Shadow-Mapping for Unsupervised Neural Causal Discovery

Matthew J. Vowels
m.j.vowels@surrey.ac.uk

Necati Cihan Camgoz
n.camgoz@surrey.ac.uk

Richard Bowden
r.bowden@surrey.ac.uk

Centre for Vision, Speech and Signal Processing
University of Surrey
Guildford, UK

Abstract

An important goal across most scientific fields is the discovery of causal structures underling a set of observations. Unfortunately, causal discovery methods which are based on correlation or mutual information can often fail to identify causal links in systems which exhibit dynamic relationships. Such dynamic systems (including the famous coupled logistic map) exhibit 'mirage' correlations which appear and disappear depending on the observation window. This means not only that correlation is not causation but, perhaps counter-intuitively, that causation may occur without correlation. In this paper we describe Neural Shadow-Mapping, a neural network based method which embeds high-dimensional video data into a low-dimensional shadow representation, for subsequent estimation of causal links. We demonstrate its performance at discovering causal links from video-representations of dynamic systems.

1. Introduction

Understanding causal structure is essential to the scientific endeavour [15]. Over recent decades, numerous methods have been proposed which seek to discover causal structure from data (for reviews, see [22, 16, 6, 5]) but the task is inherently challenging. Numerous solutions may exist which are sufficient to explain the data, and causal links estimated using associational measures, such as correlation or mutual information, may fail in systems which exhibit complex state-based dependencies [20]. These state dependent systems have been said to demonstrate *mirage* correlations which appear and vanish over time. One such system is given by the well-known coupled logistic map difference equations [12]:

$$\begin{aligned} X[n+1] &= X[n](r_x - r_x X[n] - \beta_{xy} Y[n]) \\ Y[n+1] &= Y[n](r_y - r_y Y[n] - \beta_{yx} X[n]) \end{aligned} \tag{1}$$

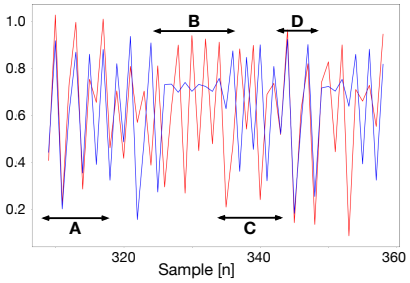

Figure 1. Illustration of *mirage* correlation for Eq. 3 with $r_x = 3.8$, $r_y = 3.8$, $\beta_{xy} = .02$, and $\beta_{yx} = .1$ (*i.e.*, there exists bi-directional causality). Region **A** exhibits positive correlation, **B** low correlation, **C** negative correlation, and **D** returns to positive correlation. Example adapted from [20].

Here, $X[n]$ and $Y[n]$ are two discrete-time varying quantities with parameters r_x and r_y and which causally influence each other via β_{xy} and β_{yx}, respectively. Figure 1 demonstrates not only that correlation is not causation, but also that causation does not necessarily imply correlation, and thus a different approach is needed.

Dynamic systems, such as those described using Eq. 3, occur frequently in nature [12], and it is therefore important that causal discovery methods can be applied. Unfortunately, it is understood that the most well-known and popular paradigm for modeling causal relations *Granger Causality* does not perform well in such systems [20]. This is because Granger causality assumes *separability*, which refers to the independence of the variables in the absence of causal interactions. In dynamic systems, where the current state of a variable may be heavily determined by the past of another, separability is unlikely to hold.

Shadow Embeddings: This failure of Granger Causality has motivated a family of causal discovery methods that operate on time-delayed coordinate embeddings. Takens [21] showed that by concatenating time-delay versions of the time series observations, one can recover the dynamics of the full system even if only one, or a limited number,

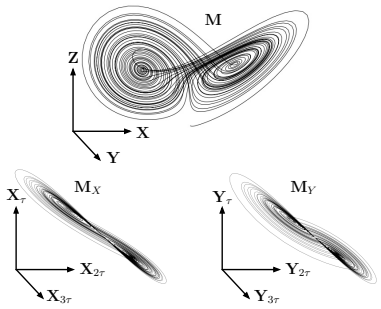

Figure 2. A 3D Lorenz attractor $\mathbf{M}$ and corresponding shadow manifolds $\mathbf{M}_X$ and $\mathbf{M}_Y$ constructed using three lags (lag degree τ).

