Causal Discovery from Sparse Time-Series Data Using Echo State Network

Haonan Chen[†1]   University of Waterloo, Waterloo, ON Canada
Bo Yuan Chang[†1]   University of Waterloo, Waterloo, ON Canada
Mohamed A. Naiel[1]   University of Waterloo, Waterloo, ON Canada
Georges Younes[1]   University of Waterloo, Waterloo, ON Canada
Steven Wardell[2]   ATS Automation, Cambridge, ON Canada
Stan Kleinikkink[2]   ATS Automation, Cambridge, ON Canada
John S. Zelek[1]   University of Waterloo, Waterloo, ON Canada
Email: [1]{haonan.chen, by2chang, mohamed.naiel,georges.younes,jzelek}@uwaterloo.ca
      [2]{swardell, skleinikkink}@atsautomation.com

Abstract

Causal discovery between collections of time-series data can help diagnose causes of symptoms and hopefully prevent faults before they occur. However, reliable causal discovery can be very challenging, especially when the data acquisition rate varies (*i.e.,* non-uniform data sampling), or in the presence of missing data points (*e.g.,* sparse data sampling). To address these issues, we propose a new system comprised of two parts, the first part fills missing data with a Gaussian Process Regression, and the second part leverages an Echo State Network, which is a type of reservoir computer (i.e., used for chaotic system modelling) for Causal discovery.

We evaluate the performance of our proposed system against three other off-the-shelf causal discovery algorithms, namely, structural expectation maximization, sub-sampled linear auto-regression absolute coefficients, and multivariate Granger Causality with vector auto-regressive using the Tennessee Eastman chemical dataset; we report on their corresponding Matthews Correlation Coefficient (MCC) and Receiver Operating Characteristic curves (ROC) and show that the proposed system outperforms existing algorithms, demonstrating the viability of our approach to discover causal relationships in a complex system with missing entries. [1]

1 Introduction

Artificially replicating and surpassing human's ability to interpret causal relationships is considered an evolutional step forward when compared to existing Artificial Intelligence systems [1]. Despite several decades of research, current state-of-the-art deep learning systems are still broadly interpreted as a family of highly nonlinear statistical models [2] or pattern detection regression engines that require a very large set of training samples to distill a meaningful outcome for a particular problem. In contrast, humans are capable of drawing conclusions and solving problems by identifying causal relationships between the different variables of a task using a very small amount of data. This inspired a large body of researchers to work on causal discovery algorithms in an attempt to answer various challenges in several fields, such as epidemiology [3], economy [4, 5], and medicine [6] to name a few. However, due to the complex relationships between a large number of hidden variables, learning causal relationships on real-world data can be very challenging as causality must be inferred from noisy data [7, 8]. Moreover, it is not uncommon to run into missing values when working with real-world data, especially when hardware limitations and human errors play an important factor in the data acquisition process (*e.g.,* time series data).

Causal discovery in time-series data is of particular importance in this work as we attempt to retrieve causal relationships from sparsely sampled time-series data. Despite the existence of several methods for causal discovery [9–11], there is currently a limited availability for methods that address causal discovery from *sparse* time-series data. While the effects of missing data points from a sparsely sampled sequence can be often solved with hardware upgrades, or better training for the employees, such solutions are often associated with a steep price tag and can be time consuming. Alternatively, one could make use of data filling interpolation methods to populate the sparse time-series sequence before performing causal learning as in [12]. To that end, we propose a system that makes use of Gaussian Process Regression (GPR) to fill missing data points, which are in turn processed with an Echo State network tailored to the causal discovery of relationships between its nodes.

