

Simultaneous statistical inference for epidemic trends

Marina Khismatullina
University of Bonn

Michael Vogt
University of Bonn

1 Introduction

There are many questions surrounding the current COVID-19 pandemic that are not well understood yet. One important question is the following: How does the outbreak pattern of COVID-19 compare across countries? Are the time trends of daily new infections comparable across countries, or is the virus spreading differently in different regions of the world? The main aim of this paper is to develop new statistical methods that help to shed light on this issue.

Let X_{it} be the number of new infections on day t in country i and suppose we observe a sample of data $\mathcal{X}_i = \{X_{it} : 1 \leq t \leq T\}$ for n different countries i . A simple way 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 $\lambda_i(t/T) = \mathbb{E}[X_{it}] = \text{Var}(X_{it})$, we can model the observations X_{it} by the nonparametric regression equation

$$X_{it} = \lambda_i\left(\frac{t}{T}\right) + u_{it} \quad (1.1)$$

for $1 \leq t \leq T$, where $u_{it} = X_{it} - \mathbb{E}[X_{it}]$ with $\mathbb{E}[u_{it}] = 0$ and $\text{Var}(u_{it}) = \lambda_i(t/T)$. 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, $\lambda_i : [0, 1] \rightarrow \mathbb{R}$ can be regarded as a function on the unit interval, which allows us to estimate it by 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 model (1.1), the outbreak pattern of COVID-19 in country i is determined by the intensity function λ_i . Hence, the question whether the outbreak patterns 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 intensity functions λ_i . More specifically, let $\mathcal{F} = \{\mathcal{I}_k \subseteq [0, 1] : k = 1, \dots, K\}$ be a family of (rescaled) time intervals and let $H_0^{(ijk)}$ be the hypothesis that the intensity

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

Our method contributes to the literature on statistical tests for comparing non-parametric regression and trend curves. Examples include the tests of Härdle and Marron (1990), Hall and Hart (1990), King et al. (1991), Delgado (1993), Kulasekera (1995), Young and Bowman (1995), Munk and Dette (1998), 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. Sections 2 and 3 lay out the statistical methodology. In Section 2, we introduce the model setting in detail which underlies our analysis, while in Section 3, we develop the multiscale test step by step. To

keep the presentation as clear as possible, the technical details are deferred to the Appendix and the Supplementary Material. Section 4 contains the empirical part of the paper. There, we apply our methods to a sample of current COVID-19 data. In addition, we run some simulation experiments to demonstrate that the multiscale test has the formal properties predicted by the theory. Even though our multiscale test is motivated by the current COVID-19 crisis, its applicability is by no means restricted to this specific issue. Indeed, it is a general method to compare nonparametric trends in epidemiological count data. In Section [number], we show that it is even useful beyond this epidemiological context, for example, in the context of financial volatility modelling. We briefly discuss this as well as some further extensions of our methods in Section [number].

2 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

$$X_{it} = \lambda_i\left(\frac{t}{T}\right) + u_{it} \quad \text{with} \quad u_{it} = \sqrt{\lambda_i\left(\frac{t}{T}\right)}\eta_{it}, \quad (2.1)$$

where η_{it} has zero mean and unit variance. In this model, both the mean and the noise variance are described by the same function λ_i . In empirical applications, however, the noise 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

$$X_{it} = \lambda_i\left(\frac{t}{T}\right) + \varepsilon_{it} \quad \text{with} \quad \varepsilon_{it} = \sigma\sqrt{\lambda_i\left(\frac{t}{T}\right)}\eta_{it}, \quad (2.2)$$

where σ is a scaling factor that allows the noise 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 (2.2), 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. For example, De Salazar et al. (2020) assume that the observed case count in country i follows a Poisson distribution with parameter λ_i being proportional to the daily air travel volume of the i -th location. In our paper, we relax the assumption of De Salazar et al. (2020) that λ_i stays constant over time. Similar, but less general

approach to modelling the number of new cases of COVID-19 was used in Pellis et al. (2020). They also consider quasi-Poisson model for the observed data with a constant parameter of overdispersion, but they parametrically restrict the mean function to be exponentially growing with a constant growth rate. Tobías et al. (2020) analyze the data on the accumulated number of cases using quasi-Poisson regression with an interaction model to estimate the change in trend. However, all these approaches lack the ability to capture day-to-day time variation, whereas our testing strategy allows us to do it without imposing any additional structure on the mean function λ_i .

In order to derive our theoretical results, we impose the following regularity conditions on model (2.2):

- (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 λ_{\min} and λ_{\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 , $\mathbb{E}[\eta_{it}] = 0$, $\mathbb{E}[\eta_{it}^2] = 1$ 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 error variance in model (2.2) equals zero if λ_i is equal to zero. Since we normalize our test statistics by an estimate of the error variance, we need the latter 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 modelling framework: The time series process $\mathcal{X}_i = \{X_{it} : 1 \leq t \leq T\}$ produced by model (2.2) is nonstationary for each i . Specifically, both the mean $\mathbb{E}[X_{it}] = \lambda_i(t/T)$ and the variance $\text{Var}(X_{it}) = \sigma^2 \lambda_i(t/T)$ 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. for example 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). 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.

3 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 intensity 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 $\mathcal{M} := \mathcal{S} \times \{1, \dots, K\}$ 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; in particular, it amounts to the number of tests we simultaneously carry out. 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.

3.1 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

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

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

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

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

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

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. As shown in the Supplementary Material, $\hat{\sigma}^2$ 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

$$\hat{\psi}_{ijk,T} := \frac{\hat{s}_{ijk,T}}{\hat{\nu}_{ijk,T}} = \frac{\sum_{t=1}^T \mathbf{1}(\frac{t}{T} \in \mathcal{I}_k)(X_{it} - X_{jt})}{\hat{\sigma} \{\sum_{t=1}^T \mathbf{1}(\frac{t}{T} \in \mathcal{I}_k)(X_{it} + X_{jt})\}^{1/2}}, \quad (3.1)$$

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

$$\hat{\psi}_{ijk,T}^0 = \frac{\sum_{t=1}^T \mathbf{1}(\frac{t}{T} \in \mathcal{I}_k) \sigma \bar{\lambda}_{ij}^{1/2}(\frac{t}{T})(\eta_{it} - \eta_{jt})}{\hat{\sigma} \{\sum_{t=1}^T \mathbf{1}(\frac{t}{T} \in \mathcal{I}_k)(X_{it} + X_{jt})\}^{1/2}} \quad (3.2)$$

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

3.2 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 at level α is

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

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. 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

$$1 - \mathbb{P}\left(\max_{(i,j,k) \in \mathcal{M}_0} |\hat{\psi}_{ijk,T}| \leq c_T(\alpha)\right) \leq \alpha. \quad (3.3)$$

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

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

where $\hat{\psi}_{ijk,T}^0$ has been introduced in (3.2). (Note that both the statistic $\tilde{\Psi}_T$ and the quantile $c_T(\alpha)$ depend on p 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 $a_k = \{\log(e/h_k)\}^{1/2}/\log \log(e/h_k)$ 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

$$\begin{aligned} \text{FWER}(\alpha) &= \mathbb{P}\left(\exists (i, j, k) \in \mathcal{M}_0 : |\hat{\psi}_{ijk,T}| > c_T(\alpha, h_k)\right) \\ &= 1 - \mathbb{P}\left(\forall (i, j, k) \in \mathcal{M}_0 : |\hat{\psi}_{ijk,T}| \leq c_T(\alpha, h_k)\right) \\ &= 1 - \mathbb{P}\left(\forall (i, j, k) \in \mathcal{M}_0 : a_k(|\hat{\psi}_{ijk,T}| - b_k) \leq q_T(\alpha)\right) \\ &= 1 - \mathbb{P}\left(\max_{(i,j,k) \in \mathcal{M}_0} a_k(|\hat{\psi}_{ijk,T}| - b_k) \leq q_T(\alpha)\right), \end{aligned} \quad (3.4)$$

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

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

in order to ensure control of the FWER at level α . Comparing (3.4) with (3.3), 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 $\hat{\psi}_{ijk,T}$ 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 with better theoretical properties (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 can not 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

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

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

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

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

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

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 $q_{T,\text{Gauss}}(\alpha)$ 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 $\alpha \in (0, 1)$:

$$\text{For each } (i, j, k) \in \mathcal{M}, \text{ reject } H_0^{(ijk)} \text{ if } |\hat{\psi}_{ijk,T}| > c_{T,\text{Gauss}}(\alpha, h_k), \quad (3.5)$$

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

and $b_k = \sqrt{2\log(1/h_k)}$.

