The Counterfactual χ-GAN: An Adversarial Method To Support Covariate Balance

Amelia J Averitt, MPH MA MPhil[1], Adler J. Perotte, MD MA[1]
[1]Columbia University Department of Biomedical Informatics, New York City, NY, USA

Is this the first time you have submitted your work to be displayed at any OHDSI Symposium? No.

ABSTRACT. *In biomedicine, causal inference often relies on the counterfactual framework, which requires that treatment assignment is independent of the outcome, known as strong ignorability. Observational data typically does not satisfy this assumption, but can be manipulated so that strong ignorability is upheld. One such manipulation is weighting. However, many weighting methods estimates can be unstable. This research proposes a generative adversarial network (GAN)-based model called the Counterfactual χ-GAN (cGAN) to learn stable, feature-balancing weights. cGAN utilizes a unique generator which balances two objective functions; the first encourages coverage over units, and the second identifies the weights and enforces strong ignorability.*

BACKGROUND. Observational data is a potentially very valuable source from which to generate causal knowledge, but it suffers from complexities such as poor quality, irregular sampling, and biases that undermine its use in causal estimation. (1,2) Methods to correct for these biases would permit for new avenues of causal inquiries to supplement experimentation.

The calculation of unbiased causal estimates from this imperfect data source is often framed as the identification of a *natural experiment*. Natural experiments are a type of observational study in which researchers do not have the ability to assign the treatment, but treatments are nonetheless assigned nearly randomly. This is analogous to enforcing the assumption of *strong ignorability,* which is often employed under the Rubin model of counterfactual inference. (3) This assumption states a unit's assignment to a treatment is independent of that unit's potential outcomes, and that treatment assignment is, therefore, ignorable. The identification of the natural experiment often focuses on manipulating the non-randomized, observational data such that the assumption of strong ignorability is upheld. Popular manipulations which support counterfactual inference and the approximation of natural experiments include (i) matching, in which treatment units are paired with similar comparator units based on the pre-treatment covariates (3–6); and (ii) weighting, in which units are disproportionally considered so that the weighted expectation of covariates are similar across arms (7–9). Both manipulations result in pseudo-populations in which covariates are equivalent between arms (*covariate-balance*) and the *unconditional form* of strong ignorability is upheld. Under a matching procedure, units may go unpaired, which is inefficient and may introduce new bias. (10,11) Weighting is a more efficient method for identifying a natural experiment. However, under many weighting techniques, downstream estimates may be unstable.

This research proposes the Counterfactual χ-GAN (cGAN), a variation on a generative adversarial network (GAN) that learns stable covariate-balancing weights for two or more treatment and comparator arms. The neural networks that underlie GANs may be both powerful and flexible enough to balance the rich covariates and relationships that lead to the unconditional form of strong ignorability. We demonstrate the effectiveness of cGAN in achieving covariate balance relative to state-of-the-art methods in simulation, in which the ground truth is known, and with real-world medical data the Observational Health Data Sciences and Informatics (OHDSI) instance of a large, academic medical center

THE MODEL. Traditional-GANs simplify the often-troublesome task of modeling a generative distribution by pitting two neural networks – one discriminator and one generator -- against each other in a zero-sum game to minimize the objective function of Jensen-Shannon divergence between the training data and the generative distribution. Assuming that units for which strong ignorability is upheld arise from a single generative distribution, we may draw upon the traditional-GAN's training process to model the distribution of strong ignorability (the shared distribution) and – as a byproduct -- aid in the learning of covariate-balancing weights to support natural experimentation.

Like the traditional-GAN, the cGAN has a generator, which parameterizes the target distribution. However, unlike the traditional-GAN, there is a discriminator for each treatment group, and each treatment group is parameterized by weights. Also unlike the traditional-GAN, our model minimizes the χ-divergence, which we show produces minimal variance importance sampling estimates. The cGAN balances two objective functions; the first identifies a target distribution that minimizes the importance sampling variance and encourages coverage, and the second identifies the importance sampling weights and enforces the unconditional form of strong ignorability.

METHODS. To evaluate the cGAN when the ground truth is known, we applied the model on simulated data of two populations. Each population was comprised of two subpopulations. Each subpopulation contained 5 covariates, drawn from a randomly generated multivariate normal distribution with a normal-Wishart prior distribution. Population 1 was composed of 250 samples from both subpopulation A and subpopulation B; and Population 2 was composed of 250 samples from both subpopulation A and subpopulation C. Because our simulation deliberately constructs populations from a shared subpopulation distribution (A); we would expect points generated from this subpopulation to have higher weights, as they should contribute more to the shared distribution, than units from either subpopulation B or subpopulation C.

