A Wiener Causality Defined by Relative Entropy

Junya Chen[1], Jianfeng Feng[2], and Wenlian Lu[2](✉)

[1] School of Mathematical Sciences, Fudan University,
Shanghai, China
junyachen15@fudan.edu.cn

[2] Institute of Science and Technology for Brain-Inspired Intelligence,
Fudan University, No. 220 Handan Road, Shanghai, China
jianfeng64@gmail.com, wenlian@fudan.edu.cn

Abstract. In this paper, we propose a novel definition of Wiener causality to describe the intervene between time series, based on relative entropy. In comparison to the classic Granger causality, by which the interdependence of the statistic moments beside the second moments are concerned, this definition of causality theoretically takes all statistic aspects into considerations. Furthermore under the Gaussian assumption, not only the intervenes between the co-variances but also those between the means are involved in the causality. This provides an integrated description of statistic causal intervene. Additionally, our implementation also requires minimum assumption on data, which allows one to easily combine modern predictive model with causality inference. We demonstrate that REC outperform the standard causality method on a series of simulations under various conditions.

Keywords: Granger causality · Time series
Relative entropy causality · Transfer entropy

1 Introduction

Causality, which describes the inferring interactions from data, has been of long-term interest to both statisticians and scientists in diverse fields [1]. In general, causality is defined by the models containing families of possible distributions of the variables observed and appropriate mathematical descriptions of causal structures in the data [2]. One of the most popular approaches, introduced originally by Wiener [3], has a philosophic idea as follows: Given sets of interdependent variables x and y, it is said that y causes x if, in certain statistical sense, including possible distribution and model, y assists in predicting the future of x, in comparison to the scenario that x already predicts its own future.

This idea was formalised in terms of multivariate auto-regression (MVAR) by Granger, and importantly, in term of identifying a causality interaction, the measurement and statistic inference was proposed based on MVAR by comparing the variances of the residual errors with and without considering y in

L. Cheng et al. (Eds.): ICONIP 2018, LNCS 11302, pp. 124–133, 2018.
https://doi.org/10.1007/978-3-030-04179-3_11

the prediction of x [4]. Information theory provided another important quantity, transfer entropy [5], to measure directed information transfer between joint processes in complex systems [6] that is theoretically model-free. Even though transfer entropy is not identical to identifying a physically instantiated causal interaction in a system in terms of prediction but of resolution of uncertainty: the difference of the conditional entropy of x with and without considering y, it has, however, been proved to be equivalent to Wiener-Granger causality in MVAR with Gaussian assumptions [7]. Thus, both definitions can be physically regarded the statistic aspect identified by comparing the model with and without considering the intervene of y and have provided deep insights into the directed connectivity of systems in a variety of fields [8], for example, in neuroscience [9,10].

However, the differences of statistic aspects, such as the conditional co-variances in Granger causality and conditional entropy in transfer entropy, are insufficient to depict the distance in statistic sense, between with and without considering intervene from y. Two statistic distributions may not coincide if they share the same co-variance and/or entropy. Hence, there might be the case that the measurements of causality cannot deliver the exact inferring interactions between processes. In terms of prediction, the intervene from the first moment, the mean, of y should not be excluded, according to Wiener's original definition of causality, unlike what Granger did. In other words, removing the mean from the time series may cause ignorance of the causal intervenes between them.

In this paper, we propose the relative entropy causality (REC), based on Wiener's original definition of causality, including all statistic aspects. Instead of the linear and stationary assumption in Granger casusality, in REC we extend it and fit it into the nonlinear and nonstationary cases. Moreover, it turns out to be very convenient in analyzing causality combined with the nonlinear predictive models like XGBoost [11]. We show that REC outperforms standard methods like transfer entropy and Granger causality under different conditions of simulation experiments.

2 Notations

2.1 Residual Errors

The history of a time series x up to p lags is denoted by $x_t^p = [x_{t-1}, \cdots, x_{t-p}]$, also x^p for simplicity if it is stationary.

We denote the residual errors in predicting x by

$$\epsilon = x - \tilde{x} \tag{1}$$

where $\tilde{x}$ is the predictor of x under certain statistic model based on $x_t^p = [x_{t-1}, \cdots, x_{t-p}]$, when as well as possibly $y_t^q = [y_{t-1}, \cdots, y_{t-q}]$.

