Kolloquium „Statistische Methoden in der empirischen Forschung“

Wann: 27. Januar 2026, 17:00 – 18:30 Uhr

Wo: Campus Charité Mitte, Raum 02.002, Virchowweg 10, 10117 Berlin

Online-Übertragung: der Link wird auf der Website zur Verfügung gestellt

Vortragssprache: Deutsch

Katja Ickstadt (Technische Universität Dortmund)

Variable Selection in High-Dimensional Omics Applications – Bayesian and Non-Bayesian Ideas

High-dimensional omics data pose significant challenges for variable selection due to far more features than samples, strong correlations among predictors, and typically modest effect sizes, leading to unstable selections and difficult scientific interpretation. This talk contrasts non-Bayesian and Bayesian approaches, emphasizing strategies for improved reliability. Non-Bayesian methods start with univariate screening combined with multiple-testing corrections as a simple baseline, progressing to multivariate penalized regression techniques such as LASSO and Elastic Net. These highlight the critical role of hyperparameter tuning and handling of correlation structures in balancing predictive accuracy and interpretability. A complementary scalable approach employs cross-leverage scores to rank variables based on design geometry, facilitating detection of interaction effects in large-scale genetic studies without exhaustive pairwise enumeration. Efficiency is enhanced through random batching, sliding windows, or sketching approximations. Bayesian frameworks leverage global-local shrinkage and sparse priors, including the Bayesian LASSO, horseshoe, spike-and-slab, and regularized horseshoe, which aggressively shrink near-zero effects while preserving large signals and enabling uncertainty quantification via posterior inference and decision rules. Future developments point toward variational inference for scaling Bayesian methods to massive datasets and integrated decision frameworks that align scientific goals, data characteristics, and computational resources to promote robust, reproducible variable selection workflows. These ideas are demonstrated through extensive simulations motivated by identifying aging-associated proteins in cerebrospinal fluid (CSF) proteomics, complemented by analyses of a real CSF proteomics dataset.