of observational variables are used. This time-delay embedding is known as a *shadow manifold*. Specifically, assuming manifold $\mathbf{M}$ and T time-based observations $\mathbf{X} : \mathbf{M} \rightarrow \mathbb{R}$, where $\mathbf{X} = \{x_1, x_2, ...x_T\}$, the shadow manifold $\mathbf{M}_X$ is a Hankel matrix of delayed segments from $\mathbf{X}$:

$$\mathbf{M}_X = \begin{bmatrix} x_1 & x_2 & \cdots & x_w \\ x_2 & x_3 & \cdots & x_{w+1} \\ \vdots & \vdots & \ddots & \vdots \\ x_p & x_{p+1} & \cdots & x_{p+w} \end{bmatrix} \quad (2)$$

where p is the number of lags, and w is the segment/window size. An example of the Lorenz dynamic system being represented using two shadow manifolds is shown in Figure 2. $\mathbf{M}_X$ and $\mathbf{M}_Y$ are diffeomorphic embeddings of the system $\mathbf{M}$.

Convergent Cross-Mapping (CCM): The embedding spaces can be used to identify causal links between the variables $\mathbf{X}$ and $\mathbf{Y}$ using CCM [20]. Figure 3 shows shadow manifolds $\mathbf{M}_X$ and $\mathbf{M}_Y$. Specifically, consider a particular time point in $\mathbf{M}_X$, illustrated with the star icon on the left hand side of the figure. The corresponding timepoint t^* in $\mathbf{M}_Y$, indicated by the star in the right hand side of the figure. Additionally, there exists a collection of nearest neighbors illustrated with circles near to the point $\mathbf{M}_X^{t^*}$. The time indices I_X of this set of nearest neighbors $\mathbf{M}_X^{I_X}$ may be far apart, despite their closeness in this shadow space. We can do the same thing in reverse for M_Y by picking a time point $t^{\square}$ (the square on the right hand side), and finding the time indices I_Y for nearest neighbors $\mathbf{M}_Y^{I_Y}$. If the neighborhood of nearest neighbor indices I_X, derived from $\mathbf{M}_X$, induces a dense neighborhood $\mathbf{M}_Y^{I_X}$ then we may conclude that there exists a causal link $\mathbf{Y} \rightarrow \mathbf{X}$. This relationship can be tested in both directions (as illustrated in Figure 3) to test whether $\mathbf{X} \rightarrow \mathbf{Y}$. Furthermore, this process can be used to iteratively test more complex causal systems. For example, if the structure is confounding (*i.e.*, a v-structure) such that $\mathbf{X} \leftarrow \mathbf{Z} \rightarrow \mathbf{Y}$, then the CCM test would establish this dependency structure.

Prior Work: A number of methods exist which leverage the principles behind CCM. Besides the original presentation of the method itself [20], it has been extended to identify lags [23], evaluated for its robustness to noise [14], improved using multivariate shadow embeddings [13], and adapted for neural network integration [11]. A related method involves the use of reservoir computing methods to improve upon the efficiency of CCM [7].

More generally, methods fall into three predominant groups: score-based, constraint based, and asymmetry based. Score-based methods are learned by evaluating the fit of the observations to a proposed structure, constraint based methods test for statistical conditional independencies, and asymmetry based methods test for differences that arise when causality occurs in one direction compared with the other. Some of the most well-known methods for causal discovery are constraint based and include the PC algorithm [19] and FCI [19], both of which test for conditional independencies in the data. Asymmetry based algorithms include LiNGAM [18] and IGCI [8], and score based methods include GES [2], and neural network methods such as NOTEARS [24].

Contributions: Shadow embedding methods for discovering causal structure are not widely used in machine learning. In a recent review of over 100 causal discovery methods [22] only a small number of methods leveraged the principles behind CCM. The goals of this paper are therefore two-fold. (1) To introduce the principles behind CCM and dynamic systems causality to the machine learning community, with hopes for its wider adoption and exploration. (2) We present *Neural Shadow-Mapping* (NSM), a method for causal discovery from video. As far as we are aware, NSM is the first application of the shadow-manifold causal discovery principles to image/video data.

2. Method

The block diagram for NSM is shown in Figure 4. The process may be broken down into 5 steps. **Step (1)**: Video frames are used to train a Pyro [1] implementation of the unsupervised scene understanding method Attend, Infer, Repeat (AIR) [4]. This method provides frame-by-frame estimations of the object's positions. In the example used in

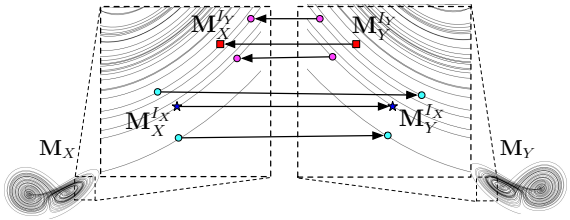

Figure 3. Illustrating corresponding neighborhoods of points in the shadow manifolds (see text for details).

