Detection of causality in nonlinear time series using extremal value theory

Bc. Juraj Bodík

Charles University, Prague
Department of Probability and Mathematical Statistics,
Supervisor: doc. RNDr. Zbyněk Pawlas, Ph.D.
Consultant: RNDr. Milan Paluš, DrSc.

Date of the Master thesis defense: 21.6.2021

Outline

1. Preliminaries
2. Main idea
3. Other approaches
4. Simulations and application

Granger causality

Let $X = (X_t, t \in \mathbb{Z})$ and $Y = (Y_t, t \in \mathbb{Z})$ be time series.
X causes Y if there exists measurable set A such that

$$P(Y_{t+1} \in A \mid \mathbb{U}_t) \neq P(Y_{t+1} \in A \mid \mathbb{U}_t \setminus \sigma(X_s, s \leq t)),$$

where $\mathbb{U}_t$ is all the information in the universe until time t.

VAR model

Bivariate time series (X, Y) follows VAR model if

$$X_t = \alpha_1 X_{t-1} + \cdots + \alpha_q X_{t-q} + \gamma_1 Y_{t-1} + \cdots + \gamma_q Y_{t-q} + \varepsilon_t^X,$$
$$Y_t = \beta_1 Y_{t-1} + \cdots + \beta_q Y_{t-q} + \delta_1 X_{t-1} + \cdots + \delta_q X_{t-q} + \varepsilon_t^Y.$$

It can be rewritten as

$$X_t = \sum_{i=0}^{\infty} a_i \varepsilon_{t-i}^X + \sum_{i=0}^{\infty} c_i \varepsilon_{t-i}^Y,$$
$$Y_t = \sum_{i=0}^{\infty} b_i \varepsilon_{t-i}^Y + \sum_{i=0}^{\infty} d_i \varepsilon_{t-i}^X.$$

VAR model

Bivariate time series (X, Y) follows VAR model if

$$X_t = \alpha_1 X_{t-1} + \cdots + \alpha_q X_{t-q} + \gamma_1 Y_{t-1} + \cdots + \gamma_q Y_{t-q} + \varepsilon_t^X,$$
$$Y_t = \beta_1 Y_{t-1} + \cdots + \beta_q Y_{t-q} + \boxed{\delta_1 X_{t-1} + \cdots + \delta_q X_{t-q}} + \varepsilon_t^Y.$$

It can be rewritten as

$$X_t = \sum_{i=0}^{\infty} a_i \varepsilon_{t-i}^X + \sum_{i=0}^{\infty} c_i \varepsilon_{t-i}^Y,$$
$$Y_t = \sum_{i=0}^{\infty} b_i \varepsilon_{t-i}^Y + \boxed{\sum_{i=0}^{\infty} d_i \varepsilon_{t-i}^X}.$$

Nonlinear model (NAR)

Let $Z = (Z_t, t \in \mathbb{Z})$ be d-dimensional time series.
General form $Z_t = f(Z_{t-1}, \ldots, Z_{t-q}, \varepsilon_t)$.
Additive form $Z_t = f_1(Z_{t-1}) + \cdots + f_q(Z_{t-q}) + \varepsilon_t$.
For bivariate $Z = (X, Y)$ and $q = 1$ is

$$
\begin{aligned}
X_t &= f_1(X_{t-1}) + f_2(Y_{t-1}) + \varepsilon_t^X, \\
Y_t &= g_1(Y_{t-1}) + g_2(X_{t-1}) + \varepsilon_t^Y.
\end{aligned}
$$

Nonlinear model (NAR)

Let $Z = (Z_t, t \in \mathbb{Z})$ be d-dimensional time series.
General form $Z_t = f(Z_{t-1}, \dots, Z_{t-q}, \varepsilon_t)$.
Additive form $Z_t = f_1(Z_{t-1}) + \cdots + f_q(Z_{t-q}) + \varepsilon_t$.
For bivariate $Z = (X, Y)$ and $q = 1$ is

$$X_t = f_1(X_{t-1}) + f_2(Y_{t-1}) + \varepsilon_t^X,$$
$$Y_t = g_1(Y_{t-1}) + \boxed{g_2(X_{t-1})} + \varepsilon_t^Y.$$

Regular variation

We will always assume $(\varepsilon_i, i \in \mathbb{Z}) \overset{\text{iid}}{\sim} RV(\theta)$, which means

$$P(\varepsilon_t > x) \sim x^{-\theta} L(x)$$

for some slowly varying L.
Examples include Cauchy or Pareto distributions.