In this paper, we will consider two predictors: $\tilde{x} = \phi(x^p, \theta)$ and $\tilde{x} = \psi(x^p, y^q, \vartheta)$ (or $\tilde{x} = \phi(x^p, z^r, \theta)$ and $\tilde{x} = \psi(x^p, y^q, z^r, \vartheta)$ when considering the conditional causality). The first one only depends on the history of x itself while the second depends on the history of y plus x's own. The realisation of these predictors are based on the estimation of the parameters θ and ϑ, which depends on the samples of x, denoted by $X = [X_1, \cdots, X_T]$, and the samples of y, denoted by $Y = [Y_1, \cdots, Y_T]$ plus X respectively.

2.2 KL Divergence

A natural quantity to describe the distance between two distributions is the celebrated Kullback-Leibler (KL) divergence that in fact defines a relative entropy between them [12,13]. Let $P(\cdot)$ and $Q(\cdot)$ be two probability distributions on a common state space $\{\Omega, \mathcal{F}\}$ with state space Ω and σ-algebra $\mathcal{F}$. Their KL divergence from $P(\cdot)$ to $Q(\cdot)$, namely, the relative entropy of $P(\cdot)$ over $Q(\cdot)$ can be formulated as

$$D_{KL}(P \parallel Q) = \int_{\Omega} \ln\left(\frac{dP(x)}{dQ(x)}\right) P(dx). \tag{2}$$

An essential nature of relative entropy is: $D_{KL}(P \parallel Q) = 0$ if and only if $P = Q$. But, KL divergence is asymmetry, i.e., $D_{KL}(P \parallel Q) \neq D_{KL}(Q \parallel P)$ in general. Also, we can denote the *conditional KL divergence* between two conditional probability $P(u|v)$ and $Q(u|z)$ for random variables (vectors) v and z as follows [14]:

$$D_{KL}[P(\cdot|v) \parallel Q(\cdot|z)] = \int_{\Omega_{v,z}} \int_{\Omega} \ln\left(\frac{dP(u|v)}{dQ(u|z)}\right) P(du|v)W(dv, dz) \tag{3}$$

where $W(dv, dz)$ denotes the joint probability distribution of (v, z) on their joint state space $\Omega_{v,z}$. It can be seen that $D_{KL}[P(\cdot|v) \parallel Q(\cdot|z)] = 0$ if and only if $P(\cdot|v) = Q(\cdot|z)$ with probability one in the sense of $W(dv, dz)$ when $v = z$.

3 REC: Relative Entropy Causality

We come back to Wiener's original definition of causality in terms of prediction by the difference of the residual errors in the statistic model with and without considering the inference of y.

Definition 1. The causality intervene from y to x is defined as the following conditional relative entropy:

$$REC_{y \to x} = D_{KL}\left[P(\cdot|x^p \oplus y^q) \parallel Q(\cdot|x^p)\right]. \tag{4}$$

named *relative entropy (RE) causality.* $P(\epsilon|x^p \oplus y^q)$ is the (stationary) conditional probability distribution of the residual error for the predictor ψ, depending on x and y's history, where the symbol $\oplus$ stands for the union of two random vectors, and $Q(\epsilon|x^p)$ is that for predictor ϕ, only depending on x's history.

Take the MVAR model for example. Let x_t and y_t be two real random processes of one-dimension[1] in discrete time. Consider the following two MVAR models:

$$x_t = x_t^p \cdot \tilde{a} + \tilde{\epsilon}_t \tag{5}$$
$$x_t = x_t^p \cdot a + y_t^q \cdot b + \epsilon_t \tag{6}$$

with the lag orders p and q, and $t = 1, 2, \cdots, T$, where $a, \tilde{a} \in \mathbb{R}^p$, $b \in \mathbb{R}^q$, $\tilde{\epsilon}_t, \epsilon_t \in \mathbb{R}$ and T is the time duration. These two models give two predictors $\phi(x_t^p, \tilde{a}) = x_t^p \cdot \tilde{a}$ and $\psi(x_t^p, x_t^q, a, b) = x_t^p \cdot a + y_t^q \cdot b$. At this stage, we assume that all time series are stationary (w.r.t measures). We should highlight that we do not include a constant term in the formulas of MVAR, which, however, is partially absorbed in the residual terms ϵ and $\tilde{\epsilon}$. In other words, ϵ and $\tilde{\epsilon}$ may have nonzero means in these models.

