Pois DGLM on Country- and County-Level Covid Data

Meini Tang

2022-07-28

1 Models

The instantaneous occurrence rate of the continuous-time Hawkes process is

$$\lambda(t) = \mu + R \sum_{t_k < t} \phi(t - t_l),$$

where ϕ is the kernel representing the distribution of the transmission delays and R is the reproduction number representing the average number of events induced by a single event. Koyama, Horie, and Shinomoto (2021) convert it to a discrete-time Hawkes process,

$$\lambda_t = \mu' + \sum_{k=1}^{t-1} \underset{\sim}{\phi_k} R_{t-k} y_{t-k},$$

where μ' is the deterministic immigration intensity (which is assumed to be zero for now), ϕ_k measures the transmission delays at lag k, R_t is the reproduction number.

We are comparing four models here. The first one is from Koyama, Horie, and Shinomoto (2021):

$$\mathcal{M}_1 : \begin{cases} y_t | \lambda_t \sim \text{Pois}(\lambda_t) \\ \lambda_t := F_t(\theta_t) = \phi_1 y_{t-1} \max(\theta_{t-1}, 0) + \cdots + \phi_L y_{t-L} \max(\theta_{t-L}, 0) \\ \theta_t = \theta_{t-1} + \omega_t, \end{cases}$$

where ω_t follows either a Cauchy distribution or normal distribution. Koyama et al. propose to use $\max(\theta_{t-1}, 0)$ as an estimator of the reproduction number R_t. The transmission delays $\phi_1, \ldots, \phi_L$ are modeled by the PMF of a lognormal distribution with mean equal to 4.7 days and standard deviation equal to 2.9 days, which is the estimated distribution of the serial intervals (Nishiura, Linton, and Akhmetzhanov 2020).

The second one is a modified version of $\mathcal{M}_2$, where $\max(\theta_{t-k}, 0)$ is substituted with $\exp(\theta_{t-k})$.

$$\mathcal{M}_2 : \begin{cases} y_t | \lambda_t \sim \text{Pois}(\lambda_t) \\ \lambda_t := F_t(\theta_t) = \phi_1 y_{t-1} \exp(\theta_{t-1}) + \cdots + \phi_L y_{t-L} \exp(\theta_{t-L}) \\ \theta_t = \theta_{t-1} + \omega_t, \end{cases}$$

where ω_t follows a normal distribution at this moment. It can also be extended to other error distributions. $\exp(\theta_{t-1})$ might be related to the reproduction number R_t.

For $\mathcal{M}_1$ and $\mathcal{M}_2$, the state space form is

$$\begin{pmatrix} \theta_t \\ \theta_{t-1} \\ \vdots \\ \theta_{t-L+1} \end{pmatrix} = \begin{pmatrix} 1 & \ldots & 0 & 0 \\ 1 & \ldots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \ldots & 1 & 0 \end{pmatrix} \begin{pmatrix} \theta_{t-1} \\ \theta_{t-2} \\ \vdots \\ \theta_{t-L} \end{pmatrix} + \begin{pmatrix} \omega_t \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

The third model uses the Pascal distributed lags proposed by Solow (1960), with more discussion by Ravines, Schmidt, and Migon (2006).

$$\mathcal{M}_3 : \begin{cases} y_t|\lambda_t \sim \text{Pois}(\lambda_t) \\ \lambda_t = \exp(\theta_t) \\ \theta_t = 2\rho\theta_{t-1} - \rho^2\theta_{t-2} + (1-\rho)^2\beta_t \log(y_{t-1}) \\ \beta_t = \beta_{t-1} + \omega_t, \end{cases}$$

$\mathcal{M}_3$ doesn't really have a discretized Hawkes process form but the β_t might be interesting.

The county-level data has many zero observations, which cause a trouble for $\mathcal{M}_3$. In this cases, β_t will go to negative values, especially when the preceding number of daily new cases is non-zero.