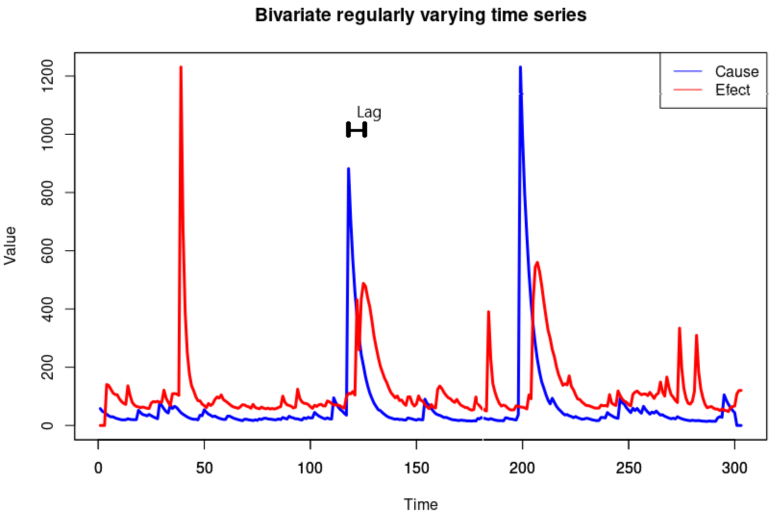
Main idea

Consider NAR model

$$
\begin{aligned}
X_t &= \frac{1}{2} X_{t-1} + \varepsilon_t^X, \\
Y_t &= \frac{1}{2} Y_{t-1} + \sqrt{X_{t-5}} + \varepsilon_t^Y,
\end{aligned}
$$

with $\varepsilon_t^X, \varepsilon_t^Y \overset{\text{iid}}{\sim} \textit{Pareto}$.

Sample realization



Causal tail coefficient for stationary time series

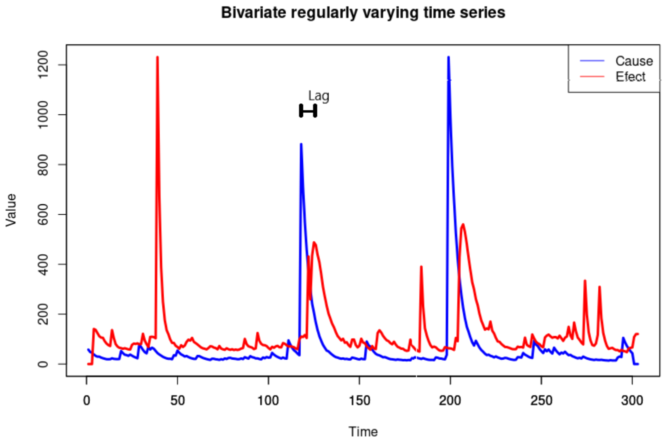
Therefore, we propose the following coefficient,

$$\Gamma^{time}_{X,Y}(q) := \lim_{u \to 1^-} \mathbb{E}[\max\{F_Y(Y_0), \dots, F_Y(Y_q)\} \mid F_X(X_0) > u],$$

where, in our case, is $q = 5$. Here, we denote by F_X, F_Y the stationary distributions of X_t and Y_t, respectively.

Causal tail coefficient for stationary time series

$$\Gamma^{time}_{X,Y}(q) = \lim_{u \to 1^-} \mathbb{E}[\max\{F_Y(Y_0), \ldots, F_Y(Y_q)\} \mid F_X(X_0) > u]$$



Main theorem

$$\Gamma^{time}_{X,Y}(q) = \lim_{u \to 1^-} \mathbb{E}[\max\{F_Y(Y_0), \dots, F_Y(Y_q)\} \mid F_X(X_0) > u]$$

Theorem

$(X, Y)^\top$ follow either heavy-tailed VAR or heavy-tailed NAR model of order q. Let "extremal causal condition" hold. Then,

- *X causes Y* $\implies \Gamma^{time}_{X,Y}(q) = 1$
- *X does not cause Y* $\implies \Gamma^{time}_{X,Y}(q) < 1$

