Bayesian Inference for a Time Series

Group 11

Yi Le

Introduction

Air pollution is a huge problem in several cities in China, and PM2.5 is widely used to evaluate air quality. It will be practical and that we can implement a Bayesian method to make prediction for PM2.5 time series. Besides basic Bayesian inference, a more advanced Bayesian method will be required to develop the relationship between different time series.

Method

Prediction

For now, we have time series of PM2.5 in five cities in China. The purpose is to generate a function that can predict next number in this series in a Bayesian approach.

First, we need to clean data, keeping only PM_US post (The PM2.5 index announced by local US Embassy) and find the longest contiguous PM2.5 time series in each city.

The Bayesian inference that will be implemented in this case is the formula:

$$P(next|current) = \frac{P(current|next)P(next)}{P(current)}$$

The PM2.5 index will be categorized into ten groups, as a result, the time series will contain integer from one to ten. Given PM2.5 in this moment as a condition, we can then predict PM2.5 in next moment, based on formula above. The key step in this naïve method is to generate a matrix to stand for the whole result for this time series, for example, we have a time series

$$2, 1, 2, 3, 1$$

Where value can only be 1, 2 and 3. A matrix P can be generated to describe how time series change

	1	2	3
1	0	1	1
2	1	0	0
3	0	1	0

With row name standing for destination while column name standing for starting point. The number in this matrix reveals how many times it happens. Furthermore, it can be a probability matrix

	1	2	3	P(next)
1	0	0.5	0.5	0.5
2	1	0	0	0.25
3	0	1	0	0.25
P(current)	0.25	0.5	0.25	

We can calculate all conditional probability based on this simple table, as all probabilities have already been listed.

Relationship between Time Series

Both PM2.5 and temperature can be regarded as time series data, they are self-related, and may be somehow related to the other. PM2.5 can be regarded as a Bayesian structural time series, because we regarded temperature as a possible influence for it. The time series can be described by three components: constant, trend and regression.

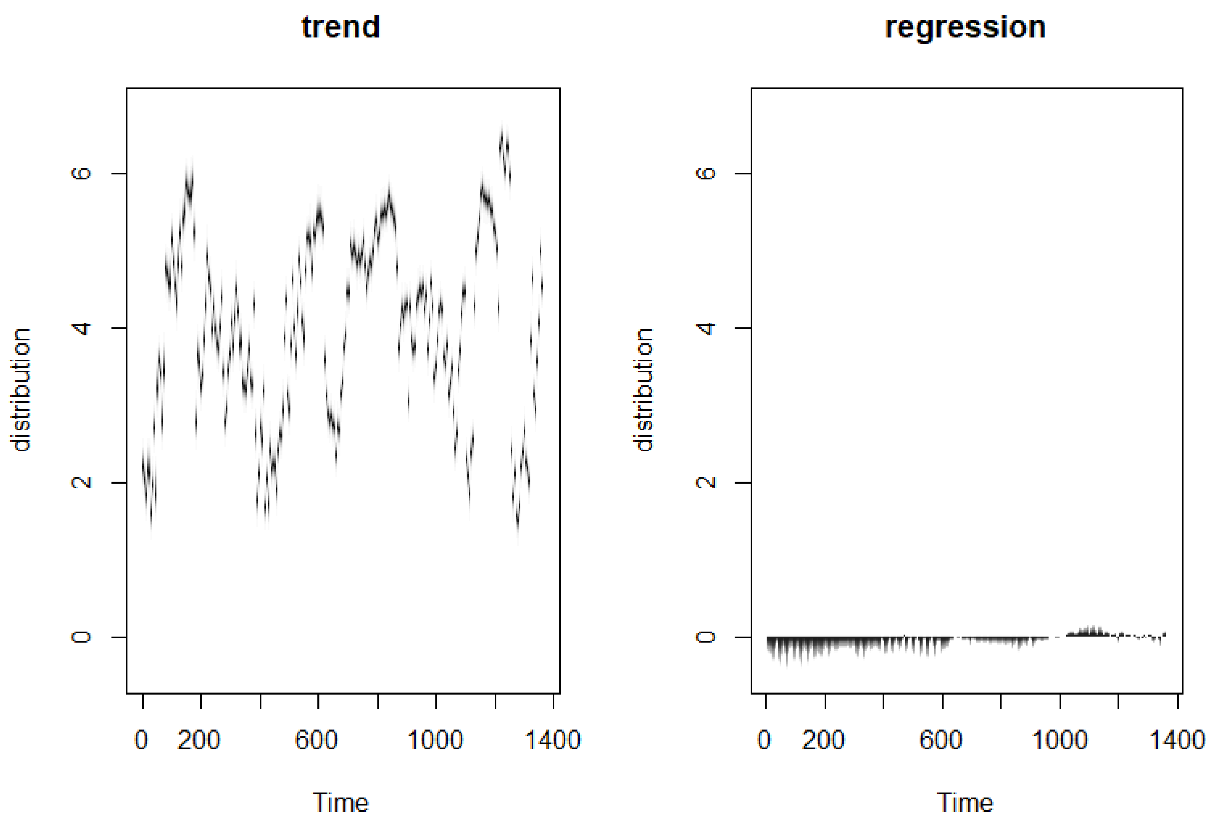
By MCMC, we can sample from the posterior distribution of PM2.5, and according to optimal Kalman forecast, we can pick the time series with the greatest ability to make forecast. Luckily a R package named bsts has already been published to solve such problem. By using this package, we can observe both self-trend and influence from other variable in a time series.

Results

For the first part of this project, the result is a function named predictpm25, you can implement it by entering current PM2.5 index and the name of city, it will work like this:

```
> predictpm25(120,"Beijing")
Expected value: 91~120
                 Pr
1~30     0.09932954
31~60    0.09954817
61~90    0.10137006
91~120   0.10260895
121~150  0.10020405
151~180  0.09954817
181~210  0.09932954
211~240  0.09940242
241~270  0.09932954
271~more 0.09932954
```

For the second part, the result can be seen below (take Beijing as an example),



The left side is the trend of log(PM2.5), while the right side is the regression relationship between log(PM2.5) and temperature.

Conclusion

Based on a naïve Bayesian inference, a predict function for PM2.5 is generated. For further development of relationship between different time series, it's obvious that PM2.5 series in Beijing is not stable, and has no regression with temperature, according to a Bayesian structure time series method.

Reference

Scott, S. L., & Varian, H. R. (2014). Predicting the present with bayesian structural time series. *International Journal of Mathematical Modelling and Numerical Optimisation*, *5*(1-2), 4-23.