Statistical Science
2019, Vol. 34, No. 4, 604–605
https://doi.org/10.1214/19-STS756
Main article: https://doi.org/10.1214/18-STS693, https://doi.org/10.1214/18-STS694

Discussion: Models as Approximations

Dalia Ghanem and Todd A. Kuffner

We congratulate the authors on these illuminating and thought-provoking articles.

1. SCORING RULES AND WELL-SPECIFICATION IN PREDICTIVE MODELING

In the context of prediction, the objective is often to minimize a particular criterion or scoring rule. If the conditional distribution is known and correctly specified, then maximum likelihood is the criterion that should be used for estimation, even if the Kullback–Leibler divergence is not the scoring rule that the forecaster has chosen to minimize. In the more likely case of misspecification, it is not clear which criterion should be used for estimation. In the context of forecasting conditional probabilities of binary outcomes, Elliott, Ghanem and Krüger (2016) examine this question and illustrate that the choice of scoring rule yields different best approximations to the true conditional probability function of the outcome of interest under misspecification, except under restrictive conditions. Interestingly, these conditions under which the choice of objective function used for estimation does not change the best approximation to the true conditional probability function imposes symmetry conditions on the regressor distribution as well as the conditional mean.

2. CAUSAL INFERENCE, WELL-SPECIFICATION AND EXTERNAL VALIDITY

In causal inference, we often consider fully nonseparable models. For simplicity, we will consider the case where $\vec{X} = T$, where T is a scalar binary variable, which we refer to as the treatment variable. The model equation is specified as

$$Y = m(T, \vec{U}). \tag{1}$$

In this simple example, the structural function could be written in terms of potential outcomes (Imbens and Rubin, 2015),

$$Y = (1 - T)Y_0 + TY_1, \tag{2}$$

where $Y_0 = m(0, \vec{U})$ and $Y_1 = m(1, \vec{U})$. We can also write the conditional distribution of Y given T in terms of the potential outcomes,

$$\begin{aligned} F_{Y|T}(y|t) &= \int 1\{m(t, \vec{u}) \le y\}\, dF_{\vec{U}|T=t} \\ &= F_{Y_t|T}(y|t), \quad \text{for } t = 0, 1. \end{aligned} \tag{3}$$

Due to potential selection, $\vec{U}$ may not be independent of T in general. In this case, we cannot identify the average treatment effect (ATE), that is,

$$\begin{aligned} \Delta(P) &= E_P[Y_1 - Y_0] \\ &= \int \big(m(1, \vec{u}) - m(0, \vec{u})\big)\, dF_{\vec{U}}, \end{aligned} \tag{4}$$

from the observed difference in mean outcomes,

$$\begin{aligned} &E_P[Y|T = 1] - E_P[Y|T = 0] \\ &\quad = \int m(1, \vec{u})\, dF_{\vec{U}|T=1} - \int m(0, \vec{u})\, dF_{\vec{U}|T=0}. \end{aligned} \tag{5}$$

Hence, the independence of U and T are critical, that is, $F_{\vec{U}|T} = F_{\vec{U}}$, for the identification of the ATE in (4) from the difference in mean outcomes in (5).

Randomized experiments or randomly assigned interventions ensure this independence assumption holds and thereby allow us to learn about *internally valid* estimands of causal impact, such as the ATE. In general, these estimands may not be *externally valid* (for further discussion of the distinction between internal and external validity, see Athey and Imbens, 2017). For instance, the ATE in a given experiment can depend on the environment. To allow for this dependence in our notation, let $e \in \mathcal{E}$ denote an environment, then

$$\begin{aligned} Y^e &= m^e(T^e, \vec{U}^e), \\ \Delta^e(P^e) &= E_{P^e}[Y_1^e - Y_0^e] \\ &= \int \big(m^e(1, \vec{u^e}) - m^e(0, \vec{u^e})\big)\, dF_{\vec{U}^e}, \end{aligned} \tag{6}$$

Dalia Ghanem is Assistant Professor, Department of Agricultural and Resource Economics, University of California, Davis, One Shields Avenue, Davis, California 95616, USA (e-mail: dghanem@ucdavis.edu). Todd A. Kuffner is Associate Professor, Department of Mathematics and Statistics, Washington University in St. Louis, 1 Brookings Dr., St. Louis, Missouri 63131, USA (e-mail: kuffner@wustl.edu).

where $\Delta^e(P^e)$ denotes the ATE for the environment e. Note that every function in the above may vary with e. The above notation also clarifies that while the independence between T and $\vec{U}^e$ ensures that we identify the ATE for the environment e, that is, an internally valid estimand, it is not sufficient for $\Delta^e(P^e) = \Delta_0$ for all $e \in \mathcal{E}$, a non-singleton set.

