Scores. We also compared results from the two R packages with those obtained when using PLINK and EIGENSOFT, showing that using our software is easier and can even be faster. We finally showed results from new methods that we easily developed thanks to our data format and combinations from functions of our packages.

Future work

My second paper will be about comparing different methods for computing PRS and assess which methods give the best predictors depending on the genetic architecture of diseases, their heritability (i.e. the proportion of variance of the trait that is explained by the genotypes) and the population structure of the dataset analyzed. We expect that the methods we implemented in our packages, especially the regularized logistic regression based on efficient rules (R. Tibshirani et al. 2012; Zeng and Breheny 2017), should better account for Linkage Disequilibrium and thus give better predictions.

Then, we will analyze the UK biobank dataset, which is the largest and most complete dataset available for genetic analyzes. First, we want to compare published estimations of heritability to our prediction estimates. Secondly, we want to see if adding environment variables can add some predictive value to the genotype variables in order to better predict risks of disease or other traits. Finally, we are also interested in how can we make PRS that can be used for the global population, not only the population they were trained in.

References

Abdellaoui, Abdel, Jouke-Jan Hottenga, Peter De Knijff, Michel G Nivard, Xiangjun Xiao, Paul Scheet, Andrew Brooks, et al. 2013. "Population structure, migration, and diversifying selection in the Netherlands." *European Journal of Human Genetics* 21 (10). Nature Publishing Group: 1277–85. doi:10.1038/ejhg.2013.48.

Abraham, Gad, and Michael Inouye. 2014. "Fast principal component analysis of large-scale genome-wide data." Edited by Yu Zhang. *PLoS ONE* 9 (4). Public Library of Science: e93766. doi:10.1371/journal.pone.0093766.

Abraham, Gad, Adam Kowalczyk, Justin Zobel, and Michael Inouye. 2012. "SparSNP: Fast and memory-efficient analysis of all SNPs for phenotype prediction." *BMC Bioinformatics* 13 (1): 88. doi:10.1186/1471-2105-13-88.

Abraham, Gad, Yixuan Qiu, and Michael Inouye. 2016. "FlashPCA2 : principal component analysis of biobank-scale genotype datasets." *BioRxiv* 12 (May): 2014–7. doi:10.1101/094714.

Aulchenko, Yurii S, Maksim V Struchalin, and Cornelia M van Duijn. 2010. "ProbABEL package for genome-wide association analysis of imputed data." *BMC Bioinformatics* 11 (1): 134. doi:10.1186/1471-2105-11-134.

Bycroft, Clare, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T Elliott, Kevin Sharp, Allan Motyer, et al. 2017. "Genome-wide genetic data on ~500,000 UK Biobank participants." *Doi.org*, July. Cold Spring Harbor Laboratory, 166298. doi:10.1101/166298.

Chang, Christopher C, Carson C Chow, Laurent CAM Tellier, Shashaank Vattikuti, Shaun M Purcell, and James J Lee. 2015. "Second-generation PLINK: rising to the challenge of larger and richer datasets." *GigaScience* 4 (1): 7. doi:10.1186/s13742-015-0047-8.

Chatterjee, Nilanjan, Bill Wheeler, Joshua Sampson, Patricia Hartge, Stephen J Chanock, and Ju-Hyun Park. 2013. "Projecting the performance of risk prediction based on polygenic analyses of genome-wide association studies." *Nature Genetics* 45 (4): 400–405, 405e1–3. doi:10.1038/ng.2579.

Danecek, Petr, Adam Auton, Goncalo Abecasis, Cornelis A. Albers, Eric Banks, Mark A. DePristo, Robert E. Handsaker, et al. 2011. "The variant call format and VCFtools." *Bioinformatics* 27 (15). Oxford University Press: 2156–8. doi:10.1093/bioinformatics/btr330.

Dudbridge, Frank. 2013. "Power and Predictive Accuracy of Polygenic Risk Scores." *PLoS Genetics* 9 (3). doi:10.1371/journal.pgen.1003348.

Euesden, Jack, Cathryn M Lewis, and Paul F O'Reilly. 2015. "PRSice: Polygenic Risk Score software." *Bioinformatics* 31 (9). Oxford University Press: 1466–8. doi:10.1093/bioinformatics/btu848.

