Causal Contrastive Learning for Counterfactual Regression Over Time

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

NeurIPS 2024 Pre-Recorded Talk

Introduction

The scope of our work

- *Assuming all confounders are observed, how can we efficiently estimate the effect of any potential sequence of future treatments on subsequent responses over extended forecasting horizons?*

- Literature Overview

Model	Backbone	Long-Term Forecast?	Learning Dependencies	Contrastive Learning	Inference Efficiency	Selection Bias Handling	Representation Invertibility
Our Model	GRU	**Yes**	Contrastive Predictive Coding	**Yes**	**High**	Balanced Rep.	**InfoMax**
Causal Transformer [1]	3 Transformers	**Yes**	Transformer	N/A	Low	Balanced Rep.	N/A
G-Net [2]	LSTM	No	N/A	N/A	Very Low	G-Computation	Covariates $\mathbf{X}_t$
CRN [3]	LSTM	No	N/A	N/A	**High**	Balanced Rep.	N/A
RMSN [4]	LSTM	No	N/A	N/A	**High**	Weighting	N/A
MSM [5]	Logistic + Linear	No	N/A	N/A	**High**	Weighting	N/A

- Research Gap
  - **Handling Long-Term Dependencies** Most models, except the Causal Transformer, struggle with capturing long-term dependencies in time-varying settings.
  - **Computational Challenge** Inference requires evaluating multiple counterfactual trajectories per individual and time step, significantly increasing test units. Efficiency is essential.
  - **Lack of Representation Invertibility** Most baseline models learn a representation of the confounding history, but none enforce its invertibility to ensure that confounding information is retained.

Contributions

- ☞ We leverage Contrastive Predictive Coding (CPC) [6], [7] with InfoNCE to capture long-term dependencies in process history, a novel approach in counterfactual regression over time.

- ☞ We adopt Information Maximization (InfoMax) [8], [9] to retain confounding information by prioritizing input reconstruction from the representation, reducing bias in counterfactual estimates.

- ☞ Using a simple GRU layer as the backbone, we show that well-designed regularization can outperform more complex transformer models.

Main setting: Potential Outcome Framework

- For each individual i and time t, we have:
- **Discrete treatment** $W_{it} \in \mathcal{W} = \{0, 1, \ldots, K-1\}$.
- **Continuous Outcome** $Y_{it} \in \mathcal{Y} \subset \mathbb{R}$.
- **Time-varying confounders** $\mathbf{X}_{it} \in \mathcal{X} \subset \mathbb{R}^{d_x}$.
- **Static confounders** $\mathbf{V} \in \mathcal{V} \subset \mathbb{R}^{d_v}$.
- **The history process** $\mathbf{H}_{t+1} = [\mathbf{V}, \mathbf{X}_{\leq t+1}, W_{\leq t}, Y_{\leq t}]$.

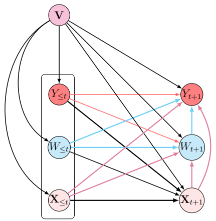


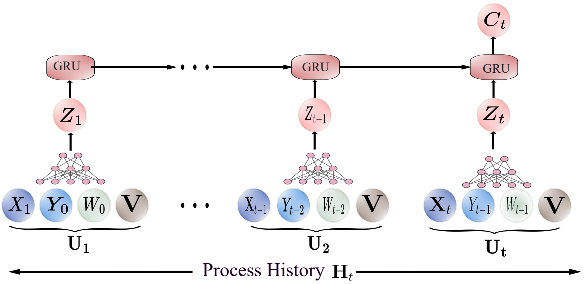
Figure: Causal graph over $\mathbf{H}_{t+1}$

★ **Goal** Assuming Sequential ignorability [10], estimate the expected counterfactual outcome for any $\omega_{t+1:t+\tau}$:

$$\mathbb{E}(Y_{t+\tau}(\omega_{t+1:t+\tau}) \mid \mathbf{H}_{t+1}) = \mathbb{E}\left(Y_{t+\tau} \mid \mathbf{H}_{t+1}, W_{t+1:t+\tau} = \omega_{t+1:t+\tau}\right).$$

