Supplementary Information for Detecting Causality from Nonlinear Dynamics with Short-term Time Series

Huanfei Ma,[1,2] Kazuyuki Aihara, [2*] Luonan Chen, [2,3*]

[1] School of Mathematical Sciences, Soochow University, China
[2] Collaborative Research Center for Innovative Mathematical Modelling, Institute of Industrial Science, The University of Tokyo, Japan
[3] Key Laboratory of Systems Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, China

* To whom correspondence should be addressed;
E-mail: aihara@sat.t.u-tokyo.ac.jp, lnchen@sibs.ac.cn.

Contents

In this Supplementary Material, we provide numerical details for the paper.

1 Preliminary of State Space Reconstruction, Cross Map and Smoothness

In this part, the preliminary knowledge of delayed embedding and cross map will be reviewed.

1.1 State Space Reconstruction

Consider an n-dimensional dynamical system and its attractor M with state variables $\boldsymbol{x}(t) = [x_1(t), x_2(t), \ldots, x_n(t)]$. Assume two scalar time series $x_1(t)$ and $x_2(t)$ are measured from two components x_1 and x_2. With appropriately chosen embedding dimension L and proper positive delay τ, one can obtain time delayed coordinate vectors $\boldsymbol{x_1}(t) = [x_1(t), x_1(t-\tau), \ldots, x_1(t-(L-1)\tau)]^T$ and $\boldsymbol{x_2}(t) = [x_2(t), x_2(t-\tau), \ldots, x_2(t-(L-1)\tau)]^T$ respectively. Generally, if $L > 2d$ where d is the box counting dimension of the attractor M, according to delayed embedding theory, the set of $\boldsymbol{x_1}(t)$ forms the reconstructed attractor M_{x1}, and one can define M_{x2} in an analogous way. Figure S1 illustrates the case of the chaotic Lorenz system with its attractor M and two reconstructed attractors M_x,M_y.

1.2 Nearest Neighbors, Mutual Neighbors and Cross Map

Consider two attractors M_x and M_y reconstructed from two scalar time series $x(t)$ and $y(t)$. For one point $\boldsymbol{y}(t_0) \in M_y$, one can find its k nearest neighbors $\boldsymbol{y}(t_{y_1}), \boldsymbol{y}(t_{y_2}), \ldots, \boldsymbol{y}(t_{y_k})$ with indices $t_{y_1}, t_{y_2}, \ldots, t_{y_k}$, and one can define the mutual neighbors for $\boldsymbol{x}(t_0) \in M_x$ as $\boldsymbol{x}(t_{y_1}), \boldsymbol{x}(t_{y_2}), \ldots, \boldsymbol{x}(t_{y_k})$, as shown in Fig.1(a). In a similar way we can define the mutual neighbors for $\boldsymbol{y}(t_0)$ from the nearest neighbors of $\boldsymbol{x}(t_0)$ as $\boldsymbol{y}(t_{x_1}), \boldsymbol{y}(t_{x_2}), \ldots, \boldsymbol{y}(t_{x_k})$, as shown in Fig.1(b). Such map from nearest neighbors to mutual neighbors is defined as a cross map.

1.3 Convergence of Cross Map and Smoothness

The fundamental idea of the mutual cross map method is that if x causally influences y or x is a driving factor of y (i.e., $x \to y$), then the information of x is encoded in the dynamics of y, and thus two close states on M_y correspond to two close states on M_x. Explicitly, for each point $\boldsymbol{y}(t_0)$ on M_y and its k nearest neighbors $\boldsymbol{y}(t_{y_1}), \boldsymbol{y}(t_{y_2}), \ldots, \boldsymbol{y}(t_{y_k})$ with time indices $t_{y_1}, t_{y_2}, \ldots, t_{y_k}$, the weighted average of the mutual nearest neighbors $\boldsymbol{x}(t_{y_1}), \boldsymbol{x}(t_{y_2}), \cdots, \boldsymbol{x}(t_{y_k})$ can be used as an estimation of $\boldsymbol{x}(t_0)$, see Fig.1(a). Inversely, if y has no influence over x, the dynamics of x is insensitive to the state of y, thus the mutual neighbors, i.e., the images of $\boldsymbol{x}(t_0)$'s nearest neighbors under the cross map $\Phi_{xy} : M_x \to M_y$, are not necessarily the neighbors of $\boldsymbol{y}(t_0)$, as illustrated in Fig.1(b).

