Supplementary Notes: Causal inference methods implementation and results

Introduction

Causal inference is one of the central challenges in statistics, machine learning, and data science. Understanding whether and how one (or more) variable influences another is fundamental for disciplines as diverse as biology, medicine, epidemiology, economics, and the social sciences [1, 2]. The aim of this study is to leverage longitudinal data to infer causal relationships. While randomized controlled trials (RCTs) are considered the gold standard for establishing causality, through randomization and control of confounding factors, they require a specific cause-effect hypothesis to be tested and an intervention to be applied on a subgroup of samples. In our cohort, we aim to apply a screen-inference approach, where we take advantage of the longitudinal design and omics data from an observational study (without any intervention) to systematically assess hundreds of variable and identify cause-effect relationships among them. To our knowledge, there are are only few approaches today that can handle longitudinal observational design or time series data [2]. Here we evaluate the cross-lagged panel models (CLPMs), a type of Structural Equation Model (SEM) designed for longitudinal data, to investigate whether changes in one variable at an earlier timepoint cause changes in another variable at a later timepoint. This framework allows us to evaluate potential temporal causality between quantitative variables where the relationship between two variables is linear. Method, simulations and results are described in detail in Supplementary Note 1. Another emerging challenge is detecting non-linear causality. Traditional methods of causal discovery, such as linear regression models, conditional independence tests, or structural equation modeling, have been widely applied and provide valuable insights. To test the causality between variables with non-linear relationships, we generalized this approach using CLPMs with the Generalized Additive Models (GAMs).

Supplementary Note 1. Cross-Lagged Panel Models

SN1.1. Simulation settings

We conducted a series of simulations with different causal structures and effect sizes. These simulations were designed to reflect realistic scenarios involving linear temporal relationships between variables under a) various sample sizes, b) presence of missing data c) influence of covariates in the model and d) presence of a third variables causally linked with the outcome. We evaluated models efficiency across multiple configurations to evaluate their ability to detect true directional effects (power), and use null simulation data sets to evaluate amount of false positive (Type 1 error) and false negatives (Type 2 error). Simulations were run for datasets of three sample sizes (n = 100, 300, and 500) and four time points; however, our simulation function is generalized to accommodate any number of samples and time points. For each setting, 200 repetitions were performed with effect sizes (beta) ranging from –2 to 2, to assess model performance under various conditions. The main simulated dataset was constructed so that variable X at time $t-1$ influences variable Y at time t, simulating a realistic temporal relationship.

We also incorporated different temporal trends in X, such as linear, quadratic, or other functional patterns. Details on simulations are given in the specific paragraphs.

SN1.2. Cross-Lagged Panel Model for linear causality

By structuring the repeated measures as a panel, we can test hypotheses such as whether hormone fluctuations predict subsequent changes in lipid levels or protein levels, providing insight into possible temporal dynamics even in the absence of experimental manipulation. We evaluated two different modeling strategies using linear regression, that can be put under the name of Cross-Lagged Panel Model:

1. Long format, to evaluate the overall influence of X on Y and exclude the vice versa:

$$\begin{cases} Y \sim X_{lag} + Y_{lag} \\ X \sim Y_{lag} + X_{lag} \end{cases}$$

thus, comparing every previous value of X with the next value of Y. We say that $X_{lag} \rightarrow Y$ if, in the first model, the coefficient associated with X_{lag} is significantly different from zero and, in the second model, Y_{lag} it's not. Since the data are simulated under the assumption that the true causal direction is $X_{lag} \rightarrow Y$, a significant effect of X_{lag} on Y is classified as a true positive (TP). Conversely, if this coefficient is not significant, the result is classified as a false negative (FN). In the second model, any significant effect of Y_{lag} on X represents a spurious reverse association. Therefore, if the coefficient of Y_{lag} is significantly different from zero, the result is classified as a false positive (FP), whereas a non-significant coefficient is classified as a true negative (TN) (Supplementary Table XX).

