State estimation of Covid-19 disease in Denmark

Kasper Schou Telkamp (s170397)

2022-12-23

Contents

Introduction **2**

Materials and method **2**

Data 2

Model formulation 4

Laplace approximation 4

Importance sampling 5

Implementations 5

Results **6**

Parameters 6

Estimating the total number of new hospital admissions in Denmark 6

Estimating the number of new hospital admissions in Denmark grouped by region 7

Discussion **10**

References **11**

Introduction

Early detection of outbreaks with communicable diseases are of great importance in order to initiate timely interventions and help prevent disease spread. In this report non-normal mixed effects models will be evaluated on their ability to identify outbreaks of Covid-19 disease using data over new hospital admissions with Covid-19 in Denmark. Different implementations of generalized linear mixed models (GLMMs) in R packages will be compared. Namely, the `glmmTMB` and `KFAS` R package available at Comprehensive R Archive Network (CRAN).

Materials and method

Data

In this project, the daily record of new hospital admissions with Covid-19 in Denmark grouped by region of residence and totals are used. The head and tail of the processed data are listed in Table 1.

Table 1: Processed dataset containing the daily record of new hospital admissions with Covid-19 in Denmark grouped by region of residence and with totals.

Dato	Hovedstaden	Sjælland	Syddanmark	Midtjylland	Nordjylland	Ukendt.Region	Total
2020-03-01	1	0	0	0	0	0	1
2020-03-02	0	0	0	0	0	0	0
2020-03-03	1	0	0	0	0	0	1
2020-03-04	0	0	0	0	0	0	0
...	...	...	...	...	...	...	...
2022-11-03	20	12	9	10	2	1	54
2022-11-04	13	15	6	3	3	0	40
2022-11-05	7	6	6	3	3	0	25
2022-11-06	8	10	7	4	2	0	31

The data is publicly available and were obtained from Statens Serum Institut (SSI) website[1]. SSI collects the data from the National Patient Registry (NPR), which contains information about outpatient contacts from Danish public as well as private hospitals. The data from NPR has some delay. Therefore, the inventory is updated daily with real-time data from the regions. The regions provide snapshot-data twice to SSI daily at 7am and 3pm. A hospital admission related to Covid-19 is defined as an admission, where a patient is admitted within 14 days after a positive SARS-CoV-2 test. Patients that are tested positive for SARS-CoV-2 during an admission is also registered as a Covid-19 related admission. Furthermore, admissions with Covid-19 are only registered for patients that are present in at least one snapshot, or if the patient have been admitted for more than 12 hours according to NPR. The total number of new admissions to the hospital in Denmark are visualized in Figure 1.