3.3 Formal properties of the test

In Theorem A.1 of the Appendix, we prove that under appropriate regularity conditions, the test defined in (3.5) (asymptotically) controls the familywise error rate $\text{FWER}(\alpha)$ for each pre-specified significance level α . As shown in Corollary A.1, this has the following implication:

$$\begin{aligned} \mathbb{P}\left(\forall (i, j, k) \in \mathcal{M} : \text{ If } |\hat{\psi}_{ijk,T}| > c_{T,\text{Gauss}}(\alpha, h_k), \text{ then } (i, j, k) \notin \mathcal{M}_0\right) \\ \geq 1 - \alpha + o(1), \end{aligned} \quad (3.6)$$

where \mathcal{M}_0 is the set of triples $(i, j, k) \in \mathcal{M}$ for which $H_0^{(ijk)}$ holds true. Verbally, (3.6) can be expressed as follows:

$$\begin{aligned} \text{With (asymptotic) probability at least } 1 - \alpha, \text{ the null } H_0^{(ijk)} \text{ is violated for} \\ \text{all } (i, j, k) \in \mathcal{M} \text{ for which the test rejects } H_0^{(ijk)}. \end{aligned} \quad (3.7)$$

In other words:

$$\begin{aligned} \text{With (asymptotic) probability at least } 1 - \alpha, \text{ the functions } \lambda_i \text{ and } \lambda_j \text{ differ} \\ \text{on the interval } \mathcal{I}_k \text{ for all } (i, j, k) \in \mathcal{M} \text{ for which the test rejects } H_0^{(ijk)}. \end{aligned} \quad (3.8)$$

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

3.4 Implementation of the test in practice

For a given significance level $\alpha \in (0, 1)$, the test procedure defined in (3.5) 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 $N = 1000$) samples of independent standard normal random variables $\{Z_{it}^{(\ell)} : 1 \leq t \leq T, 1 \leq i \leq T\}$ 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 variables $r_{ijk,T} = \mathbf{1}(|\hat{\psi}_{ijk,T}| > c_{T,\text{Gauss}}(\alpha, h_k))$ 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 $(i, j) \in \mathcal{S}$ which shows the intervals \mathcal{I}_k in the set $\mathcal{F}_{\text{reject}}(i, j) = \{\mathcal{I}_k \in \mathcal{F} : r_{ijk,T} = 1\}$, that is, the intervals \mathcal{I}_k for which the test rejects the null $H_0^{(ijk)}$. We in particular use this graphical device to present the test results of our empirical application in Section 4. According to (3.6), we can make the following simultaneous confidence statement about the intervals in $\mathcal{F}_{\text{reject}}(i, j)$:

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 $\mathcal{F}_{\text{reject}}(i, j)$. (3.9)

Hence, we can claim with statistical confidence at least $1 - \alpha$ that the functions λ_i and λ_j differ on each time interval \mathcal{I}_k which is shown in the plots of our graphical device.

As the sample size grows, the number of intervals in the $\mathcal{F}_{\text{reject}}(i, j)$ can also increase rapidly, making graphical depiction of the results of the test not particularly readable. To obtain better visual representation, we propose working with the subset of $\mathcal{F}_{\text{reject}}(i, j)$, namely, the subset of minimal intervals which is constructed as follows: As in Dümbgen (2002), we call an interval $\mathcal{I}_k \in \mathcal{F}_{\text{reject}}(i, j)$ minimal if there is no other interval $\mathcal{I}_{k'} \in \mathcal{F}_{\text{reject}}(i, j)$ with $\mathcal{I}_{k'} \subset \mathcal{I}_k$. We denote the set of minimal intervals by $\mathcal{F}_{\text{reject}}^{\min}(i, j)$.

It is easily seen that as in (3.9), we can claim with statistical confidence at least $1 - \alpha$ that the functions λ_i and λ_j differ on each time interval $\mathcal{I}_k \in \mathcal{F}_{\text{reject}}^{\min}(i, j)$. Thus, the simultaneous confidence statement about the set of minimal intervals still remains true while the graphical representation of the results becomes much more understandable. In particular, we use this way of presenting the test results in our application in Section 4.2.

4 Empirical application to COVID-19 data

We now use our test to analyze the outbreak patterns of the COVID-19 virus. We proceed in two steps. In Section 4.1, we conduct some Monte-Carlo experiments to assess the finite sample performance of our test. Specifically, we run some experiments which show that the test controls the FWER at level α as predicted by the theory. Having established this, we then apply the test to a sample of current

COVID-19 data in Section 4.2. Our multiscale test is implemented in the R package `multiscale`, available on GitHub at [\(insert the link\)](#).

4.1 Simulation experiments

We consider the following simulation setup:

- We simulate count data $\mathcal{X} = \{X_{it} : 1 \leq i \leq n, 1 \leq t \leq T\}$ by drawing X_{it} independently from the double Poisson distribution $dP_{\lambda_i(t/T), \sigma^2}$ with parameters $\lambda_i(t/T)$ and σ^2 . The double Poisson distribution dP_{λ, σ^2} with parameters λ and σ^2 was rigorously introduced by Efron (1986). As shown there, dP_{λ, σ^2} is a probability distribution on $\mathbb{N}_0 := \mathbb{N} \cup \{0\}$ with mean and variance approximately equal to λ and $\sigma^2\lambda$, respectively. Hence, by letting $X_{it} \sim dP_{\lambda_i(t/T), \sigma^2}$, we approximately obtain a nonparametric regression model of the form

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

where the noise variables η_{it} have zero mean and unit variance.

- For the first part of our simulation study, where we estimate the size properties of the test, 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

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

which has a similar shape as some of the estimated curves in the application of Section 4.2. A plot of the function λ is provided in Figure 1a.

- For the second part of our simulation study, where we estimate the power properties of the test, we consider very simple simulation scenario: we assume that only the first mean function is significantly different from the others. Specifically, we assume that $\lambda_i = \lambda_2$ for all $i \neq 1$ and that $\lambda_1 \neq \lambda_2$. In terms of testing, this implies that the hypothesis $H_0^{(ijk)}$ holds true for all (i, j, k) under consideration such that $i \neq 1$ and $j \neq 1$ and there is at least one hypothesis $H_0^{(ijk)}$ with either $i = 1$ or $j = 1$ that does not hold true. We report the estimated power for two different possibilities. In the first simulation setting, the peak of λ_1 is reached at the same point in time as of λ_2 , but the former is higher than the latter:

$$\begin{aligned}\lambda_1(u) &= 6000 \exp\left(-\frac{(10u-3)^2}{2}\right) + 1000, \\ \lambda_2(u) &= 5000 \exp\left(-\frac{(10u-3)^2}{2}\right) + 1000.\end{aligned}$$

The corresponding mean functions are depicted in Figure 2a. In the second simulation setting, the peaks are of the same height, but they are reached at different points in time, meaning that the peak of λ_1 is later in time than the peak of λ_2 :

$$\begin{aligned}\lambda_1(u) &= 5000 \exp\left(-\frac{(9u-3)^2}{2}\right) + 1000, \\ \lambda_2(u) &= 5000 \exp\left(-\frac{(10u-3)^2}{2}\right) + 1000.\end{aligned}$$

The corresponding mean functions are depicted in Figure 2b.

