eCDANs: Efficient Temporal Causal Discovery from Autocorrelated and Non-stationary Data (Student Abstract)

Muhammad Hasan Ferdous, Uzma Hasan, Md Osman Gani

Causal AI Lab, Department of Information Systems, University of Maryland, Baltimore County
h.ferdous@umbc.edu, uzmahasan@umbc.edu, mogani@umbc.edu

Abstract

Conventional temporal causal discovery (CD) methods suffer from high dimensionality, fail to identify lagged causal relationships, and often ignore dynamics in relations. In this study, we present a novel constraint-based CD approach for autocorrelated and non-stationary time series data (eCDANs) capable of detecting lagged and contemporaneous causal relationships along with temporal changes. eCDANs addresses high dimensionality by optimizing the conditioning sets while conducting conditional independence (CI) tests and identifies the changes in causal relations by introducing a surrogate variable to represent time dependency. Experiments on synthetic and real-world data show that eCDANs can identify time influence and outperform the baselines.

Introduction

Many substantial methods have been developed to estimate the underlying causal mechanism of time series data. But most of these approaches fail when the time series data is non-stationary and autocorrelated. To find causal relationships in autocorrelated data, most constraint-based approaches perform conventional CI tests between variables that may include the whole past and all contemporaneous variables, from which some of the conditioning variables are uncorrelated This results in high dimensionality, lower detection power, and incorrect results (Runge 2020). More recent works used continuous optimization to handle high-dimensionality (Pamfil et al. 2020; Sun et al. 2021). However, such methods can have multiple minima, can not handle data re-scaling, and the returned edges may or may not represent causal relationships (Kaiser and Sipos 2022). Moreover, the seasonal and cyclical nature of variables has a time influence that can change their distributions. This time influence is reflected in *changing modules* and can be represented by a surrogate variable (Zhang et al. 2017). This important component of the time series data is ignored by most of the algorithms.

To address these challenges, we propose an algorithm for efficient CD from autocorrelated and non-stationary (eCDANs) data. Our proposed approach handles high dimensionality by performing PC-stable CI tests and uses the findings to perform momentary CI (MCI) tests. This gives us the adjacent sets (skeleton graph) of contemporaneous variables. The use of MCI tests ensures the exclusion of unrelated variables in conditioning sets. We then use the skeleton graph and a surrogate variable to identify the changing modules. It utilizes the direction of time flow and changing modules, and orientation rules to identify causal directions.

Methodology: `eCDANs`

Our proposed method `eCDANs` discovers causal structure in five steps. We describe these steps below.

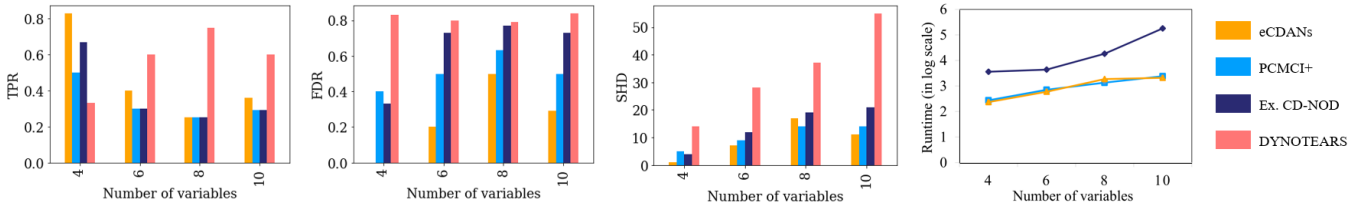
Step 1 – Detection of initial adjacent sets: Let X_t be the contemporaneous variables, X_t^j be the j^{th} observation at time t, $X_{t-\tau}^i$ be the i^{th} observation at lag τ. eCDANs begins by conducting iterative PC1 tests between X_t^j and $X_{t-\tau}^i$ for all i $(i = 1, 2, \ldots, m)$, and derives the superset of lagged adjacent set $L_a(X_t^j)$ for every X_t^j. Then it performs MCI tests to get the adjacent set $Adj(X_t^j)$ for every X_t^j. Variables in adjacency sets are stored according to the descending effect size (test statistic).

Step 2 – Construction of the undirected graph: In this phase, eCDANs creates a partially complete undirected graph G using the lagged adjacencies $L_a(X_t)$, contemporaneous adjacencies $C_a(X_t)$, and the surrogate variable C.

Step 3 – Detection of changing modules: To detect changing modules, eCDANs starts with unconditional independence tests between X_t^j and C for all $j \in (1, m)$; and keeps adding other variables in the conditioning set from $(L_a(X_t^j) \cup C_a(X_t^j))$ according to the descending effect size. This optimizes the conditioning set and the search by reducing redundant conditional tests. It removes the edge between X_t^j and C if they are independent.

Step 4 – Identification of contemporaneous adjacencies: We ignored C in *Step–1* while detecting $C_a(X_t)$ that can discover some false edges. To get rid of these edges, eCDANs performs CI tests between X_t^i and X_t^j, and removes the edge between X_t^i and X_t^j if they are independent conditional on $(L_a(X_t^i) \cup C_a(X_t^i) \cup L_a(X_t^j) \cup C_a(X_t^j) \cup C)$. This step produces a causal skeleton with contemporaneous edges, lagged edges, and the edges between contemporaneous variables and C.