We additionally applied the cGAN to two experiments using real-world clinical data from NewYork-Presbyterian Hospital's OHDSI instance. For both experiments, cohorts of comparator drugs were created using the OHDSI's ATLAS tool. The first experiment compares sitagliptin and glimepiride in elderly patients with Type II Diabetes Mellitus; and the second experiment compares atorvastatin and pravastatin in patients with a history of Acute Coronary Syndromes. We evaluate the ability of the cGAN to improve covariate balance by comparing the Absolute Standardized Difference of Means (ASDM) between the treatment and comparator cohorts under different weighting methods. The ASDM is a popular method of assessing cohort similarity, with a lower metric corresponding to improved covariate balance. The ASDM is presented for (i) the unweighted cohorts, and cohorts weighted according to (ii) the cGAN, (iii) inverse propensity scores (IPW), and (iv) clipped-IPW.

RESULTS. Results of the simulation demonstrate that points from the overlapping subpopulation A are both captured by the generator and assigned higher weights. This is confirmed upon an inspection of the average weights of data points per subpopulation. Average weights from Subpopulations 1A and 2A are substantially higher (0.0023 and 0.0023, respectively) than those from Subpopulations 1B and 2C (0.0017 and 0.0017, respectively).

Experimentation shows similarly promising results. The sitagliptin versus glimepiride experiment show that the cGAN improved mean ASDM from the unweighted data and outperformed clipped-IPW (unweighted = 0.0949; IPW = 0.0196; Clipped IPW= 0.0600; and cGAN= 0.0334). Similar results are seen in the atorvastatin versus pravastatin trial, (unweighted = 0.1017; IPW = 0.0150; Clipped IPW= 0.0319; and cGAN= 0.0251). In both experiments, IPW weighting only marginally improved covariate balance over cGAN, which may be attributable to nonoptimal parameter tuning in our model.

DISCUSSION. These results suggest that cGAN is an effective method of learning covariate balancing weights to support counterfactual inference. The cGAN may provide an alternative means to causal inference from observational data. If all potentially confounding variables are assumed to be observed and included as covariates, the performance of the model suggests that average treatment effects borne from cGAN-weighted cohorts would be less biased than those estimates generated from other weighting methods. The cGAN may, therefore, increase the confidence of claims and permit the identification of effective interventions and improve outcomes. Furthermore, the ability to determine and evaluate causal relationships from observational data may support solutions to many critical issues. Questions of causal inference are broadly applicable to domains in which there is actionable uncertainty, and may therefore benefit a wide audience. This research is enriched by the use of OHDSI tools. Our experimental cohorts are defined using the OHDSI common data model, which will allow other institutions to apply the cGAN experiments presented above to their own data. Collecting results from various institutions will aid in our understanding of the generalizability of causal claims produced by cGAN.

REFERENCES

1. Hripcsak G, Albers DJ. Next-generation phenotyping of electronic health records. JAMIA. 2012;20:117–21.
2. Weiskopf NG, Hripscak G, Swaminathan S, Weng C. Defining and measuring completeness of electronic health records for secondary use. J Biomed Inform. 2013;46:830–6.
3. Rubin DB. Matching to Remove Bias in Observational Studies. Biometrics [Internet]. 1973 Mar [cited 2018 Mar 5];29(1):159. Available from: http://www.jstor.org/stable/2529684?origin=crossref
4. Wilks S. On the distribution of statistics in samples from a normal population of two variables with matched sampling of one variable. Metron. 1932;9:87–126.
5. Cochran WG, Cox GM. Experimental Designs. New York: John Wiley & Sons, Ltd.; 1950.
6. Billewicz WZ. The efficiency of matched samples: an empirical investigation. Biometrics. 1965;21(3):623–44.
7. Czajka JL, Hirabayashi SM, Little RJA, Rubin DB. Projecting from advance data using propensity modeling: An application to income and tax statistics. J Bus Econ Stat [Internet]. 1992 Apr [cited 2018 Mar 7];10(2):117–31. Available from: http://www.jstor.org/stable/1391671?origin=crossref
8. Robins JM, Hernán MA, Brumback B. Marginal structural models and causal inference in epidemiology. Epidemiology [Internet]. 2000 Sep [cited 2018 Mar 7];11(5):550–60. Available from: http://www.ncbi.nlm.nih.gov/pubmed/10955408
9. Lunceford JKJ, Davidian M. Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study. Stat Med [Internet]. 2004 Oct 15 [cited 2018 Mar 7];23(19):2937–60. Available from: http://doi.wiley.com/10.1002/sim.1903
10. Rosenbaum PR, Rubin DB. The bias due to incomplete matching. Biometrics [Internet]. 1985 Mar [cited 2018 Mar 26];41(1):103–16. Available from: http://www.ncbi.nlm.nih.gov/pubmed/4005368
11. King G, Nielsen R, Coberley C, Pope J, Wells A, King Richard Nielsen Carter Coberley James Pope Aaron Wells GE. Comparative Effectiveness of Matching Methods for Causal Inference *. Dir Heal Res Outcomes [Internet]. 2011 [cited 2017 Dec 12]; Available from: http://people.fas.harvard.edu/