Article

Causality inference of linearly correlated variables: The statistical simulation and regression method

WenJun Zhang
School of Life Sciences, Sun Yat-sen University, Guangzhou 510275, China
E-mail: zhwj@mail.sysu.edu.cn, wjzhang@iaees.org

Received 11 October 2021; *Accepted 17 October 2021*; *Published online 20 October 2021; Published 1 December 2021*

Abstract
Causality inference of variables is a research focus in science. Due to its importance, a statistical simulation and regression method for causality inference of linearly correlated (scale or interval) variables was proposed in present study. First, a statistical simulation and regression method was developed to generate and analyze artificial data of linear correlated variables with known causality. The rule was drawn from the simulation and regression analysis on artificial data. Finally, causality inference of two linearly correlated variables was conducted based on the rule. Full Matlab codes of the method were presented.

Keywords causality; inference; linear dependency; scale or interval variables; Pearson correlation; statistical simulation; regression analysis.

Computational Ecology and Software
ISSN 2220-721X
URL: http://www.iaees.org/publications/journals/ces/online-version.asp
RSS: http://www.iaees.org/publications/journals/ces/rss.xml
E-mail: ces@iaees.org
Editor-in-Chief: WenJun Zhang
Publisher: International Academy of Ecology and Environmental Sciences

1 Introduction

Causality inference has been a frontier area in science for a long time. Past studies on causality inference have been mainly focusing on Bayesian methods. Unfortunately, few successful cases have been reported in causality inference. In the natural sense, causality will lead to correlation between two variables.

Causality inferernce can be conducted only the two variables correlate with each other. Theories and applications of correlations have been well studied, including that in biology and ecology (Qi and Zhang, 2003; Kuang and Zhang, 2011; Huang and Zhang, 2012; Jiang and Zhang, 2015a, b; Zhang, 2007, 2011b, 2012a, 2014-2018, 2021a-c; Zhang et al., 2014). There are many correlation measures, among which Pearson correlation is for interval or scale variables.

Besides conventional Bayesian methods, statistical simulation methods are widely used to make statistical inferences (Solow, 1993; Manly, 1997; Zhang and Schoenly, 1999; Zhang, 2010, 2011a). Zhang (2021a, b) has proposed some statisticall simulation methods to conduct causality inference of Boolean variables and nominal variables. However, scale or interval variables are most often used in the practical activities. In present study, I thus proposed a statistical simulation and regression method for causality inference of linearly correlated scale or interval variables. Full Matlab codes were presented for practical uses.

2 Correlation and Statistic Test of Linearly Correlated Variables

2.1 Correlation measure of linearly correlated variables

As pointed out earlier (Zhang, 2021a, b), causality will result in correlation between two variables. For statistic sample data, causality inferernce is available only the two variables are correlated with each other. In present study, I use Pearson correlation as the correlation measure of linearly correlated (scale or interval) variables x and y (Zhang, 2007, 2011b, 2012a, 2014-2018, 2021a-c; Zhang and Li, 2015):

$$r=\sum_{k=1}^{n}((x_k-\bar{x})(y_k-\bar{y}))/(\sum_{k=1}^{n}(x_k-\bar{x})^2\sum_{k=1}^{n}(y_k-\bar{y})^2)^{1/2}$$

where $-1\le r\le 1$, $\bar{x}=\sum_{k=1}^{n}x_k/n$, $\bar{y}=\sum_{k=1}^{n}y_k/n$, and n is the sample size.

2.2 Detection of correlation between two linearly correlated variables

Causality between two linearly correlated variables can be infered if their linear correlation is statistically significant. For the Pearson correlation, calculate $t=|r|/[(1-r^2)/(n-2)]^{1/2}$, and if $t>t_p(n-2)$, the linear correlation (either positive or negative correlation) between variables x and y is statistically significant.

3 Causality Inference of Two Linearly Correlated Variables

To find the general rule of causality and correlation between two linearly correlated variables, the artificial data of two linearly correlated variables, from the independent variable x to dependent variable y, can be constructed and analyzed using statistical simulation and regression method.

