Structured Learning of Compositional Sequential Interventions


Jialin Yu[1]


Andreas Koukorinis[1]


Nicolò Colombo[2]


Yuchen Zhu[1]


Ricardo Silva[1]

[1] University College London, UK [2] Royal Holloway, University of London, UK

NeurIPS 2024 (Vancouver, Canada)

Outline

Motivation

Real-world Scenarios

Daily music recommendation influence users' future listening behaviour (such as: no action, promote songs from musician "A" by 10% or demote songs from musician "B" by 20%).

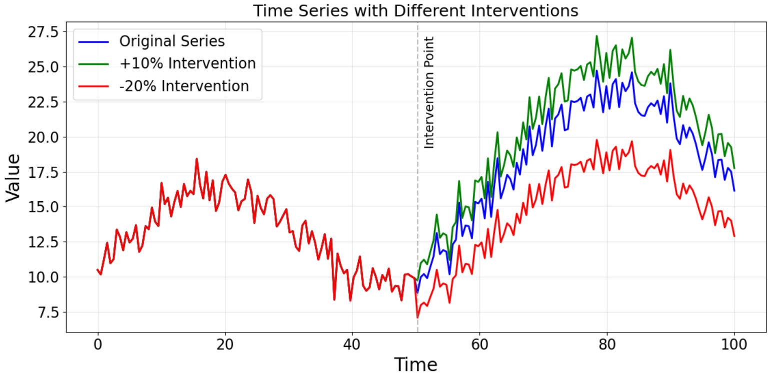


Figure 1: Toy example of time series with different interventions.

Problem

- Consider a more complex setting where the treatments may take different values for a single individual over time, often refereed to as time-varying treatments [10].
- How do we predict the effect of **combinations of categorical action/intervention sequence in the future**?

Challenges

- Can be treated as a standard prediction problem (i.e. supervised sequential modelling with LSTM [11], GRU [9], Transformers [21] etc.)
- Large categorical space, sparse interactions (mostly "default" action) and no-obvious structural assumptions between actions (in contrastive to distributed representations in natural language).

Literature

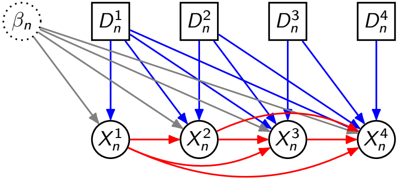
- Mostly used black-box models (LSTM, GRU and Transformers) or parametric models (often with strong Markovian assumptions [8, 20]) on short sequence and small action space.
- Some research focused on large action space, but lack of compositional identification. [12, 15, 18, 19]
- Other recent research deal with combinatorial categorical spaces [1, 5], but lack of longitudinal component.
- This work focus on **extrapolation** of **unseen interventions** in **sequential** and **compositional** setting **in the future**.

Problem statement

- Individual user n with optional time-invariant features Z_n;
- Time-series of user's behaviour $X_n^{1:T-1} := [X_n^1, \cdots X_n^{T-1}]$ and associated actions (interventions) $D_n^{1:T-1} := [D_n^1, \cdots D_n^{T-1}]$.

- Predict future behavior $X_n^{T,T+\Delta} := [X_n^T, \ldots, X_n^{T+\Delta}]$ under (hypothetical) future actions $\mathrm{do}(D_n^{1:T+\Delta})$.
- We consider the case where sequential ignorability holds, by randomization or adjustment [16, 10].

Graphical Model



Within unit n, actions D_n^t interact with (latent) variable β_n to produce behavior X_n^t represented as a dense graphical model.

Assumptions

We assume behavioral measurements X_n^t have the following conditional mean factorisation for the regime $\mathrm{do}(D_n^{1:t})$,

$$\mathbb{E}[X_n^t \mid X_n^{1:t-1}, Z_n, \mathrm{do}(D_n^{1:t})] = (\phi_n^t)^\mathsf{T}(\beta_n \odot \psi_n^t) = \sum_{l=1}^{r} \phi_{nl}^t \beta_{nl} \psi_{nl}^t, \tag{1}$$

where $\phi_{nl}^t := \phi_l(x_n^{1:t-1}, z_n)$ (evaluation of basis function) and $\psi_{nl}^t := \prod_{t'=1}^{t} \psi_l(d_n^{t'}, t', t)$ (evaluation of sequential interventions). We choose $\psi_l(d, t', t) := \sigma(w_{1dl})^{t-t'} \times w_{2dl} + w_{3dl}$ (motivated by [6])

Inspiration

- Tensor Factorisation for Causal Imputation [3, 4, 2].
- Functional analysis $f(x_a, x_b) \approx f_a^\mathsf{T}(x_a) f_b(x_b)$ (e.g. Proposition 1 [12]).

