Testing Granger Non-Causality in Panels with Cross-Sectional Dependencies

Lenon Minorics
Amazon Research
minorics@amazon.com

Caner Turkmen
Amazon Research
atturkm@amazon.com

David Kernert
Amazon Reserach
davdkern@amazon.com

Patrick Bloebaum
Amazon Research
bloebp@amazon.com

Laurent Callot
Amazon Reserach
lcallot@amazon.com

Dominik Janzing
Amazon Research
janzind@amazon.com

Abstract

This paper proposes a new approach for testing Granger non-causality on panel data. Instead of aggregating panel member statistics, we aggregate their corresponding p-values and show that the resulting p-value approximately bounds the type I error by the chosen significance level even if the panel members are dependent. We compare our approach against the most widely used Granger causality algorithm on panel data and show that our approach yields lower FDR at the same power for large sample sizes and panels with cross-sectional dependencies. Finally, we examine COVID-19 data about confirmed cases and deaths measured in countries/regions worldwide and show that our approach is able to discover the true causal relation between confirmed cases and deaths while state-of-the-art approaches fail.

1 INTRODUCTION

Within the last decade, there has been growing awareness that causal inference can improve scientific research in many disciplines as interpretability and robustness become increasingly important (Doshi-Velez and Kim, 2017; Roscher et al., 2020; Marcinkevičs and Vogt, 2020; Moraffah et al., 2020). Causality is a crucial factor for gaining insights into the decision process of algorithms, which has many use cases such as avoiding bias and discrimination (Mehrabi et al., 2019), improving user experience (Zhou and Fu, 2007) and gathering biological insights (Angermueller et al., 2016).

If the causal relation between variables is known, causality can be used to study the interaction between statistical units such as estimating the average effect of treatments (Imbens and Rubin, 2015; Holland, 1986), analyze their mediation (Berzuini et al., 2012), detect the root causes of anomalies (Janzing et al., 2019) or quantifying the causal influence of variables in a system (Janzing et al., 2013, 2020). However, knowing the causal relations between variables of interest *a priori* is important for such applications. Unfortunately, these causal relations can often only be recovered from observational data under strong assumptions, even if hidden common causes (*i.e.*, confounders) do not exist (Peters et al., 2017, Proposition 4.1). This leads to a seemingly unsolvable problem if further information about the data generating process is unavailable. For i.i.d. data, one way to mitigate this ambiguity is to impose additional assumptions about the generative process, such as linear non-Gaussian noise (LiNGAM) (Shimizu et al., 2006) or non-linear additive noise (Hoyer et al., 2008). For time series data, however, consequential additional information about the data generating process is fortunately available due to the time order. That is, the present cannot causally influence the past, and hence the causal order is known. In many cases, this information is sufficient to recover the causal direction among variables in the form of time series (Peters et al., 2017, Section 10.3.3).

A well-known approach to causal discovery in time series data was developed by Granger (Granger, 1969, 1980, 2003). Granger-causality is based on a simple definition: time series $\{X_t\}$ is said to Granger-cause

time series $\{Y_t\}$ if the past of $\{X_t\}$ improves the prediction of $\{Y_t\}$ given its own past. The presence of Granger-causality implies conditional dependence between the "present" Y_t and the past of X_t, given the past of Y_t. By assuming there are no hidden common causes, we can deduce from Reichenbach's principle of common causes (Reichenbach, 1999) and the known causal order, that $\{X_t\}$ causally influences $\{Y_t\}$. In this definition of Granger-causality, we restricted to the bi-variate case for simplicity, however note that Granger defined the concept for general multivariate settings.

Although many works have considered new approaches in time series causal discovery (Shajarisales et al., 2015; Peters et al., 2013; Hyvärinen et al., 2010; Kawahara et al., 2011), procedures based on Granger-causality are still state-of-the-art due to their simplicity and favorable applicability in practice (Berzuini et al., 2012; Bressler and Seth, 2011). That is, as stated in Bressler and Seth (2011) Chapter 22 Section 2.3, Granger Causality does not rely on specific assumptions about the data generating process (except the exclusion of instantaneous effects, if instantaneous effects exist, additional assumptions have to be made) and is therefore particularly convenient for empirical investigations.

Several researchers considered an extension of classical causal discovery in time series to panel data (Kónya, 2006; Dumitrescu and Hurlin, 2011; Holtz-Eakin et al., 1988; Juodis et al., 2021). Further, Arkhangelsky and Imbens (2021); Arkhangelsky et al. (2021); Athey et al. (2019) studied the potential outcome of treatments in panel data regimes. See also Hsiao (2003) for general analysis on panel data.

In the panel data setting, multiple variables are observed for the same members of a so-called *panel* across time. Concretely, suppose we are given tuples of time series (X_i, Y_i), $i = 1, \ldots, N$ for some $N \in \mathbb{N}$ where we succinctly refer to $X_i := \{X_{i,t}\}_{t=1}^T$. Moreover, assume that only one direction of causal influence is possible. That is, there exists $H \subseteq \{1, \ldots, N\}$ such that X_i causally influences Y_i for all $i \in H$ but there exist no $j \in \{1, \ldots, N\}$ such that Y_j causally influences X_j. In this setting, we refer to the set of tuples $\{(X_i, Y_i), i = 1, \ldots, N\}$ as (X, Y) and we say that X causally influences Y. For testing the opposite direction, we can simply interchange the role of X and Y. A simple example could be whether there is a causal link between confirmed cases and deaths resulting from COVID-19 infection in different countries. Here, we are interested in the general relation of the tuple (COVID-19 confirmed cases, COVID-19 deaths) across different members of the panel (*i.e.*, countries), see Section 5.

The paper is organized as follows. In Section 2 we give a concise exploration of the different approaches to Granger causality on panel data. Section 3 is dedicated to the problem set up and proposed method, where we also provide the main theorem of the paper. In Section 4, we present experimental results for synthetic datasets. Finally, in Section 5, we compare the results of state-of-the-art causal discovery algorithms on panel data against our approach on COVID-19 data about confirmed cases and deaths.

2 RELATED WORK

In this section, we explore the existing literature related to our work. First, we summarize different existing approaches to Granger causality on panel data. We then introduce the Dumitrescu-Hurlin test, the most popular existing Granger causality test for panel data. Finally, we summarize an existing approach to aggregating p-values for high-dimensional regression, which provides a core component our method builds upon.

2.1 Hypothesis Testing of Granger Causality on Panel Data

Existing literature distinguishes between four types of hypothesis tests for Granger-causality on panel data, *cf.* Dumitrescu and Hurlin (2011). Below, we formulate these tests for the causal influence from X to Y in terms of their null hypotheses.

Homogeneous Non-Causality (HNC): The null hypothesis of HNC test is that there is no causal relation between the variables for any individual, *i.e.*, for all i it holds that X_i does not Granger-cause Y_i.

Homogeneous Causality (HC): Under the null, X_i causes Y_i for all i. Further, it is assumed that the dynamics of X_i and Y_i do not change for different i. In particular, it is assumed that the regression parameters from X_i to Y_i given the past of Y_i are identical for all individuals.

Heterogeneous Causality (HEC): The null is the same as for HC except that the test does not assume that the dynamics of X_i and Y_i remain the same across the panel.

Heterogeneous Non-Causality (HENC): The null hypothesis is that there exists a subgroup of individuals for which X_i does Granger-cause Y_i and, hence, there exists at least 1 and at most $N-1$ individuals for which X_i does not Granger cause Y_i.

All approaches described above are based on the fol-

lowing assumption of the dynamics of X and Y:

$$X_{i,t} = \delta_{i,0} + \sum_{p=1}^{P} \delta_{i,p} X_{i,t-p} + \eta_{i,t},$$

$$Y_{i,t} = \theta_{i,0} + \sum_{p=1}^{P} \theta_{i,p} Y_{i,t-p} + \sum_{p=1}^{P} \beta_{i,p} X_{i,t-p} + \epsilon_{i,t}, \quad (1)$$

where P is the time lag order, and δ, η and β are coefficients. The concrete assumptions about the coefficient vectors δ, η and β depend on the considered hypotheses HNC, HC, HEC or HENC, but in all approaches it is assumed that the innovation processes $(\{\eta_{i,t}\}, \{\epsilon_{i,t}\})_{i,t}$ are mutually independent.

2.2 Dumitrescu–Hurlin Test

The most widely used test of Granger-causality in panel data is the DH test, developed by Dumitrescu and Hurlin (2011). The test considers the HNC null hypothesis, where no Granger-causal relationships are assumed to exist for any member i of the panel. The DH test is based on an aggregated Wald statistic of individual Granger causality tests.

There exist asymptotic and semi-asymptotic characterizations of the DH test. More precisely, as stated in the introduction, we denote by (X, Y) the panels under consideration. The DH test considers the generation process (1) where it is allowed that coefficients vary across individuals, while being time invariant. Since the DH test does not assume homogeneous coefficients across individuals, the null and alternative hypotheses read as follows:

$$\begin{aligned} H_0 &: (\beta_{i,1}, \dots, \beta_{i,p}) = (0, \dots, 0) \quad \text{for all } i \\ H_1 &: \exists\, i \ \text{ s.t. } (\beta_{i,1}, \dots, \beta_{i,p}) \neq (0, \dots, 0). \end{aligned} \quad (2)$$

To test this hypothesis, Dumitrescu and Hurlin (2011) first construct a Wald statistic $W_{i,T}$ for each individual. The *average Wald statistic* $W_{N,T}^{Hnc}$ is then given by

$$W_{N,T}^{Hnc} = \frac{1}{N} \sum_{i=1}^{N} W_{i,T}.$$

Under suitable conditions and, in particular, using the independence between individuals (since $(\{\eta_{i,t}\}_{i,t}, \{\epsilon_{i,t}\}_{i,t})$ are mutually independent), Dumitrescu and Hurlin (2011) Theorem 2 shows that the normalized average Wald statistic converges in distribution to the normal distribution with expectation 0 and variance 1. We include further details in Appendix A.

The proof relies on the central limit theorem and hence, the independence between innovation processes is required (see Assumption 2 of Dumitrescu and Hurlin

(2011)). This implies that dependencies across panel members are assumed to not exist which is a rather restrictive assumption, since it prohibits any interaction among the individual panel members. In practice, however, it is often the case that such interactions exist. In reference to our example above, the confirmed COVID-19 cases of different countries may be causally linked, *e.g.*, because infected people might travel to different countries and infect locals.

In Section 3, we will take a different approach to develop a hypothesis test where interactions between the innovation processes are taken into account. This approach will rely on a p-value aggregation idea for high dimensional regression. Although that approach is not connected to panel data, the idea can be used in panel data settings to obtain a p-value that controls the type I error by the chosen significance level even if dependencies between individuals (*i.e.*, between the innovation processes) exist.

Dumitrescu and Hurlin also suggest an alternative approach in Dumitrescu and Hurlin (2011) Section 6.2 that takes cross-sectional dependencies into account, which we call *DH block bootstrap test* in the following. As stated by the authors, the suggested algorithm results in very high computational costs; therefore new panel non-causality tests should be developed to account for cross section dependencies. For completeness, we include an explanation of the procedure in Appendix B and also compare our approach to the DH block bootstrap test in the experiments.