2 Architecture

Our proposed approach is summarized in Fig. 1 where T is a time-series input with some missing data, T^* represents the multivariate time-series that was data-filled by GPR, and $GC(x_i, x_j)$ is the causal relation between x_i and x_j. The input is a $N \times M$ multivariate time-series where N is the total number of entries for each variable (feature) and M is the number of variables. The GPR filled time-series is then processed through the GC-ESN estimator (dynamic reservoir), to generate an $M \times M$ causality matrix, which can then be displayed as a heat map of causal relations amongst variables.
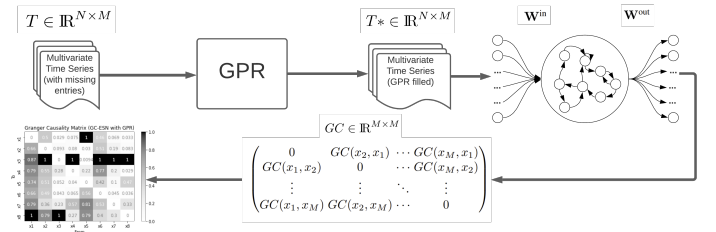

Fig. 1: Workflow of the proposed solution; a time-series with missing data is filled using GPR and fed into a GC-ESN to predict causal relations.

3 Experiments and Evaluation

3.1 Dataset Description

We evaluate our proposed system on the Tennessee Eastman's Process Dataset (or TE for short) [13]. TE is a dataset that simulates actual chemical processes, and has been widely applied in the study of fault diagnosis and root cause analysis [14] [15]. The process flow consists of 5 major physical components: the reactor, condenser, vapor-liquid separator, compressor, and the product stripper. There are several sensor measurements available (such as flow rate, temperature, pressure and feed rate) for each components. Overall, the TE dataset contains 500 operation cycles' measurement from 52 sensor readings. Each operation cycle consists of 500 entries being recorded every 3 minutes for a total duration of 1500 minutes. For further information about the TE dataset, the interested reader is referred to [16].

We select 8 sensors readings from the first 200 operation cycle's entry inside the training fault free file for our experiment. The description for the 8 selected variables are shown in Table 1. In the the experimental procedure, we aim to recover these causal relationships and compare against their ground truth values.

3.2 Experimental Setup

The selected TE data is first contaminated by removing 10% of its original entries. This provides ground truth data to validate the success of our proposed data filling process. Figure 2 shows the original *vs.* filled data for sensor 5 (recycling flow rate measurement). The contaminated data is then processed through the proposed

---

Table 1: Selected Sensor's Descriptions

Variable ID	Header in data	Description	Units
1	xmeas_5	Recycle Flow	km^3/h
2	xmeas_6	Reactor feed rate	km^3/h
3	xmeas_7	Reactor Pressure	kPa
4	xmeas_8	Reactor Level	%
5	xmeas_9	Reactor Temperature	°C
6	xmeas_12	Separator Level	%
7	xmeas_20	Compress Work	KW
8	xmeas_21	Reactor cooling water Outlet Temperature	°C

GPR method to fill out the missing entire, and subsequently fed into the Echo State Network to discover causal relationships among the variables.
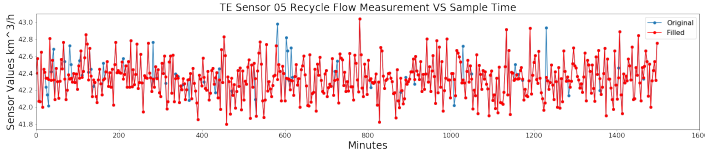

Fig. 2: Original Data VS 10% GPR Filled Data

The process is repeated on three other off-the-shelf causal discovery algorithms, namely:

- Structural Expectation Maximization (SEM) [17]: a graphical approach that is capable of learning causal relationships from sparsely sampled time series data; as such, the GPR process is not used for this algorithm, and the missing data are directly fed into the SEM algorithm for causal learning. We report on the results of *bnstruct* [18], an R implementation of SEM.
- Subsampled Linear Auto-Regression Absolute Coefficients algorithm (SLARAC): a classical approach for causal discovery written in python, and can be found at [9].
- Multivariate Granger Causality (MVGC): a classical approach written in Matlab and can be found at [11].

The GPR filling is applied to both classical approaches before performing causal discovery. To validate the impact of the data filling process, we also include our results obtained from then ESN with the original data (no missing values).