Galinsky, Kevin J., Gaurav Bhatia, Po Ru Loh, Stoyan Georgiev, Sayan Mukherjee, Nick J. Patterson, and Alkes L. Price. 2016. "Fast Principal-Component Analysis Reveals Convergent Evolution of ADH1B in

Europe and East Asia." *American Journal of Human Genetics* 98 (3): 456–72. doi:10.1016/j.ajhg.2015.12.022.

Gogarten, Stephanie M., Tushar Bhangale, Matthew P. Conomos, Cecelia A. Laurie, Caitlin P. McHugh, Ian Painter, Xiuwen Zheng, et al. 2012. "GWASTools: An R/Bioconductor package for quality control and analysis of genome-wide association studies." *Bioinformatics* 28 (24). Oxford University Press: 3329–31. doi:10.1093/bioinformatics/bts610.

Golan, David, and Saharon Rosset. 2014. "Effective genetic-risk prediction using mixed models." *American Journal of Human Genetics* 95 (4). The American Society of Human Genetics: 383–93. doi:10.1016/j.ajhg.2014.09.007.

Kane, Michael J, John W Emerson, and Stephen Weston. 2013. "Scalable Strategies for Computing with Massive Data." *Journal of Statistical Software* 55 (14): 1–19. doi:10.18637/jss.v055.i14.

Li, Heng. 2011. "A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data." *Bioinformatics* 27 (21). Oxford University Press: 2987–93. doi:10.1093/bioinformatics/btr509.

Marchini, Jonathan, and Bryan Howie. 2010. "Genotype imputation for genome-wide association studies." *Nature Reviews. Genetics* 11 (7). Nature Publishing Group: 499–511. doi:10.1038/nrg2796.

Meuwissen, T. H. E., B. J. Hayes, and M. E. Goddard. 2001. "Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps." *Genetics* 157 (4): 1819–29. http://www.genetics.org/content/157/4/1819.

Patterson, Nick, Alkes L. Price, and David Reich. 2006. "Population structure and eigenanalysis." *PLoS Genetics* 2 (12): 2074–93. doi:10.1371/journal.pgen.0020190.

Price, A L, N J Patterson, R M Plenge, M E Weinblatt, N A Shadick, and D Reich. 2006. "Principal components analysis corrects for stratification in genome-wide association studies." *Nat Genet* 38. doi:10.1038/ng1847.

Price, Alkes L, Michael E Weale, Nick Patterson, Simon R Myers, Anna C Need, Kevin V Shianna, Dongliang Ge, et al. 2008. "Long-Range LD Can Confound Genome Scans in Admixed Populations." Elsevier. doi:10.1016/j.ajhg.2008.06.005.

Privé, Florian, Hugues Aschard, and Michael G B Blum. 2017. "Efficient management and analysis of large-scale genome-wide data with two R packages: bigstatsr and bigsnpr." *BioRxiv*, September. http://biorxiv.org/content/early/2017/09/19/190926.abstract.

Purcell, Shaun M, Naomi R Wray, Jennifer L Stone, Peter M Visscher, Michael C O'Donovan, Patrick F Sullivan, Pamela Sklar, et al. 2009. "Common polygenic variation contributes to risk of schizophrenia and bipolar disorder." *Nature* 10 (AuGuST). Nature Publishing Group: 8192. doi:10.1038/nature08185.

Purcell, Shaun, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel A R Ferreira, David Bender, Julian Maller, et al. 2007. "PLINK: a tool set for whole-genome association and population-based linkage analyses." *American Journal of Human Genetics* 81 (3). Elsevier: 559–75. doi:10.1086/519795.

Tibshirani, Robert, Jacob Bien, Jerome Friedman, Trevor Hastie, Noah Simon, Jonathan Taylor, and Ryan J Tibshirani. 2012. "Strong rules for discarding predictors in lasso-type problems." *Journal of the Royal Statistical Society. Series B, Statistical Methodology* 74 (2): 245–66. doi:10.1111/j.1467-9868.2011.01004.x.

Vilhjálmsson, Bjarni J., Jian Yang, Hilary K. Finucane, Alexander Gusev, Sara Lindström, Stephan Ripke, Giulio Genovese, et al. 2015. "Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores." *The American Journal of Human Genetics* 97 (4): 576–92. doi:10.1016/j.ajhg.2015.09.001.

Zeng, Yaohui, and Patrick Breheny. 2017. "The biglasso Package: A Memory-and Computation-Efficient Solver for Lasso Model Fitting with Big Data in R." *JSS Journal of Statistical Software MMMMMM YYYY* VV (January). doi:10.18637/jss.v000.i00.