2.3 A P-Value Aggregation Method From High-Dimensional Regression

In this section, we describe the key concepts that we use for our test procedure, which is based on the work of Meinhausen et al. (2009), see also Dezeure et al. (2015). As mentioned in the previous section, this concept is not connected to panel data. Instead, it considers multiple bootstrap runs in regression, generates p-values for every bootstrap sample and aggregates them such that the corresponding test statistic (of the considered test) bounds the type I error.

More precisely, Meinhausen et al. (2009) consider the following setup. Let $\boldsymbol{Z}$ be an n-dimensional response vector and $\boldsymbol{W}$ a $n \times k$ dimensional design matrix such that

$$\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{b} + \boldsymbol{\tau},$$

with $\boldsymbol{\tau}$ being an i.i.d. n-dimensional random vector with $\tau_i \sim \mathcal{N}(0, \sigma^2)$ for some $\sigma^2 > 0$ and $\boldsymbol{b} \in \mathbb{R}^k$. Meinhausen et al. (2009) consider the problem of finding all j such that $b_j > 0$. In terms of statistical significance, Meinhausen et al. (2009) aim to assign p-values for the

null hypotheses

$$H_{0,j} : b_j = 0,$$

where they observe n samples (z_i, w_i), $i = 1, \ldots, n$. Further, the paper assumes a high dimensional setting, *i.e.*, $k \gg n$, where statistical inference is challenging. In order to alleviate this issue, Wasserman and Roeder (2009) proposes to split the data into two parts. The first part is used for feature selection where important variables are kept with high probability. The second half of the data is used to assign p-values to the kept features by using classical least squares estimation (the p-values for the dropped features can, for instance, be set to 1). Under some weak conditions, this procedure results in an approximately correct p-value.

However, Meinhausen et al. (2009) argue that this procedure relies on an arbitrary split of the n samples and hence, the results can vary significantly making the test hard to reproduce. This problem can be solved by a multi-splitting approach in which the procedure of Wasserman and Roeder (2009) is repeated m times. To understand the key idea of the multi-split approach of Meinhausen et al. (2009), we ignore the fact that we want to find all regressors for which $b_j > 0$, but rather focus on a fixed j for which we want to test $H_{0,j}$. In the multi-splitting approach, m p-values $P_j^{(1)}, \ldots P_j^{(m)}$ are generated, where each $P_j^{(l)}$, $l = 1, \ldots, m$ corresponding to each of the m splits, is an approximately correct p-value for the test $H_{0,j}$. In the next step, these p-values are aggregated to a single, approximately correct, p-value P_j. In contrast to the single split method of Wasserman and Roeder (2009), the generated p-value of Meinhausen et al. (2009) is stable and as a result, makes the experiment reproducible. To aggregate the m p-values $P_j^{(1)}, \ldots, P_j^{(m)}$, Meinhausen et al. (2009) proposes the following procedure:
For arbitrary $\gamma \in (0, 1)$, define

$$Q_j(\gamma) := \min \Big\{ 1, \\ \text{emp. } \gamma\text{-quantile}(\{P_j^{(l)}/\gamma, \; l = 1, \ldots, m\}) \Big\}. \quad (3)$$

Then, $P_j := Q_j(\gamma)$ is an asymptotically correct p-value, *i.e.*, using $Q_j(\gamma)$, the type one error of the test $H_{0,j}$ is approximately bounded by the chosen significance level $\alpha \in (0, 1)$ for each $\gamma \in (0, 1)$, see Meinhausen et al. (2009) Theorem 3.1.

Above, we focused on the case where we are interested in a single predictor for simplicity. However, this approach can easily be generalized to the case where we want to test all predictors. Corresponding FWER/FDR control procedures are also explained in Meinhausen et al. (2009).

3 TESTING FOR GRANGER NON-CAUSALITY ON PANEL DATA

3.1 Problem Setup

In this paper, we consider a test for the HNC hypothesis, introduced in the previous section. That is, we assume that under the null X_i does not Granger-cause Y_i for any i. As in the previous section, we consider the bivariate case for ease of notation but this can be easily generalized. We consider the same generation process as in (1):

$$
\begin{aligned}
X_{i,t} &= \delta_{i,0} + \sum_{p=1}^{P} \delta_{i,p} X_{i,t-p} + \eta_{i,t}, \\
Y_{i,t} &= \theta_{i,0} + \sum_{p=1}^{P} \theta_{i,p} Y_{i,t-p} + \sum_{p=1}^{P} \beta_{i,p} X_{i,t-p} + \epsilon_{i,t},
\end{aligned}
\quad (4)
$$

for $i = 1, \ldots, N$ denoting the cross-sectional units (*i.e.*, individuals or panel members) and $t = 1, \ldots, T$ the timestamps. We also assume that the stochastic processes Y_i, X_i are scalar. The parameters θ_i denote the heterogeneous, autoregressive coefficients of the processes. The stochastic processes $\epsilon_{i,t}$ and $\epsilon_{j,t}$ as well as $\eta_{i,t}$ and $\eta_{j,t}$ are not assumed to be independent in $i \neq j$, *i.e.*, cross-sectional dependencies might exist. Note that in contrast to the DH-test, we do not rely on an independence assumption across panel members. Figure 1 illustrates different types of dependencies between members of the panel and shows how this dependence is reflected in the dependence between the innovation processes. Our test relies on the same hypotheses test as the DH-test, namely (2). However, our algorithm is specifically designed for settings where the causal relation is the same for all individuals while the generation processes can be different, *i.e.*, if $\beta_i = (\beta_{i,1}, \ldots, \beta_{i,P}) \neq (0, \ldots, 0)$ for one i, then this holds for all i, but not necessarily $\beta_i = \beta_j$ for $i \neq j$. If $\beta_i = (\beta_{i,1}, \ldots, \beta_{i,P}) = (0, \ldots, 0)$ for one i, then this holds for all i. The corresponding hypothesis test reads

H_0 : X does not Granger-cause Y, *i.e.*,
$$\beta_i = (\beta_{i,1}, \ldots, \beta_{i,P}) = (0, \ldots, 0) \quad \text{for all } i \quad (5)$$

H_1 : X does Granger-cause Y, *i.e.*,
$$\beta_i = (\beta_{i,1}, \ldots, \beta_{i,P}) \neq (0, \ldots, 0) \quad \text{for all } i. \quad (6)$$

Note that X and Y can be exchanged in this test and hence, we can test whether X Granger-causes Y or Y Granger-causes X.

Compared to the DH-test described in Section 2.2, the hypothesis test (5) vs (6) is more restrictive, since we assume that all members of the panel share the
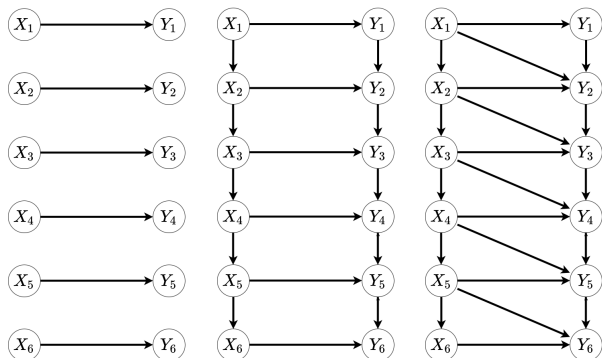
Figure 1: *Example of the summary graph (omitting self-cycles) of panel data where X causally influences Y.*

Left: *Panel data where no cross-sectional dependencies exist,* i.e., *the innovation processes are independent, namely $\epsilon_i, \eta_i, \epsilon_j, \eta_j$ are mutually independent for $j \neq i$. This structure satisfies the assumptions about the dependence between members for the DH-test and our QPPA approach (see Section 3.2).*

Middle: *Panel data with cross-sectional dependencies in the sense that ϵ_i and ϵ_j as well as η_i and η_j are dependent but ϵ_i and η_j are independent for $j \neq i$. This structure **does not** satisfy the assumptions about the panel structure for the DH-test as cross-sectional dependencies exist. However, this structure **does** satisfy the assumptions about the panel structure for the QPPA approach. Further, for QPPA, it is not required that the cross-sectional dependencies are acyclic, there might be cycles in the causal dependencies between $\{X_i : i \in N\}$ as well as $\{Y_i : i \in N\}$ and it would still satisfy the assumptions of the QPPA approach.*

Right: *Panel data where cross-sectional dependencies exist where other panel members are also confounders of X_i and Y_i. For the innovation processes this means that ϵ_i and ϵ_j as well as η_i and η_j are dependent **and** η_j and ϵ_i are dependent for all i, j as well. This structure **does not** satisfy the assumptions about the panel structure for the DH-test and our QPPA approach, seeing that X_{i-1} would be a confounder of X_i and Y_i.*

Note that this figure does not show any relations between the innovation processes in time. In general, for all settings we assume that the innovation processes are independent in time, i.e., for all $t \neq s \neq h \neq l$, $\epsilon_{i,t}$, $\eta_{i,s}$, $\epsilon_{j,h}$, $\eta_{j,l}$ are mutually independent (for all i, j).

same causes, whereas the DH-test does not rely on such a strong assumption. That is, in the generation process of the DH-test under the alternative there exists *at least* one member of the panel where $\beta_i \neq 0$. In contrast, under (6) it holds that for all i, $\beta_i \neq 0$. However, this assumption is realistic in many cases, as, for instance, in the setting of our experiments on COVID-19 data. For more details and explanations, see Section 5. Note, however, that the test procedure we will introduce can still guarantee type-I error control under the test scenario (2). We show in Appendix H that our test is robust when causal connections are sporadically missing for some individuals, and that the difference between the two tests is negligible in practice when our assumption of uniformity in the existence of causal relations nearly holds.

3.2 Quantile p-value Panel Adjustment (QPPA)

In this section, we describe the procedure to test (5) vs (6) using the idea described in Section 2.3. Instead of using an aggregated Wald statistic (see Section 2.2), we propose a procedure that is inspired by Meinhausen et al. (2009) (Section 2.3). That is, Meinhausen et al. (2009) uses the aggregation method (3) to aggregate p-values of different bootstrap samples. Translating this idea into our setting, we can aggregate the p-values of the panel members, hence, we treat individual panel members in the same way Meinhausen et al. (2009) treats different bootstrap runs. The aggregation will then control the type-I error asymptotically by the chosen significance level. Since we calculate a Granger non-causality p-value for every member, we further need the following technical assumptions that we carry over from Dumitrescu and Hurlin (2011), see also (Granger, 1969).

Time i.i.d. residuals: For each fixed $i \in \{1, \ldots, N\}$ $\epsilon_{i,t}$ are independent for all $t = 1, \ldots, T$ and normally distributed with $\mathbb{E}(\epsilon_{i,t}) = 0$ and $\mathbb{E}(\epsilon_{i,t}^2) < \infty$, where $\mathbb{E}(\epsilon_{i,t}^2)$ is assumed to be constant in t.