[1] https://covid19.ssi.dk/overvagningsdata/download-fil-med-overvaagningdata

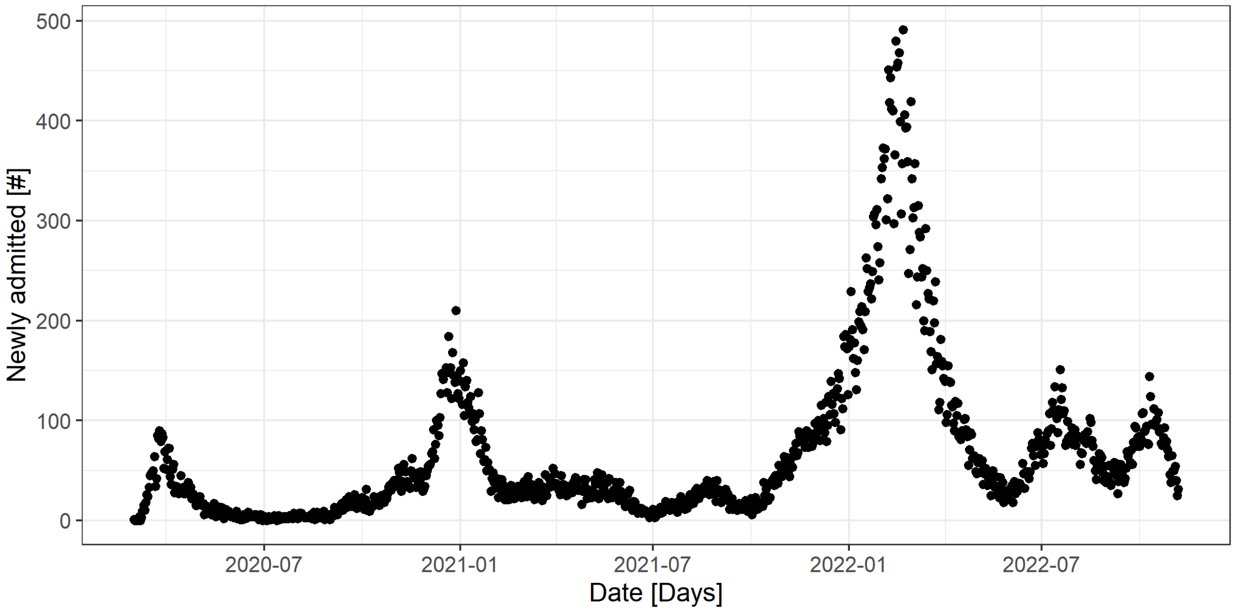

Figure 1: Daily number of new hospital admissions in Denmark.

Clearly, the observed number of new hospital admissions are correlated in time.

In Figure 2 the daily number of new hospital admissions grouped by region of residence is visualized.

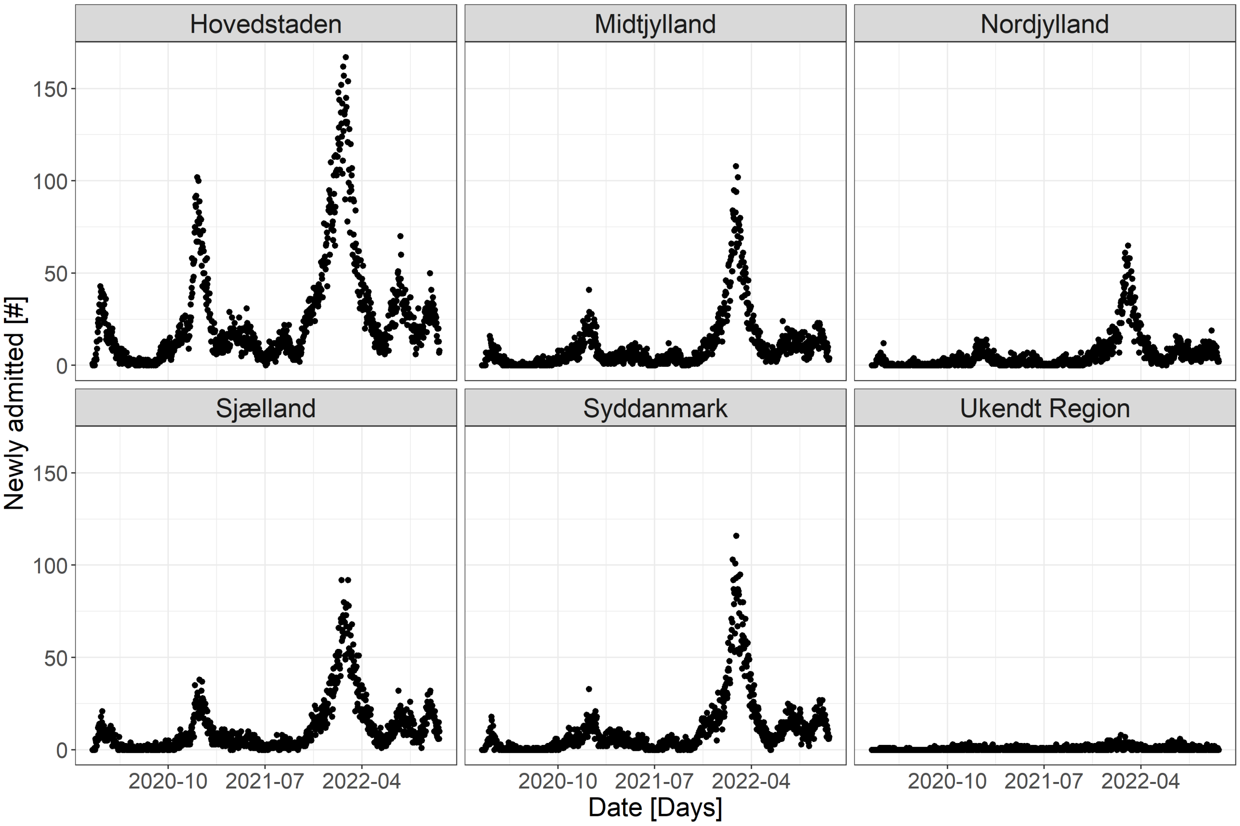

Figure 2: Total number of new hospital admissions in Denmark.

