Google Trends Covid

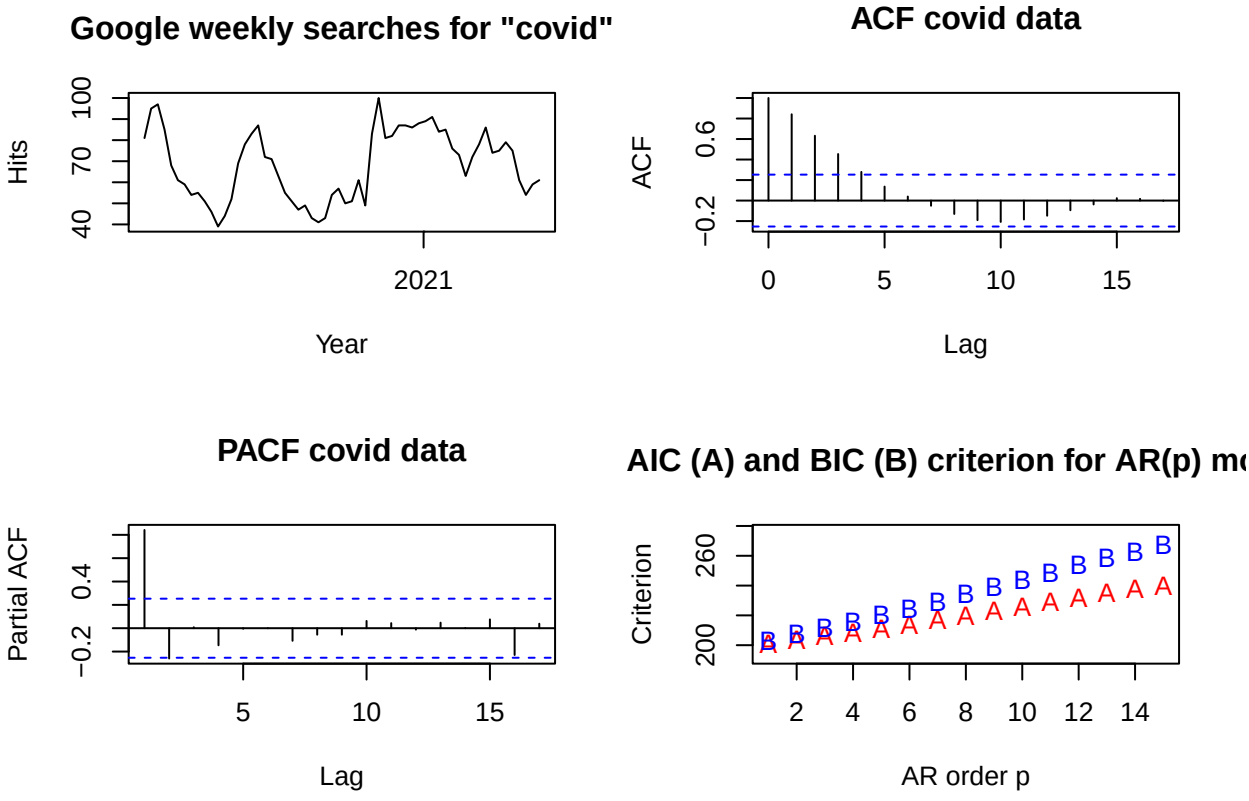
Gerardo R.

2023-11-08

Abstract. An autorregressive AR(1) model and a mixed autorregressive mAR(1,k=5) have been fitted for time-series data involving weekly Google searches for the term "covid" between the time range 03/2020 - 05/2021. Both models predict that in the last three weeks of May 2021 the number of searches for the term are likely to increase.

Introduction and data description. The data employeed is based on the number of weekly Google searches performed between 03/2020-05/2021. It can either be downloaded manually from Google or using the Rpackage Rgtrends. This project aims to practise bayesian modelling applied to AR and mAR models for this time series data, which is depicted below.

Data exploration. The ACF plot shows significant correlations between time series in the first 5 lags (so the model we are making will be with those time series). The ACF and PACF plots presents a decay, suggesting the presence of a potential **linear trend** and **seasonality** in the data.



AR(p) model

Order of the model The Akaike Informative Criterion (AIC) and Bayesian Information Criterion (BIC) were computed for 8 orders (see previous figure). The values have been regularized with a 1e-5 value to elude problems in the Cholesky decomposition (**this can make my plot different to yours!**). Based on the AIC and BIC criterions, the best model is order 1.