Model 4 is a variant of $\mathcal{M}_3$ using the identity link instead of the logarithm link.

$$\mathcal{M}_4 : \begin{cases} y_t|\lambda_t \sim \text{Pois}(\lambda_t) \\ \lambda_t = \theta_t \\ \theta_t = 2\rho\theta_{t-1} - \rho^2\theta_{t-2} + (1-\rho)^2\beta_t y_{t-1} \\ \beta_t = \beta_{t-1} + \omega_t, \end{cases}$$

exp(beta[t-1])

It is also related to the discretized Hawkes process if we rewrite the evolution equations in the following way:

$$\begin{aligned} \theta_t &= \sum_{k=0}^{\infty} (k+1)(1-\rho)^2\rho^k \beta_t y_{t-k-1}, \\ \beta_t &= \beta_{t-1} + \omega_t. \end{aligned}$$

y[t-k]

(1 - rho B)^2 theta = (1-rho)^2 exp(beta[t-1]) y[t]

The β_t is the reproduction-ish number that could be interesting. There is no restriction on the parameter space of β_t for now but we might want it to be positive.

The transmission delay $\phi_k = (k+1)(1-\rho)^2\rho^k$ is the PMF of a negative binomial distribution with mean equal to $2\rho/(1-\rho)$ and variance equal to $2\rho/(1-\rho)^2$, where ρ is the success probability in each trial and the number of failures $r = 2$. When $r = 6$ and $\rho \approx 0.4393$, it has a mean equal to 4.7 and variance equal to 8.38 (standard deviation equal to 2.9), as suggested by Nishiura, Linton, and Akhmetzhanov (2020).

For now I just try $r = 2$ and ρ around 0.7, but the derivation should be similar for $r = 6$. The comparisions of different PMF is shown in figure 1.

The state space form is

$$\begin{pmatrix} \theta_t \\ \theta_{t-1} \\ \beta_t \end{pmatrix} = \begin{pmatrix} 2\rho & -\rho^2 & (\rho)^2 x_t \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \theta_{t-1} \\ \theta_{t-2} \\ \beta_{t-1} \end{pmatrix} + \begin{pmatrix} (1-\rho)^2 x_t \omega_t \\ 0 \\ \omega_t \end{pmatrix}$$

In the following examples, $\mathcal{M}_1$ is estimated by particle filtering assuming normal error terms, others are estimated by the linear Bayes estimator. β_t might be related to the reproduction number R_t. Also, though I use the data from March 2020 to March 2021 to fit the models, I zoom into the period from May 2020 to March 2021 for visualization because the first few starting points sometimes goes wild for filtering.

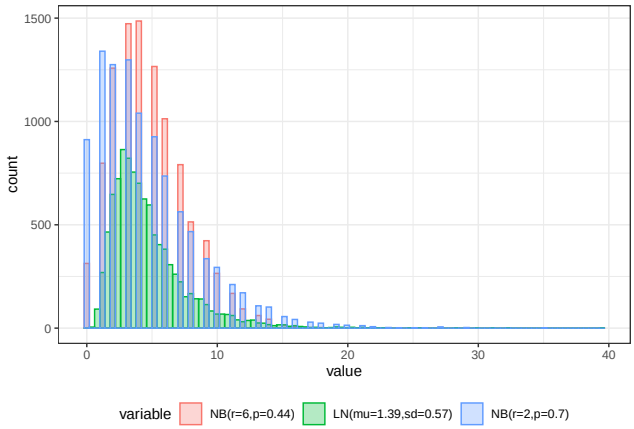
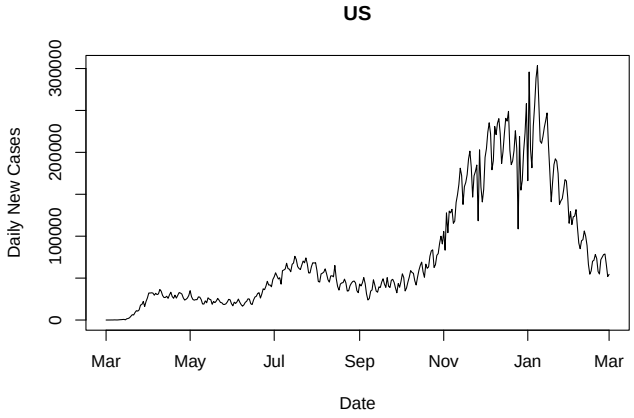


