DEPARTMENT OF MATHEMATICS

TECHNISCHE UNIVERSITÄT MÜNCHEN

Master's Thesis in Mathematics in Data Science

Machine learning for genetic risk prediction

Simon Zabrocki

DEPARTMENT OF MATHEMATICS

TECHNISCHE UNIVERSITÄT MÜNCHEN

Master's Thesis in Mathematics in Data Science

Machine learning for genetic risk prediction

Maschinelles Lernen zur Vorhersage des genetischen Risikos

Author:	Simon Zabrocki
Supervisor:	Dr. Matthias Heinig
Advisor:	Dr. Matthias Heinig
Submission Date:	30th July 2020

I hereby declare that this master's thesis in mathematics in data science is my own work and that no other sources have been used except those clearly indicated and referenced.

Munich, 30th July 2020

Simon Zabrocki

Zabrocki

Acknowledgments

First, I would like to thank my supervisor Matthias Heinig for giving me the opportunity to write this thesis at the institute for computational biology at the Munich Helmholtz Zentrum. Under his guidance I discovered the rich and fascinating field of genomics.

I would also a like to deeply thank Ines Assum for her great insights on biology and research. She was of great advice on all matters, from biology to modeling and computing.

Finally, I would like to thank all the members of the Genetic and Epigenetic Gene Regulation lab at Helmotz for the helpful questions and conversations during meetings.

Abstract

Solving the "missing heritability problem" is essential to the accurate estimation of genetic risks. Current linear models are not able to capture much of the variability observed in complex traits. Furthermore, despite being interpretable in the mathematical sense, these models provide little to no biological and mechanistic insights into the studied disease. At last, in spite of evidences of the importance of environmental factors, genetic and non genetic risks are rarely studied jointly. In parallel, the past two decades have seen the development of effective computational methods to estimate ever more complex statistical models. The goal of this thesis, is to explore how these machine learning methods can be leveraged to solve the problem at hand. The original method described in this thesis aims at providing a complex but interpretable model. Using well established deep variational models, we investigate links between obesity, genetics information and environmental factors.

Die Lösung des "Problems der fehlenden Heritabilität" ist für die genaue Schätzung genetischer Risiken unerlässlich. Aktuelle lineare Modelle sind nicht in der Lage, einen Großteil der bei komplexen Merkmalen beobachteten Variabilität zu erfassen. Darüber hinaus liefern diese Modelle, obwohl sie im mathematischen Sinne interpretierbar sind, wenig bis keine biologischen und mechanistischen Erkenntnisse über die untersuchte Krankheit. Schließlich werden genetische und nicht-genetische Risiken trotz der nachgewiesenen Bedeutung von Umweltfaktoren selten gemeinsam untersucht. Parallel dazu wurden in den letzten zwei Jahrzehnten effektive Berechnungsmethoden entwickelt, um immer komplexere statistische Modelle zu schätzen. Das Ziel dieser Arbeit ist es, zu erforschen, wie diese Methoden des maschinellen Lernens genutzt werden können, um das vorliegende Problem zu lösen. Die ursprüngliche Methode, die in dieser Arbeit beschrieben wird, zielt darauf ab, ein komplexes, aber interpretierbares Modell zu erstellen. Unter Verwendung etablierter tiefer Variationsmodelle untersuchen wir Zusammenhänge zwischen Fettleibigkeit, genetischen Informationen und Umweltfaktoren.

Contents

1. Introduction

The completion of the Human Genome Project in 2003 came with the promise to revolutionize biology and medicine. Direct access to DNA sequences, which was considered as the blue print of living organisms, would provide a seemingly infinite amount of knowledge on biological processes. These hopes would soon be confronted with reality, sequencing the genome was only the first step to deciphering DNA. First methods developed by researchers were confronted to the "missing heritability" problem. Associations were found between genetic variations and traits, but their effect size were far too small to explain most phenotypic variations. Genomics proved more complicated than "one trait equals one gene equals one location in the genome". In 2009, Manolio and Collins [1] gave their insightful thoughts about the field of genetics. A decade later, these thoughts have not aged a bit. They give a number of practical recommendations to improve genetic risk models. First, unlike Mendelian diseases where a single variation lead to disease, more common traits are the result of a large number of low effect variations. To estimates these effects, large cohorts and meta analyses are required. To improve generalizability of these analyses, they must be expanded to non European samples. Additionally, gene-gene and gene-environment interactions must be thoroughly and rigorously investigated. As of 2020, all these recommendations are still up to date. Since 2009 however, a class of statistical methods, machine learning, have displayed ground breaking performances in the fields of computer vision and natural language processing. Advances in software and hardware have made it possible to fit complex models with millions of parameters relatively easily. The goal of this thesis is to explore if and how these computational methods can solve problems encountered in genomics. First, an introduction provides general considerations to get a better understanding of the challenge at hand. Then, classical genomics methods are explained and illustrated to demonstrate their limits. With a clear objective in mind, relevant statistical methods are detailed and explored in the context of genomics. In particular, an original approach using recent advances in variational inference is described in depth. Finally, results of these methods are presented and discussed.

1.1. Machine learning in genetics

When building a machine learning model for genetics, a number of caveats need to be considered. A first constraint comes from the dimensionality of the data. The large number of variations studied leads to technical implementation challenges that forbids quick trial and error methods. As with any project involving a substantial amount of data, formats, conventions and naming vary from one dataset to another. Being hasty to feed the data in the latest "state of the art" model put all of the analysis done later at risk of failure. Unlike

other data like images, text or sound, raw genetic data is not easily visualized or interpretable on its own. It is not straightforward to see whether or not something went wrong in the pipeline at a given point. This is why the processing pipeline needs to be carefully crafted and tested before running. Furthermore, preprocessing steps can bias the results and need to be properly accounted for in the cross validation process.

Furthermore, the small cohorts size relative to the number of raw features imposes the curse of dimensionality on every analysis done. As explained later in the section on Genome Wide Association Studies (GWAS), even sound and robust methods can lead to a number of false discoveries and spurious correlations. While in some applications relying on spurious correlations is perfectly justifiable, in the field of genetics the model is merely a means to an end. Of course, predicting diseases given genetic information with good accuracy is desirable. However, this is far from sufficient for success. A black box, unexplained model predicting body mass index with great precision would have very little to no scientific or commercial value. Instead, the value of the models comes from what they tell on the pathway from genome to disease.

"Why is the model telling us that this person is at risk of obesity ?", "How are the genetic loci interacting with one another to cause an increase in cardiovascular disease risk ?" are questions far more relevant than "How to improve the accuracy of my classification model by x point of percentage ?". The ideal model would have decent predictive performance but more importantly would help identifying variants with clear and specific biological roles. Only then can biologists use these models to experiment, understand and create new data on biological systems.

As a mathematician it is tempting to leave the biology completely on the side and focus solely on some fascinating computational challenges. However, remembering where the data comes from and what is needed out of it is in my opinion crucial to steer ideas and implementations into sensible directions.

1.2. Biology prerequisites

While this thesis focuses on mathematics and not biology, a number of biological terms, definitions, and concepts are needed to correctly understand the topic and challenges at hand. Here is a minimal introduction to the basis of genetics necessary for this study inspired from general public sources [2, 3, 4, 5].

In every cell of every living organism, a chemical compound called DNA encodes the instructions needed for the organism to develop. This molecule has a specific double stranded structure called double helix. Each of these strands are made up of 4 units called nucleotides: adenine (A), thymine (T), guanine (G), and cytosine (C). On the double helix, two pairs of

bases are possible: A-T and G-C. The arrangement of these bases encodes genetic information. Much like bits in a computer, it is the lowest level of information encoding. The set of all these bases of an organism is called the genome or genotype. In human beings, the genome has roughly 3 billion base pairs.

The pairs can be grouped together in meaningful units of information called genes. The genes contain information on how to create proteins, the lower level building block of biological systems. Genes can be more or less expressed by different cells and different organisms. This expression can turn genes on and off as well as affect the quantity of proteins produced. The type, location, time and quantity of protein produced define in turn higher level traits.

Genes and gene regulation are encoded in two separate regions, coding and non coding. The coding regions encode information on how to build proteins. Only a small fraction of the genome is coding ($\sim 1\%$). The rest of the genome plays more complex roles in the regulation of gene expression. While coding regions are well understood and mapped, non coding regions still hide plenty of mysteries.

Since humans share a mostly similar genetic code, it is enough to focus on parts of the genome with enough variation across the human population. Most variation is the result of insertion, deletion and single nucleotide polymorphisms (SNPs). An individual differs from another by around 4 to 5 million SNPs. In total, hundreds of millions of SNPs are referenced. This study focuses on SNPs. When these SNPs have an impact on gene expression, they are referred to as eQTL (expression Quantitative Trait Loci). eQTLs are valuable to map what role these variants may have at the cell and tissue level. Our goal is to detect, understand and measure the effect of this variation on the phenotypes of individuals.

The whole challenge of genetics is to understand how this variation at the molecular scale translates to observable and measurable traits. These traits are referred to as phenotypes. While this may seem straightforward, a number of challenges must be overcome. The first one comes from linkage disequilibrium (LD), an essential property of genomics data. Linkage disequilibrium is defined as the non random association of two variants. While different measures of LD exist, in this thesis two SNPs are said to be in linkage disequilibrium if their correlation is above predetermined level. Natural selection or mechanisms of exchange of genetic information can lead to these non random associations between genes. One example of such mechanism is that sequences that are close together in the genome tend to be inherited together. This particular structure of genomics data leads to uncertainty on the causal nature of findings. A variant located in the vicinity of a true causal variant can have strong correlation with a trait despite being completely unrelated biologically.

Additionally, the gap from variants to trait is huge, based on genetic information, genes are expressed by cells in different ways. These cells then define the behavior of different tissues

across the body. These tissues in turn define phenotypes at the organism level. At all levels, proteins, cells and DNA interact together creating an intricate circuitry. To make matters worse, while genetic information stays relatively stable through time within an organism, gene expression and tissue behaviors can be altered by external environmental factors.

It is possible to study the problem at different scales from sub-cell level to full organisms. This thesis will focus on studying phenotypes at the organism scale. While we use knowledge from other scales such as tissue and single cells, these are not the focus of the study.

1.3. Available Data

In this section, the data used in the project is succinctly presented.

1.3.1. Presentation

As mentioned in the biology prerequisite section, the thesis focus on single nucleotide polymorphism (SNPs). SNPs are available for different cohorts. The first cohort comes from the one thousand genome project [6]. This cohort is openly accessible and contains a few thousands samples. This dataset is extremely useful for testing pipelines. However, it does not contain any phenotypic information. To built polygenic risk scores, we rely on the [7] dataset. This dataset contains the genome of 500 000 persons living in the UK as well as a large number of phenotypic and environmental factors. Some of these additional data are biomarkers measured by healthcare professionals, other are answers to questionnaires. Therefore these complementary data should not always be considered as "ground truth". Additionnaly, because of its large size, analyses on the dataset have more power. Consequently many researchers use the bio bank to fit and test their methods. This poses potential reproducibility issues. To alleviate this problem the KORA [8] cohort comprised of 3000 samples is used for testing the results.

Expression data supplement the genetic information contained in these cohorts. These data consist of linear expression models [9, 10] for different tissues. Their use is two folds. First, they may be used as prior information to filter SNPs. They also can be used in retrospect to analyze the SNP hits found by the different methods and uncover potential biological pathways linked to individual SNPs.

Finally, this study relies on existing polygenic risk scores [11]. This data is a valuable source to compare, sanity check and inspect results from different models. These scores are built by independent research groups mostly using the UK bio bank dataset.

1.3.2. Mathematical formulation

Encoding SNPs As mentioned earlier, this thesis is restricted to the study of single nucleotide polymorphisms (SNPs). To perform any kind of quantitative analysis, this genetic information needs to be encoded numerically. The standard way of doing so is through label encoding. For each variant, the alleles can have only three different states: homozygous-dominant, heterozygous, homozygous-recessive. We encode those states respectively as 0, 1 and 2 .

(For the homozygous case, "dominant" and "recessive" are sometimes called "effect" and "reference" alleles or “alternative” and “reference”. From dataset to dataset the effect and reference alleles are sometimes swapped. While this is not a difficult problem, ensuring that the encodings are consistent across data is crucial to the quality of the analysis.)

We study a cohort of N individuals. Each of those individual is represented by a vector comprised of genetic, phenotypic and additional information. Depending on the cohort N varies from 10^3 to 10^5 samples.

SNP data For each sample i in the cohort, a vector x_i holds the genetic information. The coordinates are either 0, 1 or 2. Its dimensionality d can vary from $d \sim 10^6$ when considering the full genome to $d \sim 10^2$ when looking at selected SNPs. Without prior processing, $d >> N$. We can therefore consider that x_i is highly dimensional. Additionally, x_i is sparse since the proportion of 1 and 2 is negligible compared to the one of 0.

For the rest of this thesis, $X \in \{0, 1, 2\}^{N \times d}$ refers to SNP data.

Target phenotypic Data The second type of data available is the target phenotype data. Each individual i has p phenotypes represented by y_i. The phenotypes can be discrete (presence of diabetes) or continuous (body mass index).

For the rest of this thesis, $Y \in \mathcal{R}^{N \times p}$ refers to target phenotype data.

Additional Data Finally, additional data is available for each person i. These data can be non target phenotypes (such as gender or height) as well as information on the individual (such as age or country of residence). They can also be environmental variables such as smoking or diet habits.

For the rest of this thesis, $C \in \mathcal{R}^{N \times q}$ refers to target additional data.

Snapshot of the data In order to illustrate the previous paragraph, here is a subset of the available data. It contains a random subset of SNPs from chromosome 15 from the 1000 genome project [6] as well phenotype and additional data. With the previous notation, we could define X would be columns $[0 : 9999]$, y could be "Population" and C could be "Gender".

	0	1	2	3	4	5	6	7	8	9	...	9994	9995	9996	9997	9998	9999	Gender	Population	Phase1_LC_Platform	Super_Population
0	0	0	0	0	1	0	0	0	1	0	...	0	0	0	0	0	1	male	GBR	ILLUMINA	EUR
1	0	0	0	0	1	0	0	0	0	0	...	0	0	0	0	0	0	female	GBR	ABI_SOLID	EUR
2	0	0	0	0	1	0	0	0	0	0	...	0	0	0	0	0	1	female	GBR	ABI_SOLID	EUR
3	0	0	0	0	1	0	0	0	0	0	...	0	0	0	0	0	0	female	GBR	ILLUMINA	EUR
4	0	1	0	0	0	0	0	0	1	0	...	0	0	0	0	0	0	male	GBR	ABI_SOLID	EUR

5 rows × 10004 columns

Figure 1.1.: Snapshot of genetic data from 1k genome

1.4. Ethical considerations

Genetics more than any other field deals with extremely sensitive data. DNA is widely perceived as the essence of who we are. While this is true to a degree, plenty of traits do not find causal explanations in the genome. Nevertheless this "essential" and almost transcendental nature of DNA leads to two considerations.

First, in practice, genetic data must be handled with great care. Data must be anonymized and shared securely. Naturally this adds some practical constraints such as having to access the data on dedicated servers. However, patients trust the scientific community with the most personal data there is about themselves and this is only normal to handle it with care. This is ever more true when working on generative models that have the potential to simply memorize data. As a safeguard, privacy metrics should be computed for generative models trained on sensitive data.

More importantly, interpretation of results should be done with extreme care. Announcing hastily that a given trait, biological or behavioral can be "explained by" or "associated" to variations in the genome can lead to misinformation at best and extreme harm at worst. Since the genome cannot be changed at will, any disease or undesirable trait "associated" to genetics by poor interpretation can be perceived as an irrevocable sentence. This kind of over interpreted associations is potential fuel to all kinds of dangerous statements and practices.

Until an association is found across multiple studies on multiple cohorts from different backgrounds, explaining a trait by genetics may be premature. Even then, repeatable association is not enough. Once an association is robustly observed, a clear mechanistic pathway from the variations at the DNA level to the trait at the macroscopic level needs to be experimentally identified to ensure the casual nature of the association. At last, having a causal link is not quite the final step, once this clear mechanistic explanation is unveiled, the effect size needs to be clearly estimated and stated. A variant can be statistically significantly associated with a trait, with a clear causal link and yet still be biologically and scientifically irrelevant. If the effect size is negligible before other non genetic factors this discovery may not be of much use. This process is demanding but ensures an informed statement on the link between the

genome and a trait. Appendix A explores an article that potentially over interprets results leading to questionable conclusions.

2. Motivation

2.1. Genome Wide Association Studies

2.1.1. Presentation

Genome Wide Association Studies (GWAS) aim at finding specific loci in the genome that are associated with a given trait. They are conducted on cohorts of a few hundreds to hundreds of thousands of non related samples. It is a simple but proven way of identifying potentially causal variations. For each locus, the association between the SNP and trait is measured through a statistical test (e.g χ^2 test). If the p value is low enough (e.g. $\leq 10^{-8}$), the SNPs are considered as hits and further studied.

The benefits of GWAS are its computational simplicity (linear in the number of loci) and its interpretability. Once a locus is considered a hit, it can be linked to coding and non coding regions and later to biological functions. The limitation of GWAS comes from its inability to account for potential interactions between SNPs. SNPs are tested independently of one another. However, this assumption is too strong. Biological systems are extremely intricate with plenty of interactions, cascades of signals and regulatory circuits.

A more conceptual issue of GWAS stems from the high dimensionality of genetic data. Millions of SNPs are analyzed on only thousands of samples leading to both false and missed discoveries. Because effect sizes of individual SNPs can hardly be estimated in advance, power analyses are often challenging.

At last, GWAS measures association only. On their own they do not measure effect size or reveal causal links. Still, GWAS are a useful dimensionality reduction tool to extract potentially interesting loci.

2.1.2. Simulation

To demonstrate how GWAS works, we create the following simulation. n_{snp} are drawn of n_{snp} binomial distributions with parameters $B(n_{sample}, MAF_i)$. n_{sample} is the population size. MAF_i is the minor allele frequency of SNP i. The minor allele frequency MAF_i is drawn from a uniform distribution with parameter $U(0.05, 0.3)$. To create a trait, we select a single causal SNP. Given this SNP, a sample has trait $y = 1$ with probability $P(trait|variant)$ if a variant is present on the SNP and $P(trait|no\_variant)$ otherwise. We assume that the variant increases the chance of observing the trait ie $P(trait|variant) > P(trait|no\_variant)$.

As we simulate only $10^3 \sim 10^4$ SNPs, we arbitrarily set the significance level to $p_{value} \leq 10^{-3}$ instead of the genome wide significance level of $p_{value} \leq 10^{-8}$.

This simulation has limitations. First, SNP only have 2 possible values instead of 3 in actual genetic data. Then SNPs are independent from one another. We therefore cannot recreate the correlation structure present in real data. Second, for simplicity the trait only depends on a single SNP. Finally, in real data minor allele frequencies are not distributed uniformly. Despite its simplicity, this "back of the envelope" simulation allows to explore different use cases of the GWAS method to show its strengths and limitations. Simulation parameters are summarized in table 2.1. For the sake of this argument we explore only 3 cases by varying effect size, number of SNPs and number of samples. Additionally a 4th example illustrating limitations with regards to SNP interaction is given.

case	n_{sample}	n_{snp}	P(trait | variant)	P(trait | no_variant)
1	1000	100	0.8	0.1
2	1000	100	0.15	0.05
3	1000	100000	0.15	0.05

Table 2.1.: GWAS simulation parameters

Case 1: Large number of samples and high effect

$$n_{snp} << n_{sample} \text{ and } P(trait|variant) \sim 1$$

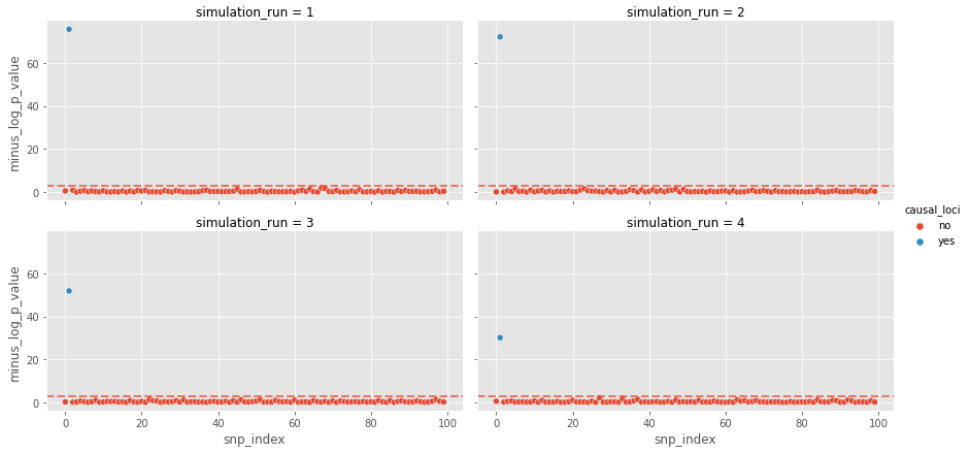

Figure 2.1.: Case 1

In the case of high number of samples and high effect, the GWAS is able to recover the causal SNP with ease every time. There is no false positive of false negative. This situation occurs for Mendelian traits.

Case 2: Large number of samples and low effect

$$n_{snp} << n_{sample} \text{ and } P(trait|variant) << 1$$

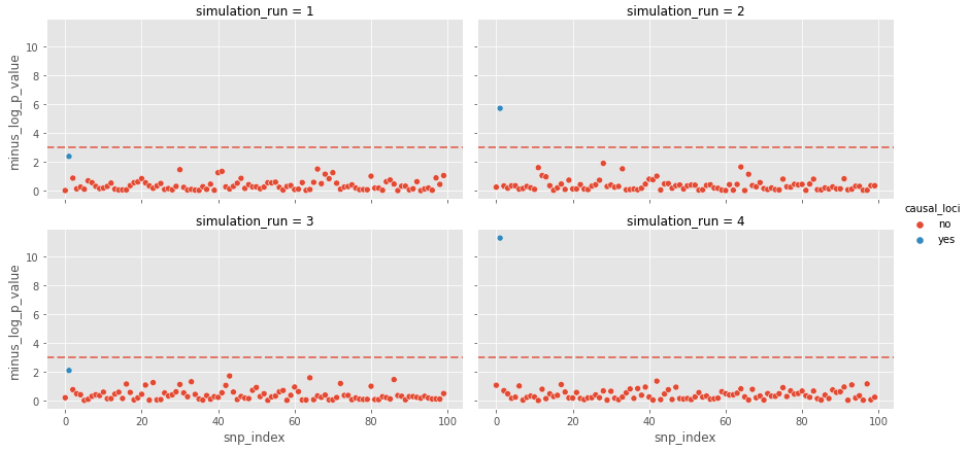

Figure 2.2.: Case 2

With enough samples, even a low effect variant can be recovered by the association study. However, compared to case 1, the result is less clear cut. On some occasion, potential false negative appear.

Case 3: Low number of samples and low effect

$$n_{snp} >> n_{sample} \text{ and } P(trait|variant) << 1$$

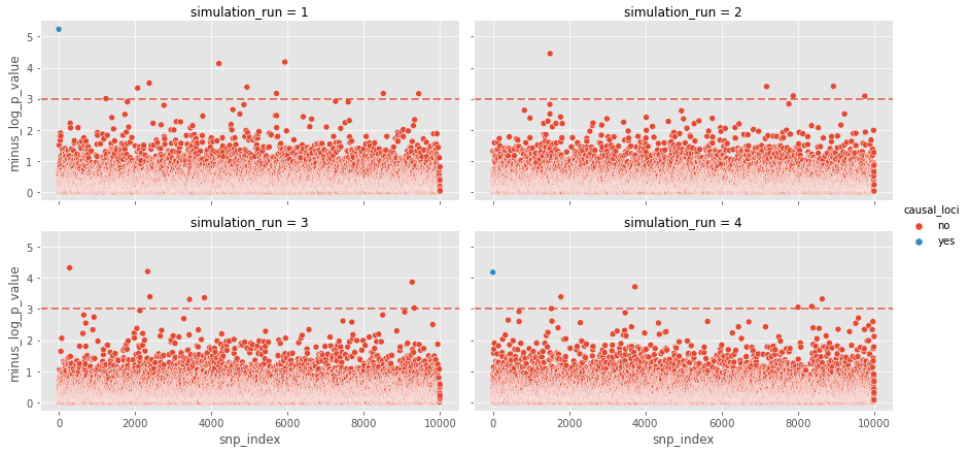

Figure 2.3.: Case 3

GWAS are often used in cases different from 1 and 2. In most datasets, the number of SNPs is in the millions whereas the number of samples is only in the thousands to hundreds of

thousands. Moreover, GWAS studies look for loci with weak effects. In this scenario, the GWAS can fail to recover the causal SNP. It also leads to a number of strong false positives as suggested in figures 2.3 and 2.4. The distribution of p values for causal SNPs and non causal SNPs is estimated by simulating 10k causal and non causal SNPs. On this plot it is apparent that while the p value distribution of causal loci is shifted compared to non causal loci, both distributions overlap.

