Simultaneous statistical inference for epidemic trends

Marina Khismatullina[1] University of Bonn

Michael Vogt[2] University of Bonn

The COVID-19 pandemic is one of the most pressing issues at present. A question which is particularly important for governments and policy makers is the following: Does the virus spread in the same way in different countries? Or are there significant differences in the development of the epidemic? In this paper, we devise new inference methods that allow to detect differences in the development of the COVID-19 epidemic across countries in a statistically rigorous way. In our empirical study, we use the methods to compare the outbreak patterns of the epidemic in a number of European countries.

Key words: simultaneous hypothesis testing; multiscale test; time trend; panel data; COVID-19.

JEL classifications: C12; C23; C54.

1 Introduction

There are many questions surrounding the current COVID-19 pandemic that are not well understood yet. A question which is particularly important for governments and policy makers is the following: How do the outbreak patterns of COVID-19 compare across countries? Are the time trends of daily new infections more or less the same across countries, or is the virus spreading differently in different regions of the world? Identifying differences between countries may help, for instance, to better understand which government policies have been more effective in containing the virus than others. The main aim of this paper is to develop new inference methods that allow to detect differences between time trends of COVID-19 infections in a statistically rigorous way.

Let X_{it} be the number of new infections on day t in country i and suppose we observe a sample of data $\mathcal{X}_i = \{X_{it} : 1 \leq 1 \leq T\}$ for n different countries i. In order to make the data comparable across countries, we take the starting date $t = 1$ to be the day of the 100th confirmed case in each country. This way of "normalizing" the data is common practice (cp. e.g. Cohen and Kupferschmidt, 2020). A simple way

[1]Corresponding author. Address: Bonn Graduate School of Economics, University of Bonn, 53113 Bonn, Germany. Email: marina.k@uni-bonn.de.

[2]Address: Department of Economics and Hausdorff Center for Mathematics, University of Bonn, 53113 Bonn, Germany. Email: michael.vogt@uni-bonn.de.

to model the count data X_{it} is to use a Poisson distribution. Specifically, we may assume that the random variables X_{it} are Poisson distributed with time-varying intensity parameter $\lambda_i(t/T)$, that is, $X_{it} \sim P_{\lambda_i(t/T)}$. Since $\lambda_i(t/T) = \mathbb{E}[X_{it}] = \mathrm{Var}(X_{it})$, we can model the observations X_{it} by the nonparametric regression equation

$$X_{it} = \lambda_i\Big(\frac{t}{T}\Big) + u_{it} \tag{1.1}$$

for $1 \leq t \leq T$, where $u_{it} = X_{it} - \mathbb{E}[X_{it}]$ with $\mathbb{E}[u_{it}] = 0$ and $\mathrm{Var}(u_{it}) = \lambda_i(t/T)$. As usual in nonparametric regression (cp. Robinson, 1989), we let the regression function λ_i in model (1.1) depend on rescaled time t/T rather than on real time t. Hence, $\lambda_i : [0,1] \to \mathbb{R}$ can be regarded as a function on the unit interval, which allows us to estimate it by standard techniques from nonparametric regression. Since λ_i is a function of rescaled time t/T, the variables X_{it} in model (1.1) depend on the time series length T in general, that is, $X_{it} = X_{it,T}$. To keep the notation simple, we however suppress this dependence throughout the paper. In Section 3, we introduce the model setting in detail which underlies our analysis. As we will see there, it is a generalized version of the Poisson model (1.1).

In model (1.1), the time trend of new COVID-19 infections in country i is described by the intensity function λ_i of the underlying Poisson distribution. Hence, the question whether the time trends are comparable across countries amounts to the question whether the intensity functions λ_i have the same shape across countries i. In this paper, we construct a multiscale test which allows to *identify* and *locate* the differences between the functions λ_i. More specifically, let $\mathcal{F} = \{\mathcal{I}_k \subseteq [0,1] : 1 \leq k \leq K\}$ be a family of (rescaled) time intervals $\mathcal{I}_k$ and let $H_0^{(ijk)}$ be the hypothesis that the functions λ_i and λ_j are the same on the interval $\mathcal{I}_k$, that is,

$$H_0^{(ijk)} : \lambda_i(w) = \lambda_j(w) \text{ for all } w \in \mathcal{I}_k.$$

We design a method to test the hypothesis $H_0^{(ijk)}$ *simultaneously* for all pairs of countries i and j under consideration and for all intervals $\mathcal{I}_k$ in the family $\mathcal{F}$. The main theoretical result of the paper shows that the method controls the familywise error rate, that is, the probability of wrongly rejecting at least one null hypothesis $H_0^{(ijk)}$. As we will see, this allows us to make simultaneous confidence statements of the following form for a given significance level $\alpha \in (0,1)$:

> *With probability at least $1-\alpha$, the functions λ_i and λ_j differ on the interval $\mathcal{I}_k$ for every (i,j,k) for which the test rejects $H_0^{(ijk)}$.*

Hence, the method allows us to make simultaneous confidence statements (a) about which time trend functions differ from each other and (b) about where, that is, in which time intervals $\mathcal{I}_k$ they differ.

Even though our multiscale test is motivated by the current COVID-19 crisis, its applicability is by no means restricted to this specific event. It is a general method to compare nonparametric trends in epidemiological (count) data. It thus contributes to the literature on statistical tests for equality of nonparametric regression and trend curves. Examples of such tests can be found in Härdle and Marron (1990), Hall and Hart (1990), King et al. (1991), Delgado (1993), Kulasekera (1995), Young and Bowman (1995), Munk and Dette (1998), Lavergne (2001), Neumeyer and Dette (2003) and Pardo-Fernández et al. (2007). More recent approaches were developed in Degras et al. (2012), Zhang et al. (2012), Hidalgo and Lee (2014) and Chen and Wu (2019). Compared to existing methods, our test has the following crucial advantage: it is much more informative. Most existing procedures allow to test *whether* the regression or trend curves under consideration are all the same or not. However, they do not allow to infer *which* curves are different and *where* (that is, in which parts of the support) they differ. Our multiscale approach, in contrast, conveys this information. Indeed, it even allows to make rigorous confidence statements about which curves λ_i are different and where they differ. To the best of our knowledge, there is no other method available in the literature which allows to make such simultaneous confidence statements. As far as we know, the only other multiscale test for comparing trend curves has been developed in Park et al. (2009). However, their analysis is mainly methodological and not backed up by a general theory. In particular, theory is only available for the special case $n = 2$. Moreover, the theoretical results are only valid under very severe restrictions on the family of time intervals $\mathcal{F}$.

The paper is structured as follows. As already mentioned above, Section 3 details the model setting which underlies our analysis. The multiscale test is developed step by step in Section 4. To keep the presentation as clear as possible, the technical details are deferred to the Appendix and the Supplementary Material. Section 5 contains the empirical part of the paper. There, we run some simulation experiments to demonstrate that the multiscale test has the formal properties predicted by the theory. Moreover, we use the test to compare the outbreak patterns of the COVID-19 epidemic in a number of European countries.

2 Literature review on COVID-19

Alfano and Ercolano (2020) analyse the effect (or lack thereof) of lockdown measures on the number of daily new cases across countries. Specifically, they consider a panel model where one of the covariates is a dummy-variable that indicates whether the country has a lockdown policy for a specific number of days or not. They find

out that this dummy variable has a negative and statistically significant coefficient for the whole range of days considered, suggesting that countries that implemented the lockdown have fewer new cases per day than countries that did not. However, when considering the European subsample the coefficient is positive and statistically significant up to day 13, statistically insignificant between days 13 and 17, and negative and statistically significant from day 17 on. These results suggest that the relationship between the number of new cases per day and the governmental response policies might be much more complicated than just linear.

3 Model setting

As already discussed in the Introduction, the assumption that $X_{it} \sim P_{\lambda_i(t/T)}$ leads to a nonparametric regression model of the form

$$X_{it} = \lambda_i\Big(\frac{t}{T}\Big) + u_{it} \quad \text{with} \quad u_{it} = \sqrt{\lambda_i\Big(\frac{t}{T}\Big)}\eta_{it}, \tag{3.1}$$

where η_{it} has zero mean and unit variance. In this model, both the mean and the variance are described by the same function λ_i. In empirical applications, however, the variance often tends to be much larger than the mean. To deal with this issue, which has been known for a long time in the literature (Cox, 1983) and which is commonly called overdispersion, so-called quasi-Poisson models (McCullagh and Nelder, 1989; Efron, 1986) are frequently used. In our context, a quasi-Poisson model of X_{it} has the form

$$X_{it} = \lambda_i\Big(\frac{t}{T}\Big) + \varepsilon_{it} \quad \text{with} \quad \varepsilon_{it} = \sigma\sqrt{\lambda_i\Big(\frac{t}{T}\Big)}\eta_{it}, \tag{3.2}$$

where σ is a scaling factor that allows the variance to be a multiple of the mean function λ_i. In what follows, we assume that the observed data X_{it} are produced by model (3.2), where the noise residuals η_{it} have zero mean and unit variance but we do not impose any further distributional assumptions on them.

Poisson and quasi-Poisson models are often used in the literature on epidemic modelling. De Salazar et al. (2020), for example, assume that the observed COVID-19 case count in country i follows a Poisson distribution with parameter λ_i being a linear function of some covariate Z_i, that is, $\lambda_i = \beta Z_i$. Pellis et al. (2020) consider a quasi-Poisson model for the number of new COVID-19 cases. They in particular examine (a) a version of the model where the mean function is parametrically restricted to be exponentially growing with a constant growth rate and (b) a version where the mean function is modelled nonparametrically by splines. Tobías et al.

(2020) analyze data on the accumulated number of cases using quasi-Poisson regression, where the mean function is modelled parametrically as a piecewise linear curve with known change points.

In order to derive our theoretical results, we impose the following regularity conditions on model (3.2):

(C1) The functions λ_i are uniformly Lipschitz continuous, that is, $|\lambda_i(u) - \lambda_i(v)| \leq L|u - v|$ for all $u, v \in [0, 1]$, where the constant L does not depend on i. Moreover, they are uniformly bounded away from zero and infinity, that is, there exist constants $\lambda_{\min}$ and $\lambda_{\max}$ with $0 \leq \lambda_{\min} \leq \min_{w\in[0,1]} \lambda_i(w) \leq \max_{w\in[0,1]} \lambda_i(w) \leq \lambda_{\max} < \infty$ for all i.

(C2) The random variables η_{it} are independent both across i and t. Moreover, for any i and t, it holds that $\mathbb{E}[\eta_{it}] = 0$, $\mathbb{E}[\eta_{it}^2] = 1$ and $\mathbb{E}[|\eta_{it}|^\theta] \leq C_\theta < \infty$ for some $\theta > 4$.

(C1) imposes some standard-type regularity conditions on the functions λ_i. In particular, the functions are assumed to be smooth, bounded from above and bounded away from zero. The latter restriction is required because the noise variance in model (3.2) equals zero if λ_i is equal to zero. Since we normalize our test statistics by an estimate of the noise variance as detailed in Section 4, we need this variance and thus the functions λ_i to be bounded away from zero. (C2) assumes the noise terms η_{it} to fulfill some mild moment conditions and to be independent both across countries i and time t. In the current COVID-19 crisis, independence across countries i seems to be a fairly reasonable assumption due to severe travel restrictions, the closure of borders, etc. Independence across time t is more debatable, but it is by no means unreasonable in our model framework: The time series process $\mathcal{X}_i = \{X_{it} : 1 \leq t \leq T\}$ produced by model (3.2) is nonstationary for each i. Specifically, both the mean $\mathbb{E}[X_{it}] = \lambda_i(t/T)$ and the variance $\mathrm{Var}(X_{it}) = \sigma^2\lambda_i(t/T)$ are time-varying. A well-known fact in the time series literature is that nonstationarities such as a time-varying mean may produce spurious sample autocorrelations (cp. e.g. Mikosch and Stărică, 2004; Fryzlewicz et al., 2008). Hence, the observed persistence of a time series (captured by the sample autocorrelation function) may be due to nonstationarities rather than real autocorrelations. This insight has led researchers to prefer simple nonstationary models over intricate stationary time series models in some application areas such as finance (cp. Mikosch and Stărică, 2000, 2004; Fryzlewicz et al., 2006; Hafner and Linton, 2010). In a similar vein, our model accounts for the persistence in the observed time series $\mathcal{X}_i$ via nonstationarities rather than autocorrelations in the error terms.

4 The multiscale test

Let $\mathcal{S} \subseteq \{(i,j) : 1 \leq i < j \leq n\}$ be the set of all pairs of countries (i,j) whose trend functions λ_i and λ_j we want to compare. Moreover, as already introduced above, let $\mathcal{F} = \{\mathcal{I}_k : 1 \leq k \leq K\}$ be the family of (rescaled) time intervals under consideration. Finally, write $\mathcal{M} := \mathcal{S} \times \{1, \ldots, K\}$ and let $p := |\mathcal{M}|$ be the cardinality of $\mathcal{M}$. In this section, we devise a method to test the null hypothesis $H_0^{(ijk)}$ simultaneously for all pairs of countries $(i,j) \in \mathcal{S}$ and all time intervals $\mathcal{I}_k \in \mathcal{F}$, that is, for all $(i,j,k) \in \mathcal{M}$. The value $p = |\mathcal{M}|$ is the dimensionality of the simultaneous test problem we are dealing with. It amounts to the number of tests that we carry out simultaneously. As shown by our theoretical results in the Appendix, p may be much larger than the time series length T, which means that the simultaneous test problem under consideration can be very high-dimensional.