2. Wide format, to evaluate the influence of X in a single time point on Y in the next time point, then consider the non-independence of the observations, since they are coming from the same individuals. This is composed of three models combined together:

   a) model 1, comparing time point 1 and time point 2:

$$\begin{cases} Y_2 \sim X_1 + Y_1 \\ X_2 \sim Y_1 + X_1 \end{cases}$$

   b) model 2, comparing time point 2 and time point 3:

$$\begin{cases} Y_3 \sim X_2 + Y_2 \\ X_3 \sim Y_2 + X_2 \end{cases}$$

   c) model 3 comparing time point 3 and time point 4:

$$\begin{cases} Y_4 \sim X_3 + Y_3 \\ X_4 \sim Y_3 + X_3 \end{cases}$$

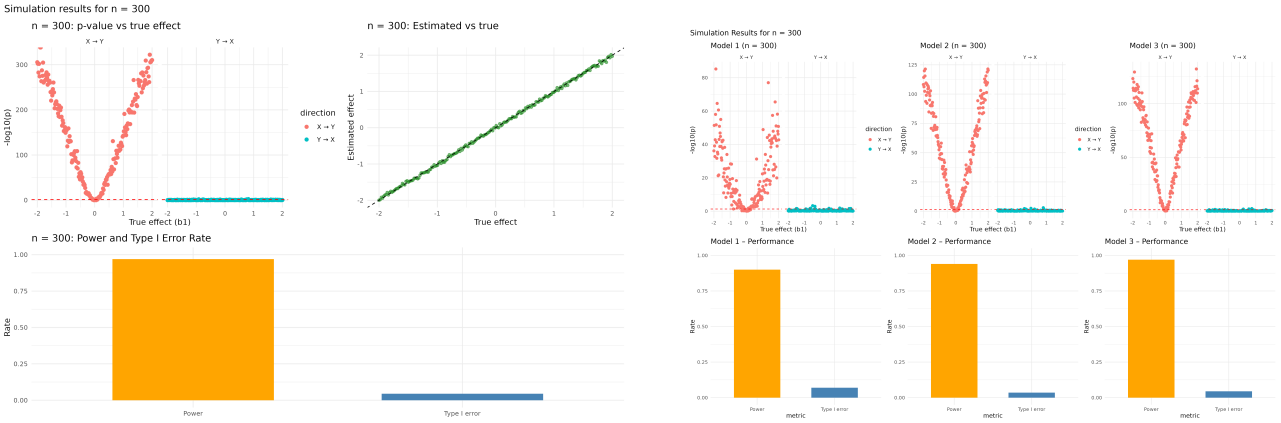
The causal relationships are defined as in the long model: $X_i \to Y_{i+1}$ when in $Y_{i+1} \sim X_i + Y_i$, the coefficient of X_i is significantly different from 0 and in $X_{i+1} \sim X_i + Y_i$ the coefficient of Y_i is not significantly different from 0. The wide-format approach can be interpreted as a time-specific implementation of the lagged cross-lagged panel model, thus keeping a similar definition for TP, FN, FP and TN.

SN1.2.1. Simulation 1 - main.

The main simulation scenario was carried out as follows:

$$\begin{cases} X_t = X_{t-1} + f(t) + \varepsilon \\ Y_t = Y_{t-1} + \beta X_{t-1} + \varepsilon \end{cases}$$

with $f(t) = a \cdot t$; where a is a constant number and t has integer values between 1 and 4, ε is drawn from a random normal distrubution, X is also random with mean 0 and sd equal to 1, β – the causal effect – is varying between -2 and 2. We observed that for values of $a = -1.2, -0.2, 0, 0.2, 1.2$, results are stable (data not shown). We therefore considered a fixed value of $a = 0.2$ for simulation results presented here. For each simulated scenario, we considered the causal effect to be significant if the p-value in that model was < 0.05. To compare estimated causal effect sizes with true values (β), we considered the estimates from all models, regardless of the p-value. The results with linear dependency between X and t are shown in Figure SN1.



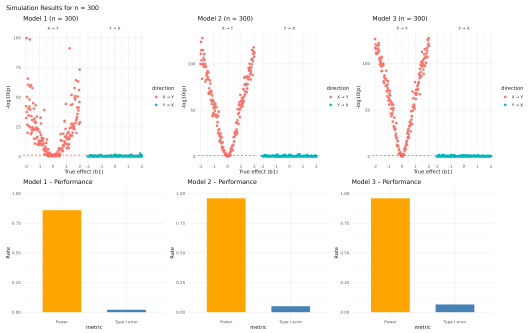
(a) Long format model

(b) Wide format models

Figure SN1: Results obtained when simulating a model with only variables X and Y, beta coefficients ranging from -2 to 2, sample size of 300, and a = 0.2 and f(t) linear. Dashed line indicates nominal p-value of 0.05. Panels: (a) long format model, (b) wide format models.