To specify this definition, we consider the scenario of one dimension by further assuming that all random variables are stationary (w.r.t measures), ergodic and Gaussian, and conduct parameter estimation by some approach, for example, the least mean square (LMS) approach or the maximum likelihood approach. Consider a column random vector u and a random matrix v of the identical column dimension. Let $\mathbb{E}(u)$ be the expectation of u, $\Sigma(u) = \mathbb{E}(uu^\top) - \mathbb{E}(u)[\mathbb{E}(u)]^\top$ be the covariance of u, $\Sigma(u, v) = \mathbb{E}(uv^\top) - \mathbb{E}(u)\mathbb{E}(v^\top)$ be the covariance between u and v, $\Sigma(u|v) = \Sigma(u) - \Sigma(u, v)\Sigma(v)^{-1}\Sigma(u, v)^\top$ be the partial covariance [7], and $\mu(u|v) = \mathbb{E}(u) - E(v)\Sigma(v)^{-1}\Sigma(v, u)$.

Let $X^p = [x_p^p, \cdots, x_T^p]$ and $Y^q = [y_q^q, \cdots, y_{T-1}^q]$ be the samples of the history of x and y picked from X and Y respectively. Then, by the least means square approach, suppose that the parameters ($\tilde{a}$, a and b) can be asymptotically perfectly estimated, i.e., the estimated values of $\tilde{a}$, a and b dependent of the samples converge to their theoretical value as the number of samples goes to infinity. Hence, we can obtain the asymptotic conditional distribution of $\tilde{\epsilon}$ and ϵ with respect to x^p and $x^p \oplus y^q$ as follows:

$$\begin{aligned} &\epsilon | x^p \oplus y^q \sim N_0 \left[\mu(x|x^p \oplus y^q), \Sigma(x|x^p \oplus y^q)\right] \\ &\tilde{\epsilon} | x^p \sim N_1 \left[\mu(x|x^p), \Sigma(x|x^p)\right]. \end{aligned} \tag{7}$$

Due to the limit of space, we omit the calculations which is not particularly difficult.

Thus, the RE causality from y to x in MVAR (5, 6) is:

$$\begin{aligned} &REC_{y \to x} = D_{KL}(N_0 \parallel N_1) \\ &= \frac{1}{2} \Big\{ \mathrm{tr}[\Sigma(x|x^p)^{-1}\Sigma(x|x^p \oplus y^q)] + [\mu(x|x^p \oplus y^q) - \mu(x|x^p)]^\top \\ &\Sigma(x|x^p)^{-1}[\mu(x|x^p \oplus y^q) - \mu(x|x^p)] \\ &-n - \ln \frac{\det[\Sigma(x|x^p \oplus y^q)]}{\det[\Sigma(x|x^p)]} \Big\}. \end{aligned} \tag{8}$$

[1] For the case of high dimensions, the similar results can be derived by the same fashion.

It can be seen that $REC_{y\to x} = 0$ if and only if $\Sigma(x/x^p) = \Sigma(x/x^p \oplus y^q)$ and $\mathbb{E}(x/x^p \oplus y^q) = \mathbb{E}(x/x^p)$. In comparison with Granger causality, in the MVAR model and under Gaussian assumption, the coefficients are estimated by the maximum likelihood approach. It can be defined as[2]

$$GC_{y\to x} = \ln\{\det[\Sigma(x|x^p)]/\det[\Sigma(x|x^p \oplus y^q)]\}.$$

Hence, $GC_{y\to x} = 0$ if and only if $\Sigma(x|x^p) = \Sigma(x|x^p \oplus y^q)$, namely, the covariance contribution of Y^q can be totally replaced by X^q. Therefore, in this scenario, the RE causality contains the y's contribution to predicting x in both covariance and mean, but the Granger causality only includes its covariance contribution.

