MEDICAL UNIVERSITY
OF VIENNA

Wiener Biometrische Sektion
der Internationalen Biometrischen Gesellschaft
Region Österreich – Schweiz

Please join the Biometric Colloquium

Gregor Buch

Johannes Gutenberg University Mainz, Germany

Overview, evaluation, and development of group variable selection methods for knowledge integration

February 15th, 2023 at 9:00 am

Seminarraum Center for Medical Data Science (previously CeMSIIS),
Spitalgasse 23, Room 88.03.513
Medical University of Vienna, 1090 Wien

Host: Georg Heinze

Abstract

Introduction

Many datasets have a natural group structure due to high correlations or contextual similarities of variables. Group variable selection methods can account for such structure in the selection process to identify variables that are related to each other and share a common and traceable relationship with the response variable. There is a need for a systematic review of implemented approaches, a fair comparison and evaluation of these techniques, and the development of improved methods for omics datasets.

Methods

A systematic literature search was conducted to identify group variable selection methods implemented in R. The selection performance of the identified methods was evaluated within simulation studies. A subset of the best performing approaches was used to select predictors in a time-to- event, regression, and classification task in bootstrap samples from a prospective cohort study (MyoVasc; NCT04064450). The Sparse Group Exponential Penalty (SGE) was proposed to address the limitations of existing methods. Its performance was compared to established techniques in simulation studies and applied to data from a randomized clinical trial (EmDia; NCT02932436).

Results

The systematic review revealed 14 methods, which were classified into knowledge-driven and data-driven approaches. The first category includes group-level and bi-level selection methods, while twostep and collinear tolerant approaches constitute the second category. Simulation studies show the advantage of bi-level selection methods over the other approaches. In the real-world scenario, the bi-level selection methods were also shown to outperform the traditional LASSO, as they were able to

treat variables of a group consistently and select correlated variables together.
SGE demonstrated superiority in variable and group selection in almost all settings where the number of observations exceeded the number of variables. In cases where there were fewer observations than variables, SGE was the best bi-level selection method when few groups contained predictive signals.

Conclusions

A variety of methods can incorporate a group structure of predictors in the selection process. The choice of the most appropriate method is dependent on the specific research question and demands careful consideration. Bi-level selection methods, particularly the SGE, appear promising for exploratory analysis of grouped omics data, such as lipidomics data.