7 Summary and discussion

The arrival of high-dimensional microarray data sets brought along a need for new approaches and methods for both significance analysis and classification tasks. We focus on the significance analysis, for which we have to deal with an extensive multiple testing problem. The expression of thousands of genes is measured simultaneously. Each gene is tested for expression changes between two or more groups of samples and thus constitutes one hypothesis. Classical approaches to multiple testing like the family-wise error rate are considered to be too strict, such that only a few genes are called significantly induced. The family-wise error rate equals the probability that we include at least one false positive into our set of rejected hypotheses. In the light of tens of thousands of hypotheses, a less conservative error rate is needed. In their seminal work, Benjamini and Hochberg (1995) rediscovered the false discovery rate as an error measure suitable for microarray or other large-scale data. The false discovery rate is defined as the expected rate of false positives among all rejected hypotheses, such that we might allow a certain level of false positives in our test procedure. For both error rate models rich research exists and various procedures were suggested to provide control of these rates. Control means that a procedure retains the desired level of significance under certain definitions of the null model. Control of error rates is important for drawing conclusions on the significance of a set of genes. This feature brings along a problematic aspect of error rate control: the false discovery rate is a global error rate that assigns a certain level of significance to sets of genes, not to individual genes. By including more highly significant genes into this set, we might allow the inclusion of less significant genes while maintaining the same level of significance. This "problem of cheating" (Finner and Roters, 2001) was discussed and illustrated in Section 3.3.

The local false discovery rate (Efron *et al.*, 2001) appears to be a variant of the global false discovery rate, yet it is what we were looking for: an error probability assigned to individual genes. It is defined as the probability that a gene is not differentially expressed given we observed a certain p-value for this gene and given the set of observed p-values of all genes in the experiment. Regarding the multiple testing issue, we might compute local false discovery rates for each gene, define a threshold and call the respective genes significantly induced. However it is not clear if we can assure control of the local false discovery rate. With the local variant of the false discovery rate we indeed leave the field of error rate *control* and focus on error rate *estimation* instead. The estimation involves the complete set of observed p-values and provides a global view on amount and mixture of differential and non-differential expression in the data set. We are aware of the problem that experimenters might want to return to control or at least to conclusion on certain gene sets. Efron *et al.* (2001) suggests to base inference on individual local false discovery rate values and to call genes with values not exceeding a certain threshold significant. We do not regard local false discovery rate estimation in this sense but use it as a sensible tool to explore the overall amount of evidence in the data set. When starting with the project, the primary idea was to discover p-value levels where an accumulation of p-values occurs. We observed that the p-value density does not always decrease monotonically. Instead, hubs of p-values appear in zones where evidence for differential expression is not supported by the observed p-value level. Using global false discovery rates, these *twilight zones* would lie far beyond our significance thresholds and we would not consider examining them. P-value hubs in twilight zones contradict the assumption of concave p-value density in Genovese and Wassermann (2004). However, only a few applications reveal twilight zones. One example included in this thesis is the Breast-cancer 1 comparison. In Figure 5.3 we observed a plateau of local false discovery rate around the p-value level of 0.2. We might classify the Breast-cancer 1 comparison as an experiment exhibiting only weak but wide-spread evidence for differential expression.

With appropriate local false discovery rate estimators we can track twilight zones. However, for every-day's analysis of microarray data sets where inference on individual genes is wanted, it is challenging to provide an accurate estimate of the percentage of non-induced genes π_0. The improved estimate can then be used for example in the positive false discovery rate adjustment procedure of Storey and Tibshirani (2003). Our first contribution to improve the significance analysis of microarray data includes both: accurate estimates of the local false discovery rate and of prior π_0. In Chapter 5 we proposed a regularized stochastic search algorithm. The approach is termed SEP for successive exclusion procedure and works by dividing the set of p-values into two parts. One part represents p-values of induced genes, the other one represents p-values of non-induced genes. Following probability theory, we assume the second part to be uniformly distributed. To force a unique separation into the two subsets, we further assume that the uniform part consists of as many p-values as possible. That is, we do not allow that the alternative subset contains any uniformly distributed remainders. These two assumptions—uniformity of the non-induced part and exclusion of any uniform fraction in the induced part—are sufficient to identify the mixture parameter π_0, that is the proportion of non-induced genes in the experiment. The separation into two sets of p-values only provides distributional information. We must not conclude that p-values in one set represent genes that are truly differentially expressed. We can only estimate the proportion of differentially expressed genes at a certain p-value level. Again, the vertical separation into genes called significantly induced or non-induced leads us back to the issue of control of multiple testing procedures, which we do not pursue in this thesis.

