Statistical Science
2020, Vol. 35, No. 3, 545–554
https://doi.org/10.1214/20-STS804
Main article: https://doi.org/10.1214/20-STS786

Rejoinder: On nearly assumption-free tests of nominal confidence interval coverage for causal parameters estimated by machine learning

Lin Liu, Rajarshi Mukherjee and James M. Robins

We thank the editors for this opportunity and the discussants Kennedy, Balakrishnan and Wasserman (2020) (abbreviated as KBW in the sequel) for their insightful commentaries on our paper (Liu, Mukherjee and Robins, 2020) (abbreviated as LMR in the sequel).

1. A BRIEF INTRODUCTION TO HIGHER ORDER INFLUENCE FUNCTIONS

We would like to start our rejoinder by responding to the philosophical comments in Section 6 of KBW's discussion before getting into the other more technical comments. In Section 6, KBW divide statistical procedures into structure-driven and methods-driven but also acknowledge that the boundary between these two categories is blurry. For example, even for the poster child of the methods-driven tools – deep neural networks – one common research direction is to prove some form of optimality or robustness under some assumptions, often quantified by smoothness, sparsity or other related complexity measures such as metric entropy (Schmidt-Hieber, 2020, Hayakawa and Suzuki, 2020, Barron and Klusowski, 2018).

The discussants then state that higher order influence function (HOIF) based methods are 'structure-driven' because 'they typically rely on carefully constructed series estimates' and achieve 'better performance over appropriate Hölder spaces potentially at the expense of being more structure driven.' This statement misunderstands the motivation and goals of HOIF estimation. Our goal has always been to make HOIF fully methods-driven. However, before we reach this goal, difficult open problems remain to be solved. Until then, we have had to make restrictive assumptions to obtain sharp mathematical results – these assumptions can make our methodology appear at least partly 'structure-driven'.

The theory of HOIF is (simplifying somewhat) a theory based only on higher order scores of finite dimensional submodels. As a consequence, the theory by itself cannot quantify the rates of convergence of a HOIF estimator and thus the bias of a HOIF estimator without additional complexity reducing model assumptions, a central point we stressed throughout LMR. To be more concrete, for now let us restrict the attention to smooth nonlinear functionals $\psi(\theta)$ of a distribution P_θ lying in an infinite dimensional model $\mathcal{M} = \{\mathsf{P}_\theta; \theta \in \Theta\}$ with a first order influence function $\mathbb{IF}_{1,\psi}(\theta)$ but (as is generally the case in infinite dimensional models) without mth order influence functions for $m > 1$. Therefore, HOIF theory often considers finite $k = k(n)$-dimensional sieves $\mathcal{M}_{\text{sub},k} = \{\mathsf{P}_\theta; \theta \in \Theta_{\text{sub},k} \subset \Theta\}$ containing an initial training sample estimator $\hat{\theta}$, an associated projection map $\theta \mapsto \tilde{\theta}_k$ from Θ onto $\Theta_{\text{sub},k}$ that is the identity for $\theta \in \Theta_{\text{sub},k}$. The projection map defines a truncated parameter $\tilde{\psi}_k(\theta), \theta \in \Theta$ by $\tilde{\psi}_k(\theta) = \psi(\tilde{\theta}_k(\theta)), \theta \in \Theta$, which will typically have HOIFs of all orders because $\Theta_{\text{sub},k}$ is finite dimensional. The theory of HOIF applied to the parameter $\tilde{\psi}_k(\theta)$ guarantees that $\{\tilde{\psi}_k(\hat{\theta}) + \mathbb{E}_\theta[\mathbb{IF}_{m,\tilde{\psi}_k}(\hat{\theta})]\} - \tilde{\psi}_k(\theta) = O(\|\hat{\theta} - \theta\|^{m+1})$ or, equivalently,

$$\mathbb{E}_\theta\big[\hat{\psi}_{m,k} - \tilde{\psi}_k(\theta)\big] \equiv \mathsf{EB}_{\theta,m,k}(\hat{\psi}_1) = O\big(\|\hat{\theta} - \theta\|^{m+1}\big)$$
$$\text{where } \hat{\psi}_{m,k} = \psi(\hat{\theta}) + \mathbb{IF}_{m,\tilde{\psi}_k}(\hat{\theta}).$$

Here $\hat{\psi}_1 = \psi(\hat{\theta}) + \mathbb{IF}_{1,\tilde{\psi}_k}(\hat{\theta})$ is a doubly robust machine learning (DRML) estimator based on the first order influence function and $\mathbb{IF}_{m,\tilde{\psi}_k}(\hat{\theta}) = \mathbb{IF}_{1,\tilde{\psi}_k}(\hat{\theta}) - \sum_{j=2}^{m} \mathbb{IF}_{jj,\tilde{\psi}_k}(\hat{\theta})$ where, under $\mathsf{P}_{\hat{\theta}}$, $\mathbb{IF}_{jj,\tilde{\psi}_k}(\hat{\theta}) \equiv \widehat{\mathbb{IF}}_{jj,k}$ is a jth order U-statistic.[1] Unless stated otherwise all expectations are conditional on the training sample.

Lin Liu is Assistant Professor, Institute of Natural Sciences, School of Mathematical Sciences and SJTU-Yale Joint Center for Biostatistics, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: linliu@sjtu.edu.cn). Rajarshi Mukherjee is Assistant Professor, Department of Biostatistics, Harvard T. H. Chan School of Public Health, Boston, Massachusetts 02115, USA (e-mail: ram521@mail.harvard.edu). James M. Robins is Professor, Department of Epidemiology and Biostatistics, Harvard T. H. Chan School of Public Health, Boston, Massachusetts 02115, USA (e-mail: robins@hsph.harvard.edu).

[1] Here we are using the same sign convention as in LMR, which reverses the sign conventions of Robins et al. (2008).