Figure 1: PMF of a LN(mu=1.39,sd=0.57) is the transmission delay used in Koyama et al. (2021), shown in green; NB(r=6,p=0.44) is the closest approximation of it, shown in red; NB(r=2,p=0.7) is the transmission delay used in the following examples, shown in blue.

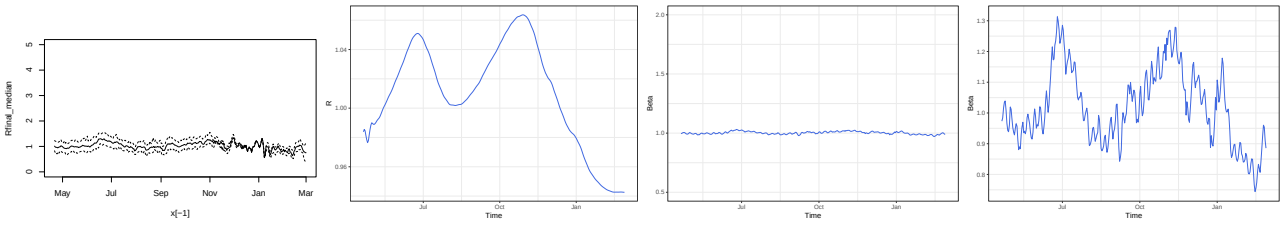
2 Country-level Covid Data

Country-level Covid data from March 1, 2020 to March 1, 2021.

2.1 US



2.1.1 Compare three different models



(a) M1 - KoyamaMax (b) M2 - KoyamaExp (c) M3 - PascalLog (d) M4 - PascalEye

Figure 2: US: M2 uses a discount factor equal of 0.99; M3 uses rho=0.7 and a discount factor of 0.7; M4 uses rho=0.7 and a discount factor of 0.8.

