A test of Granger non-causality based on nonparametric conditional independence

Sohan Seth and José C. Príncipe
Department of Electrical and Computer Engineering
University of Florida
Gainesville, FL 32608
sohan,principe@cnel.ufl.edu

Abstract—In this paper we describe a test of Granger non-causality from the perspective of a new measure of nonparametric conditional independence. We apply the proposed test on two synthetic nonlinear problems where linear Granger causality fails and show that the proposed method is able to derive the true causal connectivity effectively.

Keywords—Granger causality, conditional independence, kernel methods, regression

I. INTRODUCTION

Given stochastic processes \( \{X_t\} \) and \( \{Y_t\}, \{X_t\} \) is said to cause \( \{Y_t\} \) i.e. \( \{X_t\} \rightarrow \{Y_t\} \), in the sense of Granger, if the past values \([X_{t-1}, X_{t-2}, \ldots]\) of \( \{X_t\} \) contain additional information about the present value \( Y_t \) of \( \{Y_t\} \) that is not contained in past values \([Y_{t-1}, Y_{t-2}, \ldots]\) of \( \{Y_t\} \) alone. Separating the driving process e.g. \( \{X_t\} \) from the process being driven e.g. \( \{Y_t\} \) is of utmost importance in many applications such as machine learning [3], analysis of dynamical network [8], bioinformatics [14], neuroscience [10] and econometrics [13], [1]. The framework for causal inference in terms of predictability was first conceived by Wiener and, later, implemented by Granger with linear autoregressive models [6]. (Linear) Granger causality has drawn attention ever since and has been widely used due to its simplicity [14]. However, the drawback of this approach is its underlying assumption of linearity of stochastic processes and, therefore, this approach fails when the underlying processes are nonlinear in nature [2]. The concept of Granger causality has been generalized by several authors to circumvent this problem e.g. using nonlinear models such as kernel based approach [11], [8], and using piecewise linear models [2].

A different line of research employs test of conditional independence as a tool to detect noncausality, exploiting the fact that under non-causality i.e. \( \{X_t\} \not\rightarrow \{Y_t\} \), the present value \( Y_t \) of \( \{Y_t\} \) is conditionally independent of the past values \([X_{t-1}, X_{t-2}, \ldots]\) of \( \{X_t\} \) given past values \([Y_{t-1}, Y_{t-2}, \ldots]\) of \( \{Y_t\} \) [4]. This idea generalizes the concept of prediction and takes into account the entire conditional probability rather than just the conditional expectation. Moreover, this approach is model free and, therefore, applicable to a wide range of problems, especially for problems where linear Granger causality is inadequate [13]. However, estimating conditional independence, on its own, is a difficult problem and is still an open area of research. Recently several methods based on the estimation of the conditional probability density function or its variations [4], [13], [1] and other kernel based approach [5] have been proposed. However, the density function based approaches heavily rely on the choice of the kernel and a computationally expensive cross validation based approach needs to be employed for parameter setting. In this paper we propose a new measure of conditional independence that avoids estimation of conditional density function and uses the conditional distribution function instead [12]. We view the conditional distribution function as the expected value of an identity function and consider a regression type approach to estimate it. We prefer this approach since kernel based regression is a well studied problem and it has been shown to be more robust to the choice of kernel [7].

II. METHOD

Given random variables \((X, Y, Z), (X, Y)\) is said to be conditionally independent given \(Z\) if and only if \(P(X < x| Y = y, Z = z) = P(X < x| Z = z)\) for all values \((x, y, z)\). A measure of conditional independence is characterized by the property that it attains zero value if and only if the random variables are conditionally independent. Therefore, the following quantity,

\[
\mathcal{M}^2_{CI} = \int (P(u|v, w) - P(u|w))^2 dF_X(u)dF_YZ(v, w)
\]

is a measure of conditional independence [12].

