CDVAE: Estimating causal effects over time under unobserved adjustment variables

Mouad El Bouchattaoui [1,2] Myriam Tami [1] Benoit Lepetit [2] Paul-Henry Cournède [1]

[1] Paris-Saclay University, CentraleSupélec, MICS Lab, France

[2] Saint-Gobain, France

Colloquium: When Causal Inference meets Statistical Analysis, April 17th 2023

Motivation

Motivation

Why can providing "precise" estimates of individual treatment effects (ITE) be challenging even in RCTs?

- Treatment effect may vary conditioned on a variable not affecting the treatment!
- Adjustment variables: Variables affecting response and not treatment.
- Effect modifiers: Adjustment variables that change the causal effect.

Motivation

Do we really need adjustment variables?

- For individual treatment effects, yes!
- Needed as much as confounders (we assume sequential ignorability!).
- Know what affects the response $\rightarrow$ Estimate precise response trajectories.

Motivation

What can we do when we have?:

- Longitudinal data.
- Individual treatment effects are the target.
- Adjustment variables are not observed.

Motivation

When does causal inference meet statistics in our problem?

Motivation

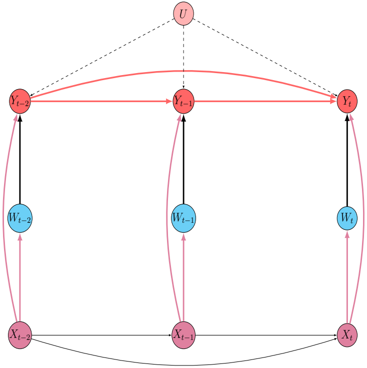
Classically, when we have:

- Unobserved sources of heterogeneity.
- Individual response trajectories are of interest.

We can perform a **mixed effect modeling**:

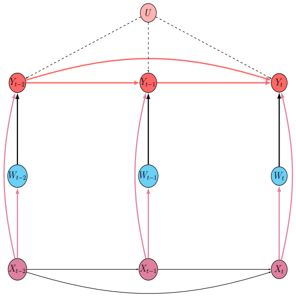
- A Parametric model over the response.
- Some parameters vary randomly among individuals.
- **Intuition**: random parameters capture heterogeneity.

Example



- Y: Vital indicator.
- W: Taking some vitamin (0 or 1).
- X: Confounder, say financial resources.
- U: unobserved effect modifier (say age).
- **Goal: Individual effect of $W_t \to Y_t$.**

Example

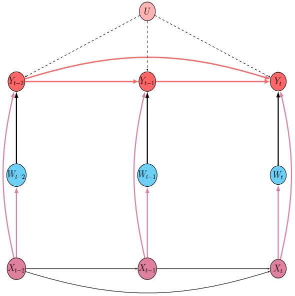


- A mixed effect modeling:

$$\mathbf{E}(Y_{i,t} \mid W_{i,t}, X_{i,\leq t}, Y_{i,<t}, \alpha_i^{(1)}, \alpha_i^{(2)}) = \gamma_1 Y_{i,t-1} + \gamma_2 Y_{i,t-2} + \underbrace{(\beta_1 Y_{i,t-1} + (\beta_2 + \alpha_i^{(1)}) X_{i,t} + \alpha_i^{(2)})}_{\text{ITE}} W_i + \beta_3 X_{i,t}.$$

- Y_{t-1} is an effect modifier!
- Y_{t-2} is not effect modifier! (still an adjustment variable).

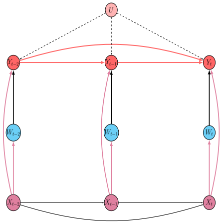
Example



$$\mathbf{E}(Y_{i,t} \mid W_{i,t}, X_{i,\leq t}, Y_{i,<t}, \alpha_i^{(1)}, \alpha_i^{(2)}) = \gamma_1 Y_{i,t-1} + \gamma_2 Y_{i,t-2} + \underbrace{(\beta_1 Y_{i,t-1} + (\beta_2 + \alpha_i^{(1)})X_{i,t} + \alpha_i^{(2)})}_{\text{ITE}} W_i + \beta_3 X_{i,t}.$$

- $\gamma = (\gamma_1, \gamma_2)$, $\beta = (\beta_1, \beta_2, \beta_3)$ Non-random parameters (fixed effects).
- $\alpha_i = (\alpha_i^{(1)}, \alpha_i^{(2)})$ random parameters (random effects).
- α_i: Accounts for unobserved factors of variation (Our U).

