Markov Random Fields And Its Application in Time Series Decomposition

Bowen Xiao

May 16, 2018

Intrduction

The project is about Markov random fields, which is a locally adaptive nonparametric curve fitting method that operates within a fully Bayesian framework. The model assumes that each data point is generated independently with some parametirc models, and the parameters are follow a Markov random fields model, which could, according to the paper[1], provide a combination of local adaptation and global control. Specificly, after removing seasonal component, my trend model used for forecasting looks like this:

$$
\begin{aligned}
y_j &\sim normal(\theta_j, \sigma); \\
\sigma &\sim exponential(1) \\
\theta_j &= \rho^1 \theta_{j-1} + \delta_{j-1}; \\
\theta_1 &= 5 * sd(y) * \theta_0 + \bar{y}; \\
\theta_0 &\sim normal(0, 1); \\
\rho^1 &\sim beta(2, 2) \\
\delta_j &= \rho^2 \delta_{j-1} + \delta_{j-1}^1; \\
\delta_j^1 &\sim normal(0, 1); \\
\delta_1 &\sim normal(0, 1); \\
\rho^2 &\sim beta(2, 2)
\end{aligned}
$$

where the trace of θ is trend component.

The idea of Bayesian Markov random fields comes from the paper in the reference[1]. This project includes the following two parts, simulation study of Gaussian Markov random fileds (GMRF) fitting and time series decomposition based on Markov random fiels. For time series decomposition, since it is count variable, in the first fitting I replace $y_i \sim normal(\theta_i, sigma)$ with $y_i \sim Poisson(e^{\theta_i})$. And different priors for δ are compared in both two fitting. As is well known, Bayes methods treat parameters as random variables. consequently, what we get is posterior distribution of θ. To compare with frequentist methods, I will simply use the median of θ.

What I want to get in time series decomposition is a nonparametric trend component, which is the main difference with classical methods, like ARIMA. In other word, I will not assume the parametric form of the trend component, like linear trend. Furthermore, I compare the results with some classical parametric techniques, like ARIMA, and machine learning techniques, like RNN/LSTM, with regrard to forecasting. Metrics for model evaluation and comparison is MSE. I make prediction straightforwardly by simulating data from Bayesian model, which is different from my original idea in proposal.