As the measured length of time series $x(t)$ and $y(t)$ increases to infinity, the points become dense over the reconstructed attractors M_x and M_y. Thus, the k nearest neighbors of $\boldsymbol{y}(t_0)$ will converge to $\boldsymbol{y}(t_0)$, and if x causally influences y, the mutual neighbors for $\boldsymbol{x}(t_0)$ under cross map Φ_{yx} will also converge to $\boldsymbol{x}(t_0)$, yielding more precise estimation. On the other hand, this is not necessary for the cross map Φ_{xy} if y has no causality over x.

Such convergence actually reflects the smoothness of the cross map, i.e., the cross map $\Phi_{yx} : M_y \to M_x$ maps the neighborhood of $\boldsymbol{y}(t_0)$ to the neighborhood of $\boldsymbol{x}(t_0)$, which actually implies that Φ_{yx} is locally smooth around $\boldsymbol{y}(t_0)$, while for Φ_{xy} this is not necessarily true. When the cross map Φ_{yx} is smooth over neighborhood of every point $\boldsymbol{y} \in M_y$, it implies that Φ_{yx} is globally smooth over M_y, as shown in Fig.1(c). While for the case where y has no causality over x, the cross map Φ_{xy} is not necessarily smooth, as illustrated in Fig.1(d).

2 CMS Algorithm

Here we assume the delayed coordinate vectors $\boldsymbol{x}_i \in \mathbb{R}^L,\ \ i = 1, 2, \ldots, n$ and $\boldsymbol{y}_i \in \mathbb{R}^L,\ \ i = 1, 2, \ldots, n$ are obtained from the measured scalar time series $x(t)$ and $y(t)$ in the preprocessing. The algorithm to detect the causality R_{xy} from x to y can be formulated as follows.

Algorithm1 Cross Map Smoothness with RBF network Algorithm

Given two data set $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n \in \mathbb{R}^L$ and $\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_n \in \mathbb{R}^L$, let $S_i = \{1, 2, \ldots, n\} \backslash i$ be the leave-one-out index set.

1. *For* $i = 1, 2, \ldots, n$
   –Train a Radial Basis Function (RBF) network $\mathcal{N}_i$ based on the condition $\mathcal{N}_i(\boldsymbol{y}_j) = \boldsymbol{x}_j, j \in S_i$
   –Calculate $\hat{\boldsymbol{x}}_i = \mathcal{N}_i(\boldsymbol{y}_i)$
   –Calculate the error as $\epsilon_i = \|\boldsymbol{x}_i - \hat{\boldsymbol{x}}_i\|$
   End For
2. Normalize the error as $\Delta = \dfrac{\langle \boldsymbol{\epsilon} \rangle_{rms}}{\langle \|\boldsymbol{x} - \bar{\boldsymbol{x}}\| \rangle_{rms}}$ and calculate the causality index as

$$R_{xy} = \frac{1}{\exp(\Delta/\sigma)},$$

where $\bar{\boldsymbol{x}}$ is the mean vector of $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$.

Here, $\|\cdot\|$ is the vector norm, and $\langle \cdot \rangle_{rms}$ is the root mean square of the set. σ is a positive constant to normalize the index so that the long tail can be cut into zero, and empirically we choose $\sigma = 5$ throughout the paper. Based on the fact that CMS index (R_{xy}) provides only the strength of the causative effectiveness from x to y with a positive value normalized in $[0, 1]$, the positive or negative regulation between the two variables should be further measured by using other criteria, e.g., correlation analysis.

3 2 Species Logistic Model

The 2 species Logistic model is formulated as

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

where $r_x = 3.7$ and $r_y = 3.8$ are self-regulation parameters; γ_{yx} and γ_{xy} are two coupling constants. In the unidirectional case, we assume the coupling constants taking values $\gamma_{yx} = 0.32, \gamma_{xy} = 0$. We use $\gamma_{yx} = 0.1$ and $\gamma_{xy} = 0.02$ for the bidirectional case. In the causality switching setting, we assume the coupling-constant pair switches values between two sets $\{\gamma_{yx} = 0.32, \gamma_{xy} = 0\}$ and $\{\gamma_{yx} = 0, \gamma_{xy} = 0.32\}$ at random intervals, indicating that the unidirectional causal relation between the two variables changes the direction from time to time. Here, the delay embedding dimension L is chosen as $L = 2$. Random numbers distributed uniformly in $[0, 1]$ are used as system initial conditions and 20 time points after transient dynamics are measured as time series.

