Of Zeros and Ones

Rauschenberger, A.

2021

document version
Publisher's PDF, also known as Version of record

Chapter 1

Introduction

Genetic, epigenetic and environmental factors determine the observable characteristics of an individual. This process involves multiple molecular profiles, and includes gene expression as an intermediary step. In this thesis, we propose novel statistical methods for analysing the connection between genotype and phenotype. Each chapter is self-contained, but the introduction gives some insights into molecular biology and high-dimensional statistics.

1.1 Molecular biology

The human body consists of trillions of cells, of which each contains the complete genetic information of the whole organism. This information, the deoxyribonucleic acid (DNA), is a sequence of about three billion nucleobases, namely adenine, cytosine, guanine, and thymine. The DNA is structured as a double helix of two complementary strands of nucleotides. Since adenine binds with thymine, and cytosine binds with guanine, knowledge of one strand implies knowledge of the other strand. In somatic cells, the DNA is normally organised in 22 autosome pairs and two sex chromosomes, that is 23 chromosome pairs in total.

All individuals have a unique DNA sequence, except monozygotic twins. Mutation and recombination, occurring naturally or through external influences, cause this genetic variation. Types of genetic variation include single-nucleotide polymorphisms (SNPs), insertion or deletion of bases (indels), copy number variation (CNV), and chromosomal rearrangements. Genetic variation in germ cells, but not in somatic cells, is heritable.

In the DNA sequence, triplets of base pairs form codons, which encode amino acids, and groups of codons form genes, which encode proteins. There are approximately 20 000 protein-coding genes in the human genome. Only a small fraction of the genome consists of protein-coding genes, the rest is non-coding DNA. Genes have regions that can be expressed (exons) and regions that cannot be

expressed (introns). All introns and some exons are removed during transcription. Through alternative splicing, the differential inclusion or exclusion of exons, a single gene can encode multiple proteins.

Although almost all cells have the same genes, they perform different functions. This is possible due to the regulation of gene expression: genes can be switched on and off, or expressed at different levels. Gene expression is the most fundamental phenotype, and it links the genotype to other phenotypes. The process of gene expression involves the transcription from DNA to ribonucleic acid (RNA). It is possible to measure exon abundance by RNA sequencing (RNA-Seq). The expression of a gene equals the sum of the abundances of the included exons.

According to the central dogma of molecular biology, genotype transcribes to gene expression, and gene expression translates to protein. Determining the function of cells, proteins contribute to the phenotype – the observable characteristics – of an individual. The epigenotype affects the transcription and translation process from genotype to protein, without altering the DNA sequence. Epigenetic variation includes DNA methylation, histone modification, and RNA interference. Epigenetic factors are heritable, but they are influenced by environmental factors.

Genome-wide association (GWA) studies search for SNPs with an effect on a quantitative trait, so-called quantitative trait loci (QTLs). Mendelian traits are determined by a single variant, but complex traits are determined by multiple genetic, epigenetic or environmental factors. If individual SNP effects on the trait are weak, it can help to dissect them into effects of SNPs on gene expression, and effects of gene expression on the trait. A SNP with an effect on the absolute expression level of a gene is an expression quantitative trait locus (eQTL), and a SNP with an effect on the relative expression levels of the exons in the gene is a splicing quantitative trait locus (sQTL). The expression or splicing of a gene might be associated with the complex trait.

1.2 High-dimensional statistics

Many research questions concern effects of covariates on a response. We are interested in the information flow from (epi)genetic factors through gene expression to other quantitative traits. This involves effects of (1) (epi)genetic factors on gene expression (eQTL and sQTL studies), (2) (epi)genetic factors on other quantitative traits (GWA studies), and (3) gene expression on other quantitative traits. Accordingly, we model (epi)genetic factors as covariates, gene expression as either covariates or responses, and other quantitative traits as responses. We want to test whether the response is associated with the covariates, or to predict the response from the covariates. With increasing levels of complexity, we could examine the effect(s) of one covariate on a univariate response, multiple covariates on a univariate response, or multiple covariates on a multivariate response.