- We set the overdispersion parameter $\sigma = 15$ (the results of the robustness checks for $\sigma = 10$ and $\sigma = 20$ are provided in the Supplementary Material). The estimated value of σ in the application takes similar values.
- We consider different values for T and n , in particular, $T \in \{100, 250, 500\}$ and $n \in \{5, 10, 50\}$.
- 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 with $h_k \in \{7/T, 14/T, 21/T, 28/T\}$. Hence, all intervals in \mathcal{F} have length either 7 days (1 week), 14 days (2 weeks), 21 days (3 weeks), or 28 days (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 $j = 1, 2, \dots$. A graphical presentation of the family \mathcal{F} for $T = 123$ (as in the application) is given in Figure 1b.

We estimate the probability of Type I error (i.e. the test size) by calculating percentage of at least one rejection of the null hypothesis $H_0^{(i,j,k)}$ for some (i, j, k) from $R = 5000$ simulated samples \mathcal{X} of count data.

The results of the size simulations are presented in Table 1. As can be seen from the table, in all scenarios the empirical FWER coincides with the nominal size α . It can also be seen that size numbers are slightly biased downwards, but this bias gets slowly larger as the number of time series increases. The bias does not significantly diminish as we increase the sample size, which can be explained by the following fact: For all the sample sizes considered we use the same family of intervals \mathcal{F} , hence, the length of the smallest interval (7 days, i.e. one week) relative to the rescaled time decreases, which leads to the effective sample size staying at the same level rather than increasing with the sample size. If we would let the length of the smallest interval be dependent on the sample size, this would probably diminish the bias in the empirical size compared to the nominal one. To summarize, (a) in finite samples, the test allows us to control empirical FWER(α) for pre-specified significance level α and (b) the test is somewhat conservative, but this issue may be

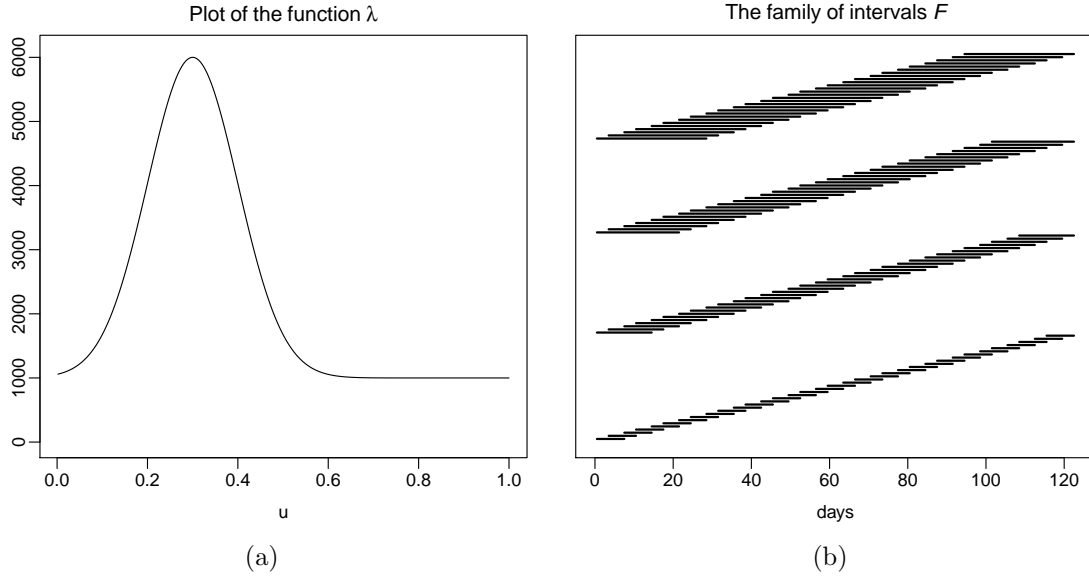


Figure 1: (a) Plot of the function λ ; (b) plot of the family \mathcal{F} .

Table 1: Size of the test for different number of time series $n \in \{5, 10, 50\}$.

	$n = 5$			$n = 10$			$n = 50$		
	nominal size α			nominal size α			nominal size 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

overcome by using a different set of intervals that depends on the sample size (and on the number of time series).

In the second part of our simulation study, we estimate the statistical power of the test. In order to do this, we calculate the percentage of the cases where we correctly found significant difference between the first mean function and one of the others while at the same time did not find any significant differences between all of the other mean functions. That is, based on $R = 5000$ simulated samples \mathcal{X} of count data we calculate the percentage of simulation runs where the set of $\mathcal{F}(1, j)$ is not empty at least for one $j \in \{2, \dots, n\}$ and at the same time all of the other sets $\mathcal{F}(i, j)$ with $\{(i, j) : 2 \leq i < j \leq n\}$ are empty.

The results of the power simulations for the different heights of the peak (see Figure 2a) are presented in Table 2. The results of the power simulations for the different locations of the peak (see Figure 2b) are presented in Table 3. As can be seen, our test is much more powerful when detecting shifts in the underlying trend curves than the difference in their intensity. This is what one would expect: In the second power simulation scenario, the mean functions are significantly different on

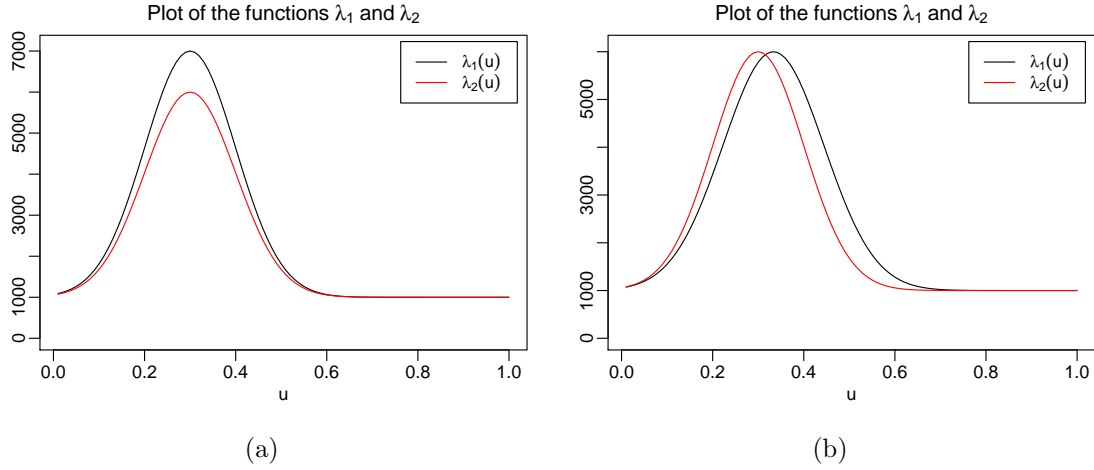


Figure 2: Plot of the functions λ_1 and λ_2 for different power simulation scenarios.

Table 2: Power of the test for different number of time series $n \in \{5, 10, 50\}$ for the first simulation scenario where the mean functions differ by the heights of the peak.

	$n = 5$			$n = 10$			$n = 50$		
	nominal size α			nominal size α			nominal size 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 number of time series $n \in \{5, 10, 50\}$ for the second simulation scenario where the mean functions differ by the locations of the peak.

	$n = 5$			$n = 10$			$n = 50$		
	nominal size α			nominal size α			nominal size 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

wider interval (from $u = 0.2$ to $u = 0.6$), and even on the boundaries the difference is substantial, while in the first power simulation scenario, the main difference is located between $u = 0.2$ and $u = 0.4$.

As can also be seen from the tables, our test is more powerful with smaller number of time series. This is not surprising, because the problem of making simultaneous confidence statements becomes much more difficult if the number of hypotheses under consideration increases substantially. For example, if we keep the family of intervals \mathcal{F} and the sample size T constant, but only increase the number of time series by the factor of two, the number of pairwise comparisons (and hence the number of null hypotheses $H_0^{(i,j,k)}$) increases more than 4 times. With increase by

the factor of 10, we get the increase in comparisons more than 100 times. However, even for $n = 50$ our test has almost 80% power for $T = 500$ and $\alpha = 0.05$.