We also simulated a quadratic dependency between X and t, namely $f(t) = a \cdot t^2$, with the same values of a as above. We observed that for $|a| < 1$, the results for the wide model still show high power and low type 1 error rate. Instead, for the long model with absolute values of a more than 1, the type 1 error drastically increases. Therefore we fixed $|a| = 1.2$ for the simulation results presented here (Figure SN2).

(a) Long format model



(b) Wide format models

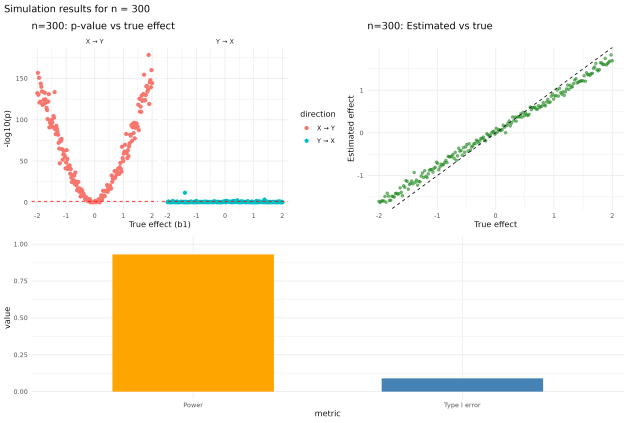
Figure SN2: Results obtained when simulating a model with only variables X and Y, beta coefficients ranging from -2 to 2, sample size of 500, and a = 1.2 and f(t) quadratic. Dashed line indicates nominal pvalue of 0.05. Panels: (a) long format model, (b) wide format models.

SN1.2.2. Simulation 2 - with covariates.

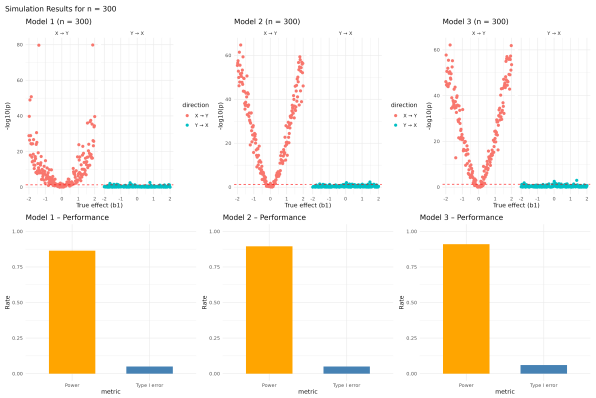
To evaluate the efficiency of the model in scenario where covariates influence the outcome, we simulated 200 datasets with sample size 100, 200 and 300, with the following model:

$$\begin{cases} X_t = X_{t-1} + f(t) + \varepsilon \\ Y_t = Y_{t-1} + \beta X_{t-1} + Z_t + \varepsilon \\ Z_t = \varepsilon \end{cases}$$

We tested the efficiency when ignoring the presence of Z, and when taking into account Z in the model. In the Figure SN3, we show results when ignoring Z for a data set of size 300.



(a) Long format model



(b) Wide format models

Figure SN3: Results obtained when simulating a model with variables X, Y and Z, beta coefficients ranging from -2 to 2, sample size of 300, and a = 0.2 and f(t) linear. The results in the figure are obtained when not considering Z in the causal inference test. Dashed line indicates nominal pvalue of 0.05. Panels: (a) long format model, (b) wide format models.