The hierarchical model equation. The AR model can be defined through the following equation:

$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \epsilon_t, \quad \epsilon_t \sim N(0, 1)$

Assumptions. The error is normally distributed with constant variance (initial 1). Also assumes a normal inverse gamma model: $x_t = \phi_0 + \phi_1 x_{t-1} + \ldots + \phi_p x_{t-p} + \sigma z_t$

Prior parameters and parameter estimation. 1000 simulations (Gibbs sampling) were run considering an AR model of order 1 and using flat priors. The next hyperparameters were chosen:

$m_0 = (0, 0)^T, C_0 = I_2, n_0 = 2, d_0 = 2$

Based on the sample trace plots the parameter estimation process converged. Their Gaussian posterior distributions are represented with histogram representations. The posterior means of the parameters and their distributions are also reported:

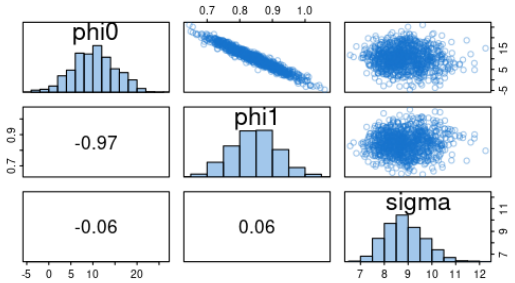


Figure 1: AR model estimates

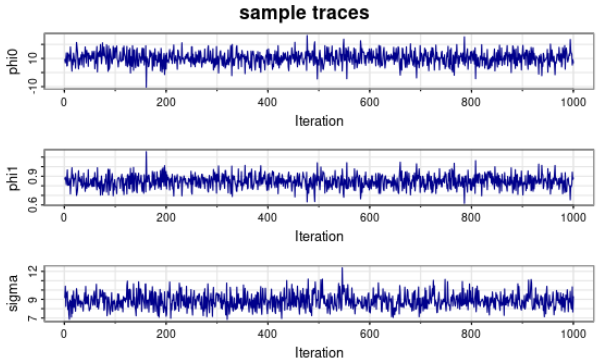
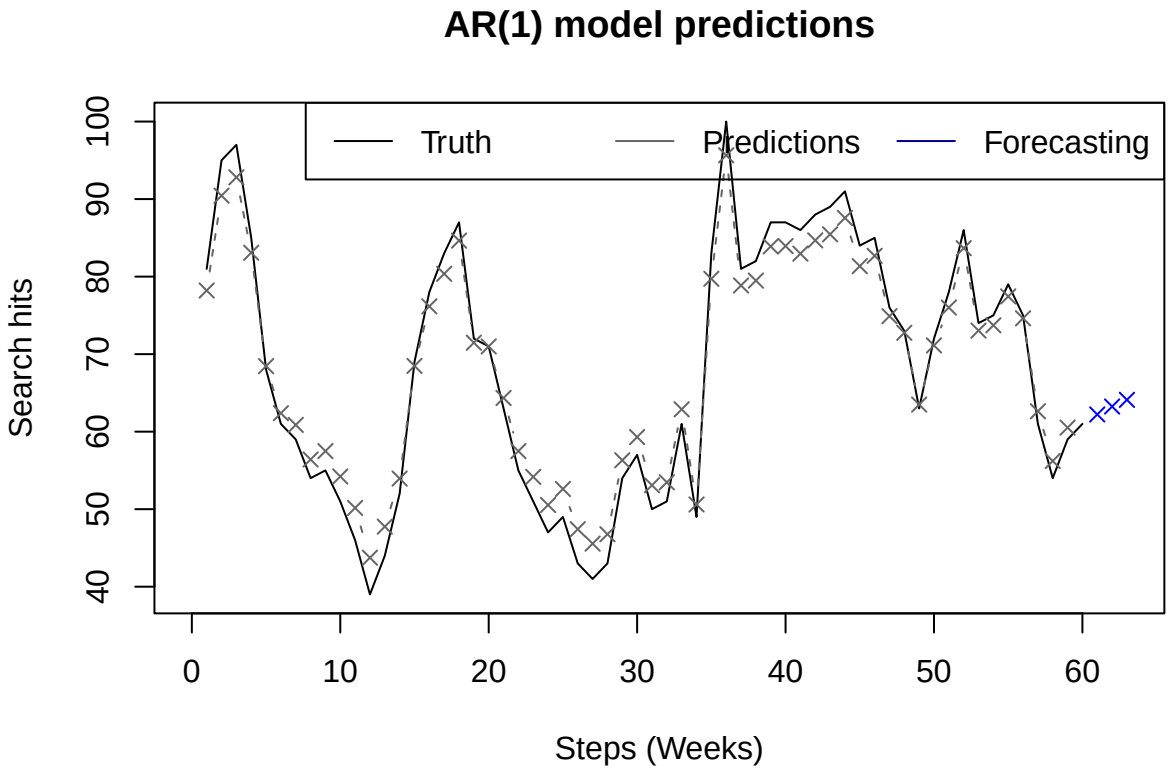