4.1 Construction of the test statistics

A statistic to test the hypothesis $H_0^{(ijk)}$ for a given triple (i, j, k) can be constructed as follows. To start with, we introduce the expression

$$\hat{s}_{ijk,T} = \frac{1}{\sqrt{Th_k}} \sum_{t=1}^{T} \mathbf{1}\Big(\frac{t}{T} \in \mathcal{I}_k\Big)(X_{it} - X_{jt}),$$

where h_k is the length of the time interval $\mathcal{I}_k$, $\mathbf{1}(\cdot)$ denotes the indicator function and $\mathbf{1}(t/T \in \mathcal{I}_k)$ can be regarded as a rectangular kernel weight. A simple application of the law of large numbers yields that $\hat{s}_{ijk,T}/\sqrt{Th_k} = (Th_k)^{-1} \sum_{t=1}^{T} \mathbf{1}(t/T \in \mathcal{I}_k)\{\lambda_i(t/T) - \lambda_j(t/T)\} + o_p(1)$ for any fixed pair of countries (i, j). Hence, the statistic $\hat{s}_{ijk,T}/\sqrt{Th_k}$ estimates the average distance between the functions λ_i and λ_j on the interval $\mathcal{I}_k$. Under (C2), it holds that

$$\nu_{ijk,T}^2 := \mathrm{Var}(\hat{s}_{ijk,T}) = \frac{\sigma^2}{Th_k} \sum_{t=1}^{T} \mathbf{1}\Big(\frac{t}{T} \in \mathcal{I}_k\Big)\Big\{\lambda_i\Big(\frac{t}{T}\Big) + \lambda_j\Big(\frac{t}{T}\Big)\Big\}.$$

In order to normalize the variance of the statistic $\hat{s}_{ijk,T}$, we scale it by an estimator of $\nu_{ijk,T}$. In particular, we estimate $\nu_{ijk,T}^2$ by

$$\hat{\nu}_{ijk,T}^2 = \frac{\hat{\sigma}^2}{Th_k} \sum_{t=1}^{T} \mathbf{1}\Big(\frac{t}{T} \in \mathcal{I}_k\Big)\{X_{it} + X_{jt}\},$$

where $\hat{\sigma}^2$ is defined as follows: For each country i, let

$$\hat{\sigma}_i^2 = \frac{\sum_{t=2}^{T}(X_{it} - X_{it-1})^2}{2\sum_{t=1}^{T} X_{it}}$$

and set $\hat{\sigma}^2 = |\mathcal{C}|^{-1} \sum_{i \in \mathcal{C}} \hat{\sigma}_i^2$ with $\mathcal{C} = \{\ell : \ell = i \text{ or } \ell = j \text{ for some } (i, j) \in \mathcal{S}\}$ denoting the set of countries that are taken into account by our test. The idea behind the estimator $\hat{\sigma}_i^2$ is as follows: Since λ_i is Lipschitz continuous,

$$X_{it} - X_{it-1} = \sigma\sqrt{\lambda_i\Big(\frac{t}{T}\Big)}(\eta_{it} - \eta_{it-1}) + r_{it},$$

where $|r_{it}| \leq C(1 + |\eta_{it-1}|)/T$ with a sufficiently large constant C. This suggests that $T^{-1}\sum_{t=2}^{T}(X_{it} - X_{it-1})^2 = 2\sigma^2\{T^{-1}\sum_{t=2}^{T}\lambda_i(t/T)\} + o_p(1)$. Moreover, since $T^{-1}\sum_{t=1}^{T} X_{it} = T^{-1}\sum_{t=1}^{T}\lambda_i(t/T) + o_p(1)$, we expect that $\hat{\sigma}_i^2 = \sigma^2 + o_p(1)$ for any i and thus $\hat{\sigma}^2 = \sigma^2 + o_p(1)$. In Lemma S.1 of the Supplementary Material, we formally show that $\hat{\sigma}^2$ is a consistent estimator of σ^2 under our regularity conditions.

Normalizing the statistic $\hat{s}_{ijk,T}$ by the estimator $\hat{\nu}_{ijk,T}$ yields the expression

$$\hat{\psi}_{ijk,T} := \frac{\hat{s}_{ijk,T}}{\hat{\nu}_{ijk,T}} = \frac{\sum_{t=1}^{T} \mathbf{1}(\frac{t}{T} \in \mathcal{I}_k)(X_{it} - X_{jt})}{\hat{\sigma}\{\sum_{t=1}^{T} \mathbf{1}(\frac{t}{T} \in \mathcal{I}_k)(X_{it} + X_{jt})\}^{1/2}}, \tag{4.1}$$

which serves as our test statistic of the hypothesis $H_0^{(ijk)}$. For later reference, we additionally introduce the statistic

$$\hat{\psi}^0_{ijk,T} = \frac{\sum_{t=1}^{T} \mathbf{1}(\frac{t}{T} \in \mathcal{I}_k)\, \sigma \overline{\lambda}_{ij}^{1/2}(\frac{t}{T})(\eta_{it} - \eta_{jt})}{\hat{\sigma}\{\sum_{t=1}^{T} \mathbf{1}(\frac{t}{T} \in \mathcal{I}_k)(X_{it} + X_{jt})\}^{1/2}} \tag{4.2}$$

with $\overline{\lambda}_{ij}(u) = \{\lambda_i(u) + \lambda_j(u)\}/2$, which is identical to $\hat{\psi}_{ijk,T}$ under $H_0^{(ijk)}$.

4.2 Construction of the test

Our multiscale test is carried out as follows: For a given significance level $\alpha \in (0,1)$ and each $(i,j,k) \in \mathcal{M}$, we reject $H_0^{(ijk)}$ if

$$|\hat{\psi}_{ijk,T}| > c_{ijk,T}(\alpha),$$

where $c_{ijk,T}(\alpha)$ is the critical value for the (i,j,k)-th test problem. The critical values $c_{ijk,T}(\alpha)$ are chosen such that the familywise error rate (FWER) is controlled at level α, which is defined as the probability of wrongly rejecting $H_0^{(ijk)}$ for at least one (i,j,k). More formally speaking, for a given significance level $\alpha \in (0,1)$, the FWER is

$$\begin{aligned}
\mathrm{FWER}(\alpha) &= \mathbb{P}\Big(\exists (i,j,k) \in \mathcal{M}_0 : |\hat{\psi}_{ijk,T}| > c_{ijk,T}(\alpha)\Big) \\
&= 1 - \mathbb{P}\Big(\forall (i,j,k) \in \mathcal{M}_0 : |\hat{\psi}_{ijk,T}| \leq c_{ijk,T}(\alpha)\Big) \\
&= 1 - \mathbb{P}\Big(\max_{(i,j,k)\in\mathcal{M}_0} |\hat{\psi}_{ijk,T}| \leq c_{ijk,T}(\alpha)\Big),
\end{aligned}$$

where $\mathcal{M}_0 \subseteq \mathcal{M}$ is the set of triples (i,j,k) for which $H_0^{(ijk)}$ holds true.

There are different ways to construct critical values $c_{ijk,T}(\alpha)$ that ensure control of the FWER at level α. In the traditional approach, the same critical value $c_T(\alpha) = c_{ijk,T}(\alpha)$ is used for all (i,j,k). In this case, controlling the FWER at the level α requires to determine the critical value $c_T(\alpha)$ such that

$$\mathrm{FWER}(\alpha) = 1 - \mathbb{P}\Big(\max_{(i,j,k)\in\mathcal{M}_0} |\hat{\psi}_{ijk,T}| \leq c_T(\alpha)\Big) \leq \alpha. \tag{4.3}$$

This can be achieved by choosing $c_T(\alpha)$ as the $(1-\alpha)$-quantile of the statistic

$$\tilde{\Psi}_T = \max_{(i,j,k)\in\mathcal{M}} |\hat{\psi}^0_{ijk,T}|,$$

where $\hat{\psi}^0_{ijk,T}$ was introduced in (4.2). (Note that both the statistic $\tilde{\Psi}_T$ and the quantile $c_T(\alpha)$ depend on the dimensionality p of the test problem in general. To keep the notation simple, we however suppress this dependence throughout the paper. We use the same convention for all other quantities that are defined in the sequel.)

A more modern approach assigns different critical values $c_{ijk,T}(\alpha)$ to the test problems (i,j,k). In particular, the critical value for the hypothesis $H_0^{(ijk)}$ is allowed to depend on the length h_k of the time interval $\mathcal{I}_k$, that is, on the scale of the test problem. A general approach to construct scale-dependent critical values was pioneered by Dümbgen and Spokoiny (2001) and has been used in many other studies since then; cp. for example Rohde (2008), Dümbgen and Walther (2008), Rufibach and Walther (2010), Schmidt-Hieber et al. (2013), Eckle et al. (2017) and Dunker et al. (2019). In our context, the approach of Dümbgen and Spokoiny (2001) leads to the critical values

$$c_{ijk,T}(\alpha) = c_T(\alpha, h_k) := b_k + q_T(\alpha)/a_k,$$

where $a_k = \{\log(e/h_k)\}^{1/2}/\log\log(e^e/h_k)$ and $b_k = \sqrt{2\log(1/h_k)}$ are scale-dependent constants and the quantity $q_T(\alpha)$ is determined by the following consideration: Since

$$\begin{aligned}
\mathrm{FWER}(\alpha) &= \mathbb{P}\Big(\exists(i,j,k)\in\mathcal{M}_0 : |\hat{\psi}_{ijk,T}| > c_T(\alpha,h_k)\Big)\\
&= 1-\mathbb{P}\Big(\forall(i,j,k)\in\mathcal{M}_0 : |\hat{\psi}_{ijk,T}| \le c_T(\alpha,h_k)\Big)\\
&= 1-\mathbb{P}\Big(\forall(i,j,k)\in\mathcal{M}_0 : a_k\big(|\hat{\psi}_{ijk,T}| - b_k\big) \le q_T(\alpha)\Big)\\
&= 1-\mathbb{P}\Big(\max_{(i,j,k)\in\mathcal{M}_0} a_k\big(|\hat{\psi}_{ijk,T}| - b_k\big) \le q_T(\alpha)\Big), \qquad (4.4)
\end{aligned}$$

we need to choose the quantity $q_T(\alpha)$ as the $(1-\alpha)$-quantile of the statistic

$$\hat{\Psi}_T = \max_{(i,j,k)\in\mathcal{M}} a_k\big(|\hat{\psi}^0_{ijk,T}| - b_k\big)$$

in order to ensure control of the FWER at level α. Comparing (4.4) with (4.3), the current approach can be seen to differ from the traditional one in the following respect: the maximum statistic $\tilde{\Psi}_T$ is replaced by the rescaled version $\hat{\Psi}_T$ which re-weights the individual statistics $\hat{\psi}^0_{ijk,T}$ by the scale-dependent constants a_k and b_k. As demonstrated above, this translates into scale-dependent critical values

$c_{ijk,T}(\alpha) = c_T(\alpha, h_k)$.

Our theory allows us to work with both the traditional choice $c_{ijk,T}(\alpha) = c_T(\alpha)$ and the more modern, scale-dependent choice $c_{ijk,T}(\alpha) = c_T(\alpha, h_k)$. Since the latter choice produces a test approach with better theoretical properties in general (cp. Dümbgen and Spokoiny, 2001), we restrict attention to the critical values $c_T(\alpha, h_k)$ in the sequel. There is, however, one complication we need to deal with: As the quantiles $q_T(\alpha)$ are not known in practice, we cannot compute the critical values $c_T(\alpha, h_k)$ exactly in practice but need to approximate them. This can be achieved as follows: Under appropriate regularity conditions, it can be shown that

$$\begin{aligned}\hat{\psi}^0_{ijk,T} &= \frac{\sum_{t=1}^T \mathbf{1}(\frac{t}{T} \in \mathcal{I}_k)\, \sigma\overline{\lambda}_{ij}^{1/2}(\frac{t}{T})(\eta_{it} - \eta_{jt})}{\hat{\sigma}\{\sum_{t=1}^T \mathbf{1}(\frac{t}{T} \in \mathcal{I}_k)(X_{it} + X_{jt})\}^{1/2}} \\ &\approx \frac{1}{\sqrt{2Th_k}} \sum_{t=1}^T \mathbf{1}\Big(\frac{t}{T} \in \mathcal{I}_k\Big)\{\eta_{it} - \eta_{jt}\}.\end{aligned}$$

A Gaussian version of the statistic displayed in the final line above is given by

$$\phi_{ijk,T} = \frac{1}{\sqrt{2Th_k}} \sum_{t=1}^T \mathbf{1}\Big(\frac{t}{T} \in \mathcal{I}_k\Big)\big\{Z_{it} - Z_{jt}\big\},$$

where Z_{it} are independent standard normal random variables for $1 \le t \le T$ and $1 \le i \le n$. Hence, the statistic

$$\Phi_T = \max_{(i,j,k)\in\mathcal{M}} a_k\big(|\phi_{ijk,T}| - b_k\big)$$

can be regarded as a Gaussian version of the statistic $\hat{\Psi}_T$. We approximate the unknown quantile $q_T(\alpha)$ by the $(1-\alpha)$-quantile $q_{T,\mathrm{Gauss}}(\alpha)$ of Φ_T, which can be computed (approximately) by Monte Carlo simulations and can thus be treated as known.

To summarize, we propose the following procedure to simultaneously test the hypothesis $H_0^{(ijk)}$ for all $(i,j,k) \in \mathcal{M}$ at the significance level $\alpha \in (0,1)$:

$$\text{For each } (i,j,k) \in \mathcal{M},\ \text{reject } H_0^{(ijk)} \text{ if } |\hat{\psi}_{ijk,T}| > c_{T,\mathrm{Gauss}}(\alpha, h_k), \tag{4.5}$$

where $c_{T,\mathrm{Gauss}}(\alpha, h_k) = b_k + q_{T,\mathrm{Gauss}}(\alpha)/a_k$ with $a_k = \{\log(e/h_k)\}^{1/2}/\log\log(e^e/h_k)$ and $b_k = \sqrt{2\log(1/h_k)}$.

4.3 Formal properties of the test

In Theorem A.1 of the Appendix, we prove that under appropriate regularity conditions, the test defined in (4.5) (asymptotically) controls the familywise error rate FWER(α) for each pre-specified significance level α. As shown in Corollary A.1, this has the following implication:

$$\mathbb{P}\Big(\forall (i,j,k) \in \mathcal{M} : \text{ If } |\hat{\psi}_{ijk,T}| > c_{T,\text{Gauss}}(\alpha, h_k), \text{ then } (i,j,k) \notin \mathcal{M}_0\Big) \geq 1 - \alpha + o(1), \tag{4.6}$$

where $\mathcal{M}_0$ is the set of triples $(i,j,k) \in \mathcal{M}$ for which $H_0^{(ijk)}$ holds true. Verbally, (4.6) can be expressed as follows:

> *With (asymptotic) probability at least $1-\alpha$, the null hypothesis $H_0^{(ijk)}$ is violated for all $(i,j,k) \in \mathcal{M}$ for which the test rejects $H_0^{(ijk)}$.* (4.7)

In other words:

> *With (asymptotic) probability at least $1-\alpha$, the functions λ_i and λ_j differ on the interval $\mathcal{I}_k$ for all $(i,j,k) \in \mathcal{M}$ for which the test rejects $H_0^{(ijk)}$.* (4.8)

Hence, the test allows us to make simultaneous confidence statements (a) about which pairs of countries (i,j) have different trend functions and (b) about where, that is, in which time intervals $\mathcal{I}_k$ the functions differ.

4.4 Implementation of the test in practice

For a given significance level $\alpha \in (0,1)$, the test procedure defined in (4.5) is implemented as follows in practice:

Step 1. Compute the quantile $q_{T,\text{Gauss}}(\alpha)$ by Monte Carlo simulations. Specifically, draw a large number N (say $N = 5000$) samples of independent standard normal random variables $\{Z_{it}^{(\ell)} : 1 \leq t \leq T, 1 \leq i \leq n\}$ for $1 \leq \ell \leq N$. Compute the value $\Phi_T^{(\ell)}$ of the Gaussian statistic Φ_T for each sample ℓ and calculate the empirical $(1-\alpha)$-quantile $\hat{q}_{T,\text{Gauss}}(\alpha)$ from the values $\{\Phi_T^{(\ell)} : 1 \leq \ell \leq N\}$. Use $\hat{q}_{T,\text{Gauss}}(\alpha)$ as an approximation of the quantile $q_{T,\text{Gauss}}(\alpha)$.

Step 2. Compute the critical values $c_{T,\text{Gauss}}(\alpha, h_k)$ for $1 \leq k \leq K$ based on the approximation $\hat{q}_{T,\text{Gauss}}(\alpha)$.