Example



- A mixed effect modeling:

$$\mathbf{E}(Y_{i,t} \mid W_{i,t}, X_{i,\leq t}, Y_{i,<t}, \alpha_i^{(1)}, \alpha_i^{(2)}) = \gamma_1 Y_{i,t-1} + \gamma_2 Y_{i,t-2} + \underbrace{(\beta_1 Y_{i,t-1} + (\beta_2 + \alpha_i^{(1)}) X_{i,t} + \alpha_i^{(2)})}_{\text{ITE}} W_i + \beta_3 X_{i,t}.$$

- Let's see α_i as a random variable.
- Let's connect α_i to unobserved U_i:

$$\alpha_i = \alpha(U_i) := (\alpha^{(1)}(U_i), \alpha^{(2)}(U_i)).$$

- $\alpha^{(1)}, \alpha^{(2)} : U_i \to \mathbb{R}$: arbitrary unknown mappings.

$$\underbrace{\mathbf{E}(Y_{i,t} \mid W_{i,t}, X_{i,\leq t}, Y_{i,<t}, U_i)}_{\mathbf{E}(Y_{i,t} | pa(Y_{i,t}))} = \gamma_1 Y_{i,t-1} + \gamma_2 Y_{i,t-2} + \underbrace{(\beta_1 Y_{i,t-1} + (\beta_2 + \alpha_i^{(1)}(U_i)) X_{i,t} + \alpha_i^{(2)}(U_i))}_{\text{ITE}} W_i + \beta_3 X_{i,t}.$$

Modeling

Assumptions

Suppose

- The causal graph is known.
- Sequential ignorability.
- Some **static effect modifiers** are unobserved.

We suggest:

- Instead of learning random parameters α_i:
  - See unobserved adjustment variables as latents.
  - Learn a representation of U_i.
  - Model the mapping by highly flexible neural networks.
  - Condition the treatment effect on the representation.

Causal framework

- Binary treatment W_{it} ➞ Two potential outcomes $Y_{it}(1), Y_{it}(0)$.
- Sequential ignorability: $Y_{it}(\omega_{it}) \perp\!\!\!\perp W_{it} \mid \mathbf{X}_{i,\leq t} = \mathbf{x}_{i,\leq t}, Y_{i,<t} = y_{i,<t}$
- Causal quantity of interest:

$$\tau_{it} := \mathbb{E}(Y_{it}(1) - Y_{it}(0) \mid \mathbf{X}_{i,\leq t} = \mathbf{x}_{i,\leq t}, Y_{i,<t} = y_{i,<t}, \underbrace{U_i = u_i}_{\text{To be constructed}})$$

- Identify:

$$\tau_{it} = \mathbb{E}(Y_{it} \mid \mathbf{X}_{i,\leq t} = \mathbf{x}_{i,\leq t}, Y_{i,<t} = y_{i,<t}, U_i = u_i, W_{it} = 1) - \mathbb{E}(Y_{it} \mid \mathbf{X}_{i,\leq t} = \mathbf{x}_{i,\leq t}, Y_{i,<t} = y_{i,<t}, U_i = u_i, W_{it} = 0)$$

Modeling

- Conditional probabilistic model:

$$p_\theta(y_{\leq T}, u \mid \mathbf{x}_{\leq T}, \omega_{\leq T}) = \prod_{t=1}^{T} \left[p_\theta(y_t \mid y_{<t}, \mathbf{x}_{\leq t}, \omega_{\leq t}, u)\right] p(u)$$

- Use d-separation to simplify $p_\theta(y_t \mid y_{<t}, \mathbf{x}_{\leq t}, \omega_{\leq t}, u)$.
- Define an inference model $q_\phi\left(u \mid y_{\leq T}, \mathbf{x}_{\leq T}, \omega_{\leq T}\right)$ that approximates $p_\theta\left(u \mid y_{\leq T}, \mathbf{x}_{\leq T}, \omega_{\leq T}\right)$.
- Consistency: Simplify $q_\phi\left(u \mid y_{\leq T}, \mathbf{x}_{\leq T}, \omega_{\leq T}\right)$ using d-separation.