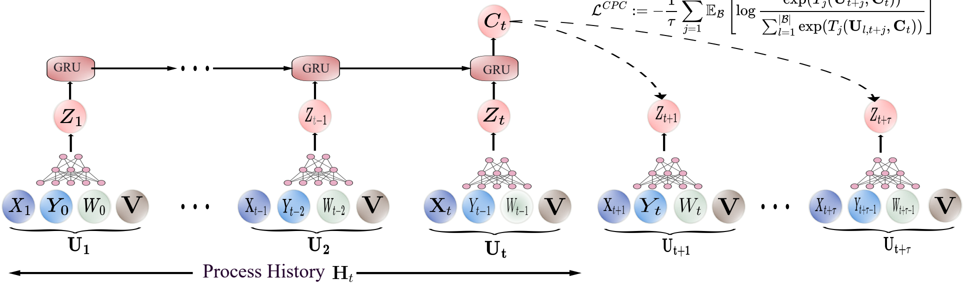
Modeling

Encoding step 1: Learn a context of the process $\mathbf{H}_t$

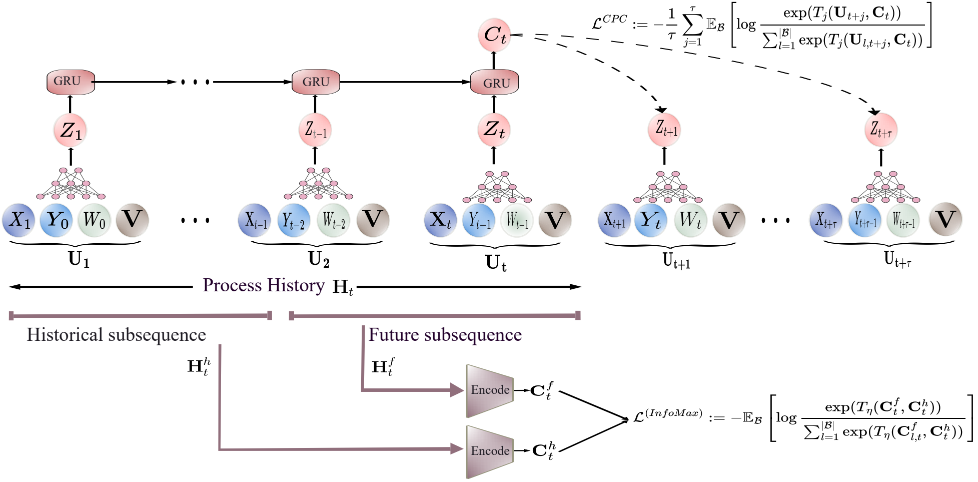


Encoding step 2: Contrastive Predictive Coding

$$\mathcal{L}^{CPC} := -\frac{1}{\tau}\sum_{j=1}^{\tau}\mathbb{E}_{\mathcal{B}}\left[\log\frac{\exp(T_j(\mathbf{U}_{t+j},\mathbf{C}_t))}{\sum_{l=1}^{|\mathcal{B}|}\exp(T_j(\mathbf{U}_{l,t+j},\mathbf{C}_t))}\right]$$



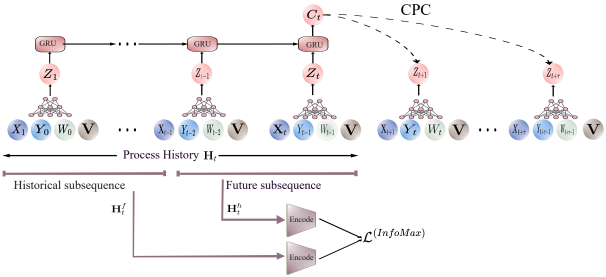
Encoding step 3: InfoMax



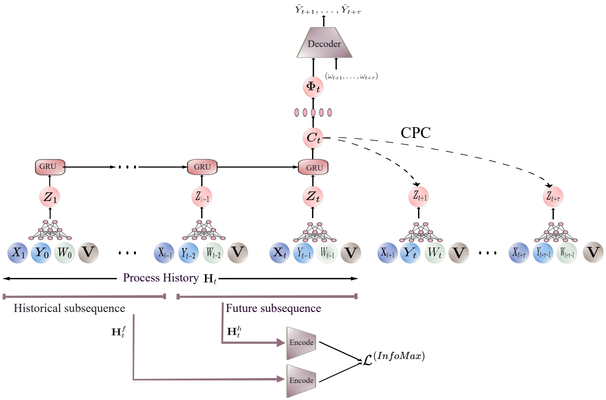
Encoder pertaining: Loss

★ **Encoder loss** $\mathcal{L}_{enc} = \mathcal{L}^{CPC} + \mathcal{L}^{(InfoMax)}$.