Idea why theorem holds

We have a heavy-tailed VAR model where X does not cause Y:

$$X_t = \sum_{i=0}^{\infty} a_i \varepsilon_{t-i}^X + \sum_{i=0}^{\infty} c_i \varepsilon_{t-i}^Y$$

$$Y_t = \sum_{i=0}^{\infty} b_i \varepsilon_{t-i}^Y.$$

Then,

$$\Gamma_{X,Y}^{time}(q) = \lim_{u \to 1^-} \mathbb{E}[\max\{F_Y(Y_0), \dots, F_Y(Y_q)\} \mid F_X(X_0) > u] < 1$$

if and only if

$$\lim_{u \to \infty} \mathbb{E}[f(\varepsilon_q^Y, \varepsilon_{q-1}^Y, \varepsilon_{q-2}^Y, \cdots) \mid \sum_{i=0}^{\infty} a_i \varepsilon_{-i}^X + \sum_{i=0}^{\infty} c_i \varepsilon_{-i}^Y > u] < 1.$$

We apply a principle of the single big jump.

Application of Theorem

Theorem

$(X, Y)^\top$ *follow either heavy-tailed VAR or heavy-tailed NAR model of order* q*. Let "extremal causal condition" hold. Then,*

- X *causes* $Y \implies \Gamma^{time}_{X,Y}(q) = 1$
- X *does not cause* $Y \implies \Gamma^{time}_{X,Y}(q) < 1$

We will apply this in the following.

Lag

Definition (Minimal lag)

Let $(X, Y)^\top$ follow stable $VAR(q)$ model, specified by

$$X_t = \alpha_1 X_{t-1} + \cdots + \alpha_q X_{t-q} + \gamma_1 Y_{t-1} + \cdots + \gamma_q Y_{t-q} + \varepsilon_t^X,$$
$$Y_t = \beta_1 Y_{t-1} + \cdots + \beta_q Y_{t-q} + \delta_1 X_{t-1} + \cdots + \delta_q X_{t-q} + \varepsilon_t^Y.$$

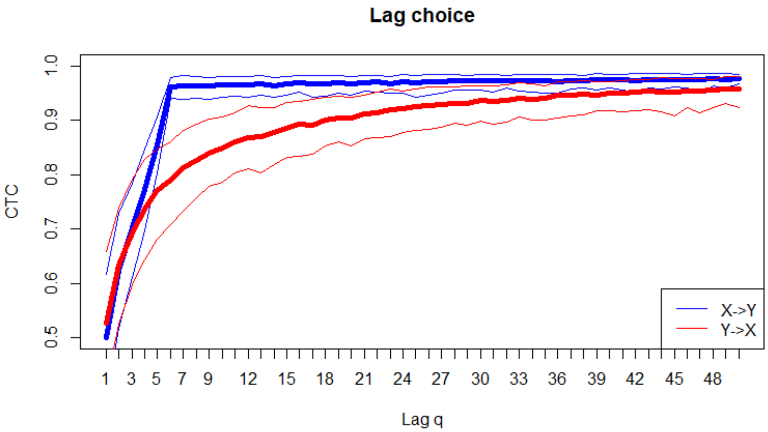
We call $p \in \mathbb{N}$ the *minimal lag*, if $\gamma_1 = \cdots = \gamma_{p-1} = \delta_1 = \cdots = \delta_{p-1} = 0$ and either $\delta_p \neq 0$ or $\gamma_p \neq 0$. If such p does not exist, we define the minimal lag as $+\infty$.

Lemma (Specification of our Theorem)

Let the assumptions from the previous theorem hold. Let p be the minimal lag. Then, $\Gamma^{time}_{X,Y}(r) < 1$ for all $r < p$, and $\Gamma^{time}_{X,Y}(r) = 1$ for all $r \geq p$.

How to estimate the minimal lag? Just take minimal p for which is $\Gamma^{time}_{X,Y}(p) = 1$.