3.1 Causality principle of linearly correlated variables

Assume that the causality exists between two linearly correlated variables, x and y, and x is the independent variable and y is the dependent variable. In a simulation, first, randomly generate a set of data for independent variable x with the random size. Second, generate a set of data for dependent variable y following the equation (1), with random a and b and a small random error ε_x:

$$y=a+bx\pm\varepsilon_x \quad (1)$$

Third, standardize the generated data of variables x and y:

$$x'=(x-min\ x)/(max\ x-min\ x)$$
$$y'=(y-min\ y)/(max\ y-min\ y)$$

And construct the regression relationships with x' and y' as independent variable and dependent variable, and y' and x' as independent variable and dependent variable, respectively:

$$y'=\alpha+\beta x' \quad (2)$$
$$x'=\alpha'+\beta' y' \quad (3)$$

It should be noted that for Pearson correlation r between x and y, and r' between x' and y', $r=r'$. Calculate and summarize the mean of absolute residuals for prediction of y' from x', $r_{x'y'}$, and for prediction of x' from y', $r_{y'x'}$:

$$r_{x'y'}=\sum_{k=1}^{n}|y'_k-(\alpha+\beta x'_k)|/n$$
$$r_{y'x'}=\sum_{k=1}^{n}|x'_k-(\alpha'+\beta' y'_k)|/n$$

At the end of the simulation, record r, $r_{x'y'}$ and $r_{y'x'}$. Repeat above procedure s times. Finally, calculate the mean of Pearson correlation r of all simulations, the ratio of means of $r_{x'y'}$ and $r_{y'x'}$, res_{mr}, and the frequency of $r_{x'y'}>r_{y'x'}$, p_{xy}:

$$\bar{r}=\sum_{k=1}^{s} r_k/s$$
$$res_{mr}= mean\ r_{x'y'} / mean\ r_{y'x'}$$
$$p_{xy}=\sum_{k=1}^{s}(r_{x'y'}>r_{y'x'})/s$$

The full Matlab codes, xyGen, of the statistical simulation and regression analysis for finding relationship between causality and statistic parameters are as follows (Fig. 1; see supplementary material also):

```
clear;
sel=1;          %sel=1: positive correlation; sel=2: negative correlation
yerr=0.2;     %Relative error of y following x in a given pattern
n=100;          %For determining the maximum size of nominal variables x and y
sim=2000;     %Number of simulations (randomizations)
for s=1:sim
nm=floor(n*rand()+30);       %The size, nm, can be fixed, e.g., nm=n
x=zeros(nm,1);
y=zeros(nm,1);
coeff=rand()*5;                 %To set regress coefficient: rand()*0.1, *0.5, *4, *10, etc.
const=rand()*10;                %To set regress constant: rand()*0.3, *0.8, *9, *20, etc.
x=rand(nm,1);
yres=rand(nm,1)*yerr;
si=floor(rand(nm,1)+0.5)+1;
if (sel==1)
y=const+coeff*x+(-1).^si.*yres.*x*coeff;                %Normal distributed random error can be used also.
elseif (sel==2)
y=max(y)-const-coeff*x-(-1).^si.*yres.*x*coeff;       %Normal distributed random error can be used also.
end
x=(x-min(x))/(max(x)-min(x));
y=(y-min(y))/(max(y)-min(y));
[bxy,bxyint,rxy,rintxy,statsxy]=regress(y,[ones(nm,1) x]);
[byx,byxint,ryx,rintyx,statsyx]=regress(x,[ones(nm,1) y]);
r=corr(x,y);
res(s,:)=[r sum(abs(rxy))/nm sum(abs(ryx))/nm] ;
end
rmean=mean(r)
resmeanratio=mean(res(:,2))/mean(res(:,3))
pxy=sum(res(:,2)>res(:,3))/sim
```