It can be seen that most of the new hospital admissions can be linked to *Region Hovedstaden*, while only a negligible amount are linked to the unknown region. This is largely due to the fact, that *Region Hovedstaden* considers more individuals.

Model formulation

In order to analyze the data a simple state space model is proposed. The count observations h_t , $t = 1, \dots, n$ in a period of $n = 981$ days starting on the 1st of March, 2020, of new hospital admissions is assumed to follow a Poisson distribution $h_t \sim \mathrm{P}(\lambda_t)$ with intensities given by

$$\log(\lambda_t) = \beta + u_t \tag{1}$$

Here β is a fixed parameter, that represents the average intensity and u_t is a random effect that is assumed to follow a first order autoregressive process

$$u_t = au_{t-i} + \epsilon_t \tag{2}$$

where $\epsilon \sim N(0, \sigma^2), t > 1$ is a white noise process, and a and σ are model parameters. Using the results from Madsen (2007), it is assumed that u_1 follow the stationary distribution of the first order autoregressive process $u_1 \sim N(0, \sigma^2/(1-a^2))$. Hence the joint likelihood becomes

$$L(a, \beta, \sigma; \mathbf{u}, \mathbf{y}) = \phi_{0,\frac{\sigma^2}{1-\sigma^2}}(u_1) \prod_{t=2}^{n} \left(\phi_{0,\sigma^2}(u_t - au_{t-1})\right) \prod_{t=1}^{n} \left(p_{\lambda_t}(h_t)\right) \tag{3}$$

where ϕ_{μ,σ^2} is the probability density function (pdf) of the normal distribution with mean μ and variance σ_2, and p_λ is the pdf of the Poisson distribution with mean λ.

In order to obtain the likelihood for the model parameters (a, β, σ) the observed random effects are integrated out. Hence, the marginal likelihood is obtained

$$L_M = (a, \beta, \sigma; \mathbf{y}) = \int_{\mathbb{R}^q} L(a, \beta, \sigma; \mathbf{u}, \mathbf{y}) d\mathbf{u} \tag{4}$$

where q is the number of random effects and a, β, and σ are the parameters to be estimated.

In order to make computation of the joint likelihood function in (4) feasible, the estimation is carried out using the multivariate Laplace approximation.

Laplace approximation

The marginal log-likelihood $l_M(a, \beta, \sigma; \mathbf{y}) = \log(L_M(a, \beta, \sigma; \mathbf{y}))$ is approximated by a second order Taylor approximation around the optimum $\tilde{\mathbf{u}} = \hat{\mathbf{u}}_\theta$ of the log-likelihood function w.r.t. the unobserved random variables $\mathbf{u}$, i.e.,

$$l(a, \beta, \sigma; \mathbf{u}, \mathbf{y}) \approx l(a, \beta, \sigma; \tilde{\mathbf{u}}, \mathbf{y}) - \frac{1}{2}(\mathbf{u} - \tilde{\mathbf{u}})^T H(\tilde{\mathbf{u}})(\mathbf{u} - \tilde{\mathbf{u}}) \tag{5}$$

where the first-order term of the Taylor expansion disappears since the expansion is done around the optimum $\tilde{\mathbf{u}}$ and $H(\tilde{\mathbf{u}}) = -l''_{uu}(a, \beta, \sigma, \mathbf{u}, \mathbf{y})|_{\mathbf{u}=\tilde{\mathbf{u}}}$ is the negative Hessian of the joint log-likelihood evaluated at $\tilde{\mathbf{u}}$.

Using the approximation in (5) on (4) the Laplace approximation of the marginal log-likelihood becomes (See Madsen & Thyregod (2011))

$$
\begin{aligned}
l_{M,LA}(a,\beta,\sigma;\mathbf{y}) &= \log \int_{\mathbb{R}^q} \exp\left(l(a,\beta,\sigma;\tilde{\mathbf{u}},\mathbf{y}) - \frac{1}{2}(\mathbf{u}-\tilde{\mathbf{u}})^T H(\tilde{\mathbf{u}})(\mathbf{u}-\tilde{\mathbf{u}})\right) d\mathbf{u} && (6)\\
&= l(a,\beta,\sigma;\tilde{\mathbf{u}},\mathbf{y}) - \frac{1}{2}\log\left|\frac{H(\tilde{\mathbf{u}})}{2\pi}\right| && (7)
\end{aligned}
$$

Importance sampling

Importance sampling is a re-weighting technique for approximating integrals w.r.t. a density f by simulation in cases where it is not feasible to simulate from the distribution with density f. Instead it uses samples from a different distribution with density g, where the support of g includes the support of f.

