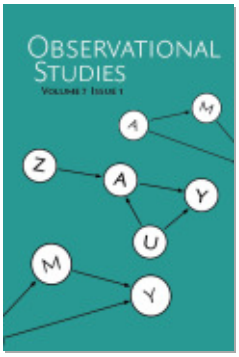
Melting together prediction and inference

Citation for the original published paper (version of record):
Daoud, A., Dubhashi, D. (2021). Melting together prediction and inference. Observational Studies, 7(1): 1-7. http://dx.doi.org/10.1353/obs.2021.0035

Melting together prediction and inference

Adel Daoud, Devdatt Dubhashi



Observational Studies, Volume 7, Issue 1, 2021, pp. 1-7 (Article)

Published by University of Pennsylvania Press
DOI: *https://doi.org/10.1353/obs.2021.0035*

Observational Studies Vol. 7.1 (2021) Published September2021

Melting together prediction and inference

Adel Daoud **adel.daoud@liu.se**
Institute for Analytical Sociology & The Division for Data Science and Artificial Intelligence
Linköping University & Chalmers University of Technology
Norrköping, Sweden & Gothenburg, Sweden

Devdatt Dubhashi **dubhashi@chalmers.se**
The Division for Data Science and Artificial Intelligence
Chalmers University of Technology
Gothenburg, Sweden

Abstract

In Leo Breiman's influential article "Statistical modeling-the two cultures" he identified two cultures for statistical practices. The *data modeling culture* (DMC) denotes practices tailored for statistical inference targeting a quantity of interest, $\widehat{\beta}$. The *algorithmic modeling culture* (AMC) refers to practices defining an algorithm, or a machine-learning (ML) procedure, that generates accurate predictions about an outcome of interest, $\widehat{Y}$. As DMC was the dominant mode, Breiman argued that statisticians should give more attention to AMC. Twenty years later and energized by two revolutions—one in data-science and one in causal inference—a hybrid modeling culture (HMC) is rising. HMC fuses the inferential strength of DMC and the predictive power of AMC with the goal of analyzing cause and effect, and thus, HMC's quantity of interest is causal effect, $\widehat{\tau}$. In combining inference and prediction, the result of HMC practices is that the distinction between prediction and inference, taken to its limit, melts away. While this hybrid culture does not occupy the default mode of scientific practices, we argue that it offers an intriguing novel path for applied sciences.

Keywords: causal inference, prediction, machine learning, data science, statistical cultures

1. Introduction

Breiman (2001a) identified two cultures for statistical modeling. The *data modeling culture* (DMC) refers roughly to practices aiming to conduct statistical inference on one or several quantities of interest. By *unbiased statistical inference*, we mean a procedure tailored to estimates a quantity $\widehat{\beta}$ such that the difference between the true quantity β is as small as possible: the procedure is unbiased when the difference $\beta - \widehat{\beta}$ is negligible in expectation. This true quantity β exists independently of the statistical model producing $\widehat{\beta}$. The *algorithmic modeling culture* (AMC) refers to practices defining a procedure, f, that generates accurate predictions, $\widehat{Y}$, about an event (outcome), Y. By *accurate*, we mean predictions that are as similar as possible to the true event that f has yet not encountered (Hastie et al., 2009). The smaller the difference $Y - \widehat{Y}$, the higher the similarity. A *procedure* is an algorithm, or a function, that takes some input X, operates on this input $f(X)$, and then, produces an output $f(X) = \widehat{Y}$. Often, this procedure are defined in terms of a

machine-learning (ML) algorithm (Hastie et al., 2009). Thus, $\widehat{Y}$-prediction problems and $\widehat{\beta}$-inference problems form distinct sets of practices in mobilizing data, algorithms, and results (Mullainathan and Spiess, 2017).

While Breiman argued that statisticians should devote more attention to $\widehat{Y}$-problems than $\widehat{\beta}$-problems, today a third culture is rising: the *hybrid modeling culture* (HMC). This third culture emerges from statistical practices where prediction and inference synthesize into new procedures (Daoud and Dubhashi, 2020; Kino et al., 2021). As the main concern of HMC is causality, its focus is on, what we denote, $\widehat{\tau}$-problems. Using the Neyman-Rubin causal framework or Pearl's do-calculus, we define causal effects as the difference between potential outcomes $\tau = Y^1 - Y^0$. The potential outcome Y^1 is the outcome when a cause W (e.g., a treatment, policy, or exposure) is active and Y^0 occurs when W is inactive. While the interest in identifying causal effects exist in DMC already, a key difference between DMC-powered models for causal inference and HMC-powered models for causal inference, is that the latter mobilizes the predictive power of ML. In other words, HMC uses tricks from AMC to achieve DMC goals. As these tricks rely on combining inference and prediction, the result of HMC is that the distinction between $\widehat{Y}$ and $\widehat{\beta}$—taken to its limit—melts away. In this commentary, we delineate our "melting away" argument.