Step 3. Carry out the test for each $(i,j,k) \in \mathcal{M}$ and store the test results in the variable $r_{ijk,T} = \mathbf{1}(|\hat{\psi}_{ijk,T}| > c_{T,\text{Gauss}}(\alpha, h_k))$ for each $(i,j,k) \in \mathcal{M}$, that is, let $r_{ijk,T} = 1$ if the hypothesis $H_0^{(ijk)}$ is rejected and $r_{ijk,T} = 0$ otherwise.

To graphically present the test results, we produce a plot for each pair of countries $(i,j) \in \mathcal{S}$ that shows the intervals $\mathcal{I}_k$ for which the test rejects the null $H_0^{(ijk)}$, that is, the intervals in the set $\mathcal{F}_{\text{reject}}(i,j) = \{\mathcal{I}_k \in \mathcal{F} : r_{ijk,T} = 1\}$. The plot is designed such that it graphically highlights the subset of intervals $\mathcal{F}^{\min}_{\text{reject}}(i,j) = \{\mathcal{I}_k \in \mathcal{F}_{\text{reject}}(i,j) :$ there exists no $\mathcal{I}_{k'} \in \mathcal{F}_{\text{reject}}(i,j)$ with $\mathcal{I}_{k'} \subset \mathcal{I}_k\}$. The elements of $\mathcal{F}^{\min}_{\text{reject}}(i,j)$ are called minimal intervals. By definition, there is no other interval $\mathcal{I}_{k'}$ in $\mathcal{F}_{\text{reject}}(i,j)$ which is a proper subset of a minimal interval $\mathcal{I}_k$. Hence, the minimal intervals can be regarded as those intervals in $\mathcal{F}_{\text{reject}}(i,j)$ which are most informative about the precise location of the differences between the trends λ_i and λ_j. In Section 5, we use the graphical device just described to present the test results of our empirical application; cp. panels (d) in Figures 3–6.

According to (4.6), we can make the following simultaneous confidence statement about the intervals in $\mathcal{F}_{\text{reject}}(i,j)$ for $(i,j) \in \mathcal{S}$:

> *With (asymptotic) probability at least* $1-\alpha$*, it holds that for every pair of countries* $(i,j) \in \mathcal{S}$*, the functions* λ_i *and* λ_j *differ on each interval in* $\mathcal{F}_{reject}(i,j)$*.* (4.9)

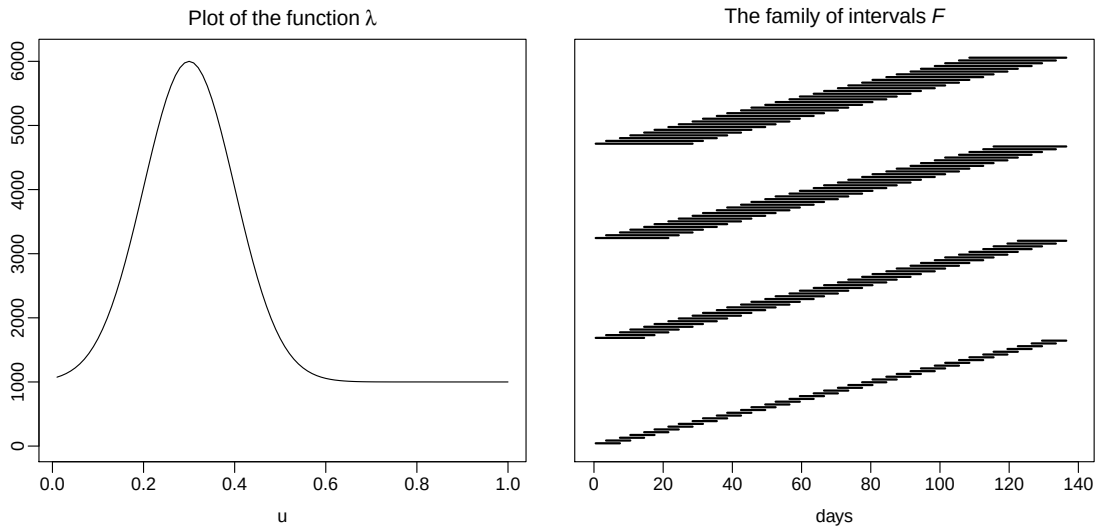
Hence, we can claim with statistical confidence at least $1-\alpha$ that the functions λ_i and λ_j differ on each time interval which is depicted in the plots of our graphical device. Since $\mathcal{F}^{\min}_{\text{reject}}(i,j) \subseteq \mathcal{F}_{\text{reject}}(i,j)$ for any $(i,j) \in \mathcal{S}$, the confidence statement (4.9) trivially remains to hold true when the sets $\mathcal{F}_{\text{reject}}(i,j)$ are replaced by $\mathcal{F}^{\min}_{\text{reject}}(i,j)$.

5 Empirical application to COVID-19 data

We now use our test to analyze the outbreak patterns of the COVID-19 epidemic. We proceed in two steps. In Section 5.1, we assess the finite sample performance of our test by Monte-Carlo experiments. Specifically, we run a series of experiments which show that the test controls the FWER at level α as predicted by the theory and that it has good power properties. In Section 5.2, we then apply the test to a sample of COVID-19 data from different European countries. Our multiscale test is implemented in the R package `multiscale`, available on GitHub at `https://github.com/marina-khi/multiscale`.

5.1 Simulation experiments

We simulate count data $\mathcal{X} = \{X_{it} : 1 \le i \le n, 1 \le t \le T\}$ by drawing the observations X_{it} independently from a negative binomial distribution with mean $\lambda_i(t/T)$ and variance $\sigma^2\lambda_i(t/T)$. By definition, X_{it} has a negative binomial distribution



(a) (b)

Figure 1: (a) Plot of the function λ; (b) plot of the family of intervals $\mathcal{F}$.

with parameters q and r if $\mathbb{P}(X_{it} = m) = \Gamma(m + r)/(\Gamma(r)m!)q^r(1 - q)^m$ for each $m \in \mathbb{N} \cup \{0\}$. Since $\mathbb{E}[X_{it}] = r(1 - q)/q$ and $\mathrm{Var}(X_{it}) = r(1 - q)/q^2$, we can use the parametrization $q = 1/\sigma^2$ and $r = \lambda_i(t/T)/(\sigma^2 - 1)$ to obtain that $\mathbb{E}[X_{it}] = \lambda_i(t/T)$ and $\mathrm{Var}(X_{it}) = \sigma^2\lambda_i(t/T)$. With this parametrization, the simulated data follow a nonparametric regression model of the form

$$X_{it} = \lambda_i\Big(\frac{t}{T}\Big) + \sigma\sqrt{\lambda_i\Big(\frac{t}{T}\Big)}\eta_{it},$$

where the noise variables η_{it} have zero mean and unit variance. The functions λ_i are specified below. The overdispersion parameter is set to $\sigma = 15$, which is similar to the estimate $\hat{\sigma} = 14.44$ obtained in the empirical application of Section 5.2. Robustness checks with $\sigma = 10$ and $\sigma = 20$ are provided in the Supplementary Material.

We consider different values for T and n, in particular, $T \in \{100, 250, 500\}$ and $n \in \{5, 10, 50\}$. Note that in the application, we have $T = 139$ and $n = 5$. We let $\mathcal{S} = \{(i, j) : 1 \leq i < j \leq n\}$, that is, we compare all pairs of countries (i, j) with $i < j$. Moreover, we choose $\mathcal{F}$ to be a family of time intervals $\mathcal{I}_k$ with length $h_k \in \{7/T, 14/T, 21/T, 28/T\}$. Hence, the intervals in $\mathcal{F}$ have length either 7, 14, 21 or 28 days (i.e., 1, 2, 3 or 4 weeks). For each length h_k, we include all intervals that start at days $t = 1 + 7(j - 1)$ and $t = 4 + 7(j - 1)$ for $j = 1, 2, \ldots$ A graphical presentation of the family $\mathcal{F}$ for $T = 139$ (as in the application) is given in Figure 1b. All our simulation experiments are based on $R = 5000$ simulation runs.

In the first part of the simulation study, we examine whether our test controls

Table 1: Empirical size of the test for different values of n and T.

	$n = 5$			$n = 10$			$n = 50$		
	significance level α			significance level α			significance level α		
	0.01	0.05	0.1	0.01	0.05	0.1	0.01	0.05	0.1
$T = 100$	0.011	0.047	0.093	0.010	0.044	0.087	0.008	0.037	0.075
$T = 250$	0.009	0.047	0.091	0.009	0.046	0.087	0.008	0.035	0.069
$T = 500$	0.010	0.044	0.083	0.008	0.048	0.093	0.007	0.035	0.077

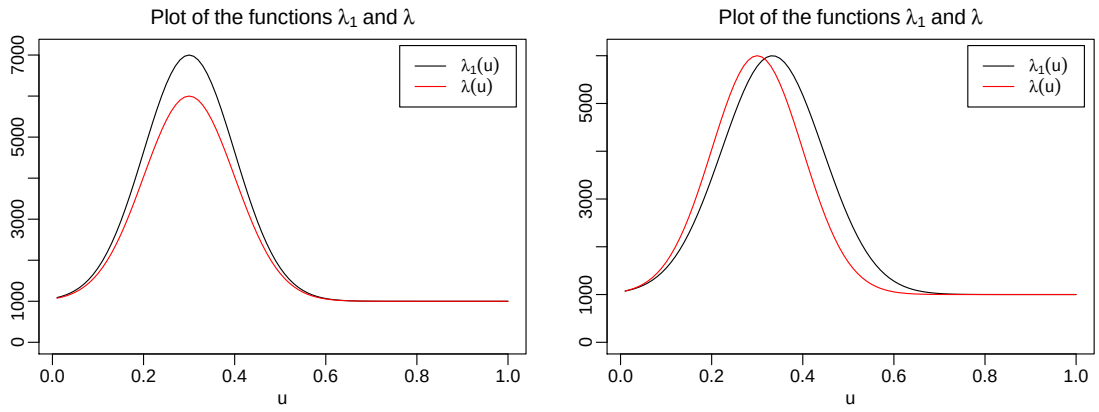
the FWER as predicted by the theory. To do so, we assume that the hypothesis $H_0^{(ijk)}$ holds true for all (i, j, k) under consideration, which implies that $\lambda_i = \lambda$ for all i. We consider the function

$$\lambda(u) = 5000 \exp\Big(- \frac{(10u - 3)^2}{2} \Big) + 1000, \tag{5.1}$$

which is similar in shape to some of the estimated trend curves in the application of Section 5.2. A plot of the function λ is provided in Figure 1a. To evaluate whether the test controls the FWER at level α, we compare the empirical size of the test with the target α. The empirical size is computed as the precentage of simulation runs in which the test falsely rejects at least one null hypothesis $H_0^{(ijk)}$.

The simulation results are reported in Table 1. As can be seen, the empirical size gives a reasonable approximation to the target α in all scenarios under investigation, even though the size numbers have a slight downward bias. This bias gets larger as the number of time series n increases, which reflects the fact that the test problem becomes more difficult for larger n. Already for $n = 5$, the number p of hypotheses to be tested is quite high, in particular, $p = 960, 2\,680, 5\,560$ for $T = 100, 250, 500$. This number increases to $p = 117\,600, 328\,300, 681\,100$ when $n = 50$. Hence, the dimensionality and thus the complexity of the test problem increases considerably as n gets larger. On first sight, it may seem astonishing that the downward bias does not diminish notably as the time series length T increases. This, however, has a simple explanation: The interval lengths h_k remain the same (7, 14, 21 or 28 days) as T increases, which implies that the effective sample size for computing the test statistics $\hat{\psi}_{ijk,T}$ does not change as well. To summarize, even though slightly conservative, the test controls the FWER quite accurately in the simulation setting at hand.

In the second part of the simulation study, we investigate the power properties of the test. To do so, we assume that $\lambda_i = \lambda$ for all $i > 1$ and that $\lambda_1 \neq \lambda$, where λ is defined in (5.1). Hence, only the first mean function λ_1 is different from the others. This implies that the hypothesis $H_0^{(ijk)}$ holds true for all (i, j, k) with $i > 1$ and $j > 1$, while there is at least one hypothesis $H_0^{(ijk)}$ with either $i = 1$ or $j = 1$



(a) Scenario A

(b) Scenario B

Figure 2: Plot of the functions λ_1 (black) and λ (red) in the simulation scenarios A and B.

that does not hold true. We consider two different simulation scenarios. In Scenario A, the function λ_1 has the form

$$\lambda_1(u) = 6000 \exp\Big(- \frac{(10u-3)^2}{2} \Big) + 1000$$

and is plotted together with λ in Figure 2a. As can be seen, the two functions λ_1 and λ peak at the same point in time, but the peak of λ_1 is higher than that of λ. In Scenario B, we let

$$\lambda_1(u) = 5000 \exp\Big(- \frac{(9u-3)^2}{2} \Big) + 1000.$$

Figure 2b shows that the peaks of λ_1 and λ have the same height but are reached at different points in time. To evaluate the power properties of the test in Scenarios A and B, we compute the percentage of simulation runs where the test (i) correctly detects differences between λ_1 and at least one of the other mean functions and (ii) does not spuriously detect differences between the other mean functions. Put differently, we calculate the percentage of simulation runs where (i) the set $\mathcal{F}_{\text{reject}}(1, j)$ is non-empty at least for one $j \in \{2, \ldots, n\}$ and (ii) all other sets $\mathcal{F}_{\text{reject}}(i, j)$ with $2 \leq i < j \leq n$ are empty. We call this percentage number the (empirical) power of the test. We thus use the term "power" a bit differently than usual.

The results for Scenario A (see Figure 2a) are presented in Table 2 and those for Scenario B (see Figure 2b) in Table 3. As can be seen, the test has substantial power in all the considered simulation settings. It is more powerful in Scenario B than in Scenario A, which is most presumably due to the fact that the differences $|\lambda_1(u) - \lambda(u)|$ are much larger in Scenario B. Moreover, it is less powerful for larger

Table 2: Power of the test for different values of n and T in Scenario A.

	$n = 5$			$n = 10$			$n = 50$		
	significance level α			significance level α			significance level α		
	0.01	0.05	0.1	0.01	0.05	0.1	0.01	0.05	0.1
$T = 100$	0.335	0.518	0.597	0.306	0.474	0.545	0.212	0.352	0.418
$T = 250$	0.615	0.790	0.836	0.580	0.764	0.800	0.470	0.648	0.705
$T = 500$	0.736	0.905	0.917	0.738	0.884	0.890	0.636	0.799	0.830

Table 3: Power of the test for different values of n and T in Scenario B.

	$n = 5$			$n = 10$			$n = 50$		
	significance level α			significance level α			significance level α		
	0.01	0.05	0.1	0.01	0.05	0.1	0.01	0.05	0.1
$T = 100$	0.824	0.910	0.903	0.812	0.893	0.890	0.738	0.847	0.857
$T = 250$	0.991	0.972	0.941	0.991	0.960	0.920	0.991	0.965	0.933
$T = 500$	0.997	0.973	0.949	0.995	0.961	0.923	0.996	0.969	0.932

numbers of time series n, which reflects the fact that the test problem gets more high-dimensional and thus more difficult as n increases. As one would expect, the power numbers tend to become larger as the time series length T and the significance level α increase. In Scenario B (mostly for $T = 250$ and $T = 500$), however, the power numbers drop down a bit as α gets larger. This reverse dependance can be explained by the way we calculate power: we exclude simulation runs where the test spuriously detects differences between the trends in countries i and j with $i, j > 1$. The number of spurious findings increases as we make the significance level α larger, which presumably causes the slight drop in power.

