Clustering of the epidemic time trends: the case of COVID-19

We consider the following nonparametric regression equation:

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

where c_i is the country-specific scaling parameter that accounts for the size of the country or population density. We introduce this additional parameter in order to be able to compare countries that differ substantially in terms of the population, i.e. Luxembourg and Russia. In what follows, we present a method that allows researchers to test the hypothesis that the time trends of new COVID-19 cases in different countries are the same up to some scaling parameter and to cluster the countries based on the differences.

For the identification purposes, we need to assume that for each $i \in \mathcal{C}$ we have $\int_0^1 \lambda_i(u)du = 1$. Only then we are able to estimate the scaling parameter c_i. Thus, the testing procedure is as follows.

Step 1

First, we estimate the scaling parameter:

$$\begin{aligned}
\widehat{c_i} &= \frac{1}{T}\sum_{t=1}^{T} X_{it} \\
&= c_i \frac{1}{T}\sum_{t=1}^{T} \lambda_i\Big(\frac{t}{T}\Big) + \sigma\frac{1}{T}\sum_{t=1}^{T}\sqrt{\lambda_i\Big(\frac{t}{T}\Big)}\eta_{it} \\
&= c_i \frac{1}{T}\sum_{t=1}^{T} \lambda_i\Big(\frac{t}{T}\Big) + o_P(1) \\
&= c_i + o_P(1),
\end{aligned}$$

where in the last inequality we used the normalization $\int_0^1 \lambda_i(u)du = 1$. Hence, for any fixed $i \in \mathcal{C}$, $\widehat{c_i}$ is a consistent estimator of c_i.

Step 2

Instead of working with X_{it}, we consider the following variables:

$$\begin{aligned}
X_{it}^{*} &= \frac{X_{it}}{\frac{1}{T}\sum_{t=1}^{T} X_{it}} \\
&= \frac{c_i}{\widehat{c_i}}\lambda_i\Big(\frac{t}{T}\Big) + \frac{\sigma}{\widehat{c_i}}\sqrt{\lambda_i\Big(\frac{t}{T}\Big)}\eta_{it}.
\end{aligned}$$

A statistic to test the hypothesis $H_0^{(ijk)}$ for a given triple (i,j,k) is then constructed as follows. Instead of $\hat{s}_{ijk,T}$, we work with

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

Then

$$\begin{aligned}\frac{\hat{s}^*_{ijk,T}}{\sqrt{Th_k}} =& \frac{1}{Th_k} \sum_{t=1}^{T} \mathbf{1}\Big(\frac{t}{T} \in \mathcal{I}_k\Big)(X^*_{it} - X^*_{jt}) \\ =& \frac{1}{Th_k} \sum_{t=1}^{T} \mathbf{1}\Big(\frac{t}{T} \in \mathcal{I}_k\Big)\Big(\lambda_i\Big(\frac{t}{T}\Big) - \lambda_j\Big(\frac{t}{T}\Big)\Big) + R_1 + R_2,\end{aligned}$$

where

$$\begin{aligned}R_1 &= \frac{1}{Th_k} \sum_{t=1}^{T} \mathbf{1}\Big(\frac{t}{T} \in \mathcal{I}_k\Big)\Big(\Big(\frac{c_i}{\widehat{c}_i} - 1\Big)\lambda_i\Big(\frac{t}{T}\Big) - \Big(\frac{c_j}{\widehat{c}_j} - 1\Big)\lambda_j\Big(\frac{t}{T}\Big)\Big), \\ R_2 &= \frac{1}{Th_k} \sum_{t=1}^{T} \mathbf{1}\Big(\frac{t}{T} \in \mathcal{I}_k\Big)\Big(\frac{\sigma}{\widehat{c}_i}\sqrt{\lambda_i\Big(\frac{t}{T}\Big)}\eta_{it} - \frac{\sigma}{\widehat{c}_j}\sqrt{\lambda_j\Big(\frac{t}{T}\Big)}\eta_{jt}\Big).\end{aligned}$$

Since $\widehat{c}_i = c_i + o_P(1)$ and $0 \le \sum_{t=1}^{T} \mathbf{1}\big(\frac{t}{T} \in \mathcal{I}_k\big)\lambda_i\big(\frac{t}{T}\big) \le h_k \lambda_{max}$, we have

$$\begin{aligned}|R_1| &\le \Big|\frac{c_i}{\widehat{c}_i} - 1\Big| \frac{1}{Th_k} \sum_{t=1}^{T} \mathbf{1}\Big(\frac{t}{T} \in \mathcal{I}_k\Big)\lambda_i\Big(\frac{t}{T}\Big) + \Big|\frac{c_j}{\widehat{c}_j} - 1\Big| \frac{1}{Th_k} \sum_{t=1}^{T} \mathbf{1}\Big(\frac{t}{T} \in \mathcal{I}_k\Big)\lambda_j\Big(\frac{t}{T}\Big), \\ &\le o_P(1) \cdot \frac{\lambda_{max}}{T} + o_P(1) \cdot \frac{\lambda_{max}}{T} = o_P\Big(\frac{1}{T}\Big). \end{aligned} \tag{0.1}$$