Assume data are available for n samples, one response and p covariates. Statistics for genomics does not primarily deal with low-dimensional ($p \ll n$) but with high-dimensional ($p \gg n$) settings. This inversion of dimensionality, with more covariates than samples, creates challenges and opportunities. If the dimensionality of the data increases, the data may contain more information, but it may become more difficult to extract this information. Many methods designed for low dimensionality perform poorly or break down under high dimensionality. Therefore, settings with more covariates than samples require novel statistical methods.

Simple regression allows us to analyse one covariate at a time. Using the generalised linear model framework (McCullagh and Nelder, 1989), we could regress the response on each covariate separately, to calculate the marginal statistical significance or to compute the individual predictivity of the covariates. If there are few covariates, we may conclude that a significant or predictive covariate has an effect on the response. However, if there are many covariates, some covariates will be highly significant or highly predictive by chance. We would need to adjust for multiple comparisons before concluding that a covariate has an effect on the response.

Multiple regression allows us to analyse all covariates at once. A single model can represent the effects of multiple covariates on the response. In low-dimensional settings, we can estimate such models by maximum likelihood, and then test the individual or joint significance of covariates, or predict the response from the covariates. However, we cannot estimate high-dimensional models by classical maximum likelihood. An alternative is to assume the regression coefficients follow a probability distribution. Both the global test (Goeman et al., 2004) and penalised regression (Zou and Hastie, 2005) have such a Bayesian interpretation. Then we can test whether at least one covariate has an effect on the response (Chapters 2 and 4), or predict the response from the covariates (Chapters 5 and 6).

The generalised linear model (McCullagh and Nelder, 1989) provides a framework for high-dimensional regression. It models the response, specifically the expected value of the linear predictor, as a linear function of the covariates. The link function renders this framework flexible, because it allows for responses from different families (e.g. Gaussian, binomial, Poisson), and the coefficients render this framework interpretable, because they represent the effects of the covariates on the linear predictor.

Prediction models should not memorise but generalise. Suppose we want to use gene expression data to distinguish between patients and controls. It is simple to fit a model that perfectly splits the observed samples into patients and controls, but it is more difficult to fit a model for classifying previously unseen samples. The aim is to estimate the coefficients, the effects of the covariates on the linear predictor, from some samples, and then use these estimates to predict the response of other samples.

Overfitting is a major problem in high-dimensional regression. Since there are more covariates than samples, we could estimate their coefficients in such a way that the observed covariates explain all variation of the observed response. Then the model fits well to the present data, but will fit badly to future data. The elastic net penalty (Zou and Hastie, 2005) prevents overfitting by shrinking the

coefficients towards zero.

The regularisation parameter of the elastic net determines the amount of shrinkage. There is underfitting (the model adapts too little to the data) if the coefficients are shrunken too much towards zero, and there is overfitting (the model adapts too much to the data) if the coefficients are shrunken too little towards zero. The optimal amount of shrinkage is usually determined by cross-validation.

The elastic net penalty combines the ridge and lasso penalties. While ridge penalises the squared coefficients, the lasso penalises the absolute coefficients. Both penalties shrink the coefficients towards zero, but there is an interesting difference. While ridge leads to dense models, i.e. all estimated coefficients are different from zero, the lasso leads to sparse models, i.e. only some estimated coefficients are different from zero. Thus, in contrast to ridge, the lasso performs variable selection. The advantage of models with few non-zero coefficients is that they are easy to interpret. One challenge in high-dimensional settings is to search for the most predictive model of a given size (Chapter 5).

Ensemble learning combines different models to make predictions, e.g. through stacking (Wolpert, 1992) or bagging (Breiman, 1996). In regression analysis, we have the covariates on the one hand and the response on the other hand. The coefficients represent the effects of the covariates on the response, with interpretable effect directions (signs) and effect sizes (absolute values). Ensemble learning includes a hidden layer between the covariates and the response. We first predict the response multiple times and then combine the predictions. Instead of a single direct effect, each covariate has multiple indirect effects on the response. This makes ensemble models typically more predictive but less interpretable than regression. But it is also possible to design interpretable ensemble models (Chapter 6).