Covariance stationarity: For all i and t, it holds that $X_{i,t}$ and $Y_{i,t}$ have finite variance and $\mathbb{E}(X_{i,t}X_{j,t+h})$, $\mathbb{E}(Y_{i,t}Y_{j,t+h})$, $\mathbb{E}(Y_{i,t}X_{j,t+h})$, $\mathbb{E}(Y_{i,t})$ and $\mathbb{E}(X_{i,t})$ do not depend on t.

Under these assumptions, we can introduce our test procedure, called **Quantile p-value Panel Adjustment** (QPPA) in two steps.

Step 1: Compute a p-value for every member of the panel. The first step is the same as for the DH-test: We apply Granger Non-causality to each individual panel member, where we use a Wald-statistic to

test for the presence of Granger causality. Corresponding to these Wald statistics, we obtain an asymptotically correct p-value $p_{X \to Y}^{(i)}$ for each panel member. For instance, an f-statistic (which belongs to the family of Wald-statistics) can be used to calculate the corresponding p-values. Note that the construction of the f-statistic requires the assumption of time i.i.d. residuals as well as covariance stationarity. Further, in order for the p-values to be (asymptotically) correct, $\eta_{j,t}$ and $\epsilon_{i,s}$ need to be independent for all t, s, j and i. This requirement is necessary in general and does not impose additional restrictions to the generation process compared to the existing literature for Granger causality (see Figure 1 right for an example where this assumption is violated).

Step 2: Aggregate p-values. Similar to the procedure in Section 2.3, we aggregate the computed p-values as follows. For $\gamma \in (0,1)$, we define

$$Q_{X \to Y}(\gamma) := \min \left\{ 1, \right.$$
$$\left. \text{emp. } \gamma\text{-quantile} \left\{ p_{X \to Y}^{(i)}/\gamma; \ i = 1, \ldots, N \right\} \right\},$$
$$(7)$$

We can now formulate the main theorem of our work, the proof of which can be found in Appendix D.

Theorem 1. *Assume the generation procedure* (4) *and let* $\alpha, \gamma \in (0,1)$. *Then,* $Q_{X \to Y}(\gamma)$ *is an asymptotically correct p-value,* i.e.,

$$\limsup_{T \to \infty} \mathbb{P}\left(Q_{X \to Y}(\gamma) \leq \alpha \right) \leq \alpha,$$

where T *denotes the number of timestamps.*

With this theorem, we obtain type-I error control for arbitrary $\gamma \in (0,1)$. Note that the proof does not require any restrictions on the dependence between p-values. In particular, the individuals do not need to be independent, *i.e.*, restrictions on the relation between $\epsilon_{i,t}$ and $\epsilon_{j,t}$ as well as $\eta_{i,t}$ and $\eta_{j,t}$ are not required.

Although Theorem 1 holds for arbitrary γ, choosing the right γ could be difficult. While Meinhausen et al. (2009) recommend $\gamma = 0.5$ as a practical choice, they also construct another p-value which only requires to specify a lower bound for γ. This approach can also be translated into the panel data setting. We include the corresponding procedure in Appendix C.

Note that, although our approach is based on the procedure of Meinhausen et al. (2009) which deals with high dimensional statistics, we do not consider a high dimensional setup here. The connection between Meinhausen et al. (2009) and our approach is the problem

that a single p-value results in an unstable procedure which is hence hard to reproduce.

Throughout this paper, we only consider the bivariate case for simplicity but this can easily be generalized and corresponding FWER/FDR procedures based on Meinhausen et al. (2009) can be constructed. Moreover, since $Q_{X \to Y}(\gamma)$ is an asymptotically correct p-value, also classical FWER/FDR control procedures (e.g. Bonferoni or Benjamini-Hochberg) can be applied in the multivariate case.

Before moving to numerical experiments, we briefly discuss some implications of our assumptions and potential limitations of our approach. Both covariance stationarity and residuals that are i.i.d. in time are standard assumptions in Granger causality tests. Although covariance stationarity is a strong assumption, different data preprocessing procedures could be applied to obtain stationarity in practice (Hyndman and Athanasopoulos, 2018), also see Appendix G. For example, we employ such methods to make the COVID-19 time series stationary in our study in Section 5. Similarly, assuming residuals are i.i.d. or Gaussian are strong assumptions. Our experiments on COVID-19 data and further experiments in Appendix H examine the robustness of our algorithm against the violation of this assumption. Finally, the generation process (4) has two further implications: assuming ϵ and η are independent implicitly assumes causal sufficiency (*i.e.*, no hidden confounders), and that there are no *instantaneous effects* between X_t and Y_t. Again, both assumptions are typical in Granger causality analysis. We discuss how to decrease the false discovery rate in the presence of confounding, subject to causal faithfulness, in Appendix E and the robustness of our approach under instantaneous effects in Appendix F.

4 EXPERIMENTS WITH SYNTHETIC DATA

In this section, we present experimental results on synthetic data. We compare the DH test, DH block bootstrap method and our approach (QPPA). To compare these methods, we use the existing `xtg-cause` package developed by Lopez and Weber (2017), which includes the DH-test and the DH block bootstrap test. For the QPPA approach presented in Section 3, we use our own implementation where we also rely on the Granger causality test implementation in `statsmodels` (Seabold and Perktold, 2010) to compute f-statistics and corresponding p-values as explained in Section 3.2. We compare these approaches in two scenarios, respectively without and with cross-sectional dependencies, detailed below.

Experiment 1: For the first experiment, we do not insert cross-sectional dependencies. The generation process, an autoregressive process of order 1, is specified as follows:

$$X_{i,t} = \delta_{i,1} X_{i,t-1} + \eta_{i,t}$$
$$Y_{i,t} = \theta_{i,1} Y_{i,t-1} + \beta_{i,1} X_{i,t-1} + \epsilon_{i,t},$$

where the innovation processes are i.i.d. Gaussian random variables with $\eta_{i,t}, \epsilon_{i,t} \sim N(0, 0.1)$ and we draw the parameters from a uniform distribution: $\delta_{i,1}, \theta_{i,1} \sim Unif(0.2, 0.8)$.

Experiment 2: For this experiment we insert cross-sectional dependencies for X and Y:

$$X_{i,t} = \delta_{i,1} X_{i,t-1} + \zeta_{i,t}$$
$$Y_{i,t} = \theta_{i,1} Y_{i,t-1} + \beta_{i,1} X_{i,t-1} + \xi_{i,t},$$

where $\boldsymbol{\zeta}_t := (\zeta_{1,t}, \dots \zeta_{N,t}) \sim N(\mathbf{0}, \Sigma)$, $\boldsymbol{\xi}_t := (\xi_{1,t}, \dots \xi_{N,t}) \sim N(\mathbf{0}, \tilde{\Sigma})$, $\Sigma = A^T A$, $\tilde{\Sigma} = \tilde{A}^T \tilde{A}$ and $A, \tilde{A}$ are random vectors where each entry is sampled from $Unif(0.5, 1.5)$. Finally, $\delta_{i,1}, \theta_{i,1} \sim Unif(0.2, 0.8)$ as in Experiment 1. The cross-sectional dependency is embedded in the multivariate normal distribution of the noise terms $\boldsymbol{\zeta}_t$ and $\boldsymbol{\xi}_t$. Note however that there is no dependence between $\boldsymbol{\zeta}_t$ and $\boldsymbol{\xi}_t$ and also no dependence of the noise terms in time.

In both experiments we either sample $\beta_{i,1}$ from $Unif(0.2, 0.8)$ if the null should be rejected and set $\beta_{i,1} = 0$ if the null should be accepted.

We report power and false discovery rates (FDR) for DH-test, DH-test with block bootstrap (`DH-test-bb`), and QPPA in Tables 1 and 2 for experiments 1 and 2 respectively. Each number reported is an average of 100 experiments. For QPPA we use $\gamma = 0.5$ in (7) and for the DH-test and DH-test with block bootstrap we use the statistic $\tilde{Z}_N^{HNC}$, see Appendices B and A.

Table 1 shows that QPPA performs equally well as DH and DH-bb in the setting without cross-sectional dependencies with a sufficiently large history (T>10). The results in Table 2 exhibit that if cross-sectional dependencies exist, the FDR of the DH-test increases significantly above the significance level 0.05 even in the large sample regime. Also DH-test-bb shows higher FDR than 0.05 even in the large sample regime. This is not the case for QPPA which remains robust against this type of dependency. Moreover, the power of both tests is 1 in the large sample regime. For low sample sizes, QPPA is rather conservative (for the choice $\gamma = 0.5$), hence it has low power and low FDR for $T = 10$. The opposite holds for the DH-test. Note that for all approaches we set $\alpha = 0.05$. However, in the experiments, we observe lower FDR than the chosen significance level for QPPA. This is due to the

Table 1: *Empirical results for Experiment 1 (no cross-sectional dependencies)*

		QPPA		DH-test		DH-test-bb	
		Power	FDR	Power	FDR	Power	FDR
T=10	N=1	0.15	0.211	0.35	0.186	0.33	0.154
	N=10	0.000	0.000	0.93	0.212	0.74	0.119
	N=30	0.000	0.000	1.0	0.359	0.97	0.110
T=50	N=1	0.77	0.038	0.83	0.117	0.83	0.126
	N=10	0.98	0.000	1.0	0.074	1.0	0.074
	N=30	1.0	0.000	1.0	0.048	1.0	0.083
T=100	N=1	0.91	0.022	0.94	0.078	0.97	0.093
	N=10	1.0	0.000	1.0	0.065	1.0	0.074
	N=30	1.0	0.000	1.0	0.074	1.0	0.107

Table 2: *Empirical results for Experiment 2 (with cross-sectional dependencies)*

		QPPA		DH-test		DH-test-bb	
		Power	FDR	Power	FDR	Power	FDR
T=10	N=1	0.170	0.320	0.420	0.236	0.340	0.261
	N=10	0.120	0.000	0.630	0.344	0.320	0.220
	N=30	0.160	0.059	0.810	0.449	0.380	0.191
T=50	N=1	0.800	0.048	0.850	0.086	0.850	0.105
	N=10	0.920	0.000	1.0	0.180	0.990	0.083
	N=30	0.960	0.000	1.0	0.408	0.990	0.075
T=100	N=1	0.940	0.051	0.970	0.049	0.980	0.110
	N=10	0.970	0.010	1.0	0.174	1.0	0.074
	N=30	1.0	0.000	1.0	0.419	1.0	0.082

fact that the p-values of QPPA are not uniformly distributed but α is only an upper bound for the false discovery rate, see Theorem 1. While we heuristically set $\gamma = 0.5$ for results reported in Figures 1 and 2, we show in Figures 2 and 3 how power and FDR vary for $\gamma = 0.01, 0.02, \dots, 0.99$. The Figure shows that in these experiments, the power decreases with increasing γ. However, the FDR is 0 for all γ due to the well separation between the distribution under the null and alternative. We add further experiments in Appendix
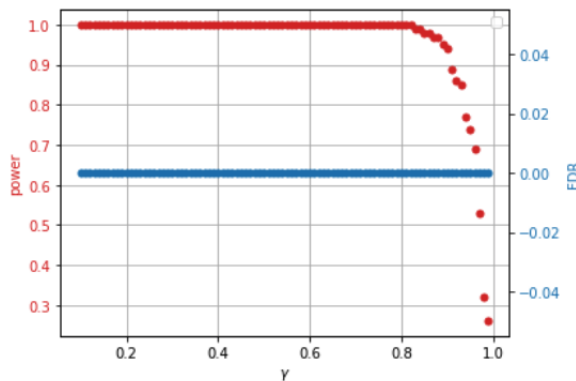

Figure 2: *Empirical results in the experimental setup 1 for QPPA for $\gamma = 0.01, 0.02 \dots 0.99$, where we set $T = 100$ and $N = 30$.*

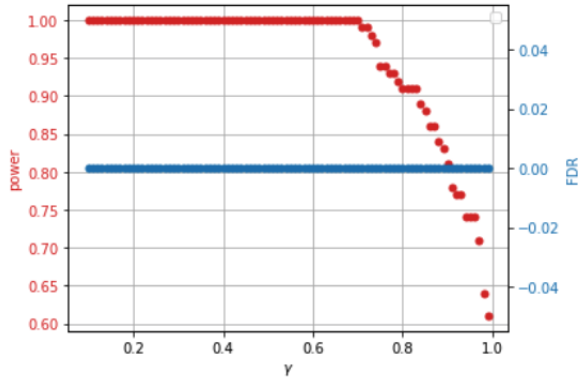H, where we also include additional results about the behavior of QPPA for varying γ.