Algorithm: CSI-VAE

We call our method *Compositional Sequential Intervention Variational Autoencoder (CSI-VAE)*.

Statistical Inference

- We optimize the (marginal) log-likelihood using a black-box amortized variational inference framework [13, 14, 17].
- We use GRU model to approximate mean-field Gaussian posterior.
- The approximate posterior at time step t is thus $\mu_{q,\beta,n} := \mathrm{MLP}(\mathrm{GRU}_{\eta_{\beta,1}}(d_n^{1:t}, x_n^{1:t}, z_n))$, and $\log \sigma_{q,\beta,n} := \mathrm{MLP}(\mathrm{GRU}_{\eta_{\beta,2}}(d_n^{1:t}, x_n^{1:t}, z_n))$.
- Prediction for $X_n^{T:T+\Delta}$ is done by sampling $M = 50$ trajectories and then use marginal Monte Carlo average.
- More details in paper.

Experiments

Data

- Fully synthetic data.
- Semi-synthetic Spotify data [7].

CSI-VAE Models

- CSI-VAE-1: proposed model.
- CSI-VAE-2: ablation, relaxed the product form of Eq. (1).
- CSI-VAE-3: ablation, relaxed the product form of ψ.

Baselines

- GRU-0: GRU uses $X_n^{1:T-1}$ and Z_n only.
- GRU-1: GRU uses $X_n^{1:T-1}$, D_n^{T-1} and Z_n only.
- GRU-2: GRU uses $X_n^{1:T-1}$, $D_n^{1:T-1}$ and Z_n.
- LSTM: LSTM uses $X_n^{1:T-1}$, $D_n^{1:T-1}$ and Z_n.
- Transformer: Transformer uses $X_n^{1:T-1}$, $D_n^{1:T-1}$ and Z_n.

Main Results

Table 1: Main experimental results, averaged mean squared root error over five different seeds.

	Full Synthetic					Semi-Synthetic Spotify				
Model	T+1	T+2	T+3	T+4	T+5	T+1	T+2	T+3	T+4	T+5
CSI-VAE-1	36.53	41.46	41.73	41.12	41.32	68.23	82.94	83.53	81.97	79.63
CSI-VAE-2	97.80	118.25	117.79	127.25	135.03	253.85	312.53	305.08	303.68	302.83
CSI-VAE-3	138.78	164.02	141.71	132.59	125.55	757.94	937.07	800.55	704.66	634.72
GRU-0	229.72	269.66	220.95	208.30	188.43	215.42	260.65	193.41	137.20	117.06
GRU-1	230.76	270.83	220.93	208.33	184.92	223.61	269.69	205.91	141.53	126.36
GRU-2	93.73	101.03	118.01	88.53	132.28	154.18	187.42	177.96	133.36	127.58
LSTM	114.71	126.65	137.12	105.22	137.19	130.35	156.02	133.28	94.35	85.92
Transformer	111.66	122.08	150.57	175.84	87.89	133.42	157.66	154.61	164.70	158.03

Further Results

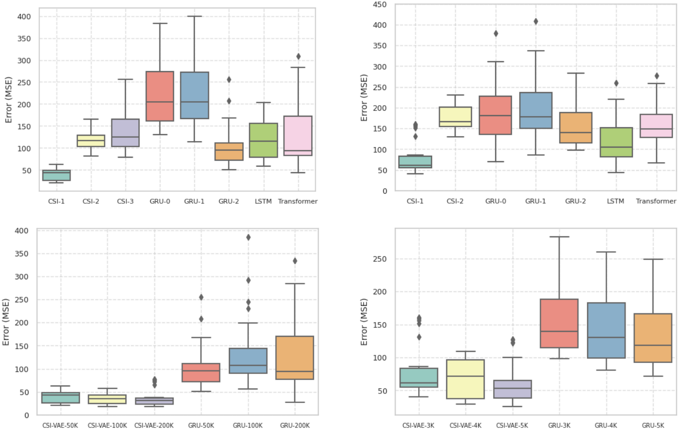


Figure 2: **Top**: Fully-synthetic (left) and semi-synthetic Spotify (right). **Bottom**: How errors change as training sizes are increased.

