Generalized Data Thinning Using Sufficient Statistics

Jacob Bien

(University of Southern California, Los Angeles)

12.06.2023, 15.00 (s.t.)

Department of Statistics, Ludwigstr. 33, Room 144
and online via Zoom (Link)
(Meeting-ID: 913-2473-4411; Password: StatsCol22)

Sample splitting is one of the most tried-and-true tools in the data scientist toolbox. It breaks a data set into two independent parts, allowing one to perform valid inference after an exploratory analysis or after training a model. A recent paper (Neufeld, et al. 2023) provided a remarkable alternative to sample splitting, which the authors showed to be attractive in situations where sample splitting is not possible. Their method, called convolution-closed data thinning, proceeds very differently from sample splitting, and yet it also produces two statistically independent data sets from the original. In this talk, we will show that sufficiency is the key underlying principle that makes their approach possible. This insight leads naturally to a new framework, which we call generalized data thinning. This generalization unifies both sample splitting and convolution-closed data thinning as different applications of the same procedure. Furthermore, we show that this generalization greatly widens the scope of distributions where thinning is possible. This work is a collaboration with Ameer Dharamshi, Anna Neufeld, Keshav Motwani, Lucy Gao, and Daniela Witten.

Biography:
Jacob Bien is Associate Professor of Data Sciences and Operations & Dean's Associate Professor in Business Administration at University of Southern California. His research focuses on statistical machine learning and in particular the development of novel methods that balance flexibility and interpretability for analyzing complex data. He combines ideas from convex optimization and statistics to develop methods that are of direct use to scientists and others with large datasets. His work has been supported by an NSF CAREER award, a three-year NSF grant on high-dimensional covariance estimation, an NIH R01 grant on methods for multi-view data, and grants from the Simons Foundation on developing new statistical methodology for oceanography. He serves as an associate editor of the Journal of the American Statistical Association and the Journal of the Royal Statistical Society (Series B), and he was previously an associate editor for Biometrika, the Journal of Computational and Graphical Statistics, and Biostatistics. Before joining USC, he was an assistant professor at Cornell.

Workshop

Writing R packages with literate programming

Jacob Bien

(University of Southern California, Los Angeles)

13.06.2023, 16.00 (s.t.)

Department of Statistics, Ludwigstr. 33, Room 144
and online via Zoom (Link)
(Meeting-ID: 913-2473-4411; Password: StatsCol22)

In this workshop, we will demonstrate a simple new approach to writing R packages. The idea is that you write a single R markdown file and then when you press knit, an R package will be created in addition to the usual html file. Using this approach means that you can present your code in a logical, well explained fashion, complete with latex equations, section headings, figures, and examples. Your research collaborators and your future self will find it easier to understand, modify, and extend code written this way. Feel free to bring a laptop if you'd like to try it out during the workshop. For more on the approach we will be using, see here: `https://jacobbien.github.io/litr-project/`.