4 Complex system with 3 nodes

A complex system consisting of 3 nodes can be described as

$$Y_j(t+1) = Y_j(t) \left(\gamma_{jj} - \sum_{i=1,2,3} \gamma_{ji} Y_i(t) \right), j = 1, 2, 3, \tag{2}$$

where γ_{ji} are coupling parameters. In the fan-out pattern, the parameters are set as $\gamma_{11} = 4, \gamma_{22} = 3.1, \gamma_{33} = 2.12, \gamma_{21} = 0.21$, and $\gamma_{31} = -0.636$ while other parameters are set to zero. In the fan-in pattern, the parameters values are $\gamma_{11} = 4, \gamma_{22} = 3.6, \gamma_{33} = 2.12, \gamma_{31} = 0.636$, and $\gamma_{32} = -0.636$, and all the other parameters are set to zero. Random numbers distributed uniformly in $[0, 1]$ are used as system initial conditions and 20 time points are measured after transient dynamics. The delay embedding dimension L is chosen as $L = 2$.

5 Coupled Rössler-Lorenz system

We follow the setting in [1] so that the driving Rössler system and the response Lorenz system are given by the following equations:

$$\begin{aligned}
&\text{Rössler system X:}\\
&\dot{x}_1 = -\alpha(x_2 + x_3),\\
&\dot{x}_2 = \alpha(x_1 + 0.2x_2),\\
&\dot{x}_3 = \alpha(0.2 + x_3(x_1 - 5.7)),\\
&\text{Lorenz system Y:}\\
&\dot{y}_1 = 10(-y_1 + y_2),\\
&\dot{y}_2 = 28y_1 - y_2 - y_1y_3 + Cx_2^2,\\
&\dot{y}_3 = y_1y_2 - 8/3y_3,
\end{aligned} \tag{3}$$

where $\alpha = 6$ is a timescale constance and $C = 2$ is the strength of the unidirectional coupling. The initial values of the system is randomly chosen and we assume that the sampling interval is $\Delta t = 0.2$. After the transient dynamics, both the driving and response systems come to chaotic attractors. We use a time series with 50 time points to detect the causality between the two systems. Here it is noted that for continuous systems, the sampling interval should be moderately large and too small sampling intervals will bring short term data too close to each other on the attractor and no geometric information can be extracted.

6 Gene regulatory network

The synthetic gene expression data sets are generated by the synthetic network generator, SynTReN [2], which produces simulated gene expression data of the associated mRNA concentrations for each gene, based on Michaelis-Menten and Hill kinetics. The topology of the regulatory network and the interactions (activating, repressing, or dual) are from the real network of *E.coli* or *S. cerevisiae*, and various kinds of noise can be considered to approximate experimental expression measurements.

We generate GRN data with a subnetwork of 50 genes for *E.coli* and subnetworks of 100 and 150 genes for *S. cerevisiae*, using the cluster addition strategy to randomly select a connected subgraph of the whole network. The cluster addition strategy randomly chooses and adds a node into the graph together with all its neighbors, which efficiently extracts a subnetwork that approximates well the topology of the source network. We generate GRN data with 10 time points for each gene, uniformly sampled in time. For the noisy data, we consider three levels of noise intensities, namely, 0.1, 0.2, and 0.3. For each level of noise intensity, we consider three kinds of noise simultaneously, namely, the biological noise, the experimental noise and noise on correlated inputs. Moreover, for the noisy data, synthetic gene expression data with 6 independent noisy technical replicates have been generated from the subnetwork, and we use the average values of the technical replicates to approximate the measurements from real

experiments. The embedding dimension L in the CMS algorithm is chosen as $L = 2$ according to the FNN test.

7 SCN data

The gene expression profiles was measured with Affimetrix microarray (Genechip Rat Genome 230 2.0), and the 16 time points were measured after the drug perturbation in the 19th hour. The circadian oscillation of *Per1* expression was immediately phase-shifted by forskolin, resembling the abrupt light-induced phase shift that takes place in SCN. The measuring time points are from 22.5 hour to 90 hour with an even time interval of 4.5 hours. The embedding dimension L in the CMS algorithm is chosen as $L = 2$.