In the second simulation scenario for sufficiently large sample sizes ($T = 250, 500$), the estimated power is always bigger than 0.9 for all significance levels α and number of time series n . Note that in this setting for $T = 250, 500$ the power numbers get smaller as the significance level α gets larger. This reverse dependance can be explained by the way we calculate power: we count only those cases where we find that the first mean function is different from the others but we do not find any significant difference among all of the others. Hence, we suppose that in this setting the main driving force is the number of false rejections, i.e. the number of cases where the test incorrectly rejects $H_0^{(i,j,k)}$ for some (i, j, k) with $i \neq 1$ and $j \neq 1$. Since this number is controlled by the significant level α , it gets higher as the significance level α increases. And since we exclude these cases from calculating the power, the test become slightly less powerful.

4.2 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 6 July 2020, more than 11.32 million cases of COVID-19 infections have been reported worldwide, resulting in more than 532,000 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 policies have been more efficient 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 statistical rigorous way, but it does not tell what caused these differences. By distinguishing statistically significant differences from artefacts of the sampling noise, the test provides the necessary basis for a further investigation into the causes. Such an investigation, however, goes beyond a mere statistical analysis.

4.2.1 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 6 July 2020. In order to make the data comparable across countries, we take the day of the 100th confirmed case in each country as the starting date $t = 1$. This way of “normalizing” the data is common practice (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 $T = 123$ observations (as of July 6). 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.

4.2.2 Test results

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

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

which was introduced in equation (2.2). The overdispersion parameter σ is estimated by the procedure described in Section 3.1, which yields the estimate $\hat{\sigma} = 14.35$. 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 4.1. 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 $H_0^{(i,j,k)}$ that $\lambda_i(\cdot) = \lambda_j(\cdot)$ 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.3 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 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. Specifically, 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 $H_0^{(i,j,k)}$. The minimal intervals in the subset $\mathcal{F}_{\text{reject}}^{\min}(i, j)$ are depicted in black. Note that according to (3.6), 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 and 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 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. Moreover, our test allows us to localise these differences using the notion of minimal intervals. Specifically, we can say that significant differences between trend curves in Germany and Italy occur between days 46 and 77. Taken together, our results provide evidence that the pandemic 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 Ger-

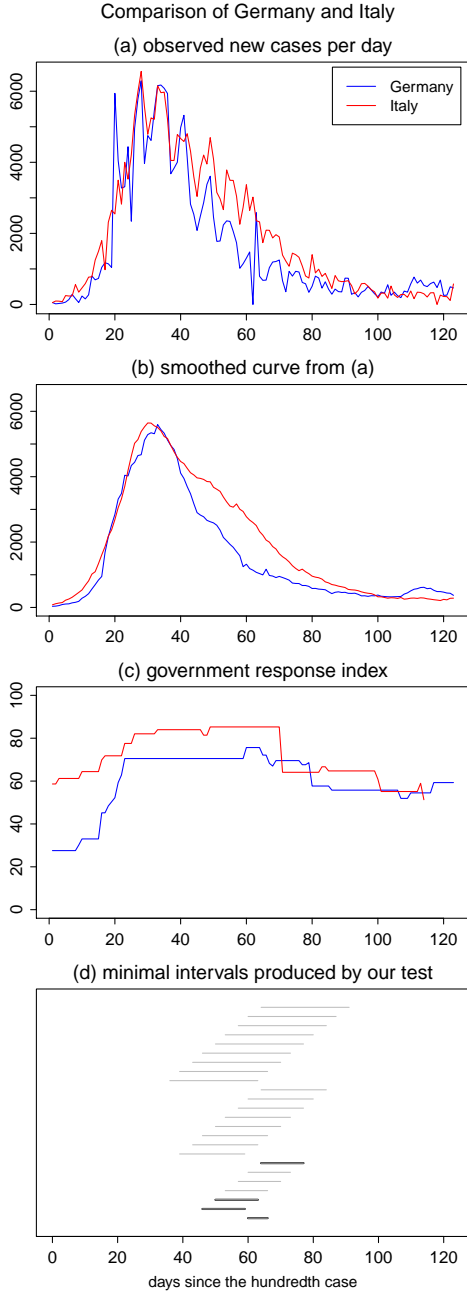


Figure 3: Test results for the comparison of Germany and Italy.

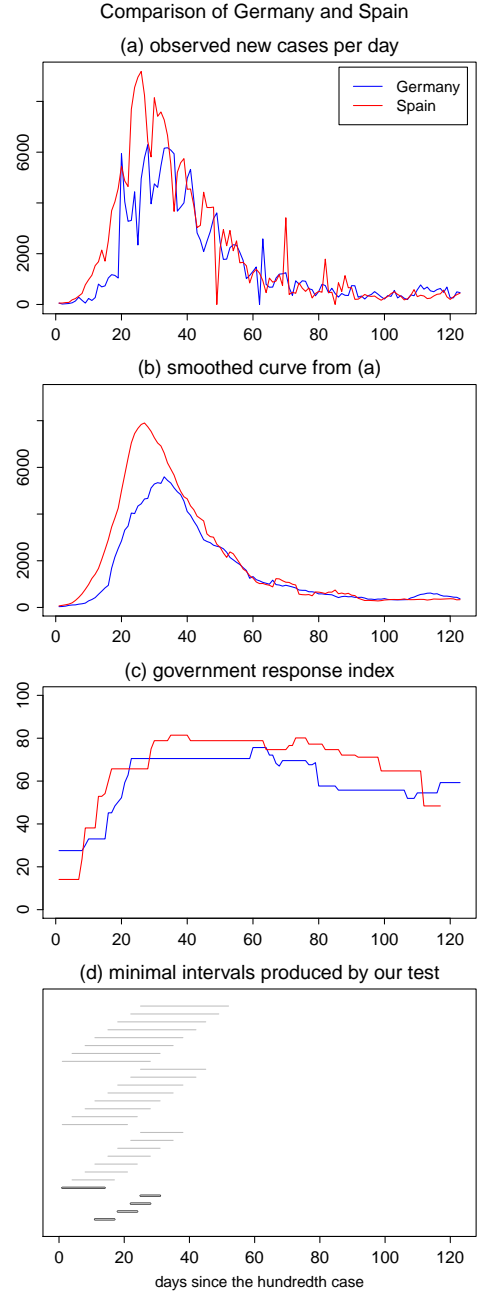


Figure 4: Test results for the comparison of Germany and Spain.

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 $\mathcal{F}_{\text{reject}}^{\min}(i, j)$ in black.

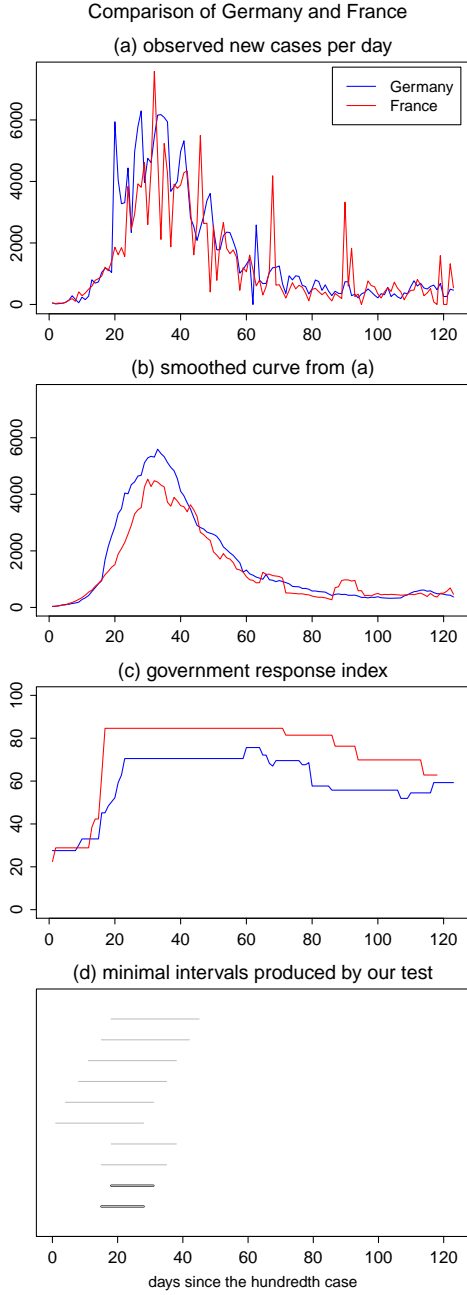


Figure 5: Test results for the comparison of Germany and France.