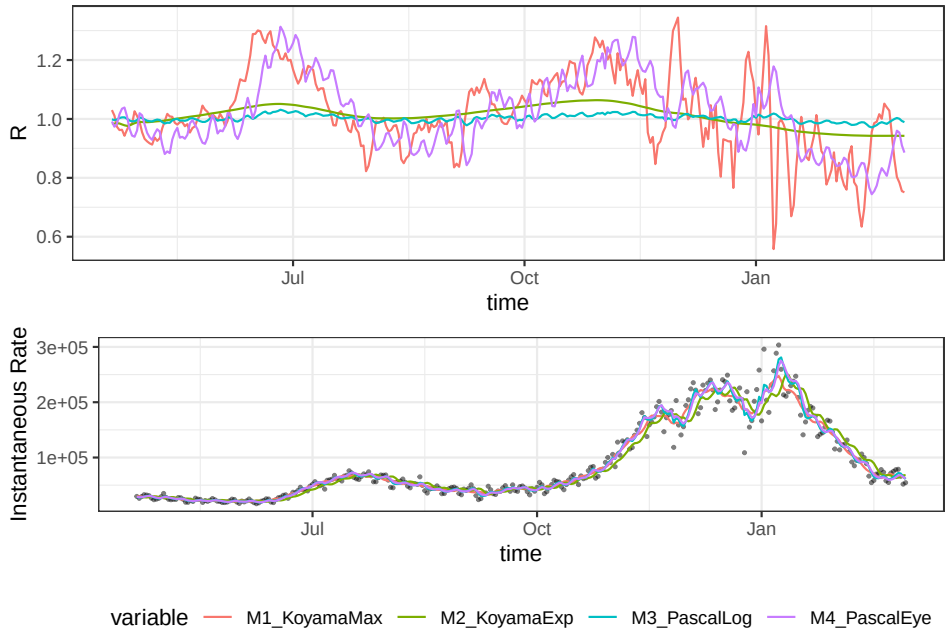
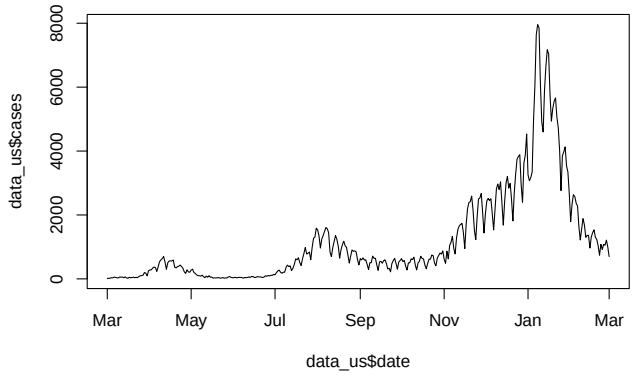
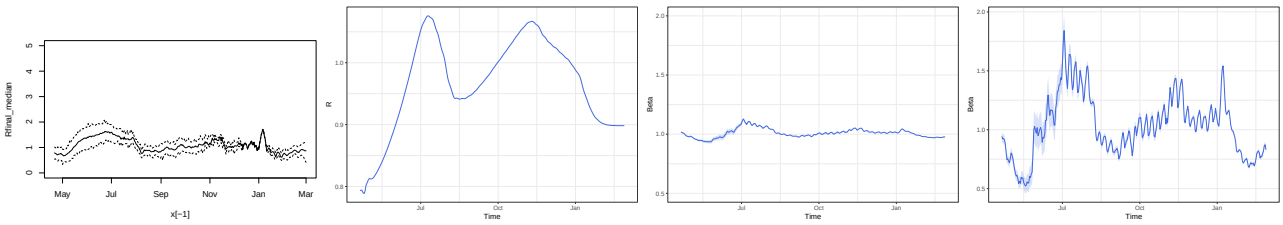


Figure 3: US: M2 uses a discount factor equal of 0.99; M3 uses rho=0.7 and a discount factor of 0.7; M4 uses rho=0.7 and a discount factor of 0.8.

2.2 Japan



2.2.1 Compare three different models



(a) M1-KoyamaMax (b) M2-KoyamaExp (c) M3-PascalLog (d) M4-PascalEye

Figure 4: Japan: M2 uses a discount factor equal of 0.99; M3 uses rho=0.7 and a discount factor of 0.8; M4 uses rho=0.7 and a discount factor of 0.9.

In our meeting on Tuesday, I actually set the ρ for $\mathcal{M}_4$ equal to 0.9, and now I change it to 0.7 so the mean of