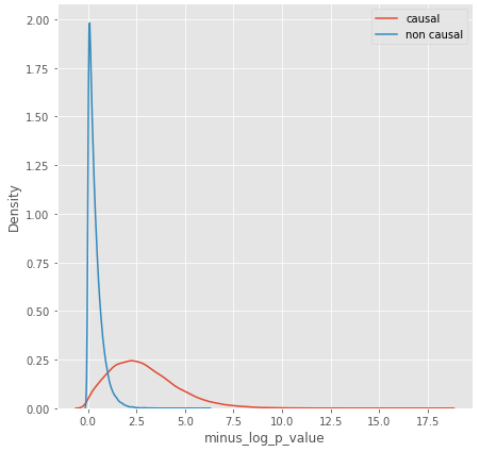

Figure 2.4.: negative log p values distributions

Case 4: SNP interactions In case 4, we look at a different limitation of GWAS. As explained earlier, the method looks at each SNP individually, potentially missing interactions between SNPs. In this simulation, we create two SNPs with $MAF = 0.5$. Instead of creating a trait using variants from a single SNP, we combine the two SNPs using the exclusive union operator: The variant pairs $(snp_0 = 1, snp_1 = 0)$ and $(snp_0 = 0, snp_1 = 1)$ both lead to the trait with probability $P(trait|variant)$. The pairs $(snp_0 = 0, snp_1 = 0)$ and $(snp_0 = 1, snp_1 = 1)$ lead to the trait with probability $P(trait|no_{variant})$. Just as before, the variant increases the chance of observing the trait, ie $P(trait|variant) > P(trait|no\_variant)$. Effectively, snp_0 and snp_1 cancel one another.

The plot below shows that snp_0 and snp_1 are not associated but snp_0 XOR snp_1 is without a doubt. We observe that the number of variants in groups with $trait = 1$ and $trait = 0$ are similar on snp_0 and snp_1. However, when investigating snp_0 XOR snp_1, the difference between groups becomes apparent. The trait appears only on individuals when snp_0 XOR $snp_1 = 1$

Unlike the previous cases, effect and sample size are not part of the challenge. Instead this is the major computational limitation of GWAS. The method is not built to capture situations that are potentially frequent among the huge possible pairs of SNPs. Since biological systems often involve large number components, we could very well imagine interactions involving even a greater number of SNPs.

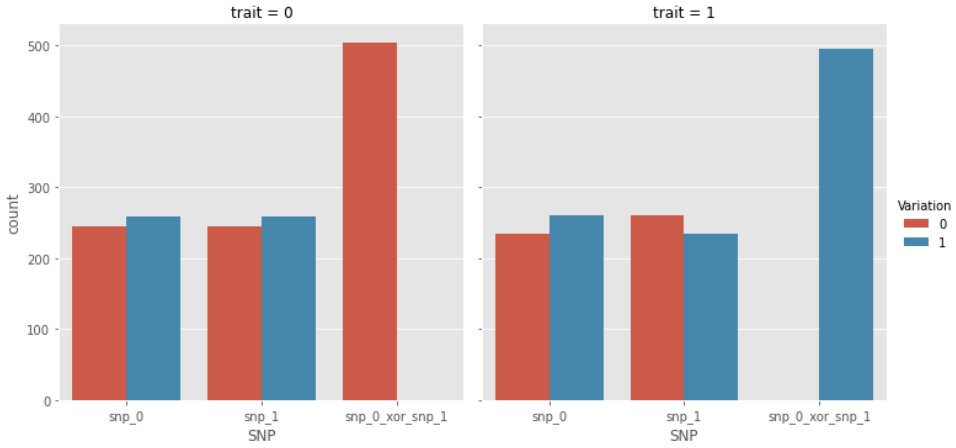

Figure 2.5.: Case 4

Remarks on the XOR operator The choice of XOR operator for this illustration is motivated by the following fact: The XOR operator creates classes that are non linearly separable. All linear models will fail to capture this behavior. This very simple example justify the need for models that are not only multivariate but also non linear.

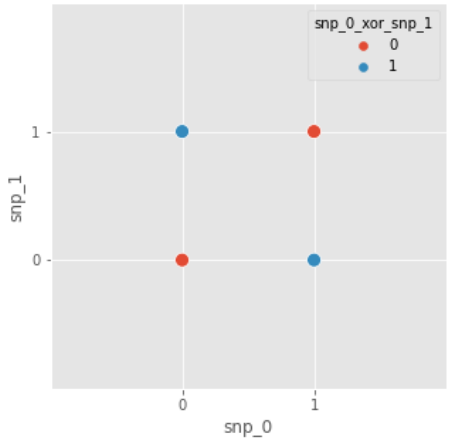

Figure 2.6.: XOR operator

2.1.3. Discussion

This introduction to GWAS showed two challenges related to the study of genetic data. In case 4, we showed that GWAS were not built to capture multivariate behaviors in genetic data. While this is a serious limitation, the full machine learning tool box can be deployed to try and fit ever more complex models to the data. These advanced models will without a doubt capture higher order behaviors. Naturally, fitting non linear, multivariate models comes at

the price of interpretability. This a fundamental challenge that will be explored further in this thesis.

In cases 3 we showed that GWAS could lack power leading to false positives and false negatives. This problem however is not computational or related to the specific GWAS approach. State of the art machine learning models would encounter similar issues as well. Instead it is a fundamental challenge stemming from the high dimensional nature of genetic data. The size of even the largest cohorts are still orders of magnitude smaller than the number of loci studied. To solve this issue, it is often relevant to look for prior knowledge elsewhere first and only then improve the computational models. For instance, it is possible to link individual SNPs to low level biological systems such as proteins or regulatory circuits. This knowledge accessible through data on specific cells and tissues helps reduce the gap between information in the genome and high level traits. While studying raw genetic information through computational methods alone can lead to discoveries, only a comprehensive approach covering all the steps of the biological pathway can lead to a full understanding of the biological circuitry.

More generally and just as in any other field, reproducibility on independent datasets is necessary to ensure veracity of findings. Independent cohorts from heterogeneous populations ensure that discoveries are solid.

2.2. Polygenic Risk Scores

While GWAS helps uncover potentially interesting loci in the genome and even assess the total heritability of a trait via LD-score regression, they do not allow for personalized risk score predictions. Polygenic risk scores attempt to provide better models by adding the effects of many risk variants. The content of this section can be found in this review article [12] on polygenic risk scoring which provides a clear presentation of the topic.

2.2.1. Presentation

Fundamentally, polygenic risk scoring involves building a predictive model that links genetic and phenotypic data. The genetic data is usually composed of single nucleotide polymorphisms (SNP). The phenotypic data is generally composed of disease markers. These markers can be continuous or categorical. In this thesis, the primary focus is on linking variation in the genome to the body mass index. At first glance, the task appears rather straightforward, yet, a number of particularities must be accounted for when building such risk score.

Classically, a polygenic risk score (PRS) is a weighted sum of SNPs. Each SNPs' effect on the observed phenotype is summed to obtain a score. The score is polygenic because it

includes many SNPs, in practice, the more SNPs available the greater the correlation between the phenotype and the score. Mathematically, the classical polygenic risk score is simply:

$$y = \sum_i \beta_i X_i \tag{2.1}$$

Where y is the phenotype, X_i is SNP i and β_i is the effect size of SNP i. This model can simply be estimated via a linear regression. When predicting a categorical output, a logistic regression is used instead. As mentioned in the previous section, high dimensionality is a key feature of genomic data. Therefore to get an effective model, shrinkage must be applied to the coefficients. This shrinkage can be done via regularization (for instance LASSO) or via significance thresholds as detailed later.

While shrinkage helps with high dimensionality of the data, linkage disequilibrium greatly complicates the estimation of a true causal polygenic risk score. As presented in the introduction, biological mechanisms can lead to non random association between pairs of variants. For this reason, polygenic risk scores can be studied in pair with additional biological evidence to confirm the causal nature of selected variants. More over, to compute a robust polygenic risk score, the genetic population structure must be carefully accounted for. Because geography plays a crucial role into how people choose their mates, genetic variations do not have the same structure from one place to the other. The geographical structure also correlates with different environmental factors (cultures, diets or access to healthcare). Therefore, keeping in mind the source of the data and correcting for population structure is an essential part of polygenic risk score modeling.

Note that theses two challenges are completely independent of the model chosen and should be addressed regardless of whether a linear regression or the latest deep learning architecture is applied.

Polygenic risk scores can have two applications. First, there might be potential clinical applications. A PRS could be used to stratify the population according to disease risks and provide targeted treatments. Given the relatively low predictive power and the experimental nature of PRS, this application is still very hypothetical. A more realistic application is to use the scores to get insights on biological processes causing a disease. Variations can be linked to specific parts of the biological circuitry. These links can help understand how and why disease occurs.

2.3. Research questions

In the previous sections, different gaps in the current methodologies were identified. First, high dimensionality and linkage disequilibrium both lead to difficulties in identifying causal SNPs and estimating their effects. Rather than taking the computational route, using biological priors to select variants and estimate their effect could solve the dimensionality challenge and

lead to interpretable models. Potential biological priors available are gene expression models. They contain eQTLs and their effects on gene expression. Selecting eQTLs in relevant tissues for a given disease could sufficiently reduce the dimensionality for the next modelling steps. To assess if this route is promising, we first check how close are existing scores to these prior. It boils down to the following hypothesis:

Hypothesis 1. *Gene expression models can be used as prior for polygenic risk scoring*

While computing risk scores is the most crucial step of the genomics pipeline, understanding how peer reviewed scores perform can help uncover potential applications and improvements. In particular, to apply risk score on individual patient we must check the following:

Hypothesis 2. *Risk scores have enough predictive power for prediction at the individual level*

Current articles mostly focus on regression metrics but other aspects can be explored to get a better grasp at what the risk scores capture. If risk score correlates with factors others than the target disease, their interpretation becomes more complex. To check this, we test the following hypothesis:

Hypothesis 3. *Risk scores depend on one another and population structure*

At the moment, established and peer reviewed risk scores are linear. Using machine learning to build non linear models is a natural next step. More than just fitting a model for the sake of performance, the idea is to check the following:

Hypothesis 4. *SNP-SNP interactions can be captured by non linear models to improve existing risk scores*

Risk scores are mostly build as if the gene-trait system was isolated. Exploring how genetics, environment interact with one another could provide a more complete picture of complex traits. To explore this path the following hypothesis can be checked for a specific trait (BMI):

Hypothesis 5. *Studying genetic and non-genetic variables jointly improves association*

A final approach is to try and model SNPs and biomarkers jointly with the environment to explore potential linkage between genetic information and biological processes. To achieve this, the following hypothesis is tested:

Hypothesis 6. *Complex interactions between SNPs, biomarkers and environment can be captured by generative models*

3. Methods

In this section, we present methods that are widely used in genomics. Each of these methods are developed in great details in the dedicated external resources available in the bibliography. As such, the goal of this section is not to provide an extensive theoretical presentation of each method. Instead the techniques are presented succinctly and illustrated by examples related to our genomics problem. As the variational model is less established in genomics and at the center of this thesis, it is developed in greater details. PCA, association testing and linear models sections provide the tool to understand existing risk scores and tackle hypotheses 1, 2, 3. Boosted models are most important for capturing non linearities solving hypothesis 4, 5 on gene-gene and gene-environment interaction. Finally, variational models are developed to create a comprehensive model of gene-environment-biomarkers and solve hypothesis 6.

3.1. Principal Component Analysis

Principal component analysis (PCA) is among the simplest and most commonly used type of exploratory method. In essence, it linearly projects vectors from a high dimensional space onto a lower dimensional one while conserving as much variance as possible. The explanations detailed below comes from the book Foundation in Data Science [13].

3.1.1. Presentation

In short, a matrix $X \in R^{n \times p}$ with centered columns, is decomposed into its singular values decomposition $X = U\Sigma V^T$. In this decomposition U and V are $n \times p$ and $p \times p$ orthogonal matrix. Σ is a diagonal matrix with the sorted singular values on the diagonal. The right singular vectors are called principal components.

We can then define $A_k = U\Sigma_k V^T$, where Σ_k only contains the top k singular values. This decomposition gives A_k, the best k-rank approximation of A.

Property. *For any matrix B of rank at most k, $\|A - A_k\|_F \leq \|A - B\|_F$.*

In practice it is convenient to study the projection of X onto the top k right singular vector space by computing $T = XV_k$. For instance, choosing $k = 2$ gives a convenient way to represent X in a two dimensional plan.

Principal component analysis can be equivalently seen as finding the eigenvector of the correlation matrix $\frac{1}{p}X^TX$. As such having centered columns is essential to interpret the SVD decomposition in terms of projection on subspace maximizing variance.

In genomics as in many other fields, PCA is an essential first step of the analysis pipeline. This methods capture the population structure of human genomes in an extremely relevant way [14]. Allele frequencies can vary greatly between persons from different ancestries. PCA gives a convenient quantitative way to capture this information. Principal components can then be used in models to control for this origin and avoid spurious correlations [15].

3.1.2. Illustration

To illustrate this fact, we apply the method from [14] at the world scale. We perform a PCA on 10^4 SNPs randomly sampled from chromosome 15 from the 1000 genome project. The principal components capture population origin very well. The first principal component is linked to origin by latitude, the second principal component by longitude. A map is overlaid on the PCA plot 3.2 to illustrate this fact. The relative position of population in the principal component space mirrors the geographical position of populations.

Figure 3.1.: PC 1 versus PC 2

While being a convenient way of measuring population stratification, PCA has some limitations that prevent more ambitious analysis. Since PCA is unsupervised in nature, the method always captures the properties responsible for a the most variance in the data. Population structure induces a large amount of variation in the genome that PCA captures very well. However, there is no guarantees that variations related to traits of interest will be captured by a given component. The variations might be extremely scarce. Instead, in order to advance further in our study, we need supervised methods that can directly estimate links between genetic information and traits of interest.

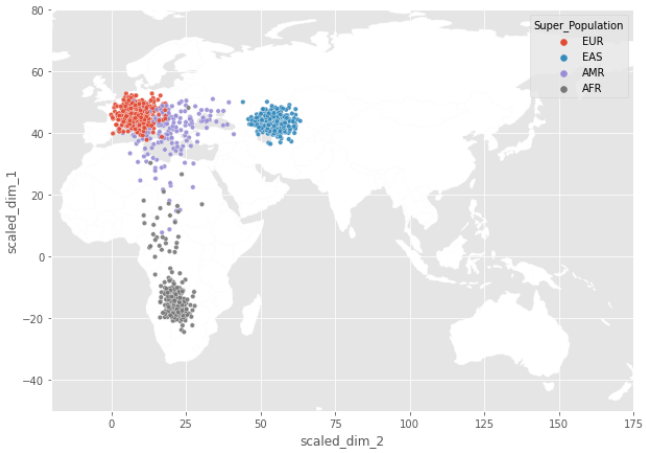

Figure 3.2.: PC 1 versus PC 2 overlaid on world map colored by population origin

3.2. Association Testing

As mentioned in the paragraph on GWAS, association testing is at the basis of genomics. In this paragraph, the methodology of computing association for a single SNP and trait pair is explained. The goal of the method is to compare quantitatively the allele frequencies between a control and a case group for a given loci in the genome. Based on the frequencies observed in the two groups, we wish to determine whether a SNP is associated to a trait. The usual framework used is hypothesis testing. The general topic of testing statistical hypothesis is explored in great details in this book [16]. This article [17] covers much of its application in genetics.

3.2.1. Presentation

We wish to determine whether to accept or reject an hypothesis H_0. If the null hypothesis is rejected, we accept an alternative hypothesis H_1. Assuming that H_0 is true, a quantity computed from the data, the test statistic, should follow a known distribution. If the observed test statistics is unlikely in this distribution we reject the null hypothesis. To quantify this, we compute the probability of obtaining the test statistics or a more extreme value under the null hypothesis' distribution. This is the p-value. The lower the p-value, the more improbable the observed data is under the null hypothesis. Below a predetermined probability called the significance level, we reject H_0 in favor of H_1.

In genetic association testing, the null hypothesis is that there is no difference in allele frequencies between the case and control group. For categorical phenotypes, the test statistic is the χ^2 Pearson's cumulative statistic which under the null hypothesis follows a χ^2 distribution of degree of freedom 1. Because of the high dimensionality of the problem, the significance level is on the order of 10^{-8}, much lower than the usual 5×10^{-2} level.

Hypothesis testing relies on the correct definition of the test statistics. For instance, the χ^2 statistics relies on the assumption that under the null hypothesis the frequencies between group fluctuates around an expected frequency following a Gaussian distribution. The test statistics is build by summing those fluctuations. Given observed frequencies O_i and excepted frequencies under null hypothesis E_i, the test statistic is:

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$$

Under the null hypothesis, this statistic is a sum of Gaussian which does follow a χ^2 distribution.

For continuous trait such as blood sugar levels, t-test on the slope of the univariate linear regression between the continuous phenotype and the SNP are used. While the test statistic is different, the rest of the method stays the same.

3.2.2. Illustration

Here is an illustrative example of the method of hypothesis testing. We wish to apply the method to determine whether a SNP is associated to a categorical trait or not. To simplify, we consider SNPs with two uniformly distributed alleles, reference (0) and effect (1). The trait is binary and distributed uniformly. Hence control and case groups have equal size. We simulate a cohort of $N = 10^3$ samples. The methods of sampling SNPs and trait is detailed in the introduction example on GWAS.

First, we sample SNPs and traits independently of one another. Looking at the contingency tables 3.3 below, we can see that in both cases, the distribution of reference and effect allele seems uniformly distributed across control and case group. Given the setup of the simulation, we would except to have roughly 250 sample in each case. Naturally, this visual inspection is not enough to determine whether the alleles are independent from the trait. The goal of hypothesis testing is to determine quantitatively whether the fluctuations observed can be considered random or not.

For the purpose of the example, we first check empirically that under the null hypothesis (SNPs and Trait are independent), the χ^2 test statistics does follow a χ^2 distribution. In order to do so, we sample 1000 independents SNPs and 1 trait that is not associated to the SNPs.

Assuming now that we have a causal and a non causal variant, we verify that the statistical test detects the causal one. Looking at the contingency table, we start to see that the frequencies of the effect allele are different in the case and control group for the causal variant.

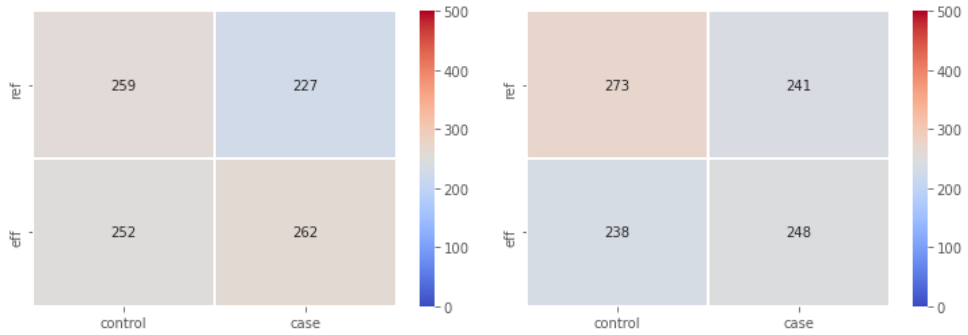

Figure 3.3.: Contingency tables of SNPs with no effect

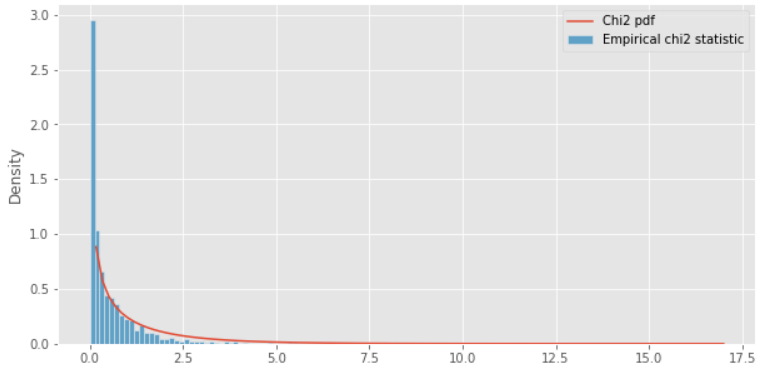

Figure 3.4.: Empirical test statistic histogram and χ^2 density

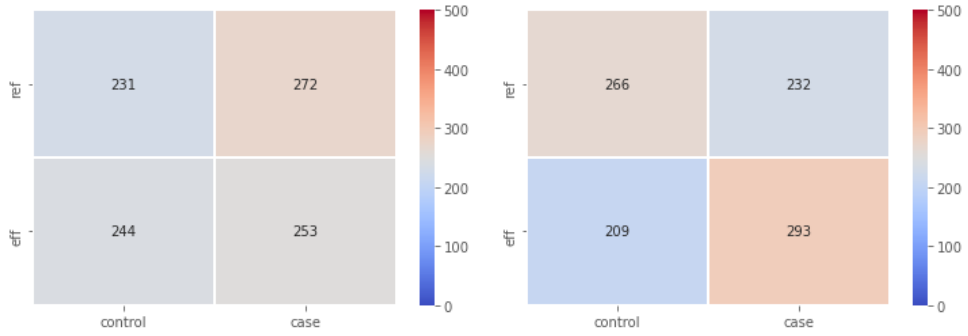

Figure 3.5.: Contingency tables of SNPs without (left) and with (right) effect

The χ^2 test confirms that this observation is very improbable under the null hypothesis for the causal variant.

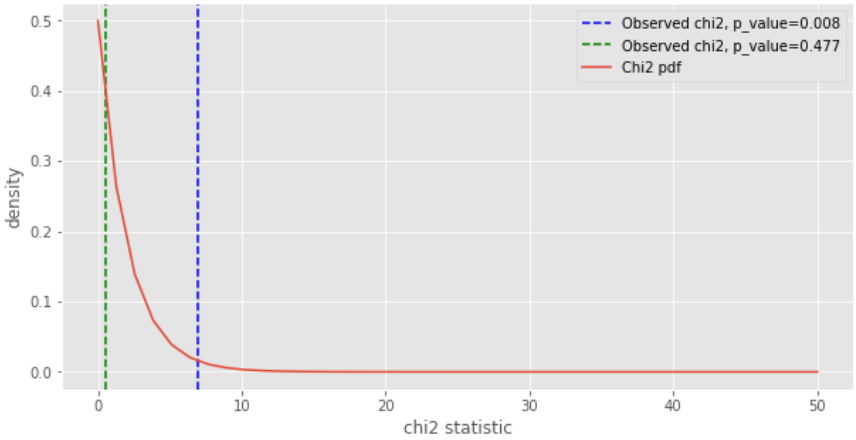

Figure 3.6.: Test statistic for effect (blue) and no effect (green) SNP and χ^2 density

While hypothesis testing approach is relatively natural and straightforward, it does have a number of limitation detailed in the introduction example. Nevertheless, its simplicity makes it a relevant tool in genetics.

3.3. Linear models

Linear models and regularization cover a wide range of fascinating subjects at the interface of statistics, optimization and algorithmic. Covering all these topics is beyond the scope of this thesis. In this section, we explain how assumptions on our genomic problem can guide us towards models with adequate mathematical properties. The content in this section follows the lecture [18] with an added practical illustrating our problem.

As illustrated in the previous chapter, genomic data is highly dimensional. Recall that genetic information is encoded into a matrix $X \in \{0,1,2\}^{N \times d}$ where $d >> N$. For the sake of this example we consider a continuous phenotype $y \in \mathcal{R}^N$. An example of such continuous phenotype could be body mass index. The considerations below can be generalized to categorical phenotypes using generalized linear models.

