Permutation-based inference for high-dimensional data.

Hemerik, J.

1 Introduction

This thesis is about statistical methods based on permutations or other transformations of data. Throughout much of this work, emphasis is on testing many hypotheses simultaneously, which is called *multiple testing*. A key message of this thesis is that by using permutation-based multiple testing, the dependence structure of the data can be taken into account, which leads to powerful methods.

In this introduction, we will first give some typical examples of *high-dimensional data*. We will then explain the key statistical issues when testing hypotheses about such data. Finally, we will introduce the new methods in this thesis, which address those issues.

1.1 High-dimensional data: examples

In many biomedical and psychological experiments, a large number of measurements are collected per individual, for example, measurements of many protein concentrations in the blood. Such measurements are referred to as *high-dimensional data*. We now give two examples of high-dimensional data that are frequently studied in respectively biomedical research and neuroscience.

Our DNA contains about 20,000 genes. A gene is a piece of DNA that codes for a molecule that has a function in the body. A gene contains a sequence of nucleotides. There are four types of nucleotides: adenine, thymine, guanine and cytosine (A, T, G and C). A gene is used to produce a string of RNA. Usually this is in turn used to code for the synthesis of a protein. Proteins are involved in various processes in our body, including our immune system and other processes related to disease. The speed at which a gene produces RNA strings to make proteins varies over time and per person and is called the *expression level* of that gene. By measuring expression levels and comparing these between different individuals, cells or tumors, we can learn whether these levels are associated with certain phenotypical traits, such as disease status. We can then build a model that predicts the disease type based on the gene expression profile. This can help to allow cheap, early and accurate diagnosis of diseases. Likewise, we can build a model that chooses the optimal treatment for a particular patient, based on his or her gene expression profile.

There are various technologies for measuring gene expression, each with their own pros and cons, related for example to financial costs and measurement accuracy. Until a decade ago the main method for measuring gene expression was to use *DNA microarrays*. A DNA microarray is a collection of microscopic DNA spots attached to a surface. A piece of RNA only binds to such a spot if it contains the complementary sequence of nucleotides. Thus, measuring how much RNA binds to a specific spot, leads to an estimate of the amount of the corresponding type of RNA in the sample. This is used to estimate the expression level of the gene that coded for this piece of RNA.

In the last decade more advanced techniques for measuring gene expression have appeared under the banner of *next-generation sequencing*. These techniques provide improvements such as lower costs and better quantification of lowly and highly expressed genes. Often the expression levels of thousands of genes are measured, leading to very high-dimensional data.

A different field of research where high-dimensional data occur, is the analysis of brain images. Researchers can measure the activity of various brain regions of a person, while that person performs some task. In this way it can be inferred which brain regions are involved in which types of human behavior.

Measurement of brain activity is often performed using *functional mag-*

netic resonance imaging (*fMRI*). In brain regions with high activity there is more deoxygenated hemoglobin in the blood, which interferes with a magnetic resonance signal, so that the active regions can be distinguished.

Often brain activity is measured of many thousands of small disjoint areas in the brain (referred to as *voxels*), which cover the whole brain. This is a good example of high-dimensional data. A large number of brain regions are investigated simultaneously, which usually means that the researcher tries to answer a lot of research questions at the same time. The more questions are asked, the more wrong answers could be given. Thus, great care should be taken in posing and addressing these questions, to limit the number of false research findings. (See (Bennett et al., 2009) for an example of how this can go wrong.)

1.2 Multiple testing and the false discovery proportion

Often many hypotheses are tested simultaneously. For example, for 20,000 genes one could test the hypothesis that their expression rate is not (substantially) associated with some phenotype. When multiple hypotheses are tested, the most classical way to avoid type-I errors, is to make sure that with high (e.g. 95%) probability, no true hypotheses are rejected. This is called strongly controlling the *Family-Wise Error Rate* (FWER). Controlling the FWER means that a hypothesis is only rejected if there is extremely strong proof to do this. However, often the proof against a single hypothesis is not extremely strong. The consequence may be that no rejections are made, even if there is (clinically relevant) signal in the data.

This way of avoiding type-I errors, i.e. controlling the probability of any type-I errors, is sometimes too strict. Indeed, it is not only important to avoid type-I errors, but also to not miss out on important scientific findings. Hence, researchers have been looking for less strict statistical approaches, which allow researchers better to detect the presence of false hypotheses. Note that such approaches allow more false findings than FWER controlling methods, so that the results have a different meaning.