Let \(g_u(v, w)\) and \(h_u(w)\) be estimates of \(P(X < u| Y = v, Z = w)\) and \(P(X < u| Z = w)\) respectively. Then an estimator of \(\mathcal{M}_{CI}\) is given by,

\[
\hat{\mathcal{M}}^2_{CI} = \int (g_u(v, w) - h_u(w))^2 dF_X(u)dF_YZ(v, w).
\]

Using the triangle inequality, it can be easily shown that the absolute difference between the actual and estimated \(\mathcal{M}_{CI}\) i.e. \(|\mathcal{M}_{CI} - \hat{\mathcal{M}}_{CI}|\) is upper bounded by the \(L_2\) distance between the actual and estimated conditional distribution functions. Therefore, we choose the estimators, \(g_u(v, w)\) and...
is a set of appropriate basis functions and \( \{ \alpha^u_i \}_{i=1}^m \) is a set of coefficients and minimize the cost function

\[
\mathcal{J}^2(u) = \int (P(X < u | Z = w) - h_u(w))^2 dF_Z(w)
\]

for all \( u \) [12]. After some simple algebraic manipulation and using the property that \( P(X < u) = EI(X < u) \) where \( I \) is the identity function, it can be shown that the minimizer of this problem is \( \alpha^u_i = (\Psi^T \Psi + \lambda_u I)^{-1} \Psi^T u \), where, given realizations \( \{(x_i, y_i)\}_{i=1}^m \) of \( (X, Y) \), \( \Psi_{ij} = \psi_j(z_i) \) is a matrix of basis functions evaluated at the realizations, \( I \) is the identity matrix, \( I_u \) and \( \alpha^u_i \) are vectors of \( I(x_i < u) \)'s and \( \alpha^u_i \)'s, respectively, and \( \lambda_u \) is a Tikhonov regularization parameter. Replacing this solution in the definition of \( \mathcal{M}_{CI} \) and using strong law of large numbers it can be shown that \( \mathcal{M}^2_{CI} \approx \frac{1}{m} \|I(u - \Psi I)\|^2 \) where \( I = A(A^T A + \lambda A)^{-1} A^T \), \( [i, j] \) is \( i \)-th row and \( j \)-th column of matrix \( A \).

Since we are estimating \( P(X < u | Y = v) \) and \( P(X < u | Y = v, Z = w) \) by minimizing the \( L_2 \) distance between these functions and the corresponding estimators, we need to ensure that the basis functions are rich enough to make this distance arbitrarily small. Therefore, we choose \( \psi_j(\cdot) \) to be positive definite kernels, often a Gaussian. We choose this model since it can be shown that under the condition \( n \to \infty \), the span of these functions can represent any function \( P(X < u | Y = v) \) (and \( P(X < u | Y = v, Z = w) \)) with arbitrary accuracy in the \( L_2 \) distance sense. Using these basis functions the matrices \( \Phi \) and \( \Psi \) become Gram matrices \( K_{Y,Y}[\cdot, \cdot] \) and \( K_{Z,Z} \), respectively, where \( [K_{U,V}]_{ij} = \kappa(u_i, v_j) \) where \( \kappa \) is a positive definite kernel.

The proposed method involves two free parameters namely the kernel and the regularization value. The kernel can further be considered as two free parameters i.e. the type of the kernel and the size of the kernel. We choose to work with Gaussian kernel since it is universal [9]. To choose an appropriate kernel size we follow a nearest neighbor based approach. To be specific, we select the kernel size to be the average distance between a sample and its \( k \)-th nearest neighbor. Next, we choose the regularization parameter to be a function of the sample size such that it drops at a certain rate as sample size is increased i.e. we select \( \lambda = n^{-\gamma} \) where \( \gamma > 0 \).

### III. Simulation

In this section we apply the proposed method to test Granger non-causality on two synthetic datasets. We focus on the effect of sample size and choice of free parameters on the accuracy of causal inference. In the following experiments, we first generate \( n \) samples from the desired time series. Then we perform a permutation test to decide the causal connectivity of the network. For the permutation test we use 100 resamplings and we set the size of the test to 0.1. We repeat this process a 100 times and record the percentage of times a particular causal connection has been established. Due to space limitation, we provide results for two different kernel sizes i.e. we set \( k = 25 \) and \( k = 50 \), two different regularizations, i.e. \( \lambda = \lambda = n^{-2} \) and two different sample sizes i.e., \( n = 50 \) and \( n = 100 \), respectively.