Step 5 – Recovery of causal direction: Using the assumption that the cause-effect relationships follow the flow



(a) TPR (b) FDR (c) SHD (d) Runtime (in sec.)

Figure 1: Performance metrics of different algorithms

of time, it orients $(X^i_{t-\tau},\ X^j_t)$ as $(X^i_{t-\tau} \rightarrow X^j_t)$ for all $\tau = 1,\ 2,\ \ldots,\ \tau_{\max}$. eCDANs orients $\left(C, X^j_t\right)$ as $C \rightarrow X^j_t$ assuming that the surrogate variable causes the distribution shift. For triple of the form $\left(C - X^i_t - X^j_t\right)$, eCDANs recalls the conditional set of the CI test between C and X^j_t. If the conditioning set does not include X^i_t, it orients the triple as $C \rightarrow X^i_t \leftarrow X^j_t$. Otherwise, orients as $C \rightarrow X^i_t \rightarrow X^j_t$. When both X^i_t and X^j_t are adjacent to C, eCDANs uses extended HSIC to orient the edge between X^i_t and X^j_t. eCDANs also uses independent changes of causal modules to determine the causal direction (Huang et al. 2020).

Evaluation

We partially optimized the conditioning sets using lagged parents and benchmarked the performance of eCDANs against three baselines on – (1) synthetic datasets (4, 6, 8, and 10 variables), and (2) a clinical dataset (Gani et al. 2020) database which contains 12 time-series variables1. We discuss the experimental findings below.

Results on synthetic datasets: Figure 1 shows the performance of eCDANs and PCMCI+ (Runge 2020), Extended CD-NOD (Huang et al. 2020), and DYNOTEARS (Pamfil et al. 2020) in terms of True positive rate (TPR), False discovery rate (FDR), and Structural hamming distance (SHD). eCDANs consistently outperformed PCMCI+ and Extended CD-NOD in terms of all matrices. DYNOTEARS has the highest TPR, however, it fails to identify true causal relationships and produces very high FDR and SHD. Hence we did not consider DYNOTEARS for further analysis.

Results on real-world dataset: We can not compare the performance on the real-world dataset due to the unavailability of the ground truth causal graph. Instead, we compared the outcomes with a non-temporal causal graph proposed by (Gani et al. 2020; Bikak et al. 2020), and found that eCDANs produces the closest graph to the non-temporal findings.

Runtime: We present runtime for eCDANs, PCMCI+ and Extended CD-NOD in (Figure 1). Our experimental results show that eCDANs is efficient compared to others and has the lowest runtime consistently. In the 8-variable setting, it has a slightly longer runtime than PCMCI+.

Future Works

This is a work in progress and we are working on finding the contemporaneous adjacencies along with the lagged adjacencies to further optimize the CI tests. This will further improve the performance and runtime of the proposed algorithm. We also plan to extend eCDANs when the time series dataset contains latent confounders along with changing modules to identify a partial ancestral graph.

Acknowledgements

This study was supported in parts under grants from NSF (Award # 2118285) and UMBC START.

References

Bikak, M.; Kethireddy, S.; Gani, M. O.; and Adibuzzaman, M. 2020. Structural Causal Model with Expert Augmented Knowledge to Estimate the Effect of Oxygen Therapy. *CHEST*, 158(4): A636.

Gani, M. O.; Kethireddy, S.; Bikak, M.; Griffin, P.; and Adibuzzaman, M. 2020. Structural Causal Model with Expert Augmented Knowledge to Estimate the Effect of Oxygen Therapy on Mortality in the ICU. *arXiv preprint arXiv:2010.14774*.

Huang, B.; Zhang, K.; Zhang, J.; Ramsey, J. D.; Sanchez-Romero, R.; Glymour, C.; and Schölkopf, B. 2020. Causal Discovery from Heterogeneous/Nonstationary Data. *J. Mach. Learn. Res.*, 21(89): 1–53.

Kaiser, M.; and Sipos, M. 2022. Unsuitability of NOTEARS for Causal Graph Discovery when Dealing with Dimensional Quantities. *Neural Processing Letters*, 1–9.

Pamfil, R.; Sriwattanaworachai, N.; Desai, S.; Pilgerstorfer, P.; Georgatzis, K.; Beaumont, P.; and Aragam, B. 2020. Dynotears: Structure learning from time-series data. In *International Conference on Artificial Intelligence and Statistics*, 1595–1605. PMLR.

Runge, J. 2020. Discovering contemporaneous and lagged causal relations in autocorrelated nonlinear time series datasets. In *Conference on Uncertainty in Artificial Intelligence*, 1388–1397. PMLR.

Sun, X.; Liu, G.; Poupart, P.; and Schulte, O. 2021. NTS-NOTEARS: Learning Nonparametric Temporal DAGs With Time-Series Data and Prior Knowledge. *arXiv preprint arXiv:2109.04286*.

Zhang, K.; Huang, B.; Zhang, J.; Glymour, C.; and Schölkopf, B. 2017. Causal discovery from nonstationary/heterogeneous data: Skeleton estimation and orientation determination. In *IJCAI: Proceedings of the Conference*, volume 2017, 1347. NIH Public Access.