Modeling

- The Evidence Lower Bound (ELBO):

$$\begin{aligned}\mathrm{ELBO}(\theta,\phi) = \sum_{t=1}^{T} \mathbb{E}_{q_\phi(u|y_{\leq T},\mathbf{x}_{\leq T},\omega_{\leq T}))} \left[\log p_\theta(y_t \mid y_{<t},\mathbf{x}_{\leq t},\omega_t,u)\right] \\ - D_{KL}(q_\phi(u \mid y_{\leq T},\mathbf{x}_{\leq T},\omega_{\leq T}) \mid\mid p(u))\end{aligned}$$

Counterfactual regression

- How to estimate individual treatment effects?:
  - Make $p_\theta(y_t \mid y_{<t}, \mathbf{x}_{\leq t}, \omega_{\leq t}, u)$ a TARNet style [1].
  - Weighting with a function of the propensity scores $\alpha_\eta(\mathbf{x}_{\leq t}) = f(p_\eta(W_t = 1 \mid \mathbf{x}_{\leq t}))$ [1], [2].
  - Write a loss, weighting ELBO:

$$\mathcal{L}_{total}(\theta, \phi, \eta) = -\sum_{t=1}^{T} \underbrace{\alpha_\eta(\mathbf{x}_{\leq t})}_{\text{Weighting}} \mathbb{E}_{q_\phi(u|y_{\leq T}, \mathbf{x}_{\leq T}, \omega_{\leq T}))} \left[\underbrace{\log p_\theta(y_t \mid y_{<t}, \mathbf{x}_{\leq t}, \omega_t, u)}_{\text{Reconstruction}} \right]$$
$$+ \underbrace{\beta D_{KL}(q_\phi(u \mid y_{\leq T}, \mathbf{x}_{\leq T}, \omega_{\leq T}) \mid\mid p(u))}_{\text{Regularization}} + \underbrace{\mathcal{L}_W(\eta)}_{\text{loss for propensity}}$$

Posterior collapse

- Avoid $D_{Kl} \approx 0 \longrightarrow$ Cyclical scheduling of β [3].

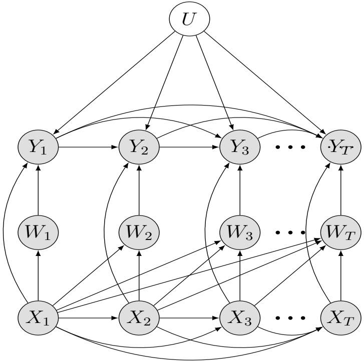
we call the model:

- CDVAE: Causal Dynamic Variational Auto-encoder, $\beta = 1$
- β_{cyc}-CDVAE: CDVAE with β updated cyclically.

EMPIRICAL STUDIES

Simulation setup

Data simulated according to:



- Treatment effect to estimate:

$$\tau(\mathbf{X}_t, U) := exp(\frac{1}{d_x}\sum_{j=1}^{d_x}\mathbf{X}_{t,j} + \frac{1}{d_u}\sum_{j=1}^{d_u}U_j)$$

Results

Benchmark

- RMSMs: Recurrent Marginal Structural Models [4].
- CRN: Counterfactual Recurrent Network [5].
- CausalForestDML:Forest Double Machine Learning model [6], [7].

Model	ϵ_{ATE}	$\mathrm{MAE}(\tau)$	$\mathrm{MAE}(y)$	$\mathrm{RMSE}(\tau)$	$\mathrm{RMSE}(y)$
β_{cyc}-**CDVAE**(ours)	0.07 $\pm$ 0.01	**0.17 $\pm$ 0.02**	**0.17 $\pm$ 0.02**	**0.25 $\pm$ 0.02**	**0.22 $\pm$ 0.02**
CDVAE(ours)	0.18 $\pm$ 0.03	0.23 $\pm$ 0.01	0.21 $\pm$ 0.01	0.29 $\pm$ 0.01	0.31 $\pm$ 0.02
CausalForestDML	**0.002 $\pm$ 0.001**	0.24 $\pm$ 0.01	0.78 $\pm$ 0.03	0.32 $\pm$ 0.02	0.95 $\pm$ 0.02
RMSM	1.18 $\pm$ 0.02	1.18 $\pm$ 0.02	0.44 $\pm$ 0.03	1.26 $\pm$ 0.03	0.64 $\pm$ 0.02
CRN	0.12 $\pm$ 0.01	0.34 $\pm$ 0.02	0.38 $\pm$ 0.01	0.46 $\pm$ 0.02	0.49 $\pm$ 0.02

Whats happens during training?

