University of Vienna
Faculty of Computer Science
RNDr. Katerina Schindlerová, CSc., Privatdoz.
Pranava Mummoju, MSc.

Vienna, November 12, 2024

Data Mining
WS 2024/25

Exercise Sheet 3: Causality

Submission until November 27, 2024, 23:59

Discussion: November 28, 2024

Exercise 3-1 *Granger causal test* (6 points)

Consider the following dataset from the R package "lmtest" to answer the age old question of what came first, "the chicken or the egg". The data was presented by Walter Thurman and Mark Fisher in the American Journal of Agricultural Economics, May 1988, titled "Chickens, Eggs, and Causality, or Which Came First?"

year	chicken (Y)	egg (X)
1930	468491	3581
1931	449743	3532
1932	436815	3327
1933	444523	3255
1934	433937	3156
1935	389958	3081

Tabelle 1: Population of chickens and eggs in U.S. egg production.

In this example, we will use only 6 entries of our data (1930 - 1935) and a single lag of 1.

a) **Auto-regression for X and Y** (1.5 point)

The auto regression problem can be written in the matrix format as:

$$\begin{pmatrix} Y(2) \\ \vdots \\ Y(n) \end{pmatrix} = \beta_0 \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} + \beta_1 \begin{pmatrix} Y(1) \\ \vdots \\ Y(n-1) \end{pmatrix} + \beta_2 \begin{pmatrix} X(1) \\ \vdots \\ X(n-1) \end{pmatrix} \tag{1}$$

The design matrix X for this case is:

$$\begin{pmatrix} 1 & Y(1) & X(1) \\ \vdots & \vdots & \vdots \\ 1 & Y(n-1) & X(n-1) \end{pmatrix} \tag{2}$$

Calculate the unknown coefficients $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)$, by using the derived formula, discussed in our lecture:

$$\hat{\boldsymbol{\beta}} = (\mathrm{X}'\mathrm{X})^{+}\mathrm{X}'\mathrm{Y}$$

where X' is the transpose of the matrix X and X^+ is the generalised inverse. Calculate $\|e\|_2^2$, the sum of squared residuals for the first regression which is our RSS_1.

b) **Auto-regression without Y** (1.5 point)
Repeat the steps from the previous task to calculate RSS_2 (i. e. the sum of squared residuals for the second regression), with the 2nd design matrix, which does not contain the egg feature, corresponding to the 2nd Granger equation.

c) **Apply statistical test** (1.5 point)
Apply the Granger Sargent test with $\alpha = 0.05$, to the computed RSS_1 and RSS_2 values, to test the Granger causal direction from eggs to chicken. We use a single lag of 1, i.e. $d = 1$. Interpret the result.

d) **Causal test for direction from chicken to eggs** (1.5 point)
Do exercises a - c analogously for the equations:

$$\begin{pmatrix} X(2) \\ \vdots \\ X(n) \end{pmatrix} = \beta_0 \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} + \beta_1 \begin{pmatrix} X(1) \\ \vdots \\ X(n-1) \end{pmatrix} + \beta_2 \begin{pmatrix} Y(1) \\ \vdots \\ Y(n-1) \end{pmatrix}$$

And interpret the results, what they mean for the answer ”what was first, eggs or chicken?”.

Exercise 3-2 *Multivariate Granger model* (1+1 points)

a) Consider graphical Granger model and causal inference by ordinary least squares with adaptive lasso penalty and regularization parameter $\lambda_n = n^{3/2}$. Is the problem solved by this model consistent? Explain your answer in detail.

b) What is the algorithm HMMLGA for and what are its hyperparameters?

Exercise 3-3 *Bivariate causal models on non-temporal data* (1+1 points)

a) Why is the causal relationship between X and Y by bivariate additive noise models for the linear-Gaussian case non-identifiable? Explain.

b) Recall the example 1 from lecture 4:

solar cell (cause) → generation of electricity (effect).

One can change $P(\text{cause})$ without affecting $P(\text{effect}|\text{cause})$: A change of intensity/angle of sun changes the power output of the cell, but not the conditional distribution of the effect given the cause. One can change $P(\text{effect}|\text{cause})$ without affecting $P(\text{cause})$: E.g. by using more efficient cells - while this changes the power output of cells, it does not affect the distribution of the incoming radiation.
To do the same in the anti-causal direction is hard: i.e. to find actions changing only P(effect) and not $P(\text{cause}|\text{effect})$ or vice versa, as due to their causal connection they are intrinsically (more) dependent on each other.

Now let us have a pair of variable X= age of a person and Y = diastolic blood presure, see e.g. pair 0040 in Mooij, J.M., Peters, J., Janzing, D., Zscheischler, J., Schölkopf, B. Distinguishing Cause from Effect Using Observational Data: Methods and Benchmarks. JMLR 17(32):1-102, 2016.

What is the causal relation between X and Y? How one could change $P(\text{effect}|\text{cause})$ without affecting $P(\text{cause})$ in the correct causal direction in this concrete case? Describe, how it would be in the anti-causal direction.