Implementations

In this section two different R packages, `glmmTMB` and `KFAS` available at CRAN are presented.

glmmTMB

This section describes the R package `glmmTMB` by Brooks et al. (2017) for linear and GLMMs using Template Model Builder (TMB). The models are estimated using maximum likelihood estimation via TMB. Random effects are assumed to be Gaussian on the scale of the linear predictor and are integrated out using Laplace approximation. Additionally, gradients are calculated using automatic differentiation.

KFAS

This section goes into detail with the R package `KFAS` by Helske (2017) for state space modelling with observations from the exponential family. The `KFAS` package can perform Kalman filtering and smoothing with exact diffuse initialization using an univariate approach.

In `KFAS` the Poisson distribution with intensity λ_t and exposure term e_t together with the log-link is supported. Thus we have $\mathrm{E}\left(h_t|\log(\lambda_t)\right) = \mathrm{Var}\left(h_t|\log(\lambda_t)\right) = e_t\lambda_t$. In this report the exposure term is assumed to be constant, i.e. $e_t = 1$. Hence, the differences are represented directly in the estimated parameters and latent state.

In order to make inferences of the Poisson model, `KFAS` finds a Gaussian model with the same conditional posterior mode as $\mathrm{P}(\lambda|\mathbf{h})$. This is done trough an iterative process with Laplace approximation of $\mathrm{P}(\lambda|\mathbf{h})$, where the updated estimates of $\log(\lambda_t)$ are computed via the Kalman filtering and smoothing from the approximating Gaussian model. The final estimates of $\log(\hat{\lambda}_t)$ correspond to the mode of $\mathrm{P}(\lambda|\mathbf{h})$. Generally, the difference between the mode and the mean is negligible. Nevertheless, our interest is focused on the intensity λ_t rather than the linear predictor $\log(\lambda_t)$.

Direct transformation from the linear predictor to the intensity introduces some bias. To solve this problem `KFAS` also contains methods based on importance sampling.

Filtering in `KFAS` is denoted as

$$
\begin{aligned}
u_{t+1} &= \mathrm{E}(u_{t+1}|h_t,\ldots,h_1)\\
P_{t+1} &= \mathrm{Var}(u_{t+1}|h_t,\ldots,h_1)
\end{aligned}
$$

Smoothing in `KFAS` is denoted by

$$
\begin{aligned}
\hat{u}_t &= \mathrm{E}(u_t | h_n, \ldots, h_1) \\
V_t &= \mathrm{Var}(u_t | h_n, \ldots, h_1)
\end{aligned}
$$

Results

Parameters

Estimating the total number of new hospital admissions in Denmark

In Figure 3 the total number of new hospitals admissions in Denmark is visualized together with the smoothed estimates from `glmmTMB` and `KFAS`.