Summary

- The results show the superiority of our model against strong baselines.
- We also observed that the structural assumptions are critical, as evidenced by the drop in performance for CSI-VAE 2 and 3.
- We show that even with more data provided, our model consistently outperforms the black-box models (cannot solve this problem by simply feeding in more data).

Take Away

- Embedding is important, but how to incorporate structures into embedding is more critical for generalisation.
- Black box models are powerful, but we can make it even more powerful with additional structural information.

References I

[1] Abhineet Agarwal, Anish Agarwal, and Suhas Vijaykumar. Synthetic combinations: A causal inference framework for combinatorial interventions. *Advances in Neural Information Processing Systems*, 36:19195–19216, 2023.

[2] Anish Agarwal, Munther Dahleh, Devavrat Shah, and Dennis Shen. Causal matrix completion. In *The thirty sixth annual conference on learning theory*, pages 3821–3826. PMLR, 2023.

[3] Anish Agarwal, Devavrat Shah, and Dennis Shen. Synthetic interventions. *arXiv preprint arXiv:2006.07691*, 2020.

[4] Susan Athey, Mohsen Bayati, Nikolay Doudchenko, Guido Imbens, and Khashayar Khosravi. Matrix completion methods for causal panel data models. *Journal of the American Statistical Association*, 116(536):1716–1730, 2021.

References II

[5] Gecia Bravo-Hermsdorff, David S Watson, Jialin Yu, Jakob Zeitler, and Ricardo Silva. Intervention generalization: a view from factor graph models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 43662–43675, 2023.

[6] Kay H Brodersen, Nicolas Remy, Steven L Scott, and Fabian Gallusser. Inferring causal impact using bayesian structural time-series models. *Annals of Applied Statistics*, 9:247–274, 2015.

[7] Brian Brost, Rishabh Mehrotra, and Tristan Jehan. The music streaming sessions dataset. In *The World Wide Web Conference*, pages 2594–2600, 2019.

[8] Bibhas Chakraborty and Erica E Moodie. Statistical methods for dynamic treatment regimes. *Springer-Verlag. doi*, 10(978-1):4–1, 2013.

References III

[9] Junyoung Chung, Caglar Gulcehre, Kyunghyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. In *NIPS 2014 Workshop on Deep Learning, December 2014*, 2014.

[10] Miguel A Hernán and James M Robins. Causal inference, 2010.

[11] S Hochreiter. Long short-term memory. *Neural Computation MIT-Press*, 1997.

[12] Jean Kaddour, Yuchen Zhu, Qi Liu, Matt J Kusner, and Ricardo Silva. Causal effect inference for structured treatments. *Advances in Neural Information Processing Systems*, 34:24841–24854, 2021.

[13] Diederik P Kingma. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

References IV

[14] Andriy Mnih and Karol Gregor. Neural variational inference and learning in belief networks. In *International Conference on Machine Learning*, pages 1791–1799. PMLR, 2014.

[15] Razieh Nabi, Todd McNutt, and Ilya Shpitser. Semiparametric causal sufficient dimension reduction of multidimensional treatments. In *Uncertainty in Artificial Intelligence*, pages 1445–1455. PMLR, 2022.

[16] Judea Pearl. *Causality*. Cambridge university press, 2009.

[17] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *International conference on machine learning*, pages 1278–1286. PMLR, 2014.

References V

[18] Yuta Saito, Qingyang Ren, and Thorsten Joachims. Off-policy evaluation for large action spaces via conjunct effect modeling. In *international conference on Machine learning*, pages 29734–29759. PMLR, 2023.

[19] Yuta Saito, Jihan Yao, and Thorsten Joachims. Potec: Off-policy learning for large action spaces via two-stage policy decomposition. *arXiv preprint arXiv:2402.06151*, 2024.

[20] Richard S Sutton. Reinforcement learning: An introduction. *A Bradford Book*, 2018.

[21] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.