Using machine learning for study planning in psychology

R. Debelak[1] and F. Zimmer[1]

[1]University of Zurich, Department of Psychology, Binzmühlestrasse 14, Zurich 8050, Switzerland

The planning of an empirical study in psychology always includes the planning of the sample collection. This requires the balancing of several factors: On the one hand, the data collection should be as inexpensive as possible, and on the other hand, the collected data set must be large enough to reliably detect suspected effects. Furthermore, planning the data collection may involve complex decisions, such as determining the number of groups and individuals per group in a multilevel design. These decisions, in turn, affect the cost and power of statistical tests. Solving these problems generally requires the use of computationally intensive Monte Carlo simulations, which can benefit tremendously from being guided by algorithms.

Based on ideas originally proposed in biostatistics (e.g., [1]), we present a new method that supports finding an optimal study design using machine learning methods. The proposed method is implemented in the R package mlpwr [2]. Possible applications are a) finding a design with fixed power that is as inexpensive as possible, and b) finding a design with as high power as possible given a fixed cost budget. The new method is demonstrated using some scenarios from classical statistics, multilevel modeling, and latent variable models. In summary, we find that the application of the new method is promising and more efficient than alternative approaches if cost information is present.

[1] D. T. Wilson, R. Hooper, J. Brown, A. J. Farrin, and R. E. Walwyn, "Efficient and flexible simulation-based sample size determination for clinical trials with multiple design parameters," *Statistical Methods in Medical Research*, vol. 30, no. 3, pp. 799–815, 2021.

[2] F. Zimmer and R. Debelak, *mlpwr: A Power Analysis Toolbox to Find Cost-Efficient Study Designs*, 2022. R package version 1.0.0.