8 Lotka-Volterra system

The chaotic Lotka-Volterra system [3] is formulated as

$$\begin{aligned} x_1(t+1) &= x_1(t)(2 - x_1(t) + 0.675x_2(t) - 0.5x_3(t)), \\ x_2(t+1) &= x_2(t)(2 - 0.5x_1(t) - x_2(t) + 0.675x_3(t)), \\ x_3(t+1) &= x_3(t)(2 + 0.675x_1(t) - 0.5x_2(t) - x_3(t)). \end{aligned} \tag{4}$$

9 Training Radial Basis Function (RBF) network

A Radial Basis Function (RBF) network can be trained using training data and the error can be calculated using validation data set [4]. Generally, a Radial Basis Function (RBF) network $\mathcal{N}$ can be expresses as

$$\mathcal{N}(\boldsymbol{x}) = \sum_i a_i \rho_i(\boldsymbol{x}, c_i),$$

where each ρ_i is a radial basis function with its center located at c_i, and each a_i is a weight. Here, we set each radial basis function in the following Gaussian form:

$$\rho_i(\boldsymbol{x}, c_i) = \exp(\frac{-\|\boldsymbol{x} - c_i\|^2}{2\sigma_i^2}),$$

where $\|\boldsymbol{x} - c_i\|$ is the L_2-norm in $\mathbb{R}^n$, c_i denotes the center for ρ_i, and σ_i denotes its width.

In Algorithm 1, the RBF network is trained using conditions $\mathcal{N}_i(\boldsymbol{x}_j) = y_{ij}, j \in S$, where y_{ij} is the ith component of the vector $\boldsymbol{y}_j$. Here we note that for short time series data, the size of training set S is usually smaller than the size of the network, i.e., there are more coefficients to be determined than the number of independent equations. Therefore, the problem will be generally catalogued into an underdetermined problem.

Even though the underdetermined problem raised above would be a difficult problem, in many situations, especially with the proper choice of radial basis functions, the number of zero or nearly zero coefficients in the network can be of majority so that the coefficient vector is effectively sparse. Therefore, the recently developed idea of compressive sensing [5] provides a practical technique to solve the problem of determining the RBF network $\mathcal{N}$.

Generally, the problem of compressive sensing can be described as to reconstruct a sparse vector $\boldsymbol{x}_0 \in \mathbb{R}^n$ with m linear measurements of the form $\boldsymbol{y} = A\boldsymbol{x}_0$, where $A \in \mathbb{R}^{m\times n}$ are m known test signals and $\boldsymbol{y} \in \mathbb{R}^m$ are the measurements. According to the compressive sensing theory, when $m \ll n$, one can still actually recover x_0 exactly by solving the convex program $\min_{\boldsymbol{x}} \|\boldsymbol{x}\|_{l_1}$ subject to $A\boldsymbol{x} = \boldsymbol{y}$, where $\|\boldsymbol{x}\|_{l_1} = \sum_{i=1}^{n} |x_i|$ with the L_1 norm of vector $\boldsymbol{x}$, provided that the matrix $A \in \mathbb{R}^{m\times n}$ obeys a uniform uncertainty principle [5].

Thus the solution of the Algorithm 1 can be casted into the compressive sensing problem. First, let $\hat{\boldsymbol{x}}_j = [\rho_1(\boldsymbol{x}_j), \rho_2(\boldsymbol{x}_j), \ldots, \rho_k(\boldsymbol{x}_j)]^T$ where ρ_l is the lth radial basis function in the network. Therefore, the solution in every iteration can be expressed as to solve

$$X\boldsymbol{\alpha}_i = \boldsymbol{y},$$

where $X = [\hat{\boldsymbol{x}}_1, \hat{\boldsymbol{x}}_2, \ldots, \hat{\boldsymbol{x}}_m]^T$, $\boldsymbol{y} = [y_{i1}, y_{i2}, \ldots, y_{im}]^T$, and $\boldsymbol{\alpha}_i = [a_1, a_2, \ldots, a_k]^T$ is a sparse vector to be determined. Moreover, to ensure the uniform uncertainty principle, we normalize X by dividing elements in each column by the L_2 norms of the column: $X_{ij} = X_{ij}/L_2(j)$ with $L_2(j) = \sqrt{\sum_{i=1}^{m}(X_{ij})^2}$. Thus, from the above setting, we can cast the solution of every iteration into a standard compressive sensing problem, and therefore, $\boldsymbol{\alpha}_i$ can be determined via some standard compressive sensing algorithm [6] with re-normalization $\boldsymbol{\alpha}_i = \boldsymbol{\alpha}_i/L_2$.

