Invited sessions for ISCB42 (Lyon, 2021)
Session 6

Selective inference after variable selection

Organizers : **Georg Heinze : Medical University of Vienna, AU**
Els Goetghebeur : Ghent University, BE

Oliver Dukes : Ghent university, BE
Valid post-selection inference for cox regression parameters, with and without the proportional hazards assumption

Michael Kammer : Section for Clinical Biometrics, Centerfor Medical Statistics, Informatics and Intelligent Systems, Medical University of Vienna, Austria
Selective inference for the Lasso in statistical practice

Lisa McShane (Biometric Research Program, National Cancer Institute, National Institutes of Health, U.S.A)
Biologically-informed development of treatment selection scores from high-dimensional omics data

Valid post-selection inference for cox regression parameters, with and without the proportional hazards assumption.

Oliver Dukes[1], Kelly Van Lancker[1], Stijn Vansteelandt[1,2].

[1] Department of Applied Mathematics, Computer Science and Statistics, Ghent University, Belgium).
[2] Department of Medical Statistics, London School of Hygiene and Tropical Medicine, UK.

The problem of how to best select variables for confounding adjustment forms one of the key challenges in the evaluation of exposure or treatment effects in observational studies. A major drawback of common selection procedures is that there is no finite sample size at which they are guaranteed to deliver tests and confidence intervals with adequate performance. In this talk, I will consider this problem in the context of estimating a conditional causal hazard ratio in an observational study where all key confounders are measured. An added complication with time-to-event outcomes is that one must adjust not only for confounders, but also variables that render censoring non-informative.

Assuming first that the hazard ratio of interest is constant, I will describe a simple procedure for obtaining valid inference for the conditional hazard ratio after variable selection. This procedure involves three different selection steps, in order to best capture variables that account for confounding and informative censoring, and can be implemented using standard software for the Lasso.

Our proposed estimator (along with common alternatives) has the disadvantage that it may not converge to something easily interpretable when the proportional hazards assumption fails. The resulting tests and confidence intervals also typically lose their validity under misspecification. I will therefore outline an alternative proposal based on a nonparametric estimand that reduces to a Cox model parameter under the proportional hazards assumption, but which continues to capture the association of interest under arbitrary types of misspecification. I will then outline how to obtain valid inference for this novel estimand whilst incorporating variable selection.

The different proposals will be illustrated in the analysis of an observational study on the predictive relationship between the initial concentration of serum monoclonal protein and the progression to multiple myeloma or another plasma-cell cancer.

Selective inference for the Lasso in statistical practice

Michael Kammer[1,2], Daniela Dunkler[1], Stefan Michiels[3], Georg Heinze[1]

[1] Section for Clinical Biometrics, Center for Medical Statistics, Informatics and Intelligent Systems, Medical University of Vienna, Austria (presenting author)
[2] Department for Internal Medicine III, Division of Nephrology and Dialysis, Medical University of Vienna, Austria
[3] Service de Biostatistique et d'Epidémiologie, Gustave Roussy; INSERM, CESP U1018, University Paris-Saclay, France

Nowadays multivariable clinical models are ubiquitous, facilitating personalized medicine and guiding therapy decisions. Such models are often developed by the use of variable selection, but the uncertainty introduced by selection is often ignored in the dissemination. Part of the reason is that classical methods for statistical inference to estimate e.g. confidence intervals are not applicable after selection. One way to facilitate proper post-selection inference is by means of the selective inference framework, which addresses inference for statistical hypotheses that are formulated and analysed using the same set of data. In recent years the methodology was developed for the widely used Lasso method, i.e. L1-penalized regression, but there are also approaches agnostic of the model selection procedure.

We will present some practical considerations when working with the selective inference framework for regression problems. We will discuss a systematic simulation study in linear and logistic Lasso regression. Our focus lies on the properties of selective confidence intervals derived from different approaches, in particular selective coverage, power to exclude zero and stability. We elaborate on the practical use and interpretation of selective inference using two real-data case-studies of typical applications. First, selective inference after the selection of anthropometric features to estimate body fat in men. Second, selective inference for the main exposure after confounder selection in a cardiology dataset.

We found selective inference to be challenging in terms of interpretation and computation. Development of corresponding user-friendly software is still in its infancy. Lasso confidence intervals tended to be very wide and quite variable, but could potentially improve model selection properties, in particular false positive findings. Simple selection agnostic methods showed unsatisfactory trade-offs between selection and inference accuracy, while modern approaches were much more conservative and computationally expensive, limiting their practical usability. In conclusion, selective inference using the Lasso is a promising tool for statistical modelling, but remains difficult to use in practice.

Biologically-informed development of treatment selection scores from high-dimensional omics data

Lisa M. McShane[1,2], Ming-Chung Li[1], George Wright[1], Ting Chen[3], Lori Long[3], Qian Xie[4], Jyothi Subramanian[3], Zhiwei Zhang[1], Yingdong Zhao[1]

[1]Biometric Research Program, National Cancer Institute, National Institutes of Health, U.S.A.

[2]Presenting author

[3]Emmes Corporation, USA (under contract to the U.S. National Cancer Institute)

[4]General Dynamics Information Technology, USA (under contract to the U.S. National Cancer Institute)

Precision medicine therapeutic approaches rely on matching mechanism of action of a therapy to biological and molecular characteristics of a patient or the patient's disease. Therapies that can correct for aberrant or missing gene products or compensate for a disrupted biological pathway hold promise for the treatment of the corresponding disease. In oncology, many predictors based on multivariable scores generated from high dimensional omics data have been developed for purposes of prognosis; some of those have been secondarily assessed for their value in informing choice between different therapy options. Repurposing a prognostic predictor may be suboptimal because there are fundamental differences between the goals of prognostication and therapy selection. The modified covariates method of Tian and colleagues (*JASA* 2014;109:2350-2358) is one approach that has been proposed specifically for development of a therapy selection predictor. Biologically informed enhancements of the modified covariates approach that use information about biological pathways significantly associated with outcome or that use pre-specified variable groupings are proposed in this talk. These biologically informed approaches are found to yield treatment selection predictors with improved performance relative to that of the original modified covariates method on some real omics data from patients with cancer. The discussion additionally highlights challenges in identification of data that are suitable for treatment selection predictor development as well as issues to consider in selection of metrics to evaluate performance of these predictors.