The authors point to an interesting connection between well-specification in the context of causal inference and invariance to regressor distributions (Peters, Bühlmann and Meinshausen, 2016). We conjecture that this connection relates to external validity. We provide a simple example to support our conjecture. To do so, we present the assumptions maintained in Peters, Bühlmann and Meinshausen (2016), while adapting their notation slightly to remain consistent with ours. We observe i.i.d. realizations of (X^e, Y^e) in each environment, where $X^e \in \mathbb{R}^p$ and $Y^e \in \mathbb{R}$ is the target variable. Peters, Bühlmann and Meinshausen (2016) assume that if a subset $S^* \subseteq \{1, \dots, p\}$ is causal, then

$$(7) \qquad Y^e = g(\vec{X}^e_{S^*}, \vec{\epsilon^e}), \quad \vec{\epsilon^e} \sim F_{\vec{\epsilon}}, \vec{\epsilon^e} \perp \vec{X}^e_{S^*}.$$

The goal of Peters, Bühlmann and Meinshausen (2016) is to use different types of interventions in different environments for causal identification. Here we will apply their assumption to our treatment effect problem, where we only consider interventions that randomly assign T^e within each environment, that is, $T^e \perp (\vec{Z}^e, \vec{\epsilon}^e)$ for all $e \in \mathcal{E}$. To do so, we let $\vec{X}^e = (T^e, \vec{Z}^e)'$, assuming (7) implies that

$$(8) \qquad Y^e = g((T^e, \vec{Z}^e)', \vec{\epsilon^e}).$$

We can identify the following:

$$\begin{aligned} &E_{P^e}[Y_1^e - Y_0^e | \vec{Z}^e = \vec{z}] \\ (9) \qquad &= \int (g((1, \vec{z}')', v) - g((0, \vec{z}')', v)\, dF_{\vec{\epsilon}}(v) \\ &= CATE(\vec{z}), \end{aligned}$$

which is the conditional average treatment effect given $\vec{Z}^e = \vec{z}$. As we can see from the first equality, this object does not vary across environments. Hence, it is not only internally but also externally valid for $e \in \mathcal{E}$. This is not surprising, since (7) assumes that the function g, the distribution of unobservables $F_{\vec{\epsilon}}$ and the regressors $\vec{Z}^e$ are the same across environments, even though the distribution of $\vec{Z}^e$ can vary arbitrarily across environments.

Even though in the above example $CATE(\vec{z})$ is externally valid, we can easily show that this is not true for the ATE. To see this, note that

$$(10) \qquad \Delta^e(P^e) = \int CATE(\vec{z})\, dF_{\vec{Z}^e}(\vec{z}).$$

It clearly varies across environments depending on the distribution of $\vec{Z}^e$. Hence, it is not externally valid.

This simple example supports our conjecture that the relationship between the invariance principle introduced in these papers is indeed related to the concept of external validity. We thank the authors for this important insight.

ACKNOWLEDGMENTS

This material is based upon work supported by NSF Grant DMS-1712940.

REFERENCES

ATHEY, S. and IMBENS, G. W. (2017). The econometrics of randomized experiments. In *Handbook of Economic Field Experiments* **1** 73–140. Elsevier, Amsterdam.

ELLIOTT, G., GHANEM, D. and KRÜGER, F. (2016). Forecasting conditional probabilities of binary outcomes under misspecification. *Rev. Econ. Stat.* **98** 742–755.

IMBENS, G. W. and RUBIN, D. B. (2015). *Causal Inference—for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge Univ. Press, New York. MR3309951

PETERS, J., BÜHLMANN, P. and MEINSHAUSEN, N. (2016). Causal inference by using invariant prediction: Identification and confidence intervals. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **78** 947–1012. MR3557186