Figure 3: *Empirical results in the experimental setup 2 for QPPA for $\gamma = 0.01, 0.02 \ldots 0.99$, where we set $T = 100$ and $N = 30$.*

5 EXPERIMENTS WITH COVID-19 DATA

While synthetic data experiments confirm our test performs in line with expectations, we carry out a second set of experiments to demonstrate both the robustness of our approach on real-world problems, and how the problem we address arises ubiquitously. To this end, we give an illustrative study of causal discovery in time series related to COVID-19. Here, we regard individual countries and regions as panel members, and we analyze the causal relation between confirmed cases and deaths within each country. Since increasing number of confirmed cases lead to more deaths (after several days to approximately two weeks), we regard

$$\text{confirmed cases} \longrightarrow \text{deaths}$$

as the ground truth causal relation, and analyze causal influences from the past two weeks to the present.

Some remarks on the ground truth causal relation: Although one might argue that deaths can causally influence confirmed cases because more vulnerable immune systems reduce as the number of deaths increase, we expect such effects to take place over longer periods of time that should not occur in the scope of our study. Further, if a huge proportion of death cases did not get tested before death, death cases can causally influence confirmed cases because died people might get tested post mortem. However, confirmed cases is a proxy for the number of infections and, additionally, COVID-19 deaths where people are not going to the hospital and die without confirmation but get tested post mortem are rather rare and should not be statistically significant.

As one might expect, cross-sectional dependencies among countries should exist, *e.g.*, through active cases

traveling between countries (see Appendix I.2 for an analysis confirming this hypothesis).

We use the Johns Hopkins CSSE COVID-19 data repository[1] (Dong et al., 2020). The data set contains confirmed cases and deaths due to COVID-19, collected globally in 280 countries or regions, where cases are recorded daily between 22nd January 2020 and 4th October 2021. After applying preprocessing steps outlined in Appendix I.1, there remain 335 days of data for 225 panel members (*i.e.*, countries or regions). In Table 3 we give p-values of the DH-test, DH-test with block bootstrap and QPPA for the causal influence from confirmed cases to deaths and visa versa, where we allow a time order in the underlying models to allow influence for up to two weeks. The results show that the DH-test

Table 3: *Results of the covid-19 causal discovery study. p-val QPPA relates to the p-values obtained by our QPPA approach with $\gamma = 0.5$ in (7), p-val DH-test to the p-values obtained by the DH-test and p-val DH-test-bb to the p-values obtained by the DH-test with block bootstrap, where we use 20 breps, see Appendix B and the statistics $\tilde{Z}_N^{HNC}$, see Appendix A. $c \rightarrow d$ is the corresponding p-value to the causal link "confirmed cases causes deaths" and $d \rightarrow c$ the p-value to the causal link "deaths causes confirmed cases".*

P (=lag order)	p-val QPPA		p-val DH-test		p-val DH-test-bb	
	c -> d	d -> c	c -> d	d -> c	c -> d	d -> c
1	0.610	0.607	0.000	0.000	0.000	0.000
2	0.323	0.343	0.000	0.000	0.000	0.000
3	0.183	0.239	0.000	0.000	0.000	0.000
4	0.091	0.133	0.000	0.000	0.000	0.000
5	0.055	0.094	0.000	0.000	0.000	0.000
6	0.036	0.084	0.000	0.000	0.000	0.000
7	0.015	0.110	0.000	0.000	0.000	0.000
8	0.005	0.064	0.000	0.000	0.000	0.000
9	0.003	0.082	0.000	0.000	0.000	0.000
10	0.002	0.080	0.000	0.000	0.000	0.000
11	0.001	0.065	0.000	0.000	0.000	0.000
12	0.001	0.056	0.000	0.000	0.000	0.000

and DH-test with block bootstrap reject the null for both directions with p-value of 0.000 for all lags, leading to a wrong conclusion (confirming that deaths also cause confirmed cases). In contrast, QPPA rejects the null only for the correct direction (confirmed cases $\rightarrow$ deaths) to the significance level of 5% after a reasonable time order is given. Further, it is reasonable that the null from confirmed cases to deaths gets only rejected after including multiple lags, because infection (more precisely confirmation of infection that is recorded in the data under consideration) with COVID-19 causes death with some time delay. Figure 4 shows power and FDR of QPPA for $\gamma = 0.01, 0.02, \ldots, 0.99$. Compared

---

[1]The data set can be downloaded at https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data/csse_covid_19_time_series, from this repository we used the global csv files

to the corresponding study in Section 4 shown in Figure 2 and 3, we can see that the FDR and power for small γ is large and monotonically decreasing with increasing γ. Hence, choosing γ amounts to a trade-off between high power and low FDR. We include more results of the behavior of QPPA for varying γ in Appendix I.4 where we use different numbers of countries/regions per run.

We conclude from our results that failing to account for cross-sectional dependencies can easily result in wrong conclusions using baseline methods, and that QPPA not only mitigates this risk but is also robust against real-world cases where some of its assumptions about the data generating process (*e.g.*, Gaussianity of innovations) do not hold.
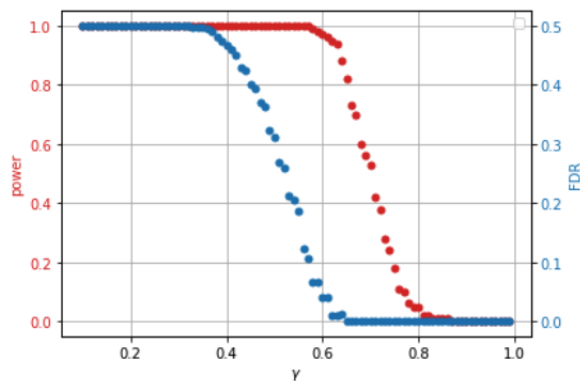
Figure 4: *Empirical results for COVID-19 data about confirmed cases and deaths using QPPA with $\gamma = 0.01, 0.02, \ldots, 0.99$. To calculate power and FDR, we randomly selected 60 countries/regions out of the 225 and checked whether QPPA detects the causal relation $c \to d$ and $d \to c$ respectively to the significance level 5% and repeat this 100 times.*

In general, if one wants to recover the unknown causal relations of a panel time series dataset, we recommend to apply QPPA for varying γ. If the results are consistent for a large interval of γ, there is strong evidence to believe that the obtained causal structure for these γ captures the relations of the underlying generative process.

6   CONCLUSION AND OUTLOOK

In this paper, we propose a new approach to causal discovery on panel data with Granger causality using a quantile based p-value adjustment approach. We calculate a Wald statistic for each individual and correspondingly generate an asymptotically correct aggregated p-value for the panel. Compared to the most widely used causal discovery method on panel data, the DH-test, we aggregate p-values instead of the individual Wald statistics. In this way, we are able to account for cross-sectional dependencies that may exist among individuals in the panel.

Numerical experiments on synthetic data confirm that our approach outperforms both the DH-test and its block bootstrap variant, designed to address cross-sectional dependencies. Notably, in contrast to baseline methods, our approach also correctly discovers causal relationships in a real world scenario – the cause-effect relationship between COVID-19 infections and deaths. Our results show that not capturing cross-sectional dependencies easily leads to incorrect conclusions, and also that our method is robust in real-world settings where some of its statistical assumptions may be violated.

Finally, we emphasize that the p-value aggregation method we employ is a general approach and does not depend on the specific generation process. Therefore, this method could be applied in other applications or causal discovery methods for panel data (*e.g.*, on non-time series data), which remains an exciting avenue for further research.

References

C. Angermueller, T. Pärnamaa, L. Parts, and O. Stegle. Deep learning for computational biology. *Molecular systems biology*, 12(7):878, 2016.

D. Arkhangelsky and G. W. Imbens. Double-robust identification for causal panel data models. *arXiv preprint arXiv:1909.09412v2*, 2021.

D. Arkhangelsky, G. W. Imbens, L. Lei, and X. Luo. Double-robust two-way-fixed-effects regression for panel data. *arXiv preprint arXiv:2107.13737v1*, 2021.

S. Athey, M. Bayati, I. G. W., and Z. Qu. Ensemble methods for causal effects in panel data settings. *AEA Papers and Proceedings*, 109:65–70, 2019.

C. Berzuini, P. Dawid, and L. Bernardinelli. *Causality: statistical perspective and applications*. Wiley, 2012.

J. Breitung. The local power of some unit root tests for panel data. *Advances in Econometrics*, pages 161–178, 2000.

J. Breitung and S. Das. Panel unit root tests under cross-sectional dependence. *Statistica Neerlandica*, pages 414–433, 2005.

S. L. Bressler and A. K. Seth. Wiener-granger causality: A well established methodology. *NeuroImage*, 58(2), 2011.

I. Choi. Unit root tests for panel data. *Journal of International Money and Finance*, pages 249–272, 2001.

R. E. De Hoyos and V. Sarafidis. Testing for cross-sectional dependence in panel-data models. *Stata Journal*, 6(4):482–496(15), 2006. URL //<?echo(www)?>.stata-journal.com/article.html?article=st0113.

R. Dezeure, P. Bühlmann, L. Meier, and N. Meinhausen. High-dimensional inference: Confidence intervals, p-values and r-software hdi. *Institute of Mathematical Statistics*, 30(4):533–558, 2015.

D. A. Dickey and W. A. Fuller. Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Associationt*, 74: 427–431, 1979.

E. Dong, H. Du, and L. Gardner. An interactive web-based dashboard to track covid-19 in real time. *The Lancet*, 2020.

F. Doshi-Velez and B. Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

E.-I. Dumitrescu and C. Hurlin. Testing for granger non-causality in heterogeneous panels. *Elsevier*, pages 1450–1460, 2011.

R. F. Engle and W. J. Granger. Co-integration and error correction: Representation, estimation and testing. *Econometrica*, 55(2):251–276, 1987.