Data Description

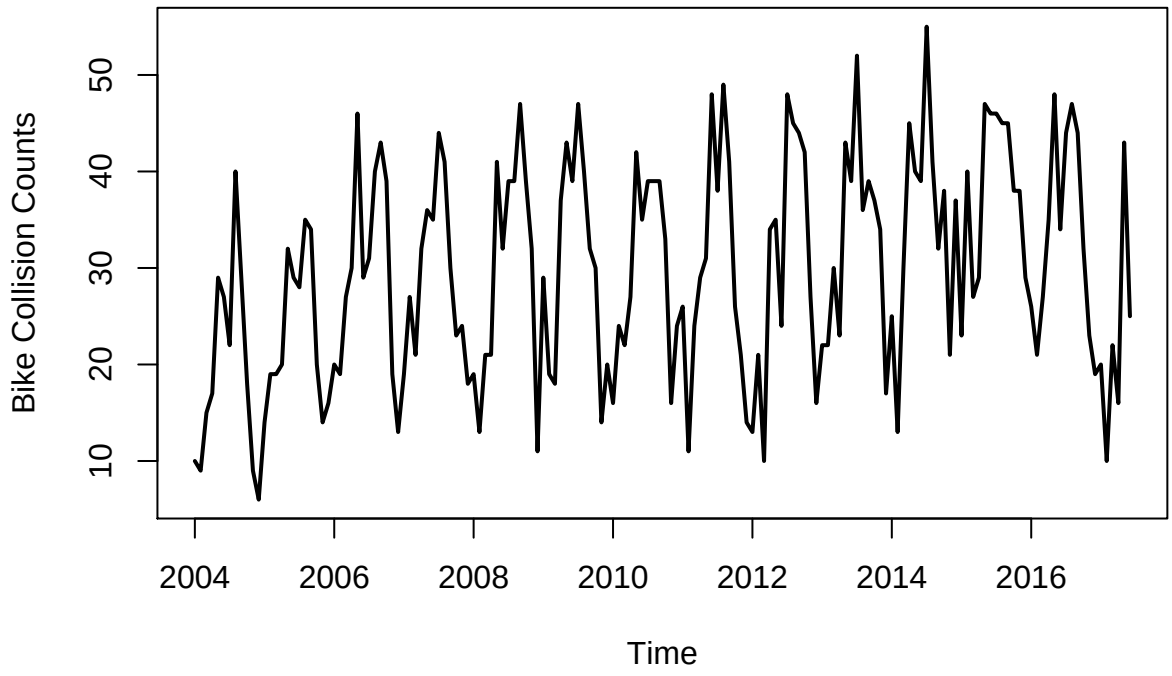
Local Smoothness

First of all, I apply Bayesian Markov random fields into nonparametric fitting and compare it with polynomial regression and NW estimators. Data will be generated randomly: $X_i \sim^{iid} Uniform(0, 1)$ and $Y_i = f(X_i) + \epsilon_i$,

where f is known, $\epsilon_i \sim^{iid} N(0,1)$ and ϵs are independent with Xs. In this project, I consider two scenarios: $f(X_i) = 2X_i$ and $f(X_i) = sin(10X_i)$.

Time Series Decomposition

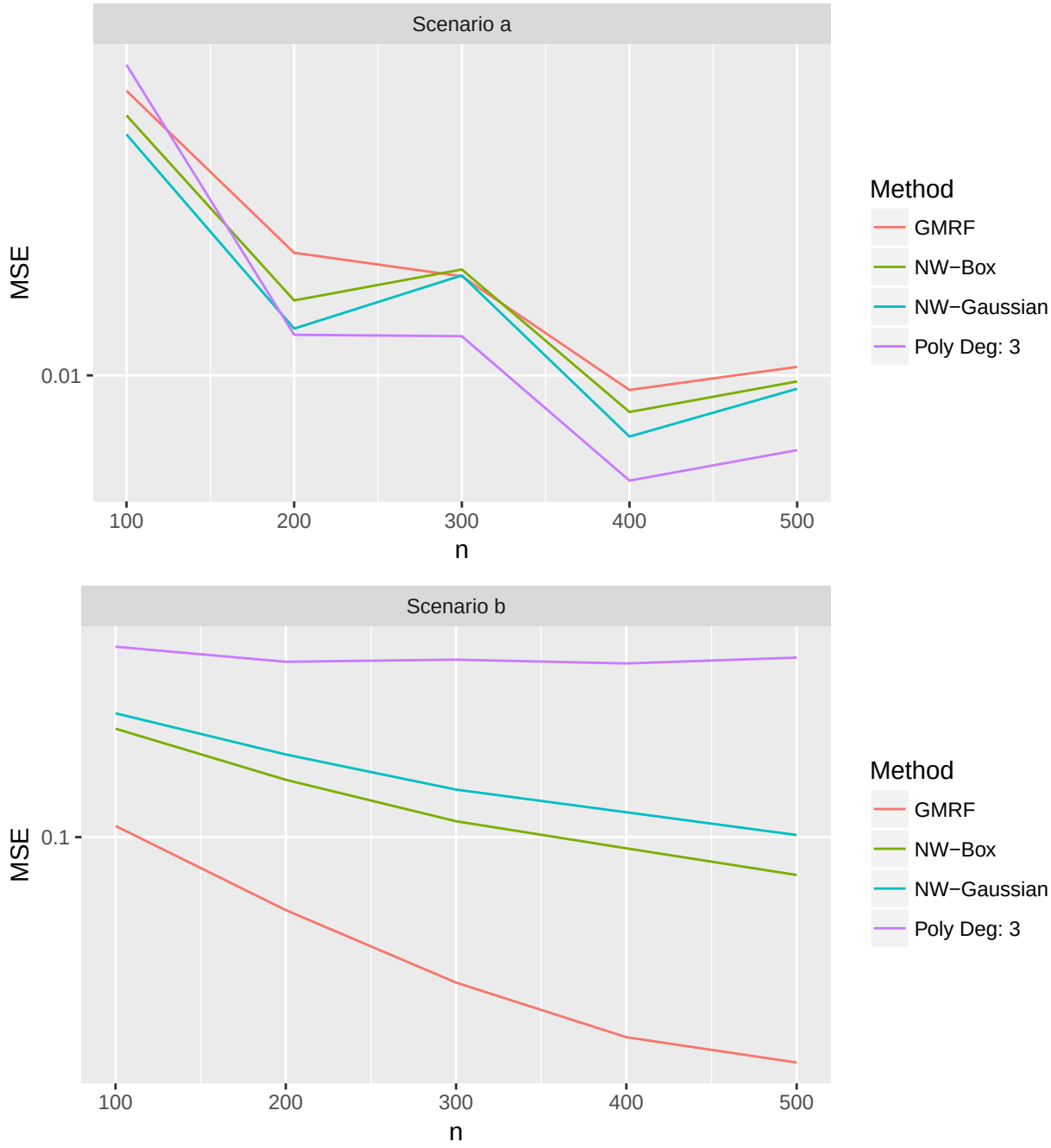
Secondly, I apply this techinique to my time series data. Specificly, I have data of bike collision records in downtown Seattle from 01/2004 to 06/2017 and I summarize them by month, so that I have 162 bike collision counts, which is shown as following. I split the data into train set (the first 150 points) and test set (the last 12 points).