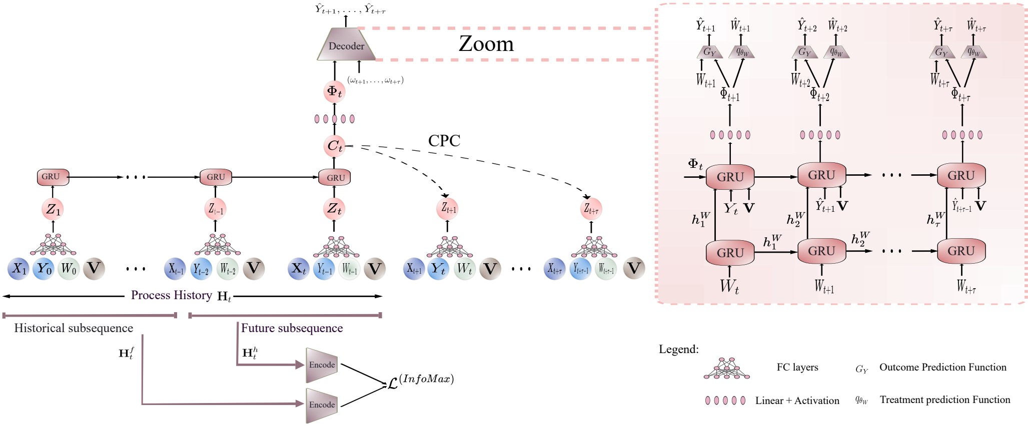
Encoder fine-tuning and Decoder training



Encoder fine-tuning and Decoder training



Encoder fine-tuning and Decoder training



Encoder Fine-Tuning and Decoder Training: Adversarial loss

- To address selection bias, we aim for $\Phi(\mathbf{H}_t) \perp\!\!\!\perp W_t$, or equivalently, $I(\Phi(\mathbf{H}_t), W_t) = 0$.
- Let I_{CLUB} represent the CLUB upper bound of mutual information [11], with $q(.)$ as a treatment classifier network.
- Let $\mathcal{L}_Y(\theta_R, \theta_Y)$ be the loss to predicting the factual responses $Y_{t+1}, \ldots, Y_{t+\tau}$ given the sequence of treatments $(W_{t+1}, \ldots, W_{t+\tau})$.

★ **Decoder Adversarial Training** We fine-tune the encoder by optimizing the factual outcome and treatment networks in the adversarial game:

$$\min_{\theta_R, \theta_Y} \mathcal{L}_{dec}(\theta_R, \theta_Y, \theta_W) = \mathcal{L}_Y(\theta_R, \theta_Y) + I_{\text{CLUB}}(\Phi_{\theta_R}(\mathbf{H}_t), W_{t+1}; q_{\theta_W}),$$

$$\min_{\theta_W} \mathcal{L}_W(\theta_W, \theta_R) = -\mathbb{E}_{\Phi_{\theta_R}(\mathbf{H}_t)}\left[\log q_{\theta_W}(W_{t+1} \mid \Phi_{\theta_R}(\mathbf{H}_t))\right].$$

Experiments

Experiments with semi-synthetic MIMIC III data

- Estimate the counterfactual blood pressure following a sequence of treatments made of vasopressors and mechanical ventilation.

Table: Results on the MIMIC III semi-synthetic reported by RMSEs. Smaller is better.

Model	$\tau = 1$	$\tau = 2$	$\tau = 3$	$\tau = 4$	$\tau = 5$	$\tau = 6$	$\tau = 7$	$\tau = 8$	$\tau = 9$	$\tau = 10$
Causal CPC (ours)	0.32±0.04	0.45±0.08	0.54±0.06	0.61 ±0.10	**0.66± 0.10**	**0.69±0.11**	**0.71± 0.11**	**0.73± 0.06**	**0.75 ± 0.05**	**0.77± 0.10**
CT	0.42 ± 0.38	**0.40± 0.06**	**0.52± 0.08**	**0.60± 0.005**	0.67±0.10	0.72 ±0.12	0.77±0.13	0.81±0.14	0.85 ±0.16	0.88 ±0.17
G-Net	0.54 ± 0.13	0.72±0.14	0.85 ±0.16	0.96 ± 0.17	1.05 ± 0.18	1.14 ±0.18	1.24± 0.17	1.33±0.16	1.41 ± 0.16	1.49±0.16
CRN	**0.27 ±0.03**	0.45±0.08	0.58 ± 0.09	0.72± 0.11	0.82± 0.15	0.92 ± 0.20	1.00 ± 0.25	1.06 ± 0.28	1.12 ± 0.32	1.17 ± 0.35
RMSN	0.40 ± 0.16	0.70 ± 0.21	0.80± 0.19	0.88 ± 0.17	0.94 ± 0.16	1.00 ± 0.15	1.05 ± 0.14	1.10 ± 0.14	1.14 ± 0.13	1.18 ± 0.13

