|  |
| --- |
| Application Note: Systems Biology  **CauseMap: Fast inference of causality from complex time series** M. Cyrus Maher1\*, Ryan D. Hernandez2,3,4  1Department of Epidemiology and Biostatistics, University of California, San Francisco, 185 Berry Street,  Lobby 5, Suite 5700 San Francisco, CA 94107  2Department of Bioengineering and Therapeutic Sciences,  3Institute for Human Genetics,  4Institute for Quantitative Biosciences (QB3),  University of California, San Francisco, 1700 4th Street San Francisco, California 94158  Received on XXXXX; revised on XXXXX; accepted on XXXXX  Associate Editor: XXXXXXX |

[[1]](#footnote-2)\*abstract

**Summary:** We present an implementation of convergent cross mapping (CCM), a model-free method for establishing causality in complex non-linear systems, even in the presence of unmeasured confounding. This method uses dense time series data to reconstruct high dimensional system dynamics and test for causal relationships among variables of interest.

**Availability and Implementation:** We implement CCM in Julia, a groundbreaking, high-level, high-performance dynamic programming language designed for technical computing. Our software package, CauseMap, is platform-independent and freely available on github (https://github.com/cyrusmaher/CauseMap). It may be installed natively in Julia using Pkg.add(“CauseMap”).

**Contact:** michael.maher@ucsf.edu

**Supplementary Information:** cyrusmaher.github.io/CauseMap

# introduction

Detecting correlations and establishing causal relationships are central goals of Systems Biology. Yet, the intricate and highly non-linear nature of most biological systems poses a significant obstacle to accurate inference. While massive datasets provide the potential for gaining insight into this complexity, methodological improvement will likely play a key role in developing broad, meaningful understanding of biological dynamics. Several metrics have been developed for detecting non-linear correlations at scale, including: spearman correlation (Spearman, 1904), distance correlation (Székely *et al.*, 2007), and mutual information content (Reshef *et al.*, 2011). Causal relationships, on the other hand, can be examined using methods such as time-lagged regression (Granger, 1969), instrumental variables (Bowden and Turkington, 1990), and dynamical Bayesian networks (Wu *et al.*, 2009).

However, these causal methods are heavily model-based, leading them to falter when examining arbitrary non-linear or context-dependent relationships at scale. Furthermore, the approaches mentioned above cannot adequately handle feedback loops, and they frequently generate both false positives and false negatives due to the influence of unmeasured confounders (Vanderweele and Arah, 2011). These are significant liabilities, particularly in fields such as systems biology, where relationships are often complex, interconnected, and embedded within a broad network of unmeasured interactions.

In this application note, we present an implementation of convergent cross mapping (CCM), a model-free approach to detecting dependencies and causality in complex non-linear systems (even in the presence of feedback loops and unmeasured confounding; Sugihara *et al.*, 2012). CCM derives this power from explicitly capturing time-dependent dynamics through a technique known as state-space reconstruction (SSR). SSR has already demonstrated utility for problems as diverse as wildlife management (Dixon *et al.*, 1999; Deyle *et al.*, 2013) and cerebral autoregulation (Heskamp *et al.*, 2013). In practice, this analysis typically requires 25 or more time points, measured with sufficient density to capture system dynamics.

# ALgorithm

CCM relies on SSR, leveraging the fact that time series can be viewed as projections of higher-dimensional system dynamics. The time series of individual variables therefore contain information about the full causal system. As a result, the full causal system (conceptualized as the state space, or manifold) can be reconstructed using individual time series.