Definition 2. The relative entropy causality from y to x at time t as the conditional KL divergence between two conditional probability distributions at time t: $REC_{y\to x}(t) = D_{KL}\left[P_t(\cdot|x_t^p \oplus y_t^q) \parallel Q_t(\cdot|x_t^p)\right]$, where $P_t(\cdot|x_t^p \oplus y_t^q)$ and $Q_t(\cdot|x_t^p)$ be the conditional distribution of ϵ_t at time t with respect to $x_t^p \oplus y_t^q$ and that of $\tilde{\epsilon}_t$ at time t with respect to x_t^p. The global RE causality as the average across the infinite time duration $[t_0, \infty)$:

$$REC_{y\to x} = \lim_{T\to\infty} \frac{1}{T} \sum_{t=t_0}^{t_0+T-1} REC_{y\to x}(t)dt, \tag{9}$$

if the limit exists and is independent of t_0, for example, the general Oscelec ergodicity conditions are satisfied [15].

4 Experiments

4.1 Stationary MVAR

To compare the RE causality with Granger causality in the MVAR model with Gaussian assumption (hence transfer entropy is equivalent to Granger causality), we consider one-order and one-dimensional version of (6) that generates the series x_t as:

$$x_t = a \cdot x_{t-1} + b \cdot y_{t-1} + \epsilon_{1,t} \tag{10}$$

where $a = 1/2$ and $b = -1/3$, and $\epsilon_{1,t}$ is a white Gaussian noise of zero mean and variance σ_1^2 independent of all y_t and all x_t. First, we consider the case that y_t is stationary: $y_t = c \cdot y_{t-1} + \epsilon_{2,t}$ where $c = 2/3$ and $\epsilon_{2,t}$ is a white Gaussian noise of zero mean and variance σ_2^2 independent of all y_t and all x_t.

The value of RE causality from y to x, denoted by $\widetilde{REC}_{y\to x}$ can be calculated by estimating the means and variances in (5) and (6) by maximum likelihood and employing equation (8). We use the bootstrap approach for statistic inference. Under the null hypothesis $b = 0$, we sample a series of $\tilde{x}_t$ by (10) with the estimated values $\tilde{a}$ of a and the estimated noise variance $\tilde{\sigma}_1^2$. Both are estimated

[2] Also as $GC_{y\to x} = \ln\{\mathrm{tr}[\Sigma(x|x^p)]/\mathrm{tr}[\Sigma(x|x^p \oplus y^q)]\}$.

by the maximum likelihood. Then, the estimated values of $\widehat{REC}^{k}_{y\to x}$ can be obtained from $\tilde{x}_t$ and y_t by Eq. (8) for the k-th realisation. Repeat this phase for $N = 1000$ times and get a collection of $\widehat{REC}^{k}_{y\to x}$, $k = 1, \cdots, N$. To avoid a zero p-value, we let $\widehat{REC}^{0}_{y\to x} = \widetilde{REC}_{y\to x}$. Thus, the p-value of the RE causality is calculated by $p_{REC} = \#\{k : \widehat{REC}^{k}_{y\to x} \geq \widetilde{REC}_{y\to x}\}/(N+1)$. Also, we calculate the RE causality from x to y by the above fashion. In comparison, the Granger causality and its p-value is calculated by its definition and F-test for both directions. When the p-value is less than a pre-given threshold p_{th}, we claim that there is a causal intervene. All the above are overlapped for $M = 100$ times and the true positive (TP) rate, i.e., the ratio (among M overlaps) that the directed causal intervene from y to x is correctly identified, and the false positive (FP) rate, i.e., the ratio (among M overlaps) that the directed causal intervene from x to y is incorrectly identified, are used to evaluate the performance of the causality definitions (Table 1).

Table 1. Performance comparison on stationary MVAR between Granger causality (Transfer entropy) and RE causality.

	True positive	False positive
Granger causality	100%	0
RE causality	100%	0

The above results show that they have similar performances to identify the intervene from y to x and exclude that from x to y for diverse p-value thresholds and model noise variances.

4.2 Non-stationary MVAR

To compare the RE causality with Granger causality in the nonstationary MVAR model, we define Granger causality in non-stationary case by the same token. Specified in the following model, by estimating the regressive parameters, one can define Granger causality and transfer entropy at each time (equivalent phase) and the whole Granger causality and transfer entropy can be defined as the averages across the time duration respectively.