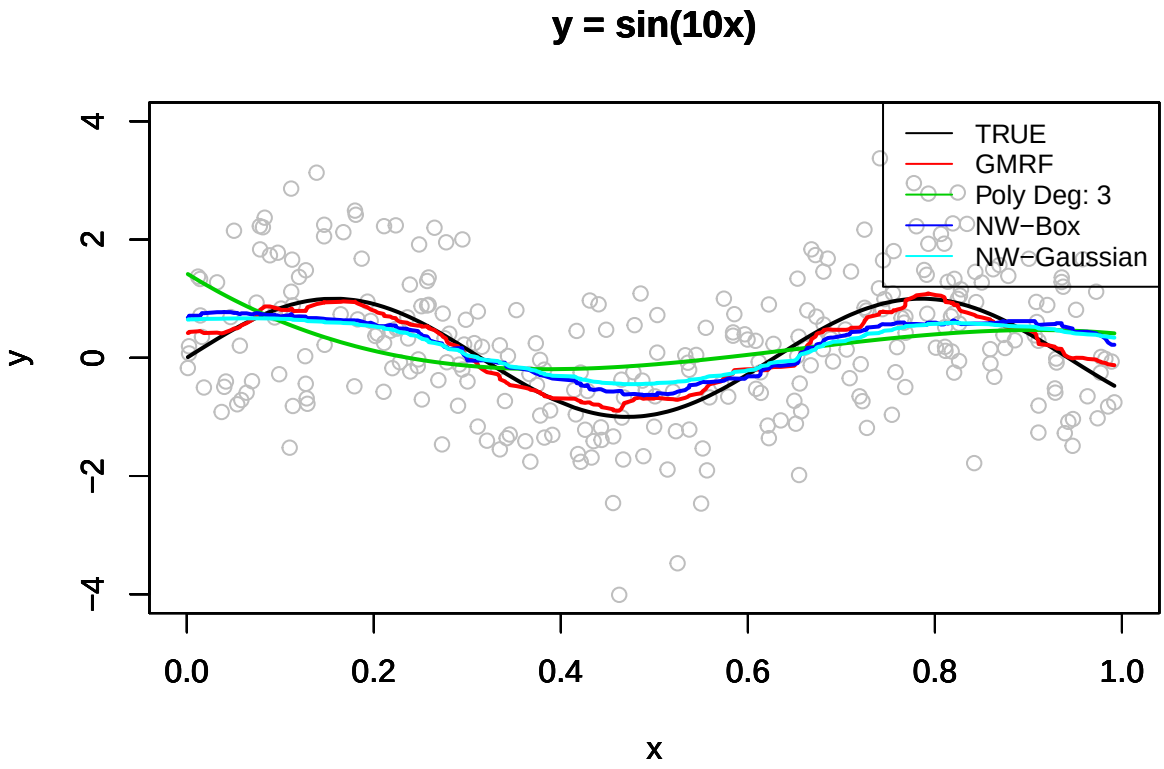
Simulation

I run the full simulation study for each of the 2 functions $f(x)$, and I calculate the MSE for $n = 100, 200, \ldots, 500$ and for each n, we report the average MSE for each n, averaged over 10 replications of the data.