We presented the main SEP algorithm along with a calibration and a fine-tuning step. The former calibrates a penalty term that safeguards against excluding more p-values than necessary from the uniform part. The more weight we give to the penalty term, the more difficult it is to remove p-values. A removal is only allowed if this leads to an increased goodness-of-fit to the uniform distribution. Since the weight on the penalty term depends on the number of values already removed, there exists a point where the exclusion of a value is more expensive than the gain of the resulting fit. Here the algorithm quits. We might think of cases where it is necessary to remove many values. If the size of the uniform part is small, a strict regularization is counterproductive. Therefore we divided the algorithm into two layers. To get a burn-in set of excluded p-values, we apply SEP without regularization until the goodness-of-fit reaches a certain threshold. Then, we fine-tune the estimates by continuing the process with regularization.

The implementation of SEP is used in-house for routine analysis of data sets. An earlier version was evaluated in a simulation study where the estimates appeared to be "stable and reliable" (Broberg, 2005). On the other hand, the study revealed shortcomings of SEP that were due to the final density estimates. Both, the in-house use and the results of the study, lead to improvements of the original version of SEP. The most visible improvement may be the fine-tuning step, other changes are more subtle. The changes brought along the need for a proper evaluation including comparison to existing methods. Since the estimation of the local false discovery rate in principle boils down to the estimation of the prior probability π_0, we compared SEP even to approaches developed for different settings than microarray data. While our initial goal was to analyze the performance of SEP in comparison to eventual competitors, we observed that many of these approaches had difficulties to estimate π_0 well. Some methods exhibited under-estimation of the true parameter consistently for all simulation settings. However, under-estimation of π_0 equals over-estimation of $\pi_1 = 1 - \pi_0$, the percentage of induced genes in the experiment. With an over-estimated π_1, we might draw overly optimistic but incorrect conclusions on the amount of differential expression. The set of observed p-values appears to carry more evidence for differential expression than there actually is. Thus, we might want to be conservative, which relates to over-estimating π_0. From the set of fourteen estimation procedures we removed three in the first step of the simulation analysis since they severely under-estimated π_0. For another four procedures including the prominent method of Storey and Tibshirani (2003) we observed increased variability of the estimates as the number of p-values increased. Since we expected a method to provide more precise estimates for higher sample sizes, we excluded these procedures from further consideration. Finally we were left with six approaches—including SEP—that performed consistently well in our simulation study. While some of these methods provided conservative upper bound estimates, other methods like SEP returned estimates of high accuracy and high precision. The current implementation of SEP still appears to be stable and reliable and performs well over the whole range of evaluated parameter combinations. The method of Nettleton and Hwang (2003) also performed well and might be used as a conservative but quick pre-estimator of π_0.

Our second contribution to an improved significance analysis targets a special artifact of gene expression data: random permutations of the class labels, which are used to model the score distribution under the null hypothesis, do not always lead to a valid null distribution. A single permutation might be correlated with an unknown covariate that triggers differential expression. Unknown confounders are for example genetic background of patients or undetected experimental artifacts. The influence of a hidden confounder is identifiable if a large number of genes is affected. By transforming the respective permutation scores into p-values, we observe an accumulation of small p-values such that the overall p-value distribution deviates substantially from a uniform distribution. Since we do not want to base inference on skewed null distributions, we propose a simple but efficient algorithm to filter for admissible permutations. To our knowledge, it is the first approach that considers the removal of whole permutations instead of removing single genes as was done for example in Xie *et al.* (2005) or down-weighing genes as was done for example in Guo and Pan (2005). We showed how the significance analysis benefits from permutation filtering: the results are longer lists of significant genes and improved accuracy when estimating the global or local false discovery rate. Along with the benefits of filtering come three problematic aspects of significance analysis of microarray data, which we will discuss in the following. In particular, we consider inclusion of known covariates, computation of p-values and dependence between genes.

There are experiments where we have to adjust for known covariates. This can be accomplished by balancing the permutations for the given covariate. For example, if patient gender is an observed covariate, which does not correlate with the variable of interest, we have to consider only those permutations balanced for gender, that is where gender does not correlate with the random assignment of the variable of interest. Including this information reduces the number of possible permutations but leaves the general procedure applicable. To date, the inclusion of known covariates into our software implementation is not possible. An exception are paired data where we observed two microarray measurements per patient, for example before and after treatment. Here the assignment to pairs is regarded as a block variable and we consider only within-block permutations. Thus we randomly assign "before" and "after" labels within each pair, which limits the number of possible permutations compared to the unpaired case. So far, our software package handles two sample paired or unpaired data sets. One can test for differences in mean expression or correlation to some variable of interest and account for paired or unpaired samples when drawing random permutations within the filtering algorithm.