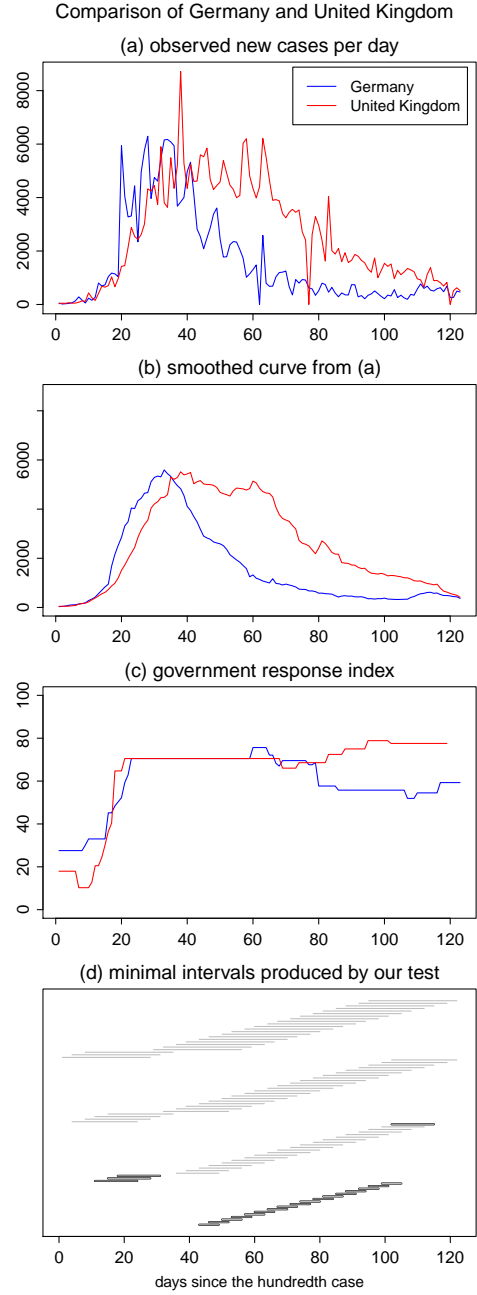


Figure 6: Test results for the comparison of Germany and the UK.

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 $\mathcal{F}_{\text{reject}}^{\min}(i, j)$ in black.

many with France and Spain. The test detects significant differences between the German trend and the trends in France and Spain up to (approximately) day 40 but not thereafter. Hence, we find that the time trends evolve differently during the outbreak of the crisis, but they 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).

4.2.3 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 52, with Spain having more new cases of infections than Germany on most days. After day 52, 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 day 36. 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.

Our discussion 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 test regimes. For example, the UK mostly tested workers on the frontline and people who needed to be hospitalized. Germany, in contrast, is often considered as a country with extensive early testing of contacts to known cases. Differences between epidemic trends may thus 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 explain why, 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 conducting more than 100,000 test 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. Throughout early March, Italy performed approximately the same number of tests per day. 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 test which does not find any significant differences before that day. Hence, the different test regimes in Germany and Italy do not appear to have an overly strong effect and to produce spurious differences between the time trends all over the sampling period. This suggests that the differences detected by our test indeed reflect differences in the way the virus spreads in Germany and Italy rather than being mere artefacts of different test regimes.

A Appendix

In what follows, we state and prove the main theoretical results on the multiscale test developed in Section 3. 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 (C2) and (C1) 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 $\alpha \in (0, 1)$,*

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

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 (C2)–(C1) 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 (C1) 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 (3.7)–(3.9).

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

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

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

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

Step 1. Let $\hat{\Psi}_T = \max_{(i,j,k) \in \mathcal{M}} a_k(|\hat{\psi}_{ijk,T}^0| - b_k)$ with $\hat{\psi}_{ijk,T}^0$ introduced in (3.2) and define $\Psi_T = \max_{(i,j,k) \in \mathcal{M}} a_k(|\psi_{ijk,T}^0| - b_k)$ with

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

To start with, we prove that

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

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, $\rho_T / r_T \rightarrow 0$ as $T \rightarrow \infty$. 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

$$\sup_{q \in \mathbb{R}} \left| \mathbb{P}(\Psi_T \leq q) - \mathbb{P}(\Phi_T \leq q) \right| = o(1). \quad (\text{A.2})$$

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

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

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

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

Analogously, we define

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

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

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

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

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

$$\begin{aligned}\mathbb{P}(\Psi_T \leq q) &= \mathbb{P}\left(\max_{(i,j,k) \in \mathcal{M}} a_k \left\{ \left| \frac{1}{\sqrt{T}} \sum_{t=1}^T V_t^{(ijk)} \right| - b_k \right\} \leq q\right) \\ &= \mathbb{P}\left(\left| \frac{1}{\sqrt{T}} \sum_{t=1}^T V_t^{(ijk)} \right| \leq q_{ijk} \text{ for all } (i,j,k) \in \mathcal{M}\right) \\ &= \mathbb{P}\left(\left| \frac{1}{\sqrt{T}} \sum_{t=1}^T \mathbf{V}_t \right| \leq \mathbf{q}\right),\end{aligned}$$

where \mathbf{q} is the \mathbb{R}^p -vector with the entries $q_{ijk} = q/a_k + b_k$. Analogously, we have

$$\mathbb{P}(\Phi_T \leq q) = \mathbb{P}\left(\left| \frac{1}{\sqrt{T}} \sum_{t=1}^T \mathbf{W}_t \right| \leq \mathbf{q}\right).$$

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) $T^{-1} \sum_{t=1}^T \mathbb{E}(V_t^{(ijk)})^2 \geq \delta > 0$ for all $(i,j,k) \in \mathcal{M}$.
- (b) $T^{-1} \sum_{t=1}^T \mathbb{E}[|V_t^{(ijk)}|^{2+r}] \leq B_T^r$ for all $(i,j,k) \in \mathcal{M}$ and $r = 1, 2$, where $B_T \geq 1$ are constants that may tend to infinity as $T \rightarrow \infty$.
- (c) $\mathbb{E}[\{\max_{(i,j,k) \in \mathcal{M}} |V_t^{(ijk)}|/B_T\}^\theta] \leq 2$ for all t and some $\theta > 4$.

Then

$$\begin{aligned}\sup_{\mathbf{q} \in \mathbb{R}^p} \left| \mathbb{P}\left(\left| \frac{1}{\sqrt{T}} \sum_{t=1}^T \mathbf{V}_t \right| \leq \mathbf{q}\right) - \mathbb{P}\left(\left| \frac{1}{\sqrt{T}} \sum_{t=1}^T \mathbf{W}_t \right| \leq \mathbf{q}\right) \right| \\ \leq C \left\{ \left(\frac{B_T^2 \log^7(pT)}{T} \right)^{1/6} + \left(\frac{B_T^2 \log^3(pT)}{T^{1-2/\theta}} \right)^{1/3} \right\}, \quad (\text{A.3})\end{aligned}$$

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

$$\sup_{\mathbf{q} \in \mathbb{R}^p} \left| \mathbb{P}\left(\left|\frac{1}{\sqrt{T}} \sum_{t=1}^T \mathbf{V}_t\right| \leq \mathbf{q}\right) - \mathbb{P}\left(\left|\frac{1}{\sqrt{T}} \sum_{t=1}^T \mathbf{W}_t\right| \leq \mathbf{q}\right) \right| = o(1),$$

which in turn implies (A.2).