My choice of bandwidth for NW estimators is motivated by the *optimal* bandwidth, which states that the optimal bandwidth should be $h_* = Cn^{-1/5}$ where the constant C depends on the function f, variance of noise, and density of x_i. In this simulation, to keep things simple, we just take $C = 1$.



As is shown, GMRF fitting performs worst in the linear scenario but appeals to have some advantages in nonlinear scenario. Actually, the idea GMRF is very similar with some well-used nonparametric estimators, like KNN and NW estimators, because they all focus on local information. The difference is GMRF always treats $\{x_i, y_i\}$ as a sequence and it seems to be more flexible with the regularization assumptions of f, which is why I consider to apply it into time series. The referenced paper shows plenty of simulation study, including smoothing functions and piecewise constant function[1]. One example of my fitting is shown as following.



Time series decomposition

I apply GMRF fitting into time series decomposition and forecasting. Firstly, I use GMRF to fit a smoothing line. I split the line by bandwidth of 12 (since I know the period is 12 month), and avarage the difference with the mean in each piece. So that I get the seasonal component. Secondly, I minus the raw data with seasonal component and do another GMRF fitting on it, which turns out to be the trend component. For the first fitting, y is Poisson distributed because it is count variable, but for the second fitting I use normal distribution.

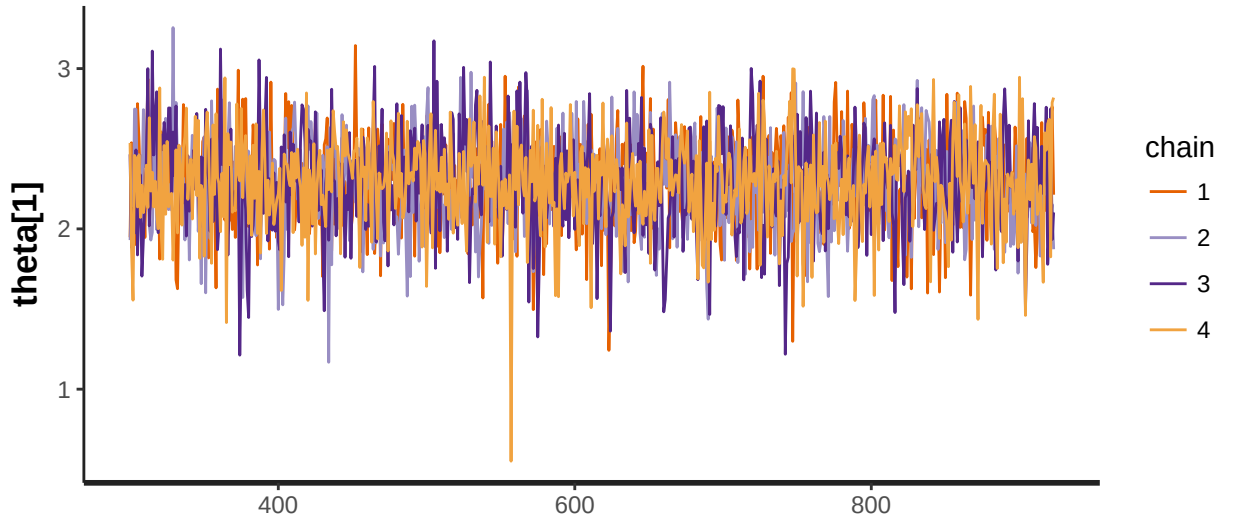
I try three priors in first fitting: Gaussian prior, Laplace prior and Horseshoe prior. There is no issue of divergence. And I compare them based on LOO-PSIS-CV (efficient approximate leave-one-out cross-validation for Bayesian models fit using Markov chain Monte Carlo, which uses Pareto smoothed importance sampling, a new procedure for regularizing importance weights.) and WAIC (a generalized version of AIC)[2].

	elpd_diff	elpd_loo	se_elpd_loo	p_loo	se_p_loo	looic	se_looic
loo2	0.000000	-531.5500	3.666170	96.01435	2.075922	1063.100	7.332340
loo3	-2.745540	-534.2955	12.194588	64.80410	5.673001	1068.591	24.389175
loo1	-9.072317	-540.6223	3.797943	103.95685	1.712851	1081.245	7.595886

	elpd_diff	elpd_waic	se_elpd_waic	p_waic	se_p_waic	waic	se_waic
waic2	0.000000	-498.7724	2.954513	63.23676	0.9022137	997.5448	5.909027
waic1	-5.209681	-503.9821	3.091133	67.31663	0.5741754	1007.9642	6.182267
waic3	-29.672419	-528.4448	11.788323	58.95339	5.2504230	1056.8897	23.576646