In the end of Section 3.2 and in Section 4.4 we investigated the influence of the p-value computation method. Using the permutation approach, we either compute the p-value of the ith gene on the permutation scores of gene i alone or we pool across genes and use the permutation scores of all genes. The first method is termed *gene-wise* approach, the second one is termed *pooled* approach. The choice of gene-wise or pooled p-values hinges on the assumption whether the scores follow the same null distribution for each gene. When using the common t-test score, the scores ideally follow a t-distribution with equal degrees of freedom for each gene. However, the t-score suffers from variability of the variance estimates in its denominator. Almost constant genes with small variances lead to variable gene rankings and a consistent gene might be among the top-scorers by chance. To safeguard against variable rankings, several *regularized* t-scores (z-scores) were proposed, which put less weight on the variances and more weight on the actual effect sizes (Efron *et al.*, 2001; Tusher *et al.*, 2001; Wu, 2005; Smyth, 2004; Cui *et al.*, 2005). These scores however might not follow the same null distribution for each gene. Xie *et al.* (2005) derived analytical results for the influence of differential and non-differential genes on the observed score distribution. They compared one-sample equivalents of log ratio, z- and t-scores introduced in Section 2.4 regarding variance and tail strength of score distributions under induction or non-induction and found that the influence of the z-score lies between those of log ratio and t-score.

In Section 4.4 we explored differences between gene-wise and pooled p-values when using the z-score of Efron *et al.* (2001), as we did throughout the thesis. Assuming gene-wise or pooled null distributions leads to substantial differences between the resulting p-values. We further examined the genes with largest differences between the two types of p-values and found in principle two subsets. One subset consisted of almost constant genes. Although the variance estimate is of less importance using z-scores, the resulting distribution of permutation scores is narrower than the pooled null distribution and the gene-wise p-value will be much smaller than the pooled p-value. Pooling might prevent from misleading results by assigning larger p-values to constant genes. The second explanation for differences between gene-wise and pooled p-values are outliers in the expression values. If one or two large expression values are present, the resulting distribution of permutation scores will have heavier tails than the pooled null distribution. The respective genes typically receive large gene-wise p-values, which might not support evidence for differential expression. However, the choice of gene-wise or pooled null distributions does not effect the results presented in this thesis. The differences lead to different rankings of the genes, which is important if one wants to conclude on individual genes. We observed little differences between the overall p-value distributions, which are the basis of our methods. Application of the local false discovery rate estimator remains valid but single-gene conclusions are certainly different if we use gene-wise p-values.

The assumption and use of gene-wise null distributions does not safeguard from the artifacts caused by hidden confounders (Section 6.2). In our first publication on permutation filtering (Scheid and Spang, 2006), we considered only pooled p-values. In an extended version we received preliminary results for gene-wise p-values (Scheid and Spang, 2007). Indeed we observed a variety of p-value density shapes very similar to the pooled cases shown in Figures 6.2 to 6.7. If a hidden variable affects a sufficiently large number of genes, the signal will distort the p-value distributions of correlated permutations. Whether the p-values were derived by gene-wise or pooled computation does not seem crucial to us. We argue that if a permutation strongly correlates with a hidden confounder and affects a large number of genes, the hidden signal is identifiable by permutation filtering. Removal of this permutation will lead to the same results as observed with pooling. To us, the key value is the number of confounded genes. Indeed, a second result of our preliminary research in Scheid and Spang (2007) is that the number of affected genes has to be large to reveal hidden signal. We repeated the simulation of Section 6.4 but confounded only 100 genes instead of 1000. Still, permutation filtering improved the sensitivity but the differences were not significant any more.

The third problematic aspect in the significance analysis of microarray data concerns the general assumption that genes are independent of each other. In fact genes are coregulated, such that two genes are expressed in parallel because they are triggered by the same regulation factor. Also, genes are connected in pathways and thus expression of one gene triggers expression of another gene further down the signal cascade. Both molecular observations—coregulation and pathway dependence—lead to the assumption that gene expression is correlated and that we might discover groups of genes, which are correlated among each other but not with genes in other groups. The covariance matrix of all genes in the experiment will then be block-structured. Storey termed this situation "local dependence" and showed that his q-value procedure provides asymptotic control of the positive false discovery rate under weak dependence assumptions (Storey, 2002, 2003). Recent publications suggest that genes are not correlated in clumps. Klebanov *et al.* (2006) assumed long-range correlation and proposed a new type of dependence. The authors investigated whether one gene is stochastically proportional to another gene, that is the first one modulates the second one such that their expression values are proportional. Indeed, they analyzed pairs of genes within several comparisons of the ALL study (Yeoh *et al.*, 2002) and found genes being stochastically proportional to the majority of genes in the experiment. In a different comparison, the same genes show common correlation to most of the other genes. The authors concluded that the clumpy dependence assumption of Storey (2003) does not hold and that a long-range correlation assumption is more appropriate. Depending on the strength of correlation, Gao (2006) suggests to use pooled p-values in case of weak correlations and gene-wise p-values in case of long-range correlations.