In modeling, it is generally useful to first fit simple models. Well implemented, understood and used models can serve as baselines before further analysis. Linear regression is an example of such a model. As a first assumption, it seems reasonable to consider that the effect of individual variants in the genome are additive. Each variant may impact a separate biological process that affects the phenotype. The basic way to build a linear model is through ordinary least squares (OLS) estimation. In our problem however the OLS falls short of expectations. Therefore, even to estimate an additive baseline model, more advanced regularization methods need to be introduced.

3.3.1. Presentation

We wish to model the phenotype as a linear function of the SNPs:

$$y = X\beta_0 + \epsilon$$

$\beta_0 \in \mathcal{R}^d$ is the vector of coefficients, and $\epsilon \in \mathcal{R}^d$ the error term.
The coefficients are obtained via least square minimization i.e:

$$\beta_0 = argmin_{\beta \in \mathcal{R}^d}(\|y - X\beta\|^2)$$

When $rank(X) = p$, the unique solution is $\beta_0 = (X^T X)^{-1} X^T y$. In our case however $rank(X) < p$ because $N << p$. Therefore the solution is no longer unique: any vector of the form $\beta_0 + \eta$ with $\eta \in ker(X)$ is also a solution. Reliable prediction and interpretation becomes impossible.

A fix to this troublesome problem is regularization. In order to find useful coefficients, we must add additional assumptions on them. Here, it is sound to look for sparse coefficients. Most of the SNPs in X involve biological processes that have nothing to do with the phenotype y. The contribution of those SNPs to the phenotype should be exactly zero. Mathematically,

this new assumption can be added to the optimization problem as a constraint on the coefficients:

$$\beta_0 = argmin_{\beta \in \mathcal{R}^d,}(\|y - X\beta\|^2 + \lambda\|\beta\|_1)$$

With λ a tuning parameter that controls the level of sparsity. This estimator is called LASSO. It selects a subset of SNPs that have the most effect on the phenotype.

Still, the LASSO has a problematic limitation coming from a theoretical property on its solution. For any $\lambda \geq 0$, the number of non zero coefficients is always less than $min(N, p)$. Concretely, given 10^3 samples and 10^6 SNPs, the LASSO would recover the effects of at most 10^3 SNPs. Given the relatively small size of cohorts and that phenotypes are most often considered highly polygenic this is a potentially serious pitfall.

Luckily, this problem can be easily overcome by tweaking the optimization problem once again. An additional constraint on the parameter is added via l_2 norm. While l_1 constraint induces sparse solution, l_2 constraint, also referred to as ridge, induces dense solutions. The mixture of both allows to fine tune the level of sparseness of our model. This model called elastic net is our baseline model.

$$\beta_0 = argmin_{\beta \in \mathcal{R}^d,}(\|y - X\beta\|^2 + \lambda_1\|\beta\|_1 + \lambda_2\|\beta\|_2)$$

Regularized linear models offer a wide range of advantages, consequently they are widely used in polygenic risk scoring. Being linear, they are highly interpretable and allows for further analysis of the selected SNPs. While the estimation of the coefficient is a topic on its own, many off the shelf implementations are available for quick experimentation and analysis.

The disadvantage of these models mainly comes from the underlying linear assumption. As discussed in the GWAS example, SNPs may interact with one another. In smaller problems, it is possible to study those interactions by expanding the covariates on polynomial bases. For instance, one could imagining creating a linear model using the SNPs as well as their product $(SNP_i \times SNP_j)_{1 \leq i,j \leq p}$. Naturally, this is out of question here as adding those first order interactions involves creating an new p^2 dimensional problem. This is simply not scalable.

3.3.2. Illustration

One could wonder what is the benefit of using regularized method such as LASSO over simple univariate feature selection using association tests followed by a separate OLS regression. The procedure for a continuous phenotype $y \in \mathcal{R}^N$ and a SNPs matrix $X \in \mathcal{R}^{N \times p}$ is the following:

1. Estimate the coefficient of $y = \beta_i x_i + \epsilon$ via OLS for each x_i

2. Test $H_0 : \beta_i = 0$ versus $H_1 : \beta_i \neq 0$ via t-test statistics for each x_i
3. Select the set of x_i with the desired significance level for β_i
4. Fit a multivariate OLS regression on the selected set of x_i

In order to demonstrate the difference between the approaches, we consider two cases. First, in a well behaved case, the columns of X are orthonormal and we have enough sample $N >> p$. In this case we could simply fit an OLS directly, but for the sake of the argument let's explore the behavior of the methods.

Here the solution to the LASSO problem is explicit. We have $\beta_i^{LASSO} = S_{N\lambda}(\beta_i^{OLS}) = \beta_i^{OLS} max(0, 1 - \frac{N\lambda}{|\beta_i^{OLS}|})$. With $\beta^{OLS} = (X^T X)^{-1} X^T y = X^T y$ by orthonormality. This solution gives the intuition on why the LASSO gives a sparse solution. The coefficients are shrunken by λ, the ones below the regularization threshold are set to zero. Selection by LASSO comes at the cost of adding bias in the estimated parameters.

The univariate selection is a threshold method as well. A co-variate passes the hypothesis test if $P(|T_i| > t_{alpha}) < p_{alpha}$. $T_i = \frac{r^2(N-2)}{1-r^2}$ is the test statistic following a T distribution. $r^2 = corr(x_i, y)$ is the correlation between x_i and y. Since the inverse of the T distribution cumulative distribution function is known, we can express the p-value threshold as a threshold on the t-score and hence on the correlation. Since the correlation is a function of β^{OLS}, we can express the threshold on the p-value as a threshold on $|\beta^{OLS}|$

The selected covariates are used to estimated the OLS model. Since the columns are orthogonal, $\beta_i^{t-test} = \beta_i^{OLS}$ for the β_i above significance level and 0 for the others.

The graph below displays those properties for different thresholds for both methods. Here X contains $N = 10^3$ samples and $p = 10^2$ independent covariates. The true model contains 10 non zero coefficients. In both graphs, we plot the index of the coefficient from 0 to 100 versus their values. The dots are colored by coefficient type. The true coefficient are red, the estimated coefficients are blue. The green area represents the value for which the coefficients are not filtered. The gray line represents the distance from the true to the estimated coefficients.

The LASSO dot plot illustrates how true coefficients above threshold (in the green area) are shrunk while the others are set to zero. The t-test/OLS plot shows that coefficient below threshold are set to zero while the other are equal to the true OLS coefficients.

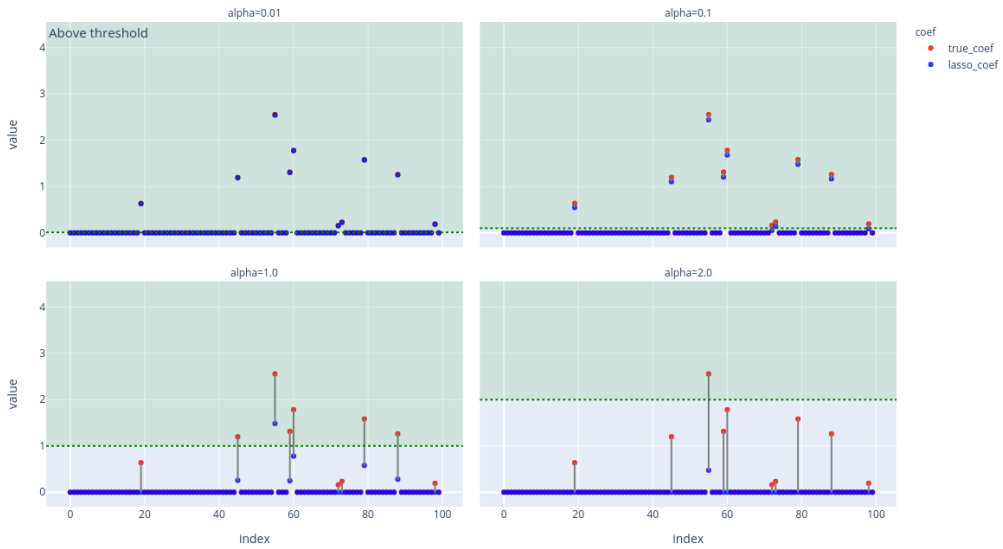

Figure 3.7.: Dot plot of LASSO coefficients with LASSO threshold, well behaved case

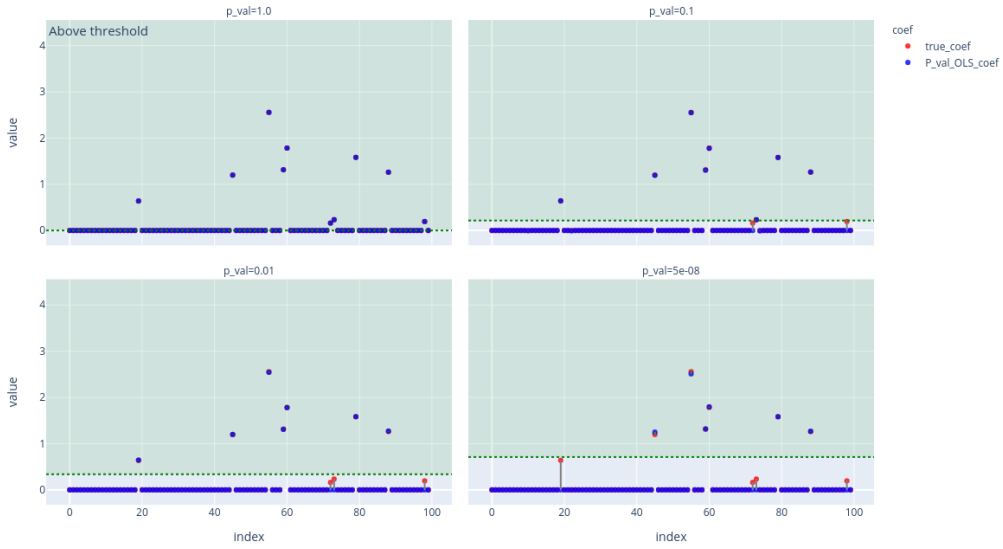

Figure 3.8.: Dot plot of t-test/OLS coefficients with p-value threshold, well behaved case

Naturally, this example is of little practical interest. In our high dimensional setting, the solutions are not explicit anymore but intuition can help understand why LASSO may be the better choice. While LASSO optimization is multivariate and take into account all the covariates at once, the t-test selection method is univariate. Doing univariate selection is risky because the covariance between co-variates is never considered, hence, we may select strongly correlated features that will put the OLS estimator in trouble.

We simulate a high dimensional problem with $N = 10^2$ samples and $p = 300$ features. There is still 10 non zero coefficients in the true model. This time however, the features are now longer orthogonal as we set $rank(X) = 10$. We no longer have an explicit solution to the LASSO.

On an high dimensional problem, the LASSO still adds some bias but robustly select the correct covariates at multiple regularization levels. The univariate selection method on the other hand not only adds bias but also misselect covariates. This problem appears even at the "99%" significance level of $p_{value} = 0.01$. In genetics, this method is sometime improved by selecting SNP based on significance as well on LD structure (covariance) in a process called "LD clumping". However this requires estimating the covariance between each pair of SNPs an often rely on arbitrary thresholds. In practice [19], regularized linear methods show increased performance compared to univariate selection for these reasons.

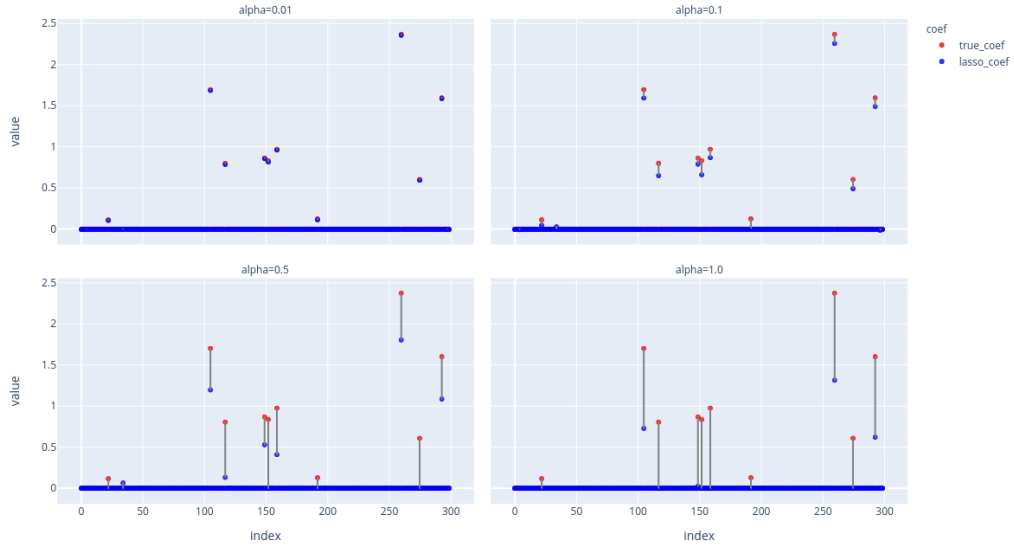

Figure 3.9.: Dot plot of LASSO coefficients, ill behaved case

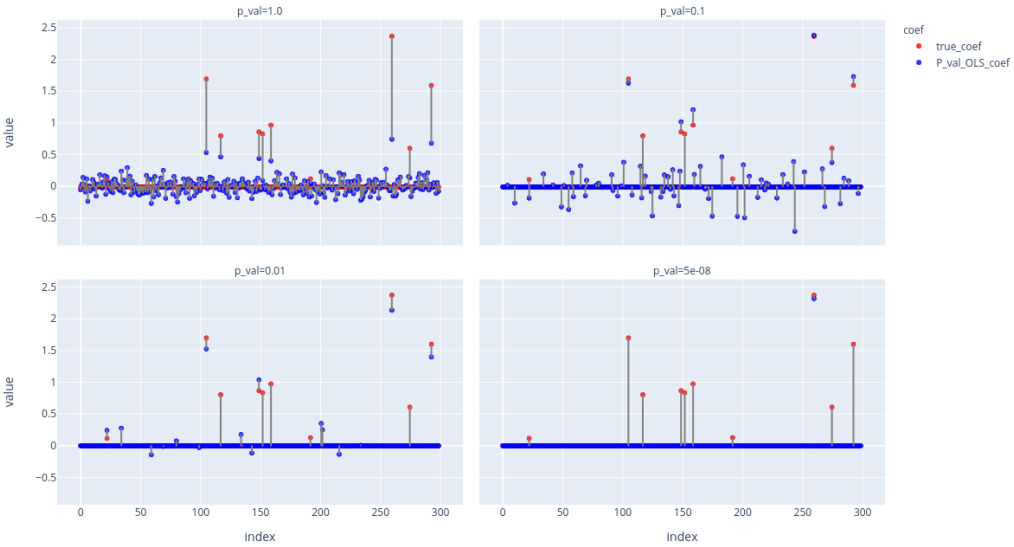

Figure 3.10.: Dot plot of t-test/OLS coefficients, ill behaved case

3.4. Boosting

To capture more complicated interactions in the data, linear models are not enough. Therefore, we must explore more advanced models. Gradient boosted models offer the possibility to create non linear models for regression or classification while being interpretable to a degree. Additionally, boosted trees are robust to outliers, handle mixed and missing data, and perform feature selection. Efficient implementations of these models are readily available. These advantages make gradient boosted models great off the shelf procedures for data mining.

While boosting is a general procedure that can use any kind of base learners, we focus here on boosted trees. Decision trees are great weak learners as they can capture non linearities. In this section we give an overview of the concept of boosting with adaptive boosting and gradient boosting following this book [20]. Finally, a small illustrative example related to genomics is given.

3.4.1. Presentation

The principle of boosting is to combine the prediction of many "weak" models to produce a reliable result. Decision trees are often used as weak models. The standard boosting algorithm is Ada-boost. We consider a binary phenotype $Y \in \{-1, 1\}$ and the SNP matrix X. Boosting methods model the output Y as a weighted sum of weak models $G(X) = sign(\sum_m \alpha_m G_m(X))$. The essential idea of boosting is to weight α_m and train each classifier more cleverly than by a bootstrap aggregation.

To do so, the weak learners G_m are fitted sequentially on modified versions of the data. The data modification is central to the algorithm, initially, all points are weighted equally. When fitting the classifier G_i, the observation misclassified by G_{i-1} have higher weights. The algorithm is as follow:

Algorithm 1 Adaptive Boosting

procedure FitAdaBoost(X, y)
  Initialize observation weights $\{w_i\}$ uniformly
  for $m = 1$ to M **do**
    Fit $G_m(x)$ on y
    Compute $err_m = \frac{1}{\sum_i w_i} \sum_i w_i I(y \neq G_m(x_i))$
    Compute $\alpha_m = log(\frac{1-err_m}{err_m})$
    Set $w_i \leftarrow w_i e^{\alpha_m I(y_i \neq G_m(x_i))}$
  end for
  Return $G(X) = sign(\sum_m \alpha_m G_m(X))$
end procedure

While boosting may look like a very empirical method to improve model performance, its success can be justified mathematically. AdaBoost is part of a wider class of models called, forward stage wise additive models. The data is modeled as an additive expansion over a set of basis functions that are sequentially added. In this study, we use Gradient Boosting instead of Adaptive boosting. While they both belong to the same class of models, gradient boosting is a more general approach to estimate such models, it is also well implemented.

The gradient boosting tree models the data as a sum of trees:

$$f_M(x) = \sum_{m=1}^{M} T(x, \Theta_m)$$

Just as in AdaBoost, the model is built sequentially, each boosting step finds $\Theta_m = \{R_{j,m}, \gamma_{j,m}\}$ such that:

$$\Theta_m = argmin_{\Theta_m} \sum_{m=1}^{M} L(y_i, f_{m-1}(x_i) + T(x_i, \Theta_m)$$

Where L is a loss function relevant to the task at hand. As the name explains, gradient boost relies on the gradient of the error to fit successive models. The algorithm goes as follow:

Algorithm 2 Gradient Boosting

procedure FitGradientBoost(X, y)
  Initialize $f_0(x) = argmin_\gamma \sum_{i=1}^{N} L(y_i, \gamma)$ ▷ Initialize a constant model
  for $m = 1$ to M **do**
    for $i = 1$ to N **do**
      Compute $r_{i,m} = -[\partial_{f(x_i)} L(y_i, f(x_i))]_{f=f_{m-1}}$
    end for
    Fit a regression tree h_m to the targets $\{r_{i,m}\}$
    Compute $\gamma_m = argmin_\gamma \sum_{i=1}^{n} L(y_i, f_{m-1}(x_i) + \gamma h_m(x_i))$
    Set $f_m(x) = f_{m-1}(x) + \gamma_m h_m(x)$
  end for
  Return $f_M(X)$
end procedure

While individual decision trees are fully transparent, interpreting their aggregation is not straightforward. In order to understand the impact of each variable on the prediction, the concept of importance is introduced. For a single tree, the importance of a covariate X_l is given by

$$I_l^2(T) = \sum_{t=1}^{J-1} i_t^2 I(v(t) = l)$$

For each node t of tree T, a variable $X_{v(t)}$ is used for splitting. Each of these splits improves the losses by i_t^2. In essence, the importance takes into account how often the variable is used for splitting and how much it improves the prediction. To compute the importance of X_l for the whole boosted model, the importances of each tree are averaged.

$$I_l^2 = \frac{1}{M} \sum_{m=1}^{M} I_l^2(T_m)$$

While this measure is extremely helpful in determining the importance of each feature in the model, it does not gives a precise picture of how the features interact with one another. It also does not give the effect of X_l on y explicitly. As such, to get a better understanding of the role of each variables, we must go further and study partial dependence plots.

3.4.2. Illustration

In genomics, the use of gradient boosted models is mainly motivated by performance. The promise of gradient boosted models is to offer an "off the shelf" method for fitting complex models. Before rushing into applying Gradient boosting on a real data set, let's make sure that the model fulfills its promise on a simple example. In order to be of any use, boosted models must solve the problem encountered in the introduction on GWAS. We consider again a cohort of $N = 10^3$ and $p = 100$ SNPs. As before SNPs have two variations: reference (0) and effective (1) sampled uniformly. A binary trait is generated by a non linear interaction of two SNPs. A person with SNP_0 XOR $SNP_1 = 1$ will have an increased risk of having the trait. We then fit the data to two models: a non linear boosted model and a linear logistic regression. Once fitted, we look at the importance of each SNP given by each model. For the logistic regression the importance is defined as the regression coefficient. For the boosted model, the importance is defined as explained earlier.

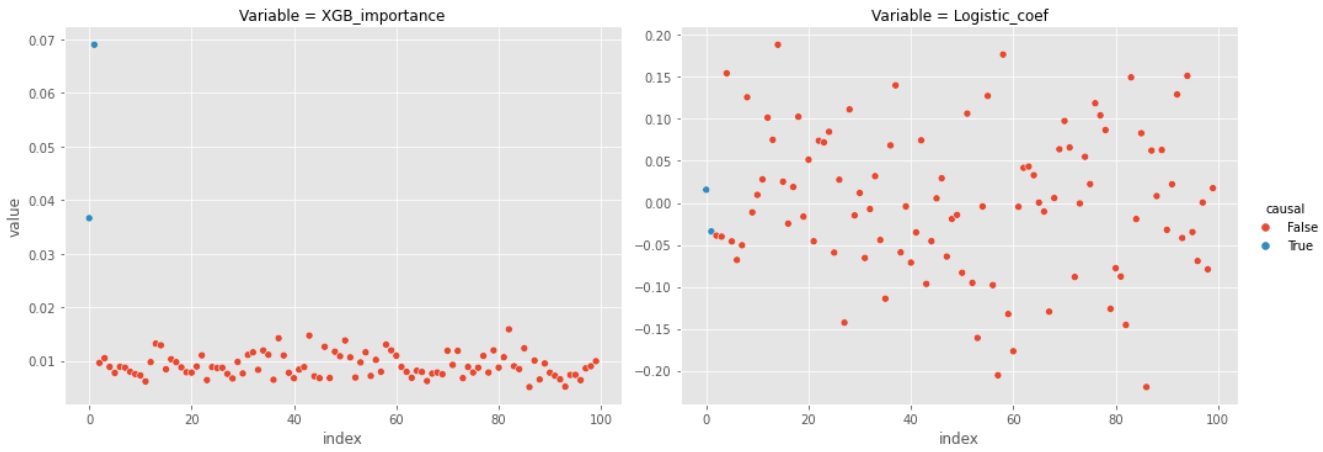

Figure 3.11.: Variable importance, Gradient Boosted model and Logistic regression

On the figure 3.11, we plot the importance of each SNP according to the boosted and model (left) and the logistic regression (right). The causal SNPs are colored in blue, the others are red. This plot shows that the boosted model is able to fit the data in a way that recovers the true causal SNPs while the logistic regression is not. The boosted model acts as promised and should be helpful in detecting SNP interaction in polygenic risk scoring.

3.5. Variational Inference

While boosted models are a proven and powerful method, they still lack some potentially useful properties. The notion of genetic risk is inherently probabilistic. Much of the biological processes are hidden and most of environmental factors are not accessible. Probabilistic graphical models can represent this random nature in a convenient way. To build complex and non linear probabilistic models, we rely on the fields of variational inference and deep learning. Since variational models are at the heart of this work and are less common in this field than the previous methods, we will present and illustrate them in greater detail. Our approach follows a slightly simplified version of the problem solved in the article [21]. Derivations and general considerations on variational inference comes from the lecture notes [22].

3.5.1. Presentation

Setup Mathematically, the goal of the study is to understand the **conditional dependencies** between three variables X, Y and Z where X and Y are observed and Z is unobserved. In our case, X is the **genotype**, Y is the **phenotype**, Z is the **latent representation** of the genotype. Two probabilities are of interest:

- $p(Z, Y|X)$ which allows to embed our X into a latent space and a phenotype space.
- $p(X|Y)$ which allows to sample X given Y.

