Model Selection Criteria for Gene-Expression Data

In toxicology, concentration-response curves are used to describe the relationship between a concentration of a certain compound applied to cells and its response. This is done simultaneously for thousands of genes. One approach is the selection and estimation of an individual concentration-response curve for each gene. Under appropriate assumptions, the resulting estimators will be unbiased, but very complex (as we have thousands of genes). Additionally, this approach can be wasteful, since the different genes might have similar properties and certain aspects of the concentration-response curves for different genes might be similar, suggesting a borrowing of strength. For instance, using the similarity of different genes could reduce the complexity of the joint model and improve the quality of the resulting estimators.

In this thesis, classical model selection criteria (AIC and BIC) shall be applied to the setting of gene expression data (VPA data set) in order to select a model with appropriate complexity. More precisely, the conservative approach where the different concentration-response relationships are considered separately should be compared to approaches where the curves of the different genes share certain parameter values. Furthermore, the (asymptotic) properties (like the probability to select for the true model) of the model selection criteria under investigation should be analysed. This can either be done by simulation or by analytical methods.

Further information: Please contact schorning@statistik.tu-dortmund.de