10 Significance Test

We carry out the random permutation tests [7] as significance tests in our work. The basic assumption is that a randomly rearrangement of time series x can rarely have causal relation with time series y. Thus, with a large number of shuffling of time series x according to random permutations, one can get the empirical distribution of the causality index R over these shuffled x and original y. Then we can finally get the quantile value for the given p-value. In this paper, for the system in Fig. 4, we run 1000 independent permutations uniformly at random, shuffle the time points according to the permutations, and run CMS on the shuffled data. With the empirical distribution shown in Fig.S2, we estimate the threshold as $\xi = 0.001$ at a significance level $p < 0.05$, i.e., we treat a causality index below 0.001 as zero. Therefore, we can determine the threshold with this significance test.

Moreover, to validate the significance test result, we also carry out an additional bootstrap test. The premise of bootstrap resampling is that a single observation can stand in for a distribution if it is resampled with replacement [8]. Using 1000 independent resampling, we also come to the conclusion that $\xi = 0.001$ has the p-value $p < 0.05$.

11 Embedding Dimension

How to determine the embedding dimension L and positive delay τ is an important topic in the state space reconstruction process, and several criteria have been proposed to the time series [9]. In this paper, we use false nearest neighbor (FNN) criterion to determine the embedding dimension [9, 10], and due to the short length of the time series, we simply choose delay τ as one lag in the time series.

Here we stress that when applied to very short-term data, FNN may not yield the optimal embedding dimension. This is mainly due to the fact that FNN also depends on finding nearest neighbors. However, we find that even if the embedding dimension L determined by FNN is not optimal, the result of causality detection will not be significantly influenced. Actually, the condition $L > 2d$ for delayed embedding reconstruction is required so that the map is one to one. Here L is the dimension of the reconstructed attractor M and d is the box-counting dimension of the original attractor. This condition is used to make sure there is no self-intersection on the reconstructed attractor [11], which is sufficient but not necessary. Moreover, even when this rule is not satisfied, the map can still be guaranteed to be an embedding outside a subset A of dimension at most $2d - L$ (see Theorem 2.10 in [11]). Therefore, in our work, even if the condition $L > 2d$ is not satisfied, only points located on the exceptional subset will yield false results, while all the points outside that specific subset still give correct results. Then note the fact that the measured points are effectively sparse over the reconstructed attractor (due to the short-term property), the probability that one measured point locates in the subset A is very small (e.g., located on an n_1-dimensional subset in a n_2-dimensional space where $n_1 << n_2$). Furthermore, we use leave-one-out strategy to train neural networks and finally use the averaged error to calculate the index. Thus a small portion of deviation will not essentially change the averaged result. Therefore, we can come to the conclusion that even when the condition $L > 2d$ is not strictly satisfied, our index will not essentially be affected. To this end, we consider the benchmark system (1) with embedding dimension $L = 2, 3, 4, 5$, and the result is shown in Fig.S3. Here the attractor of system (1) definitely has a box counting dimension larger than 1 and theoretically the optimal embedding dimension should be at least 3, but we see that for different embedding dimensions from 2 to 5, the result does not change significantly.

On the other hand, practical time series always have length limitation, and therefore the tests like FNN will not always work. Thus, in many works, particularly when applied to real data, low embedding dimensions (say, larger than 1 but no more than 4) are always adopted [1, 12].

References

[1] Quyen, M. L. V., Martinerie, J., Adam, C. & Varela, F. J. Nonlinear analyses of interictal EEG map the brain interdependences in human focal epilepsy. *Physica D* **127**, 250-266 (1999).

[2] Van den Bulcke, T. et al. SynTReN: a generator of synthetic gene expression data for design and analysis of structure learning algorithms. *BMC bioinformatics* **7**, 43 (2006).