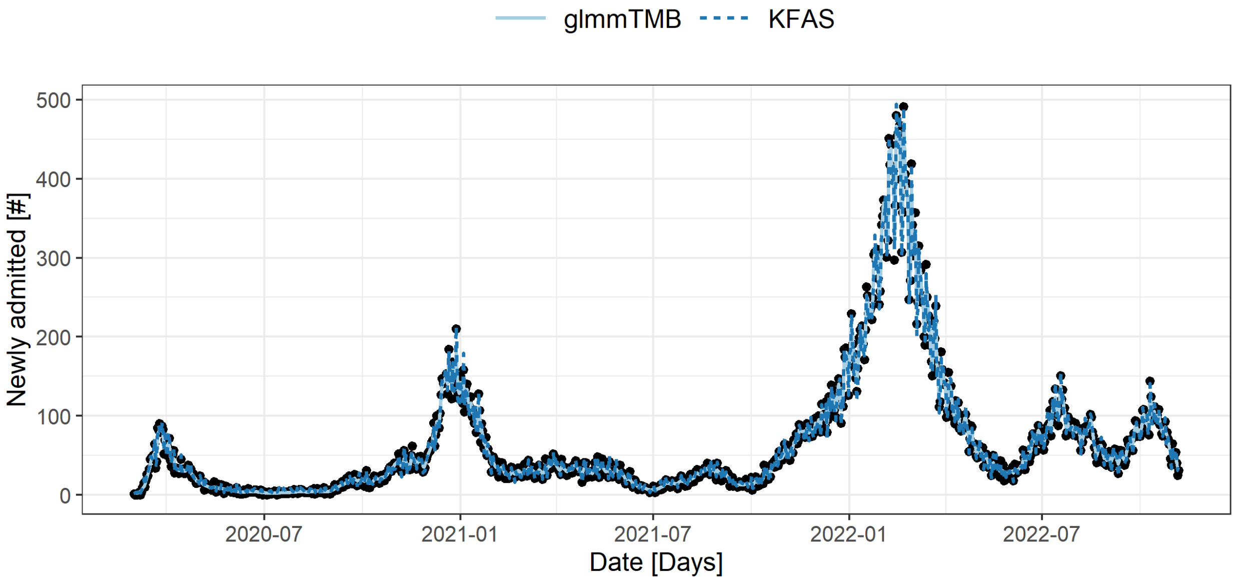

Figure 3: Total number of new hospital admissions in Denmark with smoothed estimates from glmmTMB and KFAS.

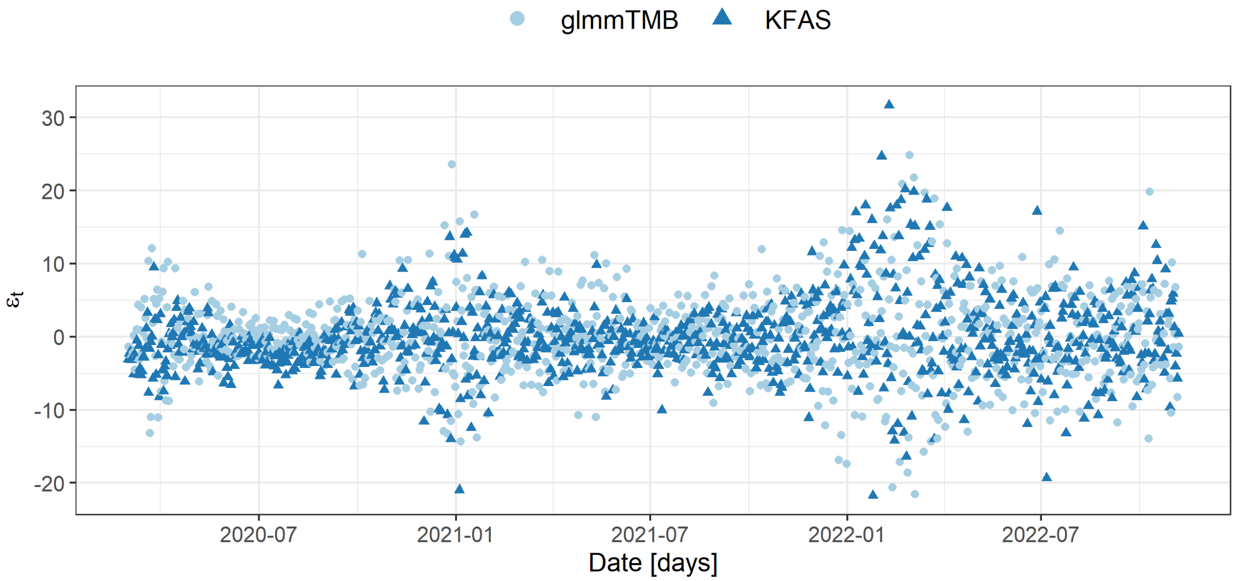

Figure 4: Model residuals from glmmTMB and KFAS.

The smoothed estimate of the latent state estimated from `glmmTMB` and `KFAS` is visualized in Figure 5.