We consider a *hidden periodic model* ([16]) to generate y_t as: $y_t = d \cdot [\sin(2\pi \cdot t/T_1) + \cos(2\pi \cdot t/T_2) + \nu_t]$, where d is a positive scale, $T_1 = 5$, $T_2 = 7$, and ν_t is a white Gaussian noise with zero mean I and variance σ^2, independent of all x_t and all $\epsilon_{1,t}$. The noise $\epsilon_{1,t}$ is white Gaussian with zero means and a static variance $\sigma_0 = 5$. x_t is generated by the MVAR (10). Their sampling time series are demonstrated in Fig. 1(A). The periods and other parameters in the hidden periodic model can be estimated from the data. Here, for simplicity, we assume that the periods are precisely estimated. To calculate $REC_{y\to x}(t)$

at a specific time t, we have multiple recording series of x_t accompanied with the same single record series of y_t, with different random noises, denoted by x_t^q for time $t = 1, \cdots, T$ and multiple index $q = 1, \cdots, Q$. After estimating the parameters in (5) and (6), we obtain the multiple series of the residual errors, denoted by $\tilde{\epsilon}_t^q$ and ϵ_t^q, respectively. Then, at each $t = t_0$, we collect $\tilde{\epsilon}_{t_0}^q$ and $\epsilon_{t_0}^q$ with $q = 1, \cdots, Q$, to calculate $\widetilde{REC}_{y\to x}(t)$ by formula (8). Then, averaging $\widetilde{REC}_{y\to x}(t)$ across $t = 1, \cdots, T$ leads $\widetilde{REC}_{y\to x}$. Alternatively, if we do not have multiple recordings of time series and the periods are known, we can segment a long time series of x_t with a period equal to the least common multiple of all periods ($T_0 = T_1 \times T_2 = 35$). After estimating the parameters, we collect the residual errors $\epsilon_l^q = \epsilon_t$ with $l = mod(t, T_0)$ and $q = \lfloor t/T_0 \rfloor$, where $mod(a, b)$ stands for the remainder of a driven by b and $\lfloor r \rfloor$ stands for the largest integer less than r. Then, by the same way, we can calculate $\widetilde{REC}_{y\to x}(l)$ for each $l = 0, \cdots, T_0 - 1$ and average them to obtain $\widetilde{REC}_{y\to x}$. The p-values can be estimated by the bootstrap approach in the analogical fashion mentioned above.

To calculate the Granger causality in this scenario, we perform two methods. One is the classic Granger causality approach without considering the hidden periodic property in the MVAR model. The other is analogy to the calculation process of RE causality mentioned above: using multiple recordings of time series of x_t or segmenting x_t with a period T_0, estimating the co-variances at each time t (or equivalent phase), calculating the Granger causality at t, and averaging them across the time duration or the whole period. TP and FP are calculated by $M = 100$ overlaps.

As Fig. 1(B) and (C) show, the periodic Granger causality failed to identify any causal intervene from y to x when σ_y is small, since the intervene of hidden periodic y_t does not cause significant variation in the variances in the residual error in (6) in comparison to (5). However, the classic Granger causality can probe this intervene when b is not very small, since the hidden periodicity in y_t that leads fluctuation can cause variance in (6) different from (5). However, the RE causality is clearly more efficient than Granger causality to identify the intervene from y to x (larger TP) and less incorrect identification (lower FP).

4.3 Time-Varing Nonstationary MVAR

We consider a more complicated time-varying means with y_t generated by a *hidden chaotic model*: Let

$$y_t = \sigma \cdot \zeta(t) + \nu_t \tag{11}$$

where $\zeta(t)$ is generated by the logistic iteration map, namely, $\zeta(t) = \beta\zeta(t-1)[1-\zeta(t-a)]$ with $\beta = 3.9$ and $\zeta(1) \in (0, 1)$, which implies that $\zeta(t)$ possesses a chaotic attractor, and ν_t is an independent identical distributed process with the uniform distribution in $[0, 0.5]$, as illustrated in Fig. 2(A). As above, x_t is generated by the model (10). With the knowledge of the distribution with unknown parameters, we can employ the method above to calculate the RE causality of non-stationary time series by $Q = 100$ multiple recordings of the time series up to $T = 200$.