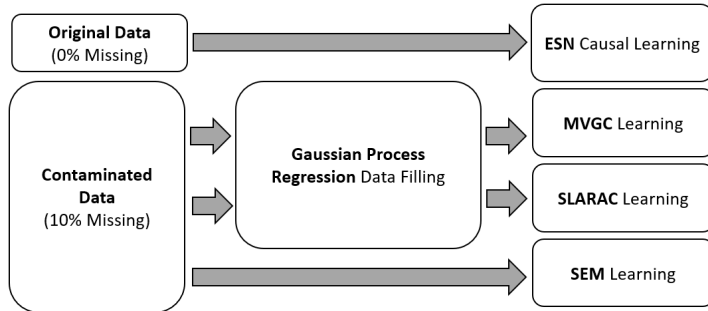The experimental comparison setup for the various systems is summarized in Fig. 3.

Fig. 3: The Proposed Experiment Setup for Performance Comparison.

3.3 Evaluation Metrics

We validate the performance of the various systems using the Matthews Correlation Coefficient (MCC) score:

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}, \quad (1)$$

that reflects the *usefulness* of each causal prediction. The MCC score provides a balanced measure between the True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) cases; its output is $\in [-1, 1]$, where a +1 score indicates a perfect prediction, 0 indicates that the prediction made is no better than random guessing, and -1 indicates complete disagreement between prediction and observation [19].

In addition to the MCC index comparison, we also report on the ROC (Receiver Operating Characteristic) curves for the various causal discovery systems. The ROC curves offer valuable insights into the specificity and sensitivity of each model at different cutoff thresholds in the causal matrices. We also report on the AUC (Area Under Curve) score for all ROC curves. Note that due to the binary matrix output nature of the SEM algorithm, we were unable to plot its ROC curve.

3.4 Results and Discussion

Figure 4 shows the causal relations recovered from our proposed GC-ESN system (a) side-by-side with its corresponding ground truth causality matrix (b). On the other hand, table 2 summarizes the MCC index of the proposed system (GPR + ESN 10% Missing) against the other algorithms.
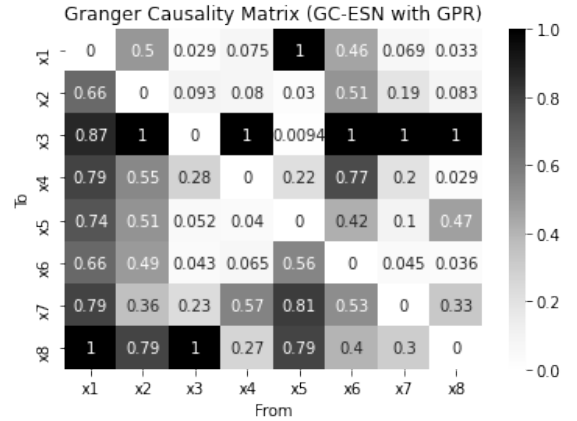
(a) Causal Matrix of GC-ESN with 10% GPR
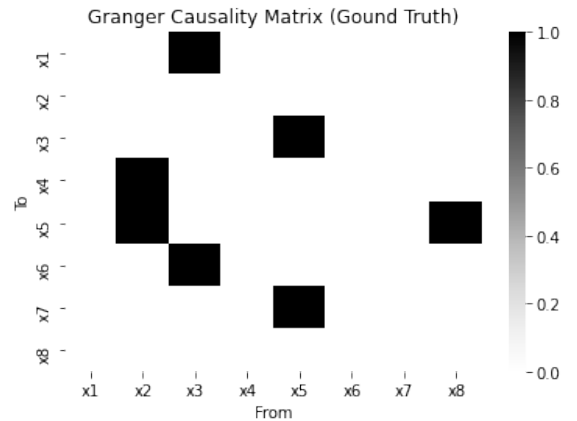
(b) The Ground Truth Causal Matrix

Fig. 4: Causal Matrix Comparison between GC-ESN with 10% Missing Data Filling using GPR and the Ground Truth.