2. Machine learning for causal inference

Before describing how ML aids in inferring causality, we will refine our definition of what we mean by causal inference. We define a cause of interest as a binary variable, W. Instead of merely recording each individual's outcome as observed by the data, Y_i, we assume that each individual i has two potential outcomes (Imbens and Rubin, 2015). One potential outcome records the outcome when the individual takes the treatment Y_i^1 and one where he or she does not take it Y_i^0. The causal effect for each individual i is then the difference between these two potential outcomes:

$$\tau_i = Y_i^1 - Y_i^0$$

If we could observe both potential outcomes, we could then directly compute τ_i and thus identify individual-level causal effects. However, the observed outcome—as supplied by the data—is a function of both the treatment and the two potential outcomes, $Y_i = (W-1)\,Y_i^1 + WY_i^0$. This function shows that the observed data reveals only one of these two potential outcomes, yet both are required to identify causal effects. Table 1 exemplifies an observed-data matrix of four individuals with fictitious variable values, and their missing potential outcomes. This impossibility of observing both potential outcomes is known as the *fundamental problem of causal inference*. Much of the causal-method development pertains to defining procedures for when the causal effect is identified from observational data (Hernan and Robins, 2020; Imbens and Rubin, 2015; Pearl, 2009; Peters et al., 2017). By *identified*, we mean a causal effect calculable from measured data.

Table 1: A toy dataset illustrating the fundamental problem of causal inference

	Y	Y^1	Y^0	W	τ	X
Jane	20	20	?	1	?	10
John	30	30	?	1	?	11
Joe	25	?	25	0	?	10
Jan	22	?	22	0	?	11

Vibrant literature in the overlap between computer science, econometrics, and statistics combine ML and causal methodology to develop new estimators, used in various domains (Angrist and Pischke, 2014; Athey and Imbens, 2017; Athey et al., 2019; Chernozhukov et al., 2018; Daoud et al., 2020, 2019; Daoud and Johansson, 2020; Hedström and Manzo, 2015; Hernan and Robins, 2020; Hill, 2011; Hirshberg and Zubizarreta, 2017; Imai, 2018; Kraamwinkel et al., 2019; van der Laan and Rose, 2011; Morgan and Winship, 2014; Pearl and Mackenzie, 2018; Shiba et al., 2021; VanderWeele, 2015). A recurring theme in these methods is the many creative combinations where predictive AMC-type algorithms are used in DMC-type of causal inference. There are several ways in which ML algorithms help the scientific endeavour (for an overview see Daoud and Dubhashi, 2020), but the most important of them is the use of ML to predicting the missing-potential outcomes (Athey and Imbens, 2017).

As observed data only reveal one-half of the potential outcomes, the other half is regarded as missing data. One way of handling this fundamental problem is to cast it as a missing-data problem and proceed to identify conditions for imputing these data to populate all the Y_i^1 and Y_i^0cells, based on covariates X (for a critique of this missing-data definition see Pearl and Mackenzie, 2018). These imputation procedures rely on common identifiability assumptions. One such central assumption is conditional independence (also known as conditional ignorability and conditional exchangability), $W \perp Y_i^1, Y_i^0 | X$. This mathematical statement means that the treatment is as-if randomly assigned conditional on one or more covariates.

Because ML excels in prediction tasks compared to commonly used parametric models, HMC-influenced scholars have developed many different procedures to predict potential outcomes (Künzel et al., 2018). For example, the *T-learner*—"T" stands for *two*—procedure defines one ML-algorithm $f_{w=1}(x_i) = E[Y = y_i | W = 1,\ X = x_i]$ trained on the treated group and another algorithm $f_{w=0}(x_i) = E[Y = y_i | W = 0,\ X = x_i]$ trained on the control group. A Lasso, a random forest, or a collection of algorithms (an ensemble) are often used to define $f_{w=1}$ and $f_{w=0}$. The SuperLearner provides a well-tested framework to mobilize ensembles for causal inference (van der Laan and Rose 2011). After training, $f_{w=1}$ imputes potential outcomes for the control group and $f_{w=0}$ imputes these outcomes for the treated group. Based on the toy data of Table 1, $f_{w=1}$ trains on Jane and John, and imputes Y_i^1 of Joe and Jan; likewise, $f_{w=0}$ trains on Joe and Jan, and imputes Y_i^0 of Jane and John. This procedure culminates by calculating the difference $\widehat{\tau}_i = \widehat{Y}_i^1 - Y_i^0$ for the control group and $\widehat{\tau}_i = Y_i^1 - \widehat{Y}_i^0$ for the treated group, and then averaging over all groups $\widehat{\tau} = E[E[\widehat{\tau}_i | W = w_i]]$ to calculate the average treatment effect.