E. W. Frees. Assessing cross-sectional correlation in panel data. *Journal of Econometrics*, 1995.

M. Friedman. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, 1937.

C. W. J. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3):424–438, 1969.

C. W. J. Granger. Testing for causality. *Journal of Economic Dynamics and Control*, 2:329–352, 1980.

C. W. J. Granger. Some aspects of causal relationships. *Journal of Econometrics*, 112:69–71, 2003.

W. J. Granger. Some properties of time series data and their use in econometric model specification. *Journal of Econometrics*, pages 121–130, 1981.

K. Hadri. Testing for stationarity in heterogeneous oanel data. *Econometrics Journal*, pages 148–161, 2000.

R. D. F. Harris and E. Tzavalis. Inference for unit roots in dynamic panels where the time dimension is fixed. *Journal of Econometrics*, pages 201–226, 1999.

P. W. Holland. Statistics and causal inference. *J. Amer. Statist. Assoc.*, 81(396):945–960, 1986.

D. Holtz-Eakin, W. Newey, and H. S. Rosen. Estimating vector autoregressions with panel data. *Econometrica*, pages 799–809, 1988.

P. Hoyer, D. Janzing, J. M. Mooij, J. Peters, and B. Schölkopf. Nonlinear causal discovery with additive noise models. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems*, volume 21. Curran Associates, Inc., 2008. URL https://proceedings.neurips.cc/paper/2008/file/f7664060cc52bc6f3d620bcedc94a4b6-Paper.pdf.

C. Hsiao. Analysis of panel data. *Cambridge University Press*, 2003.

R. J. Hyndman and G. Athanasopoulos. *Forecasting: principles and practice.* Melbourne, Australia. OTexts.com/fpp2. Accessed on Aug. 20, 2021, 2 edition, 2018.

A. Hyvärinen, K. Zhang, S. Shimizu, and P. O. Hoyer. Estimation of a structural vector autoregression model using non-gaussianity. *Journal of Machine Learning Research*, 11:1709–1731, 2010.

K. S. Im, M. H. Pesaran, and Y. Shin. Testing for unit roots in heterogeneous panels. *Journal of Econometrics*, pages 53–74, 2003.

G. W. Imbens and D. B. Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction.* Cambridge University Press, 2015.

D. Janzing, D. Balduzzi, M. Grosse-Wentrup, and B. Schölkopf. Quantifying causal influence. *The Annals of Statistics*, (5):2324–2358, 2013.

D. Janzing, K. Budhathoki, L. Minorics, and P. Blöbaum. Causal structure based root cause analysis of outliers. *arXiv preprint arXiv:1912.02724*, 2019.

D. Janzing, P. Bloebaum, L. Minorics, and P. Faller. Quantifying causal contributions via structure preserving interventions. *arXiv preprint arXiv:2007.00714*, 2020.

A. Juodis, Y. Karavias, and V. Sarafidis. A homogeneous approach to testing for granger non-causality in heterogeneous panels. *Empirical Economics*, pages 93–112, 2021.

Y. Kawahara, S. Shimizu, and T. Washio. Analyzing relationships among arma processes based on non-gaussianity of external influences. *Neurocomputing*, 74(12-13):2212–2221, 2011.

L. Kónya. Export and growth: Granger causality analysis on oecd countries with a panel data approach. *Econometrica*, 56:1371–1396, 2006.

A. Levin, C.-F. Lin, and C.-S. J. Chi. Unit root tests in panel data: Asymptotic and finite-sample properties. *Journal of Econometrics*, pages 1–24, 2002.

L. Lopez and S. Weber. Testing for granger causality in panel data. *The Stata Journal*, 17(4):972–984, 2017.

R. Marcinkevičs and J. E. Vogt. Interpretability and explainability: A machine learning zoo mini-tour. *arXiv preprint arXiv:2012.01805*, 2020.

A. A. Mastakouri, B. Schölkopf, and D. Janzing. Necessary and sufficient conditions for causal feature selection in time series with latent common causes. In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 7502–7511. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/mastakouri21a.html.

N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan. A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635*, 2019.

N. Meinhausen, L. Meier, and P. Bühlmann. p-values for high dimensional regression. *Journal of the American Statistical Association*, 104:1671–1681, 2009.

R. Moraffah, M. Karami, R. Guo, A. Raglin, and H. Liu. Causal interpretability for machine learning-problems, methods and evaluation. *ACM SIGKDD Explorations Newsletter*, 22(1):18–33, 2020.

M. H. Pesaran. General diagnostic tests for cross section dependence in panels. *University of Cambridge, Faculty of Economics, Cambridge Working Papers in Economics*, 2004.

J. Peters, D. Janzing, and B. Schölkopf. Causal inference on time series using restricted structural equation models. *Proceedings of the 26th International Conference on Neural Information Processing Systems*, 1:154–162, 2013.

J. Peters, D. Janzing, and B. Schölkopf. *Elements of Causal Inference*. MIT Press, 2017.

H. Reichenbach. *The Direction of Time*. Dover Publications Inc., 1999.

R. Roscher, B. Bohn, M. F. Duarte, and J. Garcke. Explainable machine learning for scientific insights and discoveries. *IEEE Access*, 8:42200–42216, 2020.

S. Seabold and J. Perktold. Statsmodels: Econometric and statistical modeling with python. In *Proceedings of the 9th Python in Science Conference*, volume 57, page 61. Austin, TX, 2010.

N. Shajarisales, D. Janzing, B. Schölkopf, and M. Besserve. Telling cause from effect in deterministic linear dynamical systems. *Proceedings of the 32nd International Conference on Machine Learning*, 37:285–294, 2015.

S. Shimizu, P. O. Hoyer, A. Hyvärinen, and A. Kerminen. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7.10, 2006.

L. Wasserman and K. Roeder. High dimensional variable selection. *Annals of Statistics*, 37(5A):2178–2201, 2009.

H. Zhou and X. Fu. Understanding, measuring, and designing user experience: The causal relationship between the aesthetic quality of products and user affect. In *International Conference on Human-Computer Interaction*, pages 340–349. Springer, 2007.

Supplementary Material: Testing Granger Non-Causality in Panels with Cross-Sectional Dependencies

A FURTHER EXPLANATIONS OF THE DH TEST

Corresponding to the average Wald-static

$$W_{N,T} := \frac{1}{N} \sum_{i=1}^{N} W_{i,T},$$

Dumitrescu and Hurlin (2011) show that for the normalized statistics under some regularity conditions it holds

$$Z_{N,T}^{Hnc} := \sqrt{\frac{N}{2P}}(W_{N,T} - P) \overset{T,N \to \infty}{\longrightarrow} N(0,1)$$

$$\tilde{Z}_{N}^{Hnc} := \frac{\sqrt{N}\left[W_{N,T}^{Hnc} - N^{-1}\sum_{i=1}^{N} \mathbb{E}(W_{i,T})\right]}{\sqrt{N^{-1}\sum_{i=1}^{N} Var(W_{i,T})}} \overset{N \to \infty}{\longrightarrow} N(0,1)$$

(8)

in distribution, where $N(0,1)$ denotes the normal distribution with expectation 0 and variance 1 (where P denotes the lag order, T the number of timestamps and N the number of individuals). Eq. (8) can now be used to test (2): If the probability of obtaining the realization of $\tilde{Z}_{N}^{Hnc}$, $Z_{N,T}^{Hnc}$ respectively w.r.t. the standard normal distribution is low (corresponding to the chosen significance level), H_0 is rejected. Juodis et al. (2021) constructed a different test statistic in the same setting with the additional benefit that, in contrast to the DH-test, it accounts for "Nickell" bias which occurs if $N/T^2 \to 0$ does not hold.

B DH BLOCK BOOTSTRAP TEST

The DH block bootstrap procedure relies on a resampling idea, *cf.* Dumitrescu and Hurlin (2011) Section 6.2.:

1. Define the model for each panel member to test Granger causality. Here, the model under consideration is

$$Y_{i,t} = \theta_{i,0} + \sum_{p=1}^{P} \theta_{i,p} Y_{i,t-p} + \sum_{p=1}^{P} \beta_{i,p} X_{i,t-p},$$

   see (1).

2. Estimate the model and compute the corresponding test statistics $Z_{N,T}^{Hnc}, \tilde{Z}_{N}^{Hnc}$ for each panel member.

3. Estimate the model under the null (no Granger causality, see 2), *i.e.*, estimate a model

$$Y_{i,t} = \tilde{\theta}_{i,0} + \sum_{p=1}^{P} \tilde{\theta}_{i,p} Y_{i,t-p}$$

   for each panel member and compute the residual vectors of size $(T, 1)$.

4. Resample the residuals with replacement for each panel member with "blocks" of size 1, if we want to take dependencies across time into account, we could increase the block size.

5. Construct resampled time series $\tilde{Y}_{i,t}$ under the null:

$$\tilde{Y}_{i,t} = \tilde{\theta}_{i,0} + \sum_{p=1}^{P} \tilde{\theta}_{i,p}\tilde{Y}_{i,t-p} + \tilde{\epsilon}_{i,t}$$

where $(\tilde{\epsilon}_{i,t})_t$ denotes the resampled noise of the i-th panel member.

6. Estimate the model defined in step 1 for the resampled time series $\{\tilde{Y}_{i,t}\}_t$ and construct the statistics of step 2 for this model and resampled time series.

7. Repeat steps 5 and 6 a large amount of times.

8. Compare the test statistics of step 2 against the test statistics obtained from steps 5 - 7.

Since this procedure relies on resampling, we create dependencies especially in low sample regimes. Further, it relies on the generation of a new dataset in step 5 which could be not robust to violation of assumptions on real data. The DH block bootstrap approach is implemented in the `xtg-cause` library developed by Lopez and Weber (2017). Different parameters can be specified, *e.g.*, the number of lags (this parameter can also be specified for the DH-test in the same library) and the number of *breps* which denotes the number of repetitions of step 7.

C   QPPA WITH LOWER BOUND ON γ

As explained in Section 3.2, our aggregation procedure relies on the choice of a $\gamma \in (0,1)$. Since a proper selection of γ is difficult, a different aggregation method is proposed. Specify a lower bound for γ, which we denote by $\gamma_{\min} \in (0,1)$ and correspondingly define

$$P_{X \to Y} := \min\left((1 - \log\gamma_{\min}) \inf_{\gamma \in (\gamma_{\min},1)} Q_{X \to Y}(\gamma), \ 1 \right) \tag{9}$$

for some fixed $\gamma_{\min} \in (0,1)$, where a recommended choice is $\gamma = 0.05$. Then, it holds:

Theorem 2. *Assume the generation procedure* (4) *and let* $\alpha, \gamma \in (0,1)$. *Then,* $P_{X \to Y}(\gamma)$ *is a asymptotically correct p-value,* i.e.,

$$\limsup_{T \to \infty} \mathbb{P}\left(P_{X \to Y}(\gamma) \leq \alpha\right) \leq \alpha,$$

where T denotes the number of timestamps.

Similar to Theorem 1, the aggregation idea (9) relies on an aggregation idea from high dimension statistics (see Meinhausen et al. (2009) Theorem 3.2).