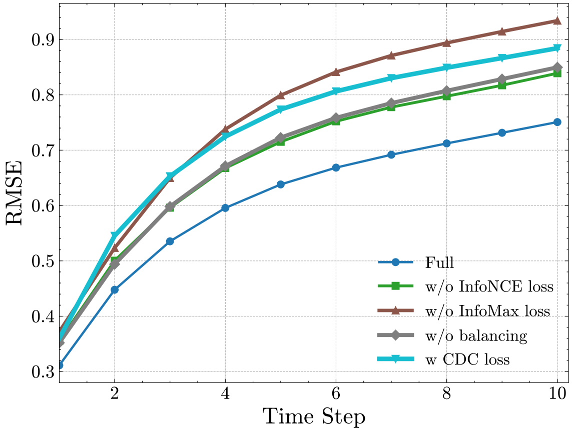
Computational Efficiency and Model Complexity

Computational Efficiency and Model Complexity

Table: Models complexity and the running time averaged over five seeds. Results are reported for tumor growth simulation ($\gamma = 1$). Hardware: GPU-1xNVIDIA Tesla M60.

Model	Trainable parameters (k)	Training time (min)	Prediction time (min)
Causal CPC (encoder + decoder)	8.2	16$\pm$ 3	4 $\pm$ 1
CT	11	12$\pm$ 2	30$\pm$ 3
G-Net	1.2	2 $\pm$ 0.5	35 $\pm$ 3
CRN	5.2	13$\pm$ 2	4$\pm$ 1
RMSN	1.6	22$\pm$ 2	4$\pm$ 1
MSM	$<$**0.1**	**1$\pm$0.5**	**1$\pm$0.5**

Ablation Study Results on MIMIC III



Conclusion

Discussion

- **Conclusion**
  - Proposed a novel, computationally efficient approach to long-term counterfactual regression by combining RNNs with contrastive learning, achieving SOTA performance without complex transformer models.
- **Future work**
  - While our model is designed for long-term predictions, it may not consistently outperform SOTA for short-horizon tasks. A trade-off could be achieved by adjusting the contrastive term weights across time steps, which we leave for future work.

Scan for paper link

References I

[1] V. Melnychuk, D. Frauen, and S. Feuerriegel, "Causal transformer for estimating counterfactual outcomes," *ArXiv*, vol. abs/2204.07258, 2022.

[2] R. Li, S. Hu, M. Lu, *et al.*, "G-net: A recurrent network approach to g-computation for counterfactual prediction under a dynamic treatment regime," in *ML4H@NeurIPS*, 2021.

[3] 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.

[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] J. M. Robins, M. A. Hernan, and B. Brumback, *Marginal structural models and causal inference in epidemiology*, 2000.

[6] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[7] O. Henaff, "Data-efficient image recognition with contrastive predictive coding," in *International conference on machine learning*, PMLR, 2020, pp. 4182–4192.

[8] R. Linsker, "Self-organization in a perceptual network," *Computer*, vol. 21, no. 3, pp. 105–117, 1988.

[9] R. D. Hjelm, A. Fedorov, S. Lavoie-Marchildon, *et al.*, "Learning deep representations by mutual information estimation and maximization," in *International Conference on Learning Representations*, 2019. [Online]. Available: https://openreview.net/forum?id=Bklr3j0cKX.

References III

[10] J. M. Robins and M. A. Hernán, "Estimation of the causal effects of time-varying exposures," *Longitudinal data analysis*, vol. 553, p. 599, 2009.

[11] P. Cheng, W. Hao, S. Dai, J. Liu, Z. Gan, and L. Carin, "Club: A contrastive log-ratio upper bound of mutual information," in *International conference on machine learning*, PMLR, 2020, pp. 1779–1788.