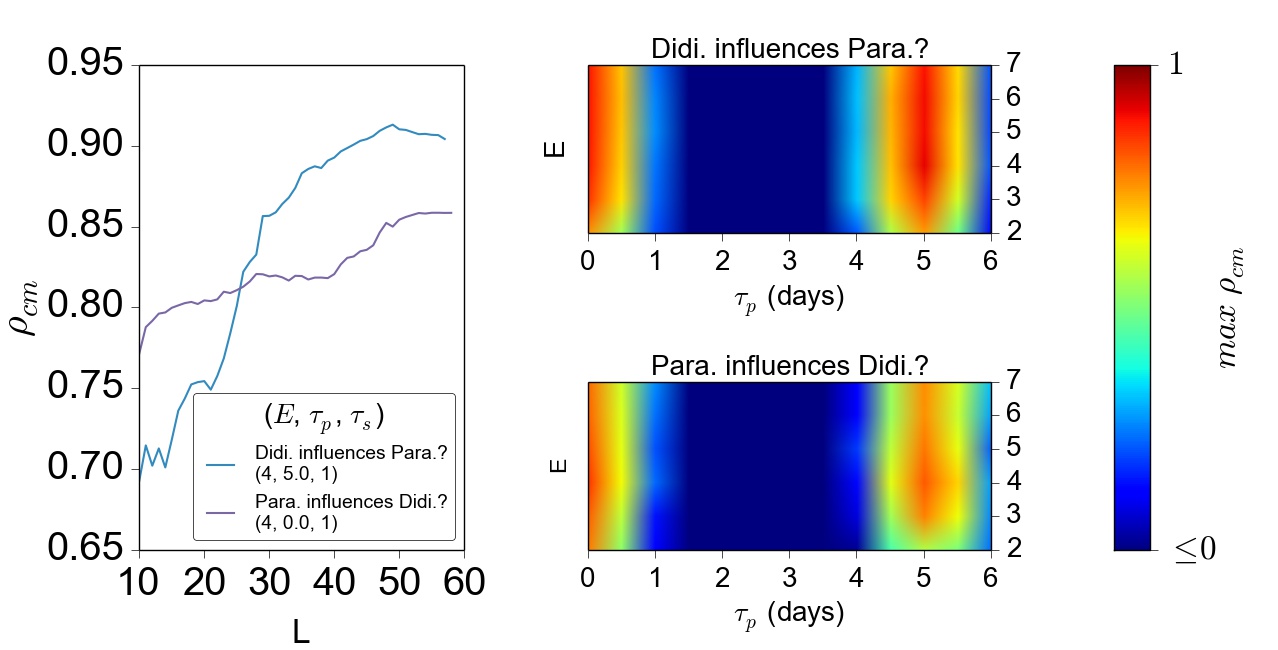
Consider time series of hypothetical variables *X* and *Y.* CCM extends SSR using time-lagged coordinates of each of these variables to produce shadow versions of their respective source manifolds. We will refer to these projection manifolds as *Mx* and *My*. To test whether *X* causes *Y*, CCM leverages the following logic: Because manifold reconstruction preserves the Lyapunov exponents of the original system, if *X* causes *Y*, then time points that are close in *My* should also be close in *Mx*. Since *Mx* is constructed from lags of the observations of *X*,the points that are close in *Mx* will also have similar values in the corresponding time series*.* Therefore, if *X* causes *Y*, then *My* can tell us which observations of *X* should best predict a given point from *X.* Furthermore, predictability should increase with the number of manifold points that are considered.

To test whether *X* causes *Y*, *My* is used to infer the points in *X* that will best predict a given held-out point from X*.* We measure this performance using predictive skill (ρccm). Intuitively, this procedure works as follows: A point is held out from *X*. We then use *My*to infer the points in *Mx* that will be closest to that point of interest*.* Using exponential weights derived from the relative pairwise distances of corresponding points in *My,* we predict the held-out point using other observations from *X.* Finally, ρccm is calculated as the Pearson correlation between observed and predicted points. To examine whether the signal converges as expected for a causal relationship, these steps are repeated using increasing time series length (*L*). Further details are available in the supplementary material of Sugihara *et al.* 2013.

# Implementation

CauseMap is our implementation of CCM in Julia, a groundbreaking high-level, high-performance dynamic programming language designed for technical computing (Bezanson *et al.*, 2012). Via intelligent JIT (just in time) compilation, Julia offers much of the speed of low-level, low-productivity languages like C, while also providing the ease of use and platform independence of much slower high-level languages like Python, R, or Matlab.

Beyond implementation of the CCM algorithm itself, CauseMap offers a number of conveniences and performance enhancements. For CCM, it is particularly important to optimize two tuning parameters: *E* and τp. *E* is related to the assumed dimensionality of the full causal system, while τp denotes the time delay of the causal effect of interest.