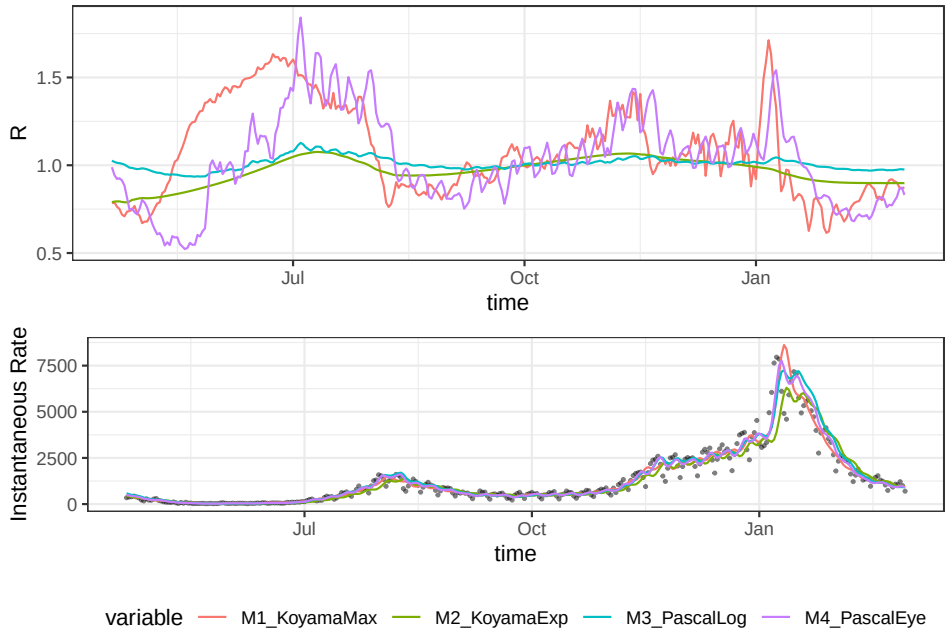


Figure 5: Japan: M2 uses a discount factor equal of 0.99; M3 uses rho=0.7 and a discount factor of 0.8; M4 uses rho=0.7 and a discount factor of 0.9.

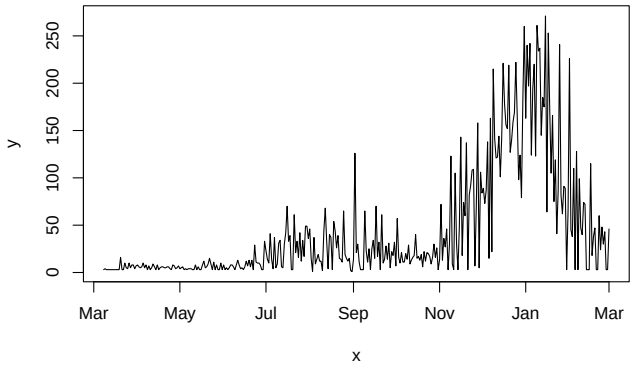
the negative binomial is the same as the lognormal for the transmission delays. It seems the reproduction-ish number is more similar to Koyama, Horie, and Shinomoto (2021) by doing so.

3 County-level Covid Data

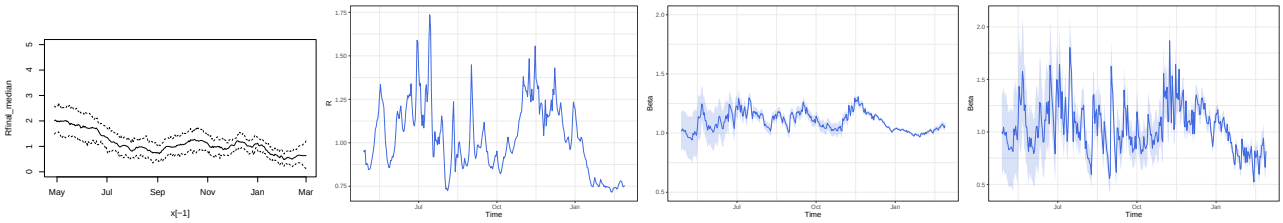
County-level Covid data from March 01, 2020 to March 01, 2021. The county-level data actually has a lot of zeros and even negative values. Therefor, I shifted the data as follows: $y \Leftarrow y - \min(y) + 1$, so that it has a minimum value of one. We need to consider preprocessing or different source of data if we want to use this data set.

In the following results, we can see some similarity between $\mathcal{M}_2$ and $\mathcal{M}_4$ for Santa Cruz and Monterey. Also, $\mathcal{M}_1$ has different behaviors compared to other models, probably some tuning is missing.

3.1 Santa Cruz



3.1.1 Compare three different models



(a) M1-KoyamaMax (b) M2-KoyamaExp (c) M3-PascalLog (d) M4-PascalEye

Figure 6: Santa Cruz: M2 uses a discount factor equal of 0.88; M3 uses rho=0.7 and a discount factor of 0.9; M4 uses rho=0.7 and a discount factor of 0.8.