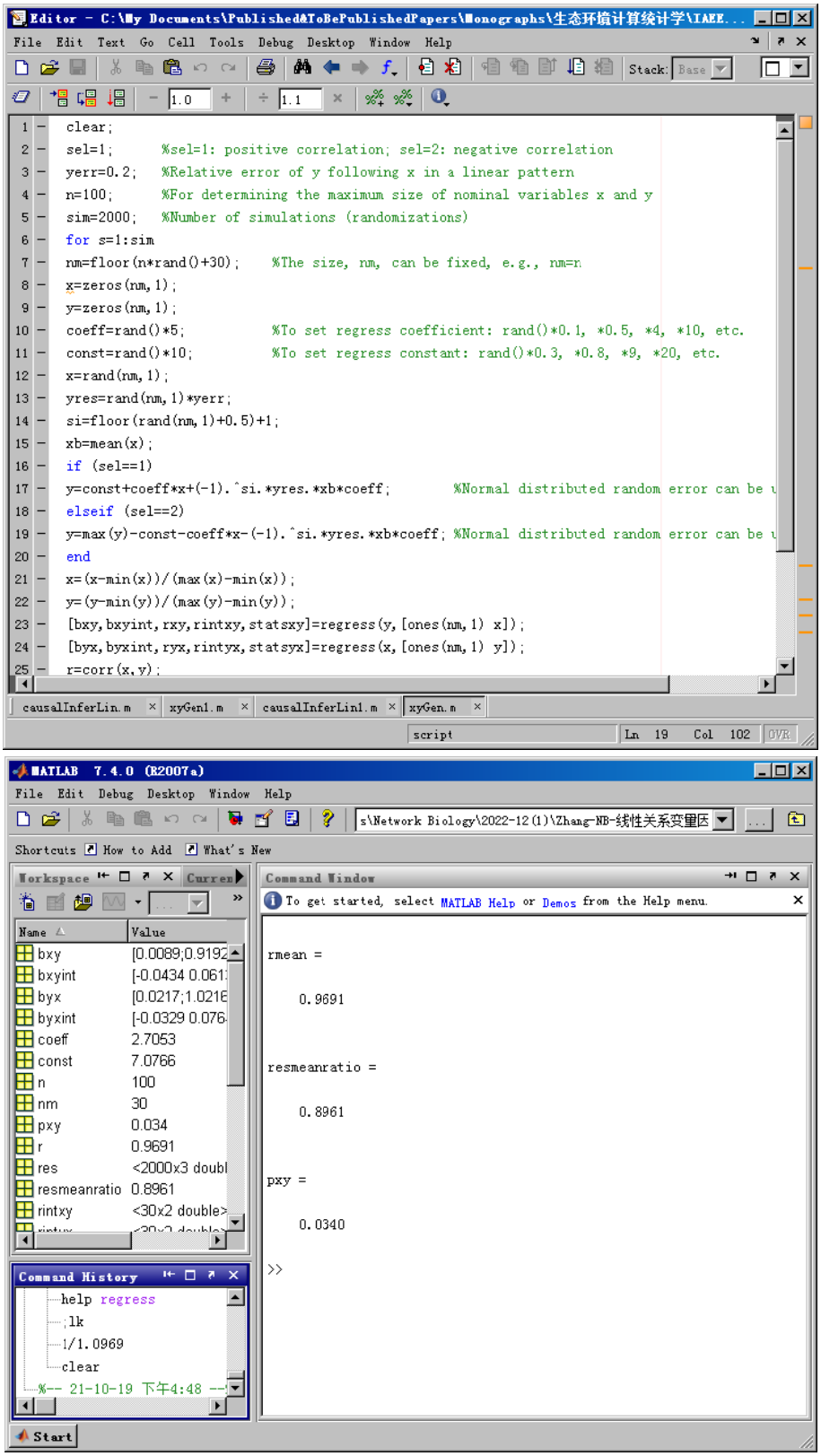

Fig. 1 Matlab interface for statistical simulation.

3.2 Causality inference of linearly correlated variables

3.2.1 Rule found from statistical simulation and regressional analysis

The results of statistical simulation and regression analysis demonstrated that for the causality of two linearly correlated variables with x as independent variable and y as dependent variable, $res_{mr}<1$, $p_{xy}<p$, p=0.01, 0.05, 0.1, etc (Fig. 1).

3.2.2 Causality inference based on observed data of linearly correlated variables

According to the rule drawn above, $res_{mr}<1$ can be used as the criteria for possible causality of linearly correlated variables x (independent variable) and y (dependent variable), and $res_{mr}>1$ can be used as the criteria for possible causality of linearly correlated variables y (independent variable) and x (dependent variable).