CauseMap precomputes all necessary manifolds and pairwise distances using a state-of-the-art, BLAS-based protocol (for benchmarks, see: https://github.com/JuliaStats/Distance.jl). *E* and τp are then optimized by multiple iterations of cyclic coordinate descent (Bertsekas, 1999). Note that while convergence of the cross map signal as a function of the time series length (*L*) is taken as a practical criterion for causality, the dependence of this signal on *E* and τp is also important for understanding the specificity of the observed signal. Therefore, CauseMap provides visualizations for the dependency of the cross map correlation on *L, E,* τp (Fig. 1) using python (www.python.org) and matplotlib (Hunter, 2007). Approximately 300 CCM evaluations were conducted to produce Figure 1. These calculations finished in less than 30 seconds on a single 2.6 GHz processor.

**B]**

**A**

# Conclusions

CauseMap provides a fast, user-friendly implementation of CCM, a powerful new method for exploring dependencies and even establishing causality in complex, highly non-linear datasets with many unobserved variables. We believe that CCM holds a great deal of promise, particularly in Systems Biology.

**Fig. 1.** An example visualization from CauseMap using abundances of *Paramecium aurelia* and *Didinium nasutum* (see S1 for more information on this system). A.) For optimal parameter values, the convergence of the cross-map correlation with library size. B.) The dependence of the maximum cross-map correlation on assumed dimensionality (measured by E) and the time lag of the causal effect (measured by τp). Note that the second maximum at τp=5 corresponds to the principal amplitudes of the *P. aurelia* and *D. nasutum* time series, as determined by fourier transform analysis.

acknowledgements

We would like to thank George Sugihara, Hao Ye, and Ethan Deyle for their invaluable help in understanding the core details of the CCM algorithm, and Lawrence Uricchio and Raul Torres for comments on this manuscript.

*Funding*: This work was partially supported by the National Institutes of Health (grant numbers P60MD006902, UL1RR024131, 1R21HG007233, 1R21CA178706, and 1R01HL117004) and an Alfred P. Sloan Foundation Research Fellowship to R.D.H.M.C.M. was supported by the National Institutes of Health F31 Predoctoral Fellowship (grant number 1 F31 CA180609-01), and a University of California, San Francisco Lloyd M. Kozloﬀ Fellowship.

References

Bertsekas,D. (1999) Nonlinear Programming Athena Scientific; 2nd edition.

Bezanson,J. *et al.* (2012) Julia: A Fast Dynamic Language for Technical Computing.

Bowden,R.J. and Turkington,D.A. (1990) Instrumental Variables Cambridge University Press.

Deyle,E.R. *et al.* (2013) Predicting climate effects on Pacific sardine. *Proc. Natl. Acad. Sci. U. S. A.*, **110**, 6430–5.

Dixon,P.A. *et al.* (1999) Episodic Fluctuations in Larval Supply. *Science (80-. ).*, **283**, 1528–1530.

Granger,C.W.J. (1969) Investigating Causal Relations by Econometric Models and Cross-spectral Methods Title. *Econometrica*, **37**, 424–438.

Heskamp,L. *et al.* (2013) Convergent cross mapping: a promising technique for future cerebral autoregulation estimation. *Cerebrovasc. Dis.*, **35**, 15–16.

Hunter,J.D. (2007) Matplotlib: A 2D graphics environment. *Comput. Sci. Eng.*, **9**, 90–95.

Reshef,D.N. *et al.* (2011) Detecting novel associations in large data sets. *Science*, **334**, 1518–24.

Spearman,C. (1904) The proof and measurement of association between two things. *Am. J. Psychol.*, **15**, 72–101.

Sugihara,G. *et al.* (2012) Detecting causality in complex ecosystems. *Science*, **338**, 496–500.

Székely,G.J. *et al.* (2007) Measuring and Testing Dependence by Correlation of Distances. *Ann. Stat.*, **35**, 2769–2794.

Vanderweele,T.J. and Arah,O.A. (2011) Bias formulas for sensitivity analysis of unmeasured confounding for general outcomes, treatments, and confounders. *Epidemiology*, **22**, 42–52.

Wu,J. *et al.* (2009) Impact of environmental inputs on reverse-engineering approach to network structures. *BMC Syst. Biol.*, **3**, 113.

1. \*To whom correspondence should be addressed. [↑](#footnote-ref-2)