Documentation `KIT-simple_nowcast`

Johanne Bracher, johannes.bracher@kit.edu

August 23, 2022

Abstract This note documents the reference nowcast procedure `KIT-simple_nowcast` implemented for the German COVID-19 Hospitalization Nowcast Hub.

Notation Denote by $X_{t,d}, d = 0, \dots, D$ the number of hospitalizations for *Meldedatum* t which appear in the data set at day $t + d$ and by

$$X_{t,\leq d} = \sum_{i=0}^{d} X_{t,i}$$

the number of hospitalizations reported for *Meldedatum* t up to day $t + d$. Moreover denote by

$$X_t = X_{t,\leq D} = \sum_{i=0}^{D} X_{t,i}$$

the total number of reported hospitalizations for t, where D denotes an assumed maximum possible delay. In the following we denote by X_t etc. a random variable and by x_t the corresponding observation.

The observed $x_{t,d}$ as available at a given time point t^* can be arranged into the so-called *reporting triangle*, see Table **??**.

Table 1: Illustration of the reporting triangle for time t^* and $D = 5$. Quantities known at time t are shown in black, yet unknown quantities are shown in grey.

day	$d=0$	$d=1$	$d=2$	$d=3$	$d=4$	$d=5$	total
1	$x_{1,0}$	$x_{1,1}$	$x_{1,2}$	$x_{1,3}$	$x_{1,4}$	$x_{1,5}$	x_1
2	$x_{2,0}$	$x_{2,1}$	$x_{2,2}$	$x_{2,3}$	$x_{2,4}$	$x_{2,5}$	x_2
$\vdots$							
t^*-5	$x_{t^*-5,0}$	$x_{t^*-5,1}$	$x_{t^*-5,2}$	$x_{t^*-5,3}$	$x_{t^*-5,4}$	$x_{t^*-5,5}$	x_{t^*-5}
t^*-4	$x_{t^*-4,0}$	$x_{t^*-4,1}$	$x_{t^*-4,2}$	$x_{t^*-4,3}$	$x_{t^*-4,4}$	$x_{t^*-4,4}$	x_{t^*-4}
t^*-3	$x_{t^*-3,0}$	$x_{t^*-3,1}$	$x_{t^*-3,2}$	$x_{t^*-3,3}$	$x_{t^*-3,4}$	$x_{t^*-3,5}$	x_{t^*-3}
t^*-2	$x_{t^*-2,0}$	$x_{t^*-2,1}$	$x_{t^*-2,2}$	$x_{t^*-2,3}$	$x_{t^*-2,4}$	$x_{t^*-2,5}$	x_{t^*-2}
t^*-1	$x_{t^*-1,0}$	$x_{t^*-1,1}$	$x_{t^*-1,2}$	$x_{t^*-1,3}$	$x_{t^*-1,4}$	$x_{t^*-1,5}$	x_{t^*-1}
t^*	$x_{t^*,0}$	$x_{t^*,1}$	$x_{t^*,2}$	$x_{t^*,3}$	$x_{t^*,4}$	$x_{t^*,5}$	x_{t^*}

As we will focus on seven-day hospitalization incidences we moreover need to consider rolling sums over windows of length W (usually $W = 7$)

$$Y_t = \sum_{w=0}^{W-1} X_{t-w}.$$

Goal Our aim is to estimate or *nowcast* Y_t based on the information available at time $t^* \geq t$. We do not take into account any information other than data on hospitalizations and their reporting delays, meaning that we model

$$Y_t \mid X_{s,d} : s + d \leq t^*, d \geq 0.$$

Point nowcast The following describes a simple heuristic to obtain a point prediction of Y_t based on information available at time t^*.