Several less strict approaches exist, one of which is to select the rejected hypotheses in such a way that the *false discovery rate* (FDR) is controlled. The FDR is the expected value of the *false discovery proportion* (FDP). This is the fraction of false findings (type-I errors) among the rejected

hypotheses. It is defined to be 0 when there are no rejections. The most well-known FDR controlling method is by Benjamini and Hochberg (1995). Controlling the FDR means to keep the FDR below a certain value, e.g. 0.05. This guarantees that on average, the FDP is at most 0.05. (This does not mean that a rejected hypothesis is false with probability at least 0.95. Indeed, there can be a positive correlation between the FDP and the number of rejections.)

Compared to FWER control, controlling the FDR usually enables the user to reject more false hypotheses. On the other hand, with FDR controlling methods there are also more rejected true hypotheses. (Rather than 'rejected', we may use the term 'selected', since we are not extremely sure that a rejected hypothesis is false.) Correspondingly, controlling the FDR represents a trade-off between keeping the number of true positives large and keeping the number of false positives small.

When a researcher tests several hypotheses, he or she may want to be confident that the FDP is smaller than a certain value. However, when the FDR is smaller than 0.05, this does not mean that the FDP is likely smaller than 0.05. Thus, the researcher has an estimate of the FDP (namely, the nominal FDR), but not a confidence interval around the estimate. In many cases, it would be relevant to provide such a confidence interval. For example, it indicates how accurate the estimate is.

Therefore, statistical methods are being developed that provide a confidence interval for the FDP. Focus is usually on confidence upper bounds for the FDP. Similarly, methodology has appeared that allows controlling the FDP with confidence. This means to keep the FDP below some given, small value with some given, large probability. The present thesis provides several contributions in this field (chapters 3 and 4), as we will discuss in section 1.4.

1.3 Dependence and permutation methods

Most multiple testing methods perform many hypothesis tests separately and then combine the results to decide on which hypotheses to reject. The result of an individual test is usually a test statistic. In particular this could be a p-value. When m hypotheses are tested, m test statistics are obtained. These need to be used somehow to decide which hypotheses are rejected.

Often there are dependencies in the data. For example, the expression levels of different genes are correlated. The consequence is that the m test statistics are also dependent. The dependence structure influences the behavior of a multiple testing method. Hence, when choosing such a method, the dependence structure in the data should be taken into account.

In many situations, however, there is limited knowledge about the dependence structure in the data. Consequently, it is not known which multiple testing methods will control the considered error rate (e.g. the FWER or FDR) and which methods will not. This problem is often remedied by choosing the most conservative multiple testing method, to be sure that the error rate is controlled. The downside of this is that the rejection criterion is then likely to be more conservative than the user wanted.

A different approach to the problem of an *a priori* unknown dependence structure, is to use permutation-based multiple testing methods. Some of these methods have known, exact, finite-sample properties in several simple but important scenarios, such as randomized clinical trials. In multiple testing, permutation-based methods take into account the correlation structure in the data while making few assumptions on that structure. The only assumption made usually comes down to the following: the joint distribution of the part of the data under the null hypothesis should be invariant under permutation, e.g. shuffling cases and controls. (This usually implies that the hypotheses are point hypotheses, but extensions to interval hypotheses are sometimes possible.)

As a first example of a permutation-based multiple testing procedure, we consider the $maxT$ method by Westfall and Young (1993). This method strongly controls the FWER. It is the permutation-based analog of (the *closed testing*-based improvement of) Bonferroni's method. At confidence level 0.05, Bonferroni's method rejects all hypotheses with p-values smaller than $0.05/m$. The reason is that for any dependence structure of the p-values, $0.05/m$ does not exceed the 0.05-quantile of the minimum of the null p-values. The method by Westfall and Young (1993) however, makes use of the data to determine the rejection threshold. It permutes the data many times, each time calculating all m p-values and recording the smallest p-value. (Rather than p-values, other test statistics could be used, hence the name $\max T$.) The 0.05-quantile of these minimums is used as the rejection threshold. (This threshold, just as Bonferroni's threshold, can be further improved by using closed testing.) Especially when the p-values

are strongly positively correlated, this threshold can be much higher than $0.05/m$.

Another popular permutation-based multiple testing method is *Significance Analysis of Microarrays* (*SAM*) by Tusher et al. (2001). This method estimates the FDP. It is a general method, which is not at all limited to microarray data. The method works as follows. A rejection cut-off is chosen, such that each hypothesis is rejected only if its test statistic lies above the rejection cut-off. The data are permuted many times. Each time it is recorded how many hypotheses would have been rejected if the permuted data had been observed. The mean or median number of rejections for the permuted data, is an estimate of the number of false positives. Dividing this by the number rejections, gives an estimate of the FDP. In section 1.4 it is discussed how this method is improved in this thesis.

1.4 This thesis

Chapter 2 of this thesis is an introduction to the fundamentals of permutation testing. Although we will usually write 'permutation methods', most methods in this thesis are not restricted to permutations, but can use any group of transformations, such as rotations or sign-flipping.