Table 2: MCC Score Results

	GC-ESN 0% Missing	GPR + GC-ESN 10% Missing	GPR + SLARAC 10% Missing	MVGC 10% Missing	SEM 10% Missing
TP	5	5	7	7	0
FP	13	15	47	48	13
TN	44	42	8	9	44
FN	2	2	0	0	7
MCC	0.34	0.31	0.15	0.14	-0.18

As shown in Table. 2, our system was capable of recovering five of the seven causal relations by applying thresholds that yield the highest F1 scores, and despite the 10 % missing entries, our proposed system still achieved an MCC index (Table 2) of 0.31 which is slightly lower (by 0.03) than the MCC index obtained with the original data (using ESN causal learning). This indicates that the GPR filling process we adapted inside our system is very effective and did restore a satisfactory amount of information comparable to that of the original data. Furthermore, compared to the remaining algorithms on missing data, the proposed GC-ESN estimator achieved the highest MCC score. This indicates that our proposed system is

capable of offering more precise and reliable causal links suggestions than the other systems.
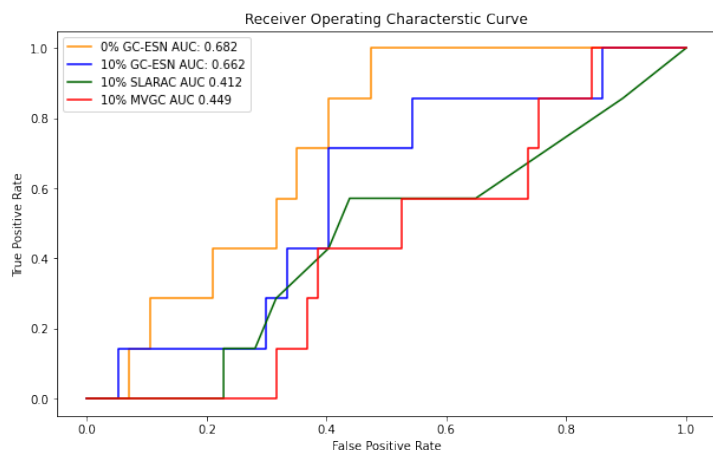

Fig. 5: ROC Curve for Different Missing Percentage GC-ESN and other estimators. The proposed system is able to outperform the rest of the estimators by a noticeable margin in their AUC scores.

The results of the ROC curves (shown in Fig. 5) further solidify our claims as our proposed GC-ESN system reports a significantly larger AUC score than the rest(other than ESN on the original data). Moreover, by varying the threshold values, GC-ESN has the highest true positive rate, proving that among the three estimators, GC-ESN exhibits the best performance.

4 Conclusion

We have proposed a system capable of performing causal discovery for sparsely sampled multivariate time series data. The system consists of two parts: (1) Data filling with Gaussian Process Regression, and (2) causal learning with an Echo State Network. The proposed system is evaluated on the Tennessee Eastman (TE) process dataset with 10 percent missing entries. In order for us to evaluate the performance, the proposed system is compared and shown to out perform several other methods including Structural Expectation Maximization (SEM), Subsampled Linear Auto-Regression Absolute Coefficients (SLARAC), and Multivariate Granger Causality (MVGC).
We also perform an ablation study to evaluate the effectiveness of the GPR in recovering causal information by comparing its results to that of causal discovery using the original (uncontaminated) data, and found that the proposed data filling process is capable of recovering causal relationships reliably and performed only marginally worse had the full original data was used. This work shows promise in recovering causal relationships from imperfect data better than current SOTA (State Of The Art) methods.
The obtained results show great potential in applying the proposed system in more complicated real world scenarios as it outperforms all other methods by a comfortable margin in both AUC scores and MCC indices.
That being said, the proposed system still fall short in comparison to a human subject-expert in identifying causal relationships; as such, causal discovery remains an ongoing research topic.

Acknowledgments

References

[1] J. Pearl, "Theoretical impediments to machine learning with seven sparks from the causal revolution," *CoRR*, vol. abs/1801.04016, 2018.