5.2 Analysis of COVID-19 data

The COVID-19 pandemic is one of the most pressing issues at present. The first outbreak occurred in Wuhan, China, in December 2019. On 30 January 2020, the World Health Organization (WHO) declared that the outbreak constitutes a Public Health Emergency of International Concern, and on 11 March 2020, the WHO characterized it as a pandemic. As of 22 July 2020, more than 14.56 million cases of COVID-19 infections have been reported worldwide, resulting in more than 607 000 deaths.

There are many open questions surrounding the current COVID-19 pandemic. A question which is particularly relevant for governments and policy makers is whether the pandemic has developed similarly in different countries or whether there are notable differences. Identifying these differences may give some insight into which government policies have been more effective in containing the virus than others. In

what follows, we use our multiscale test to compare the development of COVID-19 in several European countries. It is important to emphasize that our test allows to identify differences in the development of the epidemic across countries in a statistically rigorous way, but it does not tell what causes these differences. By distinguishing statistically significant differences from artefacts of the sampling noise, the test provides the basis for a further investigation into the causes. Such an investigation, however, presumably goes beyond a mere statistical analysis.

5.2.1 Data

We analyze data from five European countries: Germany, Italy, Spain, France and the United Kingdom. For each country i, we observe a time series $\mathcal{X}_i = \{X_{it} : 1 \leq t \leq T\}$, where X_{it} is the number of newly confirmed COVID-19 cases in country i on day t. The data are freely available on the homepage of the European Center for Disease Prevention and Control (`https://www.ecdc.europa.eu`) and were downloaded on 22 July 2020. As already mentioned in the Introduction, we take the day of the 100th confirmed case in each country as the starting date $t = 1$, which is a common way of "normalizing" the data and making them comparable across countries (cp. Cohen and Kupferschmidt, 2020). The time series length T is taken to be the minimal number of days for which we have observations for all five countries. The resulting dataset consists of $n = 5$ time series, each with $T = 139$ observations (as of July 22). Some of the time series contain negative values which we replaced by 0. Overall, this resulted in 6 replacements. Plots of the observed time series are presented in the upper panels (a) of Figures 3–6.

To interpret the results produced by our multiscale test, we consider the Government Response Index (GRI) from the Oxford COVID-19 Government Response Tracker (OxCGRT) (Hale et al., 2020b). The GRI measures how severe the actions are that are taken by a country's government to contain the virus. It is calculated based on several common government policies such as school closures and travel restrictions. The GRI ranges from 0 to 100, with 0 corresponding to no response from the government at all and 100 corresponding to full lockdown, closure of schools and workplaces, ban on travelling, etc. Detailed information on the collection of the data for government responses and the methodology for calculating the GRI is provided in Hale et al. (2020a). Plots of the GRI time series are given in panels (c) of Figures 3–6.

5.2.2 Test results

We assume that the data X_{it} of each country i in our sample follow the nonparametric trend model

$$X_{it} = \lambda_i\Big(\frac{t}{T}\Big) + \sigma\sqrt{\lambda_i\Big(\frac{t}{T}\Big)}\eta_{it},$$

which was introduced in equation (3.2). The overdispersion parameter σ is estimated by the procedure described in Section 4.1, which yields the estimate $\hat{\sigma} = 14.44$. Throughout the section, we set the significance level to $\alpha = 0.05$ and implement the multiscale test in exactly the same way as in the simulation study of Section 5.1. In particular, we let $\mathcal{S} = \{(i,j) : 1 \leq i < j \leq 5\}$, that is, we compare all pairs of countries (i,j) with $i < j$, and we choose $\mathcal{F}$ to be the family of time intervals plotted in Figure 1b. Hence, all intervals in $\mathcal{F}$ have length either 7, 14, 21 or 28 days.

With the help of our multiscale method, we simultaneously test the null hypothesis $H_0^{(ijk)}$ that $\lambda_i = \lambda_j$ on the interval $\mathcal{I}_k$ for each $(i,j,k) \in \mathcal{M}$. The results are presented in Figures 3–6, each figure comparing a specific pair of countries (i,j) from our sample. For the sake of brevity, we only show the results for the pairwise comparisons of Germany with each of the four other countries. The remaining figures can be found in Section S.2 of the Supplementary Material. Each figure splits into four panels (a)–(d). Panel (a) shows the observed time series for the two countries i and j that are compared. Panel (b) presents smoothed versions of the time series from (a), that is, it shows nonparametric kernel estimates (specifically, Nadaraya-Watson estimates) of the two trend functions λ_i and λ_j, where the bandwidth is set to 7 days and a rectangular kernel is used. Panel (c) displays the Government Response Index (GRI) of the two countries. Finally, panel (d) presents the results produced by our test: it depicts in grey the set $\mathcal{F}_{\text{reject}}(i,j)$ of all the intervals $\mathcal{I}_k$ for which the test rejects the null $H_0^{(ijk)}$. The minimal intervals in the subset $\mathcal{F}_{\text{reject}}^{\min}(i,j)$ are highlighted by a black frame. Note that according to (4.6), we can make the following simultaneous confidence statement about the intervals plotted in panels (d) of Figures 3–6: we can claim, with confidence of about 95%, that there is a difference between the functions λ_i and λ_j on each of these intervals.

We now have a closer look at the results in Figures 3–6. Figure 3 presents the comparison of Germany with Italy. The two time series of daily new cases in panel (a) can be seen to be very similar until approximately day 40. Thereafter, the German time series appears to trend downwards more strongly than the Italian one. The smoothed data in panel (b) give a similar visual impression: the kernel estimates of the German and Italian trend curves λ_i and λ_j are very close to each other until approximately day 40 but then start to differ. It is however not clear

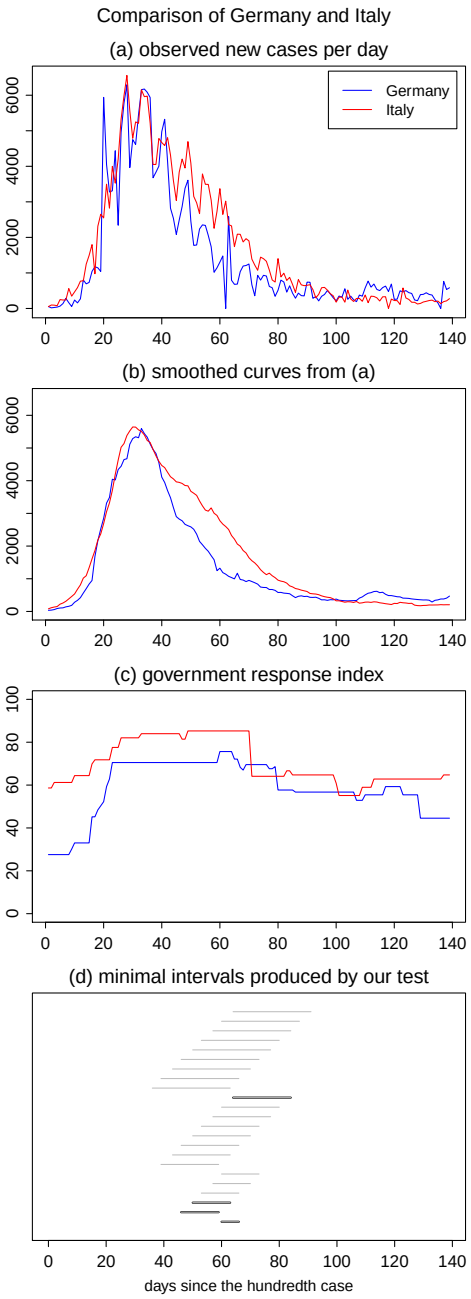


Figure 3: Test results for the comparison of Germany and Italy.

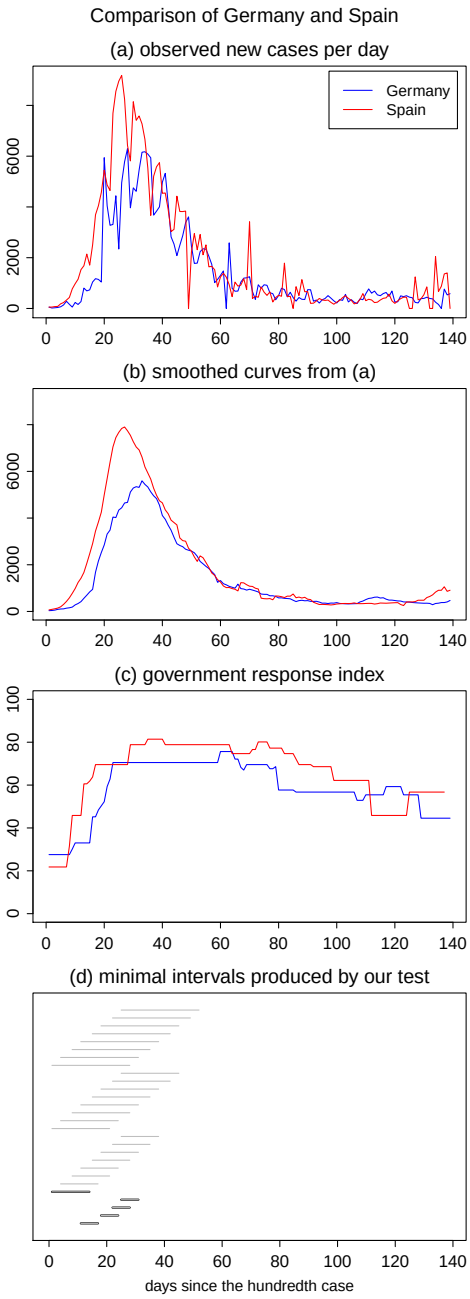


Figure 4: Test results for the comparison of Germany and Spain.

Note: In each figure, panel (a) shows the two observed time series, panel (b) smoothed versions of the time series, and panel (c) the corresponding Government Response Index (GRI). Panel (d) depicts the set of intervals $\mathcal{F}_{\text{reject}}(i,j)$ in grey and the subset of minimal intervals $\mathcal{F}^{\text{min}}_{\text{reject}}(i,j)$ with a black frame.

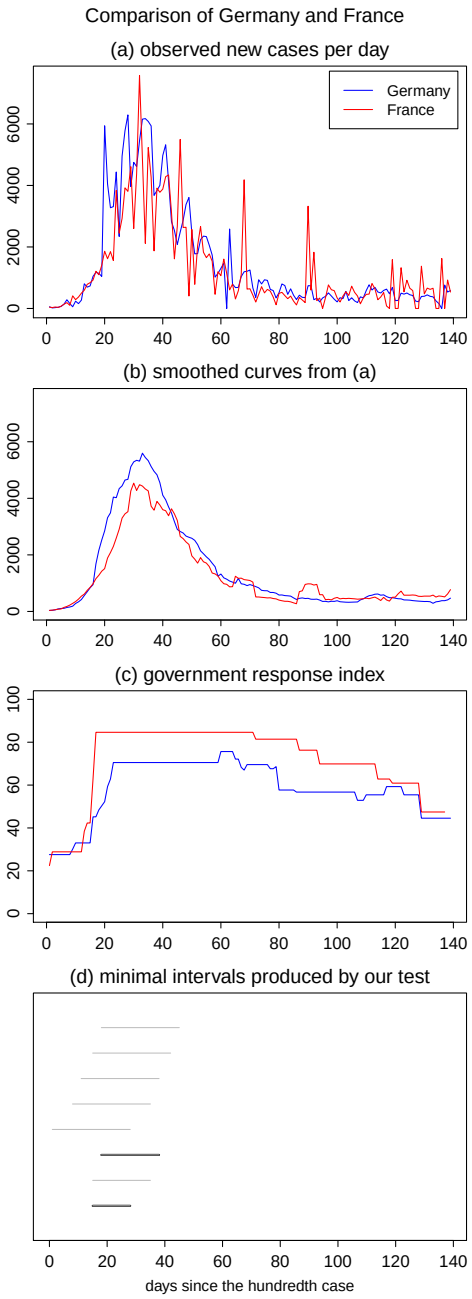


Figure 5: Test results for the comparison of Germany and France.

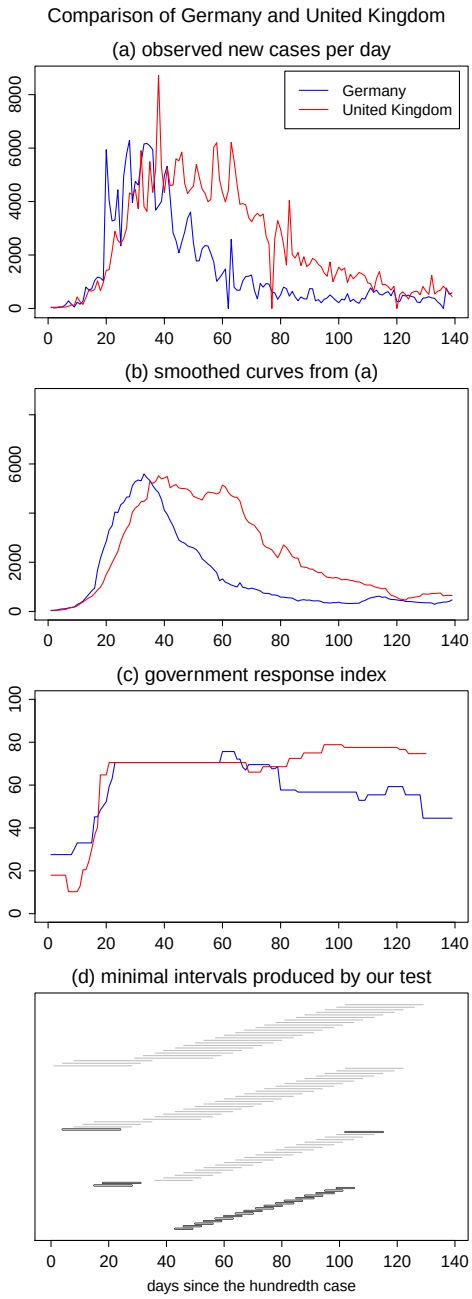


Figure 6: Test results for the comparison of Germany and the UK.

Note: In each figure, panel (a) shows the two observed time series, panel (b) smoothed versions of the time series, and panel (c) the corresponding Government Response Index (GRI). Panel (d) depicts the set of intervals $\mathcal{F}_{\text{reject}}(i,j)$ in grey and the subset of minimal intervals $\mathcal{F}^{\text{min}}_{\text{reject}}(i,j)$ with a black frame.

whether the differences between the two curve estimates reflect differences between the underlying trend curves or whether these are mere artefacts of sampling noise. Our test allows to clarify this issue. Inspecting panel (d), we see that the test detects significant differences between the trend curves in the time period between day 36 and 91. However, it does not find any significant differences up to day 36. Taken together, our results provide evidence that the epidemic developed very similarly in Germany and Italy until a peak was reached around day 40. Thereafter, however, the German time series exhibits a significantly stronger downward trend than the Italian one.