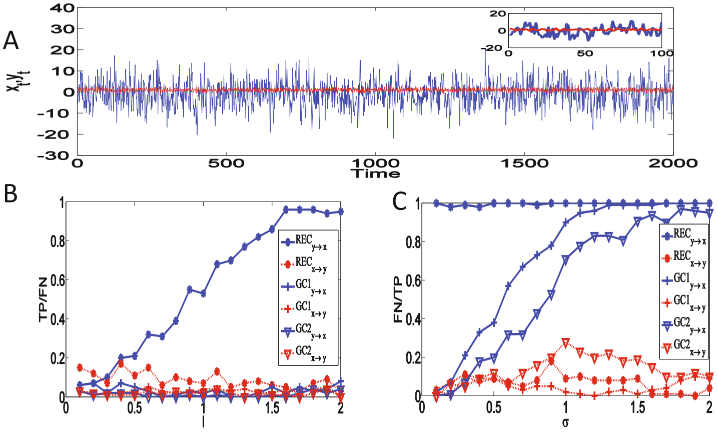

Fig. 1. Performance comparison among the RE causality (REC, $-*-$), classic Wiener-Granger causality (GC1, $-+-$) and time-varying Wiener-Granger causality (GC2, $-\nabla-$) by the TP rate (blue solid lines), and FN rate (red dash lines) for a causal intervene signal generated by (11). (A) the sampling hidden chaotic time series of X_1 (blue line) and X_2 (red line) as well as the inset plots for them in a small time intermal [1, 100]; (B) performance comparison of hidden periodic model with respect to I with fixed $\sigma = 0.2$; (C) performace comparison of hidden periodic model with respect σ with fixed $I = 2$. (Color figure online)

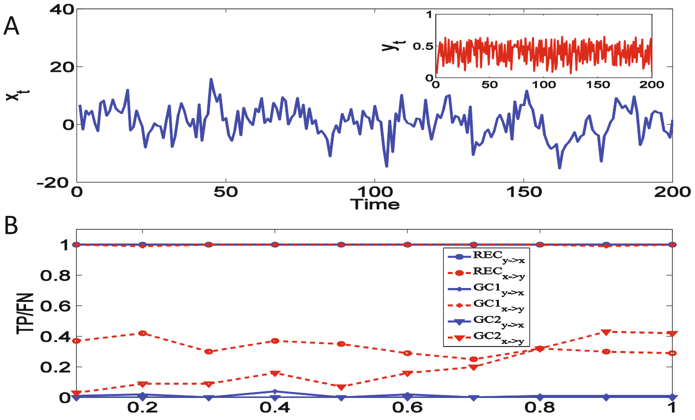

Fig. 2. Performance comparison among the RE causality (REC, $-*-$), classic Wiener-Granger causality (GC1, $-+-$) and time-varying Wiener-Granger causality (GC2, $-\nabla-$) by the TP rate (blue solid lines), and FN rate (red dash lines). (A) the sampling hidden chaotic time series of X_1 (blue line) and X_2 (red line); (B) the performance comparison of hidden chaotic model with respect to σ. (Color figure online)

Figure 2(B) show that the RE causality performs very well in identifying the causal intervene from y to x and excluding that from x to y. In comparison the classic Granger and non-stationary Granger methods, which are deployed still with Gaussian assumption for statistic inference, RE causality is clearly better than them.

4.4 Predictive Models

With the help of modern predictive models, the prediction accuracy can be improved obviously compared with linear regression. As a consequence, we want to compare the performance of RE causality and Granger causality combining with some predictive model.

In this example, x_t is generated by a nonlinear model:

$$x_t = a\sin(x_t - 1) + by_t - 1 + cy_{t-2} + \epsilon_{1,t} \tag{12}$$

with y_t generated by: $y_t = c - dy_{t-1} + \sin(\epsilon_{2,t})$.

Applying XGBoost on the sampling time series up to $T = 1000$ to predict x_t using x and y's history and x's history respectively, we obtain the error time series ϵ and $\tilde{\epsilon}$. Assuming that the error time series follow Gaussian Mixture Model (GMM), we use EM algorithm for fitting, and thus obtain the distributions of ϵ and $\tilde{\epsilon}$. Since there is no closed form for the KL divergence between GMMs, we choose to do Monte Carlo to approximate the causalities. The last step is to calculate the TP, FP rate of RE causality and transfer entropy by $P = 100$ multiple recordings of time series.

