Wiebke Werft
Dr. sc. hum.

Identification of Predictive Factors in High Dimensions for the Therapeutic Success in Solid Tumours

Promotionsfach: Biostatistik (DKFZ)
Doktormutter: Prof. Dr. Annette Kopp-Schneider

This dissertation investigates statistical methods for the identification of predictive factors in high-dimensional data. The research has been motivated by a companion study to a neoadjuvant breast cancer study whose primary objective was to identify predictive genes for the response to two specific treatments. Prior to treatment, breast cancer tissue of the patients has been collected for microarray experiments. Specific genes should be identified that can potentially predict the patients response to treatment (i.e. presence or absence of pathological complete response in the breast) specifically to each of the two neoadjuvant regimens.

A predictive factor has to be distinguished from a prognostic factor. While a prognostic factor is an influential covariate independent of treatment, a predictive factor holds a treatment/covariate interaction. Hence, the model which is used in this thesis for the identification of predictive factors is a gene-wise generalised linear model including an interaction between the gene expression covariate and the treatment assignment. Testing the regression parameter of the interaction term will indicate if a gene is a predictive factor or not. Microarray experiments dealing with gene-wise screening approaches pose the problem of testing thousands of hypotheses simultaneously. Therefore, p-value adjustment procedures for control of the false discovery rate (FDR) are highly required for such multiple testing scenarios.

Initial analyses of the interaction term were based on Wald and likelihood-ratio test (LR) statistics which are standard methods for inference on coefficients in generalised linear models. Resulting p-values were adjusted for control of the FDR according to the Benjamini-Hochberg (BH) procedure. Simulations for a binary treatment response on this approach revealed that the true error rate is either overestimated (LR test) or underestimated (Wald test) in small sample sizes.

In further approaches, the methods of inference on the interaction term were modified to account for deficits due to small sample sizes. In order to prevent the occurrence of infinite parameter estimates in case of monotone likelihood, an effect which might occur due to small sample sizes, the Wald and LR test statistics are computed based on estimates derived from a penalised maximum likelihood function. Moreover, the asymptotic distributions assumption of the LR test statistics might not hold in small sample sizes, and hence, a permutation of regressor residuals (PRR) test is considered. In a separate investigation, the interaction was modelled via multivariable fractional polynomials and was then tested by means of the LR test statistic.

Additionally, adaptive modifications of the BH adjustment procedures were considered. To take into account the dependency structure of the gene expression data and hence of the test statistics, the Benjamini-Yekutieli (BY) and the Blanchard-Roquain (BR) procedures were included in the analyses. These multiple testing procedures (MTPs) are referred to as marginal

as they do not consider the joint distribution of the test statistics. Moreover, resampling-based joint MTPs also suitable for arbitrarily dependent test statistics were extended for the logistic regression model incorporating shift and scale-transformed (SST) and quantile-transformed (QT) joint null distributions for the step-down minP and maxT procedures.

Comprehensive simulation studies were performed to compare the different approaches for the logistic regression model. The conclusions can be summarised as follows:
Within marginal MTPs, the BH procedure and adaptive modifications of it behave less conservative than the BY and the BR procedures if the tests statistics are independent. As inference method the LR test statistic in combination with the PRR test should be used, especially when small sample sizes are considered. When dependent data is present, only the PRR test in combination with the BY and the BR procedures provides adequate control of the FDR. Within resampling-based joint MTPs, the step-down maxT procedure based on QT joint null distributions of the LR test statistics is recommended in practice when sample sizes are sufficiently large ($n \leq 75$ in the considered simulation scenarios).

Applications of methodologies to the motivating breast cancer study data correspond to the results of the simulation studies. The BH adjustment procedure applied to the asymptotic LR test detected 175 potentially predictive genes, and the application of the PRR test reduced this list to 53 genes. Corresponding to the simulations, the FDR for the PRR test is reduced. Thereby, the list of predictive genes is reduced to a list of potentially more reliable candidates. When focusing on marginal MTPs suitable for any dependency structure, no significant gene was found when applied to the asymptotic LR test. This reflects the too conservative simulation results of the BY and BR procedures. Only the PRR test detected nine predictive genes if used together with the BY or the BR methods. These reduced lists of predictive genes are also observed when using joint MTPs. No predictive gene could be detected when applying the maxT procedure, which provides better control of the FDR than the minP procedure according to the simulation results.

Sample size issues and the correct choice of test statistics together with an appropriate multiple testing procedure play a major role for controlling the FDR for the identification of predictive factors. If statistical analyses were based on the above recommendations the identified lists of potential predictive genes will generally have less false positive detections. Hence, the present thesis provides guidance in establishing predictive factors for usage in personalised medicine.