Modelling Epidemic Trends

Statistical problem

Suppose we observe a sample of data $\mathcal{N}_i = \{N_{it} : 1 \leq 1 \leq T\}$ for n different countries i, where N_{it} is the number of new infections on day t in country i.[1] We assume that $N_{it} \sim \mathcal{P}_{\lambda_i(t/T)}$, that is, N_{it} is Poisson distributed with (time-varying) intensity parameter $\lambda_i(t/T)$. Since $\lambda_i(t/T) = \mathbb{E}[N_{it}] = \mathrm{Var}(N_{it})$, we can model the observations N_{it} by the nonparametric regression equation

$$N_{it} = \lambda_i\Big(\frac{t}{T}\Big) + u_{it}, \tag{1}$$

where $u_{it} = N_{it} - \mathbb{E}[N_{it}]$ with $\mathbb{E}[u_{it}] = 0$ and $\mathrm{Var}(u_{it}) = \lambda_i(t/T)$. For simplicity, we suppose that the noise terms u_{it} in model (1) are independent both across countries i and time t. In the current Covid-19 crisis, independence across countries i seems justified as borders between countries are effectively closed. Independence across time t is more debatable, but may be justified as follows: [Discuss relationship between nonstationarities and autocorrelations. Refer to literature on simple volatility models.]

An important question in the current Covid-19 crisis is whether the intensity function λ_i has the same shape across countries i. If the intensity function of country i differs from that of country j, the dynamics of the epidemic are different in the two countries, that is, the virus spreads differently. Differences in the way the virus spreads are caused, among other things, by different policies and measures that countries take against the virus. Hence, a better understanding of how the intensity functions differ across countries may help to learn which measures against the virus are effective and which are not.

In this paper, we construct a statistical procedure which allow to identify differences between the intensity functions λ_i. More specifically, let $\{\mathcal{I}_k : k = 1, \ldots, K\}$ be a family of (rescaled) time intervals and let $H_0^{(ijk)}$ be the hypothesis that the intensity functions λ_i and λ_j differ 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 and for all intervals $\mathcal{I}_k$ under consideration. For a given significance level $\alpha \in (0, 1)$, the method produces a vector r with the entries r_{ijk}, where $r_{ijk} = 1$ if the test rejects the hypothesis $H_0^{(ijk)}$ and $r_{ijk} = 0$ otherwise. We derive theory

[1] We can also work with other kinds of count data, e.g., with the accumulated number of infections and the number of deaths (per day or accumulated).

which shows that the method allows to make uniform confidence statements of the following form: With statistical confidence at least $1-\alpha$, there is a difference between the two intensity functions λ_i and λ_j on the interval $\mathcal{I}_{ijk}$ for all (i,j,k) for which our test rejects, that is, for which $r_{ijk}=1$. Hence, our method allows to make uniform confidence statements (i) about which intensity functions differ from each other and (ii) about where, that is, in which time intervals they differ.

Model

Assuming that $N_{it} \sim \mathcal{P}_{\lambda_i(t/T)}$ leads to a nonparametric regression model of the form (1), which can be rewritten as

$$N_{it} = \lambda_i\Big(\frac{t}{T}\Big) + \sqrt{\lambda_i\Big(\frac{t}{T}\Big)}\eta_{it}, \tag{2}$$

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 that implied by the Poisson distribution. To deal with this issue, so-called quasi-Poisson models are frequently used. In our context, a quasi-Poisson model of N_{it} has the form

$$N_{it} = \lambda_i\Big(\frac{t}{T}\Big) + \sigma\sqrt{\lambda_i\Big(\frac{t}{T}\Big)}\eta_{it}, \tag{3}$$

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 work with this model, where the noise residuals η_{it} are assumed to have zero mean and unit variance but we do not impose any further distributional assumptions on them.

Test procedure

Construction of the test statistic:

(1) Let $\{\mathcal{I}_k : 1 \le k \le K\}$ be a family of subintervals of $[0,1]$. We first define a statistic to test the hypothesis $H_0^{(ijk)}$. To do so, we introduce the expression

$$\hat{s}_{ijk,T} = \sum_{t=1}^{T} w_k\Big(\frac{t}{T}\Big)(N_{it} - N_{jt}),$$

where $w_k(t/T)$ is a (rectangular) kernel weight defined by

$$w_k\Big(\frac{t}{T}\Big) = \frac{\mathbf{1}(\frac{t}{T} \in \mathcal{I}_k)}{\{\sum_{s=1}^{T} \mathbf{1}(\frac{s}{T} \in \mathcal{I}_k)\}^{1/2}}.$$