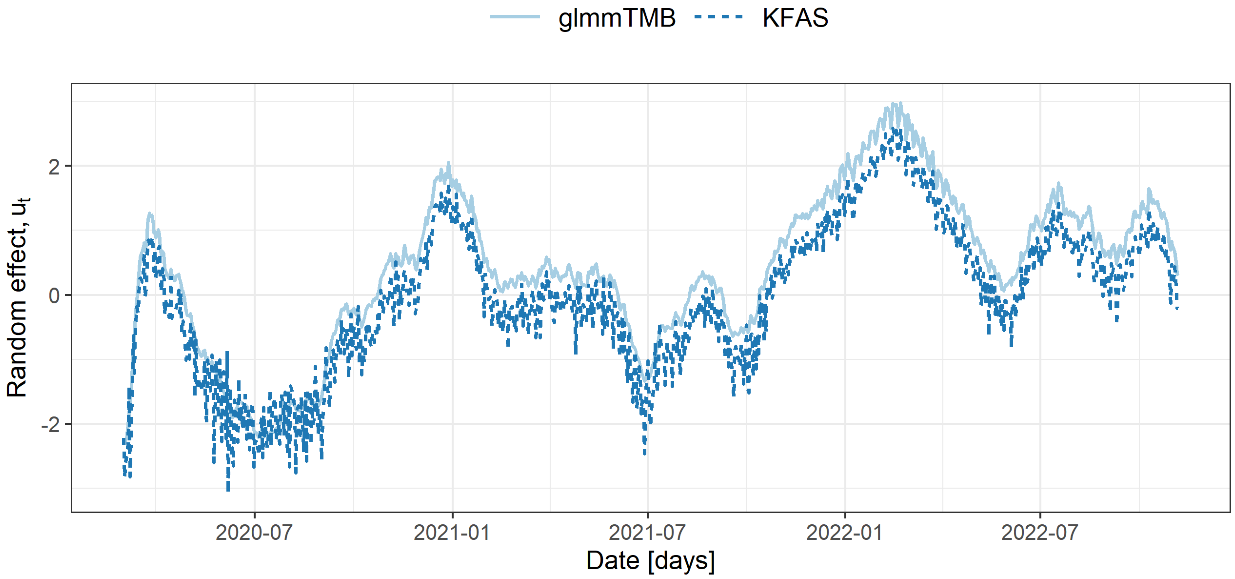

Figure 5: Smoothed estimates of the latent states from `glmmTMB` and `KFAS`.

Estimating the number of new hospital admissions in Denmark grouped by region

In Figure 6 the total number of new hospitals admissions in Denmark grouped by region is visualized together with the smoothed estimates from `glmmTMB` and `KFAS`.