[2] P. L. Bartlett, A. Montanari, and A. Rakhlin, "Deep learning: a statistical viewpoint," *arXiv preprint arXiv:2103.09177*, 2021.

[3] M. Hernán, B. Brumback, and J. Robins, "Marginal structural models to estimate the causal effect of zidovudine on the survival of hiv-positive men." *Epidemiology*, vol. 11 5, pp. 561–70, 2000.

[4] J. Hicks *et al.*, *Causality in economics*. Australian National University Press, 1980.

[5] J. J. Heckman, "Econometric causality," *International statistical review*, vol. 76, no. 1, pp. 1–27, 2008.

[6] R. P. Thompson, "Causality, mathematical models and statistical association: dismantling evidence-based medicine," *Journal of Evaluation in Clinical Practice*, vol. 16, no. 2, pp. 267–275, 2010.

[7] R. Guo, L. Cheng, J. Li, P. R. Hahn, and H. Liu, "A survey of learning causality with data," *ACM Computing Surveys*, vol. 53, no. 4, p. 1–37, Sep 2020.

[8] L. Yao, Z. Chu, S. Li, Y. Li, J. Gao, and A. Zhang, "A survey on causal inference," 2020.

[9] S. Weichwald, M. E. Jakobsen, P. B. Mogensen, L. Petersen, N. Thams, and G. Varando, "Causal structure learning from time series: Large regression coefficients may predict causal links better in practice than small p-values," in *Proceedings of the NeurIPS 2019 Competition and Demonstration Track*, ser. Proceedings of Machine Learning Research, H. J. Escalante and R. Hadsell, Eds., vol. 123. PMLR, 08–14 Dec 2020, pp. 27–36. [Online]. Available: http://proceedings.mlr.press/v123/weichwald20a.html

[10] M. Nauta, D. Bucur, and C. Seifert, "Causal discovery with attention-based convolutional neural networks," *Machine Learning and Knowledge Extraction*, vol. 1, no. 1, pp. 312–340, Jan. 2019.

[11] L. Barnett and A. K. Seth, "The mvgc multivariate granger causality toolbox: A new approach to granger-causal inference," *Journal of Neuroscience Methods*, vol. 223, pp. 50–68, 2014.

[12] B. Chang, M. Naiel, S. Wardell, S. Kleinikkink, and J. Zelek, "Time-series causality with missing data," *Journal of Computational Vision and Imaging Systems*, vol. 6, no. 1, pp. 1–4, Jan. 2021. [Online]. Available: https://openjournals.uwaterloo.ca/index.php/vsl/article/view/3552

[13] X. Chen, "Tennessee eastman simulation dataset," 2019. [Online]. Available: https://dx.doi.org/10.21227/4519-z502

[14] X. Chen, J. Wang, and J. Zhou, "Probability density estimation and bayesian causal analysis based fault detection and root identification," *Industrial & Engineering Chemistry Research*, vol. 57, no. 43, pp. 14 656–14 664, 2018.

[15] H. Gharahbagheri, S. A. Imtiaz, and F. Khan, "Root cause diagnosis of process fault using kpca and bayesian network," *Industrial & Engineering Chemistry Research*, vol. 56, no. 8, pp. 2054–2070, 2017.

[16] C. A. Rieth, B. D. Amsel, R. Tran, and M. B. Cook, "Issues and advances in anomaly detection evaluation for joint human-automated systems," in *Advances in Human Factors in Robots and Unmanned Systems*, J. Chen, Ed. Cham: Springer International Publishing, 2018, pp. 52–63.

[17] N. Friedman, "The bayesian structural em algorithm," 2013.

[18] A. Franzin, F. Sambo, and B. di Camillo, "bnstruct: an r package for bayesian network structure learning in the presence of missing data," *Bioinformatics*, vol. 33, no. 8, pp. 1250–1252, 2017.

[19] B. Matthews, "Comparison of the predicted and observed secondary structure of t4 phage lysozyme," *Biochimica et Biophysica Acta (BBA) - Protein Structure*, vol. 405, no. 2, pp. 442–451, 1975.