The T-learner algorithm is one of several, but common to most of these ML algorithms is the procedure of imputing potential outcomes (Künzel et al., 2018) or imputing the treatment effect directly (Athey et al., 2019). Although much research is devoted to analyzing biases arising from ML regularization, these HMC algorithms demonstrate how $\widehat{\tau}_i$-problems subsume $\widehat{\beta}$-problems originating from DMC by mobilizing the algorithmic power of AMC used for $\widehat{Y}$-problems. Thereby, the original distinction between $\widehat{\beta}$ and $\widehat{Y}$ has dissipated—melted away.

3. Conclusions

It is perhaps a historical irony that one of the most popular HMC algorithms, the generalized random forest (Athey et al., 2019), uses a random-forest algorithm as a key ingredient for causal inference; the same algorithm that Breiman (2001a) used to exemplify what AMC-type of predictions had to offer the scientific endeavour. It is the same algorithm he devoted much research in developing (Breiman, 2001b).

Evidently, Breiman's work has opened up new a new perspective not only for statistics but also for applied sciences. This perspective direct our attention towards the possibilities of our time—the era of data science. Twenty years later, based on the advances in machine learning and causal inference, scholars are enabled to move one step further. As identifying causal effects is one of the core goals of the scientific endeavor, we conclude that instead of retaining the dichotomy between AMC-prediction and DMC-inference, this endeavor gains more by embracing both synthetically. HMC provides a way to think about how this synthesis is possible in the era of data science (Meng, 2020) while still maintaining the scientific endeavor's higher goal: explaining reality. Although this hybrid culture does not occupy the default mode of scientific practices, we argue that it offers an intriguing novel path forward for applied sciences (Daoud and Dubhashi, 2020).

Acknowledgments

We would like to thank Fredrik Johansson for constructive comments. The usual disclaimer applies.

References

Joshua D. Angrist and Jörn-Steffen Pischke. *Mastering 'Metrics: The Path from Cause to Effect*. Princeton University Press, Princeton ; Oxford, with french flaps edition edition, 12 2014. ISBN 978-0-691-15284-4.

Susan Athey and Guido W. Imbens. The state of applied econometrics: Causality and policy evaluation. *Journal of Economic Perspectives*, 31(2):3–32, 5 2017. ISSN 0895-3309. doi: 10.1257/jep.31.2.3.

Susan Athey, Julie Tibshirani, and Stefan Wager. Generalized random forests. *The Annals of Statistics*, 47(2):1148–1178, 4 2019. ISSN 0090-5364, 2168-8966. doi: 10.1214/18-AOS1709.

Leo Breiman. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, 16(3):199–231, 8 2001a. ISSN 0883-4237, 2168-8745. doi: 10.1214/ss/1009213726. MR: MR1874152 Zbl: 1059.62505.

Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 5 2001b. ISSN 1573-0565. doi: 10.1023/A:1010933404324.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2 2018. ISSN 1368-4221. doi: 10.1111/ectj.12097. publisher: Oxford Academic.

Adel Daoud and Devdatt Dubhashi. Statistical modeling: the three cultures. *arXiv:2012.04570 [cs, stat]*, 12 2020. URL http://arxiv.org/abs/2012.04570. arXiv: 2012.04570.

Adel Daoud and Fredrik Johansson. Estimating treatment heterogeneity of international monetary fund programs on child poverty with generalized random forest. *SocArXiv*, 2020. doi: osf.io/preprints/socarxiv/awfjt.

Adel Daoud, Rockli Kim, and S. V. Subramanian. Predicting women's height from their socioeconomic status: A machine learning approach. *Social Science and Medicine*, 238: 112486, 10 2019. ISSN 0277-9536. doi: 10.1016/j.socscimed.2019.112486.