Inspecting Figures 4 and 5, a quite different picture arises when comparing Germany with France and Spain. The test detects significant differences between the German trend and the trends in France and Spain up to (approximately) day 50 but not thereafter. Hence, we find that the time trends evolve differently during the outbreak of the crisis, but they appear to decrease in more or less the same fashion after a peak was reached. Finally, the comparison of Germany with the UK in Figure 6 reveals significant differences between the time trends over essentially the whole observation window. Inspecting the time series in panel (a), it is quite obvious that the UK trend evolves differently from the German one after day 40. However, our test also detects differences between the trends during the onset of the crisis, which is not obvious from the time series plot in panel (a).

5.2.3 Discussion

Having identified significant differences between the epidemic trends in the five countries under consideration, one may ask next what are the causes of these differences. As already mentioned at the beginning of this section, this question cannot be answered by our test. Rather, a further analysis which presumably goes beyond pure statistics is needed to shed some light on it. We here do not attempt to provide any answers. We merely discuss some observations which become apparent upon considering our test results in the light of the Government Response Index (GRI). For reasons of brevity, we focus on the comparison of Germany with Italy and Spain in Figures 3 and 4.

According to our test results in Figure 4, there are significant differences between the trends in Germany and Spain during the onset of the epidemic up to about day 50, with Spain having more new cases of infections than Germany on most days. After day 50, the trends become quite similar and start to decrease at approximately the same rate. This may be due to the fact that Spain in general introduced more severe measures of lockdown than Germany (as can be seen upon inspecting the GRI in panel (c) of Figure 4), which may have helped to battle the spread of infection.

However, a much more thorough analysis is of course needed to find out whether this is indeed the case or whether other factors were mainly responsible.

Turning to the comparison of Germany and Italy, we found that the German trend drops down significantly faster than the Italian one after approximately day 40. Interestingly, the GRI of Italy almost always lies above that of Germany. Hence, even though Italy has in general taken more severe and restrictive measures against the virus than Germany, it appears that the virus could be contained better in Germany (in the sense that the trend of daily new cases went down significantly faster in Germany than in Italy). This suggests that there are indeed important factors besides the level of government response to the pandemic which substantially influence the trend of new COVID-19 cases.

This brief discussion already indicates that it is extremely difficult to determine the exact causes of the differences in epidemic trends across countries. Since even similar countries such as those in our sample differ in a variety of aspects that are relevant for the spread of the virus, it is very challenging to pin down these causes. One issue that is often discussed in the context of cross-country comparisons are country-specific strategies to test for the coronavirus. The argument is that differences between epidemic trends may be spuriously produced by country-specific test procedures.

Even though we can of course not fully exclude this possibility, our test results are presumably not driven by different test regimes in the countries under consideration. To see this, we consider again the comparison of Germany and Italy: The test regimes in these two countries are arguably quite different. Germany is often cited as the country that employed early, widespread testing with more than 100 000 tests per week even in the beginning of the pandemic (Cohen and Kupferschmidt, 2020), while testing in Italy became widespread only in the late stages of the pandemic. Nevertheless, visual inspection of the raw and smoothed data in panels (a) and (b) of Figure 3 suggest that the underlying time trends are very similar up to day 36. This is confirmed by our multiscale test which does not find any significant differences before that day. Hence, the different test regimes in Germany and Italy towards the beginning of the pandemic do not appear to have an overly strong effect and to produce spurious differences between the time trends. This suggests that the differences detected by our multiscale test indeed reflect differences in the way the virus spread in Germany and Italy rather than being mere artefacts of different test regimes.

6 Possible issues

- Adjusting to the size of the country: dividing by population.

  Obviously, the size of the country matters. However, it is not straightforward to normalise the data properly. Simply dividing by the population of the given country would not be sufficient in our case, because the outbreak of the virus in a specific country is a quite local event (Wuhan in China, Lombardy in Italy, etc.). Clearly, this locality does not play any role in the later stages of the epidemic because additional local outbreaks occur over time and their number will presumably be larger in countries with a larger population. So, overall, the relationship between the population size and the epidemic curve can be very complicated. As for our paper, in the application sections we consider the European countries that are fairly similar (in population size but also in other characteristics), we opted for not normalising the data by population size at all rather than "wrongly normalising".

- Building in some policy changes.

  Need to think how to do that correctly. Look at Alfano and Ercolano (2020) and Courtemanche et al. (2020).

- Measurement error.

- Why not case-fatality rate?

A Appendix

In what follows, we state and prove the main theoretical results on the multiscale test developed in Section 4. Throughout the Appendix, we let C be a generic positive constant that may take a different value on each occurrence. Unless stated differently, C depends neither on the time series length T nor on the dimension p of the test problem. We further use the symbols $h_{\min} := \min_{1\leq k\leq K} h_k$ and $h_{\max} := \max_{1\leq k\leq K} h_k$ to denote the smallest and largest interval length in the family $\mathcal{F}$.

Theorem A.1. *Let (C1) and (C2) be satisfied. Moreover, assume that (i)* $h_{\max} = o(1/\log T)$*, (ii)* $h_{\min} \geq CT^{-b}$ *for some* $b \in (0,1)$*, and (iii)* $p = O(T^{(\theta/2)(1-b)-(1+\delta)})$ *for some small* $\delta > 0$*. Then for any given* $\alpha \in (0,1)$*,*

$$\mathrm{FWER}(\alpha) := \mathbb{P}\Big(\exists (i,j,k) \in \mathcal{M}_0 : |\hat{\psi}_{ijk,T}| > c_{T,\mathrm{Gauss}}(\alpha, h_k)\Big) \leq \alpha + o(1),$$

where $\mathcal{M}_0 \subseteq \mathcal{M}$ *is the set of all* $(i,j,k) \in \mathcal{M}$ *for which* $H_0^{(ijk)}$ *holds true.*

According to Theorem A.1, the multiscale test asymptotically controls the FWER at level α under conditions (C1)–(C2) and the restrictions (i)–(iii) on $h_{\min}$, $h_{\max}$ and p. Restriction (i) allows the maximal interval length $h_{\max}$ to converge to zero very slowly, which means that $h_{\max}$ can be picked very large in practice. According to restriction (ii), the minimal interval length $h_{\min}$ can be chosen to go to zero as any polynomial T^{-b} with some $b \in (0,1)$. Restriction (iii) allows the dimension p of the test problem to grow polynomially in T. Specifically, p may grow at most as the polynomial T^{γ} with $\gamma = (\theta/2)(1-b) - (1+\delta)$. As one can see, the exponent γ depends on the number of error moments θ defined in (C2) and the parameter b that specifies the minimal interval length $h_{\min}$. In particular, for any given $b \in (0,1)$, the exponent γ gets larger as θ increases. Hence, the larger the number of error moments θ, the faster p may grow in comparison to T. In the extreme case where all error moments exist, that is, where θ can be made as large as desired, p may grow as any polynomial of T, no matter how we pick $b \in (0,1)$. Thus, if the error terms have sufficiently many moments, the dimension p can be extremely large in comparison to T and the minimal interval length $h_{\min}$ can be chosen very small.

The following corollary is an immediate consequence of Theorem A.1. It provides the theoretical justification needed to make simultaneous confidence statements of the form (4.7)–(4.9).

Corollary A.1. *Under the conditions of Theorem A.1,*

$$\mathbb{P}\Big(\forall (i,j,k) \in \mathcal{M} : \text{ If } |\hat{\psi}_{ijk,T}| > c_{T,\mathrm{Gauss}}(\alpha, h_k), \text{ then } (i,j,k) \notin \mathcal{M}_0\Big) \geq 1 - \alpha + o(1)$$

for any given $\alpha \in (0,1)$*.*

Proof of Theorem A.1. The proof proceeds in several steps.

Step 1. Let $\hat{\Psi}_T = \max_{(i,j,k)\in\mathcal{M}} a_k(|\hat{\psi}^0_{ijk,T}| - b_k)$ with $\hat{\psi}^0_{ijk,T}$ introduced in (4.2) and define $\Psi_T = \max_{(i,j,k)\in\mathcal{M}} a_k(|\psi^0_{ijk,T}| - b_k)$ with

$$\psi^0_{ijk,T} = \frac{1}{\sqrt{2Th_k}} \sum_{t=1}^{T} \mathbf{1}\Big(\frac{t}{T} \in \mathcal{I}_k\Big)(\eta_{it} - \eta_{jt}).$$

To start with, we prove that

$$\big|\hat{\Psi}_T - \Psi_T\big| = o_p(r_T), \tag{A.1}$$

where $\{r_T\}$ is any null sequence that converges more slowly to zero than $\rho_T = \sqrt{\log T}\{\log p/\sqrt{Th_{\min}} + h_{\max}\sqrt{\log p}\}$, that is, $\rho_T/r_T \to 0$ as $T \to \infty$. Since the proof of (A.1) is rather technical and lengthy, the details are provided in the Supplementary Material.

Step 2. We next prove that

$$\sup_{q\in\mathbb{R}} \Big|\mathbb{P}\big(\Psi_T \le q\big) - \mathbb{P}\big(\Phi_T \le q\big)\Big| = o(1). \tag{A.2}$$

To do so, we rewrite the statistics Ψ_T and Φ_T as follows: Define

$$V_t^{(ijk)} = V_{t,T}^{(ijk)} := \sqrt{\frac{T}{2Th_k}}\mathbf{1}\Big(\frac{t}{T} \in \mathcal{I}_k\Big)(\eta_{it} - \eta_{jt})$$

for $(i,j,k) \in \mathcal{M}$ and let $\boldsymbol{V}_t = (V_t^{(ijk)} : (i,j,k) \in \mathcal{M})$ be the p-dimensional random vector with the entries $V_t^{(ijk)}$. With this notation, we get that $\psi^0_{ijk,T} = T^{-1/2}\sum_{t=1}^{T} V_t^{(ijk)}$ and thus

$$\begin{aligned}\Psi_T &= \max_{(i,j,k)\in\mathcal{M}} a_k\big(|\psi^0_{ijk,T}| - b_k\big)\\ &= \max_{(i,j,k)\in\mathcal{M}} a_k\Big\{\Big|\frac{1}{\sqrt{T}}\sum_{t=1}^{T} V_t^{(ijk)}\Big| - b_k\Big\}.\end{aligned}$$

Analogously, we define

$$W_t^{(ijk)} = W_{t,T}^{(ijk)} := \sqrt{\frac{T}{2Th_k}}\mathbf{1}\Big(\frac{t}{T} \in \mathcal{I}_k\Big)(Z_{it} - Z_{jt})$$

with Z_{it} i.i.d. standard normal and let $\boldsymbol{W}_t = (W_t^{(ijk)} : (i,j,k) \in \mathcal{M})$. The vector $\boldsymbol{W}_t$ is a Gaussian version of $\boldsymbol{V}_t$ with the same mean and variance. In particular, $\mathbb{E}[\boldsymbol{W}_t] = \mathbb{E}[\boldsymbol{V}_t] = 0$ and $\mathbb{E}[\boldsymbol{W}_t\boldsymbol{W}_t^\top] = \mathbb{E}[\boldsymbol{V}_t\boldsymbol{V}_t^\top]$. Similarly as before, we can write

$\phi_{ijk,T} = T^{-1/2}\sum_{t=1}^{T} W_t^{(ijk)}$ and

$$\begin{aligned}\Phi_T &= \max_{(i,j,k)\in\mathcal{M}} a_k\big(|\phi_{ijk,T}| - b_k\big)\\ &= \max_{(i,j,k)\in\mathcal{M}} a_k\Big\{\Big|\frac{1}{\sqrt{T}}\sum_{t=1}^{T} W_t^{(ijk)}\Big| - b_k\Big\}.\end{aligned}$$

For any $q \in \mathbb{R}$, it holds that

$$\begin{aligned}\mathbb{P}\big(\Psi_T \le q\big) &= \mathbb{P}\Big(\max_{(i,j,k)\in\mathcal{M}} a_k\Big\{\Big|\frac{1}{\sqrt{T}}\sum_{t=1}^{T} V_t^{(ijk)}\Big| - b_k\Big\} \le q\Big)\\ &= \mathbb{P}\Big(\Big|\frac{1}{\sqrt{T}}\sum_{t=1}^{T} V_t^{(ijk)}\Big| \le c_{ijk}(q) \text{ for all } (i,j,k)\in\mathcal{M}\Big)\\ &= \mathbb{P}\Big(\Big|\frac{1}{\sqrt{T}}\sum_{t=1}^{T} \boldsymbol{V}_t\Big| \le \boldsymbol{c}(q)\Big),\end{aligned}$$

where $\boldsymbol{c}(q) = (c_{ijk}(q) : (i,j,k)\in\mathcal{M})$ is the $\mathbb{R}^p$-vector with the entries $c_{ijk}(q) = q/a_k + b_k$, we use the notation $|v| = (|v_1|,\dots,|v_p|)^\top$ for vectors $v \in \mathbb{R}^p$ and the inequality $v \le w$ is to be understood componentwise for $v, w \in \mathbb{R}^p$. Analogously, we have

$$\mathbb{P}\big(\Phi_T \le q\big) = \mathbb{P}\Big(\Big|\frac{1}{\sqrt{T}}\sum_{t=1}^{T} \boldsymbol{W}_t\Big| \le \boldsymbol{c}(q)\Big).$$

With this notation at hand, we can make use of Proposition 2.1 from Chernozhukov et al. (2017). In our context, this proposition can be stated as follows:

Proposition A.1. *Assume that*

(a) $T^{-1}\sum_{t=1}^{T}\mathbb{E}(V_t^{(ijk)})^2 \ge \delta > 0$ *for all* $(i,j,k)\in\mathcal{M}$.

(b) $T^{-1}\sum_{t=1}^{T}\mathbb{E}[|V_t^{(ijk)}|^{2+r}] \le B_T^r$ *for all* $(i,j,k)\in\mathcal{M}$ *and* $r = 1,2$, *where* $B_T \ge 1$ *are constants that may tend to infinity as* $T\to\infty$.

(c) $\mathbb{E}[\{\max_{(i,j,k)\in\mathcal{M}}|V_t^{(ijk)}|/B_T\}^\theta] \le 2$ *for all* t *and some* $\theta > 4$.

Then

$$\begin{aligned}\sup_{\boldsymbol{c}\in\mathbb{R}^p}\Big|\mathbb{P}\Big(\Big|\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\boldsymbol{V}_t\Big| \le \boldsymbol{c}\Big) - \mathbb{P}\Big(\Big|\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\boldsymbol{W}_t\Big| \le \boldsymbol{c}\Big)\Big|&\\ \le C\Big\{\Big(\frac{B_T^2\log^7(pT)}{T}\Big)^{1/6} + \Big(\frac{B_T^2\log^3(pT)}{T^{1-2/\theta}}\Big)^{1/3}\Big\},&\quad (A.3)\end{aligned}$$

where C *depends only on* δ *and* θ.