The graphical model is defined as follow:

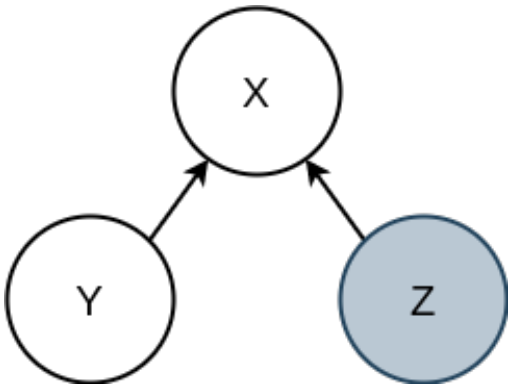

Figure 3.12.: Probabilistic graphical model

With joint probability :

$$p(X, Y, Z) = p(X|Y, Z) \times p(Y) \times p(Z)$$

In the model Z and Y are independent. Note that this is not to be confused with a causal model. Y does not cause X, the phenotype has no impact on the genotype in this model.

However, estimating $p(X|Y)$ allows to sample from a population with a given phenotype. The challenge of this approach is that we do not have access to Z, it is a hidden variable. Therefore to use this model, we must first estimate $P(Z|X,Y)$. Naturally, we wish to capture intricate relations between variables, therefore, $P(Z|X,Y)$ may be arbitrarily complex and may not have an analytical expression. To solve this, variational inference is used.

Variational inference Variational inference [22] can approximate the intractable distribution $P(Z|X,Y)$ by a more simple variational distribution $Q(Z|X,Y)$. To start, we must link the likelihood of the observed variable X and Y to the variational distribution $Q(Z|X,Y)$:

$$\begin{aligned} logp(X,Y) &= log \int p(X,Y,z)dz \\ &= log \int p(X,Y,z)\frac{q(z|X,Y)}{q(z|X,Y)}dz \\ &= logE_q\left[\frac{p(X,Y,z)}{q(z|X,Y)}\right] \end{aligned} \tag{3.1}$$

While $logE_q\left[\frac{p(X,Y,z)}{q(z|X,Y)}\right])$ can not be expressed easily, $E_q\left[log\frac{p(X,Y,z)}{q(z|X,Y)}\right]$ can. Hence, Jensen inequality with $x \mapsto -log(x)$ is applied get the following inequation:

$$logp(X,Y) \geq E_q\left[log\frac{p(X,Y,z)}{q(z|X,Y)}\right] = ELBO \tag{3.2}$$

$ELBO$ stands for Evidence Lower Bound because it bounds the evidence (also called log likelihood). To understand the relevance of this bound, we must first rearrange it. Using the property of the model $p(X,Y,Z) = p(X|Y,Z) \times p(Y) \times p(Z)$, and $Q(Z|X,Y) = Q(Z|X)$, we get:

$$\begin{aligned} ELBO &= E_q\left[logp(X|Y,z) - log\frac{q(z|X)}{p(z)}\right] + logp(Y) \\ &= E_q[logp(X|Y,z)] - E_q[log\frac{q(z|X)}{p(z)}] + logp(Y) \end{aligned} \tag{3.3}$$

The bound is maximized to estimate the variational distribution as well as the true posterior. A way to see this is to derive the ELBO starting from the KL divergence between the variational and true distribution. Using $p(z|X,Y) = \frac{p(z,X,Y)}{p(X,Y)}$, we get:

$$\begin{aligned} D_{KL}(q(Z|X,Y)||p(Z|X,Y))) &= E_q\left[log\frac{q(z|X,Y)}{p(z|X,Y)}\right] \\ &= E_q[logq(z|X,Y)] - E_q[logp(z|X,Y)] \\ &= E_q[logq(z|X)] - E_q[logp(z,X,Y)] + logp(X,Y) \\ &= -ELBO + logp(X,Y) \end{aligned} \tag{3.4}$$

Since, $logp(X, Y)$ is independent of the variational distribution, maximizing the ELBO is equivalent to minimizing the KL divergence, i.e approximating $P(Z|X, Y)$ by $Q(Z|X, Y)$. Recall that the KL divergence gives information on how similar two distributions are. $D_{KL}(P|Q) = 0$ if and only if, $P = Q$ almost everywhere.

Likelihood maximization The variational model does not allow to estimate $p(Y|X)$ easily. In order to access $p(Y|X)$ we could estimate $p(X, Y)$ and $p(X)$ via Monte Carlo methods using sampling on the latent variable Z. However, this may require a large number of samples and be highly variable. Instead, we can directly estimate $p(Y|X)$ via a likelihood maximization approach jointly with the variational model. The model $p(Y|X)$ can be any kind of classification or regression model.

Objective function We combine the variational inference objective with the log likelihood maximization objective to get the objective function of the problem. This was first derived in [21]:

$$\mathcal{L}(\theta, \phi, \beta|X, Y), \beta = ELBO(X, Y, \theta, \phi) + \gamma logL(Y|X, \beta) \tag{3.5}$$

With γ a tuning parameter between the two objectives. θ, ϕ and β represents the true model parameters, the variational distribution parameters and the supervised model parameters respectively.

Remarks In the original article [21], this loss function is improved to take into account missing values in Y. For simplicity we assume that the phenotype Y has no missing values. Additionally, the model above is improved by adding a conditioning variable C on X. This variable C can be used to add information to the model. For instance, C could be age, gender or environmental factors. As C is not central to the derivation of the equations it has been omitted above for clarity. It can be added back to the model by replacing X by $X|C$. This conditioning is detailed in article [23].

Optimization preliminary The next challenge is to minimize the loss function given a model. Neural networks are a class of models that have shown success on many tasks. Since this thesis focuses on the variational model itself rather than the neural network, background on neural networks is omitted but can be found in this book [20]. To understand the following paragraph, it suffices to know that neural networks are parametric differentiable functions whose parameters are estimated by gradient descent methods. Fundamentally, to fit any neural network model with gradient methods, one must have a differentiable objective function. Finding an efficient differentiable estimator of the $ELBO$ is far from obvious. In the general case, we wish to take the gradient of an evidence lower bound of the form:

$$\begin{aligned} ELBO &= E_{q_\phi}[logp_\theta(X|z)] - E_{q_\phi}[log\frac{q_\phi(z|X)}{p_\theta(z)}] \\ &= E_{q_\phi}[f(z)] - E_{q_\phi}[log\frac{q_\phi(z|X)}{p_\theta(z)}] \end{aligned} \tag{3.6}$$

Looking at the gradient of the first term of the equation, we get:

$$\begin{aligned}\nabla_\phi E_{q_\phi}[f(z)] &= \nabla_\phi \int f(z) q_\phi(z) dz \\ &= \int f(z) \nabla_\phi q_\phi(z) \\ &= \int f(z) q_\phi(z) \nabla_\phi ln q_\phi(z) \\ &= E_{q_\phi}[f(z) \nabla_\phi ln q_\phi(z)]\end{aligned} \tag{3.7}$$

This derivation from article [24] assumes that we can swap the integral and derivative and uses the relation $\nabla_\phi q_\phi(z) = q_\phi(z) \nabla_\phi ln q_\phi(z)$. Unfortunately, the literature [25] shows that using this empirical expectation as an estimator does not give good results in practice. The reason is that we need to sample a large number of points to get a good estimate.

Optimization solution Article [25] introduces a simple and elegant solution to get a better estimator, the reparametrization trick. It gives a differentiable estimator that does not require direct sampling over Z. For certain distribution it is possible to rewrite $z \sim q_\phi(Z|X)$ as:

$$z = g_\phi(X, \epsilon) \tag{3.8}$$

With g_ϕ a differentiable function and ϵ a random variable following a fixed distribution $p(\epsilon)$.

The problematic expectation now becomes:

$$E_{q_\phi}[f(z)] = E_{q_{p(\epsilon)}}[g_\phi(x, \epsilon)] \tag{3.9}$$

The expectation is now taken over a random variable that is independent of the estimated parameters. The final estimator of the ELBO is now:

$$\widehat{ELBO} = \frac{1}{L} \sum_{i=1}^{L} log p_\theta(x_i, z_i) - log q_\phi(z_i|x_i) \tag{3.10}$$

With $z_i = g_\phi(x_i, \epsilon_i)$ and $\epsilon_i \sim p(\epsilon)$. Since this estimator is differentiable it can be used to estimate neural network parameters via gradient descent methods. The paper [25] introducing this estimator is the foundation of the development of deep variational auto encoder.

Reparametrization While reparametrization can be done for a number of different distributions, we restrict ourselves to the most common case. A gaussian random variable $z \sim N(\mu_x, \sigma_x)$ can be written as:

$$z = \mu_x + \sigma_x \epsilon \tag{3.11}$$

With $\epsilon \sim N(0, 1)$. μ_x and σ_x are differentiable functions parametrized by ϕ (for instance neural networks). Hence, the reparametrization is differentiable.

Full specification example Here is a possible model specification for modeling SNPs and a categorical phenotype. As we use neural networks to model our data, the deep learning vocabulary is applied to define our model. $p_\theta(X|Y,z)$ is referred to as the decoder, $q_\phi(z,y|X)$ as the encoder/classifier. The models are parameterized by θ, ϕ for the decoder and encoder/classifier respectively.

Recall from the introduction that $X \in \{0,1,2\}^{N\times d}$. The following output distribution is choosen:

- $p(x_i = j|y,z) = Categorical_{K_p}(p_{ij})$

Where, $K_p = 3$ and $(p_{ij})_{1\leq i\leq d, j\in\{0,1,2\}} = decoder(z,y)$. *decoder* is a neural network with a multidimensional softmax output layer.

As a side note, in case X includes a subset of continuous variables $X_{continuous}$, the distribution for those would be:

- $p(x_{continuous}|y,z) = \mathcal{N}(\mu_{z,y}, \sigma_{z,y}\mathbb{1})$

Where $\mu_{z,y}$ and $\sigma_{z,y}$ are neural network with linear output layers.

To get a simple latent space, the following distributions are chosen:

- $q(z|x) = \mathcal{N}(\mu_x, \sigma_x\mathbb{1})$
- $p(z) = \mathcal{N}(0,\mathbb{1})$

Where $\mu_x, \sigma_x = encoder(x)$. Here *encoder* is a neural network with a linear output layer.

Because the phenotype is categorical, we choose:

- $q(y|x) = Categorical_{K_q}(p_x)$
- $p(y) = Categorical_{K_q}(\frac{1}{K_q})$

K_q is the number of possible classes for y and $p_x = classifier(x)$. Here $classifier$ is a neural network with a softmax output layer. Note that $p(y)$ is independent from the encoder and decoder and can therefore be omitted from the final loss.

As explained above, the *ELBO* estimator is defined as:

$$ELBO = \sum_{i=1}^{N}(logp_\theta(x_i|y_i,z) - D_{KL}(q_\phi(z|x_i)||p(z))) \tag{3.12}$$

On the other hand, since Y is categorical, the log likelihood loss can be defined as the negative cross entropy:

$$logL(Y,X,\beta) = -CE(Y,\hat{Y}) \tag{3.13}$$

Summing the terms, we get:

$$\mathcal{L} = -\sum_{i=1}^{N}(logp_\theta(x_i|y_i,z) - D_{KL}(q_\phi(z|x_i)||p(z)) - y_i log(q_\phi(y_i|x_i))) \tag{3.14}$$

Practical considerations While the loss function is justified mathematically, in practice, getting good results can be challenging. In essence, we combine three terms, a reconstruction task, a regularization via KL divergence, and a classification task. The classical VAE can fail at giving informative embeddings when the KL divergence and reconstruction loss are not correctly balanced. This limitation is called the "posterior collapse". Sampling any Z in the latent space produces the same X in the original space. A solution to this issue is to weight the terms of the loss empirically [26]. These weights can be kept constant or can be modified as training goes on. Naturally, this comes at the cost of adding new hyper-parameters to the model. Additionally, just as for linear models, it is possible to regularize the neural network by adding constraints on the parameters.

Loss function example Combining all the consideration above, we get the following loss function for our example. It is expressed as the weighted sum of four terms. A reconstruction objective, a classification objective, a regularization on the latent space, and a regularization on the parameters.

$$\mathcal{L} = \sum_{i=1}^{N} (-\alpha \times log p_\theta(x_i|y_i, z) + \beta \times D_{KL}(q_\phi(z|x_i)||p(z)) + \gamma \times y_i log(q_\phi(y_i|x_i))) + \lambda \mathcal{R}(\theta, \phi, \beta)$$

3.5.2. Implementation

Deriving the loss function is only half the work. To get results, the model must be carefully implemented. The encoder and decoder can be any kind of neural network. In this study, the data is neither spatial nor temporal, therefore densely connected layers are chosen. To improve optimization, batch normalization [27] layers are added between dense layers.

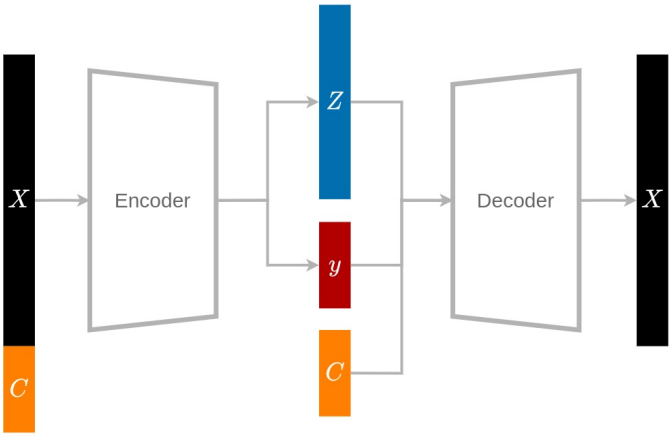

Figure 3.13.: Variational model architecture

In this figure, X is the SNP data, C is the conditional data (environmental factors), y is the phenotype and Z is the latent representation of X. Using the notations from above, the encoder is $q_\phi(z|X,C)$ and $q_\phi(y|X,C)$. The decoder is $p_\theta(X|y,Z,C)$.

The PyTorch [28] library is used to implement the model and solve the optimization problem. This library allows for auto differentiating functions defined as computational graphs. The parameters are estimated via the ADAM optimizer [29]. Additionally, a set of visualizations helps validation and debugging of the model using Tensorboard [30]. These visualizations are explained in the model validation section.

3.5.3. Applications

The purpose of implementing the rather complex model above is to have an investigable and flexible method. The essential property of the model is that it is generative. X can be sampled given C and Y and be investigated. Because of the supervised task, we can precisely assess how much information in X and C can be extracted to determine y. With usual C-VAE this is not possible.

The latent space The objective of this model is motivated by the creation of a disentangled and interpretable latent space. To control the generative process, the latent space must be controlled. Classical auto encoders may capture insightful embeddings, however there is no guarantee that we can either sample from them or interpret them. This model solves those issues.

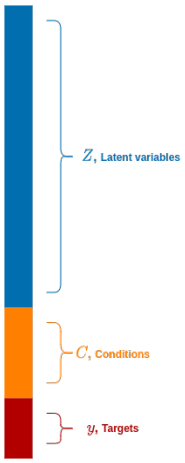

Figure 3.14.: Factorized latent space given by the model

The model precisely defines the latent space. Assuming that the model was properly fitted, the embeddings have the following properties. Z is a spherical gaussian, hence, its components are independent from one another. Each Z_i captures a different property of X. Theses properties can easily be observed by sampling along a given Z dimension. The

graphical model defining the relation between variable ensures that Z and Y are independent. Therefore it should be possible to perform "all other things equal" type of analysis. Holding Z fixed and modifying Y should give relevant information about the link between X and Y only. Finally, any confounding factors can be added via C to avoid polluting Z with irrelevant information. For instance, population structure to C can be added to remove potential biases. As a whole, this structure removes the need for any additional analysis on the latent space. It is factorized and interpretable by construction.

Interpretability Disentangled latent space is a fundamental property that allows for an interpretable model. Since the latent space is controlled, samples generated from the model are controlled as well. Provided that samples are coherent with the original data, we can study properties of the data by sampling from different places in the latent space. These samples can recreate synthetic counterfactual experiments.

Synthetic counterfactual experiments Counterfactual analysis[1] is at the heart of causal inference. It answers the question "What if ?". Using the model, the following experiment is performed. First, given a z, we sample X_0 from $P(X|Y=0,z)$ and X_1 from $P(X|Y=1,z)$. Since Y and Z are independent and z is the same in both experiments the only differences between X_0 and X_1 are linked to Y. This answers the question "What would X_0 look like if it was classified as $y=1$ instead of $y=0$?". The difference $X_1 - X_0$ allows to recover what caused the model to classify X as Y. Repeating the process for a great number of z and averaging allows to recover the most influential components of X for the classification of y.

$$AE = \frac{1}{N}\sum_i X_{1,i} - X_{0,i} \tag{3.15}$$

The figure 3.15 displays why the counterfactual experiment is made possible by the latent space. The architecture on figure 3.16 is inspired by a method [31] for assessing causal effect using variational auto encoders.

3.5.4. Illustrations

Synthetic problem In order to illustrate the previous section, we simulated a synthetic problem using the MNIST dataset [32]. Rather than predicting the class of the digit, we modify some digits to emulate a "phenotype". The modification is as follows: a circle is added somewhere on the surface of the digit. The circle can be at different places depending on the digit's shape.

We perform this modification on a random subset of the dataset. Once a digit is modified, the original is removed from the dataset.

The goal is two folds:

- Identify if a number has been modified or not.

[1] Note that here the counterfactual analysis is not done on the process itself but on the model, we are not looking at what caused phenotype y but rather at what caused the model to classify X as y

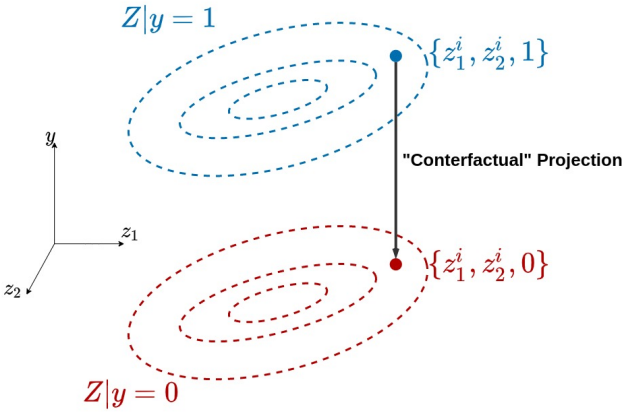

Figure 3.15.: Disentangled latent space with counterfactual projection

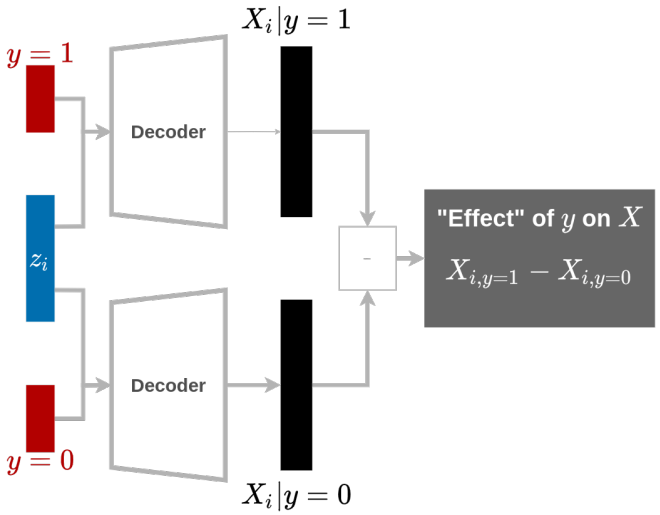

Figure 3.16.: Counterfactual architecture implementation

- Identify what the modification looks like on each class of digit.

Here are examples of synthetic modifications on digits. Note that the circle is positioned at different places depending on the class of the digit.

Figure 3.17.: Example of Modification

Below are samples from the modified dataset. On the top row modified digits ($Y = 1$) are displayed. On the bottom original samples ($Y = 0$) are shown. Notice that the position of the circle changes slightly for each digit.

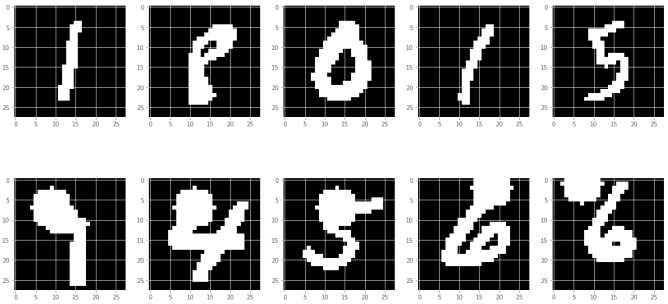

Figure 3.18.: Modified dataset samples

For a given digit, we need to identify whether it was modified or not. Then, we need to identify where the modification occurred. For each number we have access to its class (whether it is a one, a five etc...). Using the previous section notation. X is the image, C is the type of digit and y represents presence of the modification.

After fitting, the semi-supervised variational model can classify each pictures in class $y = 0$ or $y = 1$ and generate samples. By sampling digits given $y = 0$ or $y = 1$, we can perform the artificial "counterfactual" experiment displayed below. On the figure, the model acts as expected, the fives sampled from $P(X|y = 1, z)$ and $P(X|y = 0, z)$ are similar everywhere except where the modification occurred. Naturally, there is some noise the model that needs to be dealt with. A simple way to get a better understanding of the "phenotype" y is to sample a great number of difference and average them. In the figure below, we show the average difference between phenotype 1 and 0 for digit six and zero.

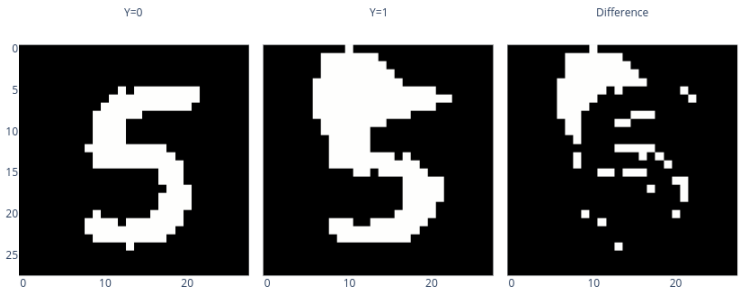

Figure 3.19.: Counterfactual experiment

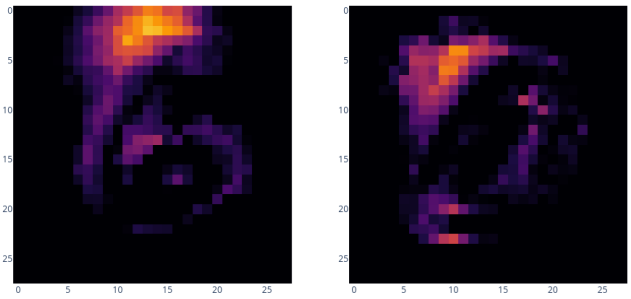

Figure 3.20.: Average counterfactual differences for six (left) and zero (right)

Why is this relevant ? We tested our implementation of the model by fitting it on the MINST dataset. We then created a synthetic problem to emulate what we wish to do on the real genetic data. In this visual example it is easy to identify the link between X, Y. Nevertheless, the model can both predict the class of X and recover the modification for each digit. In this synthetic problem we solved the following task:

> For **a given digit**, we can identify **whether the digit is corrupted or not**. Then, we can identify **which pixels are corrupted**.

On genomics data this approach could solve the following problem:

> For **a given person**, we are able to to identify **whether a disease is likely or not**. Then, we can identify **which SNPs are link to the disease**.

Bottlenecks Many challenges are still ahead however. The largest one being that the genetic data is of high dimension leading to computational challenges. There are also bottle necks on the approach. If the classifier performs poorly the approach is not valid as it means the model found no link between phenotype and genotype. If the reconstruction is too noisy, the variants linked with the phenotypes may be impossible to recover. Since the SNP dataset is not visual by nature plotting the result will not allow for quick identification as it was the case here. Finally, from a technical standpoint the hyper parameter tuning have been ignored until now. As the model combines different approaches, the success of the model relies on

an accurate tuning of the many parameters (loss terms weights, regularization and training parameters). Nevertheless this first step allows us to approach the next step of solving our problem with some confidence in both the model and the approach chosen.

3.6. Model validation

Model validation is the most critical step of genetic risk scoring. Models may follow a sound approach, use plenty of regularization and still provide disappointing results. As everywhere else in machine learning, methods must be cross validated rigorously from the SNP selection to prediction. While cross validation within a cohort provides hints about a model's performance, testing on independent datasets is critical. Biases from genetic, environmental and methodological factors can only be taken into account through meticulous testing on cohorts from different population and studies.