D   PROOFS

D.1   Proof of Theorem 1

Proof. We follow the idea of the proof of Meinhausen et al. (2009) Theorem 3.1. Therefore, let

$$f_{X \to Y}(u) := \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}\{p_{X \to Y}^{(i)} \leq u\}, \quad u \in (0,1). \tag{10}$$

Notice that

$$Q_{X \to Y}(\gamma) \leq \alpha \quad \Leftrightarrow \quad f_{X \to Y}(\alpha\gamma) \geq \gamma \tag{11}$$

and hence

$$\mathbb{P}\left(Q_{X \to Y}(\gamma) \leq \alpha\right) = \mathbb{P}\left(f_{X \to Y}(\alpha\gamma) \geq \gamma\right).$$

Using the Markov inequality, we have

$$
\mathbb{P}\left(f_{X \to Y}(\alpha \gamma) \geq \gamma\right) \leq \frac{1}{\gamma} \mathbb{E}\left(f_{X \to Y}(\alpha \gamma)\right)
$$

$$
= \frac{1}{\gamma} \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}\left(\mathbb{1}\{p_{X \to Y}^{(i)} \leq \alpha \gamma\}\right).
$$

$$
= \frac{1}{\gamma} \frac{1}{N} \sum_{i=1}^{N} \mathbb{P}\left(p_{X \to Y}^{(i)} \leq \alpha \gamma\right)
$$

Since $p_{X \to Y}^{(i)}$ is by assumption asymptotically correct for each i, it holds

$$
\limsup_{T \to \infty} \mathbb{P}\left(p_{X \to Y}^{(i)} \leq \alpha \gamma\right) \leq \alpha \gamma \quad \text{for all } i
$$

and hence

$$
\limsup_{T \to \infty} \mathbb{P}\left(f_{X \to Y}(\alpha \gamma) \geq \gamma\right) \leq \lim_{T \to \infty} \frac{1}{\gamma} \frac{1}{N} \sum_{i=1}^{N} \mathbb{P}\left(p_{X \to Y}^{(i)} \leq \alpha \gamma\right)
$$

$$
\leq \frac{1}{\gamma} \frac{1}{N} \sum_{i=1}^{N} \alpha \gamma = \alpha
$$

which completes the proof. $\qquad\square$

D.2 Proof of Theorem 2

Proof. We follow the idea of the proof of Meinhausen et al. (2009) Theorem 3.2. Therefore, first note that for a uniformly distributed random variable U, it holds

$$
\sup_{\gamma \in (\gamma_{\min}, 1)} \frac{\mathbb{1}\{U \leq \alpha \gamma\}}{\gamma} = \begin{cases} 0 & U \geq \alpha \\ \alpha/U & \alpha \gamma_{\min} \leq U < \alpha \\ 1/\gamma_{\min} & U < \alpha \gamma_{\min} \end{cases} \tag{12}
$$

And therefore

$$
\mathbb{E}\left[\sup_{\gamma \in (\gamma_{\min}, 1)} \frac{\mathbb{1}\{U \leq \alpha \gamma\}}{\gamma}\right] = \alpha(1 - \log \gamma_{\min}).
$$

Since $p_{X \to Y}^{(i)}$ is an asymptotically correct p-value, it holds that for the cdf of $p_{X \to Y}^{(i)}$ denoted by $K^{(i)}$ and the cdf of U denoted by G it holds that $\lim_{T \to \infty} K^{(i)}(x) \leq G(x)$ for all x and i and therefore, for every weakly decreasing bounded function u and all i it holds that

$$
\limsup_{T \to \infty} \int u(x)\, dK^{(i)}(x) = \int u(x)\, d\left(\limsup_{T \to \infty} K^{(i)}(x)\right) \leq \int u(x) dG(x). \tag{13}
$$

Seeing that for arbitrary fixed i, it holds

$$
\sup_{\gamma \in (\gamma_{\min}, 1)} \frac{\mathbb{1}\{p_{X \to Y}^{(i)} \leq \alpha \gamma\}}{\gamma} = \begin{cases} 0 & p_{X \to Y}^{(i)} \geq \alpha \\ \alpha/p_{X \to Y}^{(i)} & \alpha \gamma_{\min} \leq p_{X \to Y}^{(i)} < \alpha \\ 1/\gamma_{\min} & p_{X \to Y}^{(i)} < \alpha \gamma_{\min}, \end{cases} \tag{14}
$$

the only difference between $\mathbb{E}\left[\sup_{\gamma \in (\gamma_{\min}, 1)} \frac{\mathbb{1}\{U \leq \alpha \gamma\}}{\gamma}\right]$ and $\mathbb{E}\left[\sup_{\gamma \in (\gamma_{\min}, 1)} \frac{\mathbb{1}\{p_{X \to Y}^{(i)} \leq \alpha \gamma\}}{\gamma}\right]$ is in the second case in (12) vs (14), but since $u(x) := \alpha/x$ is monotonically decreasing in $[\alpha \gamma_{\min}, \alpha]$ and because of (13), we obtain

$$
\limsup_{T \to \infty} \mathbb{E}\left[\sup_{\gamma \in (\gamma_{\min}, 1)} \frac{\mathbb{1}\{p_{X \to Y}^{(i)} \leq \alpha \gamma\}}{\gamma}\right] \leq \mathbb{E}\left[\sup_{\gamma \in (\gamma_{\min}, 1)} \frac{\mathbb{1}\{U \leq \alpha \gamma\}}{\gamma}\right] = \alpha(1 - \log \gamma_{\min})
$$

Taking the mean over all panel member, we obtain

$$\limsup_{T \to \infty} \mathbb{E}\left[\sup_{\gamma \in (\gamma_{\min}, 1)} \frac{\frac{1}{N}\sum_{i=1}^N \mathbb{1}\{p_{X \to Y}^{(i)} \le \alpha\gamma\}}{\gamma} \right] \le (1 - \log \gamma_{\min}).$$

Using again the Markov inequality, we obtain

$$\limsup_{T \to \infty} \mathbb{P}\left[\sup_{\gamma \in (\gamma_{\min}, 1)} \mathbb{1}\{f_{X \to Y}(\alpha\gamma) \ge \alpha\gamma\} \right] \le \alpha(1 - \log \gamma_{\min}),$$

where we use the definition of $f_{X \to Y}$ from (10). Using (11), we obtain

$$\limsup_{T \to \infty} \mathbb{P}\left[\inf_{\gamma \in (\gamma_{\min}, 1)} Q_{X \to Y}(\gamma) \le \alpha \right] \le \alpha(1 - \log \gamma_{\min})$$

and hence

$$\limsup_{T \to \infty} \mathbb{P}\left[\inf_{\gamma \in (\gamma_{\min}, 1)} Q_{X \to Y}(\gamma)(1 - \log \gamma_{\min}) \le \alpha \right] \le \alpha$$

which completes the proof. □

E   HOW TO DEAL WITH CONFOUNDING

Since we apply Granger causality on every individual of the panel to test whether X_i causes Y_i, we have to deal with the problem that Granger causality disregards hidden confounding. In this section, we want to give a practical solution to this problem. In the following, we denote by $past(t)$ the past timestamps of t. We now give sufficient conditions for confounding vs actual causal influence. For that, we consider two arbitrary time series W and V without the context of panel data. The following Proposition is closely related to Mastakouri et al. (2021) Theorem 1.b. and requires partially the same assumptions which we list bellow (note that we always assume that the present cannot causally influence the past):

Assumptions Appendix E:

1. The Causal Markov condition in the full time graph holds.

2. Causal Faithfulness in the full time graph

3. Stationary full time graph: the full time graph is invariant under a joint time shift of all variables

Proposition 3. *Assume that Assumption Appendix E holds, $W_{past(t)} \not\perp\!\!\!\perp V_t | V_{past(t)}$, $W_t \perp\!\!\!\perp V_{past(t)} | W_{past(t)}$ for all t and that W is not causing V. Then, there exists a (potentially high dimensional) memoryless (i.e. it does not hold that $Z_{t-1} \to Z_t$) confounder Z such that the triplet $\{W, V, Z\}$ is causally sufficient and there exists at least one $t' \in past(t)$ such that $Z_{t'} \to V_t$ and $Z_{t'} \to W_{t-n}$ for some $n > 0$ but there exists no $t'' \in past(t)$ such that $Z_{t''} \to W_t$ and $Z_{t''} \to V_{t-n}$ for some $n > 0$.*

Proof. Assume that W is not causing V, then according to Reichenbachs principle of common causes, there exists a confounder Z between W and V. Further, without loss of generality, we can assume causal sufficiency for the triplet $\{W, V, Z\}$ since, if another confounder exists we include it to the (potentially high dimensional) confounder Z. It then follows, that there exists no $t'' \in past(t)$ such that $Z_{t''} \to W_t$ and $Z_{t''} \to V_{t-n}$ for some $n > 0$ because otherwise $W_t \not\perp\!\!\!\perp V_{past(t)} | W_{past(t)}$ which is a contradiction to the assumption. Further, there exists at least one $t' \in past(t)$ such that $Z_{t'} \to V_t$ and $Z_{t'} \to W_{t-n}$ for some $n > 0$ because otherwise $W_{past(t)} \perp\!\!\!\perp V_t | V_{past(t)}$ which is a contradiction to the assumption.

It remains to show that Z has no memory effect. If Z would have a memory effect, then

$$V_t \leftarrow Z_{past(t)} \to Z_t \to Z_{t+1} \to W_{t+1}$$

and hence $W_{t+1} \not\perp\!\!\!\perp V_{past(t+1)} | W_{past(t+1)}$ which is a contradiction to the assumption. □

In particular, this proposition shows that if we observe W causing V but not V causing W, then this can only be due to hidden confounding if the confounder has the following structure: Z has to causally influence the present of V with larger time delay than the present of W. Further, Z cannot have a memory effect.

Hence, a practical way of dealing with hidden confounding is to exclude bi-directional influence between W and V. In that way, we might decrease the true positive rate but also remove false positives due to hidden confounding. Since our QPPA approach relies on Granger causality for the individual panel members, the same also holds for QPPA.

Here, we want to mention that Peters et al. (2017) argue after Figure 10.7.(a) that such a confounder structure occurs in real data using the example of price of butter and cheese. They state that the price of butter and cheese are confounded by the price of milk but the influence from milk to cheese has a larger time delay than from milk to butter because it takes longer to produce cheese.

However, this is only an example for such a confounder structure if the price of milk has no memory effect, otherwise we observe significant influence (of course depending on the strength of influence and memory effect) from the past of cheese to the present of butter **and** from the past of butter to the present of cheese.