Adel Daoud, Anders Herlitz, and S. V. Subramanian. Combining distributive ethics and causal inference to make trade-offs between austerity and population health. *arXiv:2007.15550 [econ, q-fin]*, 8 2020. URL http://arxiv.org/abs/2007.15550. arXiv: 2007.15550.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer New York, 2 edition edition, 8 2009. 00007.

Peter Hedström and Gianluca Manzo. Recent trends in agent-based computational research a brief introduction. *Sociological Methods and Research*, 44(2):179–185, 5 2015. ISSN 0049-1241, 1552-8294. doi: 10.1177/0049124115581211. 00002.

Miguel A. Hernan and James M. Robins. *Causal Inference*. CRC Press, 1st edition edition, 2020. ISBN 978-1-4200-7616-5.

Jennifer L. Hill. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 1 2011. ISSN 1061-8600. doi: 10.1198/jcgs.2010.08162.

David A. Hirshberg and José R. Zubizarreta. On two approaches to weighting in causal inference. *Epidemiology*, 28(6):812, 11 2017. ISSN 1044-3983. doi: 10.1097/EDE.0000000000000735.

Kosuke Imai. *Quantitative Social Science: An Introduction*. Princeton University Press, Princeton, illustrated edition edition, 2 2018. ISBN 978-0-691-17546-1.

Guido W. Imbens and Donald B. Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, New York, 1 edition edition, 4 2015. ISBN 978-0-521-88588-1. 00005.

Shiho Kino, Yu-Tien Hsu, Koichiro Shiba, Yung-Shin Chien, Carol Mita, Ichiro Kawachi, and Adel Daoud. A scoping review on the use of machine learning in research on social determinants of health: Trends and research prospects. *SSM - Population Health*, 15: 100836, 9 2021. ISSN 2352-8273. doi: 10.1016/j.ssmph.2021.100836.

Sören R. Künzel, Jasjeet S. Sekhon, Peter J. Bickel, and Bin Yu. Meta-learners for estimating heterogeneous treatment effects using machine learning. *arXiv:1706.03461 [math, stat]*, 10 2018. URL http://arxiv.org/abs/1706.03461. arXiv: 1706.03461.

Nadine Kraamwinkel, Hans Ekbrand, Stefania Davia, and Adel Daoud. The influence of maternal agency on severe child undernutrition in conflict-ridden nigeria: Modeling heterogeneous treatment effects with machine learning. *PLOS ONE*, 14(1):e0208937, 1 2019. ISSN 1932-6203. doi: 10.1371/journal.pone.0208937.

Xiao-Li Meng. What is your list of 10 challenges in data science? *Harvard Data Science Review*, 7 2020. ISSN ,. doi: 10.1162/99608f92.a3e88876. URL https://hdsr.mitpress.mit.edu/pub/9ptj5iu7/release/2. publisher: PubPub.

Stephen L. Morgan and Christopher Winship. *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. Cambridge University Press, New York, NY, 2 edition edition, 11 2014. ISBN 978-1-107-69416-3. 00000.

Sendhil Mullainathan and Jan Spiess. Machine learning: An applied econometric approach. *Journal of Economic Perspectives*, 31(2):87–106, 5 2017. ISSN 0895-3309. doi: 10.1257/jep.31.2.87.

Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, Cambridge, U.K. ; New York, 2nd edition edition, 9 2009. ISBN 978-0-521-89560-6.

Judea Pearl and Dana Mackenzie. *The Book of Why: The New Science of Cause and Effect*. Basic Books, New York, 1 edition edition, 5 2018. ISBN 978-0-465-09760-9.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. The MIT Press, Cambridge, Massachuestts, 11 2017. ISBN 978-0-262-03731-0.

Koichiro Shiba, Jacqueline M. Torres, Adel Daoud, Kosuke Inoue, Satoru Kanamori, Taishi Tsuji, Masamitsu Kamada, Katsunori Kondo, and Ichiro Kawachi. Estimating the impact of sustained social participation on depressive symptoms in older adults. *Epidemiology*, 9 2021. ISSN 1044-3983. doi: 10.1097/EDE.0000000000001395. URL https://journals.lww.com/epidem/abstract/9000/estimating_the_impact_of_sustained_social.98254.aspx. [Online; accessed 2021-09-15].

Mark van der Laan and Mark Rose. *Targeted Learning - Causal Inference for Observational*. Springer, 2011. 00310.

Tyler VanderWeele. *Explanation in Causal Inference: Methods for Mediation and Interaction*. Oxford University Press, Oxford, New York, 4 2015. ISBN 978-0-19-932587-0.