The modeling process involves continuous and categorical data types, classifications and regressions, supervised and unsupervised task. As such, a number of visualizations and metrics must be used to assess the model's performances. In this section, the validation workflow developed for the variationnal model is detailed.

Regression The go-to metric for validating a regression is the R^2 coefficient defined as $1 - \frac{SS_{res}}{SS_{tot}}$. Where $SS_{res} = \sum_i (y_i - \hat{y}_i)^2$ and $SS_{tot} = \sum_i (y_i - \bar{y}_i)^2$. In essence, the R^2 quantifies how a models compare to a baseline mean model. If the model's prediction are exactly equal to the fitted data, $R^2 = 1$. Since an over fitted model can perform worse than modeling the data by the mean, the coefficient can become negative. Using a single metric for validation is convenient but often not informative enough. Metrics can be accompanied by appropriate visualizations. For instance, a scatter plot representing fitted versus predicted values along with their respective distribution plot can highlight how the model behaves.

Classification The variational model can perform classification both for the supervised and reconstruction task. The classification task involve multiple and imbalanced classes. SNPs are ternary and imbalanced, BMI can be divided into six imbalanced classes. In this context, accuracy is a poor metric. Instead, it is more suitable to directly observe the relation between the fitted and the predicted values using confusion matrices. Heat maps can represent these matrices concisely.

As explained earlier, in the SNPs matrix most elements are zeros. Hence, a model predicting zero all the time would already have an accuracy close to one. Using the confusion matrix, it is possible observe the behavior of the model for the different alleles. In the illustrative example 3.21, the reference alleles are well reconstructed but the variations less so. Since the variational model relies on individual SNPs reconstructions for interpretability, confusion matrix can also be checked SNP wise.

Generation The generative process can be validated by comparing the properties of the sampled and fitted data. For instance, in the case of a generative model of SNPs, the allele frequencies and the linkage structure must be similar to the one from the original data. Depending on the application, privacy metrics may be needed to ensure that the samples are different enough from the data to avoid. In the case of genetic where most dataset are

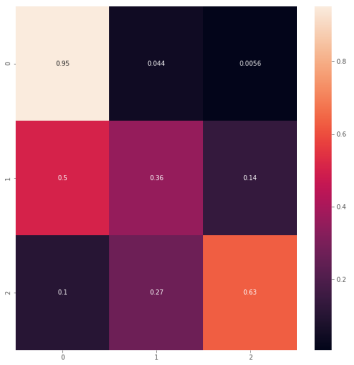

Figure 3.21.: SNPs reconstruction confusion matrix

protected, publishing a generative model that simply recreates the training set may cause worrisome privacy leakage.

Regularization As detailed before, regularizing involves adding assumptions to the model. To make sure that the fitting process went accordingly, theses assumptions should be checked. In the case of LASSO, one should check that the coefficient are sparse. In the case of the gaussian variational auto encoder, the latent variables should approximately follow a spherical gaussian distribution. To check these assumptions on the latent space, the covariance matrix is displayed as a heat map, the dimensions distribution are shown using violin plots.

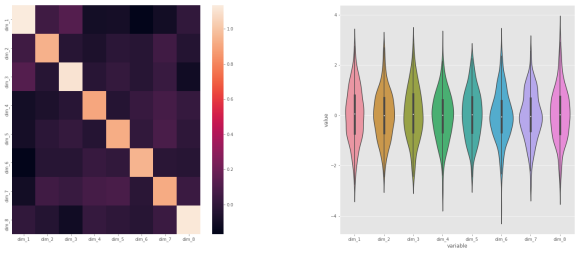

Figure 3.22.: Visualization of modeled embeddings

Implementation Because the variational model can involve all these topics, a convenient dashboard regrouping all the visualizations mentioned above was developed using Tensor Board[30]. This library allows to directly log losses, metrics, and visualization while training the model.

4. Results

Using the methods developed until now and in light of the challenges raised earlier in the motivation section, this chapter details results of experiments on risk scoring for BMI. BMI is a complex trait affected by a mix of genetic and environmental factors. As such, it is a great candidate for illustrating the different points made until now. More over, this trait is accurately measured in all cohorts, making it readily available.

First, existing peer reviewed risk scores are analyzed to understand their interpretability and applicability (hypotheses 1 to 3). Then boosted models are build to mine for both gene-gene and gene-environment interactions as stated in hypotheses 4, 5. Finally, experiments on variational models concludes on the feasibility of a complete gene-environment-bio marker model (hypothesis 6).

4.1. Existing scores analysis

After an in depth analysis of existing risk scores, this section explores the first three hypothesis 1, 2, 3 on some of their properties.

4.1.1. Existing risk scores presentation

Because of the sheer number of SNPs and the relatively small effect between genetic information and traits such as BMI, jumping into fitting models to the data proves to be perilous. To avoid this complication, it is critical to first study and explore existing scores to get a feeling of what to expect. The polygenic risk score catalog [11] lists weights for BMI risk score models from 5 different articles. The authors use different strategies to estimate the model's weights. Smaller ones use a p-value threshold followed by a linear regression. PRS 27 and 320 use softwares called LDPred [33] and PRSice [34]. Both compute scores from GWAS statistics but differ in how they take into account LD structure. While none of them directly use regularization as mentioned in the methods section 3.3, their outcomes are comparable. PRSice performs LD clumping which selects the most informative SNPs and removes the correlated ones. It results in sparse model comparable to LASSO. Instead of directly removing the correlated SNPs, LDPred uses a Gaussian mixture prior giving results more comparable to a ridge regression. General properties of the PRS are detailed in table 4.1.

Some simple but interesting facts can already be observed. First the number of SNP across scores covers 5 orders of magnitudes, from 10^2 to 10^6. This naturally leads to question the

PGS	Source	SNP count	Cohort	Cohort Size
27	[35]	2,100,302	UK Biobank	(∼500k)
34	[36]	97	NHS HPFS	(∼15k)
320	[37]	263,640	UK Biobank	(∼500k)
717	[38]	557	UK Biobank	(∼500k)
298	[39]	941	TRAILS + meta-GWAS	(∼1k)

Table 4.1.: Existing polygenic scores summary table

causal nature of selected SNPs. Furthermore, the scores are mostly fitted to UK bio bank or European cohorts. Unfortunately, obesity affects individuals from other ancestries as well. Until further analysis is conducted on cohorts from other ancestries, the generalizability of these scores should not be taken for granted.

Looking at the two extremes (score 27 and 34) highlights the difficulty of the task at hand. Score number 27 uses 2 millions SNPs. To put this number into perspective, individuals usually differ from one another at 4 to 5 millions SNPs. While the two number are not directly comparable, this highlight the level of diversity that 2 millions SNPs represents. More practically, 2 millions is still a large number. As such, score 27, regardless of its performances (detailed later) probably lacks the specificity needed to identify causal SNPs. On the other extreme, score 34 has a very low number of SNPs making it highly interpretable, however, this size may be insufficient to get robust performance (detailed later). This score most probably lacks sensitivity.

To get a quantitative comparison of scores, we analyze their SNP supports and weights. Table 4.2 summarizes the co-occurrences of SNPs across models. The overlap varies from score to score. For instance, PGS 717 and 27 only have 166 SNPs in common. This is a relatively small overlap since PGS 717 and 27 select 557 SNPs and 2 millions SNPs respectively. These scores disagree with one another in terms of SNP selection. Score 27, 320 and 34 give more concurrent results. PGS 298 agrees well with 27 only. Overall, this short analysis demonstrates the current scores are far from reaching a consensus on SNP selection.

On the other hand, the models reach a more stable agreement on the effect of SNPs. Figure 4.1 displays the relations between the weights of pair of scores. Each graph displays the weights of one model versus the weights of another for the SNP they have in common. Beside score 27, weights across models are correlated. For instance, high effect SNPs in score 34 also have a high effect in score 298. Compared to the others, score 27 is built on a dense assumption rather than a sparse one. This change in assumption explains both its large number of SNPs and the different weight distribution observed.

	PGS_27	PGS_298	PGS_320	PGS_34	PGS_717
PGS_27	2095834	797	261172	91	166
PGS_298	797	941	385	35	43
PGS_320	261172	385	263639	83	51
PGS_34	91	35	83	97	3
PGS_717	166	43	51	3	557

Table 4.2.: Count of co-occurrence of SNPs across PRS models

In total, while the scores do not come to an agreement in terms of SNP selection, they do agree on the effect of individual SNPs. This highlights the importance of the selection step in polygenic risk scoring.

4.1.2. Gene expression as a prior to risk score

This subsection explores if expression models are good priors for polygenic risk scoring (Hypothesis 1). To do so, the existing risk scores are compared with gene expression models.

An important step toward biological interpretability of scores is to look whether selected SNPs are linked to gene expression. Most variations occurs in non coding regions, therefore SNPs could play an essential role into the observed phenotype via gene expression regulation rather than direct alterations of protein sequences. To do so, we use the elastic net expression model from article [10]. Table 4.3 shows that for all scores beside score 717, a large fraction of SNPs are indeed linked to gene expression.

	SNP count	eQTL count	eQTL %
PGS_27	2095834	1282285	61.2
PGS_298	941	629	66.8
PGS_320	263639	180672	68.5
PGS_34	97	67	69.1
PGS_717	557	138	24.8

Table 4.3.: Fraction of SNPs reported as eQTLs [10]

To understand if the scores select SNPs with potentially relevant biological functions, we perform a small enrichment analysis. Simply put, we analyze how the eQTLs in each PRS are distributed across tissues. For the case of BMI, we could expect that the risk scores favor

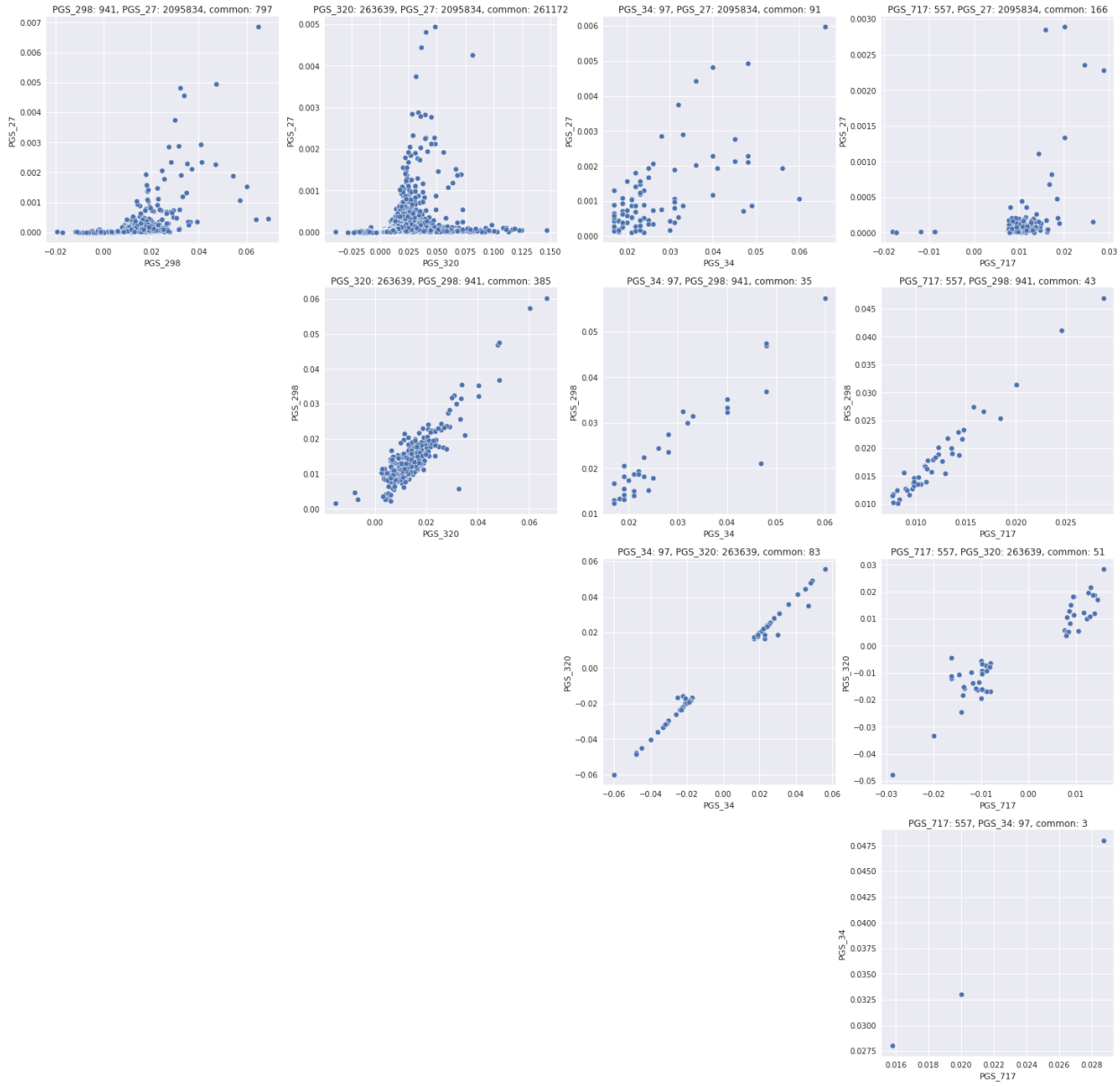

Figure 4.1.: PRS models' weight pair plot

SNPs linked to gene expression in metabolically relevant tissues such as adipose, liver or hypothalamus tissues for instance. To check for enrichment, we compare the distribution of eQTL by tissue within the risk score and within the expression model. More specifically, let's assume that x% of eQTLs in the risk score play a role in tissue A. Because eQTLs are not evenly distributed across tissues in expression models, we must compare x to the prior distribution of eQTLs in tissue A. This prior is simply the number of eQTLs in tissue A divided by the total number eQTL in the expression model. Figure 4.2 shows that for both scores 320 (left) and 27 (right), the eQTLs are mostly distributed the same way in the risk score and in the expression model. Some potentially relevant tissues for BMI are highlighted on the plots. From this, we conclude that the data from this expression model does not provide insights on potential underlying gene expression changes in the risk scores.

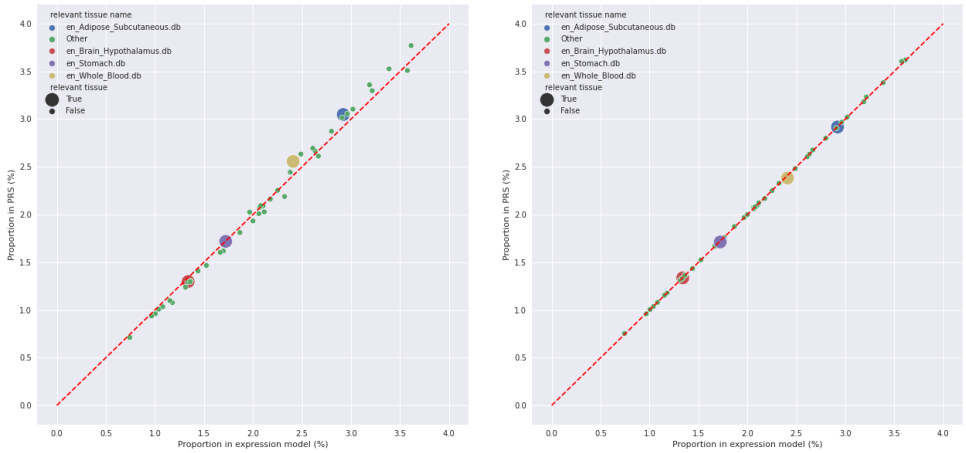

Figure 4.2.: eQTL proportions by tissue, PRS versus expression model. PRS 320 (left), PRS 27 (right)

To go further, we can study the relation between expression weights and polygenic risk score weights. A first simple assumption would be that more extreme gene expression leads to increased trait expression. As a consequence, expression weights and risk score weights would be expected to correlate. Unfortunately, this assumption is not verified empirically. Figure 4.3 displays the relation between SNPs expression and risk score weights for PGS 320. Here, SNPs with high effects do not lead to high expression. This analysis is consistent across all the risk scores.

Given that expression models do not give insight on biological functions of scores and that expression and risk weights are not related, we conclude with regards to hypothesis 1:

> **Available gene expression models are not relevant priors to current genetic risk scores.**

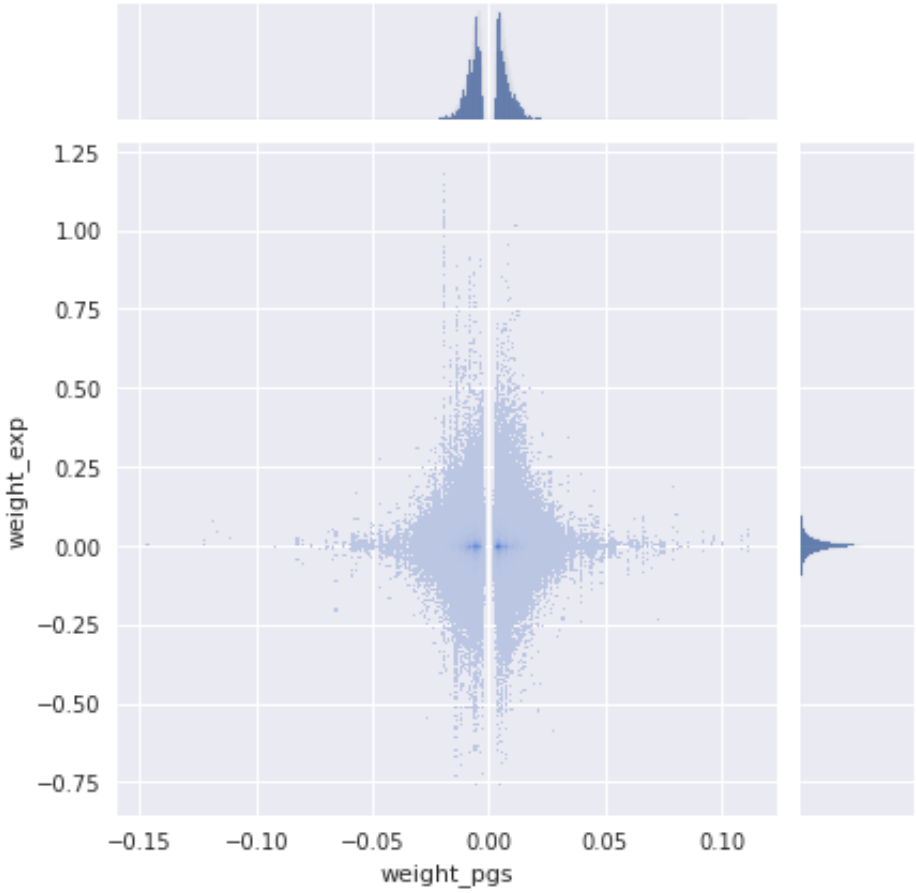

Figure 4.3.: Expression weight vs PRS weights for SNPs in PGS 320

4.1.3. Risk scoring at the individual level

In this section, the applicability of risk score for prediction at the individual level (hypothesis 2) is checked by testing the performance of all risk scores on an independent data set.

Up until now, the predictive performance of scores has not been considered. Because performances are reported using different metrics (R^2, correlation) on different datasets, self reported results are hardly comparable across studies. Furthermore, even on identical datasets, research group sometime use different imputation influencing the results. For these reasons, to get a fairer comparison, all the scores are recomputed using the same pipeline on the same dataset [8]. Because scores have different scales, the metric used to assess the models is correlation. Table 4.4 shows the correlation between pairs of scores and BMI.

Looking at the right most column of table 4.4, we can see that PRS_27, PRS_320 have the best performance at 0.36. They also happen to use the most SNPs with two millions and two hundred thousands SNPs respectively (see table 4.1). The less SNP intensive scores have significantly lower performance. The correlation between scores varies from 0.18 to 0.83. This is excepted as scores do not share the same SNPs. While PRS_27, PRS_320 have similar performances they are different in terms of number of SNPs and they do not correlate perfectly. Given that PRS_320 contains almost all SNPs from PRS_27, one could wonder whether PRS_27 contains superfluous information.

	PRS_298	PRS_34	PRS_717	PRS_320	PRS_27	BMI
PRS_298	1.00	0.53	0.62	0.41	0.58	0.18
PRS_34	0.53	1.00	0.43	0.19	0.46	0.10
PRS_717	0.62	0.43	1.00	0.32	0.43	0.13
PRS_320	0.41	0.19	0.32	1.00	0.84	0.37
PRS_27	0.58	0.46	0.43	0.84	1.00	0.37
BMI	0.18	0.10	0.13	0.37	0.37	1.00

Table 4.4.: Cross correlations between PRS scores, BMI on KORA [8]

Figure 4.4 displays the raw data summarized by 4.4 to get a better feel at the relation between risk scores and BMI. In particular, it is apparent that even the best risk scores have a high variance. They can hardly be used for accurate prediction at the individual level.

More worryingly, while all risk scores follow a gaussian distribution, BMI does not. Since risk scores are computed as the sum of mostly independent SNP, the central limit theorem states that they must follow Gaussian distribution. On the other hand, because of physiological reasons, BMI tend to be skewed. The healthy BMI lies between 18.5 and 25. A BMI lower

than 16 is an immediate life threat while a BMI over 40 is a longer term life threat. As a result, the distribution tails is heavy on the extremely obese side and nonexistent below 16. This distribution mismatch between the risk score and the trait observed inevitably leads to lower predictive performance especially for values on extreme ends of the spectrum.

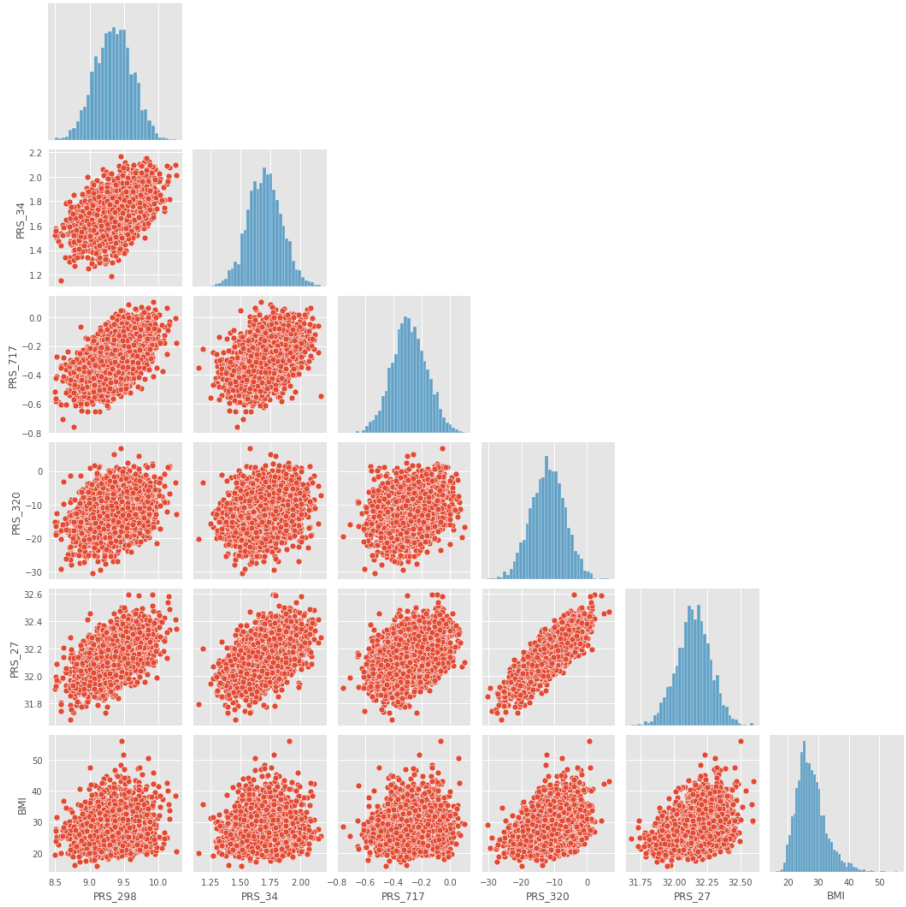

Figure 4.4.: Risk scores and BMI correlation pair plot on KORA [8]

As seen in figure 4.4, raw plots are hard to interpret and compare. Instead it is common to plot summary statistics rather than raw risk scores. Figure 4.5 presents this information in a summarized but clearer way.