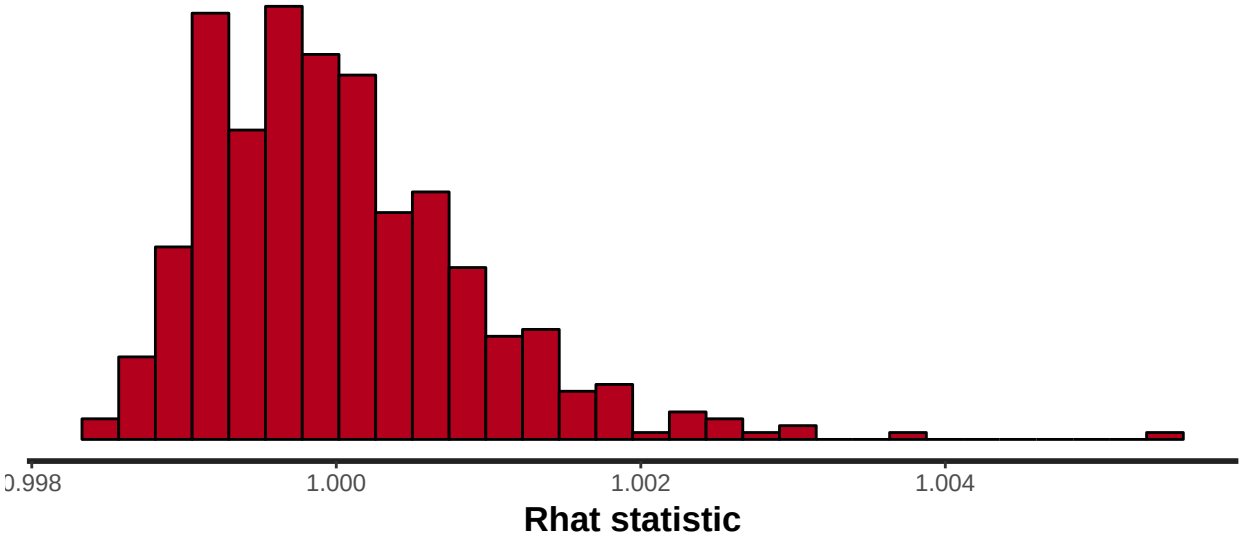
Both LOO-PSIS-CV and WAIC indicate model based on Laplace prior is the best. Thus, I will go on with this model.

The trace of $theta_1$ in the sampling can be shown as following.