The full Matlab codes, causalInferLin, of the method for causality inference based on observed data of linearly correlated variables are as follows (Fig., 2; see supplementary material also):

```
clear
xyd=input('Input the Excel file name of raw data (e.g., xyd.xls: xyd=(dij)n×2, i=1,2,...,n; j=1,2. In the file, column 1 is for variable 1 and column 2 is for variable 2): ','s');
p=input('Input the statistical significance level p for correlation inference (e.g., 0.001): ');
xyd=xlsread(xyd);
n=size(xyd,1);
x=xyd(:,1);
y=xyd(:,2);
r=corr(x,y);
tvalue=abs(r)/sqrt((1-r^2)/(n-2));
alp=(1-tcdf(tvalue,n-2))*2;
id=0;
if (alp<p)
sprintf(['There is a significant Pearson correlation (r=',num2str(r),') between two linearly correlated variables (p=',num2str(alp),')\n'])
id=1;
else sprintf(['There is not significant Pearson correlation (r=',num2str(r),') between two variables (p=',num2str(alp),')\n'])
sprintf(['So, causality may not exist between two variables based on Pearson correlation.\n'])
end
xs=(x-min(x))/(max(x)-min(x));
ys=(y-min(y))/(max(y)-min(y));
[bxy,bxyint,rxy,rintxy,statsxy]=regress(ys,[ones(n,1) xs]);
[byx,byxint,ryx,rintyx,statsyx]=regress(xs,[ones(n,1) ys]);
resxy=sum(abs(rxy))/sum(abs(ryx));
if (resxy<1)
sprintf(['The variable for column 1 is most likely the independent variable, and the variable for column 2 is most likely the dependent variable.\n'])
elseif (resxy>1)
sprintf(['The variable for column 2 is most likely the independent variable, and the variable for column 1 is most likely the dependent variable.\n'])
end
```

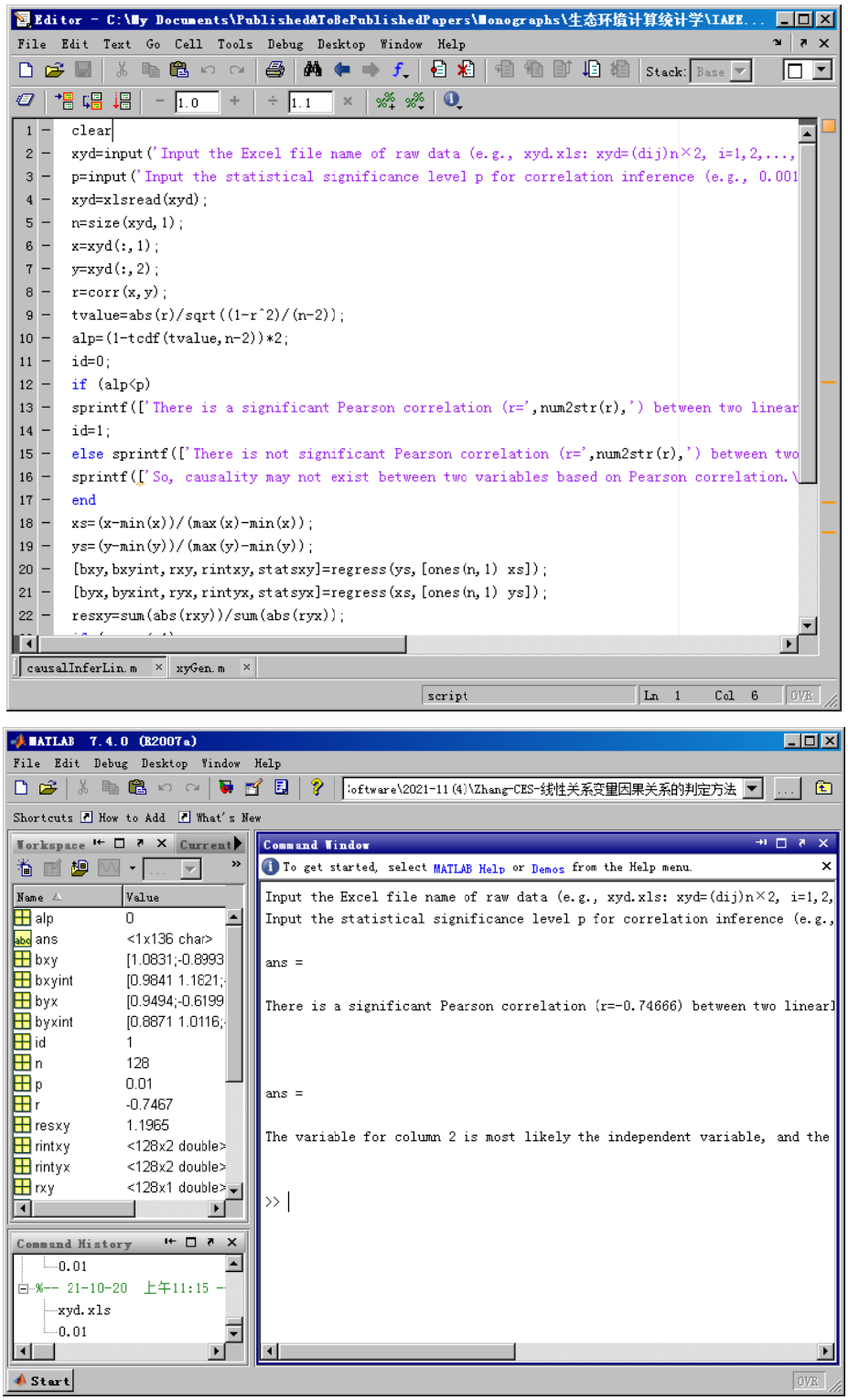

Fig. 2 Matlab interface for making causality inference.