Table 2. Performance comparison between Transfer entropy and RE causality combining with XGBoost.

	True positive	False positive
Transfer entropy	75%	0
RE causality	80%	0

Table 2 show that RE causality slightly improves the performance of identifying the causal intervene from y to x and both causality exclude that from x to y.

5 Discussions and Conclusions

In conclusion, we propose a novel definition of Wiener causality by the relative entropy of the residual errors of prediction with and without considering the other time series. In comparison to the well-known Granger causality, which defines causality as the difference of the conditional co-variances, this RE causality is in the normal track to realise the Wiener's causal philosophy. For example, the intervene of the means in the time series cannot be identified by the Granger causality; however, it has contribution to prediction, in other words, span the future of x. We illustrate this argument by several models as intervenes with presenting the non-stationary RE causality. The fluctuation of the first-order statistics definitely causes information flow to x, so a causality, according to Wiener's arguments [3]. At the meantime, we may employ some predictive models, such as XGBoost, to provide high prediction accuracy, and we discover that RE causality can lead to further improvement of the reliability of causal relations inferred from data.

Acknowledgments. This work is jointly supported by the National Natural Sciences Foundation of China under Grant No. 61673119, the Key Program of the National Science Foundation of China No. 91630314, the Laboratory of Mathematics for Nonlinear Science, Fudan University, and the Shanghai Key Laboratory for Contemporary Applied Mathematics, Fudan University.

References

1. Hu, S., Wang, H., Zhang, J., Kong, W., Cao, Y., Kozma, R.: Comparison analysis: granger causality and new causality and their applications to motor imagery. IEEE Trans. Neural Netw. Learn. Syst. **27**(7), 1429–1444 (2016)
2. Pearl, J.: Causality: Models, Reasoning, and Inference. Cambridge University Press, Cambridge (1999)
3. Wiener, N.: The Theory of Prediction, Modern Mathematics for Engineers. McGraw-Hill, New York (1956)
4. Granger, C.: Investigating causal relations by econometric models and cross-spectral methods. Econ. J. Econ. Soc. **37**(3), 424–438 (1969)
5. Schreiber, T.: Measuring information transfer. Phys. Rev. Lett. **85**(2), 461 (2000)
6. Liang, X.S., Kleeman, R.: Information transfer between dynamical system components. Phys. Rev. Lett. **95**(24), 244101 (2005)
7. Bernett, L., Barrett, A.B., Seth, A.K.: Granger causality and transfer entropy are equivalent for Gaussian variables. Phys. Rev. Lett. **103**(23), 238701 (2009)
8. Sobrino, A., Olivas, J.A., Puente, C.: Causality and imperfect causality from texts: a frame for causality in social sciences. In: International Conference on Fuzzy Systems, Barcelona, pp. 1–8 (2010)
9. Ding, M., Chen, Y., Bressler, S.: Handbook of Time Series Analysis: Recent Theoretical Developments and Applications. Wiley, Wienheim (2006)
10. Seth, A.K., Barrett, A.B., Barnett, L.: Granger causality analysis in neuroscience and neuroimaging. J. Neurosci. **35**(8), 3293–3297 (2015)
11. Chen, T.Q., Guestrin, C.: XGBoost: a scalable tree boosting system. arXiv:1603.02754v3 (2016)
12. Kullback, S., Leibler, R.A.: On information and sufficiency. Ann. Math. Stat. **22**(1), 79 (1951)
13. Cliff, O.M., Prokopenko, M., Fitch, R.: Minimising the Kullback-leibler divergence for model selection in distributed nonlinear systems. Entropy **20**(2), 51 (2018)
14. Cover, T.M., Thomas, J.A.: Elements of Information Theory. Wiley, New York (1991)
15. Oseledec, V.I.: A multiplicative ergodic theorem: Liapunov characteristic number for dynamical systems. Trans. Moscow Math. Soc. **19**, 197 (1968)
16. He, S.Y.: Parameter estimation of hidden periodic model in random fields. Sci. China A **42**(3), 238 (1998)