Furthermore, applying the law of large numbers, we get:

$$\frac{1}{Th_k} \sum_{t=1}^{T} \mathbf{1}\Big(\frac{t}{T} \in \mathcal{I}_k\Big)\sqrt{\lambda_i\Big(\frac{t}{T}\Big)}\eta_{it} = o_P(1).$$

Hence, if we uniformly bound the scaling parameters away from 0, i.e. $\exists\, c_{min}$ such that for all $i \in \mathcal{C}$ we have $0 < c_{min} \le c_i$, we can use the fact that $\frac{\sigma}{\widehat{c}_i} = O_P(1)$ to get that

$$\begin{aligned}R_2 &= \frac{\sigma}{\widehat{c}_i} \frac{1}{Th_k} \sum_{t=1}^{T} \mathbf{1}\Big(\frac{t}{T} \in \mathcal{I}_k\Big)\sqrt{\lambda_i\Big(\frac{t}{T}\Big)}\eta_{it} - \frac{\sigma}{\widehat{c}_j} \frac{1}{Th_k} \sum_{t=1}^{T} \mathbf{1}\Big(\frac{t}{T} \in \mathcal{I}_k\Big)\sqrt{\lambda_j\Big(\frac{t}{T}\Big)}\eta_{jt} \\ &= o_P(1). \end{aligned} \tag{0.2}$$

Combining (0.1) and (0.2) together, we get $\hat{s}^*_{ijk,T}/\sqrt{Th_k} = (Th_k)^{-1} \sum_{t=1}^{T} \mathbf{1}(t/T \in \mathcal{I}_k)\{\lambda_i(t/T) -$

$\lambda_j(t/T)\} + o_p(1)$ for any fixed pair of countries (i,j). Hence, the statistic $\hat{s}^*_{ijk,T}/\sqrt{Th_k}$ estimates the average distance between the functions λ_i and λ_j on the interval $\mathcal{I}_k$. The variance of $\hat{s}^*_{ijk,T}$ can not be easily calculated:

$$\begin{aligned}
\text{Var}(\hat{s}^*_{ijk,T}) =& \frac{1}{Th_k}\text{Var}\Big(\sum_{t=1}^T \mathbf{1}\Big(\frac{t}{T}\in\mathcal{I}_k\Big)(X^*_{it}-X^*_{jt})\Big) \\
=& \frac{1}{Th_k}\text{Var}\Big(\sum_{t=1}^T \mathbf{1}\Big(\frac{t}{T}\in\mathcal{I}_k\Big)X^*_{it}\Big) + \frac{1}{Th_k}\text{Var}\Big(\sum_{t=1}^T \mathbf{1}\Big(\frac{t}{T}\in\mathcal{I}_k\Big)X^*_{jt}\Big) \\
=& \frac{1}{Th_k}\text{Var}\Bigg(\frac{\sum_{t=1}^T \mathbf{1}\big(\frac{t}{T}\in\mathcal{I}_k\big)X_{it}}{\frac{1}{T}\sum_{t=1}^T X_{it}}\Bigg) + \frac{1}{Th_k}\text{Var}\Bigg(\frac{\sum_{t=1}^T \mathbf{1}\big(\frac{t}{T}\in\mathcal{I}_k\big)X_{jt}}{\frac{1}{T}\sum_{t=1}^T X_{jt}}\Bigg),
\end{aligned}$$

hence, we 'normalize' $\hat{s}^*_{ijk,T}$ intuitively by dividing it by the following value:

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

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

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

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

$$\hat{\psi}^{*,0}_{ijk,T} = \frac{\sum_{t=1}^T \mathbf{1}(\frac{t}{T}\in\mathcal{I}_k)\Big(\big(\frac{c_i}{\hat{c}_i}-\frac{c_j}{\hat{c}_j}\big)\overline{\lambda}_{ij} + \big(\frac{\sigma}{\hat{c}_i}-\frac{\sigma}{\hat{c}_j}\big)\overline{\lambda}_{ij}^{1/2}(\frac{t}{T})(\eta_{it}-\eta_{jt})\Big)}{\hat{\sigma}\{\sum_{t=1}^T \mathbf{1}(\frac{t}{T}\in\mathcal{I}_k)(X^*_{it}+X^*_{jt})\}^{1/2}}$$

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