Minimal lag



Lag choice
CTC
1.0
0.9
0.8
0.7
0.6
0.5
1
3
5
7
9
12
15
18
21
24
27
30
33
36
39
42
45
48
Lag q
X->Y
Y->X

Estimation

$$\Gamma^{time}_{X,Y}(q) = \lim_{u \to 1^-} \mathbb{E}[\max\{F_Y(Y_0), \ldots, F_Y(Y_q)\} \mid F_X(X_0) > u]$$

Definition

$$\hat{\Gamma}^{time}_{X,Y}(q) := \frac{1}{k} \sum_{i: X_i \geq \tau^X_k} \max\{\hat{F}_Y(Y_i), \ldots, \hat{F}_Y(Y_{i+q})\},$$

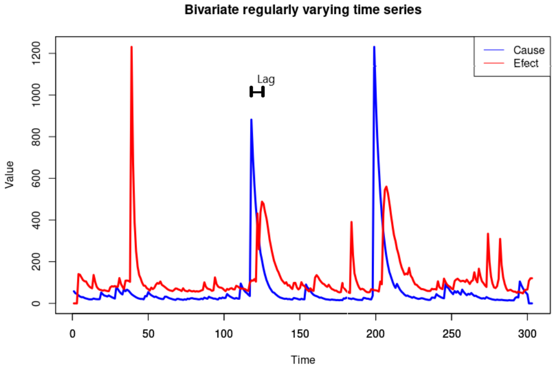
where $\tau^X_k = X_{(n-k+1)}$ is the $k-$ th largest value of X_i, and $\hat{F}_Y(Y_i) = \frac{1}{n} \sum_{j=1}^{n} 1\{Y_j \leq Y_i\}$.

Here,

$$k_n \to \infty, \frac{k_n}{n} \to 0, \text{ as } n \to \infty.$$

Bootstrap

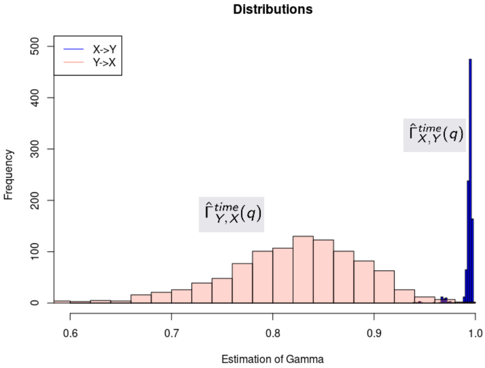
Distribution of $\hat{\Gamma}^{time}_{X,Y}(q)$ is hard to compute. We used block bootstrap method for computing confidence intervals [1].



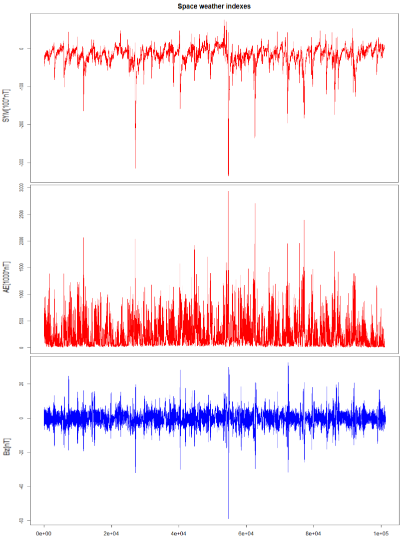
[1] More precisely so-called Reverse Bootstrap Percentile Interval.

Simulations

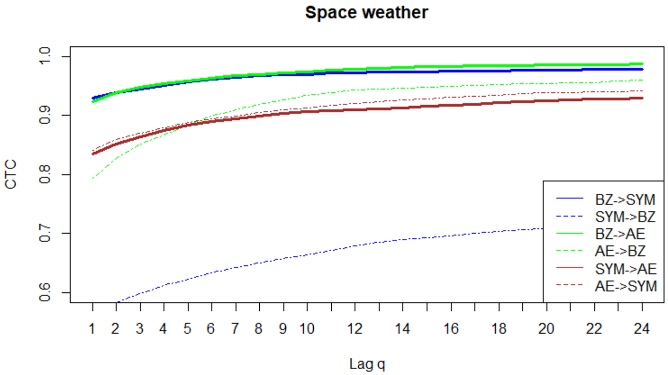
$$X_t = 0.5X_{t-1} + \varepsilon_t^X,$$
$$Y_t = 0.5Y_{t-1} + 0.5X_{t-2} + \varepsilon_t^Y.$$



Application



Application



Thank you!