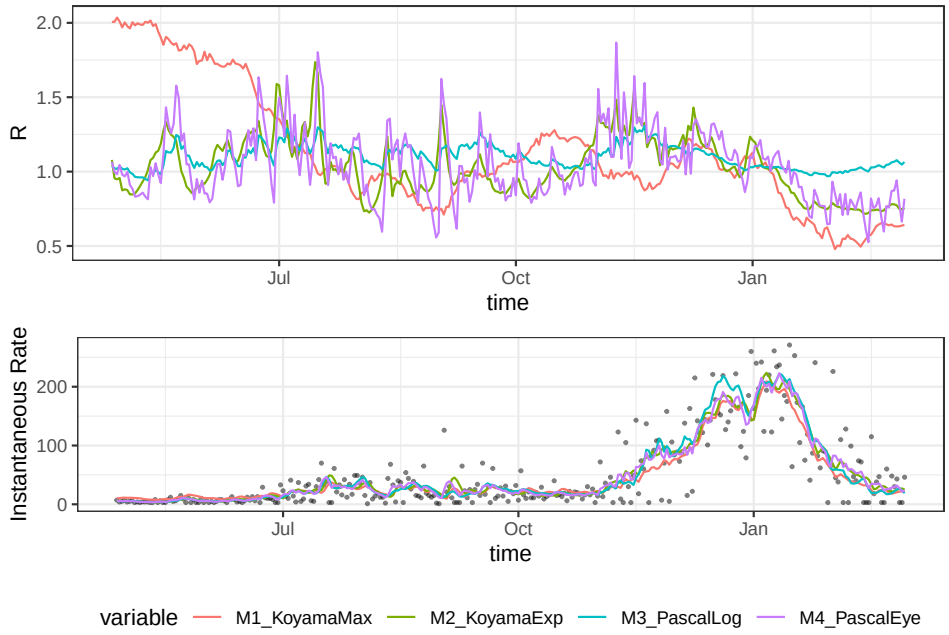
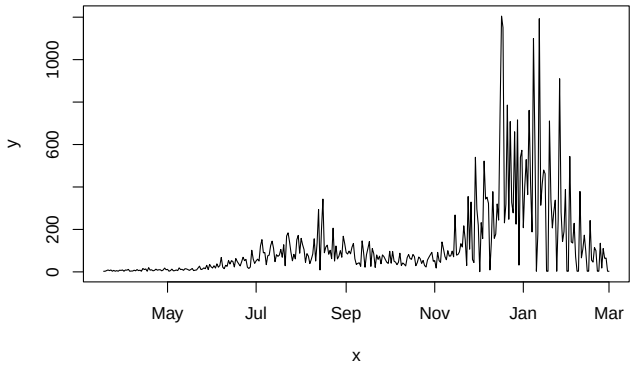
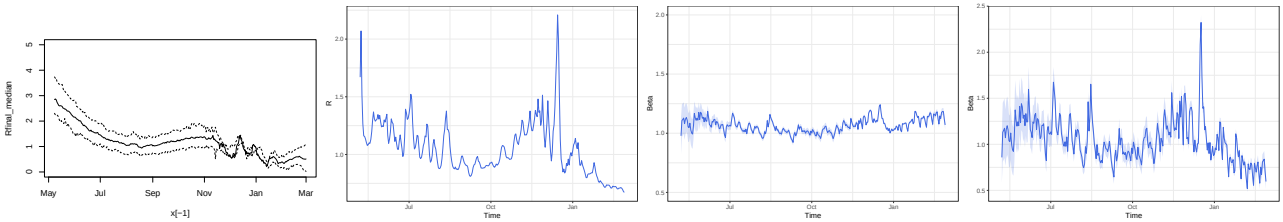


Figure 7: Santa Cruz: M2 uses a discount factor equal of 0.88; M3 uses rho=0.7 and a discount factor of 0.9; M4 uses rho=0.7 and a discount factor of 0.8.