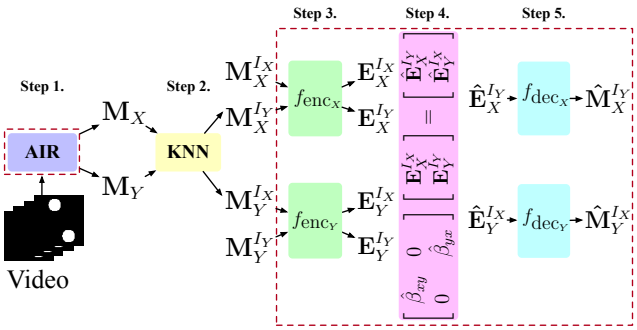

Figure 4. Block diagram for NSM on a 2D example. Functions in red-dashed boxes are trained using gradient descent. Functions $f_{(.)}$ are fully connected neural networks.

Video Graph	p-val. ($\hat{\beta}_{XY}$)	p-val. ($\hat{\beta}_{YX}$)	Identified?
$X \quad Y$	0.61	0.24	✓
$X \rightarrow Y$	0.24	1.88e-6	✓
$X \leftarrow Y$	3.31e-8	0.45	✓
$X \leftrightarrow Y$	3.16e-19	1.66e-12	✓
Time-Series Graph		**Threshold**	**Identified?**
$X \rightarrow Y \quad Z$		0.25	✓
$X \leftrightarrow Y \quad Z$		0.25	✓
$X \rightarrow Z \rightarrow Y \rightarrow X$		0.25	✓
$Y \leftarrow X \rightarrow Z$		0.25	✓
$X \rightarrow Z \leftarrow Y$		0.25	✓
$X \leftrightarrow Z \rightarrow Y$		0.25	✓
$X \leftrightarrow Z \leftrightarrow Y$		0.25	✓

Table 1. Upper: p-values from a one sided, 2-sample KS test between estimated path coefficients on video embeddings of bivariate data with the structure represented by corresponding graphs, and the IAAFT surrogates of those data ($p > 0.01$ means no effect). Lower: the lower portion is for trivariate time-series data.

the figure, there are two objects which each vary in their horizontal positions. There are therefore two system dimensions, and AIR provides time series for these dimensions $\mathbf{X}$ and $\mathbf{Y}$. **Step (2)**: Shadow embeddings $\mathbf{M}_X$ and $\mathbf{M}_Y$ are formed, and fed into a k-nearest neighbors algorithm [3] to yield the neighborhoods indexed with the nearest neighbor indices I from each embedding $\mathbf{M}_X$ and $\mathbf{M}_Y$, at each timepoint. For an example neighborhood at a single timepoint, we have *e.g.*, $\mathbf{M}_X^{I_Y} \in \mathbb{R}^{p \times k}$, where p is the number of lags used to form the shadow embedding, and k is the number of nearest neighbors. We can form a batch over each indexed manifold by randomly selecting time points t^* around which to form neighborhoods. **Step (3)**: $\mathbf{M}_X^{I_X}$ and $\mathbf{M}_X^{I_Y}$ are encoded using three fully-connected layers with non-linearity f_{enc_X}, whilst $\mathbf{M}_Y^{I_X}$ and $\mathbf{M}_Y^{I_Y}$ are fed through an equivalent fully-connected encoder f_{enc_Y}. Separate networks were used for each neighborhood-index/variable combination to maintain independence. **Step (4)**: The encoders yield new embeddings $\mathbf{E}_{(.)}^{(.)}$ of the indexed shadow embeddings. The motivation for the encoder is yield a lower-dimensional representation ameanable to interpretable linear regression: $\mathbf{A}\bar{\mathbf{E}} = \hat{\mathbf{E}}$ where $\mathbf{A}$ is a learnable, diagonal weight matrix intended to discover the causal links β_{xy} and β_{yx}. $\bar{\mathbf{E}}$ is a stacked matrix of row vectors $\mathbf{E}_X^{I_X}$ $\mathbf{E}_Y^{I_Y}$, that we use to predict $\hat{\mathbf{E}}$, which constitutes $\mathbf{E}_X^{I_Y}$ and $\mathbf{E}_Y^{I_X}$, respectively. This step (4) therefore represents a modified version of the CCM element of the process, because traditional cross-mapping would use e.g. $\mathbf{E}_X^{I_X}$ to predict $\mathbf{E}_Y^{I_X}$ - the superscript indices I_X would be used to index Y. In practice we found that mapping 'within-variable' using different indices yields better results. The resulting $\hat{\beta}$ coefficients were biased due to the inherent correlation for within-variable mapping, but this can be accounted for in the surrogate testing process. An l_2 loss between $\hat{\mathbf{E}}$ and $[\mathbf{E}_X^{I_Y}, \mathbf{E}_Y^{I_X}]$ is used to optimize the non-zero parameters in $\mathbf{A}$. **Step (5)**: The final step involves reconstruction of the indexed manifolds $\mathbf{M}_X^{I_Y}$ and $\mathbf{M}_Y^{I_X}$ from the estimated $\mathbf{E}_X^{I_Y}$ and $\mathbf{E}_Y^{I_X}$ via fully-connected decoders f_{dec_X} and f_{dec_Y}, respectively. The error on the reconstructions is measured using the l_2 loss which is also used to drive learning via backpropagation for steps (3)-(5).