It is straightforward to verify that assumptions (a)–(c) are satisfied under the conditions of Theorem A.1 for sufficiently large T, where B_T can be chosen as $B_T = Cp^{1/\theta}h_{\min}^{-1/2}$ with C sufficiently large. Moreover, it can be shown that the right-hand side of (A.3) is $o(1)$ for this choice of B_T. Hence, Proposition A.1 yields that

$$\sup_{\boldsymbol{c}\in\mathbb{R}^p}\Big|\mathbb{P}\Big(\Big|\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\boldsymbol{V}_t\Big|\le\boldsymbol{c}\Big)-\mathbb{P}\Big(\Big|\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\boldsymbol{W}_t\Big|\le\boldsymbol{c}\Big)\Big|=o(1),$$

which in turn implies (A.2).

Step 3. With the help of (A.1) and (A.2), we now show that

$$\sup_{q\in\mathbb{R}}\Big|\mathbb{P}\big(\hat{\Psi}_T\le q\big)-\mathbb{P}\big(\Phi_T\le q\big)\Big|=o(1). \tag{A.4}$$

To start with, the above supremum can be bounded by

$$\begin{aligned}
&\sup_{q\in\mathbb{R}}\Big|\mathbb{P}\big(\hat{\Psi}_T\le q\big)-\mathbb{P}\big(\Phi_T\le q\big)\Big|\\
&=\sup_{q\in\mathbb{R}}\Big|\mathbb{P}\Big(\Psi_T\le q+\big\{\Psi_T-\hat{\Psi}_T\big\}\Big)-\mathbb{P}\big(\Phi_T\le q\big)\Big|\\
&\le\sup_{q\in\mathbb{R}}\max\Big\{\Big|\mathbb{P}\Big(\Psi_T\le q+\big|\Psi_T-\hat{\Psi}_T\big|\Big)-\mathbb{P}\big(\Phi_T\le q\big)\Big|,\\
&\qquad\qquad\Big|\mathbb{P}\Big(\Psi_T\le q-\big|\Psi_T-\hat{\Psi}_T\big|\Big)-\mathbb{P}\big(\Phi_T\le q\big)\Big|\Big\}\\
&\le\sup_{q\in\mathbb{R}}\max\Big\{\Big|\mathbb{P}\Big(\Psi_T\le q+r_T\Big)-\mathbb{P}\big(\Phi_T\le q\big)\Big|+\mathbb{P}\Big(\big|\Psi_T-\hat{\Psi}_T\big|>r_T\Big),\\
&\qquad\qquad\Big|\mathbb{P}\Big(\Psi_T\le q-r_T\Big)-\mathbb{P}\big(\Phi_T\le q\big)\Big|+\mathbb{P}\Big(\big|\Psi_T-\hat{\Psi}_T\big|>r_T\Big)\Big\}\\
&\le\max_{\ell=0,1}\sup_{q\in\mathbb{R}}\Big|\mathbb{P}\Big(\Psi_T\le q+(-1)^\ell r_T\Big)-\mathbb{P}\big(\Phi_T\le q\big)\Big|+\mathbb{P}\Big(\big|\Psi_T-\hat{\Psi}_T\big|>r_T\Big)\\
&=\max_{\ell=0,1}\sup_{q\in\mathbb{R}}\Big|\mathbb{P}\Big(\Psi_T\le q+(-1)^\ell r_T\Big)-\mathbb{P}\big(\Phi_T\le q\big)\Big|+o(1),
\end{aligned}\tag{A.5}$$

where the last line is by (A.1). Moreover, for $\ell=0,1$,

$$\begin{aligned}
&\sup_{q\in\mathbb{R}}\Big|\mathbb{P}\Big(\Psi_T\le q+(-1)^\ell r_T\Big)-\mathbb{P}\big(\Phi_T\le q\big)\Big|\\
&\le\sup_{q\in\mathbb{R}}\Big|\mathbb{P}\Big(\Psi_T\le q+(-1)^\ell r_T\Big)-\mathbb{P}\Big(\Phi_T\le q+(-1)^\ell r_T\Big)\Big|\\
&\quad+\sup_{q\in\mathbb{R}}\Big|\mathbb{P}\Big(\Phi_T\le q+(-1)^\ell r_T\Big)-\mathbb{P}\big(\Phi_T\le q\big)\Big|\\
&=\sup_{q\in\mathbb{R}}\Big|\mathbb{P}\Big(\Phi_T\le q+(-1)^\ell r_T\Big)-\mathbb{P}\big(\Phi_T\le q\big)\Big|+o(1),
\end{aligned}\tag{A.6}$$

the last line following from (A.2). Finally, by Nazarov's inequality (cp. Nazarov,

2003 and Lemma A.1 in Chernozhukov et al., 2017), we have that for $\ell = 0, 1$,

$$\begin{aligned}
&\sup_{q\in\mathbb{R}} \Big|\mathbb{P}\Big(\Phi_T \le q + (-1)^\ell r_T\Big) - \mathbb{P}\big(\Phi_T \le q\big)\Big| \\
&= \sup_{q\in\mathbb{R}} \Big|\mathbb{P}\Big(\Big|\frac{1}{\sqrt{T}}\sum_{t=1}^{T} \boldsymbol{W}_t\Big| \le \boldsymbol{c}(q + (-1)^\ell r_T)\Big) - \mathbb{P}\Big(\Big|\frac{1}{\sqrt{T}}\sum_{t=1}^{T} \boldsymbol{W}_t\Big| \le \boldsymbol{c}(q)\Big)\Big| \\
&\le C r_T \sqrt{\log(2p)}, \qquad \text{(A.7)}
\end{aligned}$$

where C is a constant that depends only on the parameter δ defined in condition (a) of Proposition A.1. Inserting (A.6) and (A.7) into equation (A.5) completes the proof of (A.4).

Step 4. By definition of the quantile $q_{T,\mathrm{Gauss}}(\alpha)$, it holds that $\mathbb{P}(\Phi_T \le q_{T,\mathrm{Gauss}}(\alpha)) \ge 1 - \alpha$. As shown in the Supplementary Material, we even have that

$$\mathbb{P}(\Phi_T \le q_{T,\mathrm{Gauss}}(\alpha)) = 1 - \alpha \qquad \text{(A.8)}$$

for any $\alpha \in (0, 1)$. From this and (A.4), it immediately follows that

$$\mathbb{P}\big(\hat{\Psi}_T \le q_{T,\mathrm{Gauss}}(\alpha)\big) = 1 - \alpha + o(1), \qquad \text{(A.9)}$$

which in turn implies that

$$\begin{aligned}
\mathrm{FWER}(\alpha) &= \mathbb{P}\Big(\exists (i,j,k) \in \mathcal{M}_0 : |\hat{\psi}_{ijk,T}| > c_{T,\mathrm{Gauss}}(\alpha, h_k)\Big) \\
&= \mathbb{P}\Big(\max_{(i,j,k)\in\mathcal{M}_0} a_k\big(|\hat{\psi}_{ijk,T}| - b_k\big) > q_{T,\mathrm{Gauss}}(\alpha)\Big) \\
&= \mathbb{P}\Big(\max_{(i,j,k)\in\mathcal{M}_0} a_k\big(|\hat{\psi}^0_{ijk,T}| - b_k\big) > q_{T,\mathrm{Gauss}}(\alpha)\Big) \\
&\le \mathbb{P}\Big(\max_{(i,j,k)\in\mathcal{M}} a_k\big(|\hat{\psi}^0_{ijk,T}| - b_k\big) > q_{T,\mathrm{Gauss}}(\alpha)\Big) \\
&= \mathbb{P}\big(\hat{\Psi}_T > q_{T,\mathrm{Gauss}}(\alpha)\big) = \alpha + o(1).
\end{aligned}$$

This completes the proof of Theorem A.1. □

Proof of Corollary A.1. By Theorem A.1,

$$\begin{aligned}
1 - \alpha + o(1) &\le 1 - \mathrm{FWER}(\alpha) \\
&= \mathbb{P}\Big(\nexists (i,j,k) \in \mathcal{M}_0 : |\hat{\psi}_{ijk,T}| > c_{T,\mathrm{Gauss}}(\alpha, h_k)\Big) \\
&= \mathbb{P}\Big(\forall (i,j,k) \in \mathcal{M} : \text{ If } |\hat{\psi}_{ijk,T}| > c_{T,\mathrm{Gauss}}(\alpha, h_k), \text{ then } (i,j,k) \notin \mathcal{M}_0\Big),
\end{aligned}$$

which gives the statement of Corollary A.1. □

References

Alfano, V. and Ercolano, S. (2020). The efficacy of lockdown against covid-19: A cross-country panel analysis. *Applied Health Economics and Health Policy* 1.

Chen, L. and Wu, W. B. (2019). Testing for trends in high-dimensional time series. *Journal of the American Statistical Association*, **114** 869–881.

Chernozhukov, V., Chetverikov, D. and Kato, K. (2017). Central limit theorems and bootstrap in high dimensions. *Annals of Probability*, **45** 2309–2352.

Cohen, J. and Kupferschmidt, K. (2020). Countries test tactics in 'war' against COVID-19. *Science*, **367** 1287–1288.

Courtemanche, C., Garuccio, J., Le, A., Pinkston, J. and Yelowitz, A. (2020). Strong social distancing measures in the united states reduced the covid-19 growth rate: Study evaluates the impact of social distancing measures on the growth rate of confirmed covid-19 cases across the united states. *Health Affairs* 10–1377.

Cox, D. R. (1983). Some remarks on overdispersion. *Biometrika*, **70** 269–274.

De Salazar, P. M., Niehus, R., Taylor, A., Buckee, C. and Lipsitch, M. (2020). Using predicted imports of 2019-nCoV cases to determine locations that may not be identifying all imported cases. *medRxiv*.

Degras, D., Xu, Z., Zhang, T. and Wu, W. B. (2012). Testing for parallelism among trends in multiple time series. *IEEE Transactions on Signal Processing*, **60** 1087–1097.

Delgado, M. A. (1993). Testing the equality of nonparametric regression curves. *Statistics & Probability Letters*, **17** 199–204.

Dümbgen, L. and Spokoiny, V. G. (2001). Multiscale testing of qualitative hypotheses. *Annals of Statistics*, **29** 124–152.

Dümbgen, L. and Walther, G. (2008). Multiscale inference about a density. *Annals of Statistics*, **36** 1758–1785.

Dunker, F., Eckle, K., Proksch, K. and Schmidt-Hieber, J. (2019). Tests for qualitative features in the random coefficients model. *Electronic Journal of Statistics*, **13** 2257–2306.

Eckle, K., Bissantz, N. and Dette, H. (2017). Multiscale inference for multivariate deconvolution. *Electronic Journal of Statistics*, **11** 4179–4219.

Efron, B. (1986). Double exponential families and their use in generalized linear regression. *Journal of the American Statistical Association*, **81** 709–721.

Fryzlewicz, P., Sapatinas, T. and Subba Rao, S. (2006). A Haar-Fisz technique for locally stationary volatility estimation. *Biometrika*, **93** 687–704.

Fryzlewicz, P., Sapatinas, T. and Subba Rao, S. (2008). Normalized least-squares estimation in time-varying ARCH models. *Annals of Statistics*, **36** 742–786.

Hafner, C. M. and Linton, O. (2010). Efficient estimation of a multivariate multi-

plicative volatility model. *Journal of Econometrics*, **159** 55–73.

Hale, T., Petherick, A., Phillips, T. and Webster, S. (2020a). Variation in government responses to COVID-19. *Blavatnik school of government working paper*, **31**.

Hale, T., Webster, S., Petherick, A., Phillips, T. and Kira, B. (2020b). Oxford COVID-19 government response tracker. Blavatnik school of government. `http://www.bsg.ox.ac.uk/covidtracker`.

Hall, P. and Hart, J. D. (1990). Bootstrap test for difference between means in nonparametric regression. *Journal of the American Statistical Association*, **85** 1039–1049.

Härdle, W. and Marron, J. S. (1990). Semiparametric comparison of regression curves. *Annals of Statistics*, **18** 63–89.

Hidalgo, J. and Lee, J. (2014). A CUSUM test for common trends in large heterogeneous panels. In *Essays in Honor of Peter C. B. Phillips*. Emerald Group Publishing Limited, 303–345.

King, E. C., Hart, J. D. and Wehrly, T. E. (1991). Testing the equality of regression curves using linear smoothers. *Statistics & Probability Letters*, **12** 239–247.

Kulasekera, K. B. (1995). Comparison of regression curves using quasi-residuals. *Journal of the American Statistical Association*, **90** 1085–1093.

Lavergne, P. (2001). An equality test across nonparametric regressions. *Journal of Econometrics*, **103** 307–344.

McCullagh, P. and Nelder, J. (1989). *Generalized linear models*. Chapman and Hall.

Mikosch, T. and Stărică, C. (2000). Is it really long memory we see in financial returns? In *Extremes and Integrated Risk Management* (P. Embrechts, ed.). 149–168.

Mikosch, T. and Stărică, C. (2004). Non-stationarities in financial time series, the long-range dependence, and IGARCH effects. *The Review of Economics and Statistics*, **86** 378–390.

Munk, A. and Dette, H. (1998). Nonparametric comparison of several regression functions: exact and asymptotic theory. *Annals of Statistics*, **26** 2339–2368.

Nazarov, F. (2003). On the maximal perimeter of a convex set in $\mathbb{R}^n$ with respect to a Gaussian measure. In *Geometric Aspects of Functional Analysis*, vol. 1807 of *Lecture Notes in Mathematics*. Springer, 169–187.

Neumeyer, N. and Dette, H. (2003). Nonparametric comparison of regression curves: an empirical process approach. *Annals of Statistics*, **31** 880–920.

Pardo-Fernández, J. C., Van Keilegom, I. and González-Manteiga, W. (2007). Testing for the equality of k regression curves. *Statistica Sinica*, **17** 1115–1137.

Park, C., Vaughan, A., Hannig, J. and Kang, K.-H. (2009). SiZer analysis for the comparison of time series. *Journal of Statistical Planning and Inference*, **139** 3974–3988.

Pellis, L., Scarabel, F., Stage, H. B., Overton, C. E., Chappell, L. H., Lythgoe, K. A., Fearon, E., Bennett, E., Curran-Sebastian, J., Das, R. et al. (2020). Challenges in control of Covid-19: short doubling time and long delay to effect of interventions. *arXiv:2004.00117*.

Robinson, P. M. (1989). Nonparametric estimation of time-varying parameters. In *Statistical Analysis and Forecasting of Economic Structural Change* (P. Hackl, ed.). Springer, 253–264.

Rohde, A. (2008). Adaptive goodness-of-fit tests based on signed ranks. *Annals of Statistics*, **36** 1346–1374.

Rufibach, K. and Walther, G. (2010). The block criterion for multiscale inference about a density, with applications to other multiscale problems. *Journal of Computational and Graphical Statistics*, **19** 175–190.

Schmidt-Hieber, J., Munk, A. and Dümbgen, L. (2013). Multiscale methods for shape constraints in deconvolution: confidence statements for qualitative features. *Annals of Statistics*, **41** 1299–1328.

Tobías, A., Valls, J., Satorra, P. and Tebé, C. (2020). COVID19-Tracker: A shiny app to produce to produce comprehensive data visualization for SARS-CoV-2 epidemic in Spain. *medRxiv*.

Young, S. G. and Bowman, A. W. (1995). Nonparametric analysis of covariance. *Biometrics*, **51** 920–931.