[3] Bischi G I & Tramontana F. Three-dimensional discrete-time LotkaCVolterra models with an application to industrial clusters. *Commun. Nonlinear Sci. Numer. Simul*, 15(10),3000-3014 (2010).

[4] Ancona, N., Marinazzo, D. & Stramaglia, S. Radial basis function approach to nonlinear Granger causality of time series., *Phys. Rev. E* **70**, 056221 (2004).

[5] Candes, E., Romberg, J. & Tao, T. Stable signal recovery from incomplete and inaccurate measurements. *Commun. Pur. Appl. Math* **59**, 1207 (2006).

[6] http://users.ece.gatech.edu/∼justin/l1magic/

[7] Nichols, T. E. & Holmes, A. P. Nonparametric permutation tests for functional neuroimaging: A primer with examples. *Hum. Brain Mapp.* **15** 1-25. (2002).

[8] Efron, B. & Tibshirani, R. J. Introduction to the Bootstrap. Monographs on Statistics and Applied Probability, Chapman & Hall.(1994).

[9] Kantz, H. & Schreiber, T. Nonlinear time series analysis, Cambridge university press.(2004).

[10] Kennel, M. B., et al. Determining embedding dimension for phase-space reconstruction using a geometrical construction. *Phys. Rev. A* **45**, 3403-3411. (1992).

[11] Sauer, T., Yorke, J. A. & Casdagli, M. Embedology, *Journal of statistical Physics* **65**, 579-616.(1991).

[12] Sugihara, G. et al. Detecting causality in complex ecosystems., *Science* **338**, 496-500 (2012).