We start by imputing

$$x_{t^*,1} = x_{t^*,0} \times \frac{\sum_{i=1}^{t^*-1} x_{t^*-i,1}}{\sum_{i=1}^{t^*-1} x_{t^*-i,0}},$$

i.e. use a simple multiplication factor computed from the complete rows of our data set. Next we compute

$$x_{t^*,2} = x_{t^*,\leq 1} \times \frac{\sum_{i=1}^{t^*-1} x_{t^*-i,2}}{\sum_{i=1}^{t^*-1} x_{t^*-i,\leq 1}},$$

where in the computation of

$$x_{t^*-i,\leq 1} = x_{t^*-i,\leq 0} + x_{t^*-i,1}$$

we just treat the $x_{t^*-i,1}$ imputed in the first step as if it was a known value. The same can be done for

$$x_{t^*-1,2} = x_{t^*-1,\leq 1} \times \frac{\sum_{i=1}^{t^*-1} x_{t^*-i,2}}{\sum_{i=1}^{t^*-1} x_{t^*-i,\leq 1}}.$$

We repeat this same procedure to fill in the missing values of the reporting triangle step by step, moving from the left to the right and the bottom to the top.

This is equivalent to the following slightly more formal formulation: We denote by π_d the probability that a hospitalization with *Meldedatum* t appears in the data on day $t + d$ and by

$$\pi_{\leq d} = \sum_{i=0}^{d} \pi_i$$

the probability that such a hospitalization appears in the data no later than $t+d$. We introduce

$$\theta_d = \frac{\pi_d}{\pi_{\leq d-1}},$$

which allows us to formulate the recursion

$$\pi_{\leq d} = (1 + \theta_d)\pi_{\leq d-1}.$$

To estimate the θ_d for $d = 1, \ldots, D < t$ based on quantities available at time t^* we use

$$\hat{\theta}_d(t^*) = \frac{\sum_{j=d}^{J} X_{t^*-j,d}}{\sum_{j=d}^{J} X_{t^*-j,\leq d-1}},$$

where J is the number of past observations to include into the estimation (in practice it is often helpful to use only a recent subset rather than the entire available history). Note that we treat this estimate as a function t^* as it may change over time. Estimates of the probabilities $\pi_{\leq d}$ can then be obtained as

$$\hat{\pi}_{\leq d}(t^*) = (1 + \hat{\theta}_d)\hat{\pi}_{\leq d-1}.$$

These can subsequently serve to estimate the total number X_t of hospitalizations with *Meldedatum* t based on the $X_{t,\leq t^*-t}$ hospitalizations already reported by time t^*:

$$\hat{X}_t(t^*) = \frac{X_{t,\leq t^*-t}}{\hat{\pi}_{\leq t^*-t}(t^*)}.$$

We can also compute the estimates for the respective number of hospitalizations reported with a given delay $d > t^* - t$, which is given by

$$\hat{X}_{t,d}(t^*) = \hat{\pi}_d(t^*)\hat{X}_t(t^*).$$

In a last step we move to the rolling sum Y_t, which we estimate as

$$\hat{Y}_t(t^*) = \sum_{w=0}^{W-1} \hat{X}_{t-w}(t^*).$$

Uncertainty quantification Our general idea to quantify the nowcast uncertainty for $\hat{Y}_t(t^*)$ is to generate point forecasts $\hat{Y}_{t-1}(t^*-1), \hat{Y}_{t-2}(t^*-2), \ldots, \hat{Y}_{t-K}(t^*-K)$ for $K > D$ past time points, each based on the information available at the respective time point. These could then be compared to the corresponding observations $Y_{t^*-1}, \ldots, Y_{t^*-K}$, and nowcast dispersion could be based on a simple parametric model. However, two aspects need to be taken into account:

- The information available at t^*, on which the nowcast $\hat{Y}_t(t^*)$ is based, already implies a lower bound for Y_t, namely the hospitalizations which have already been observed. Only the hospitalizations for *Meldedatum* t which will be reported after t^* need to be modelled probabilistically. We thus introduce the decomposition

  $$Y_t = Y_{t,\leq t^*-t} + Y_{t,>t^*-t}.$$

  Here,

  $$Y_{t,\leq t^*-t} = \sum_{w=0}^{W-1}\sum_{d=0}^{D} X_{t-w,d} \times \mathbb{I}(-w+d \leq t^*-t)$$

  are those already observed by t^* (i.e., the lower bound) and

  $$Y_{t,>t^*-t} = \sum_{w=0}^{W-1}\sum_{d=0}^{D} X_{t-w,d} \times \mathbb{I}(-w+d > t^*-t)$$

  are those yet to be observed. We only need to quantify the uncertainty about the latter.