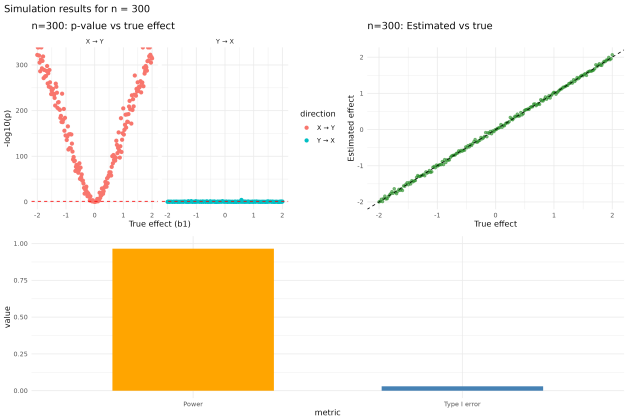
To account for Z in our statistical model, we used the following equations:

$$\begin{cases} Y \sim X_{lag} + Y_{lag} + Z \\ X \sim Y_{lag} + X_{lag} + Z \end{cases} \quad \begin{cases} Y_2 \sim X_1 + Y_1 + Z_2 \\ X_2 \sim Y_1 + X_1 + Z_2 \end{cases} \quad \begin{cases} Y_3 \sim X_2 + Y_2 + Z_3 \\ X_3 \sim Y_2 + X_2 + Z_3 \end{cases} \quad \begin{cases} Y_4 \sim X_3 + Y_3 + Z_4 \\ X_4 \sim Y_3 + X_3 + Z_4 \end{cases}$$

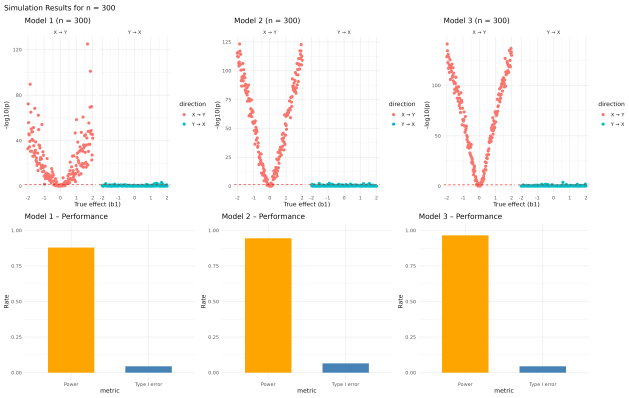
We observed that including the Z in the model (Figure SN4) the p-values are in general lower (thus significance is stronger) than not including Z (Figure SN3).

We then simulated a scenario where a covariate Z wasn't related to Y, simulating Z as a random variable as follows:

$$\begin{cases} X_t = X_{t-1} + f(t) + \varepsilon \\ Y_t = Y_{t-1} + \beta X_{t-1} + \varepsilon \\ Z_t = \varepsilon \end{cases}$$



(a) Long format model



(b) Wide format models

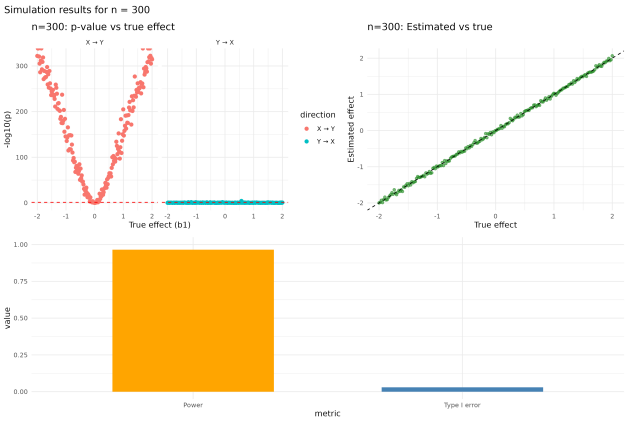
Figure SN4: Results obtained when simulating a model with variables X, Y and Z, beta coefficients ranging from -2 to 2, sample size of 300, and a = 0.2 and f(t) linear. The results in the figure are using a model testing Z. Dashed line indicates nominal pvalue of 0.05. Panels: (a) long format model, (b) wide format models.

Performing causal inference analyses with or without taking into account Z in the statistical framework did not affect model performance.