Neural networks can include multiple hidden layers between the covariates and the response. The information flows from the covariates to the first hidden layer, from each hidden layer to the next hidden layer, and then from the last hidden layer to the response. Non-linear activation functions allow neural networks to capture

non-linear effects of the covariates on the response. Neural networks might have advantages in predictivity but have disadvantages in interpretability ("black-box" models). For clinical prediction problems with omics data, however, more interpretable models might be equally predictive.

1.3 Experimental data

Molecular data are not only high-dimensional but also have special properties. Typically, they are strongly correlated and contain much noise. Although covariate groups also appear in low-dimensional settings, such as dummy variables for different levels of a categorical variable, they are more relevant in high-dimensional settings. If there are many covariates, some covariates may form natural groups, because of the experimental design (Goeman et al., 2004) or external information (van de Wiel et al., 2016). One example are paired covariate settings (Chapter 5). If multiple molecular profiles are available, each profile leads to one variable group, and the integration of multiple molecular profiles may improve the understanding of genotype-phenotype relationships. Accounting for multiple molecular profiles can improve testing (Menezes et al., 2016, Chapter 2) and prediction (van de Wiel et al., 2016; Boulesteix et al., 2017). Besides, sharing information among related variables may be beneficial (Robinson and Smyth, 2007; van de Wiel et al., 2013, Chapters 2, 3 and 5). In random effect models, the results depend on the scaling of the covariates. Therefore, standardisation of covariates influences the global test (Goeman et al., 2004) and penalised regression (Zou and Hastie, 2005). Covariates which are significant or predictive are not always the same (Lo et al., 2015). This difference might be more important for covariate groups in high-dimensional settings, because a stronger correlation among the covariates, if not accounted for, increases significance but decreases predictivity.

Sequencing experiments lead to count variables with large variances and frequent zeros. Many statistical methods assume the

response follows a Gaussian distribution. Since sequencing data violate the Gaussian assumption, we must either adapt the data to the distribution (Law et al., 2014) or adapt the distribution to the data (Robinson and Smyth, 2007; van de Wiel et al., 2013). Also after adjusting the raw counts for different library sizes, the pseudo-counts retain the characteristic mean-variance relationship. Although the Poisson distribution may capture technical variability, biological variability may lead to overdispersion (Anders and Huber, 2010; Robinson et al., 2010). The negative binomial distribution has a more flexible mean-variance relationship, which allows the variance to exceed the mean. Even more flexible is the zero-inflated negative binomial distribution (van de Wiel et al., 2013). If a test assuming the Poisson distribution is applied to data following the (zero-inflated) negative binomial distribution, it might make too many errors. Many methods exploit the distribution of the response, but few methods exploit the distribution of the covariates. However, a special treatment of zero counts in the covariates might improve the global test and penalised regression (cf. Telonis et al., 2017, Chapter 5).

1.4 Overview

The five main chapters of this thesis present novel methods for association (Chapters 2, 3 and 4) and prediction (Chapters 5 and 6) in genomics, followed by a discussion of common aspects of these methods (Chapter 7). Each main chapter corresponds to one scientific paper:

Chapter 2 (Rauschenberger et al., 2016, "`globalSeq`") presents the global test for sequencing data. Testing for association between RNA-Seq and other genomic data is challenging due to high variability of the former and high dimensionality of the latter. Using the negative binomial distribution and a random-effects model, we develop an omnibus test that overcomes both difficulties. It may be

conceptualised as a test of overall significance in regression analysis, where the response variable is overdispersed and the number of explanatory variables exceeds the sample size. The proposed test can detect genetic and epigenetic alterations that affect gene expression. It can examine complex regulatory mechanisms of gene expression.

- Rauschenberger, A., Jonker, M.A., van de Wiel, M.A. & Menezes, R.X. (2016). "Testing for association between RNA-Seq and high-dimensional data." *BMC Bioinformatics*, 17(1):118. doi: 10.1186/s12859-016-0961-5.