At the top, for each score and each PRS decile, we show a BMI box plot. For all scores, there is a tendency of BMI to increase with PRS. As expected by the previous analysis on correlation, PRS 27 and 320 display a much stronger link to BMI than the other scores. PRS 717 and 34 show a much less convincing link. The distribution mismatch detailed earlier explains the larger BMI variance for the top PRS quantile. For this reason, the scores require improvement to study effectively the effects of genetics on extreme obesity.

At the bottom of figure 4.5, a point plot displays the mean (and its uncertainty) for each score across PRS deciles. This plot allows direct comparison of the different PRS. Again, BMI increases on average faster with PRS 27 and 320 than other scores. While BMI increases on average with most PRS, the models have their limits. For instance for PRS 27, the confidence interval on the mean of quantile 4 includes the mean of both quantile 3 and 5.

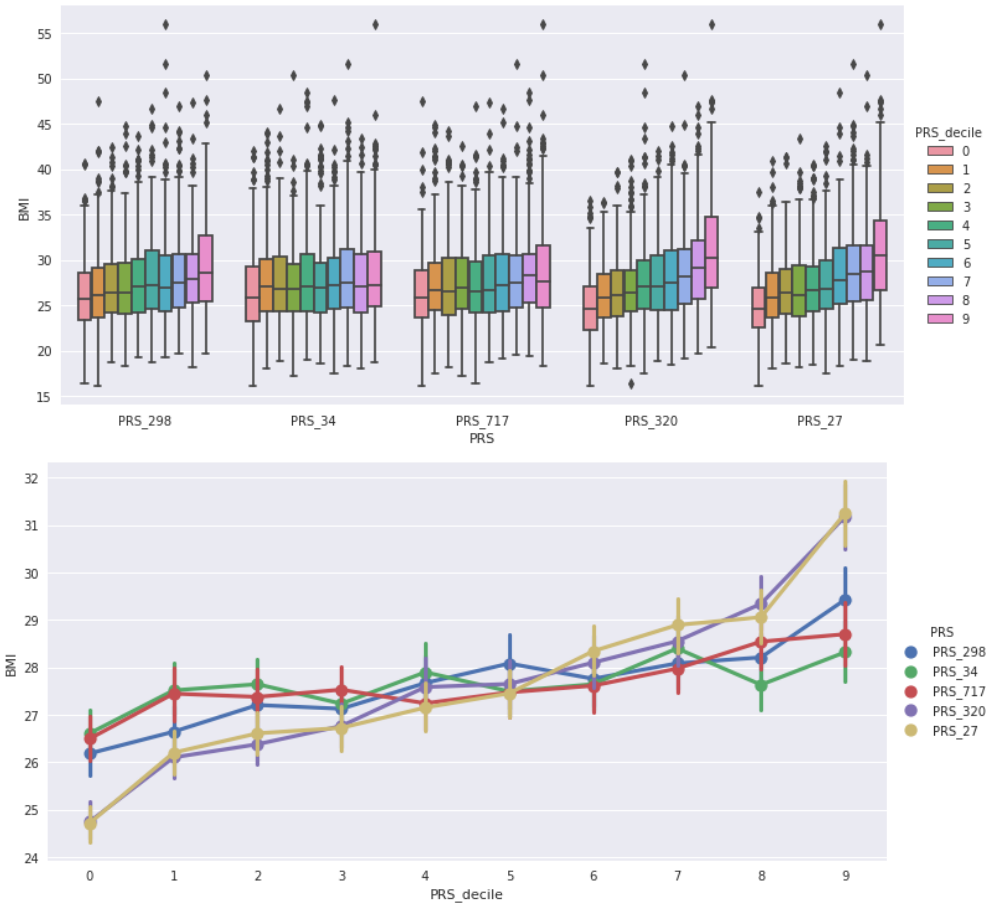

Figure 4.5.: PRS distributions by PRS decile on KORA [8]

Note that while these plots seem to show a very strong relation between the PRS and BMI, it is crucial to remember that they display summary statistics. In particular the overlap between PRS quantiles is large for all scores. Figure 4.6 gives a more complete picture of the scores and the plots above taken individually. In this figure we can observe the clear increase averages of BMI using the boxen plots while still getting a sense of the high variability of the scores using the scatter plot to the left.

While the scores are built as regression models, it is still possible to use them as binary classifiers to predict obesity rather than BMI. To do so, it suffices to define a PRS threshold

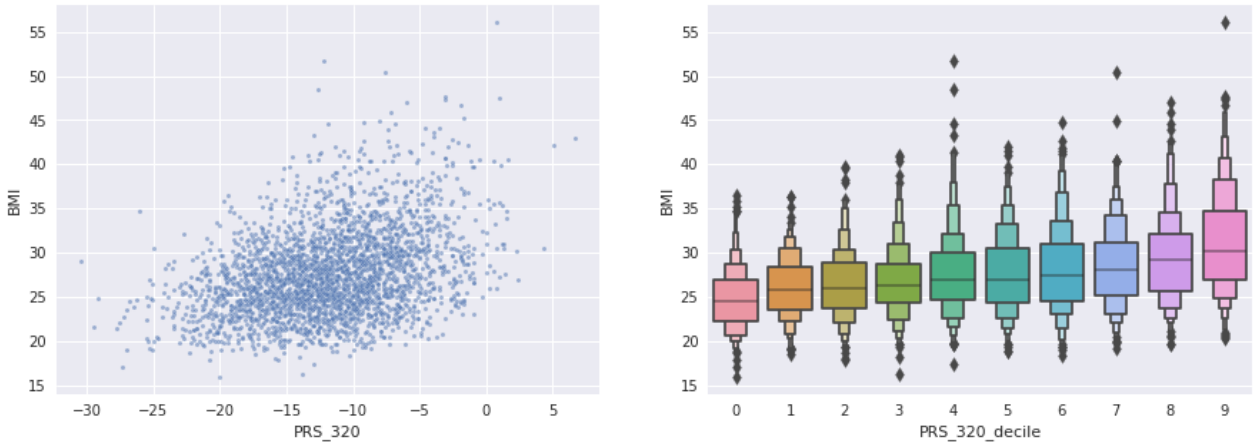

Figure 4.6.: PRS 320 computed on KORA [8], raw and boxen by BMI classes

above which all samples are considered obese. The resulting classification models are evaluated using the precision recall curves on figure 4.7. While the scores performs better than a random prediction model, precision drops rapidly as recall increases. Because of this, scores for individual samples should be interpreted with caution.

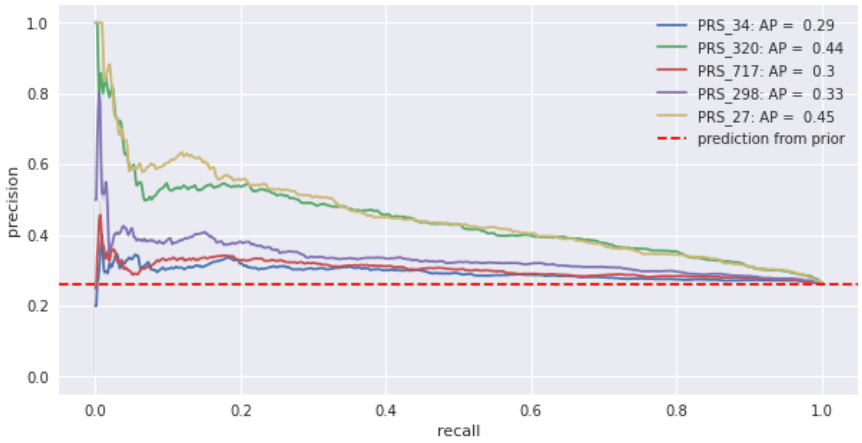

Figure 4.7.: Precision recall curves for all PRS

Because of large residual variance and relatively low predictive performance, we reject hypothesis 2:

> **The available models do not allow for individual prediction**

4.1.4. An ensemble polygenic risk score

Building an ensemble model of existing score is simple way to whether risk scores are independent of one another (hypothesis 3).

An interesting characteristic from the existing scores is that their SNP support is somewhat different, yet they all correlate to BMI. An hypothesis is that the scores do not capture the identical information about BMI. Hence, we could imagine building a first polygenic risk score that simply combines the risk scores. To do so, we create a linear and a boosted model using the PRS 298, 34, 717 and 320 as well as age and gender as feature. PRS 27 is omitted as it is highly correlated to PRS 320, contains more SNPs but do not show increased predictive power. To perform independent validation, the score are fitted and cross validated on UKB and tested on KORA. The test correlation stand at 0.33 and 0.29 for the linear model and the boosted model respectively. This slightly lower than PRS 320 and 27. Figure 4.8 shows how comparable the three scores are.

Since combining the risk scores do not improve predictive performance, we conclude that they mostly hold the same information. Hence, hypothesis 3 is confirmed:

> **Risk scores are not independent of one another.**

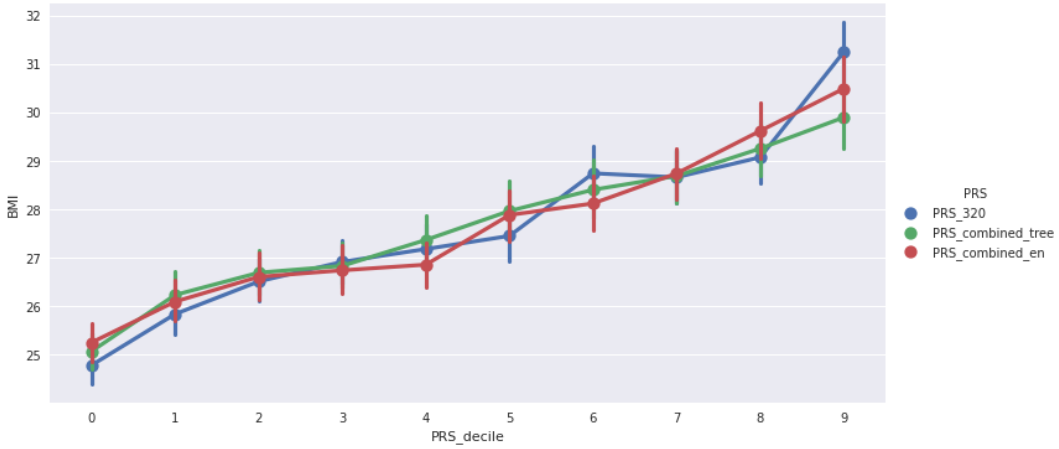

Figure 4.8.: Pointplot, PRS 320 and combined PRS

4.1.5. The importance of population structure

In this section, we check the second aspect of hypothesis 3 by looking at the distribution of risk scores for population of different ancestries.

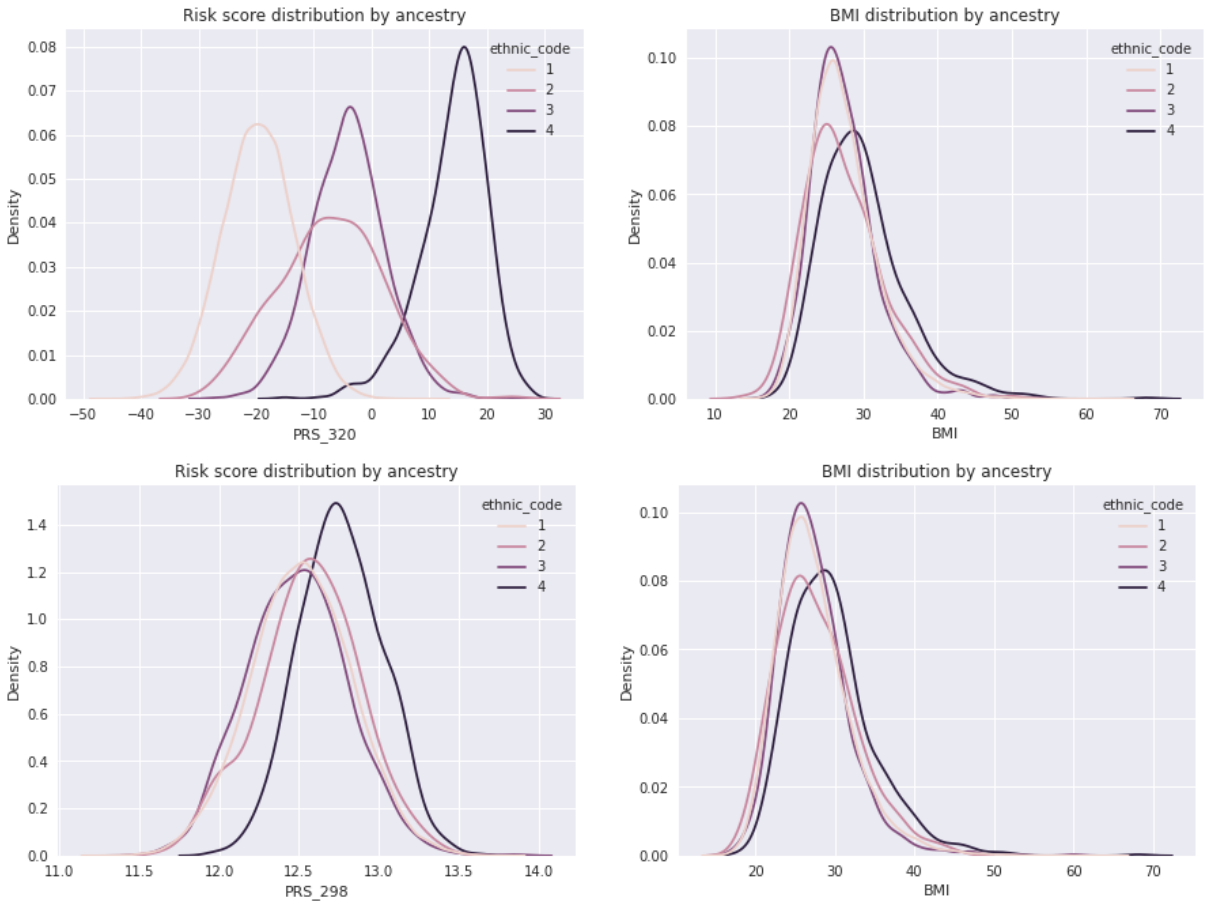

Figure 4.9.: PRS 320 (top), PRS 298 (bottom) and BMI distributions by ancestry

As mentioned previously, population structure is a large source of biases in genomic studies. Genetic information varies from population to population even at the European scale. Ancestry also correlates with cultural and environmental factors. For this reason, risk scores should always include genetic principal components to capture some of this structure. On figure 4.9, the distribution of PRS 320 and PRS 298 computed on the UK biobank dataset are displayed along the distribution of BMI. These scores are "raw" sums of SNPs that do not include any gender or population covariates. In this setting, the risk scores distributions are dramatically shifted across ancestries especially for score 320.

Because the risk score distribution are shifted by population, the second part of hypothesis 3 is confirmed:

> **Risk scores depends heavily on ancestry.**

4.2. Non linear polygenic risk scores

Hypothesis 4 on SNP-SNP interactions is checked by building a boosted model using the SNPs present in existing risk scores. To capture the interactions but to limit over fitting, the maximum depth of individual trees is set to 3. Detecting such interaction could improve the predictive performance of risk scores.

We gather the 264698 SNPs present in all the risk score but 27. Then, we fit and cross validate the data on the UK biobank. To fit the model, we use the highly effective XGBoost python library [40]. Unfortunately, this model does not come close to the best risk scores in terms of correlation. Figure **??** shows the regression diagnostic plots and metrics for the model. Despite efforts to regularize the model using l1 penalty, sub-sampling on rows and columns, the model over fits. Because of this poor performance, the next steps will focus on exploring links between these SNPs and non genetic factors.

As explained before, the selection is also crucial to the success of the model. Figure 4.11 shows that the SNP importances and the absolute weight in the PRS scores are radically different. Some important SNPs for the boosted tree have near zero coefficient in PRS 320 and vice versa.

This result is consistent at smaller scale. When combining all PRS but 27 and 320 we get 1527 SNPs. Again, the regression boosted model shows worst performance than a regularized linear model and a vastly different SNP importance structure.

As final note, the variationnal model showed no better performance than the random forest on either configuration. While the non linear models were not able to perform better than linear models on these particular set of SNPs, they can be leveraged to to explore genetic and non-genetic interactions.

Figure 4.10.: Diagnostics and metrics, 260k SNPs boosted tree risk score

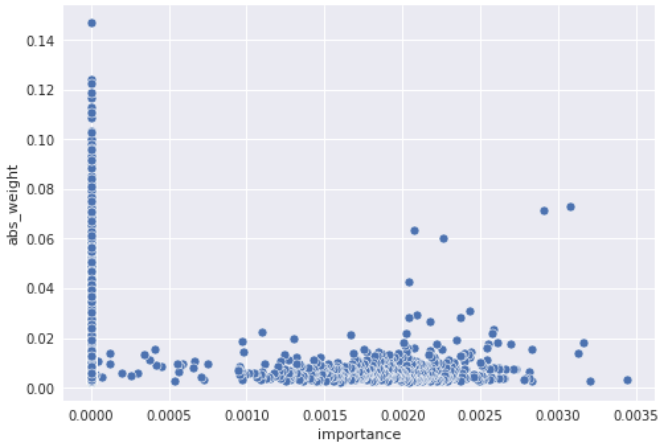

Figure 4.11.: boosted tree importance vs PRS 320 absolute weights

Because the predictive performance is not improved relative to the linear risk score, we reject hypothesis 4:

> **Our non linear models did not improve risk scores by capturing SNP-SNP interactions**

4.3. Genetic and non genetic interactions

To explore interactions between genetic and non genetic variables (hypothesis 5), we first assess the association between non genetic variables and complex traits. Then, we build a model containing both a genetic signal and environmental variables.

4.3.1. Non-genetic associations

To understand the mechanistic link between genetic variants and observed BMI, studying the non genetic factors is necessary. It is possible to imagine many possible pathways from SNPs to increased risks of obesity. For instance, a well studied variant linked to obesity is located on the fat mass and obesity-associated protein (FTO). Article [41] reviews potential mechanistic pathways between this gene and obesity. Eating habits, energy expenditures, or circadian rhythms are mentioned as potential factors influenced by the FTO gene. The UK biobank provides a number of environmental factors and bio markers that can be used to explore potential mechanistic pathways towards obesity.

Studying non genetic associations adds some challenges compared to studying genetic risks alone. First of all, environmental data is harder to measure in an unbiased way. Much of the data in the UK bio bank relies on questionnaire filled by patients. Bio markers on the other ends are objectively measurable and comparable. Therefore, the non genetic model includes data of very different nature. Comparison between the two must be done carefully. A more problematic issue comes from the nature of non genetic data. While genetic information is invariant to usual external factors for a given patient, bio markers and other variables are not. As such, a genetic association is either a causal link or a spurious correlation confounded by population or LD structure for instance. This greatly simplifies interpretation. In the case of external factors, biological or environmental, causal links can be harder to disentangle.

Here is an illustrative example for the case of BMI. To simplify the example, only genetic information and sedentarity [1] are considered. We could imagine different causal and spurious links. Sedentary could reduce energy consumption thus leading to weight gain. Weight gain could limit movement thus leading to sedentary. Both traits could also be caused by a common genetic factor. The true explanation could also be a bit of all of the above. Figure 4.12 displays possible causal links. While this example may seems elementary, this riddle must be solved for each and every bio marker and environmental factor considered. A possible

[1] spending too much seated

way to test for causality between those variables would be through Mendelian randomization. Because this analysis goes beyond the scope of this thesis, the non genetic risks model is not used as predictive analytic tool but as a way to mine for associations between obesity and non genetic factors. The strongest associations can later be analyzed in pair with genetic factors.

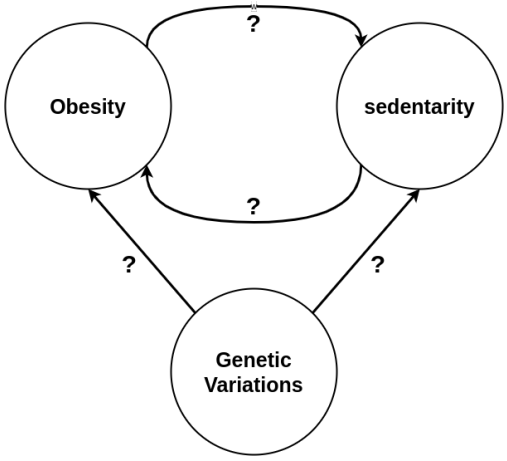

Figure 4.12.: Possible causal links between obesity, sedentarity and genetic variations

To find these associations, we fit and cross validate a boosted tree and linear regression on the UK bio-bank cohort. The covariates that directly indicate obesity such as weight or body fat percentage are removed from the model. To get more interpretable results, covariates that are too granular are removed. For instance, all variables measuring food intakes (poultry, beef, fruits, etc ...) are filtered in favor of the more general "variation in diet". Variables related to feelings and emotions are not included as this topic is too complex to be analyzed here. Bear in mind that these choices are made to get a simple enough picture rather than a full understanding of non genetic association to BMI.

The cross validation results are displayed in table 4.5. In total, 113 variables are analyzed. Given the considerations on causality described previously, these numbers are not easily comparable with the PRS[2]. Still, The R2 metric have low variances across folds attesting for robustness of the models on this dataset. The gradient boosted model shows better performance than the linear one, revealing potential interaction between variables.

model	Fold 0	Fold 1	Fold 2	Fold 3	Fold 4
Boosted Tree	0.425	0.419	0.419	0.415	0.412
Linear Regression	0.358	0.365	0.333	0.356	0.351

Table 4.5.: R2 scores by cross validation folds

[2]Beware that here, the metric used is R2 and not correlation as earlier

The models reveal that environmental factors and bio markers are strongly linked to BMI. Table 4.6 displays the 20 most important variables selected by the gradient boosted model. Bio markers related to white and red blood cells show high association with body mass index. Top environmental factors include diet and sedentarity indicators. Other variables such as heart problems and diabetes diagnosed by doctors describe diseases that co-occur with obesity. While this topic deserves a deeper analysis, the next section will only focus on exploring links between "important" variables and polygenic risk scores.

feature	importance
High light scatter reticulocyte count	0.107
Usual walking pace	0.102
Vascular/heart problems diagnosed by doctor	0.095
Weight change compared with 1 year ago	0.073
Snoring	0.046
Diabetes diagnosed by doctor	0.041
Reticulocyte count	0.034
High light scatter reticulocyte percentage	0.028
Overall health rating	0.027
Variation in diet	0.019
Time spent watching television (TV)	0.017
Major dietary changes in the last 5 years	0.017
Current tobacco smoking	0.015
Creatinine (enzymatic) in urine	0.014
Immature reticulocyte fraction	0.014
Number of treatments/medications taken	0.013
Red blood cell (erythrocyte) count	0.013
Sex	0.013
Sodium in urine	0.010
Lymphocyte count	0.009

Table 4.6.: Top 20 most important variables for boosted model

4.3.2. Environment genetic interactions

Because interpreting the association between bio markers and obesity requires more biology knowledge, the next section will focus on more intuitive environmental variables only.

To get a better idea of how the genetic signal compares to environmental associations, we fit a boosted model and a linear model using PRS 320 and environmental variable as features. These models allow us to compare the importance of the genetic risk score relative to the other variables. Table 4.7 shows that both models, linear and boosted tree performs slightly

better when the genetic risk score is added to environmental features. Still, the increase is rather modest and stands at around 0.03 R^2 unit.

model	Environment	Environment + PRS
Boosted Tree	0.31	0.34
Linear Regression	0.24	0.28

Table 4.7.: R2 validation scores by model

More interestingly, it is possible to compare the importance of variables in each model. Tables 4.6 and 4.9 show the top 20 variables for the boosted and linear model. While for the linear model the polygenic risk score comes at the very top, in the boosted model the risk score is less important relative to other variables. Remarkably, the PRS is higher than environmental factors. In this model, variation in diet is "less important" than the risk score.

