ESANN 2020 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 2-4 October 2020, i6doc.com publ., ISBN 978-2-87587-074-2.
Available from http://www.i6doc.com/en/.

Estimating Individual Treatment Effects through Causal Populations Identification

Céline Beji[1], Eric Benhamou[1,2], Michaël Bon[3], Florian Yger[1], Jamal Atif[1]

1- Paris-Dauphine University - PSL, LAMSADE, CNRS, MILES
Place du Maréchal de Lattre de Tassigny, 75016 Paris - FRANCE
2- Ai Square Connect 3- AdWay, Groupe Square

[1] **Abstract**.
Estimating the Individual Treatment Effect from observational data, defined as the difference between outcomes with and without treatment or intervention, while observing just one of both, is a challenging problems in causal learning. In this paper, we formulate this problem as an inference from hidden variables and enforce causal constraints based on a model of four exclusive causal populations. We propose a new version of the EM algorithm, coined as Expected-Causality-Maximization (ECM) algorithm and provide hints on its convergence under mild conditions. We compare our algorithm to baseline methods on synthetic and real-world data and discuss its performances.

1 Introduction

Estimating *Individual Treatment Effect* (ITE) from observational data is central in many application domains. For instance in healthcare, where the treatment is a proper medical treatment and the desired effect is the recovery of the patient. Being able to target accurately and demonstrably the population responding to a treatment has strong beneficial consequences in terms of public health by boosting personalized medicine. That would indeed allow a precise distribution of drugs to the profile of patients they can address, whereas at present, such drug may be prohibited because it does not demonstrate a positive effect at the whole population level through the lens of current standard testings. In the vein of the Rubin's causality framework [1, 2], we cast this counterfactual learning problem as an inference problem with incomplete data. We consider X the $\mathcal{X}$-valued random variable ($\mathcal{X} \subseteq \mathbb{R}^d$) representing the features of an individual and T the treatment assignment binary indicator stating whether the treatment was assigned ($T = 1$) or not ($T = 0$). We denote $Y_i(1)$ the binary outcome that would be observed if we assigned the treatment to individual i and $Y_i(0)$ the one that would be observed if we did not (e.g. $Y_i(1) = 1$ meaning that an effect was observed after treating individual i). ITE of individual $X_i = x$ is defined as the conditional mean difference in potential outcomes, $\tau(x) = \mathbb{E}[Y_i(1) - Y_i(0) \mid X_i = x]$. The fundamental problem is that for any individual i, we only observe the *factual outcome* $Y_i(t)$ corresponding to the outcome of the assignment, whereas

[1]ESANN 2020 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 22-24 April 2020.

ESANN 2020 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 2-4 October 2020, i6doc.com publ., ISBN 978-2-87587-074-2.
Available from http://www.i6doc.com/en/.

the *counterfactual* $Y_i(1-t)$ remains unknown [2]. As summarized in Table 1, from each couple $Y_i = \{Y_i(0), Y_i(1)\}$, called the *potential outcome*, we can define four mutually exclusive categories of response to the treatment [3]: responders who display a positive outcome only when treated, anti-responders who display a positive outcome only when they are *not* treated, doomed and survivors, who respectively never and always display a positive outcome.

Responder (R)	Doomed (D)	Survivor (S)	Anti-responder (A)
$\{Y(1)=1, Y(0)=0\}$	$\{Y(1)=0, Y(0)=0\}$	$\{Y(1)=1, Y(0)=1\}$	$\{Y(1)=0, Y(0)=1\}$

Table 1: Potential outcome of each causal population

Based on this typology of behaviors, we write the counterfactual learning problem as a parametric estimation of latent variables constrained by the causal groups properties. In this holistic approach, not only do we efficiently evaluate the ITE, but we are able to identify under mild assumptions the causality classes.

2 Related Work

The literature on causal inference is abundant, and it is beyond the scope of this paper to cover it exhaustively, although [4] is a good reference for a broad overview of this topic. Within Rubin's framework, baseline approaches consist in using treatment as a feature, or in learning two independent classifiers on the control and on the test datasets. This latter approach has the advantage of simplicity and versatility, but may lead to a selection bias. To overcome this problem, more sophisticated methods have been proposed. A first group on method consists in adaptations of classical machine learning methods. Examples are: (i) an SVM-like approach [5] where two hyperplanes are introduced and properly optimized to separate class behaviors. (ii) a parametric Bayesian method [6] for learning the treatment effects using a multi-task Gaussian process. (iii) several random forest-based approaches with split criteria specifically adapted to the problem [7]. Deep learning methods have also been put to good use with examples such as: (i) a deep neural network architecture [8] able to learn classifiers on the test and control populations while enforcing the minimization of an integral probability metric between the distributions of these classifiers. This work builds upon [9] where counterfactual inference has been tackled from the perspective of domain adaptation and representation learning. (ii) A number of methods using neural networks have also recently emerged [10, 11]. Interestingly, when the test and control populations have the same size, it is shown in [12] that a variable change could be used, leading to the estimation of a unique probability distribution (allowing the straightforward use of classical methods).