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

$$\sup_{q \in \mathbb{R}} \left| \mathbb{P}(\hat{\Psi}_T \leq q) - \mathbb{P}(\Phi_T \leq q) \right| = o(1). \quad (\text{A.4})$$

To start with, the above supremum can be bounded by

$$\begin{aligned} & \sup_{q \in \mathbb{R}} \left| \mathbb{P}(\hat{\Psi}_T \leq q) - \mathbb{P}(\Phi_T \leq q) \right| \\ &= \sup_{q \in \mathbb{R}} \left| \mathbb{P}\left(\Psi_T \leq q + \{\Psi_T - \hat{\Psi}_T\}\right) - \mathbb{P}(\Phi_T \leq q) \right| \\ &\leq \max \left\{ \sup_{q \in \mathbb{R}} \left| \mathbb{P}\left(\Psi_T \leq q + |\Psi_T - \hat{\Psi}_T|\right) - \mathbb{P}(\Phi_T \leq q) \right|, \right. \\ &\quad \left. \sup_{q \in \mathbb{R}} \left| \mathbb{P}\left(\Psi_T \leq q - |\Psi_T - \hat{\Psi}_T|\right) - \mathbb{P}(\Phi_T \leq q) \right| \right\} \\ &\leq \max \left\{ \sup_{q \in \mathbb{R}} \left| \mathbb{P}\left(\Psi_T \leq q + r_T\right) - \mathbb{P}(\Phi_T \leq q) \right| + \mathbb{P}\left(|\Psi_T - \hat{\Psi}_T| > r_T\right), \right. \\ &\quad \left. \sup_{q \in \mathbb{R}} \left| \mathbb{P}\left(\Psi_T \leq q - r_T\right) - \mathbb{P}(\Phi_T \leq q) \right| + \mathbb{P}\left(|\Psi_T - \hat{\Psi}_T| > r_T\right) \right\}. \end{aligned}$$

Moreover,

$$\begin{aligned} & \sup_{q \in \mathbb{R}} \left| \mathbb{P}\left(\Psi_T \leq q \pm r_T\right) - \mathbb{P}(\Phi_T \leq q) \right| + \mathbb{P}\left(|\Psi_T - \hat{\Psi}_T| > r_T\right) \\ &\leq \sup_{q \in \mathbb{R}} \left| \mathbb{P}\left(\Psi_T \leq q \pm r_T\right) - \mathbb{P}\left(\Phi_T \leq q \pm r_T\right) \right| \\ &\quad + \sup_{q \in \mathbb{R}} \left| \mathbb{P}\left(\Phi_T \leq q \pm r_T\right) - \mathbb{P}(\Phi_T \leq q) \right| + \mathbb{P}\left(|\Psi_T - \hat{\Psi}_T| > r_T\right) \\ &= \sup_{q \in \mathbb{R}} \left| \mathbb{P}\left(\Phi_T \leq q \pm r_T\right) - \mathbb{P}(\Phi_T \leq q) \right| + o(1), \end{aligned}$$

where the last line follows from (A.1) and (A.2). Finally, by Nazarov's inequality (cp. Nazarov, 2003 and Lemma A.1 in Chernozhukov et al., 2017),

$$\begin{aligned} & \sup_{q \in \mathbb{R}} \left| \mathbb{P}\left(\Phi_T \leq q \pm r_T\right) - \mathbb{P}(\Phi_T \leq q) \right| \\ &= \sup_{\mathbf{q} \in \mathbb{R}^p} \left| \mathbb{P}\left(\left|\frac{1}{\sqrt{T}} \sum_{t=1}^T \mathbf{W}_t\right| \leq \mathbf{q} \pm r_T\right) - \mathbb{P}\left(\left|\frac{1}{\sqrt{T}} \sum_{t=1}^T \mathbf{W}_t\right| \leq \mathbf{q}\right) \right| \leq Cr_T \sqrt{\log(2p)}, \end{aligned}$$

where C is a constant that depends only on the parameter δ defined in condition (a) of Proposition A.1.

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

$$\mathbb{P}(\Phi_T \leq q_{T,\text{Gauss}}(\alpha)) = 1 - \alpha \quad (\text{A.5})$$

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

$$\mathbb{P}(\hat{\Psi}_T \leq q_{T,\text{Gauss}}(\alpha)) = 1 - \alpha + o(1), \quad (\text{A.6})$$

which in turn yields the statement of Theorem A.1. \square

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

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

which gives the statement of Corollary A.1. \square

References

- 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. <https://science.sciencemag.org/content/367/6484/1287.full.pdf>, URL <https://science.sciencemag.org/content/367/6484/1287>.
- 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. (2002). Application of local rank tests to nonparametric regression. *Journal of Nonparametric Statistics*, **14** 511–537.
- 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.
- 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.

- 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 preprint 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.
- TOBIÁS, 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.
- 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 semi-parametric panel data models with fixed effects. *The Econometrics Journal*, **15** 56–100.

S Supplementary Material

S.1 Technical proofs

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*

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

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

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

where

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

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

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

where we let $g(x) = x$, $g(x) = |x|$ or $g(x) = x^2$, and $w_i(t/T)$ are 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} . Using this uniform convergence result along with conditions

(C2) and (C1), we obtain that

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

and

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

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

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

By similar but simpler arguments, we additionally get that

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

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. \square

Proof of (A.1). Since

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

it suffices to prove that

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

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

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

where

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

With this notation, we can establish the bound

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

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

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

$$\max_{(i,j,k) \in \mathcal{M}} |\psi_{ijk,T}^0| = O_p(\sqrt{\log p}) \quad (\text{S.6})$$

$$|\hat{\sigma}^2 - \sigma^2| = O_p \left(\sqrt{\frac{\log p}{T}} \right). \quad (\text{S.7})$$

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

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

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 intensity functions λ_i are Lipschitz continuous by (C1), it can be shown that

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

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

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

and

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

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

Proof of (A.5). The proof is by contradiction. Suppose that (A.5) 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,

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

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

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

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

Table 4: Size of the test for different number of time series $n \in \{5, 10, 50\}$, $\sigma = 10$.

	$n = 5$			$n = 10$			$n = 50$		
	nominal size α			nominal size α			nominal size 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 5: Size of the test for different number of time series $n \in \{5, 10, 50\}$, $\sigma = 20$.

	$n = 5$			$n = 10$			$n = 50$		
	nominal size α			nominal size α			nominal size 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

S.2 Robustness checks for Section 4.1

In what follows, we carry out some robustness checks to assess how sensitive our multiscale test is. First, we repeat the size simulations from Section 4.1 for different values of the overdispersion parameter σ . This overdispersion parameter is responsible for noise-to-signal ratio, hence, varying it, we are able to assess the performance of our test when the trend is less or more pronounced compared to the noise. The results of the size simulations for $\sigma = 10$ and $\sigma = 20$ are presented in Tables 4 and 5 respectively. As can be seen from the tables, in all scenarios the empirical FWER is either smaller or approximately equal to the nominal size α . However, as one can expect, our test becomes more conservative with higher values of σ .

We next repeat the simulation exercises of Section 4.1 with the different mean function. Specifically, we take the following function:

$$\lambda(u) = 5000 \exp\left(-\frac{(10u - 3)^2}{2}\right) + 500.$$

The results for different values of the overdispersion parameter σ are presented in Tables 6, 7 and 8. As can be seen, our test appears to be barely influenced by the choice of different mean function.