It holds that

$$\nu^2_{ijk,T} := \mathrm{Var}(\hat{s}_{ijk,T}) = \sigma^2 \sum_{t=1}^{T} w_k^2\Big(\frac{t}{T}\Big)\Big\{\lambda_i\Big(\frac{t}{T}\Big) + \lambda_j\Big(\frac{t}{T}\Big)\Big\}$$
$$\approx \sigma^2 h_k^{-1} \int_{u\in\mathcal{I}_k} \{\lambda_i(u) + \lambda_j(u)\} du,$$

where h_k is the length of the interval $\mathcal{I}_k$. 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 use the expression $\hat{\psi}_{ijk,T} := \hat{s}_{ijk,T}/\hat{\nu}_{ijk,T}$ as a test statistic, where

$$\hat{\nu}^2_{ijk,T} = \hat{\sigma}^2 \sum_{t=1}^{T} w_k^2\Big(\frac{t}{T}\Big)\{N_{it} + N_{jt}\}$$

and $\hat{\sigma}^2$ is an estimator of σ^2 introduced below. The overall test statistic thus has the form

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

(2) An estimator $\hat{\sigma}^2$ can be constructed as follows:

(a) Let $\hat{\lambda}_{i,h}$ be a Nadaraya-Watson estimator of the regression function λ_i, compute the residuals

$$\hat{r}_{it} = \frac{N_{it} - \hat{\lambda}_{i,h}(\frac{t}{T})}{\sqrt{\hat{\lambda}_{i,h}(\frac{t}{T})}},$$

and define $\hat{\sigma}_i^2 = T^{-1}\sum_{t=1}^{T} \hat{r}_{it}^2$. Finally, set $\hat{\sigma}^2 = n^{-1}\sum_{i=1}^{n} \hat{\sigma}_i^2$.

(b) Another (bandwidth-free) estimator is given by $\hat{\sigma}^2 = n^{-1}\sum_{i=1}^{n} \hat{\sigma}_i^2$ with

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

which is motivated by the fact that under certain smoothness conditions on the functions λ_i,

$$N_{it} - N_{it-1} = \lambda_i\Big(\frac{t}{T}\Big) - \lambda_i\Big(\frac{t-1}{T}\Big) + \sigma\sqrt{\lambda_i\Big(\frac{t}{T}\Big)}(\eta_{it} - \eta_{it-1})$$
$$+ \sigma\Big(\sqrt{\lambda_i\Big(\frac{t}{T}\Big)} - \sqrt{\lambda_i\Big(\frac{t-1}{T}\Big)}\Big)\eta_{it-1}$$
$$\approx \sigma\sqrt{\lambda_i\Big(\frac{t}{T}\Big)}(\eta_{it} - \eta_{it-1})$$

and thus

$$\frac{1}{T}\sum_{t=1}^{T}(N_{it}-N_{it-1})^2 \approx \sigma^2 \int_0^1 \lambda_i(u)du.$$

(3) We next combine the test statistics $\hat{\psi}_{ijk,T}$ for all intervals $\{\mathcal{I}_k : 1 \le k \le K\}$ and for all pairs of countries i and j to obtain the multiscale test statistic

$$\hat{\Psi}_{n,T} = \max_{1\le i<j\le n} \max_{1\le k\le K} \Big\{ \big|\hat{\psi}_{ijk,T}\big| - p(h_k) \Big\},$$

where as above h_k is the length of the interval $\mathcal{I}_k$ and $p(h) = \sqrt{2\log(1/h)}$.

Computation of the critical value:

(1) Under $H_0^{(ijk)}$ and appropriate regularity conditions, we obtain that

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

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

(2) We now define a Gaussian version of the statistic displayed in the last line above. In particular, we let

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

where Z_{it} are i.i.d. standard normal random variables for $1 \le t \le T$ and $1 \le i \le n$. With this, we define the statistic

$$\Phi_{n,T} = \max_{1\le i<j\le n} \max_{1\le k\le K} \Big\{ \big|\phi_{ijk,T}\big| - p(h_k) \Big\}$$

and denote its $(1-\alpha)$-quantile by $q_{n,T}(\alpha)$.

Our simultaneous test of the hypotheses $H_0^{(ijk)}$ is now carried out as follows: For a given significance level $\alpha \in (0,1)$ and each (i,j,k), reject $H_0^{(ijk)}$ if $|\hat{\psi}_{ijk,T}| - p(h_k) > q_{n,T}(\alpha)$.