Our approach is different from the ones above. In our case, with binary treatment and outcome, the population has a clear causal structure. We use this fact and model the population as a mixture of four causal groups. Then, from the general knowledge of the causal structure obtained by our method, we can derive the ITE and thus compare our results with ITE-specific methods.

ESANN 2020 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 2-4 October 2020, i6doc.com publ., ISBN 978-2-87587-074-2. Available from http://www.i6doc.com/en/.

3 A Parametric Model for Causal Populations

We model the whole population as a mixture of mutually exclusive causal groups coined as responders (R), doomed (D), survivors (S) and anti-responders (A). We denote their respective distributions as $\{f_k(.|\theta_k)\}_{k\in\{R,D,S,A\}}$, and π_k their mixing probability.

For an individual, the specific group to which he belongs is determined by his outcome with and without treatment. Since we can never observe both simultaneously, we introduce $Z_i = \{z_{ik}\}_{k\in\{R,D,S,A\}}$ the discrete latent variable that represents the class probability of individual i.

Our model implies several constraints on the distribution of this latent variable, obviously excluding two causal populations according to the factual outcome and the assigned treatment. For example, it appears from Table 1 that an individual i with $Y_i(0) = 0$ cannot be a survivor or an anti-responder. The probability distribution of these two classes can then be set to zero ($z_{iS} = z_{iA} = 0$). Similar constraints can be applied for every value of the factual outcomes and are summarized in Table 2. Our goal is to estimate the latent distribution, from which we can in particular derive the ITE.

	$Y_i(0)=0$	$Y_i(0)=1$	$Y_i(1)=0$	$Y_i(1)=1$
Causality constraints	$\begin{cases} z_{iS}=z_{iA}=0 \\ z_{iR}+z_{iD}=1 \end{cases}$	$\begin{cases} z_{iR}=z_{iD}=0 \\ z_{iS}+z_{iA}=1 \end{cases}$	$\begin{cases} z_{iR}=z_{iS}=0 \\ z_{iD}+z_{iA}=1 \end{cases}$	$\begin{cases} z_{iD}=z_{iA}=0 \\ z_{iR}+z_{iS}=1 \end{cases}$

Table 2: The causality constraints (C^*)

Proposition 1. *Knowing the latent distribution , ITE writes as*

$$\tau(x) = (l_R(x) + l_S(x))\mathbb{E}[\mathbb{1}_{Y_i(1)=1}] - (l_S(x) + l_A(x))\mathbb{E}[\mathbb{1}_{Y_i(0)=1}] \tag{1}$$

where $l_C(x) = \frac{\pi_C f_C(X_i=x|\theta_C)}{\sum_{G\in\{R,D,S,A\}} \pi_G f_G(X_i=x|\theta_G)}$, $C \in \{R, D, S, A\}$.

Proof. $\mathbb{E}[Y_i(1) = 1|X_i = x] = \frac{\mathbb{E}[X_i(1)=x|Y_i(1)=1]P(Y_i(1)=1)}{P(X_i=x)}$
$= \frac{\mathbb{E}[X_i(1)=x|X_i\in\{R,S\}]P(Y_i(1)=1)}{P(X_i=x)} = \frac{\sum_{C\in\{R,S\}} \pi_C f_C(X_i=x|\theta_C)P(Y_i(1)=1)}{\sum_{C\in\{R,D,S,A\}} \pi_C f_C(X_i=x|\theta_C)}$ □

4 ECM Algorithm

Our learning problem amounts to estimate the mixing coefficients $\{\pi_k\}_{k\in\{R,D,S,A\}}$, the distributions parameters $\theta = \{\theta_k\}_{k\in\{R,D,S,A\}}$ and the latent distribution $q(z)$. For that matter, we consider the Expectation-Maximization (EM) algorithm, originally introduced in [13], which is known to be an appropriate optimization algorithm for estimating the data distribution of hidden variables. We provide the EM algorithm with extra information about the possible groups for every observation (in spirit similarly to [14] where a concept of authorized label set is used or to [15] which uses partial information). However, contrary

ESANN 2020 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 2-4 October 2020, i6doc.com publ., ISBN 978-2-87587-074-2. Available from http://www.i6doc.com/en/.