- R package `globalSeq`: "Negative Binomial Global Test". `https://bioconductor.org/packages/globalSeq/` `https://github.com/rauschenberger/globalSeq/`

Chapter 3 (Rauschenberger et al., 2020b, "`semisup`") is about semi-supervised mixture modelling for detecting genotype-phenotype associations. Suppose we want to detect single nucleotide polymorphisms (markers) with main or interactive effects on a quantitative trait, which is a challenging problem due to the high dimensionality. Analysing one marker at a time, we split the individuals into two groups, based on the number of minor alleles. If the quantitative trait differs in mean between the two groups, the marker has a main effect. If the quantitative trait differs in distribution between *some* individuals in one group and *all* other individuals, it has a main or an interactive effect. We propose a mixture test to detect both types of effects, without distinguishing them. In simulations and applications, we show that the proposed test is statistically powerful, maintains the type I error rate, and detects meaningful signals.

- Rauschenberger, A., Menezes, R.X., van de Wiel, M.A., van Schoor, N.M. & Jonker, M.A. (2020). "Semi-supervised mixture test for detecting markers associated with a quantitative trait." *In preparation.*

- R package `semisup`: "Semi-Supervised Mixture Test".
  `https://bioconductor.org/packages/semisup/`
  `https://github.com/rauschenberger/semisup/`

Chapter 4 (Menezes et al., 2020, "`spliceQTL`") presents the global test for association between alternative splicing and the genotype. Instead of testing for association between (i) the absolute expression level of a gene or exons in a gene and (ii) SNPs in or around the gene (eQTLs), we want to test for association between (i) the relative expression levels of the exons in a gene and (ii) SNPs in or around the gene (sQTLs). The test accounts for the absolute expression level of the gene and the correlation between SNPs. It renders one p-value for each gene and all SNPs (not one p-value for each exon-SNP combination) and thereby decreases the multiple testing burden.

- Menezes, R.X., Rauschenberger, A., 't Hoen, P.A.C. & Jonker, M.A. (2020). "A powerful test for spliceQTL effects". *In preparation.*

- R package `spliceQTL`: "Multinomial Global Test".
  `https://github.com/rauschenberger/spliceQTL`

Chapter 5 (Rauschenberger et al., 2019, "`palasso`") introduces the paired lasso: a generalisation of the lasso for paired covariate settings. Our aim is to predict a single response from two high-dimensional covariate sets. We assume a one-to-one correspondence between the covariate sets, with each covariate in one set forming a pair with a covariate in the other set. Paired covariates arise, for example, when two transformations of the same data are available. It is often unknown which of the two covariate sets leads to better predictions, or whether the two covariate sets complement each other. The paired lasso addresses this problem by weighting the covariates to improve the selection from the covariate sets and the covariate

pairs. It thereby combines information from both covariate sets and accounts for the paired structure. We tested the paired lasso on more than 2 000 classification problems with experimental genomics data, and found that for estimating sparse but predictive models, the paired lasso outperforms the standard and the adaptive lasso.

- Rauschenberger, A., Ciocănea-Teodorescu, I., Jonker, M.A., Menezes, R.X. & van de Wiel, M.A. (2019). "Sparse classification with paired covariates." *Advances in Data Analysis and Classification*. doi: 10.1007/s11634-019-00375-6.

- R package `palasso`: "Paired Lasso".
  `https://`CRAN.R`-project.org/package=palasso`
  `https://github.com/rauschenberger/palasso/`

Chapter 6 (Rauschenberger et al., 2020a, "`starnet`") introduces the stacked elastic net. Machine learning in the biomedical sciences should ideally provide predictive and interpretable models. When predicting outcomes from clinical or molecular features, applied researchers often want to know which features have effects, whether these effects are positive or negative, and how strong these effects are. Regression analysis includes this information in the coefficients but typically renders less predictive models than more advanced machine learning techniques. Here we propose an interpretable meta-learning approach for high-dimensional regression. The elastic net provides a compromise between estimating weak effects for many features and strong effects for some features. It has a mixing parameter to weight between ridge and lasso regularisation. Instead of selecting one weighting by tuning, we combine multiple weightings by stacking. We do this in a way that increases predictivity without sacrificing interpretability.

- Rauschenberger, A., Glaab, E. & van de Wiel, M.A. (2020). "Predictive and interpretable models via the stacked elastic net." *Bioinformatics*, btaa535. doi: 10.1093/bioinformatics/btaa535.

- R package `starnet`: "Stacked Elastic Net".
  `https://CRAN.R-project.org/package=starnet`
  `https://github.com/rauschenberger/starnet/`