F   DETECTING INSTANTANEOUS CAUSAL INFLUENCES

According to the dynamics (4), instantaneous effects are excluded. However, as Peters et al. (2017) Chapter 10.3 argue that, under some assumptions, Granger causality is able to detect causal influence even if it is purely instantaneous. Namely, if faithfulness holds and time series X has a memory effect (note that here we specifically consider the bi-variate case, in multi variate cases, instantaneous effects can lead to non-identifiability), *i.e.*, $X_{t-1} \to X_t$ for all t, then if X causes Y, the present of Y is not independent of the past of X given the past of Y because of the causal influence $X_{t-1} \to X_t \to Y_t$. The following figure, taken from Peters et al. (2017) illustrates this dependence. Hence, including the past of X into the prediction of Y decreases the prediction error and
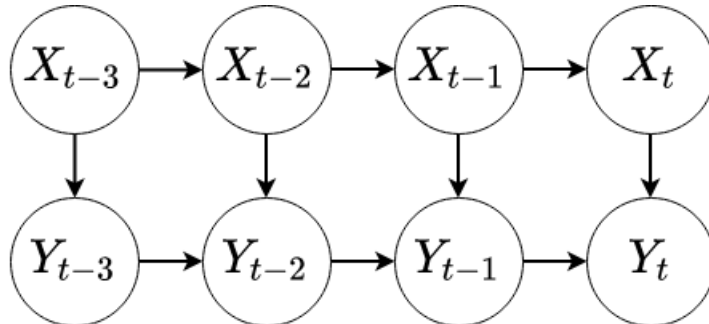

Figure 5: *Example of a causal relation with only instantaneous influences taken from Peters et al. (2017) Figure 10.8.(b) where Granger causality is able to detect the causal relation from X to Y. In this example, X and Y denote single time series, contrary to our panel data notation where X, Y denote panels.*

therefore Granger causality could be detected. Since our panel Granger Non-causality test relies on Granger causality on every individual panel member, our algorithm is thus capable of detecting causal influence even if it is purely instantaneous. This is shown in the following experiment.

Experiment for instantaneous effects: The generation process with only instantaneous causal effects (which is, as the processes in Section 4, an autoregressive process of order 1) is specified by:

$$X_{i,t} = \delta_{i,1} X_{i,t-1} + \eta_{i,t},$$
$$Y_{i,t} = \theta_{i,1} Y_{i,t-1} + \beta X_{i,t} + \epsilon_{i,t},$$

where the innovation processes are i.i.d. Gaussian random variables with $\eta_{i,t}, \epsilon_{i,t} \sim N(0, 0.1)$ and i.i.d. $\delta_{i,1}, \theta_{i,1} \sim Unif(0.2, 0.8)$. Further, we either draw β from $Unif(0.2, 0.8)$ if the null should be rejected or set $\beta = 0$ if the null should be accepted. The instantaneous effect comes from the influence $\beta X_{i,t}$ on $Y_{i,t}$ and this generation process

results in the causal structure corresponding to Figure 5. Since the dependence of X_{t-1} and Y_t given Y_{t-1} is indirect through the memory effect and the instantaneous effect, we need a stronger signal to detect the relation via QPPA, therefore we also include results of QPPA where we draw β from $Unif(0.6, 0.8)$. The results in Figure 6 show that for $T = 50$, the power decreases relatively fast. Also, for $T = 100$, the power is smaller than for causal influence with time delay. However, we also see in in Figure 6.(d) that QPPA is still able to recover the true causal relation in most cases for γ between 0.01 and approx. 0.5 even if the influence is purely instantaneous.

(a) $T = 50$, $\beta \sim Unif(0.2, 0.8)$

(b) $T = 100$, $\beta \sim Unif(0.2, 0.8$

(c) $T = 50$, $\beta \sim Unif(0.6, 0.8)$
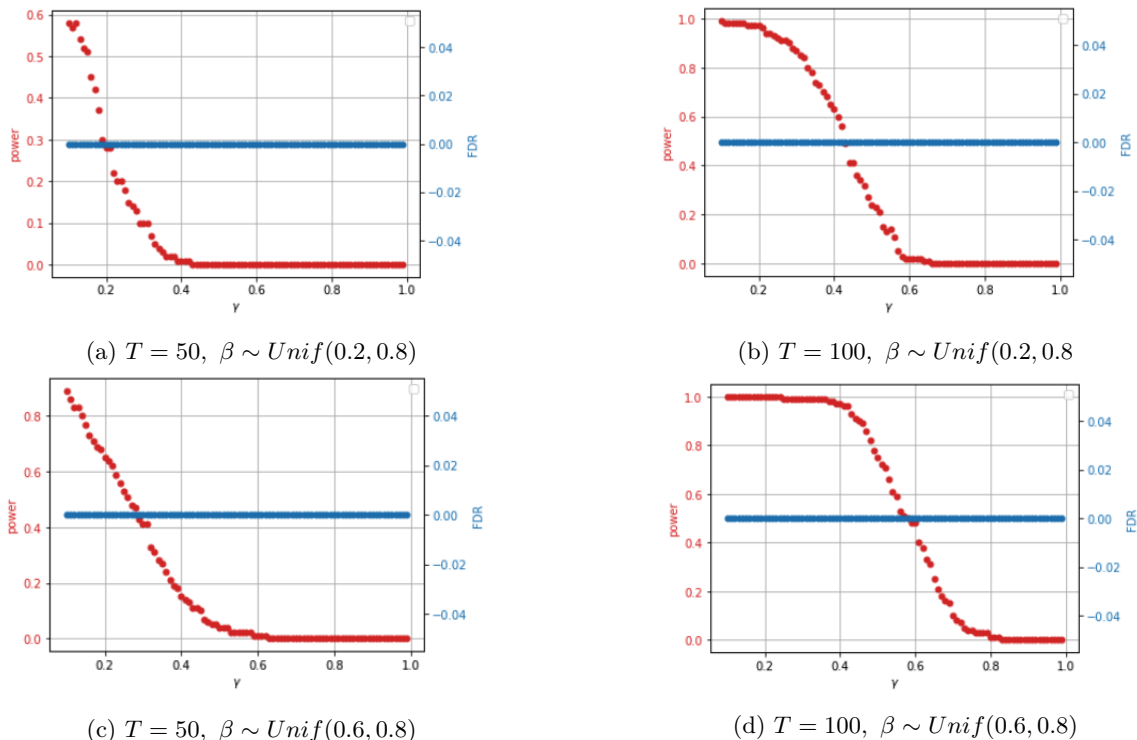
(d) $T = 100$, $\beta \sim Unif(0.6, 0.8)$

Figure 6: *Empirical results of Experiment for instantaneous effects for QPPA with $\gamma = 0.01, 0.02, \ldots, 0.99$. We choose $N = 30$ and $T = 50, 100$.*

G   DEALING WITH NON-STATIONARITIES

A common way to deal with non-stationarities to test for Granger Non-causality is to difference the processes. More precisely, following Granger (1981); Engle and Granger (1987), we say that a time series W is integrated of order d, denoted by $I(d)$, if $\{(1 - L)^d W_t\}_t$ is (covariance) stationary, where L denotes the lag operator. Different test procedures can be applied to find the order of integration, we examine one option for the COVID-19 study in Appendix I.1.

If we want to test the existence of Granger causality between two time series W and V, we can first search for the order of integration using stationarity tests on the n-th difference process where we stop as soon as stationarity gets accepted. If W and V have different orders of integration, say d_W and d_V, we take the maximum and difference both time series $\max(d_W, d_V)$-times. The difference processes often lead to stationary time series in practice. However, note that such a d does not necessarily exist and hence this approach is not applicable for all datasets.

H FURTHER EXPERIMENTS WITH SYNTHETIC DATA

Experiment for sporadically missing connections: For this experiment we again consider the model without cross-sectional dependencies:

$$X_{i,t} = \delta_{i,1} X_{i,t-1} + \eta_{i,t},$$
$$Y_{i,t} = \theta_{i,1} Y_{i,t-1} + \beta X_{i,t-1} + \epsilon_{i,t},$$

where the innovation processes are i.i.d. Gaussian random variables with $\eta_{i,t}, \epsilon_{i,t} \sim N(0, 0.1)$ and i.i.d. $\delta_{i,1}, \theta_{i,1} \sim Unif(0.2, 0.8)$. Further, we either draw β from $Unif(0.2, 0.8)$ if the null should be rejected or set $\beta = 0$ if the null should be accepted. To show robustness against sporadically missing connections, we set β to zero with probability a in the case where the null should be rejected where we let a range from 0.1 to 0.9. The proportion of missing connections when the null should be rejected is a, *i.e.*, for every panel member there is the chance of a that $\beta = 0$ although the null should be rejected. The results in Figure 7 show that especially in the large sample
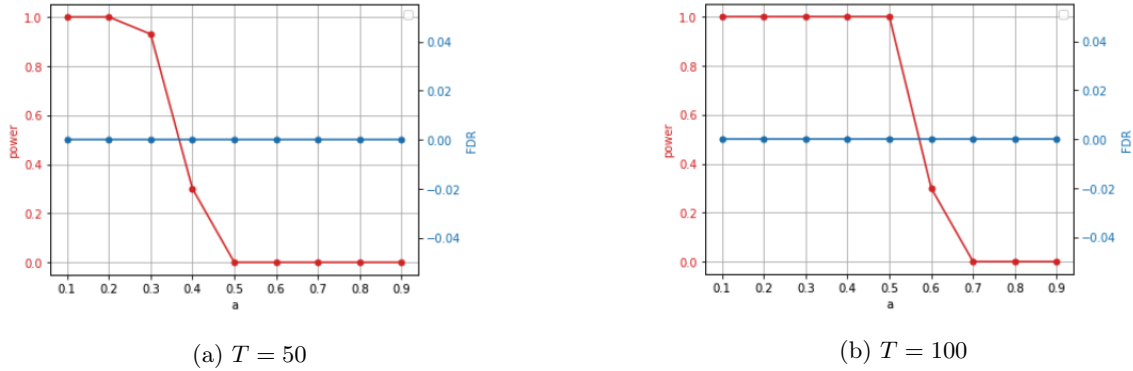

(a) $T = 50$

(b) $T = 100$

Figure 7: *Empirical results of Experiment for sporadically missing edges for QPPA, where we randomly set β to 0 in the case where the null should be rejected with probability a where we let a range from 0.1 to 0.9. We choose $N = 100$ and $T = 50, 100$ and $\gamma = 0.5$.*

regime, QPPA is robust against sporadically missing edges since the power drops only after the probability that an edges is missing although it should be there is higher than 50%.

Experiment non-Gaussian noise: Again, we consider the model without cross-section dependencies:

$$X_{i,t} = \delta_{i,1} X_{i,t-1} + \eta_{i,t},$$
$$Y_{i,t} = \theta_{i,1} Y_{i,t-1} + \beta X_{i,t-1} + \epsilon_{i,t},$$

where the innovation processes are i.i.d. **uniformly distributed** random variables with $\eta_{i,t}, \epsilon_{i,t} \sim Unif(0.2, 0.8)$ and i.i.d. $\delta_{i,1}, \theta_{i,1} \sim Unif(0.2, 0.8)$ and $\beta \sim Unif(0.2, 0.8)$ if the null should be rejected and $\beta = 0$ if the null should be accepted. The results in Figure 8 show that non-Gaussian noise does not decrease the power/FDR of QPPA, *cf.* Figure 2.