Algorithm 1 Expectation-Causality-Maximisation

Initialisation: initialise q_0 and compute π_0 and θ_0 (M-step).
While(Not Converged) do
  Expected step: $q_{t+1} = \arg\max_q(\mathcal{L}(q, \theta_t, \pi_t))$
  Causality step: Constraints on q_{t+1} with C^* (Table 2)
  Maximization step: $(\theta_{t+1}, \pi_{t+1}) = \arg\max_{\theta,\pi}(\mathcal{L}(q_{t+1}, \theta, \pi))$
End While

to [14, 15], we enrich this extra information with causal constraints derived from the structure of the problem.

In Algorithm 1, the Expectation step estimates the latent variables, the Causality step projects the solution on the causality constraints[2] displayed in Table 2, while the Maximization step maximises the likelihood $\mathcal{L}(q, \theta, \pi)$ as if the latent variables were not hidden. For a faster convergence, we initialize q_0 with a probability of half on each of the two remaining causal populations. Our algorithm converges as by construction it necessarily increases the log-likelihood at each iterations and remains bounded by an evidence lower bound similar to EM given by $\sum_{z|q(z|x,\theta,\pi)\neq 0} q(z|x,\theta,\pi) \log \frac{p(x,z|\theta,\pi)}{q(z|x,\theta,\pi)}$. Note that without information on the input features X, the distribution $q(z)$ is uniformly distributed between the two authorized groups. In addition, the causality constraints enforce that two population labels are ruled out as they are not admissible. Under some specific assumptions, we can do even better and recover the true label (cf Proposition 2). Thanks to this true label, the maximum likelihood problem is cast into four decoupled single-density maximum likelihood problems. Under concavity of the likelihood for every distribution in the mixture, the log-likelihood converge not only locally but to a unique global maximum (corollary 1). We can summarize these findings by saying that the unsupervised learning problem, implied by our model of mixture, is transformed into a semi-supervised problem.

Proposition 2. *Under causality constraints which excludes two populations (depending on treatment and observed outcome), and assuming the feature distribution conditionally to the group is the same and independent of the treatment (i.e. $p(x|t, y) = p(x)$), each group will be identified to a unique causal population.*

Proof. (sketch) Under only the information of the causal constraints, $q(z|y, t)$ is distributed with the probability $\frac{1}{2}$ between the two possible populations. If we note $p(.) = p(.|\theta, \pi)$, then $q(z|x, y, t) = \frac{p(x|z,y,t)q(z|y,t)}{p(x|y,t)} \propto \frac{p(x|z)}{p(x)}$ under causal constraints and assumption of uniform features. □

Corollary 1. *Under causality constraints, if the log-likelihood of the distribution for a single mixture is concave, the ECM algorithm reaches the global optimum.*

Proof. (sketch) Preserving the properties of the EM algorithm, ECM converges to a local maximum. As a result of the identifiability (Prop. 2) and the concavity for a single mixture, the local optimum is also global. □

[2]Forcing to zero the probabilities z_{ik} of the two impossible groups and normalizing the sum.

ESANN 2020 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 2-4 October 2020, i6doc.com publ., ISBN 978-2-87587-074-2. Available from http://www.i6doc.com/en/.

5 Experiments

Gaussian distributions (for which parameters are the mean and variance denoted by $\theta = (\mu, \Sigma)$) are a natural choice for a mixture model. Once the model is learned, interpreting the μ_k as average elements of each causal population could be of great interest and could help to answer questions like "what does an average anti-responder looks like?" to improve treatment policies. Because of the fundamental impossibility to access the true counterfactual label of any given observation, the true ITE cannot be known and thus it is unclear how to best assess the relevance of any model in real-world conditions. Hence, we need to test our model on synthetic and semi-synthetic datasets.

We first designed a synthetic dataset with two covariates distributed as a mixture of four overlapping Gaussian distributions. We use two metrics standard for causal problems: the ϵ_{PEHE} [16] the AUUC (Area Under the Uplift Curve) [17]. We compute these metrics for the optimal model (since it does not necessarily score maximally according to these metrics) and compare its values to the ones of the models we test. We also use the IHDP semi-synthetic dataset, compiled for causal effect estimation in [16]. Here the underlying distribution of each causal population remains unknown, but the outcomes with and without treatment are both available. In that case, we use our model to predict the most likely counterfactual of any individual.