Given the metric improvement and the relative importance of the genetic risk score in the model, hypothesis 5 is confirmed:

> **Studying environmental and genetic variables jointly improves association with BMI.**

top 20 feature boosted tree	boosted tree importance
Vascular/heart problems diagnosed by doctor	0.159
Usual walking pace	0.111
Overall health rating	0.058
Diabetes diagnosed by doctor	0.048
Snoring	0.047
Weight change compared with 1 year ago	0.039
Time spent watching television (TV)	0.026
PRS_320	0.026
Variation in diet	0.019
PC_2	0.017
Time spent driving	0.017
Sex	0.014
Current tobacco smoking	0.013
Time spent using computer	0.013
Smoking/smokers in household	0.013
Own or rent accommodation lived in	0.012
Major dietary changes in the last 5 years	0.012
Number of days/week of moderate physical activity	0.012
Exposure to tobacco smoke at home	0.011
Frequency of stair climbing in last 4 weeks	0.011

Table 4.8.: Boosted tree top 20 features (Environment and PRS)

top 20 feature linear model	linear model importance
PRS_320	1.254
Usual walking pace	-0.724
PC_1	-0.707
Vascular/heart problems diagnosed by doctor	0.683
Overall health rating	0.538
Weight change compared with 1 year ago	0.490
PC_2	0.468
Time spent watching television (TV)	0.362
Snoring	-0.354
Hand grip strength (left)	0.335
Variation in diet	0.331
Diabetes diagnosed by doctor	0.303
Smoking/smokers in household	0.283
Frequency of stair climbing in last 4 weeks	-0.257
Time spent using computer	0.206
Current tobacco smoking	-0.198
Frequency of friend/family visits	-0.183
Number of days/week of moderate physical activity	-0.178
Getting up in morning	0.174
Time spent driving	0.173

Table 4.9.: Linear model top 20 features (Environment and PRS)

4.4. Variational model

In this section, we perform two experiments to demonstrate how variational methods can create models including SNPs, environment and biomarkers (hypothesis 6).

In the following section models are fitted using the ADAM optimizer with a learning rate of 0.001 and a batch size of 64. The early stopping criterion is used to stop training once performance no longer improves on the validation set. The models are fitted using a single GPU on a random 100k subset of the UK Biobank. The hyper parameters are manually tuned.

4.4.1. First experiment

To make sure that the variational approach is sound on this particular SNP data, we conduct a first experiment. The goal of this is twofold. First we must check that regression results are on par with known risk score performance. Most importantly, we need to check that the generative process is working as expected. To keep this experiment simple, we simply fit the variational model the subset of SNP from the smallest risk scores (all but 27 and 320). The target variable is the BMI. As usual, we add gender and principal components to the model to account for population and sex biases. The network's architecture and parameters are described in table 4.10. These parameters are hand tuned by trial and error.

	value
X_shape	1515
latent_shape	10
activation	Sigmoid()
hidden_shapes	[50]
learning_rate	0.01
alpha	100
beta	0.01
gamma	10
lambda_l1	0.001
X_continuous_shape	0
X_categorical_shape	1515
X_categorical_n_classes	3
y_continuous_shape	1
C_shape	12

Table 4.10.: 1st experiment parameters

As excepted given the previous analysis, in terms of predictive performance, the model give approximately similar results as the smaller risk scores with a low correlation standing

around 0.20. Still, it is interesting to assess the generative and reconstructive performances of the model.

First, the reconstruction of SNPs is evaluated using the confusion matrix **??**. Each cell displays the number of SNPs predicted in a particular class divided by the total number of SNPs in that class. Note that this table must be interpreted with care as this is a multi class, multidimensional, imbalanced classification problem. Most of the 0 and 1 SNP are reconstructed correctly. Additionally, SNPs with value 2 are almost never classified as 0. The model still struggles to recognize 0 from 1 and 2 from 1.

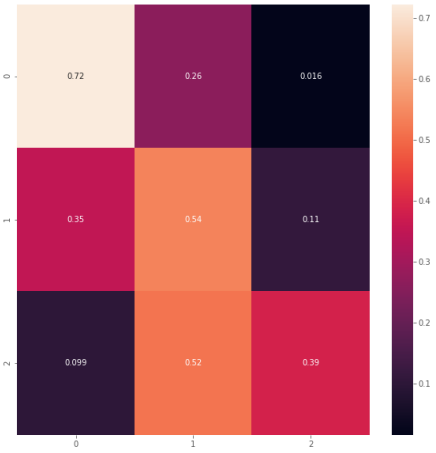

Figure 4.13.: SNPs reconstruction confusion matrix (experiment 1)

Naturally, just as with any metrics, the table above is a summary. To get a better idea of how the model performs, we can plots the reconstruction for specific samples of specific SNPs. To do so, it is helpful to look directly at the classifier outputs, the predicted probabilities. Since the classification problem has three classes, the reconstruction probabilities can be conveniently plotted on a simplex. In deed, for a given sample and SNP a prediction consists of 3 probabilities summing to one, as such, the point belong to a simplex. Each vertex of the simplex represents a SNP class, top corner is 0, left corner is 1, right corner is 2. The closer a point is to corner, the more certain classifier is about the prediction. Figure 4.14 displays this simplex for the probabilities of the reconstruction of SNP rs12033257 and rs6700816 for each sample in the validation set. On the left plot, most points are located near the 0, 1 boundary. Additionally, the probability of predicting 2 is non negligible for all points. This reveals that for each sample, there is a great uncertainty about the reconstruction of this given SNP. On the right plot however, all points are in the top corner (predicted as 0) despite having different true classes.

It is also possible to look at a similar representation from a sample wise perspective. In this case, for a given sample, we display the reconstruction probabilities for all SNPs. The results

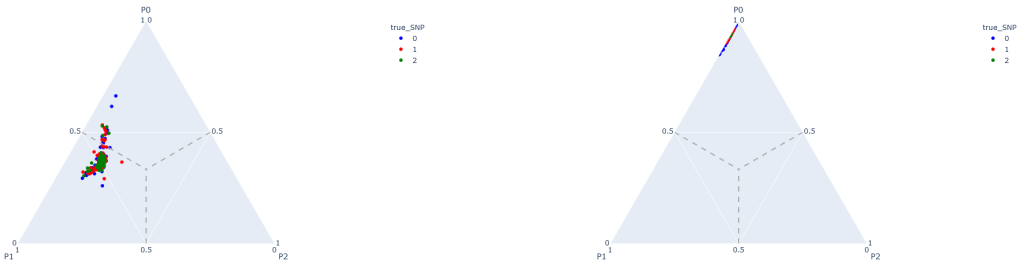

Figure 4.14.: SNP rs12033257 (left) and rs6700816 (right) probability reconstruction on validation set colored by true class

are displayed in figure 4.15. For clarity, one plot is shown per SNP class. Again, this figure shows the uncertainty on the prediction. The first observation is that many probabilities do not lie in their "correct" area. For instance, a few points encoded as 0 in the data lie in the area where the maximum probability of reconstruction is 2. Ideally, blue points (0) should be mostly in the top corner, red points (1) in the left and green point (2) in the right. More interestingly, the plots also show that, while SNP 0 and 2 can have relatively certain predictions, ones are almost always uncertain. This concurs with results already seen in the confusion matrix. Finally, no points lie at the frontier of 0 and 2.

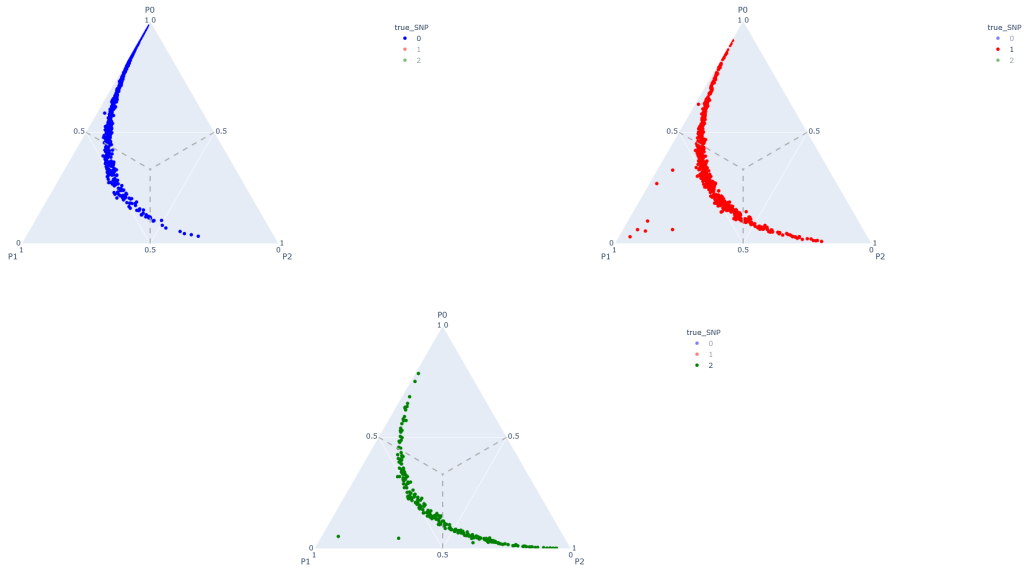

Figure 4.15.: Single sample reconstructions probabilities for all SNPs colored by true class

Now that the reconstruction performances are clearly established, it is possible to look at the generative performance. While the two are related, it is possible to get "realistic" sample

without perfect reconstructive performance. Actually, a goal of auto encoder is to capture relevant feature and smooth potential noise. For our purpose, we simply check that the population of reconstructed samples have correct allele frequencies. To do so we sample a population from the model. To get comparable populations, individuals are sampled from $p(X|c_{val}, y_{val})$ with y_{val} the BMIs and c_{val} the age, gender and PCs of the of the validation set. Figure 4.16 shows that for the majority of SNPs the sampled and test set have near identical allele frequencies.

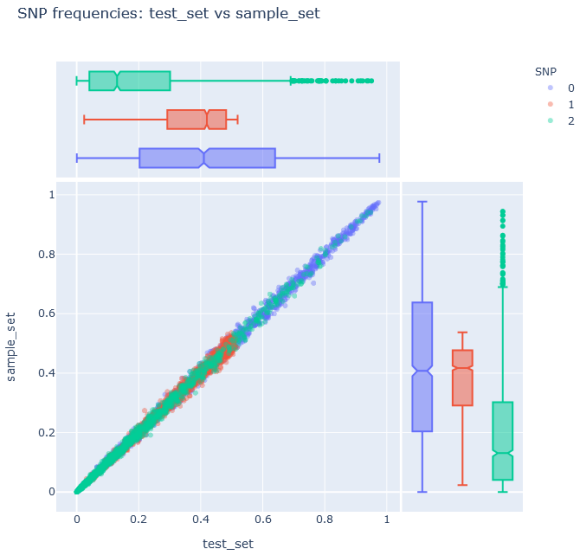

Figure 4.16.: SNP allele frequencies per SNP class, sample from model vs validation set

Since the regressive, reconstructive and generative performance are evaluated, we can move on to using the method to explore relevant variants. Of course, given the limited regressive performance of the model, results should be interpreted cautiously. Nevertheless, the method can be applied and checked against the polygenic risk scores weights. In short, we sample pairs of individual whose embeddings are rigorously the same except for the BMI value (a case and a control). Then, we look at the difference for each SNP between the case and control. As seen above, looking at the raw probabilities is more informative than looking at the SNP reconstructed by taking the max probability. Hence, in this particular case the "difference" between case and control for a given SNP probabilities is the Jensen–Shannon divergence which define a distance between probabilities [3]. This difference is analyzed on a great number of different case control pairs for different BMI level on figure 4.17. On this figure, we can see that, as we increase the difference in BMI between the case and control group, some the average distance between SNPs in the case and control group increases.

[3] In this method section, we used the absolute difference (l_1 norm) as we were comparing binary pixels instead of probabilities

Overall, the difference seems rather noisy for most SNPs, some however seem consistently above the the noise. Figure 4.18 shows the weights of PGS 298 versus the distance found by the model. Again, the result is rather noisy. Some SNPs have a high weight for both the variational model and the PGS. However, the relation seem mostly noisy. This echos the results previously found with with the boosted model.

Figure 4.17.: Sampled case-control difference for multiple BMI values

4.4.2. Second experiment

In this second experiment, we add bio markers and environment data to the model. Biomarkers are modelled (like SNPs) rather than used as conditions. This is simply done by splitting X into X_{snp} and $X_{biomarkers}$ and using the distribution detailed in the methods section. On the other hand, environmental data is used as conditional variable. The motive is to avoid conditioning on unlikely bio marker-BMI-environment combinations as the method rely on comparing realistic counterfactual samples. Paramaters of the model are detailed in 4.11, they are again hand tuned. The goal of this experiment is to understand if the model can sample

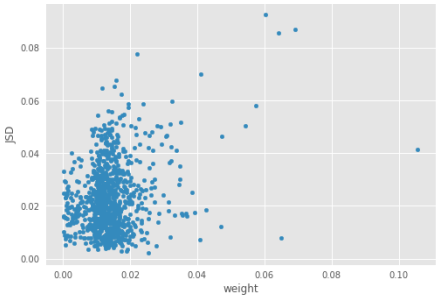

Figure 4.18.: PGS 298 weights vs case-control (20-60) difference

realistic SNPs and biomarkers at the same time. In this setting, the association between SNPs-biomarker-environment and BMI stands at an R^2 of 0.42.

	value
X_shape	1537
latent_shape	20
activation	Sigmoid()
hidden_shapes	[100, 50]
learning_rate	0.001
alpha	150
beta	125
gamma	20
lambda_l1	0.001
X_continuous_shape	22
X_categorical_shape	1515
X_categorical_n_classes	3
y_continuous_shape	1
C_shape	82

Table 4.11.: 2nd experiment parameters

First, the model reconstructs SNPs with the same performance as the previous one as shown by the confusion matrix 4.19.

From a bio marker perspective, figure 4.20 shows that most of the biomarkers are sampled realistically. Beside the last three variables, the sampled and validation distribution are nearly identical.

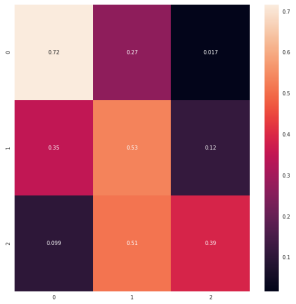

Figure 4.19.: SNPs reconstruction confusion matrix (experiment 2)

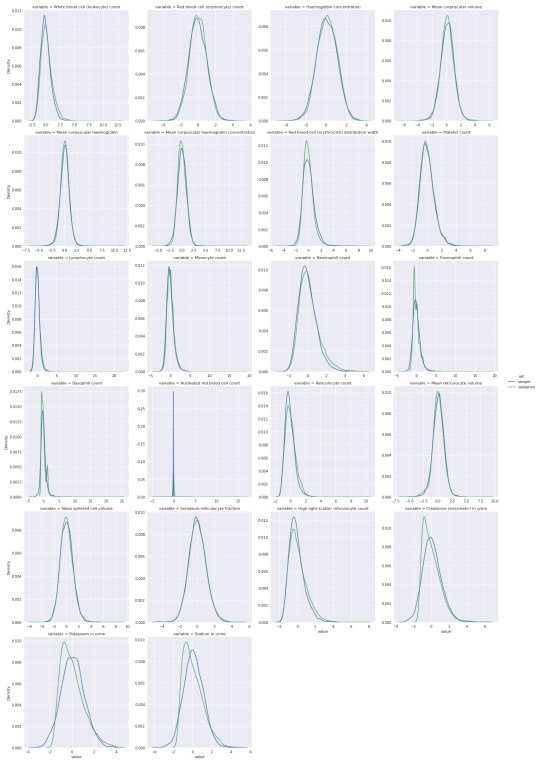

Figure 4.20.: Biomarker distribution sampled vs validation

More interestingly, figure 4.21 shows the correlation matrices for sampled (upper triangle) and validation (lower triangle) biomarkers. Beside the last three biomarkers, the matrix is symmetric. This shows that the correlation of sampled variables mimics the observed correlation well.

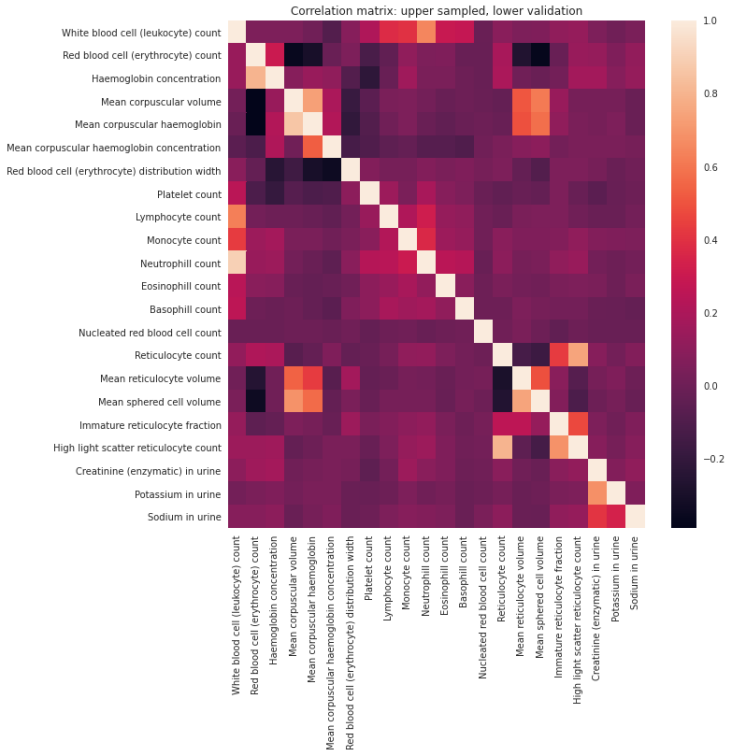

Figure 4.21.: Biomarkers correlation matrices. Validation (lower) Sampled (upper)

Figure 4.22 shows an example of sampled and validation immature reticulocyte fraction versus BMI. The relation between the synthetic points and the trait of interest is accurately captured.

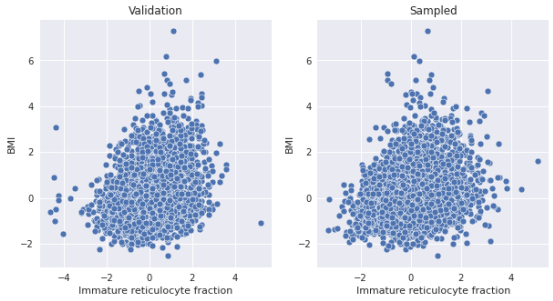

Figure 4.22.: Immature reticuloctye fraction validation and sampled

In this setting, the counterfactual experiment do not provide the same result as in the previous experiment. Manhattan plots 4.17 and 4.23 display different profiles. In particular, the top 3 SNPs between index 1200 and 1400 in figure 4.17 no longer appear in the new experiment figure 4.23.

Figure 4.23.: Sampled case-control difference for multiple BMI values. 2nd experiment

Given that the variational model sampled realistic SNPs and biomarkers simultaneously given environmental variables, we conclude with regards to hypothesis 6:

> **Complex interactions between SNPs, biomarkers and environment can be captured by variational models**

5. Discussion

5.1. Existing polygenic risk scores analysis

The results on the existing risk score provide a first overview of the topic of genetic risk estimation. The analysis of five scores from different research groups using a variety of methodologies reveal that at the moment, there is still a lot to uncover. The methodologies used by the studies are either closed to regularized linear models or linear models combined significance thresholds. As predicted in the simulation in the methods section, the models obtained do not concur on the SNPs selected. To accurately describe the influence of genetics on a trait such as BMI, causal variants must be selected. The small overlap of SNP support across model, particularly when fitted on different cohorts with different methods show that the selection process is far from solved. Additionally, enrichment analysis failed to uncover significant links between expression weights and risk score weights. We conclude with regards to hypothesis 1:

> **Available expression data do not give reasonable priors for polygenic risk score models scores**

In terms of performance, the largest models give the best association. Still, low precision/recall metrics of the risk scores forbid potential concrete applications. To illustrate this, let's imagine using score 320 to deploy preventive treatments as it is sometimes discussed. Even with a conservative recall threshold such as 0.2, the precision is only around 0.5. Concretely, to diagnose 20% of obesity cases would have to be wrong on half of our diagnostics and preemptively treating plenty of people that do not need it. More worryingly, the risk scores are fitted either on the UK Bio bank or cohorts of European ancestry leading to poor generalizability as discussed later. As a result, across all scores we conclude with regards to hypothesis 2:

> **The available models do not have the performance to allow for individual predictions.**

Because of the variety in size, SNP support and performance of scores, merging them into a single score seemed like an attractive idea. However, the results show that combining risk scores into a single predictive model does not add benefits to estimation of risk scores whether a linear or non linear model is used. A possible cause of this is that, while being different across scores, the selected SNPs are correlated, hence scores carry mostly the same information. First, this idea is supported by the empirical correlations computed in the first

section. It is also supported by a methodological argument. For instance, in a lasso model, take two SNPs x_1 and x_2 that have correlations $r_1, r_2 \sim 1$ with a trait and $r_{x_1,x_2} \sim 1$ with each others such that $r_1 > r_2$. In this scenario, x_1 will be selected in instead of x_2. Naturally, at fitting time, r_1 and r_2 are estimated based on the available data that can vary from cohort to cohort. In some cohorts, r_2 can be greater than r_1 and x_2 and be selected instead. The conclusion to the first part of hypothesis 3 is:

> **Different risk scores do not seem to contain complementary information about the complex trait**

While we assessed the models on an independent data set, at the moment there is no evidence that these scores can be generalized to populations other than European. In particular population proved to bias scores. This bias can be corrected relatively easily by adding an ancestry dependent bias[1]. However, this rough correction is not a panacea. For instance, the UK biobank contains around 95% and 2% samples from white and Asian ancestry respectively. Because of this lack of diversity, this risk scores will fit variants that are present in individuals from European background and potentially miss or misestimate variants present in other populations. The second part of hypothesis 3 is confirmed:

> **Risk scores must be corrected for population structure to be relevant**

5.2. Non linear polygenic risk score

Improving the polygenic risk score is not as straightforward as imagined. Despite exploring different models with different hyper parameters on different SNP supports exceeding the performance of the linear risk score proved challenging. Few reasons can explain this setback. The first issue is over-fitting. As discussed earlier, even linear models, which are relatively robust to over-fitting need plenty of regularization. This problem gets worse with non linear models. Without regularization, the gap between metrics on training and testing set are immense. Only stringent regularization, to the point of not capturing any relevant information reduces this gap. The reason for this is that the signal to noise ratio is so low that finding the right bias-variance trade off becomes extremely challenging. As seen in the non genetic association section, with a stronger association signal, non linear models clearly outperform linear ones. The second reason is directly linked to the SNP support. The important SNPs for the linear regression and the boosted model are vastly different. For instance in our particular experiment the highest weight for the PRS has a zero importance for the boosted model and vice versa. While finding different SNP support is one of the reasons of using boosted trees, this large difference is rather concerning. This also highlights that the SNP selection process prior to modeling is crucial. By construction a subset of SNPs selected by regularized linear models or association studies will select only SNPs that are linearly dependent to the trait. To

[1] In the studies, models add principal components to capture this structure. Unfortunately, only SNP weights were given

get better results, two more rigorous approaches could be considered. First, one could find a way to fit a non linear model directly to the full SNP space. This solution would of course be computationally challenging as we would still be faced with over-fitting problems. The other solution would be to have a better biological prior than the current ones. In light of these elements, hypothesis 4 can not be confirmed:

> **Our non linear models did not improve risk score estimation by capturing SNP-SNP interaction**

5.3. Genetic and non genetic interactions

Non genetic factors, while being crucial to the analysis of obesity, are harder to interpret. Still, off the shelf models fitted on the data provide robust and clear associations. The higher metrics recorded for boosted trees show the benefit of adding complexity to models. It seems clear that a linear model alone cannot capture the full behavior of obesity with regard to non genetic factors. More over, the boosted tree gave a simple way to mine for associations with BMI for the next steps. Despite this good performance, a more comprehensive analysis of the model's results is needed. For instance, a collaboration with physicians could give a better understanding of relation between bio markers, environment and obesity. It is possible, that some bio marker variance is a direct consequence rather than a cause of obesity. Removing those from the model could give more informative results. Additionally, the study of variable interactions and effect is not sufficient at the moment. Partial dependence plot would give a clearer view at the effect of each variable. Applying clustering algorithms in pair with medical domain knowledge on the covariates could further improve the relevance of the model. Just as with the previous risk scores, a limitation of this study comes from the lack of independent datasets to test the results on. This poses a major challenge as some of the variables are measured using questionnaires that are self reported and study specific. Hence, while the results appear sensible and are robust within the UK bio-bank, other datasets could lead to different findings.