Next, we repeat the power simulations from Section 4.1 for different values of the overdispersion parameter σ for both simulation scenarios considered in 4.1. The results of the power simulations for $\sigma = 10$ and $\sigma = 20$ for the first setting, where two mean functions are only different by the height of their peak, are presented in Tables 9 and 10 respectively. The results of the power simulations for $\sigma = 10$ and

Table 6: Size of the test for different number of time series $n \in \{5, 10, 50\}$ for different mean function and $\sigma = 10$.

$n = 5$			$n = 10$			$n = 50$		
nominal size α			nominal size α			nominal size α		
0.01	0.05	0.1	0.01	0.05	0.1	0.01	0.05	0.1

Table 7: Size of the test for different number of time series $n \in \{5, 10, 50\}$ for different mean function and $\sigma = 15$.

$n = 5$			$n = 10$			$n = 50$		
nominal size α			nominal size α			nominal size α		
0.01	0.05	0.1	0.01	0.05	0.1	0.01	0.05	0.1

Table 8: Size of the test for different number of time series $n \in \{5, 10, 50\}$ for different mean function and $\sigma = 20$.

$n = 5$			$n = 10$			$n = 50$		
nominal size α			nominal size α			nominal size α		
0.01	0.05	0.1	0.01	0.05	0.1	0.01	0.05	0.1

Table 9: Power of the test for different number of time series $n \in \{5, 10, 50\}$ for the first simulation scenario where the mean functions differ by the heights of the peak with $\sigma = 10$.

	$n = 5$			$n = 10$			$n = 50$		
	nominal size α			nominal size α			nominal size 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

$\sigma = 20$ for the second setting, where two mean functions have the same heights of the peak but reach their respective peak at different points in time, are presented in Tables 11 and 12 respectively. As can be seen, our test is much more powerful for the smaller values of the σ . This is what one would expect, because higher values of σ means that the noise is getting stronger compared to the trend function and the problem of correctly determining the intervals where the mean functions differ become much more complicated. However, even in the presence of the strong noise (in case of $\sigma = 20$), we are able to correctly determine that the first mean function is different from the others but all of the others are the same in more than half of the cases for sufficiently large sample size and small number of time series.

Table 10: Power of the test for different number of time series $n \in \{5, 10, 50\}$ for the first simulation scenario where the mean functions differ by the heights of the peak with $\sigma = 20$.

	$n = 5$			$n = 10$			$n = 50$		
	nominal size α			nominal size α			nominal size 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 11: Power of the test for different number of time series $n \in \{5, 10, 50\}$ for the second simulation scenario where the mean functions differ by the location of the peak with $\sigma = 10$.

	$n = 5$			$n = 10$			$n = 50$		
	nominal size α			nominal size α			nominal size 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 12: Power of the test for different number of time series $n \in \{5, 10, 50\}$ for the second simulation scenario where the mean functions differ by the locations of the peak with $\sigma = 20$.

	$n = 5$			$n = 10$			$n = 50$		
	nominal size α			nominal size α			nominal size 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

S.3 Additional graphs for Section 4.2

In what follows, we provide the results of the pairwise comparison between Italy, Spain, the United Kingdom and France that were omitted in Section 4.2. The results are presented in Figures 7 - 12.

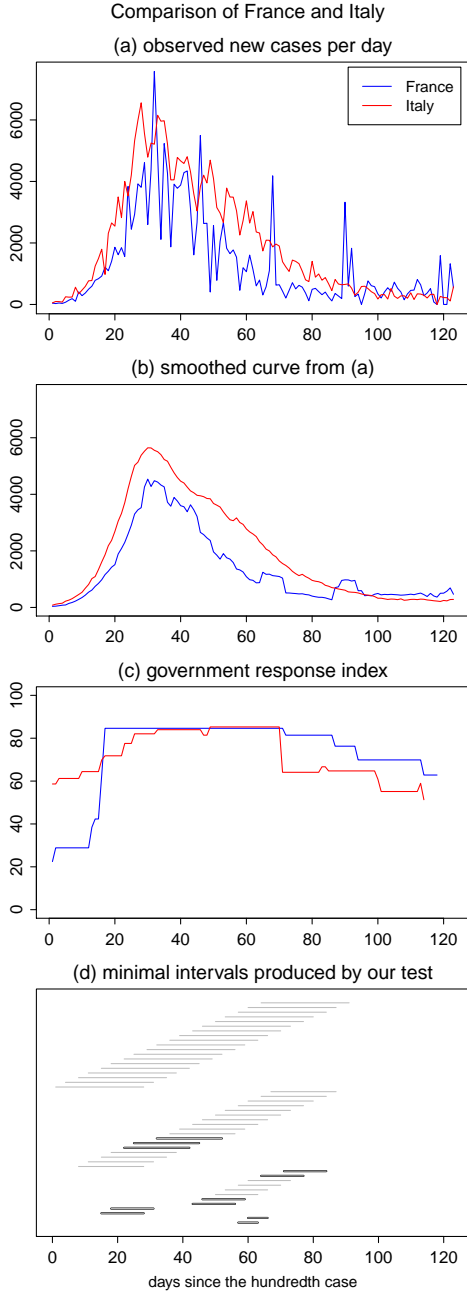


Figure 7: Test results for the comparison of France and Italy.

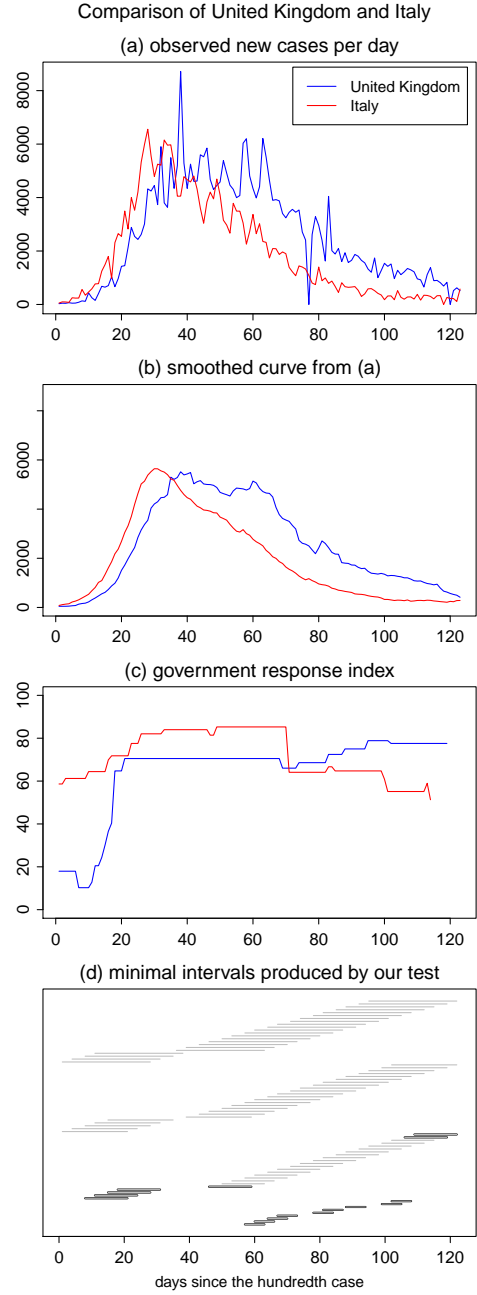


Figure 8: Test results for the comparison of the UK and Italy.

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 $\mathcal{F}_{\text{reject}}^{\min}(i, j)$ in black.