3. Experiments

In order to demonstrate this method, two synthetic datasets were created. The first is a video dataset, whereby the horizontal positions of two objects varied according to the coupled logistic map in Equation 3. Four time steps are given as example in Figure 5. To explore the 3-variable case, we also create tri-variate time series data according to the equation below:

$$X_i[n+1] = X_i[n](r_i - r_i X_i[n] - \beta_{ij} X_j[n] - \beta_{ik} X_k[n]) \tag{3}$$

for $i, j, k = \{1, 2, 3\}$.[1] We set all $r_{(.)} = 3.9$ and all non-zero $\beta_{(.,.)} = 0.25$. For both datasets, data are generated using random $\sim U[0, 1]$ initial conditions, and the learned parameter matrix $\mathbf{A}$ is amortized over all these generations. The training process is repeated to acquire distributions over the estimated parameters $\hat{\boldsymbol{\beta}}$ in $\mathbf{A}$, which are then inspected to discover causal connections. The discovery process itself involves testing for significant increase above a non-causal baseline. This baseline is established by creating time series surrogates based on the Iterative Amplitude Adjusted Fourier Transform (IAAFT) method [10, 17]. IAAFT creates surrogates which match the original data in

[1] In Table 1, $X_{\{1,2,3\}}$ are equivalent to X, Y, Z.

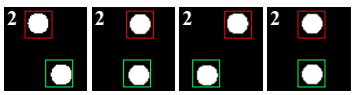

Figure 5. Four timesteps from the two-variable video, overlaid bounding boxes from AIR [4].

terms of their power spectra and therefore also their autocorrelation.[2] The set of resulting baseline parameters can be used as a 'null' distribution and tested against the parameters of the original data using a one-sided, two-sample Kolmogorov Smirnoff (KS) test. If the KS test statistic has a p-value below the false-positive threshold α, then one may infer statistical significance. We found that the KS test was not required for the trivariate time-series data, and that a simple fixed threshold of 0.25 could be used to infer the presence of an effect. For the experiments, we set the number of runs $N_r = 100$, $\alpha = 0.01$, $p = 10$, $w = 790$, $k = 10$, time series length $= \{20, 1000\}$, embedding dimension for $\mathbf{E} = 6$, batch size $= 20$, number of training iterations $3e5$, learning rate $3e-4$, and use an Adam [9] optimizer. No hyperparameter tuning was performed.

The results for the two variable and three variable datasets are shown in Table 1 where 'identified?' indicates whether a link was discovered (but not the magnitude of the link). NSM successfully recovered the true graph in all cases.

4. Discussion and Limitations

NSM was able to discover causal links from video and time-series data generated from dynamic systems. This is a notable result, particularly because visually identifying causality from the video is non-trivial. As is the case for traditional CCM, the discovery of causal links is sensitive to measurement noise and stochasticity. Furthermore, the disadvantages associated with nearest-neighbor methods, include the need for longer time series. Future work should identify ways to improve its robustness to noise, possibly by using existing techniques, and to operate with shorter time series [14]. The aim should then be to test the generalization of the method in discovering causal links in more challenging problems such as those involving human interaction or traffic data.

