Integration of multiple genomic data sources in a Bayesian proportional hazards model with variable selection

Tabea Treppmann[1], Manuela Zucknick[2] and Katja Ickstadt[1]

[1] Faculty of Statistics, Technical University Dortmund
tabea.treppmann@uni-dortmund.de, ickstadt@statistik.tu-dortmund.de

[2] German Cancer Research Center (DKFZ), Heidelberg
m.zucknick@dkfz-heidelberg.de

Abstract. The search for an optimal prediction model is very important in statistics. Especially in high dimensions the question arises, which variables mainly influence prediction. Bayesian modeling provides a good way to deal with this problem. For linear models George & McCulloch (1993) [1] developed a promising approach, where the variable selection takes place within the Markov Chain Monte Carlo (MCMC) sampling in an intuitive way. In survival analysis, for example, gene expression data constitute a common example of a high dimensional setting where only a few variables show significant influence on the prediction. Lee et al. (2011) [2] introduced a proportional hazards model for survival which is able to deal with such applications. Aiming to use the method of George & McCulloch (1993) [1] within survival analysis, we combined their approach with the survival model of Lee et al. (2011) [2]. This demonstrated promising results in an initial simulation study. To test the practical behavior and quality of the model we used data of medulloblastoma patients and incorporated different extensions, e.g. the use of copy number variation data attempting to guide the variable search. Furthermore, we used parallel tempering to improve the mixture of the MCMC chains and attain more precise predictions [3]. In the course of the analyses it could be seen that the results were marked with a high uncertainty due to the quality and structure of the data. For future research it would therefore be essential to use data of better quality, but apart from that, there is also still a lot of potential for additional changes and extensions to the model. The MCMC sampler was implemented in the statistical computing environment R [4].

References

[1] George E.I. & McCulloch R.E. (1993). Variable Selection Via Gibbs Sampling. *Journal of the American Statistical Association*. 88(423):881-889.

[2] Lee K.H., Chakraborty S. & Sun J. (2011). Bayesian Variable Selection in Semiparametric Proportional Hazards Model for High Dimensional Survival Data. *The International Journal of Biostatistics*. 7(1):Article 21.

[3] Geyer, C.J. (1991). Markov Chain Monte Carlo Maximum Likelihood. In E.M. Keramidas, editor, *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface*. Interface Foundation, Fairfax Station.

[4] R Development Core Team. (2009). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. *http://www.R-project.org*. ISBN 3-900051-07-0, Version 2.10.1.