Zhang, Y., Su, L. and Phillips, P. C. B. (2012). Testing for common trends in semiparametric panel data models with fixed effects. *The Econometrics Journal*, **15** 56–100.

S Supplementary Material

S.1 Technical details

In what follows, we provide the technical details omitted in the Appendix. To start with, we prove the following auxiliary lemma.

Lemma S.1. *Under the conditions of Theorem A.1, it holds that*

$$\big|\hat{\sigma}^2 - \sigma^2\big| = O_p\Big(\sqrt{\frac{\log p}{T}}\Big).$$

Proof of Lemma S.1. By definition, $\hat{\sigma}^2 = |\mathcal{C}|^{-1}\sum_{i\in\mathcal{C}}\hat{\sigma}_i^2$ and $\hat{\sigma}_i^2 = \{\sum_{t=2}^T (X_{it} - X_{it-1})^2\}\{2\sum_{t=1}^T X_{it}\}$. It holds that

$$\frac{1}{T}\sum_{t=2}^T (X_{it} - X_{it-1})^2 = \frac{\sigma^2}{T}\sum_{t=2}^T \lambda_i\Big(\frac{t}{T}\Big)(\eta_{it} - \eta_{it-1})^2 + \big\{R_{i,T}^{(1)} + \ldots + R_{i,T}^{(5)}\big\}, \quad \text{(S.1)}$$

where

$$\begin{aligned}
R_{i,T}^{(1)} &= \frac{2\sigma}{T}\sum_{t=2}^T \Big(\lambda_i\Big(\frac{t}{T}\Big) - \lambda_i\Big(\frac{t-1}{T}\Big)\Big)\sqrt{\lambda_i\Big(\frac{t}{T}\Big)}(\eta_{it} - \eta_{it-1}) \\
R_{i,T}^{(2)} &= \frac{2\sigma^2}{T}\sum_{t=2}^T \Big(\sqrt{\lambda_i\Big(\frac{t}{T}\Big)} - \sqrt{\lambda_i\Big(\frac{t-1}{T}\Big)}\Big)\sqrt{\lambda_i\Big(\frac{t}{T}\Big)}\eta_{it-1}(\eta_{it} - \eta_{it-1}) \\
R_{i,T}^{(3)} &= \frac{1}{T}\sum_{t=2}^T \Big(\lambda_i\Big(\frac{t}{T}\Big) - \lambda_i\Big(\frac{t-1}{T}\Big)\Big)^2 \\
R_{i,T}^{(4)} &= \frac{2\sigma}{T}\sum_{t=2}^T \Big(\lambda_i\Big(\frac{t}{T}\Big) - \lambda_i\Big(\frac{t-1}{T}\Big)\Big)\Big(\sqrt{\lambda_i\Big(\frac{t}{T}\Big)} - \sqrt{\lambda_i\Big(\frac{t-1}{T}\Big)}\Big)\eta_{it-1} \\
R_{i,T}^{(5)} &= \frac{\sigma^2}{T}\sum_{t=2}^T \Big(\sqrt{\lambda_i\Big(\frac{t}{T}\Big)} - \sqrt{\lambda_i\Big(\frac{t-1}{T}\Big)}\Big)^2\eta_{it-1}^2.
\end{aligned}$$

With the help of an exponential inequality and standard arguments, it can be shown that

$$\max_{i\in\mathcal{C}}\Big|\frac{1}{T}\sum_{t=2}^T w_i\Big(\frac{t}{T}\Big)\big\{g(\eta_{it},\eta_{it-1}) - \mathbb{E}g(\eta_{it},\eta_{it-1})\big\}\Big| = O_p\Big(\sqrt{\frac{\log p}{T}}\Big),$$

where we let $g(x,y) = x$, $g(x,y) = y$, $g(x,y) = |x|$, $g(x,y) = |y|$, $g(x,y) = x^2$, $g(x,y) = y^2$ or $g(x,y) = xy$, and $w_i(t/T)$ are deterministic weights with the property that $|w_i(t/T)| \le w_{\max} < \infty$ for all i, t and T and some positive constant $w_{\max}$. Using

this uniform convergence result along with conditions (C1) and (C2), we obtain that

$$\max_{i\in\mathcal{C}}\Big|\frac{1}{T}\sum_{t=2}^{T}\lambda_i\Big(\frac{t}{T}\Big)(\eta_{it}-\eta_{it-1})^2-\frac{2}{T}\sum_{t=1}^{T}\lambda_i\Big(\frac{t}{T}\Big)\Big|=O_p\Big(\sqrt{\frac{\log p}{T}}\Big)$$

and

$$\max_{1\le\ell\le 5}\max_{i\in\mathcal{C}}|R_{i,T}^{(\ell)}|=O_p(T^{-1}).$$

Applying these two statements to (S.1), we can infer that

$$\max_{i\in\mathcal{C}}\Big|\frac{1}{T}\sum_{t=2}^{T}(X_{it}-X_{it-1})^2-\frac{2\sigma^2}{T}\sum_{t=1}^{T}\lambda_i\Big(\frac{t}{T}\Big)\Big|=O_p\Big(\sqrt{\frac{\log p}{T}}\Big). \tag{S.2}$$

By similar but simpler arguments, we additionally get that

$$\max_{i\in\mathcal{C}}\Big|\frac{1}{T}\sum_{t=1}^{T}X_{it}-\frac{1}{T}\sum_{t=1}^{T}\lambda_i\Big(\frac{t}{T}\Big)\Big|=O_p\Big(\sqrt{\frac{\log p}{T}}\Big). \tag{S.3}$$

From (S.2) and (S.3), it follows that $\max_{i\in\mathcal{C}}|\hat{\sigma}_i^2-\sigma^2|=O_p(\sqrt{\log p/T})$, which in turn implies that $|\hat{\sigma}^2-\sigma^2|=O_p(\sqrt{\log p/T})$ as well. □

Proof of (A.1). Since

$$\begin{aligned}\big|\hat{\Psi}_T-\Psi_T\big|&\le\max_{(i,j,k)\in\mathcal{M}}a_k\big|\hat{\psi}_{ijk,T}^0-\psi_{ijk,T}^0\big|\\&\le\max_{1\le k\le K}a_k\max_{(i,j,k)\in\mathcal{M}}\big|\hat{\psi}_{ijk,T}^0-\psi_{ijk,T}^0\big|\\&\le C\sqrt{\log T}\max_{(i,j,k)\in\mathcal{M}}\big|\hat{\psi}_{ijk,T}^0-\psi_{ijk,T}^0\big|,\end{aligned}$$

it suffices to prove that

$$\max_{(i,j,k)\in\mathcal{M}}\big|\hat{\psi}_{ijk,T}^0-\psi_{ijk,T}^0\big|=o_p\Big(\frac{r_T}{\sqrt{\log T}}\Big). \tag{S.4}$$

To start with, we reformulate $\hat{\psi}_{ijk,T}^0$ as

$$\hat{\psi}_{ijk,T}^0=\hat{\psi}_{ijk,T}^*+\Big(\frac{\sigma}{\hat{\sigma}}-1\Big)\hat{\psi}_{ijk,T}^*,$$

where

$$\hat{\psi}_{ijk,T}^*=\frac{\sum_{t=1}^{T}\mathbf{1}(\frac{t}{T}\in\mathcal{I}_k)\overline{\lambda}_{ij}^{1/2}(\frac{t}{T})(\eta_{it}-\eta_{jt})}{\{\sum_{t=1}^{T}\mathbf{1}(\frac{t}{T}\in\mathcal{I}_k)(X_{it}+X_{jt})\}^{1/2}}.$$

With this notation, we can establish the bound

$$\begin{aligned}\max_{(i,j,k)\in\mathcal{M}} \big|\hat{\psi}^0_{ijk,T} - \psi^0_{ijk,T}\big| \leq \max_{(i,j,k)\in\mathcal{M}} \big|\hat{\psi}^*_{ijk,T} - \psi^0_{ijk,T}\big| \\ + \Big|\frac{\sigma}{\hat{\sigma}} - 1\Big| \max_{(i,j,k)\in\mathcal{M}} \big|\hat{\psi}^*_{ijk,T} - \psi^0_{ijk,T}\big| \\ + \Big|\frac{\sigma}{\hat{\sigma}} - 1\Big| \max_{(i,j,k)\in\mathcal{M}} \big|\psi^0_{ijk,T}\big|,\end{aligned}$$

which shows that (S.4) is implied by the three statements

$$\max_{(i,j,k)\in\mathcal{M}} \big|\hat{\psi}^*_{ijk,T} - \psi^0_{ijk,T}\big| = O_p\Big(\frac{\log p}{\sqrt{Th_{\min}}} + h_{\max}\sqrt{\log p}\Big) \tag{S.5}$$

$$\max_{(i,j,k)\in\mathcal{M}} \big|\psi^0_{ijk,T}\big| = O_p\big(\sqrt{\log p}\big) \tag{S.6}$$

$$\big|\hat{\sigma}^2 - \sigma^2\big| = O_p\Big(\sqrt{\frac{\log p}{T}}\Big). \tag{S.7}$$

Since (S.7) has already been verified in Lemma S.1, it remains to prove the statements (S.5) and (S.6).

We start with the proof of (S.6). Applying an exponential inequality along with standard arguments yields that

$$\max_{i\in\mathcal{C}} \max_{1\leq k\leq K} \Big|\frac{1}{\sqrt{Th_k}} \sum_{t=1}^{T} \mathbf{1}\Big(\frac{t}{T} \in \mathcal{I}_k\Big) w_i\Big(\frac{t}{T}\Big)\eta_{it}\Big| = O_p\big(\sqrt{\log p}\big), \tag{S.8}$$

where $w_i(t/T)$ are general deterministic weights with the property that $|w_i(t/T)| \leq w_{\max} < \infty$ for all i, t and T and some positive constant $w_{\max}$. This immediately implies (S.6).

We next turn to the proof of (S.5). As the functions λ_i are uniformly Lipschitz continuous by (C1), it can be shown that

$$\max_{i\in\mathcal{C}} \max_{1\leq k\leq K} \Big|\frac{1}{Th_k} \sum_{t=1}^{T} \mathbf{1}\Big(\frac{t}{T} \in \mathcal{I}_k\Big)\lambda_i\Big(\frac{t}{T}\Big) - \frac{1}{h_k}\int_{w\in\mathcal{I}_k} \lambda_i(w)dw\Big| \leq \frac{C}{Th_{\min}}. \tag{S.9}$$

From this, the uniform convergence result (S.8) and condition (C1), we can infer that

$$\begin{aligned}\max_{(i,j,k)\in\mathcal{M}} \Big|\frac{1}{Th_k} \sum_{t=1}^{T} \mathbf{1}\Big(\frac{t}{T} \in \mathcal{I}_k\Big)(X_{it} + X_{jt}) \\ - \frac{1}{h_k}\int_{w\in\mathcal{I}_k} \big\{\lambda_i(w) + \lambda_j(w)\big\}dw\Big| = O_p\Big(\sqrt{\frac{\log p}{Th_{\min}}}\Big)\end{aligned} \tag{S.10}$$

and

$$\max_{(i,j,k)\in\mathcal{M}} \Big| \frac{1}{\sqrt{Th_k}} \sum_{t=1}^{T} \mathbf{1}\Big(\frac{t}{T} \in \mathcal{I}_k\Big) \overline{\lambda}_{ij}^{1/2}\Big(\frac{t}{T}\Big)(\eta_{it} - \eta_{jt}) \\ - \Big\{ \frac{\int_{w\in\mathcal{I}_k} \overline{\lambda}_{ij}(w)dw}{h_k} \Big\}^{1/2} \frac{1}{\sqrt{Th_k}} \sum_{t=1}^{T} \mathbf{1}\Big(\frac{t}{T} \in \mathcal{I}_k\Big)(\eta_{it} - \eta_{jt}) \Big| \\ = O_p\Big(h_{\max}\sqrt{\log p}\Big). \qquad \text{(S.11)}$$

The claim (S.5) follows from (S.10) and (S.11) along with straightforward calculations. □

Proof of (A.8). The proof is by contradiction. Suppose that (A.8) does not hold true, that is, $\mathbb{P}(\Phi_T \leq q_{T,\text{Gauss}}(\alpha)) = 1 - \alpha + \xi$ for some $\xi > 0$. By Nazarov's inequality,

$$\mathbb{P}\big(\Phi_T \leq q_{T,\text{Gauss}}(\alpha)\big) - \mathbb{P}\big(\Phi_T \leq q_{T,\text{Gauss}}(\alpha) - \eta\big) \leq C\eta\sqrt{\log(2p)}$$

for any $\eta > 0$ with C depending only on the parameter δ specified in condition (a) of Proposition A.1. Hence,

$$\begin{aligned} \mathbb{P}\big(\Phi_T \leq q_{T,\text{Gauss}}(\alpha) - \eta\big) &\geq \mathbb{P}\big(\Phi_T \leq q_{T,\text{Gauss}}(\alpha)\big) - C\eta\sqrt{\log(2p)} \\ &= 1 - \alpha + \xi - C\eta\sqrt{\log(2p)} > 1 - \alpha \end{aligned}$$

for $\eta > 0$ sufficiently small. This contradicts the definition of the quantile $q_{T,\text{Gauss}}(\alpha)$ according to which $q_{T,\text{Gauss}}(\alpha) = \inf_{q\in\mathbb{R}}\{\mathbb{P}(\Phi_T \leq q) \geq 1 - \alpha\}$. □

S.2 Additional graphs for Section 5.2

Here, we provide the pairwise comparisons between Italy, France, Spain and the UK that were omitted in Section 5.2. The plots have the same format as Figures 3–6.

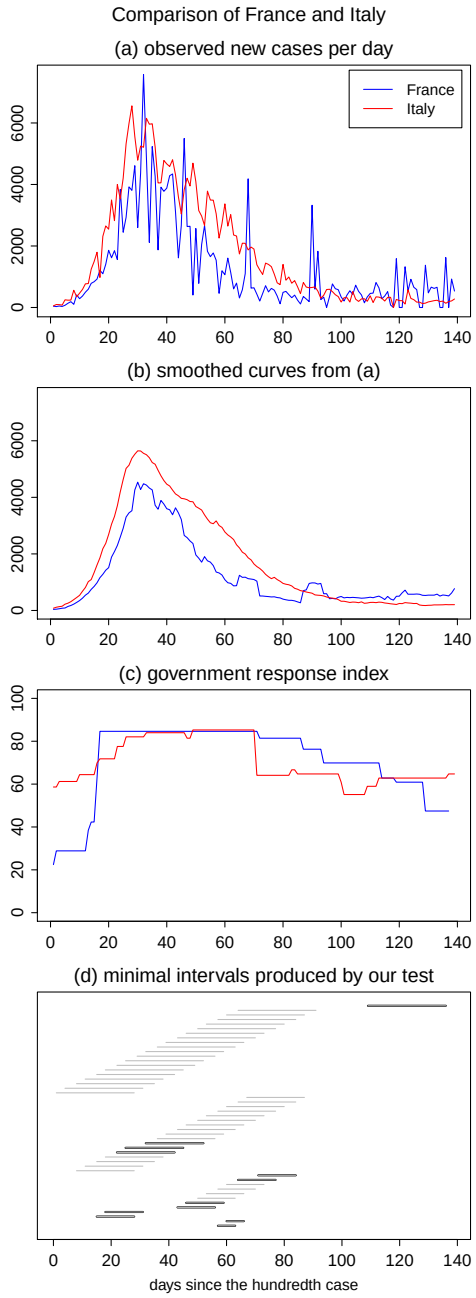


Figure S.1: Test results for the comparison of France and Italy.