3.2 Monterey



3.2.1 Compare three different models



(a) M1-KoyamaMax (b) M2-KoyamaExp (c) M3-PascalLog (d) M4-PascalEye

Figure 8: Monterey: M2 uses a discount factor equal of 0.88; M3 uses rho=0.7 and a discount factor of 0.8; M4 uses rho=0.7 and a discount factor of 0.8.

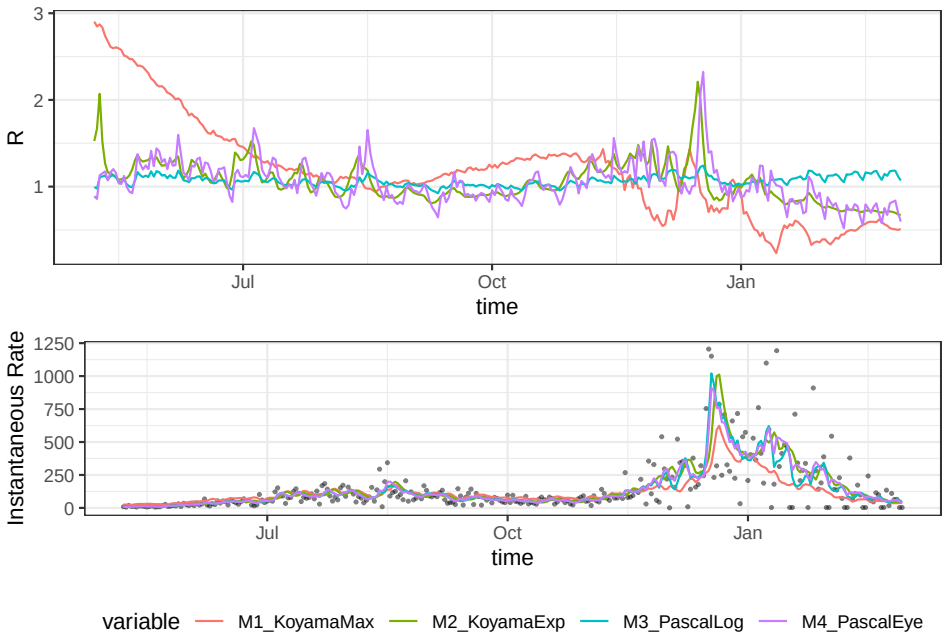
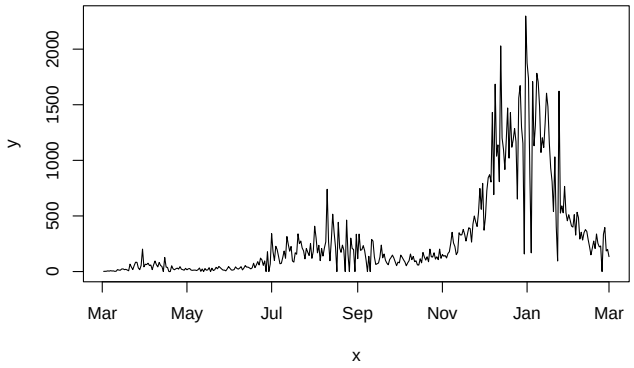
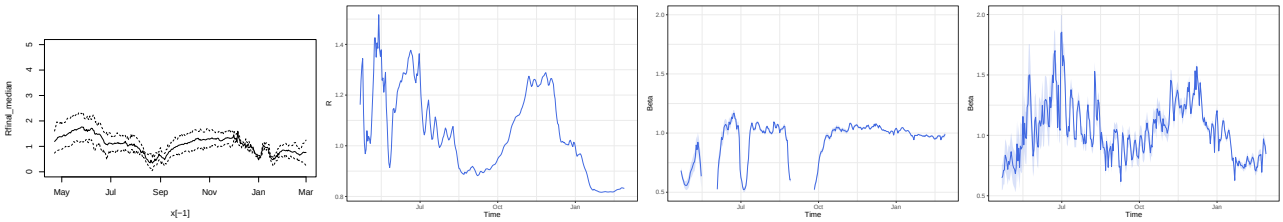


Figure 9: Monterey: M2 uses a discount factor equal of 0.88; M3 uses rho=0.7 and a discount factor of 0.8; M4 uses rho=0.7 and a discount factor of 0.8.