- At time t^*, the realizations of $Y_{t,>t^*-t}$ are only available for $t \leq t^* - D$. If we only want to use complete observations we would need to discard a lot of recent information.

We therefore construct a set of observations $Z_{t-j}, j = 1, \ldots, K$ and corresponding point predictions $\hat{Z}_{t-j}, (t^*-j)$ as follows:

- For $j = D, \ldots, K$ we can simply set

  $$Z_{t-j,>t^*-t} = Y_{t-j,>t^*-t}$$

  and point predictions $\hat{Z}_{t-j,>t^*-t}(t^*-j) = \hat{Y}_{t-j,>t^*-t}(t^*-j)$ as all relevant information are already available at t^*.

- For $j = 1, \ldots, D-1$ we use partial observations

$$\begin{aligned} Z_{t-j,>t^*-t} &= \sum_{w=0}^{W-1} \sum_{d=0}^{D} X_{t-j-w,d} \times \mathbb{I}(\underbrace{t-j-w+d \leq t^*}_{\text{"already observed at } t^*\text{"}}), \\ &= Y_{t-j,>t^*-t} - Y_{t-j,>t^*-t+j} \end{aligned}$$

  which are restricted to hospitalizations already reported by time t^*, so that $Z_{t-j,>t^*-t}$ can be evaluated. The corresponding point forecasts are given by

$$\hat{Z}_{t-j,>t^*-t} = \sum_{w=0}^{W-1} \sum_{d=0}^{D} \hat{X}_{t-j-w,d} \times \mathbb{I}(\underbrace{t-j-w+d \leq t^*}_{\text{"already observed at } t^*\text{"}}).$$

We then pragmatically assume that

$$Z_{t-j} \mid \hat{Z}_{t-j}(t^*-j) \sim \text{NegBin}(\text{mean} = \hat{Z}_{t-j}(t^*-j), \text{disp} = \psi_{t^*-t}),$$

where we parameterize the negative binomial distribution via its mean and the dispersion (size) parameter ψ_{t^*-t}. Note that the dispersion parameter depends on how far back into the past we nowcast (i.e., how much information has already accummulated between $t-j$ and t^*-j). The parameters $\psi_0, \ldots, \psi_D$ are then estimated via maximum likelihood. To avoid issues with zero expectations we add 0.1 to the expected values when feeding them into the maximum likelihood procedure.

The predictive distributions for Y_t are then set to $\text{NegBin}(\text{mean} = \hat{Y}_{t,>t^*-t}(t^*), \text{size} = \psi_{t^*-t})$, shifted by $Y_{t,\leq t^*-t}$.

As a motivation for the use of partial observations in the estimation of the overdispersion parameters we note that if

$$A \sim \text{NegBin}(\text{mean} = \hat{A}, \text{disp} = \psi)$$

and

$$B \mid A \sim \text{Bin}(A, \pi)$$

one gets

$$B \sim \text{NegBin}(\text{mean} = \pi\hat{A}, \text{disp} = \psi).$$

The negative binomial distribtion with a given dispersion parameter is thus closed to binomial subsampling, with only the expectation, but not the size parameter changing. It is thus defendable to assume the same size parameter for the constructed partial observations $Z_{t-j,>t^*-t}$ and the actual $Y_{t-j,>t^*-t}$ which we would use if they were already available.

Parameter choices To apply the suggested method, the numbers J and K of past observations used to estimate the nowcast mean and dispersion parameters. Here one needs to strike a balance between a sufficient amount and recency of training data. We set both J and K to 60 days without further assessing the impact on nowcast quality. The maximum delay D was set to 40 days.

Code The implementation of the described procedure can be found at https://github.com/KITmetricslab/hospitalization-nowcast-hub/tree/main/code/baseline.