Chapter 7

Discussion

7.1 Common themes

While the first part of this thesis is about association (`globalSeq`, `spliceQTL`, `semisup`), the second part is about prediction (`palasso`, `starnet`). Although the proposed methods address different problems, they are closely related. They all deal with multiple effects, either effects of multiple covariates (`globalSeq`, `spliceQTL`, `palasso`, `starnet`) or multiple effects of one covariate (`semisup`, `palasso`). And they deal with specific effect types, including global effects (`globalSeq`, `spliceQTL`), main and interactive effects (`semisup`), linear and non-linear effects (`palasso`), or dense and sparse effects (`starnet`). The proposed methods are suitable for analysing SNP genotype data, because they can account for the strong correlation between SNPs (`globalSeq`, `spliceQTL`) and the binary/ternary nature of SNPs (`semisup`, `palasso`). And they are suitable for analysing RNA-Seq gene expression data, because they can model the absolute expression level as a response (`globalSeq`, `semisup`), the relative exon abundance as a response (`spliceQTL`), a numerical and a binary representation of gene expression in the covariates (`palasso`), and dense or sparse effects of gene expression (`starnet`).

7.2 Performance evaluation

The development of statistical methods includes the evaluation of their performance. However, for hypothesis tests and prediction models, we do not measure but estimate the performance.

An ideal *hypothesis test* rejects all false hypotheses and fails to reject all true hypotheses. If we knew the truth, we could count the number of true positives, true negatives, false positives (type I errors), and false negatives (type II errors). Given a false positive rate, we could measure the false negative rate. This works if we know the truth, but not if we do not know it. Although we typically do not know the truth in experiments, we know the truth in simulations. If the simulated data realistically reflects the experimental data, the measured performance in the simulation is a reliable estimate of the performance in the experiment.

An ideal *prediction model* correctly predicts unknown responses. If we knew the truth, we could quantify the differences between the predicted and the observed values with a loss function, for which the choice depends on the error distribution. Although we cannot calculate the performance for samples with an unknown response, we can repeatedly hide known responses from the prediction model, and then compare them to the predictions (cross-validation). If the available data are representative of the future data, and the excluded samples are independent of the included samples, the measured performance for the available data is a reliable estimate of the performance for future data.

7.3 Approximate permutation testing

We implemented `globalSeq`, `spliceQTL` and `semisup` as approximate permutation tests, because the asymptotic distributions of their test statistics are unknown. Each permutation involves shuffling the samples (resampling without replacement). If the sample size is large, we cannot conduct exact permutation tests, which require all possible orderings of the samples. An approximate alternative is to draw a random sample from the permutations, either with or without replacement. If we permute many times, some orderings of the samples might occur more than once, but it would require much bookkeeping to prevent repetitions.

If p-values are to be corrected for multiple testing, permutation tests suffer from "diseconomies of scale", meaning that the computation time of a single test increases with the number of tests. This is because more tests imply a more stringent multiple testing correction, while more permutations are necessary to reach the lower adjusted significance level. Different strategies exist to shift computational resources from insignificant to significant p-values. A generic solution is to continue permutation while the significance level is attainable, and interrupt permutation when it is unattainable.

Approximate permutation tests lead to estimated p-values, of which the precision increases with the number of permutations. If the estimated p-value is close to the significance level and has a low precision, it may be on the other side of the significance level than the underlying p-value. Instead of ignoring this uncertainty, we could calculate confidence intervals around the estimated p-values (van Wieringen et al., 2008).

7.4 Distribution of p-values

The analysis of high-dimensional data often generates numerous p-values (Chapters 2, 3 and 4). One possibility is to examine one p-value at a time, compare it with the adjusted significance level, and decide whether it is significant or not. Another possibility is to examine all p-values at once, by displaying them in a Manhattan plot or a heatmap, and by plotting their empirical cumulative distribution. Similar to the global test (Chapters 2 and 4), the distribution of p-values summarises information on multiple effects. We believe that the distribution of p-values is often more informative than individual p-values. For example, several insignificant but relatively low p-values might indicate a stronger signal than one significant p-value. If the p-values follow a standard uniform distribution, we may conclude the signal is sparse or weak. If their distribution is positively skewed, with more p-values close to zero than expected under the uniform distribution, there is evidence against the null hypotheses. Unknown patterns might indicate overfitting or omitted-variable bias (cf. van Iterson et al., 2010).

7.5 Variable selection

We can identify the most significant or predictive covariates by pairwise testing or simple regression, respectively, and the most influential covariates in global testing or ridge regression from the contributions to the global test statistic or the estimated regression coefficients, respectively. Especially if the data are high-dimensional, the selected covariates may be strongly correlated. Although they are individually more informative than other covariates, they probably do not form the most informative covariate group, because of the overlapping information. Variable selection is a compromise between modelling all covariates at once and each covariate separately, reducing the number of covariates while ideally loosing as little information as possible. Variable selection methods can be classified into filter, wrapper and embedded methods (Saeys et al., 2007). A wrapper method (e.g. genetic algorithm) or an embedded method (e.g. lasso) might select the most *predictive* covariate group of a given size, but special care is required for the most *significant* covariate group. Similarly to canonical correlation analysis, we might want to maximise the correlation between a response and a combination of covariates. However, calculating a p-value for the covariate group is non-trivial because of variable selection. Neither *predictive* nor *significant* variables are necessarily causal. To provide some evidence for causal relationships based on observational high-dimensional data, more sophisticated analyses are required (e.g. fine-mapping, graphical modelling).

7.6 Association and prediction

When examining the effects of covariates on a response, testing and predicting are different but related problems. In low-dimensional regression, we can test the individual or joint significance of covariates with the Wald test or the likelihood-ratio test, respectively, but high-dimensional regression requires special methods for significance testing (e.g. Cule et al., 2011; Bühlmann, 2013; Lockhart et al., 2014).

Although significance and predictivity are different, some problems may suggest testing in low dimensions but prediction in high dimensions. Instead of testing whether the alternative model fits significantly better to the data than the null model, we could test whether the alternative model leads to significantly better predictions than the null model. van de Wiel et al. (2009) proposed a generic framework for testing the prediction error difference.

Combining association and prediction could improve our understanding of biological processes. Suppose multiple molecular profiles are available for patients with and without a disease, and we want to predict a clinical outcome for patients with the disease. If we combine gene expression with genetic or epigenetic profiles, gene expression will likely dominate the predictions. Instead, we could predict the clinical outcome from gene expression only, while improving the estimation with other molecular data. One possibility is a two-step procedure, which first identifies groups of genes, and then predicts from the genes. In the first step, we could use patients with and without the disease to test for association between the expression of each gene and local (epi)genetic alterations. If the expression of some genes is significantly associated with local alterations in one or more molecular profiles, they might tend to be more important in determining the clinical outcome than other genes. In the second step, we could predict the clinical outcome for patients with the disease, while taking into account this prior information. In this case the p-values serve as external data for the prediction problem (van de Wiel et al., 2016; Münch et al., 2019).

7.7 Future research

One finding of this thesis is that stacked generalisation can lead to predictive and interpretable models (Chapter 6). Cross-validation selection is about selecting one learner but stacked generalisation is about combining multiple learners. It is thus possible to combine the interpretability of a single learner with the predictivity of multiple learners. We believe that this approach has a wide applicability, including paired covariates (Chapter 5), multivariate responses (Waegeman et al., 2019), multiple sources of co-data (van de Wiel et al., 2016), and random regressions (Song et al., 2013).