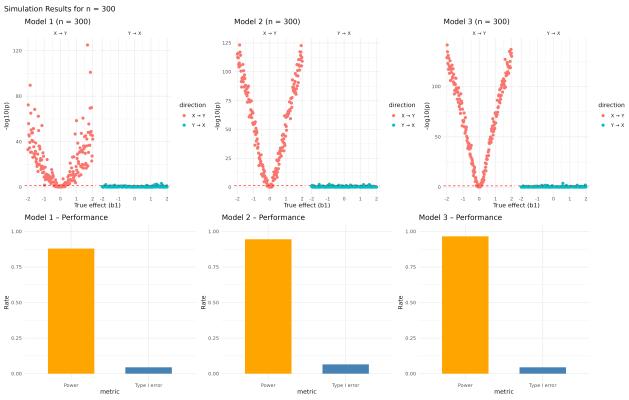
Finally, we simulated a scenario where Z has an effect on Y that is 100 times larger than X, with the following:

$$\begin{cases} X_t = X_{t-1} + f(t) + \varepsilon \\ Y_t = Y_{t-1} + \beta X_{t-1} + \varepsilon + 100 \cdot \beta Z_t \\ Z_t = \gamma Z_{t-1} + \delta f(t) + \varepsilon \end{cases}$$

We observe that a covariate with a magnitude 100 times larger than X has a negligible effect on the results in detecting the causality between X and Y (Figure SN5).



(a) Long format model

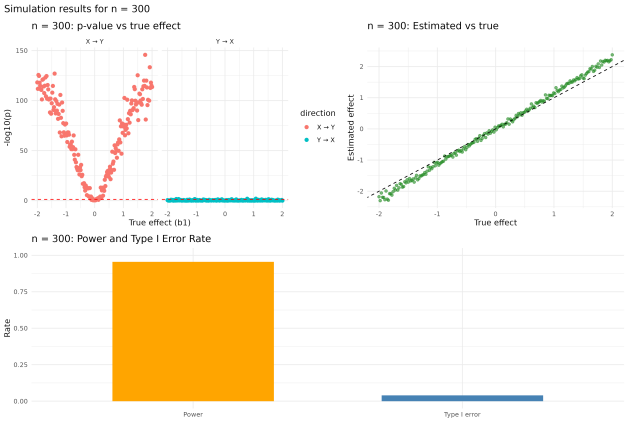


(b) Wide format models

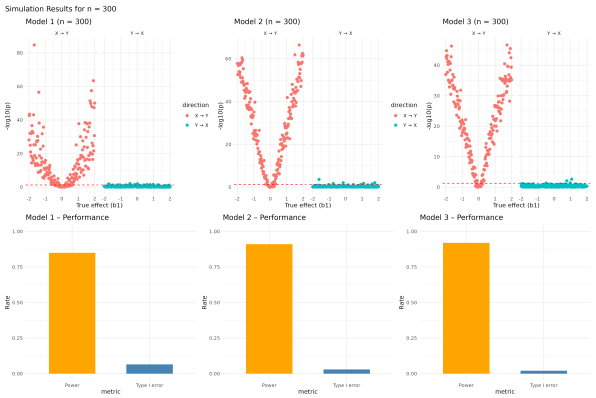
Figure SN5: Results obtained when simulating a model with variables X, Y and Z with a larger size ($Y = lag(Y) + \beta \cdot lag(X) + 100 \cdot \beta \cdot Z + \varepsilon$), beta coefficients ranging from -2 to 2, sample size of 300, and a = 0.2 and f(t) linear. The results in the figure are using a model testing Z. Dashed line indicates nominal pvalue of 0.05. Panels: (a) long format model, (b) wide format models.

SN1.2.3. Simulation 3 - with missing values

We simulated two types of missing values patterns: missing at random (MAR) and missing at not random (MNAR). The latter was simulated in a way that the more the study proceeds, the higher is the probability to find missing values, simulating 30% of NAs in the third visit and 40% in the fourth visit. This is a typical scenario for longitudinal data, as due to personal circumstances a volunteer may drop out may drop out with higher probability at the end of the observational period.