While the signal of environmental variables is noisy due to the nature of the data, it is much stronger than one from the polygenic risk score. Of course, this is partly due to the fact that the causal link between BMI and environment goes both way. Still, this highlights the importance of taking environment into account for this kind of study. Overall, adding genetic information has a little impact on the R^2. Still, in comparison to environmental variables, the polygenic risk score is relatively "important". It is on equal footing with factors commonly associated to obesity such as sedentarity and eating habits. However, coming to the conclusion that "when it comes to obesity, genetics is more important than dieting" is surely misleading. The importance as defined in the boosted tree is a rather limited metric that does not provide any quantitative information on interaction between variables. Instead the main take away, is simply that studying polygenic risk scores and environment at the same time is more informative than evaluating either one alone. To get a better picture at how the risk scores and environment interact, we would need to dig deeper into partial

dependence plots. In conclusion, given the metric improvement and the relative importance of the genetic risk score in the model hypothesis 5 is confirmed:

> **Studying non-genetic and genetic variables jointly improves association with BMI**

5.4. Variational model

Despite not improving performance of the polygenic risk score, the variational model works as intended. In terms of reconstruction, the model provides reasonable but improvable results. Since the reconstruction problem has multi class, multi dimensional and highly imbalanced properties the task is rather difficult. SNPs with high minor allele frequency are reconstructed with high uncertainty, SNPs with low minor allele frequencies on the other hand are sometimes only predicted as a single class. More worryingly, only extreme classes (0 and 2) are reconstructed fairly correctly. The 1s on the other hand are almost always uncertain. Still, the generative performance of the model is encouraging. Allele frequencies, the essential property of genetic data is conserved. Adding priors to the reconstruction to improve results was considered but not implemented. Such a prior would take the form of adding class weights equal to allele frequencies as measured on the 1000 genome project to the reconstruction loss. Of course, this would imply having separate priors for each population present in the cohort.

Results on the counterfactual experiment are harder to interpret. The SNPs "importance" are vastly different in the variational and in the linear model. Some SNPs with heavy weights in the risk scores show mild importance in this approach and vice versa. Still, a few top SNPs are consistent across models. This result is somewhat coherent with the findings of the boosted model. Non linear models give radically different SNP importances than linear models.

In terms of modeling biomarkers and SNPs jointly with environment variables, results are encouraging. Naturally, the association within this model in higher than in other models due to the fact that it includes much more features. As stated before, the goal of this experiment is not to improve performance but to check that the method can model mixed data types at once. From this perspective, adding continuous variables to the model does not affect the quality of the reconstructed SNPs. This is a bit surprising as we could excepted that biomarkers and environmental data had links to SNPs that could have improved the reconstructions. More over, both the distributions and correlations in sampled biomarkers are realistic. The link between individual variables and BMI is also well captured. In this setting, the method gives yet again different "importances" to different SNPs. While it is excepted that results from various models would not be exactly similar, findings given by boosted, linear and variational models are barely comparable even when the predictive performance is on the same order of magnitude. One reason to explain this discrepancy, is that individual SNPs have extremely

low effect, hence this estimation is sensitive to noise, methods, etc. While this may hold true for most SNPs, we could expect that models at least agree on the most and less important ones. In light of these results, this experiment points to the fact that it is possible to model the full system at once. With regards to 6 we confirm that:

> **The variational method can model biomarkers, SNPs, environment and trait jointly.**

A more general remark on the model must be made. While we were able to get some reasonable result with the model, it proved to be relatively fragile. Given the amount of hyper parameters to control, the low effects explored and the number of variables modelled, the robustness of the result is not guaranteed. In particular, hyper parameters search can often be automatized using algorithms such as HyperBand [42] to provide better and more consistent results. In this setting however, the number of tasks solved complicates this approach. Reconstruction, regression and generation tasks all use different metrics that are hard to combine in a meaningful one. Because of this absence of single metric to optimize, arbitrary choices must be made. Moreover, it is probable that this method is a "jack of all trades master of none" approach that does not provide the best results on either of the task. For this reason, the model showed more promising on better defined problems such as the MNIST dataset. In hindsight, given that linear risk scores already largely disagree with one another, betting on this method was a rather ambitious leap of faith. Still, despite these practical constraints the model provides an interesting framework for generative and investigable non linear modelling. Assuming that the model fits the data well enough endless simulation possibilities are offered. In particular, in our 2nd experiment, it could be possible to explore a large number of BMI-environment combinations and their effect on SNPs and biomarkers. In spite of these possibilities, a more established approach to interpreting deep learning models is SHAPELY [43] which can work with more flexible architectures.

6. Conclusion

Through this work on machine learning for genetic risks prediction we answered a number of questions. As illustrated by our results on existing risk scores, linking genetic information and complex traits is not straightforward. While current methods provides predictive models for traits, their performance, consistency and interpretability still call for major improvements. In particular, in the case of BMI current scores do not have the power for applications at individual level. Even simple and similar methods mostly disagree on the variants included in their models. Most importantly, the existing scores do not appear to fulfill their exploratory goal as they give little insights on biological processes leading to disease. While machine learning can be useful to improve some aspects of genomics, it is not the solution to everything. For this reason, we spent a great deal of time exploring if biological priors could be used instead of computational methods to reduce the dimensionality of the problem. Unfortunately this approach proved unsuccessful as most of our assumptions were not confirmed by available data. Instead of developing a full end-to-end pipeline, the rest of the work focused on exploring if machine learning methods could improve existing scores. Methods were motivated by the need for non linear and interpretable models. While machine learning models did not directly improve risk scores, they proved useful at bridging the gap between genetic and non genetic information. More specifically, the variational model implemented and developed for this thesis demonstrated that it was possible to model genetic variants, bio-markers and environmental factors using a single method.

Despite not directly improving risk scores prediction using machine learning, this thesis showcases the key aspects and challenges of genomics. While it is tempting to try and solve the problem using the latest deep learning models, it should be quite apparent by now that the biggest chunk of the problem is not computational. Simple and fully transparent methods do not appear to provide insights on biological processes leading to disease and do not come to a consensus on the causal variants. Hence, it is not so surprising that the alternative methods tested fail to a degree. Adding more and more sophisticated models seem to only add noise to the noise. Moreover the usual "get more data" advice is not only currently impractical but also doomed to fail as it would only solve the high dimensionality and not the linkage disequilibrium one. Instead, a better low level understanding of the genome seems required to select relevant variants. Unfortunately, existing gene expression models did not fulfill this role as expected. On another level the absence of attempts to link genetic risk scores to non-genetic variables in the analyzed studies is rather intriguing given the available knowledge on complex traits. Because regulation of gene expression is a likely pathway from genome to disease, environmental factors that also regulate genes can no longer be ignored. Finally, much of the challenges pointed out pointed by Manolio and Collins in 2009 are still

not addressed in the recent (2017-2020) studies analyzed in this thesis. In particular, the absence of diversity in available cohorts is not only ethically questionable, it limits the power of computational methods.

Some of the problems encountered in this thesis must be solved using alternative approaches. Single cell genomics [44] is among the popular ones. Instead of looking directly at full individuals and complex traits, this approach restricts itself to specific types of cells. While this approach does not directly link the genome to complex trait, it aims at understanding biological circuitry, a missing key element of the current analysis. Moreover, it allows for targeted and controlled experiments on specific functionality of cells. Another completely different but interesting approach is the analysis of trio studies [45]. While being smaller and harder to collect, this type of data provides causal guarantees on findings. Interestingly, it also circles back to historical ways of doing genomics by studying how traits where inherited within families.

In light of this element, studying genetic risks from the computational point of view only seems rather limited. Instead, leveraging machine learning to explore how genome, biological processes, complex traits, environmental factors interact together as a single system seems more promising. I tried at the modest scale of this thesis exploring this path.

A. Questionable GWAS

As explained in the introduction, genetic data is extremely sensitive and should be treated with great care. This study [46] is an example of how over interpretation can lead to debatable conclusions.

Three economists, none of which have formal training or experience in genomic conducted the study. The abstract announces: "We show that genetic endowments linked to educational attainment strongly and robustly predict wealth at retirement". This bold statement immediately calls for caution. Later, authors adopt a more cautious attitude throughout the article. As in any scientific work, the researchers mention some limitations to their genetic score referred to as "EA". They write: "One of the largest challenges in interpreting variation in the EA score comes from gene environment correlations." and add "Thus, an important limitation of our analyses is that we are not able to cleanly separate the association between the EA score and wealth into biological and environmental components". It appears that the authors are aware that they have multiple signals with potentially numerous confounding variables and spurious correlations. Furthermore when it comes to mechanistic interpretation they start off with "The evidence in this section is only suggestive" before stating "we do find a strong relationship between the EA score and portfolio choices".

The contrast between the disclaimers scattered in the article and the striking conclusions reveals a flaw of the study. The is more profound than potential technical flaws in their analysis. In fact, the authors display limits of their analysis and mention key challenges of genetics, yet they make extremely eye catching interpretations of their results. Whether what is written in this article is correct or false requires additional analysis [47] that others have made. Still, it is clear that arguing for a mechanistic link between variations in the genome and portfolio choices is more of a leap of faith than a sound scientific conclusion. Even at the cell level, mechanistic links between DNA and biological processes are still a research topic where much is to be uncovered.

From this article, the unsuspecting reader will only remember one thing from this study published in a prestigious journal: wealth is caused by genetic variations to a non negligible degree. All the author's comments on the limits of their analysis will be washed away by this astonishing but misleading statement. The topic of wealth distribution is extremely sensitive politically and ideologically driven. Such ideas, relayed in the general press [1] may drastically change society's conception of wealth and equality for the better or worse.

[1] `https://fortune.com/2017/02/28/rich-gene-billionaire/`

List of Figures

List of Tables

Bibliography

[1] T. A. Manolio, F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorff, D. J. Hunter, M. I. McCarthy, E. M. Ramos, L. R. Cardon, A. Chakravarti, J. H. Cho, A. E. Guttmacher, A. Kong, L. Kruglyak, E. Mardis, C. N. Rotimi, M. Slatkin, D. Valle, A. S. Whittemore, M. Boehnke, A. G. Clark, E. E. Eichler, G. Gibson, J. L. Haines, T. F. C. Mackay, S. A. McCarroll, and P. M. Visscher. "Finding the missing heritability of complex diseases". eng. In: *Nature* 461.7265 (2009). 19812666[pmid], pp. 747–753. ISSN: 1476-4687. DOI: 10.1038/nature08494. URL: https://pubmed.ncbi.nlm.nih.gov/19812666.

[2] URL: https://www.genome.gov/about-genomics/fact-sheets/A-Brief-Guide-to-Genomics.

[3] URL: https://www.genome.gov/genetics-glossary/Gene-Expression.

[4] URL: https://www.genome.gov/genetics-glossary/Gene-Regulation.

[5] URL: https://medlineplus.gov/genetics/understanding/basics/noncodingdna/.

[6] A. Auton, G. R. Abecasis, D. M. Altshuler, et al. "A global reference for human genetic variation". In: *Nature* 526.7571 (2015), pp. 68–74. ISSN: 1476-4687. DOI: 10.1038/nature15393. URL: https://doi.org/10.1038/nature15393.

[7] C. Sudlow, J. Gallacher, N. Allen, V. Beral, P. Burton, J. Danesh, P. Downey, P. Elliott, J. Green, M. Landray, B. Liu, P. Matthews, G. Ong, J. Pell, A. Silman, A. Young, T. Sprosen, T. Peakman, and R. Collins. "UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age". In: *PLOS Medicine* 12.3 (Mar. 2015), pp. 1–10. DOI: 10.1371/journal.pmed.1001779. URL: https://doi.org/10.1371/journal.pmed.1001779.

[8] R. Holle, M. Happich, H. Löwel, and H. E. Wichmann. "KORA–a research platform for population based health research". In: *Gesundheitswesen* 67 Suppl 1 (2005), pp. 19–25.

[9] A. N. Barbeira, M. Pividori, J. Zheng, H. E. Wheeler, D. L. Nicolae, and H. K. Im. "Integrating predicted transcriptome from multiple tissues improves association detection". In: *PLOS Genetics* 15.1 (Jan. 2019), pp. 1–20. DOI: 10.1371/journal.pgen.1007889. URL: https://doi.org/10.1371/journal.pgen.1007889.

[10] A. N. Barbeira, R. Bonazzola, E. R. Gamazon, Y. Liang, Y. Park, S. Kim-Hellmuth, G. Wang, Z. Jiang, D. Zhou, F. Hormozdiari, B. Liu, A. Rao, A. R. Hamel, M. D. Pividori, F. Aguet, L. Bastarache, D. M. Jordan, M. Verbanck, R. Do, M. Stephens, K. Ardlie, M. McCarthy, S. B. Montgomery, A. V. Segrè, C. D. Brown, T. Lappalainen, X. Wen, and H. K. Im. "Widespread dose-dependent effects of RNA expression and splicing on complex diseases and traits". In: *bioRxiv* (2019). Ed. by F. Aguet, K. Ardlie, A. N.

Barbeira, et al. DOI: 10.1101/814350. eprint: https://www.biorxiv.org/content/early/2019/10/22/814350.full.pdf. URL: https://www.biorxiv.org/content/early/2019/10/22/814350.

[11] S. A. Lambert, L. Gil, S. Jupp, S. C. Ritchie, Y. Xu, A. Buniello, A. McMahon, G. Abraham, M. Chapman, H. Parkinson, J. Danesh, J. A. L. MacArthur, and M. Inouye. "The Polygenic Score Catalog as an open database for reproducibility and systematic evaluation". In: *Nature Genetics* 53.4 (2021), pp. 420–425. ISSN: 1546-1718. DOI: 10.1038/s41588-021-00783-5. URL: https://doi.org/10.1038/s41588-021-00783-5.

[12] S. W. Choi, T. S.-H. Mak, and P. F. O'Reilly. "Tutorial: a guide to performing polygenic risk score analyses". In: *Nature Protocols* 15.9 (2020), pp. 2759–2772. ISSN: 1750-2799. DOI: 10.1038/s41596-020-0353-1. URL: https://doi.org/10.1038/s41596-020-0353-1.

[13] A. Blum, J. Hopcroft, and R. Kannan. *Foundations of Data Science*. Cambridge University Press, 2020. DOI: 10.1017/9781108755528.

[14] J. Novembre, T. Johnson, K. Bryc, Z. Kutalik, A. R. Boyko, A. Auton, A. Indap, K. S. King, S. Bergmann, M. R. Nelson, M. Stephens, and C. D. Bustamante. "Genes mirror geography within Europe". eng. In: *Nature* 456.7218 (2008). 18758442[pmid], pp. 98–101. ISSN: 1476-4687. DOI: 10.1038/nature07331. URL: https://pubmed.ncbi.nlm.nih.gov/18758442.

[15] A. L. Price, N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, and D. Reich. "Principal components analysis corrects for stratification in genome-wide association studies". In: *Nature Genetics* 38.8 (2006), pp. 904–909. ISSN: 1546-1718. DOI: 10.1038/ng1847. URL: https://doi.org/10.1038/ng1847.

[16] E. L. Lehmann and J. P. Romano. *Testing statistical hypotheses*. Third. Springer Texts in Statistics. New York: Springer, 2005, pp. xiv+784. ISBN: 0-387-98864-5.

[17] G. M. Clarke, C. A. Anderson, F. H. Pettersson, L. R. Cardon, A. P. Morris, and K. T. Zondervan. "Basic statistical analysis in genetic case-control studies". eng. In: *Nature protocols* 6.2 (2011). 21293453[pmid], pp. 121–133. ISSN: 1750-2799. DOI: 10.1038/nprot.2010.182. URL: https://pubmed.ncbi.nlm.nih.gov/21293453.

[18] R. Tibshirani. *Sparsity, the Lasso, and Friends*. URL: https://www.stat.cmu.edu/~ryantibs/statml/lectures/sparsity.pdf.

[19] J. Qian, Y. Tanigawa, W. Du, M. Aguirre, C. Chang, R. Tibshirani, M. A. Rivas, and T. Hastie. "A Fast and Scalable Framework for Large-scale and Ultrahigh-dimensional Sparse Regression with Application to the UK Biobank". In: *bioRxiv* (2020). DOI: 10.1101/630079. eprint: https://www.biorxiv.org/content/early/2020/05/31/630079.1.full.pdf. URL: https://www.biorxiv.org/content/early/2020/05/31/630079.1.

[20] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. New York, NY, USA: Springer New York Inc., 2001.

[21] D. P. Kingma, D. J. Rezende, S. Mohamed, and M. Welling. *Semi-Supervised Learning with Deep Generative Models*. 2014. arXiv: 1406.5298 [cs.LG].

[22] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. "An Introduction to Variational Methods for Graphical Models". In: *Machine Learning* 37.2 (1999), pp. 183–233. ISSN: 1573-0565. DOI: 10.1023/A:1007665907178. URL: https://doi.org/10.1023/A:1007665907178.

[23] K. Sohn, H. Lee, and X. Yan. "Learning Structured Output Representation using Deep Conditional Generative Models". In: *Advances in Neural Information Processing Systems*. Ed. by C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett. Vol. 28. Curran Associates, Inc., 2015. URL: https://proceedings.neurips.cc/paper/2015/file/8d55a249e6baa5c06772297520da2051-Paper.pdf.

[24] J. Paisley, D. Blei, and M. Jordan. *Variational Bayesian Inference with Stochastic Search*. 2012. arXiv: 1206.6430 [cs.LG].

[25] D. P. Kingma and M. Welling. *Auto-Encoding Variational Bayes*. 2014. arXiv: 1312.6114 [stat.ML].

[26] C. P. Burgess, I. Higgins, A. Pal, L. Matthey, N. Watters, G. Desjardins, and A. Lerchner. *Understanding disentangling in β-VAE*. 2018. arXiv: 1804.03599 [stat.ML].

[27] S. Ioffe and C. Szegedy. *Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift*. 2015. arXiv: 1502.03167 [cs.LG].

[28] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. "PyTorch: An Imperative Style, High-Performance Deep Learning Library". In: *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 2019, pp. 8024–8035. URL: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

[29] D. P. Kingma and J. Ba. *Adam: A Method for Stochastic Optimization*. 2017. arXiv: 1412.6980 [cs.LG].

[30] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. Software available from tensorflow.org. 2015. URL: https://www.tensorflow.org/.

[31] C. Louizos, U. Shalit, J. Mooij, D. Sontag, R. Zemel, and M. Welling. *Causal Effect Inference with Deep Latent-Variable Models*. 2017. arXiv: 1705.08821 [stat.ML].

[32] Y. LeCun and C. Cortes. "MNIST handwritten digit database". In: (2010). URL: http://yann.lecun.com/exdb/mnist/.

[33] F. Privé, J. Arbel, and B. J. Vilhjálmsson. "LDpred2: better, faster, stronger". In: *bioRxiv* (2020). DOI: 10.1101/2020.04.28.066720. eprint: https://www.biorxiv.org/content/early/2020/06/25/2020.04.28.066720.full.pdf. URL: https://www.biorxiv.org/content/early/2020/06/25/2020.04.28.066720.

[34] S. W. Choi and P. F. OReilly. "PRSice-2: Polygenic Risk Score software for biobank-scale data". In: *GigaScience* 8.7 (July 2019). DOI: 10.1093/gigascience/giz082. URL: https://doi.org/10.1093/gigascience/giz082.

[35] A. V. Khera, M. Chaffin, K. H. Wade, S. Zahid, J. Brancale, R. Xia, M. Distefano, O. Senol-Cosar, M. E. Haas, A. Bick, K. G. Aragam, E. S. Lander, G. D. Smith, H. Mason-Suares, M. Fornage, M. Lebo, N. J. Timpson, L. M. Kaplan, and S. Kathiresan. "Polygenic Prediction of Weight and Obesity Trajectories from Birth to Adulthood". In: *Cell* 177.3 (Apr. 2019), 587–596.e9. DOI: 10.1016/j.cell.2019.03.028. URL: https://doi.org/10.1016/j.cell.2019.03.028.

[36] M. Song, Y. Zheng, L. Qi, F. B. Hu, A. T. Chan, and E. L. Giovannucci. "Longitudinal Analysis of Genetic Susceptibility and BMI Throughout Adult Life". In: *Diabetes* 67.2 (Dec. 2017), pp. 248–255. DOI: 10.2337/db17-1156. URL: https://doi.org/10.2337/db17-1156.

[37] N. Chami, M. Preuss, R. W. Walker, A. Moscati, and R. J. F. Loos. "The role of polygenic susceptibility to obesity among carriers of pathogenic mutations in MC4R in the UK Biobank population". In: *PLOS Medicine* 17.7 (July 2020). Ed. by K. Clément, e1003196. DOI: 10.1371/journal.pmed.1003196. URL: https://doi.org/10.1371/journal.pmed.1003196.

[38] T. G. Richardson, E. Sanderson, B. Elsworth, K. Tilling, and G. D. Smith. "Use of genetic variation to separate the effects of early and later life adiposity on disease risk: mendelian randomisation study". In: *BMJ* (May 2020), p. m1203. DOI: 10.1136/bmj.m1203. URL: https://doi.org/10.1136/bmj.m1203.

[39] T. Xie, B. Wang, I. M. Nolte, P. J. van der Most, A. J. Oldehinkel, C. A. Hartman, and H. Snieder. "Genetic Risk Scores for Complex Disease Traits in Youth". In: *Circulation: Genomic and Precision Medicine* 13.4 (Aug. 2020). DOI: 10.1161/circgen.119.002775. URL: https://doi.org/10.1161/circgen.119.002775.

[40] T. Chen and C. Guestrin. "XGBoost: A Scalable Tree Boosting System". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '16. San Francisco, California, USA: ACM, 2016, pp. 785–794. ISBN: 978-1-4503-4232-2. DOI: 10.1145/2939672.2939785. URL: http://doi.acm.org/10.1145/2939672.2939785.

[41] K. A. Fawcett and I. Barroso. "The genetics of obesity: FTO leads the way". eng. In: *Trends in genetics : TIG* 26.6 (2010). 20381893[pmid], pp. 266–274. ISSN: 0168-9525. DOI: 10.1016/j.tig.2010.02.006. URL: https://pubmed.ncbi.nlm.nih.gov/20381893.

[42] L. Li, K. G. Jamieson, G. DeSalvo, A. Rostamizadeh, and A. Talwalkar. "Efficient Hyperparameter Optimization and Infinitely Many Armed Bandits". In: *CoRR* abs/1603.06560 (2016). arXiv: 1603.06560. URL: http://arxiv.org/abs/1603.06560.

[43] S. M. Lundberg and S.-I. Lee. "A Unified Approach to Interpreting Model Predictions". In: *Advances in Neural Information Processing Systems 30*. Ed. by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Curran Associates, Inc., 2017, pp. 4765–4774. URL: http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf.

[44] T. Kalisky and S. R. Quake. "Single-cell genomics". In: *Nature Methods* 8.4 (Mar. 2011), pp. 311–314. DOI: 10.1038/nmeth0411-311. URL: https://doi.org/10.1038/nmeth0411-311.

[45] S. Bates, M. Sesia, C. Sabatti, and E. Candès. "Causal inference in genetic trio studies". In: *Proceedings of the National Academy of Sciences* 117.39 (2020), 24117–24126. ISSN: 1091-6490. DOI: 10.1073/pnas.2007743117. URL: http://dx.doi.org/10.1073/pnas.2007743117.

[46] D. Barth, N. W. Papageorge, and K. Thom. *Genetic Endowments and Wealth Inequality*. Working Paper 24642. National Bureau of Economic Research, 2018. DOI: 10.3386/w24642. URL: http://www.nber.org/papers/w24642.

[47] S. E. Black, P. J. Devereux, P. Lundborg, and K. Majlesi. "Poor Little Rich Kids? The Role of Nature versus Nurture in Wealth and Other Economic Outcomes and Behaviours". In: *The Review of Economic Studies* 87.4 (July 2019), pp. 1683–1725. ISSN: 0034-6527. DOI: 10.1093/restud/rdz038. eprint: https://academic.oup.com/restud/article-pdf/87/4/1683/33461634/rdz038.pdf. URL: https://doi.org/10.1093/restud/rdz038.