References

[1] Eli Bingham, Jonathan P. Chen, and Martin Jankowiak et al. Pyro: Deep universal probabilistic programming. *JMLR*, 20, 2019.

[2] D.M. Chickering. Optimal structure identification with greedy search. *JMLR*, 3(Nov), 2002.

[3] T.M. Cover and P.E. Hart. Nearest neighbor pattern classification. *IEEE Trans. on Inf. Thr.*, 13(1):21–27, 1967.

[4] S. M. A. Eslami, N. Heess, T. Weber, Y. Tassa, D. Szepesvari, K. Kavukcuoglu, and G. E. Hinton. Attend, infer, repeat: fast scene understanding with generative models. *arXiv:1603.08575v3*, 2016.

[2] We also validated these surrogates by creating non-causal system data (*i.e.*, by setting $\beta_{xy} = \beta_{yx} = 0$) and comparing the results with the IAAFT results.

[5] C. Glymour, K. Zhang, and P. Spirtes. Review of causal discovery methods based on graphical models. *Frontiers in Genetics*, 10, 2019.

[6] C. Heinze-Deml, M.H. Maathuis, and N. Meinshausen. Causal structure learning. *Ann. Rev. Stat. App.*, 5, 2018.

[7] Y. Huang, Z. Fu, and C.L.E. Franzke. Detecting causality from time series in a machine learning framework. *Chaos*, 20, 2020.

[8] D. Janzing, J. Mooij, J. Zhang, K.and Lemeire, J. Zscheischler, P. Daniusis, B. Steudel, and B. Schölkopf. Information-geometric approach to inferring causal directions. *Artificial Intelligence*, 182-183, 2012.

[9] D. P. Kingma and J. L. Ba. Adam: a method for stochastic optimization. *arXiv:1412.6980v9*, 2017.

[10] J.H. Lucio, R. Valdes, and L.R. Rodriguez. Improvements to surrogate data methods for nonstationary time series. *Phys Rev E Stat Nonlin Soft Matter Phys*, 85, 2012.

[11] H. Ma, K. Aihara, and L. Chen. Detecting causality from nonlinear dynamics with short-term time series. *Scientific Reports*, 4(7464), 2014.

[12] R. May. Simple mathematical models with very complicated dynamics. *Nature*, 26:459–467, 1976.

[13] J.M. McCracken and R.S. Weigel. Convergent cross-mapping and pairwise asymmetric inference. *arXiv:1407.5696v1*, 2014.

[14] D. Monster, R. Fusaroli, K. Tylen, A. Roepstorff, and J.F. Sherson. Causal inference from noisy time-series data - testing the convergent cross-mapping algorithm in the presence of noise and external influence. *Fut. Gen. Comp. Sys.*, 73, 2016.

[15] J. Pearl. *Causality*. Cambridge University Press, Cambridge, 2009.

[16] J. Runge, S. Bathiany, E. Bollt, G. Camps-Valls, D. Coumou, E. Deyle, C. Glymour, and M. et al. Kretschmer. Inferring causation from time series in earth system sciences. *Nat. Comm.*, 10(2553), 2019.

[17] T. Schreiber and A. Schmitz. Improved surrogate data for nonlinearity tests. *Physical Review Letters*, 77(635), 1996.

[18] S. Shimizu, P.O. Hoyer, A. Hyvärinen, and A. Kerminen. A linear non-Gaussian acyclic model for causal discovery. *JMLR*, 7, 2006.

[19] P. Spirtes, C. Glymour, and R. Scheines. *Causation, prediction, and search*. MIT Press, Cambridge, MA, 2nd edition, 2000.

[20] G. Sugihara, R. May, C.-h. Hsieh, E. Deyle, M. Fogarty, and S. Munch. Detecting causality in complex ecosystems. *Science*, 338, 2012.

[21] F. Takens. *Dynamical Systems and Turbulence, Lecture notes in Mathematics 898*, chapter Detecting strange attractors in turbulence. Springer, Berlin, Heidelberg, 1981.

[22] M.J. Vowels, N.C. Camgoz, and R. Bowden. D'ya like DAGs? *arXiv:2103.02582*, 2021.

[23] H. Ye, E. Deyle, L.J. Gilarranz, and G. Sugihara. Distinguishing time-delayed causal interactions using convergent cross mapping. *Scientific Reports*, 5(14750), 2015.

[24] X. Zheng, B. Aragam, P. Ravikumar, and E. P. Xing. DAGs with NO TEARS: Continuous optimization for structure learning. *arXiv:1803.01422v2*, 2018.