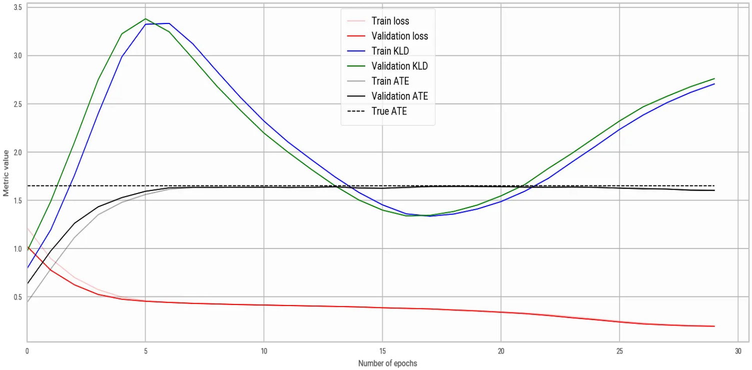


Figure: CDVAE With cycling: Good balancing.

Whats happens during training?

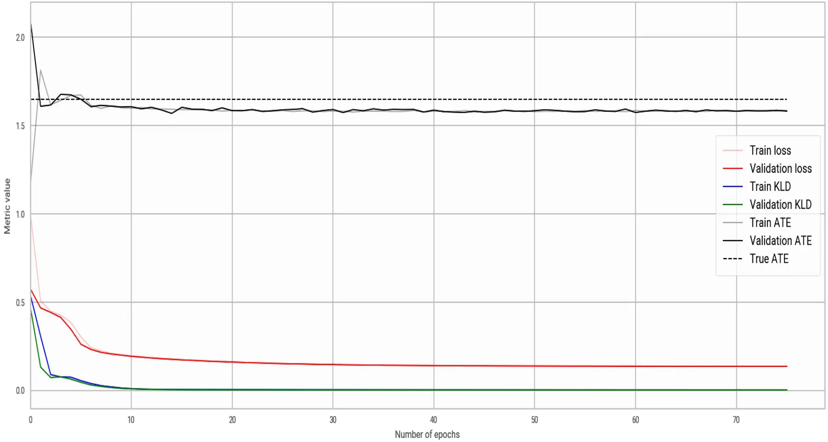


Figure: CDVAE Without cycling: Bad balancing.

Conclusion

Conclusion

Pros:

- ITEs are better estimated.
- A good trade-off: being predictive of both responses and causal effects.
- Handling responses of different nature: continuous, discrete, ...

Cons

- Difficulty in calibration: cycling strategy.

Prospects:

- How about unobserved time-varying adjustment variables?
- How about the individual effect of a sequence of interventions?

References I

[1] U. Shalit, F. D. Johansson, and D. Sontag, "Estimating individual treatment effect: Generalization bounds and algorithms," in *International Conference on Machine Learning*, PMLR, 2017, pp. 3076–3085.

[2] N. Hassanpour and R. Greiner, "Learning disentangled representations for counterfactual regression," in *International Conference on Learning Representations*, 2019.

[3] H. Fu, C. Li, X. Liu, J. Gao, A. Celikyilmaz, and L. Carin, "Cyclical annealing schedule: A simple approach to mitigating kl vanishing," *arXiv preprint arXiv:1903.10145*, 2019.

[4] B. Lim, "Forecasting treatment responses over time using recurrent marginal structural networks," *advances in neural information processing systems*, vol. 31, 2018.

References II

[5] I. Bica, A. M. Alaa, J. Jordon, and M. van der Schaar, "Estimating counterfactual treatment outcomes over time through adversarially balanced representations," *arXiv preprint arXiv:2002.04083*, 2020.

[6] S. Wager and S. Athey, "Estimation and inference of heterogeneous treatment effects using random forests," *Journal of the American Statistical Association*, vol. 113, pp. 1228–1242, 2018.

[7] S. Athey, J. Tibshirani, and S. Wager, "Generalized random forests," *The Annals of Statistics*, 2019.

Thanks for your attention!

Are there questions?