Chapters 3 and 4 make use of some of the theory in chapter 2. Chapter 3 constructs confidence bounds for the FDP. Chapter 4 generalizes chapter 3 to simultaneous confidence bounds, which in particular allows for FDP control with confidence. Chapter 5 is quite different from chapters 2-4. It lets go of the fundamental assumption of distributional invariance under transformations. Rather, it provides tests which are only asymptotically exact. It presents methods for robust testing in generalized linear models (GLMs). Apart from being important on its own, chapter 5 broadens the applicability of the earlier chapters, allowing those methods to be used in the context of GLMs.

1.4.1 Permutation-based confidence on the false discovery proportion

Chapter 2 of this thesis lays the foundations for permutation testing based on randomly sampled permutations. Chapter 3 builds further on chapter 2 by improving the permutation-based multiple testing method SAM, which

was discussed above. The orginal SAM procedure only provided an estimate of the FDP, with unknown properties. We extend the method by providing a confidence upper bound for the FDP. We show that for $\alpha \in [0, 1)$, the $(1 - \alpha)$-quantile of the numbers of rejections for the permuted verisons of the data is a $(1 - \alpha)100\%$-confidence upper bound for the number of false positives. In particular, if we take $\alpha = 0.5$, we obtain a median unbiased estimate of the number of false positives.

Moreover, by using a method related to closed testing (Goeman and Solari, 2011), we obtain an improved confidence bound, which is at least as small as the basic bound. Although we derive a computational shortcut to compute the improved bound, when there are many hypotheses, the calculations are still computationally infeasible. Hence we provide a method which approximates the improved bound. The approximation method can be used when there are many thousands of hypotheses. In simulations, this method provided valid $(1 - \alpha)100\%$-confidence upper bounds.

Since our method is based on permutations, it tends to provide better (lower) FDP bounds than parametric methods. Another important advantage of the permutation approach is that the p-values (or test statistics) that the methods uses, need not be exact. The reason is that any test statistics can be used; the only assumption made by the method, is permutation-invariance of the part of the data under the null.

Chapter 4 generalizes the method of chapter 3 to simultaneous confidence bounds. For a range of rejection thresholds of interest, it provides confidence upper bounds for the FDP. These bounds are simultaneous: with probability at least $1-\alpha$, these bounds are all valid. This allows for *post hoc* selection of the rejection threshold: if the threshold is chosen based on the data, then still a valid confidence bound is obtained. Thus, based on the data, the user could choose a threshold that leads to a desired confidence bound for the FDP.

A similar method was already published by Meinshausen (2006). This method served as an inspiration for chapters 3 and 4 of this thesis. In chapter 4 we improve the method of Meinshausen by 1. modifying it to be exact, 2. generalizing it and 3. improving its power. In particular, we construct an iterative method, which has more power than the single-step method by Meinshausen (2006) and is still exact. We also provide a fast approximation of this method, which provided valid confidence bounds in our simulations.

1.4.2 Robust testing in generalized linear models

It is often of interest to test if a coefficient in a GLM equals 0 or some other value, or to construct a confidence interval around an estimated coefficient. When the specified model is correct, this can be done with classical parametric tests such as the likelihood ratio test or score test. In many situations however, the assumed model is not even approximately exact, due to e.g. overdispersion, heteroscedasticity or ignored nuisance parameters. The traditional parametric tests then lose their properties.

In chapter 5 we propose a novel type of test, based on sign-flipping score contributions. (Note that the score, the derivative of the log-likelihood, is a sum of n individual contributions.) The basic idea underlying this test is the following. Under a point null hypothesis of the form $H_0 : \beta = 0$, the score contributions have mean 0. By recalculting the score many times after randomly multiplying the summands by -1 with probability 0.5, we obtain a reference distribution, that we compare with the original score. We reject the null hypothesis if the original score lies in the 5%-tails of this distribution. This method is closely related to the permutation test. It is often robust against the mentioned forms of model misspecification. For example, if there is constant overdispersion, then the variance of the score is misspecified, so that the classical parametric tests fail. Our test remains asymptotically exact, however, since both the observed score and flipped scores are misspecified by the same factor.

A key assumption underlying this test, is that the n score contributions are independent. When nuisance parameters are estimated, this is no longer the case. As a consequence, the variance of the original score is shrunk and our test becomes conservative. To solve this issue, we consider the *effective score* (Marohn, 2002). This is the part of the score that is orthogonal to the nuisance score. It is asymptotically unaffected by the nuisance estimation. As a consequence, we again obtain an asymptotically exact test.

If the scores were independent and symmetric, then our sign-flipping test would be a special case of the exact test in chapter 2. Due to its close relationship to permutation tests, our sign-flipping test can be combined with the multiple testing methods from chapters 3 and 4. This allows using those methods in the context of GLMs.