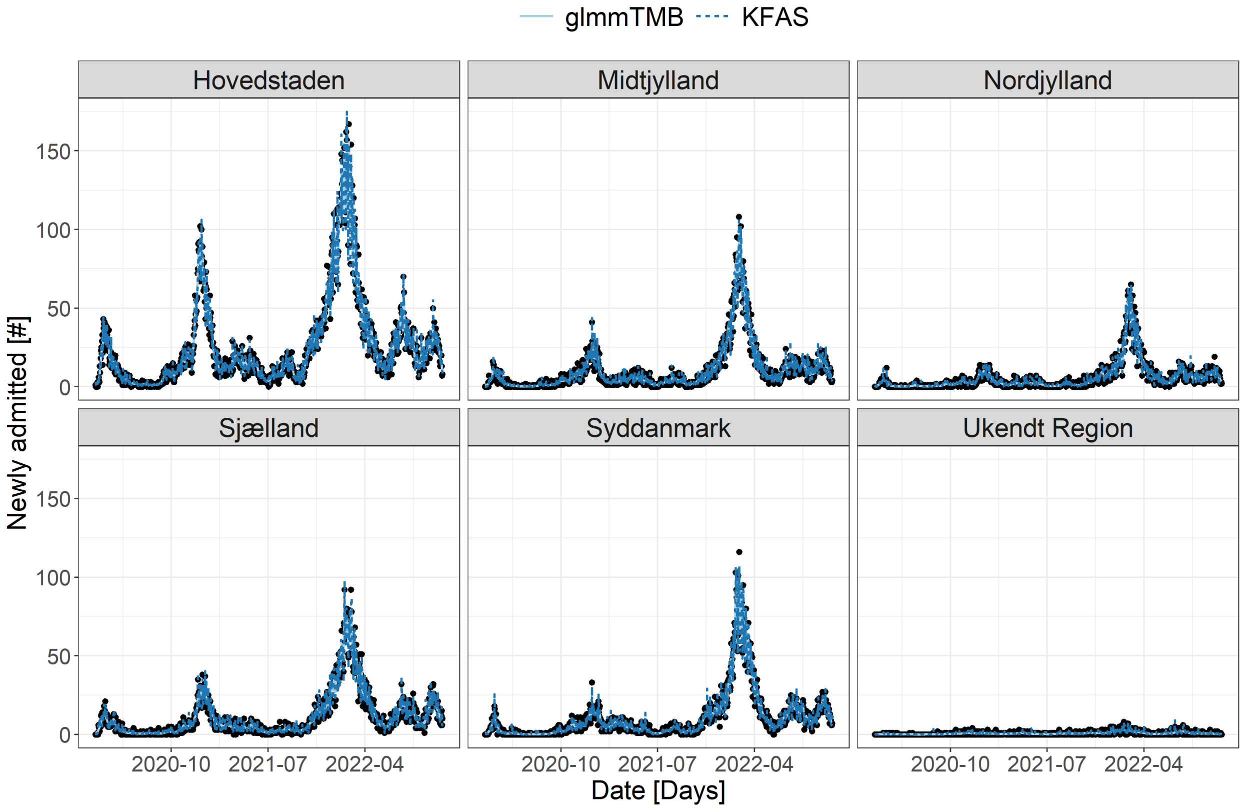

Figure 6: Total number of new hospital admissions in Denmark grouped by region with smoothed estimates from glmmTMB and KFAS.

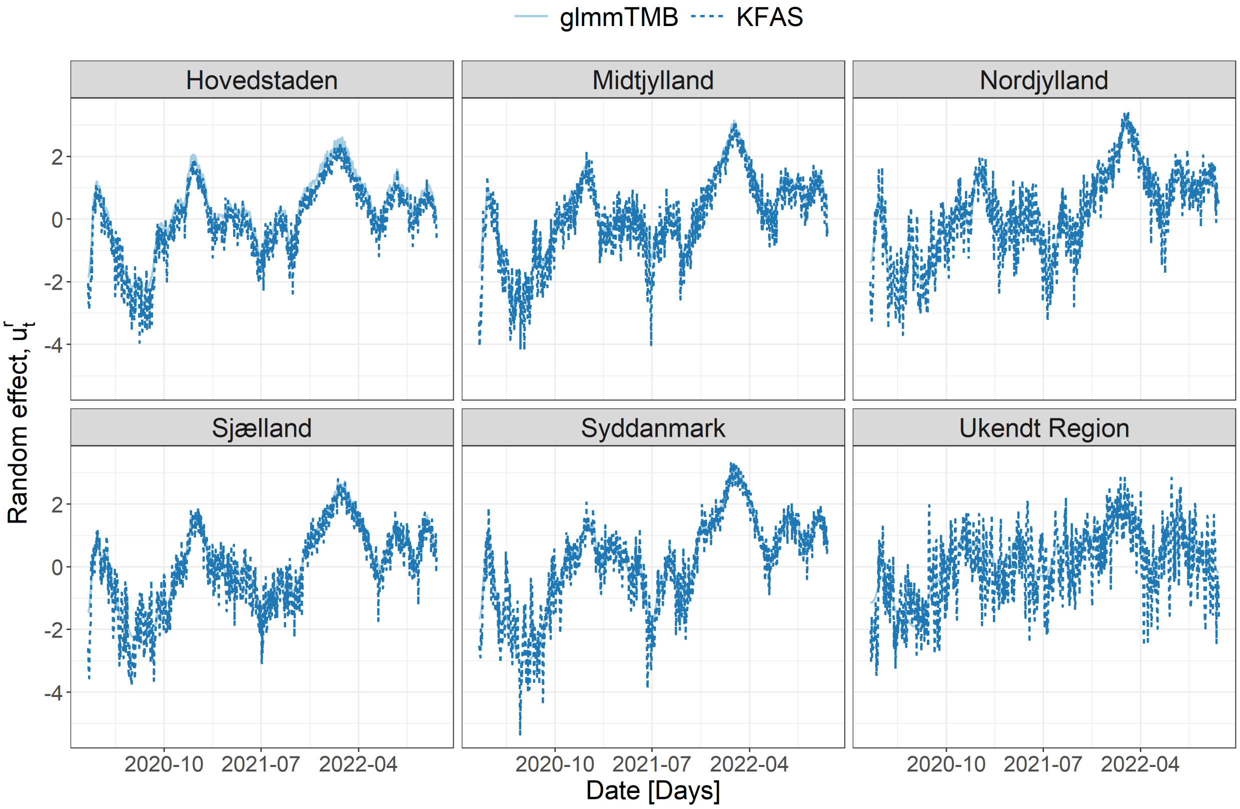

Figure 7: Smoothed estimates of the latent states from `glmmTMB` and `KFAS`.