The results are reported out of a sample over 20 trials and a Wilcoxon signed-rank test (with a confidence level of 5%) is used to confirm the significance of the results. We compare our method to standard baselines that provide competitive results with respect to the state of the art [18]: the approach using two separate classification models (LR2), the approach using the treatment variable as feature (LR1) and the model based on the class variable transformation [12] (LRZ), each using logistic regressions as classifiers.

	Synthetic dataset		IHDP	
	ϵ_{PEHE}	AUUC	ϵ_{PEHE}	AUUC
Ref.	0.24	1488	.	3149
LR1	0.57 +/- 0.08	742 +/- 175	0.66 +/- 0.08	2202 +/- 625
LR2	0.79 +/- 0.08	943 +/- 206	0.67 +/- 0.07	2168 +/- 618
LRZ	.	939 +/- 208	.	2191 +/- 558
ECM	**0.27 +/- 0.04**	**1512 +/- 203**	**0.59 +/- 0.09**	**2226 +/- 580**

Table 3: Experimental results on synthetic and real datasets.

6 Results and Conclusion

Compared to the baselines, our results are clearly the ones closest to optimality on both synthetic and real datasets. Moreover, our model is intrinsically more interpretable than the compared baselines as the parameters of the distributions

ESANN 2020 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 2-4 October 2020, i6doc.com publ., ISBN 978-2-87587-074-2. Available from http://www.i6doc.com/en/.

of the causal groups provide information about the causal mechanism at play. Finally, our model is versatile and can be adapted to multiple treatments [19], non-compliance to treatment cases [2] or separate labels [20].

References

[1] D.B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701, 1974.

[2] Guido W Imbens and Donald B Rubin. Bayesian inference for causal effects in randomized experiments with noncompliance. *The Annals of Statistics*, pages 305–327, 1997.

[3] Larry Wasserman. Causal Inference. In *All of statistics : a concise course in statistical inference*, chapter 16. Springer, 2004.

[4] Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2009.

[5] Łukasz Zaniewicz and Szymon Jaroszewicz. Support vector machines for uplift modeling. In *ICDMW*, pages 131–138, 2013.

[6] Ahmed M Alaa and Mihaela van der Schaar. Bayesian inference of individualized treatment effects using multi-task gaussian processes. In *Advances in Neural Information Processing Systems*, pages 3424–3432, 2017.

[7] Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.

[8] Uri Shalit, Fredrik D Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *ICML*, 2017.

[9] Fredrik Johansson, Uri Shalit, and David Sontag. Learning representations for counterfactual inference. In *ICML*, 2016.

[10] Jinsung Yoon, James Jordon, and Mihaela van der Schaar. Ganite: Estimation of individualized treatment effects using generative adversarial nets. In *ICLR*, 2018.

[11] Christos Louizos, Uri Shalit, Joris M Mooij, David Sontag, Richard Zemel, and Max Welling. Causal effect inference with deep latent-variable models. In *NIPS*, 2017.

[12] Maciej Jaskowski and Szymon Jaroszewicz. Uplift modeling for clinical trial data. In *ICML Workshop on Clinical Data Analysis*, 2012.

[13] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society*, 39(1):1–38, 1977.

[14] C Ambroise and G Govaert. Em algorithm for partially known labels. In *Data analysis, classification, and related methods*, pages 161–166. Springer, 2000.

[15] Etienne Côme, Latifa Oukhellou, Thierry Denoeux, and Patrice Aknin. Learning from partially supervised data using mixture models and belief functions. *Pattern recognition*, 42(3):334–348, 2009.

[16] Jennifer L Hill. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011.

[17] Eustache Diemert, Artem Betlei, Christophe Renaudin, and Massih-Reza Amini. A large scale benchmark for uplift modeling. In *Proceedings of the AdKDD and TargetAd Workshop, KDD*. ACM, 2018.

[18] Ahmed Alaa and Mihaela Schaar. Limits of estimating heterogeneous treatment effects: Guidelines for practical algorithm design. In *ICML*, 2018.

[19] Markus Frölich. Programme evaluation with multiple treatments. *Journal of Economic Surveys*, 18(2):181–224, 2004.

[20] Ikko Yamane, Florian Yger, Jamal Atif, and Masashi Sugiyama. Uplift modeling from separate labels. In *NIPS*, 2018.