A set of theoretical data were used to test the method and the validity of the method was overally confirmed (Fig. 2). Most of the data have been correctly predicted (Fig. 1, Fig. 2).

4 Discussion

It is suggested that data size (sample size, *n*) of variables should be large enough in order to enhance the reliability of causality inference. The present method is expected to be fundamental because more complex functional relationship can be approximated as the linear relationship in the local domain. Further improvement of the method includes the estimation of statistic confidence degree for causality inference.

Acknowledgment

I am thankful to the support of The National Key Research and Development Program of China (2017YFD0201204), from China, and Discovery and Crucial Node Analysis of Important Biological and Social Networks (2015.6-2020.6), from Yangling Institute of Modern Agricultural Standardization, China.

References

Huang JQ, Zhang WJ. 2012. Analysis on degree distribution of tumor signaling networks. Network Biology, 2(3): 95-109

Jiang LQ, Zhang WJ. 2015a. Determination of keystone species in CSM food web: A topological analysis of network structure. Network Biology, 5(1): 13-33

Jiang LQ, Zhang WJ. 2015b. Effects of parasitism on robustness of food webs. Selforganizology, 2(2): 21-34

Kuang WP, Zhang WJ. 2011. Some effects of parasitism on food web structure: a topological analysis. Network Biology, 1(3-4): 171-185

Qi YH, Zhang WJ. 2003. CorreDetector: A network sharing software used in correlation analysis of information. Journal of Information, 22(Suppl.): 266-268

Zhang WJ. 2007. Computer inference of network of ecological interactions from sampling data. Environmental Monitoring and Assessment, 124: 253–261

Zhang WJ. 2010. Computational Ecology: Artificial Neural Networks and Their Applications. World Scientific, Singapore

Zhang WJ. 2011a. A Java program to test homogeneity of samples and examine sampling completeness. Network Biology, 1(2): 127-129

Zhang WJ. 2011b. Constructing ecological interaction networks by correlation analysis: hints from community sampling. Network Biology, 1(2): 81-98

Zhang WJ. 2012a. Computational Ecology: Graphs, Networks and Agent-based Modeling. World Scientific, Singapore

Zhang WJ. 2012b. How to construct the statistic network? An association network of herbaceous plants constructed from field sampling. Network Biology, 2(2): 57-68

Zhang WJ. 2014. Interspecific associations and community structure: A local survey and analysis in a grass community. Selforganizology, 1(2): 89-129

Zhang WJ. 2015. Calculation and statistic test of partial correlation of general correlation measures. Selforganizology, 2(4): 65-77

Zhang WJ. 2016. Selforganizology: The Science of Self-Organization. World Scientific, Singapore

Zhang WJ. 2017. Network pharmacology of medicinal variables and functions of Chinese herbal medicines: (II) Relational networks and pharmacological mechanisms of medicinal variables and functions of Chinese herbal medicines. Network Pharmacology, 2(2): 38-66

Zhang WJ. 2018. Fundamentals of Network Biology. World Scientific Europe, London, UK

Zhang WJ. 2021a. A statistical simulation method for causality inference of Boolean variables. Network Biology, 11(4): 263-273

Zhang WJ. 2021b. Causality inference of nominal variables: A statistical simulation method. Computational Ecology and Software, 11(4): 142-153

Zhang WJ. 2021c. A web tool for generating user-interface interactive networks. Network Biology, 11(4): 247-262

Zhang WJ, Li X. 2015. Linear correlation analysis in finding interactions: Half of predicted interactions are undeterministic and one-third of candidate direct interactions are missed. Selforganizology, 2(3): 39-45

Zhang WJ, Schoenly KG. 1999. IRRI Biodiversity Software Series. III. BOUNDARY: A Program for Detecting Boundaries in Ecological Landscapes. IRRI Technical Bulletin No.3. International Rice Research Institute, Manila, Philippines

Zhang WJ, Wang R, Zhang DL, et al. 2014. Interspecific associations of weed species around rice fields in Pearl River Delta, China: A regional survey. Selforganizology, 1(3-4): 143-205