Questioning common multiple testing practice, Klebanov, Yakovlev and coauthors published an interesting paper series on the effects of "pooling" (Klebanov and Yakovlev, 2006; Qiu *et al.*, 2005, 2006). Here pooling does not refer to p-value computation but to the general permutation test approach. The authors discuss whether one should pool across scores to estimate the mixture distribution and derive global or local false discovery rate values. Reasons against pooling are again non-equal null distributions and long-range correlation. The authors argue that even if we can estimate the mixture distribution with high accuracy, the implications drawn from this estimate do not provide reasonable single-gene information. Qiu *et al.* (2005) showed in simulations how strong correlation affects the local false discovery rate procedure of Efron *et al.* (2001). Likewise, Qiu *et al.* (2006) observed that the number of genes selected with multiple testing procedures is highly variable in presence of correlation. The authors propose a resampling-based strategy to explore whether the analysis of the data set at hand is influenced by gene correlations. Long-range correlations are indeed a source of confounding signals. Together with the assumption of non-equal null distributions, we have to be aware of biased implications.

Throughout the thesis, we assumed independence between genes. We were well aware that this assumption is too naive in case of biological data sets. In particular, we believe that the permutation artifacts shown in Figures 6.2 to 6.7 are not only caused by hidden confounders but also by highly correlated expression values. The border between both phenomena is fluent. If clumps of genes are correlated, they might receive similar p-values, which constitutes deviation from uniformity. Permutation filtering compensates for confounding variables but is less effective under strong correlation. To examine the performance of our method under correlation, we repeated the simulation experiment without a confounding variable but with certain correlation structures. First, we simulated clumpy dependence as in Storey *et al.* (2004). The values for 2500 genes were drawn from a multivariate normal distribution with mean 0 and a block-diagonal covariance matrix, such that blocks of 100 genes had a correlation of ±0.4. The first 500 genes were induced by adding a value of 2. Second, we simulated data with global correlation as in Qiu *et al.* (2005), such that the correlation between all 2500 genes was 0.4. The first 500 genes were induced as above. Details of these preliminary simulations can be found in Scheid and Spang (2007). The introduction of clumpy or global correlation increased the variability of sensitivity and specificity of the positive false discovery rate estimates. In case of global correlation the FDR procedure failed, which was in accordance to the results in Qiu *et al.* (2005). In addition, filtering did not lead to substantial improvements—or, in other words—the performance based on random permutations did not decrease as much as with a single confounding variable being present, as we observed in Section 6.4.

Inclusion of correlated genes into the simulation model shows the limitation of our filtering approach. We might compensate for hidden covariates but not for correlation. Also, the number of confounded genes has to be large enough to be detected by our method. If only a few genes are influenced by a hidden signal, the deviation from uniformity is not strong enough and filtering does not substantially improve the significance analysis. Another shortcoming of our method so far is that it is a purely heuristic approach. We believe that the problem of confounding signal is present in most microarray comparisons and that it manifests itself in permutation p-value distributions. Permutation filtering can be considered as a first step towards a novel concept in multiple testing, yet theoretical research is necessary to further understand its implications on microarray data.

Throughout the thesis, we illustrated the proposed methods on simulated data and on several biological data sets. The examples were limited to comparisons of two clinical classes. Our current software implementation can only handle two-sample tests on equality of means and tests on correlation to a clinical variable. In principle, our methods are not restricted to two-sample cases but apply to a variety of statistical hypothesis tests. If we have more than two classes at hand we might compute F-statistics to test whether one class shows differential expression. More sophisticated settings might require the use of regression models to explore a contrast of interest. In any case it is possible to run a permutation test on the chosen score and derive a set of empirical p-values. From here on, we apply our methods to estimate the mixture parameter π_0 and the local false discovery rate. To include the permutation filtering, we would have to return to the level of expression values and class labels. The estimation approach and the filtering are independent of the chosen score and are applicable whenever large-scale permutation is possible.

The algorithmic approaches presented in this thesis have another feature in common: both are not limited to gene expression data. The estimation of the local false discovery rate and the amount of true null hypotheses π_0 make sense for any kind of large-scale multiple hypothesis testing. Filtering for admissible permutations is an elegant way of reverse engineering when the true null distribution is unknown or covered by signals of hidden covariates. We believe that the benefits brought by our methods are not limited to the significance analysis of gene expression data but may support the analysis in many fields of large-scale applications.