3.3 Santa Clara



3.3.1 Compare three different models



(a) M1-KoyamaMax (b) M2-KoyamaExp (c) M3-PascalLog (d) M4-PascalEye

Figure 10: Santa Clara: M2 uses a discount factor equal of 0.95; M3 uses rho=0.7 and a discount factor of 0.9; M4 uses rho=0.7 and a discount factor of 0.8.

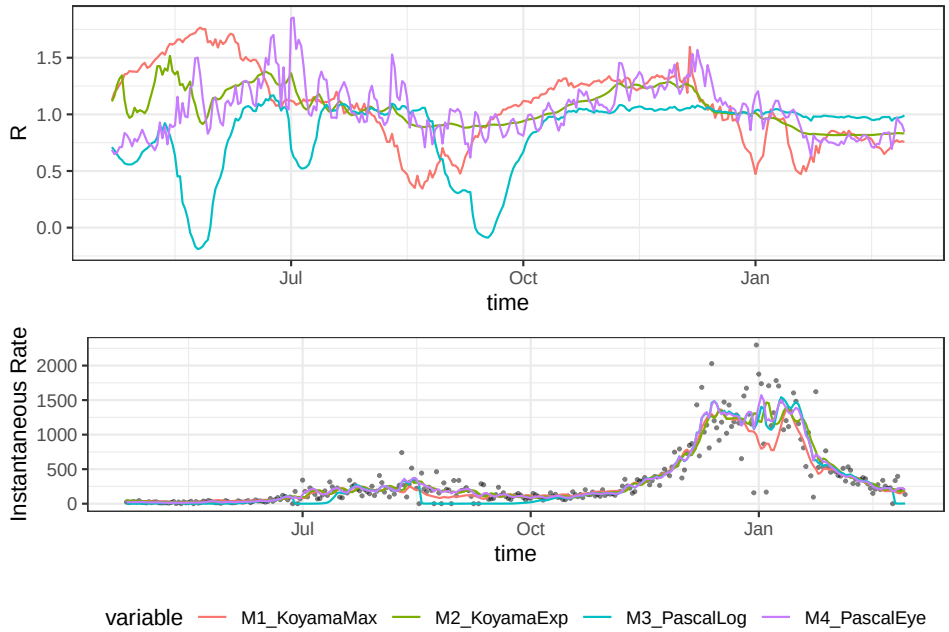


Figure 11: Santa Clara: M2 uses a discount factor equal of 0.95; M3 uses rho=0.7 and a discount factor of 0.9; M4 uses rho=0.7 and a discount factor of 0.8.

4 References

Koyama, Shinsuke, Taiki Horie, and Shigeru Shinomoto. 2021. "Estimating the Time-Varying Reproduction Number of COVID-19 with a State-Space Method." Edited by Roger Dimitri Kouyos. *PLOS Computational Biology* 17 (1): e1008679. https://doi.org/10.1371/journal.pcbi.1008679.

Nishiura, Hiroshi, Natalie M. Linton, and Andrei R. Akhmetzhanov. 2020. "Serial Interval of Novel Coronavirus (COVID-19) Infections." *International Journal of Infectious Diseases* 93 (April): 284–86. https://doi.org/10.1016/j.ijid.2020.02.060.

Ravines, Romy R., Alexandra M. Schmidt, and Helio S. Migon. 2006. "Revisiting Distributed Lag Models Through a Bayesian Perspective." *Applied Stochastic Models in Business and Industry* 22 (2): 193–210. https://doi.org/10.1002/asmb.628.

Solow, Robert M. 1960. "On a Family of Lag Distributions." *Econometrica* 28 (2): 393. https://doi.org/10.2307/1907729.