### A. Bivariate time series

We first discuss the following bivariate time series with nonlinear coupling as described in [2].

\[
x_1(t) = 3.4x_1(t-1)(1-x_1^2(t-1))e^{-x_1^2(t-1)} + 0.8x_2(t-2) + \epsilon_1
\]

\[
x_2(t) = 3.4x_2(t-1)(1-x_2^2(t-1))e^{-x_2^2(t-1)} + 0.5x_2(t-2) + cx_1^3(t-2) + \epsilon_2
\]

where \( 0 \leq c \leq 1 \) is the coupling strength. In this particular example \( x_1 \) causes \( x_2 \) for \( c > 0 \) but not the other way around. We consider this particular example since linear Granger causality fails to detect the true causation. Chen et al. used a modification of linear Granger causality exploiting local linearity to successfully derive the true cause and effect relationship [2]. However, this approach requires a large number (\( \sim 10^3 \) - \( 10^4 \)) of samples. We test whether \( x_1(t) \) is conditionally independent of \( [x_2(t-1), x_2(t-2)] \) given \( [x_1(t-1), x_1(t-2)] \) to determine if \( x_2 \not\rightarrow x_1 \). Figure 1 shows the result of the proposed approach for different sample sizes and parameter settings. We observe that the performance of the algorithm improves with the number of samples. However, even with 50 samples we have been able to correctly detect the true connections with almost 80% accuracy. Note that, the variation due to the different parameter settings is also not overwhelming. For both sample sizes we observe an almost similar pattern for different parameter settings. However, for \( n = 50 \) we observe slightly better performance for \( k = 25, \lambda = n^{-2} \) whereas for \( n = 100 \) we observe slightly worse performance \( k = 50, \lambda = n^{-1} \).
B. Multivariate time series

Next, we consider a multivariate time series model described by the following equation,

\[
x_1(t + 1) = \left(1 + \frac{x_1(t)}{1 + x_1^2(t)} \right) \sin x_2(t)
\]

\[
x_2(t + 1) = x_1(t) \exp \left(\frac{-x_2^2(t) + x_3^2(t)}{8}\right) + x_2(t) \cos x_2(t)
\]

\[
x_3(t + 1) = \frac{x_1(t)}{1 + 0.5 \sin x_2(t)} + \frac{x_2(t)}{1 + 0.5 \sin x_1(t)}
\]

where \(x_4(t)\) is white Gaussian noise of variance 1. In this time series we have 4 individual elements that interact with each other. Here \(x_1\) causes both \(x_2\) and \(x_3\), \(x_2\) causes \(x_1\) and \(x_3\), and \(x_4\) causes \(x_2, x_3\) does not influence anybody whereas \(x_4\) is not influenced by anybody. We select this particular example since linear Granger causality fails to detect the true causal connectivity. Our experiment shows that linear Granger causality detects two spurious connections in terms of \(x_3\) causing both \(x_1\) and \(x_2\). We test whether \(x_3\) is conditionally independent of \([x_j (t - 1)]\) given \([x_1 (t - 1), x_2 (t - 1), x_1 (t - 1)]\) to determine if \(x_j \perp x_i\). Figure 2 depicts the performance of the proposed method. We again observe that the performance of the algorithm improves with the number of samples. However, with only 50 samples we have been able to identify the true connectivity with around 80% accuracy. We also observe that a larger kernel widths works better in this particular example. We expect this since a broader kernel is able to estimate the conditional distribution function better in higher dimensions. For \(n = 100\) we observe that the causal network has been successfully recovered with almost 100% accuracy.

IV. Conclusion

In this paper we discuss some preliminary results of a test of Granger non-causality using a newly introduced measure of conditional independence. We discuss the properties of the proposed measure, propose some simple rules to select the free parameters and apply it on two synthetic dataset to demonstrate its performance. We observe that the proposed method can successfully recover the causal connectivity from nonlinearly coupled system where the linear Granger causality approach fails. We observe that the algorithm can infer the correct causality with substantially small sample size and it is significantly robust to the choice of free parameters.

However, in the simulations described we have used the appropriate number of lags since we know the ground truth, whereas, in practical problems the true number of lags are unknown and it must be chosen via appropriate statistical test. The proposed method can also be used to detect the appropriate lag [12]. However, we leave combining these two approaches to build a more realistic test of non-causality as future work.

REFERENCES


Figure 1. The figure depicts the estimated connections for different sample sizes, kernel widths and regularization values for the time series described in III-A.

Figure 2. The figure depicts the estimated connections for different sample sizes, kernel widths and regularization values for the time series described in III-B.