(a) Long format model

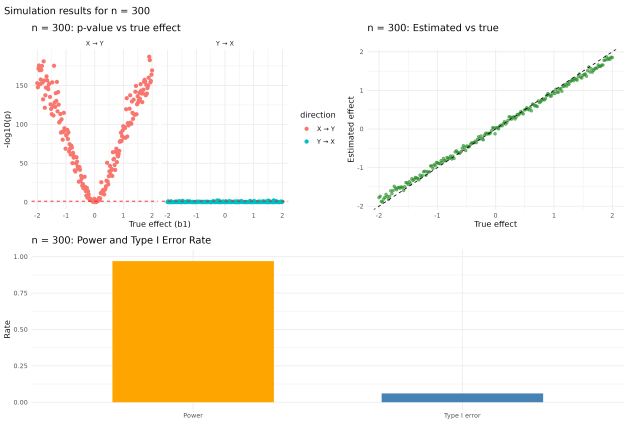


(b) Wide format models

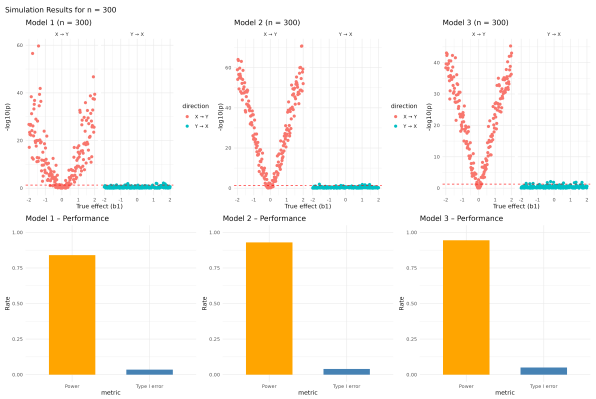
Figure SN6: Results obtained when simulating a model with variables X and Y, beta coefficients ranging from -2 to 2, sample size of 300, and a = 0.2 and f(t) linear. We simulated MNAR values and then we omitted them before fitting the model. Dashed line indicates nominal pvalue of 0.05. Panels: (a) long format model, (b) wide format models.

We then investigated the impact of these missing values using two different approaches, either omitting or imputing them. As imputation method, we used the linear interpolation, applied to each individual through time.

We present results for a scenario with an overall proportion of 30% missing values, distributed unevenly across visits according to the previously described MNAR mechanism (approximately 15%, 15%, 30%, and 40% missingness from the first to the fourth visit, respectively), with a sample size of 300, omitting them (Figure SN6) or imputing them (Figure SN7) before to apply the model .



(a) Long format model



(b) Wide format models

Figure SN7: Results obtained when simulating a model with variables X and Y, beta coefficients ranging from -2 to 2, sample size of 300, and a = 0.2 and f(t) linear. We simulated MNAR values and then we imputed them using linear interpolation before fitting the model. Dashed line indicates nominal pvalue of 0.05. Panels: (a) long format model, (b) wide format models.

We observed that the impact of omiting values slightly decreases power, while imputation slightly increases type 1 error rate. However, differences were negligible.

SN1.3. Cross-Lagged Panel Model for non-linear causality

Moreover, we tested the performance of the CLPMs using lagged linear regression for some simulated non-linear dataset using different types of function: cosine, quadratic, logarithmic, exponential, cubic. The simulations are as follows:

$$\begin{cases} X_t = X_{t-1} + f(t) + \varepsilon \\ Y_t = Y_{t-1} + g(X_{t-1}) + \varepsilon \end{cases}$$

where $f(t)$ is the function that determine the dependency of X on time and $g(X_{t-1})$ the retarded dependency of Y on X, with one of the functions specified above.

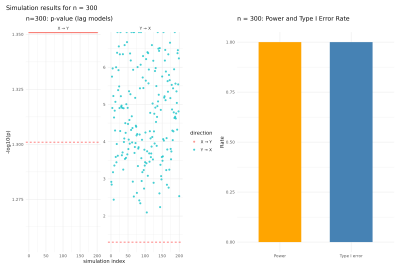
(a) Cos (b) Quadratic (c) Logarithmic

(d) Exponential (e) Cubic (f) Linear

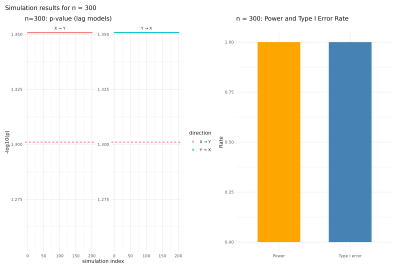
Figure SN8: Results obtained when simulating a model with variables X and Y, with non-linear relationships, sample size of 300, where X depends linearly on time with $a = 0.2$ and $f(t)$ linear. Dashed line indicates nominal pvalue of 0.05.