Experiment non i.i.d noise: Also for this experiment we use the model without cross-sectional dependencies:

$$X_{i,t} = \delta_{i,1} X_{i,t-1} + \eta_{i,t},$$
$$Y_{i,t} = \theta_{i,1} Y_{i,t-1} + \beta X_{i,t-1} + \epsilon_{i,t},$$

where we sample the innovation processes from 3-blocks, *i.e.*, $\eta_{i,1}, \eta_{i,2}, \eta_{i,3}$ are dependent, $\eta_{i,4}, \eta_{i,5}, \eta_{i,6}$ are dependent etc. and similar for $\epsilon_{i,t}$. Further, $\delta_{i,1}, \theta_{i,1}$ are i.i.d. $Unif(0.2, 0.8)$

In Figure 9, we see similar results as in Figure 8. It shows that QPPA is robust against this type of dependent residuals.
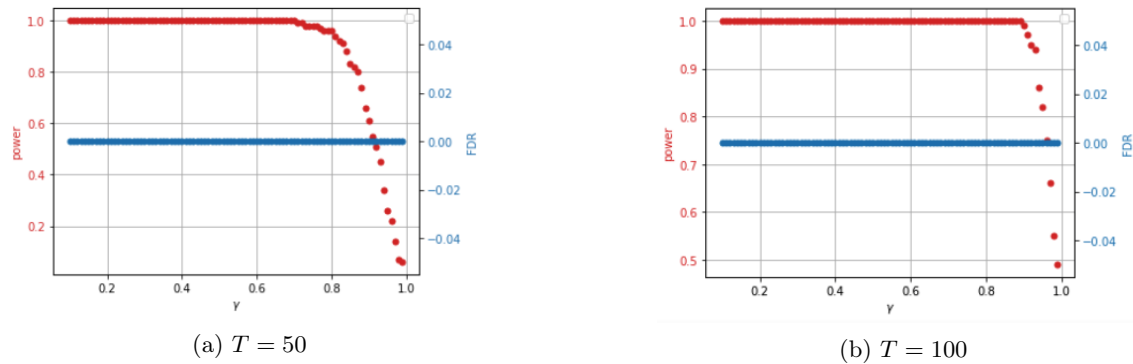
(a) $T = 50$

(b) $T = 100$

Figure 8: *Empirical results of experiment for non-Gaussian noise effects for QPPA with $\gamma = 0.01, 0.02, \ldots, 0.99$. We choose $N = 30$ and $T = 50, 100$.*

(a) $T = 50$

(b) $T = 100$

Figure 9: *Empirical results of Experiment for non-Gaussian noise effects for QPPA with $\gamma = 0.01, 0.02, \ldots, 0.99$. We choose $N = 30$ and $T = 50, 100$.*
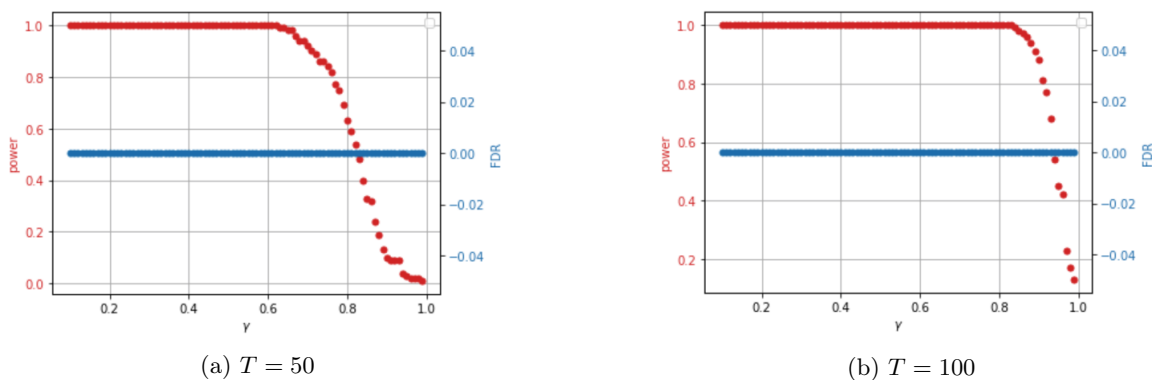
I   ADDITIONAL RESULTS TO COVID STUDY

I.1   COVID-19 Data Preprocessing Steps

The preprocessing steps we apply on the COVID-19 data about confirmed cases and deaths are as follows:

1. We remove all members of the panel (*i.e.*, Countries/Regions) which have at least one missing value in confirmed cases or deaths since neither the Granger causality implementation of `statsmodel` nor `xtg-cause` can deal with missing values. However, note that QPPA could deal with missing values here, it would only require an implementation of Granger causality that can deal with missing values.

2. Although the dataset contains records starting from 22nd January 2020, we only consider the time range 1st November 2020 until 4th October 2021 because many countries did not have confirmed COVID-19 cases (and therefore also no death cases) in the beginning of the pandemic.

3. We standardize the data.

4. Since Granger causality can only deal with stationary time series, we first apply a stationarity test for panel data that we developed based on the augmented Dickey-Fuller test (adf test) (Dickey and Fuller, 1979), combined with the p-value aggregation method explained in Section 2.3 to test for unit root for each member of the panel. We include more explanation and the test results in Appendix I.3. If the null gets accepted to the significance level of 5% (where the null is that there exists a unit root, *i.e.*, the time series are non-stationary), we generate the first difference of the panel members and apply the test again on the difference process. If the test again accepted the null (again to the significance level of 5%), we generate the second order difference processes. We continue this procedure until the null gets rejected. The results are

also given in Appendix I.3. Our test procedure rejects the null for the second order difference process which we will use from now on for the following analysis.

5. From the panel we remove those members for which either the stationary version of confirmed cases or deaths is constant since causality cannot be conducted from constant time series.

I.2 Cross-Sectional Dependence Test

To test for cross-sectional dependence, we use the `Stata`-implementation (De Hoyos and Sarafidis, 2006) which contains the test procedures of Pesaran (2004); Frees (1995); Friedman (1937). There, we apply every test on the preprocessed data according to Appendix I.1. The null of these tests is that the individual panel members are independent. For every test, we obtain the p-value 0.000 and therefore all three tests strongly reject the null. Hence, we accept the alternative that cross-sectional dependencies exist.

I.3 Non-Stationarity Test

There are several existing non-stationarity tests for panel data, see Breitung (2000); Breitung and Das (2005); Choi (2001); Hadri (2000); Harris and Tzavalis (1999); Im et al. (2003); Levin et al. (2002). However, most of them assume independence across panel members. Here, we want to present another panel stationarity test that relies on the same procedure as our QPPA approach, Section 3.2, except that in step 1 we instead of applying Granger causality on the individual panel members, we apply a stationarity test and then aggregate the corresponding p-values with the procedure described in step 2. For the stationarity test, we use the augmented Dickey-Fuller test (adf), see Dickey and Fuller (1979). Hence, we apply the adf test on each individual panel member on the COVID-19 data about confirmed cases and deaths, where we use 12 time lags. Afterwards, we aggregate these p-values using the aggregation method

$$\min\{1, \text{emp. } \gamma\text{-quantile}(p_j/\gamma : j \in 1, \ldots, N)\},$$

where p_j denotes the p-value of the adf test for the j-th panel member. Our hypothesis test reads

$$H_0 : \text{The panel has a unit root}$$
$$H_1 : \text{The panel has no unit root.}$$

Similar to the hypotheses test of QPPA, we assume that either each member of the panel has a unit root or none of them.

To find the order of integration (see Section G), we apply the adf test combined with the second step of QPPA. Table 4 show the results for different γ. We see consistent rejection of the null for the second order difference processes, whereas for the first order we only reject for $\gamma = 0.1$ and without generating the difference process, the null is not rejected for any $\gamma \in \{0.1, 0.25, 0.5, 0.75, 0.9\}$. Hence, we have clear indication that the COVID-19 data is second difference order stationary. Since we need stationary data for Granger causality, we will henceforth use the second order processes for the analysis.

Table 4: *Empirical results for the stationarity test of COVID-19 using adf combined with the second step of QPPA. With a slight abuse of notation, we say that the panel is d-order integrated (denoted by $I(d)$) if the null of our panel adf test is rejected for the panel $\{\{(1-L)^d X_{i,t}\}_t\}_i$ and $\{\{(1-L)^d Y_{i,t}\}_t\}_i$ respectively where X denotes confirmed cases and Y denotes deaths, where L denotes the back shift operator. For more details see Section G.*

Order of integration	γ	p-val confirmed cases	deaths
$I(0)$	0.1	0.414	1.0
	0.25	1.0	1.0
	0.5	1.0	1.0
	0.75	1.0	1.0
	0.9	1.0	1.0
$I(1)$	0.1	0.001	0.003
	0.25	0.337	0.098
	0.5	0.569	0.426
	0.75	0.745	0.641
	0.9	0.957	0.843
$I(2)$	0.1	9.883e-14	1.877e-12
	0.25	2.297e-12	2.457e-10
	0.5	2.543e-08	1.702e-06
	0.75	5.574e-05	0.001
	0.9	0.004	0.006

I.4 Additional COVID-19 Experiments

We repeat the experiment in Figure 4 with different number of countries/regions. The results, given in Figure 10, show that the accuracy of QPPA increases with increasing number of countries/regions per run.
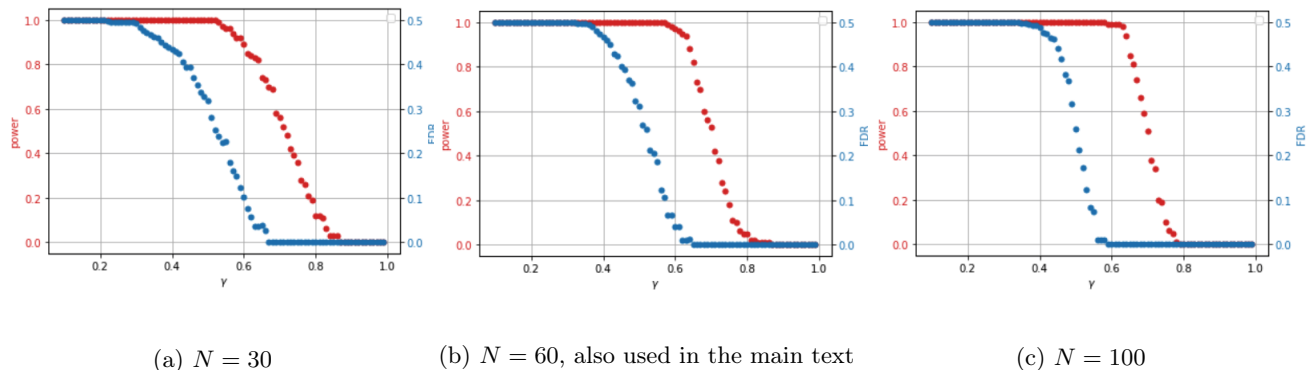

(a) $N = 30$ (b) $N = 60$, also used in the main text   (c) $N = 100$

Figure 10: Empirical results for COVID-19 data about confirmed cases and deaths using QPPA with $\gamma = 0.01$, $0.02, \ldots, 0.99$. To calculate power and FDR, we randomly selected N countries/regions out of the 225 (here, N is specified in the subfigures) and checked whether QPPA detects the causal relation c $\rightarrow$ d and d $\rightarrow$ c respectively to the significance level 5% and repeat this 100 times. Here, we use the complete time series length available which is $T = 335$.