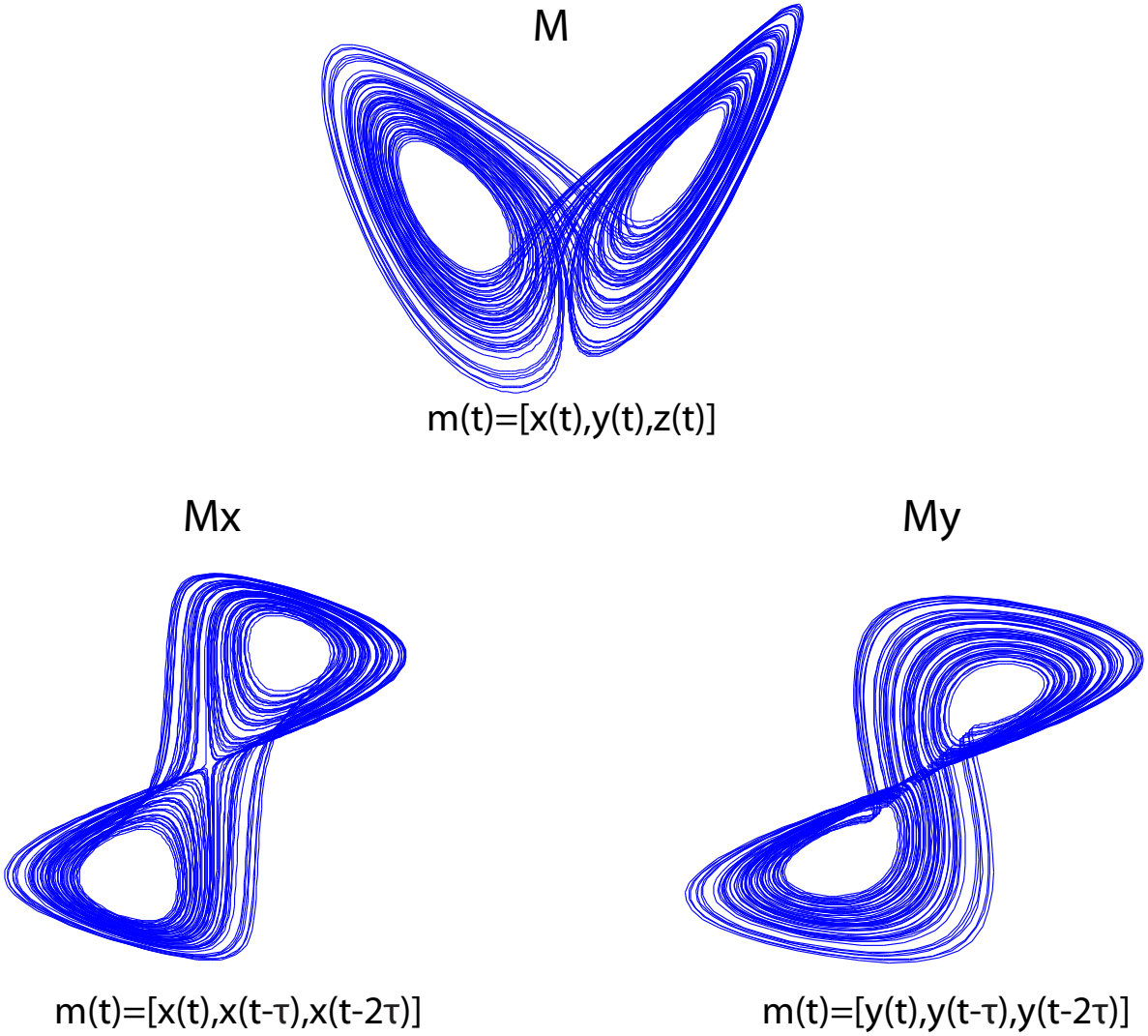

Figure S 1: Illustration of State Space Reconstruction using delayed embedding technique. This example is based on the chaotic Lorenz system. The attractor M is constructed using x, y, z and the reconstructed attractors M_x and M_y are reconstructed using the delayed coordinates of x and y respectively.

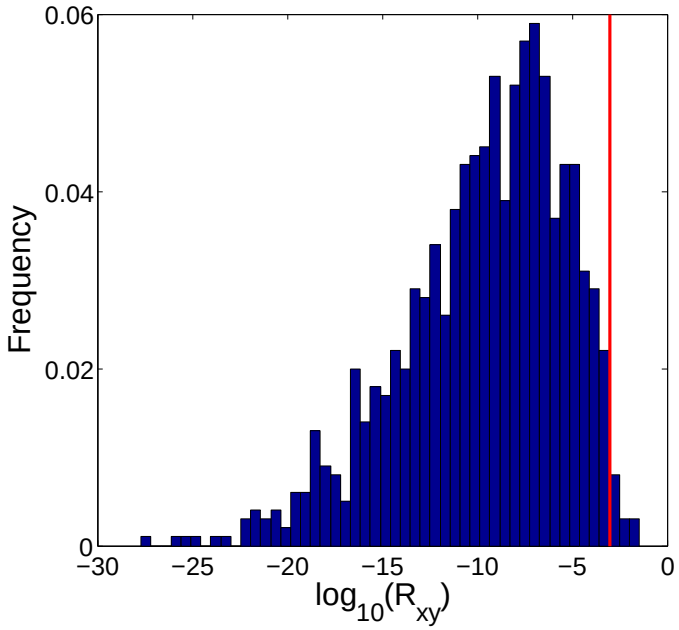

Figure S 2: Empirical distribution of the random shuffle test. Here the area on the right side of threshold is 0.013.

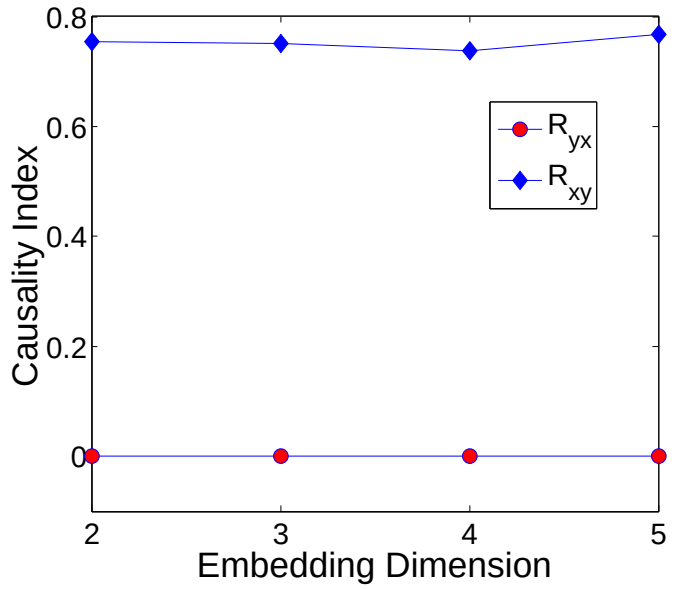

Figure S 3: Detecting causality for system (1) with unidirectional coupling using different embedding dimensions.