Figure SN8 clearly shows that for most functions, the method in the long format has extremely high type 1 error rate, while the wide model has very limited power on quadratic and logarithm functions.

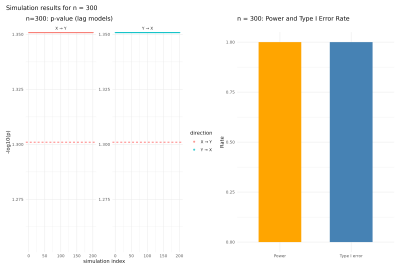
Then to generalize our algorithm on non-linear data we decided to test the performance of the cross-lagged Generalized Additive Models (GAMs), simply using GAMs instead of LM in the test. Furthermore, as shown in Figure SN9, when the X depends on time, the long models give a high type 1 error, while the wide models are working properly.



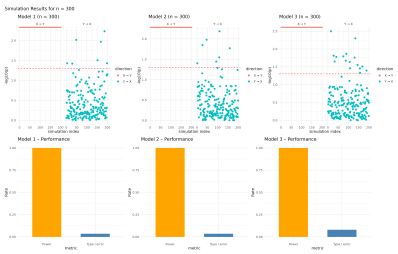
(a) Cosine function with GAMs



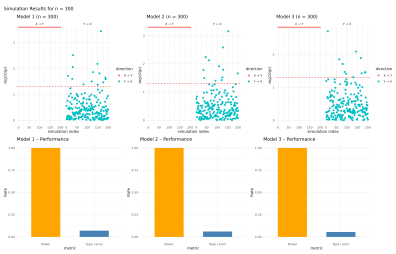
(b) Quadratic function with GAMs



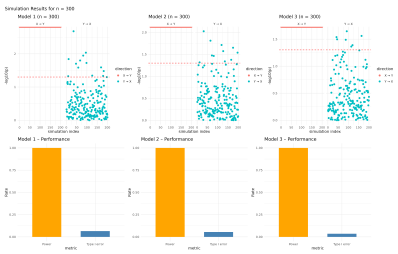
(c) Logarithmic function with GAMs



(d) Exponential function with GAMs



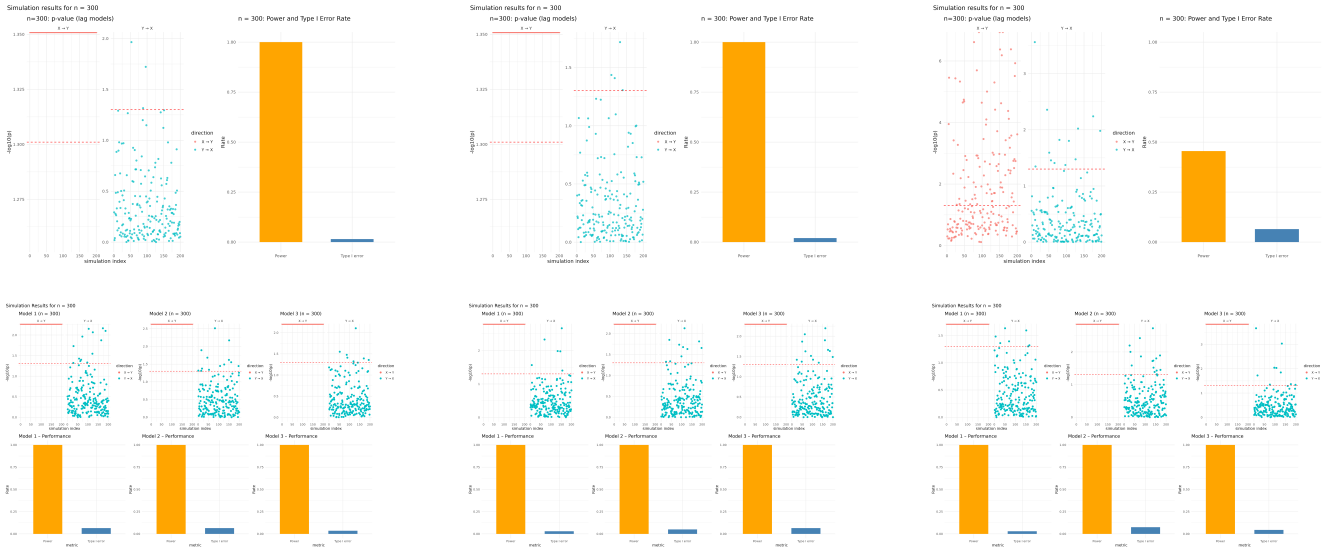
(e) Cubic function with GAMs



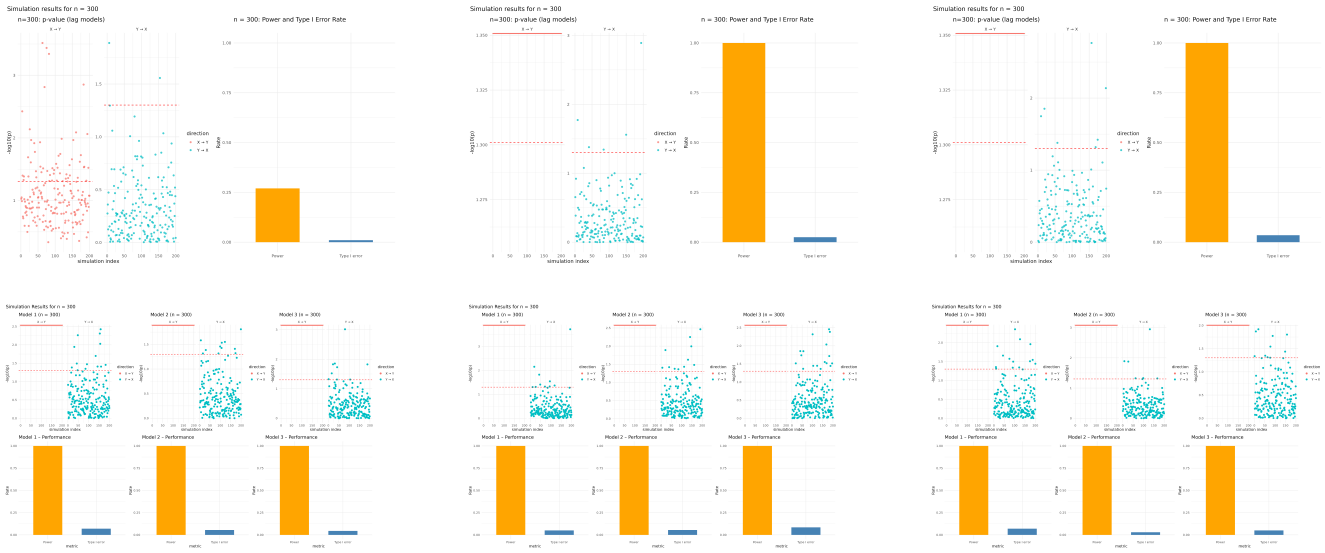
(f) Linear function with GAMs

Figure SN9: Results obtained when simulating a model with variables X and Y, with non-linear relationships and X depends linearly from time with $a = 0.2$ and $f(t)$ linear, sample size of 300, and tested with model GAM. Dashed line indicates nominal pvalue of 0.05.

In contrast, when X is randomly generated and does not depends on time, the results have low proportion of type 1 error and high power (Figure SN10). However, the wide models still give more reliable results for all the functions tested.



(a) Cosine function with GAMs (b) Linear function with GAMs (c) Quadratic function with GAMs



(d) Logarithmic function with GAMs (e) Exponential function with GAMs (f) Cubic function with GAMs

Figure SN10: Simulated model with variables X and Y, with non-linear relationships and X not time-dependent, sample size of 300, and tested with model GAM.

Conclusion

The linear regression CLPMs with wide format data is an efficient approach to assess temporal causal inference for linear relationship. Furthermore, the regression can be generalized to non-linear relationships using GAMs while maintaining proper performance.

References

1. Pearl, J. *Causality: Models, Reasoning, and Inference* (Cambridge University Press, 2009).
2. Hernán, M. A. & Robins, J. M. *Causal Inference: What If* `https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/` (Chapman & Hall/CRC, Boca Raton, 2020).