Figure 2: Sample traces

Model predictions and forecasting. Using the AR(1) model we now perform model predictions and forecast 3 weeks ahead of the results. Let's see how good it does.



As we can see in the picture this simple AR(1) is already pretty good approximating the time series curve. Additionally we can see that for the predictions in May 2021 it predicts an increase in the number of search hits for the word covid. Let's see how a mixture model will perform instead!

Mixture AR(p,k) model

Order of the model We select the same order of the model as before based on our previous results obtained with BIC and AIC estimations; p=1.

Number of components estimation. We evaluated with the DIC (Deviance Information Criterion) based on 10,000 simulations to determine the number of components that our model should have, putting an arbitrary limit of 5 (as we don't want our model to get too complex). Based on DIC, the number of components should be k=5.

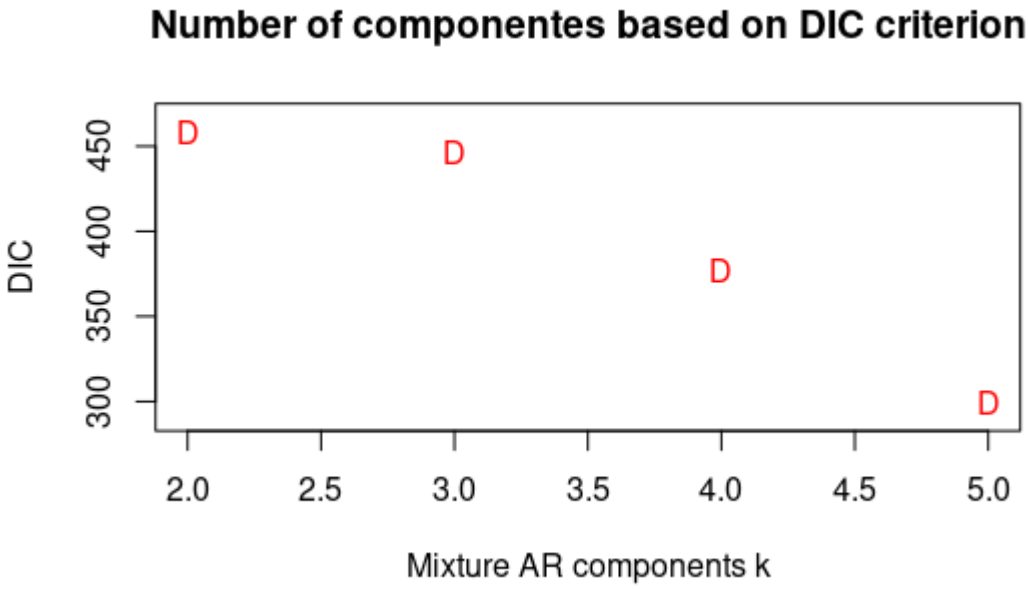


Figure 3: DIC values based on K components for the AR model

The hierarchical model equation. The model can hierarchially be written by the next equations:

$t_t \sim \sum^K w_k N(f_t^T \beta_k, \nu_k)$, $ f_t^{T=(y_{t-1},\ldots,y_{t-p})}T$, $ t=p+1,\ldots,T$

$\omega_k \sim Dir(a_1, ..., a_k)$ $\beta_k \sim N(m_0, \nu_k c_0)$ $\nu_k \sim IG(n_0/2, d_0/2)$

Assumptions and simplifications. The full posterior distribution can be written as:

$p(\omega, \beta, \nu, L|y) \propto p(y|\omega, \beta, \nu, L) \mathring{u} p(L\omega) \mathring{u} p(\omega) \mathring{u} p(\beta) \mathring{u} p(\nu)$

