

Optimal Design for Linear and Non-linear Models

R. Schwabe

This tutorial is meant to provide an introductory course to basic principles in the theory of optimal design of experiments. We start with a general linear model setup, which includes regression and analysis of variance as special cases as well as more sophisticated functional relationships. The objective of optimal design is to determine experimental settings which optimize the quality of statistical analyses to be performed after data collection. For measuring the quality the information matrix plays a crucial role. In order to make the information comparable obtained from competing designs, various design criteria are introduced.

The design problem is embedded in a convex optimization setting ("approximate theory"), which allows for characterizations of optimal designs by so-called equivalence theorems. This approach also provides numerical algorithms to generate nearly optimal designs and bounds for measuring the efficiency of a given design. In this context common principles like invariance ("symmetry") and majorization can be applied to reduce the complexity of the optimization problem.

In non-linear models usually the quality of the statistical analyses can only be measured by asymptotic criteria. To this end the information matrix is either obtained from a linearized model or, if available, as the exact Fisher information matrix, which is proportional to the inverse of the asymptotic covariance matrix of the maximum likelihood estimator. If appropriate, estimation equations (e.g. quasi-likelihood) may be used instead. The coincidence of these approaches is established for the class of generalized linear models including logistic and Poisson regression. Due to the non-linearity optimal designs may depend on the true values of the parameters, i.e. on the true shape of the response. To avoid this complication weighted and minimax criteria are proposed.

While so far the theory was developed for models, in which only fixed effects are present, we introduce some simple random effects, as they typically occur in statistical applications in biosciences, where individuals (humans or animals) are involved and multiple measurements are obtained from these individuals. This means, in general, that each individual has its own mean response curve,

which randomly deviates from a mean population response curve. Starting from models with random intercept we switch over to general random coefficients and derive optimality concepts for population parameters ("averaged" over the individuals), for prediction of individual response curves and for the analysis of the variance components. We end up with some critical remarks on frequently used optimality criteria.