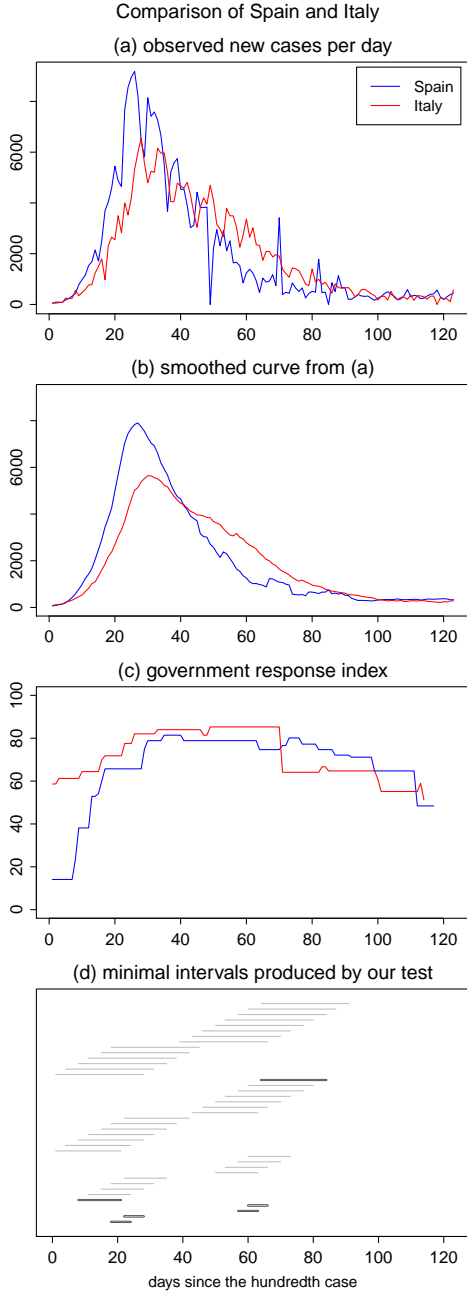


Figure 9: Test results for the comparison of Spain and Italy.

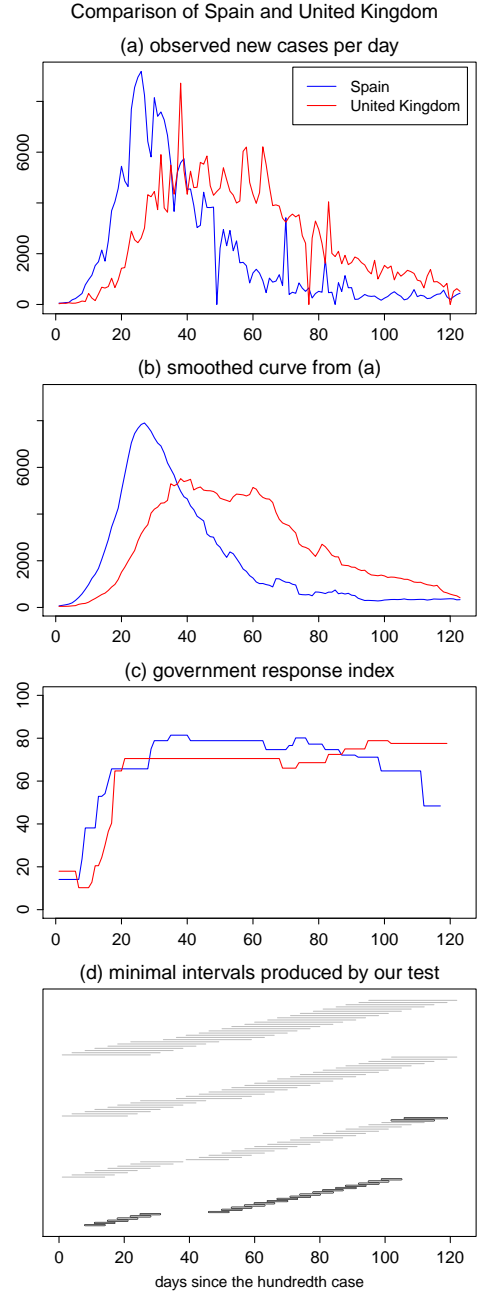


Figure 10: Test results for the comparison of Spain and the UK.

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 $\mathcal{F}_{\text{reject}}^{\min}(i, j)$ in black.

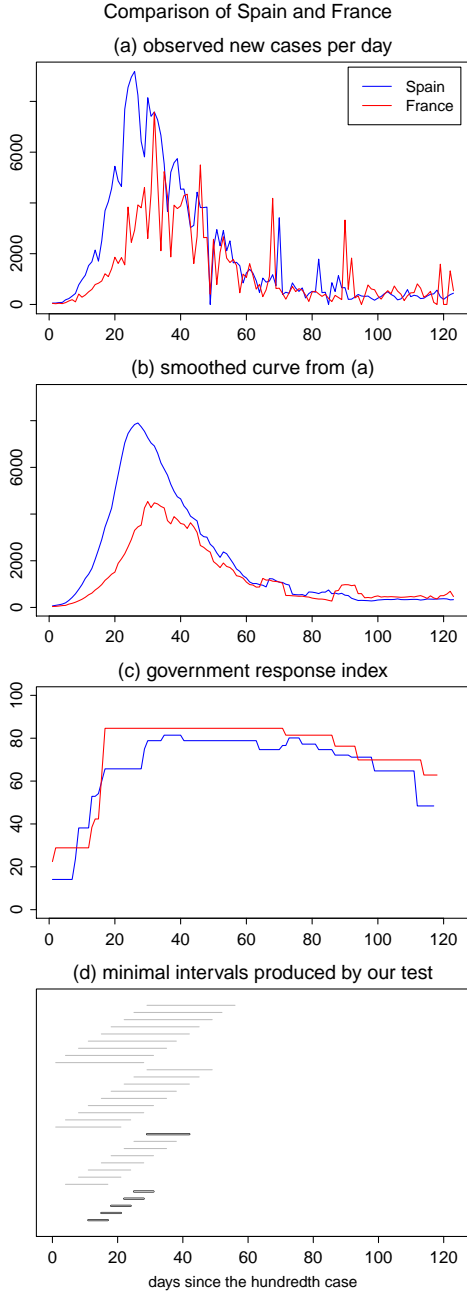


Figure 11: Test results for the comparison of Spain and France.

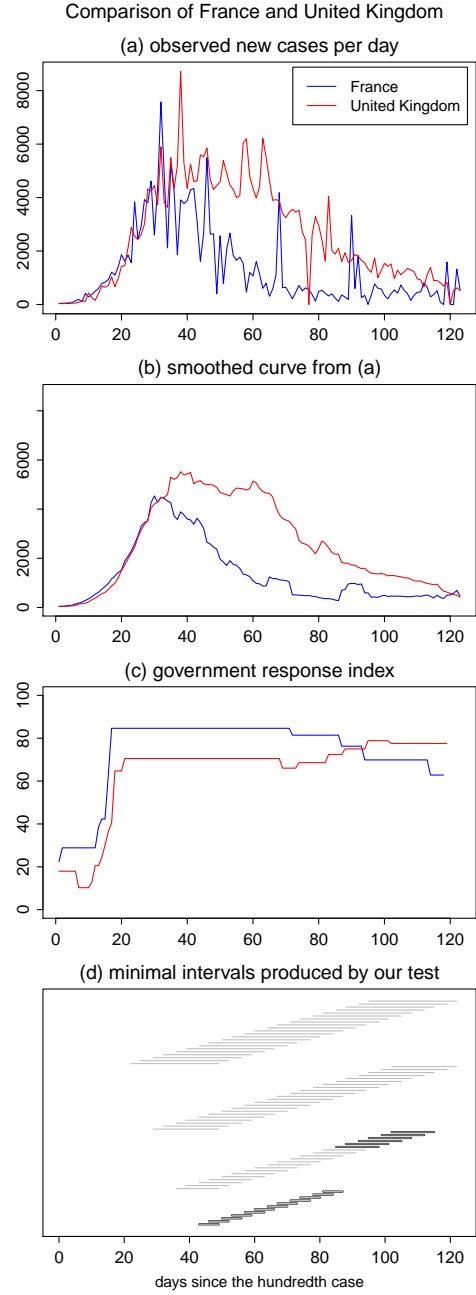


Figure 12: Test results for the comparison of France and the UK.

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 $\mathcal{F}_{\text{reject}}^{\min}(i, j)$ in black.