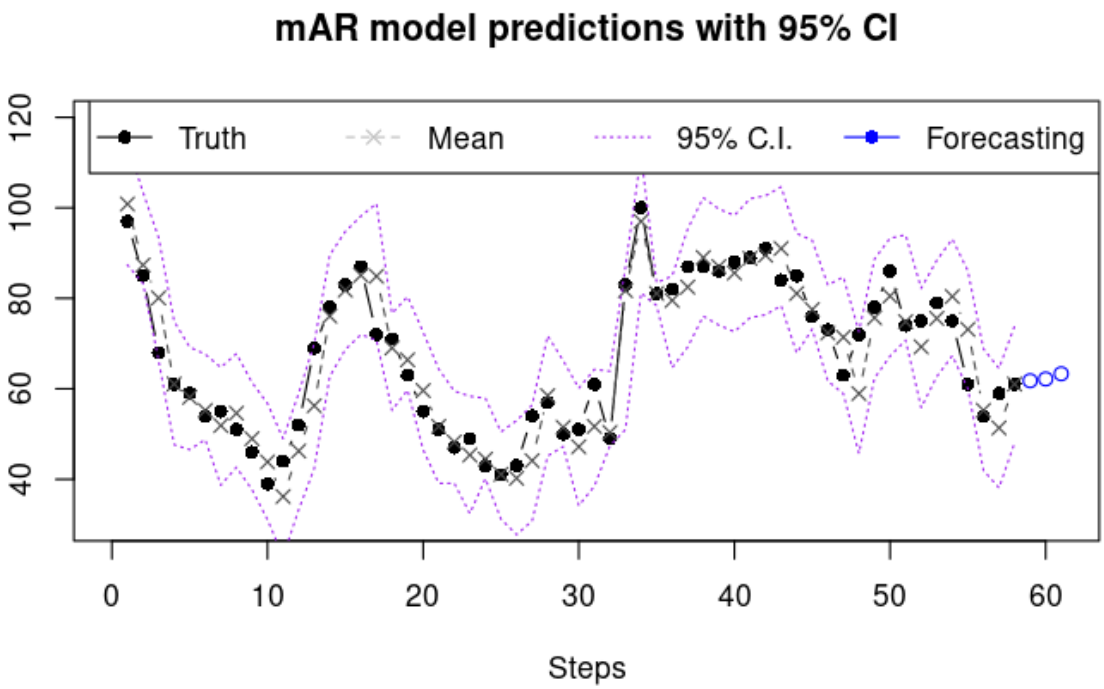
And the full conditional distribuitions of all model parameters is reduced to:

$\omega|... \sim Dir(a_1 + \sum^T 1(L_t = 1), ..., a_K + \sum^T 1(L_t = K))$

$p(L_t = k|...) \propto \omega_k N(y_t|f_t^T \beta_k, \nu)$

Priors and parameter estimation Considering the assumptions defined in the previous equations, we fitted an AR mixture model with location p=1 and k=3. The model hyperparameters were $a_1 = a_2 = a_3 = 1, m_0 = (0)^T, c_0 = 10$ and $n_0 = d_0 = 100$. The Gibbs sampling function was again used also for this model and the parameter estimation, which was computed using 10,000 runs.

Model accuracy and predictions. The original points (black) and model predictions (purple) are represented with a 95% Confidence Interval (gray). The predictions of the 3 next steps/weeks for the model are also represented in blue.



Which model performs better?. We can see that the predictions are very accurate and the predicitons for May 2021 (blue dots) are similar to the ones predicted by our AR(1). Now here comes the big question. Which one performs better? Well, based on merely the residuals the mixture mAR model ($R_{mAR} = 223.6$) manages to fit better the data than the simple AR model ($R_{AR} = 418.7$). In this regard, the mixture model performs better. However, we should also consider that is much more complex in the same way.

Conclusion. Both models perform very well based on the graphical representation of the predictions. Despite of showing a better accuracy based on its residuals, both models; AR(p=1) and mAR(p=1, k=5) seems to fit the data fairly well. Both models predict an increase in the number of Google searches of the term "covid" in the last 3 weeks of May 2021.

Additional notes for replication

Software details. The whole project was run using the R packages: MCMCpack, mtvnorm, gtrendsR, dlm. For replication purposes, you can just introduce the parameters specified in the model section. The data obtaiend was downloaded through the R package: gtrends. All code written was done using R programming.