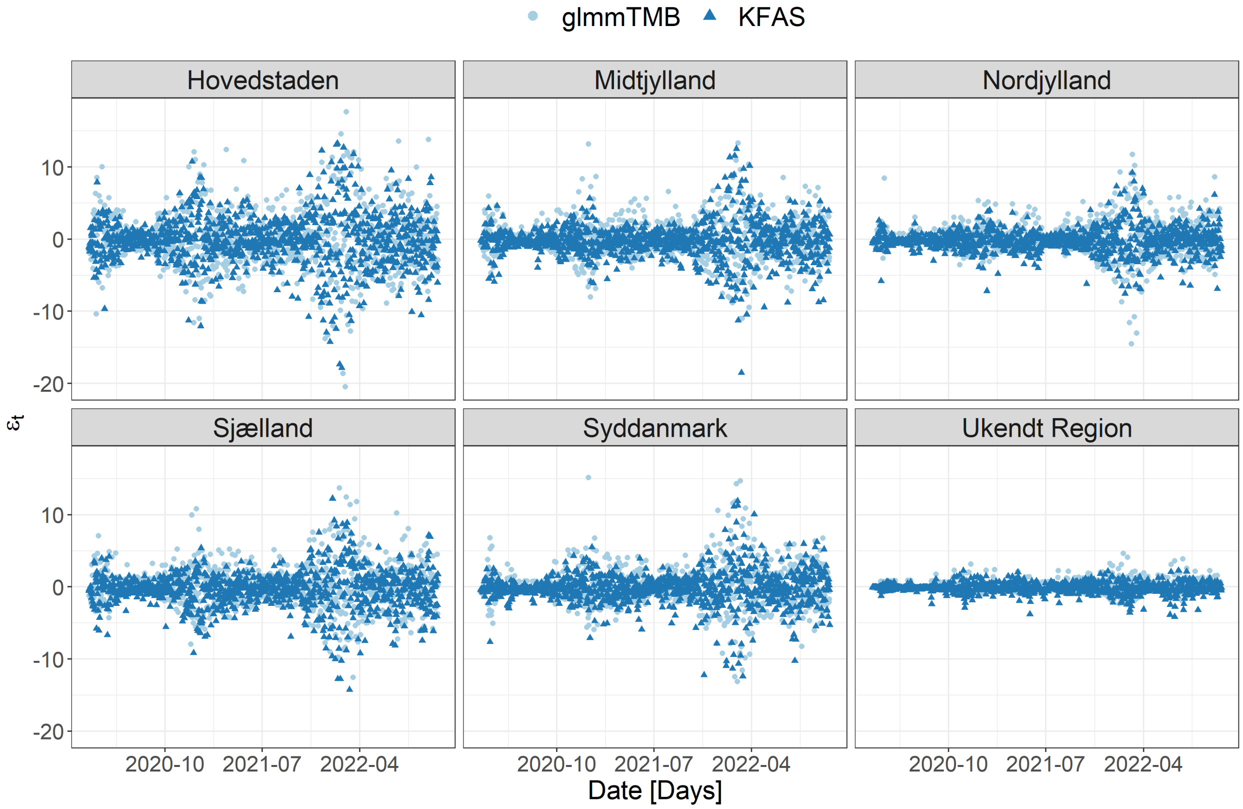

Figure 8: Model residuals from `glmmTMB` and `KFAS`.

Discussion

Write it later

References

Brooks, M. E., Kristensen, K., van Benthem, K. J., Magnusson, A., Berg, C. W., Nielsen, A., Skaug, H. J., Maechler, M., & Bolker, B. M. (2017). glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *The R Journal*, *9*(2), 378–400. https://doi.org/10.32614/RJ-2017-066

Helske, J. (2017). KFAS: Exponential family state space models in R. *Journal of Statistical Software*, *78*(10), 1–39. https://doi.org/10.18637/jss.v078.i10

Madsen, H. (2007). *Time series analysis*. Chapman & Hall. https://doi.org/10.1201/9781420059687

Madsen, H., & Thyregod, P. (2011). *Introduction to general and generalized linear models*. CRC Press.