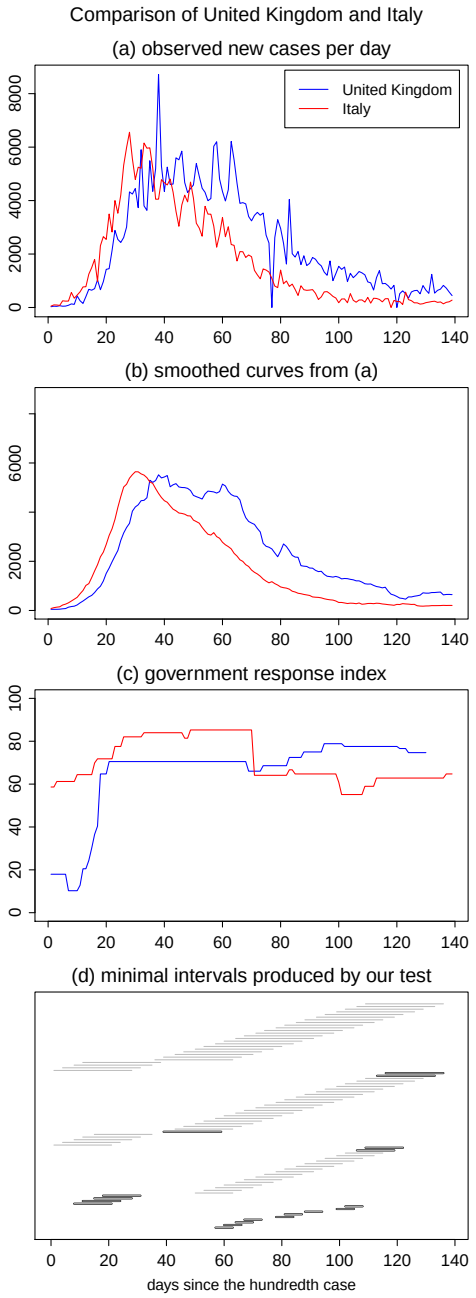


Figure S.2: Test results for the comparison of the UK and Italy.

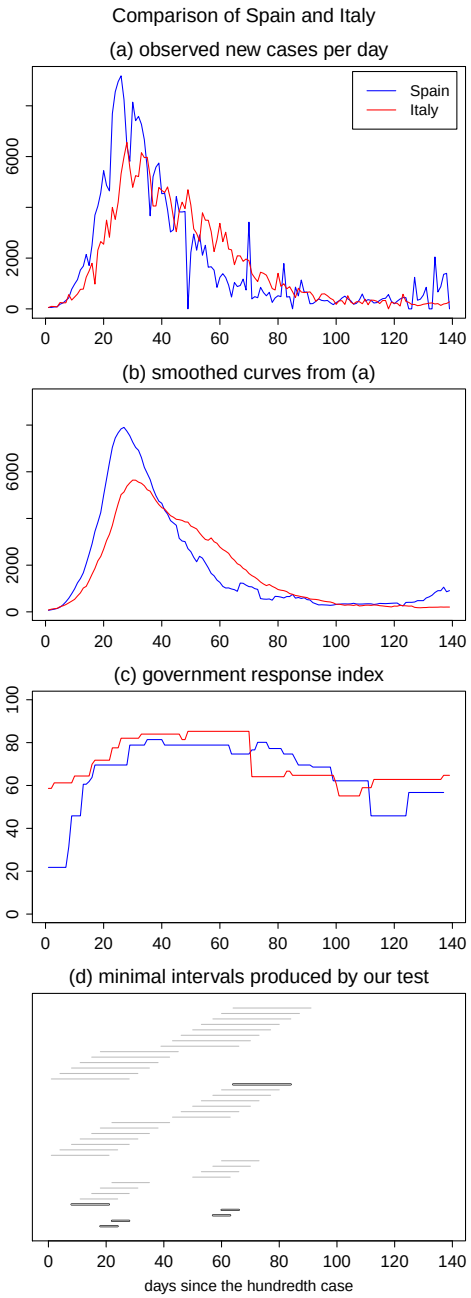


Figure S.3: Test results for the comparison of Spain and Italy.

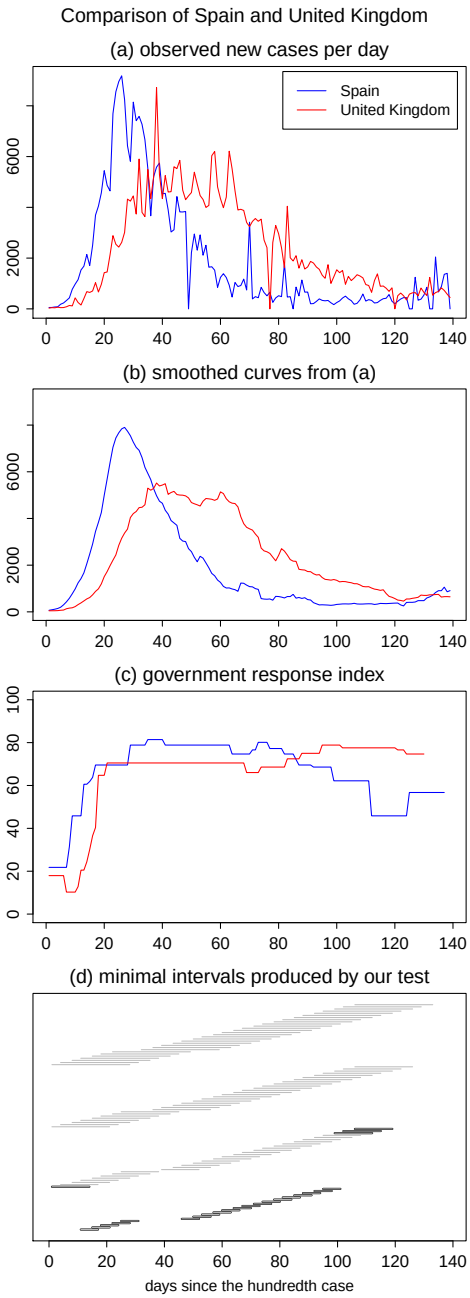


Figure S.4: Test results for the comparison of Spain and the UK.

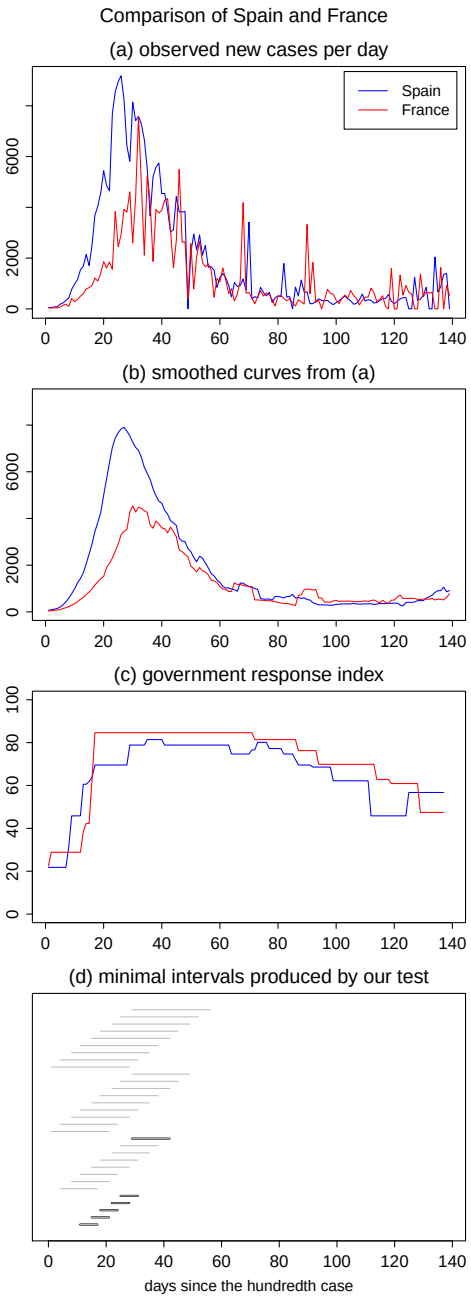


Figure S.5: Test results for the comparison of Spain and France.

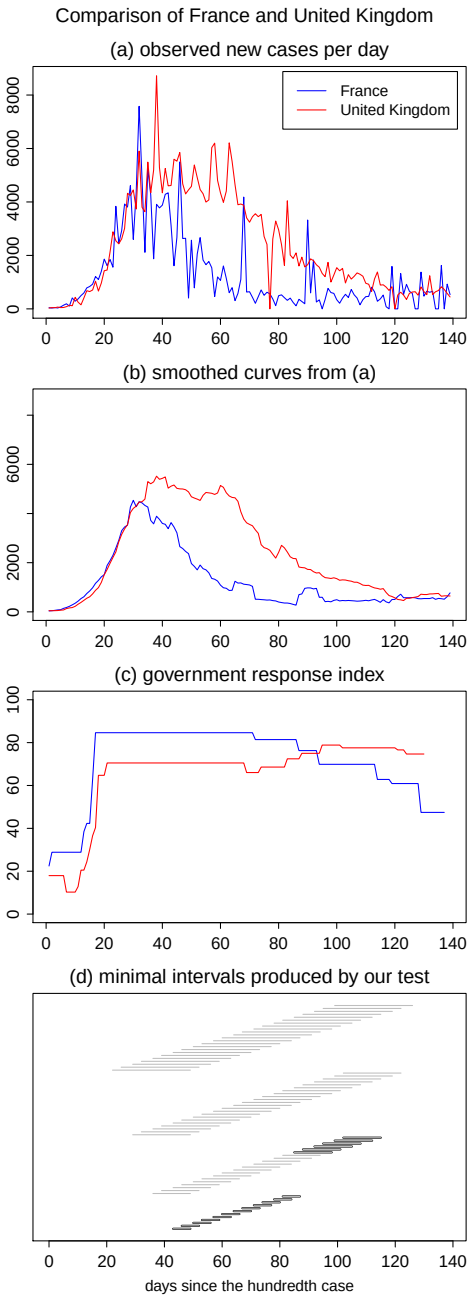


Figure S.6: Test results for the comparison of France and the UK.

S.3 Robustness checks for Section 5.1

In what follows, we supplement the simulation experiments of Section 5.1 by some robustness checks. Specifically, we repeat the experiments with different values of the overdispersion parameter σ. The larger we choose σ, the more noise we put on top of the time trend, that is, on top of the underlying signal. Hence, by varying σ, we can assess how sensitive our test is to changes in the noise-to-signal ratio. We first repeat the size simulations for $\sigma = 10$ and $\sigma = 20$. The results are presented in Tables S.1 and S.2, respectively. As can be seen, the empirical size numbers are very similar to those for $\sigma = 15$ in Table 1. We next rerun the power simulations for $\sigma = 10$ and $\sigma = 20$, where we consider the two Scenarios A and B as in Section 5.1. The results can be found in Tables S.3–S.6. They show that the test is much more powerful for $\sigma = 10$ than for $\sigma = 20$. This is what one would expect, since a higher value of σ corresponds to a higher noise-to-signal ratio. In particular, the higher σ, the more noisy the data, and thus the more difficult it is to identify differences between the trend curves. Nevertheless, even in the very noisy case with $\sigma = 20$, our test has quite some power, which tends to increase swiftly as T gets larger.

Table S.1: Empirical size of the test for $\sigma = 10$.

	$n = 5$			$n = 10$			$n = 50$		
	significance level α			significance level α			significance level α		
	0.01	0.05	0.1	0.01	0.05	0.1	0.01	0.05	0.1
$T = 100$	0.009	0.043	0.085	0.008	0.039	0.075	0.005	0.023	0.055
$T = 250$	0.011	0.047	0.095	0.010	0.050	0.094	0.009	0.039	0.079
$T = 500$	0.009	0.052	0.101	0.013	0.049	0.101	0.010	0.039	0.084

Table S.2: Empirical size of the test for $\sigma = 20$.

	$n = 5$			$n = 10$			$n = 50$		
	significance level α			significance level α			significance level α		
	0.01	0.05	0.1	0.01	0.05	0.1	0.01	0.05	0.1
$T = 100$	0.011	0.050	0.094	0.010	0.047	0.092	0.009	0.034	0.070
$T = 250$	0.009	0.047	0.088	0.008	0.044	0.085	0.006	0.032	0.062
$T = 500$	0.008	0.038	0.081	0.006	0.039	0.079	0.006	0.025	0.060

Table S.3: Power of the test in Scenario A for $\sigma = 10$.

	$n = 5$			$n = 10$			$n = 50$		
	significance level α			significance level α			significance level α		
	0.01	0.05	0.1	0.01	0.05	0.1	0.01	0.05	0.1
$T = 100$	0.836	0.915	0.911	0.833	0.903	0.898	0.777	0.874	0.882
$T = 250$	0.986	0.971	0.938	0.984	0.956	0.918	0.980	0.961	0.924
$T = 500$	0.996	0.975	0.946	0.994	0.965	0.927	0.992	0.963	0.918

Table S.4: Power of the test in Scenario A for $\sigma = 20$.

	$n = 5$			$n = 10$			$n = 50$		
	significance level α			significance level α			significance level α		
	0.01	0.05	0.1	0.01	0.05	0.1	0.01	0.05	0.1
$T = 100$	0.144	0.275	0.352	0.115	0.231	0.304	0.048	0.120	0.163
$T = 250$	0.244	0.434	0.538	0.204	0.403	0.486	0.133	0.247	0.305
$T = 500$	0.296	0.563	0.662	0.273	0.511	0.603	0.175	0.338	0.433

Table S.5: Power of the test in Scenario B for $\sigma = 10$.

	$n = 5$			$n = 10$			$n = 50$		
	significance level α			significance level α			significance level α		
	0.01	0.05	0.1	0.01	0.05	0.1	0.01	0.05	0.1
$T = 100$	0.991	0.973	0.946	0.994	0.970	0.935	0.994	0.971	0.940
$T = 250$	0.993	0.969	0.941	0.993	0.959	0.919	0.991	0.960	0.925
$T = 500$	0.996	0.976	0.948	0.993	0.966	0.928	0.993	0.962	0.917

Table S.6: Power of the test in Scenario B for $\sigma = 20$.

	$n = 5$			$n = 10$			$n = 50$		
	significance level α			significance level α			significance level α		
	0.01	0.05	0.1	0.01	0.05	0.1	0.01	0.05	0.1
$T = 100$	0.438	0.636	0.704	0.404	0.598	0.669	0.277	0.449	0.526
$T = 250$	0.864	0.934	0.927	0.850	0.923	0.915	0.811	0.891	0.898
$T = 500$	0.960	0.968	0.949	0.961	0.964	0.935	0.945	0.961	0.941