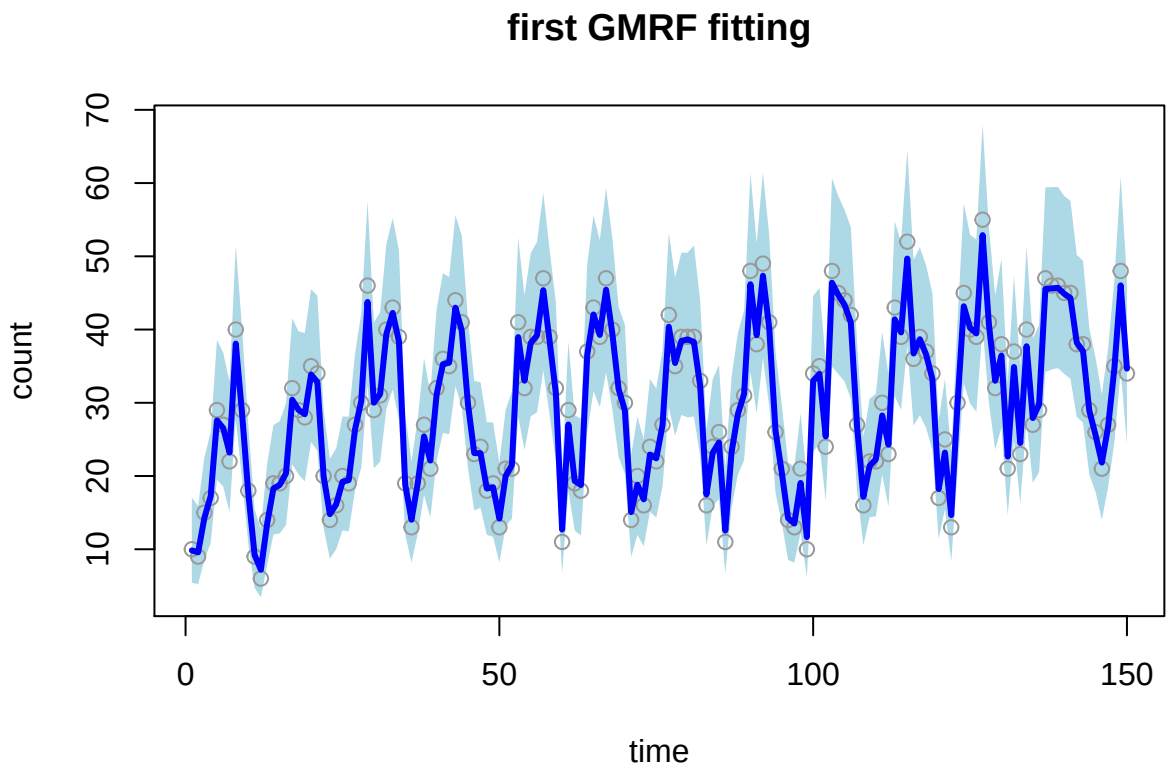


The overlapping lines show that the samples from 4 different chains are comming from one common distribution, or that, there is no violation for $theta_1$. The conclusion is also true for other parameters, because as we can see in the following, all the $\hat{R}s$ are close to 1 (The degree of convergence of a random Markov Chain can be estimated using the Gelman-Rubin convergence statistic, $\hat{R}$, based on the stability of outcomes between and within m chains of the same length, n. Values close to one indicate convergence to the underlying distribution. Values greater than 1.1 indicate inadequate convergence.)[3].

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

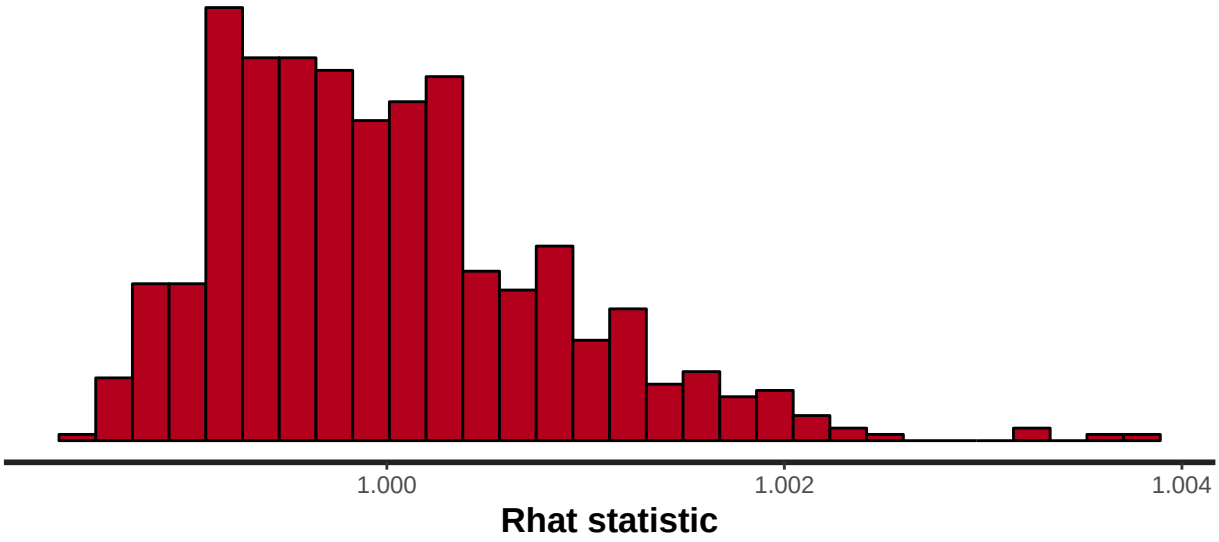


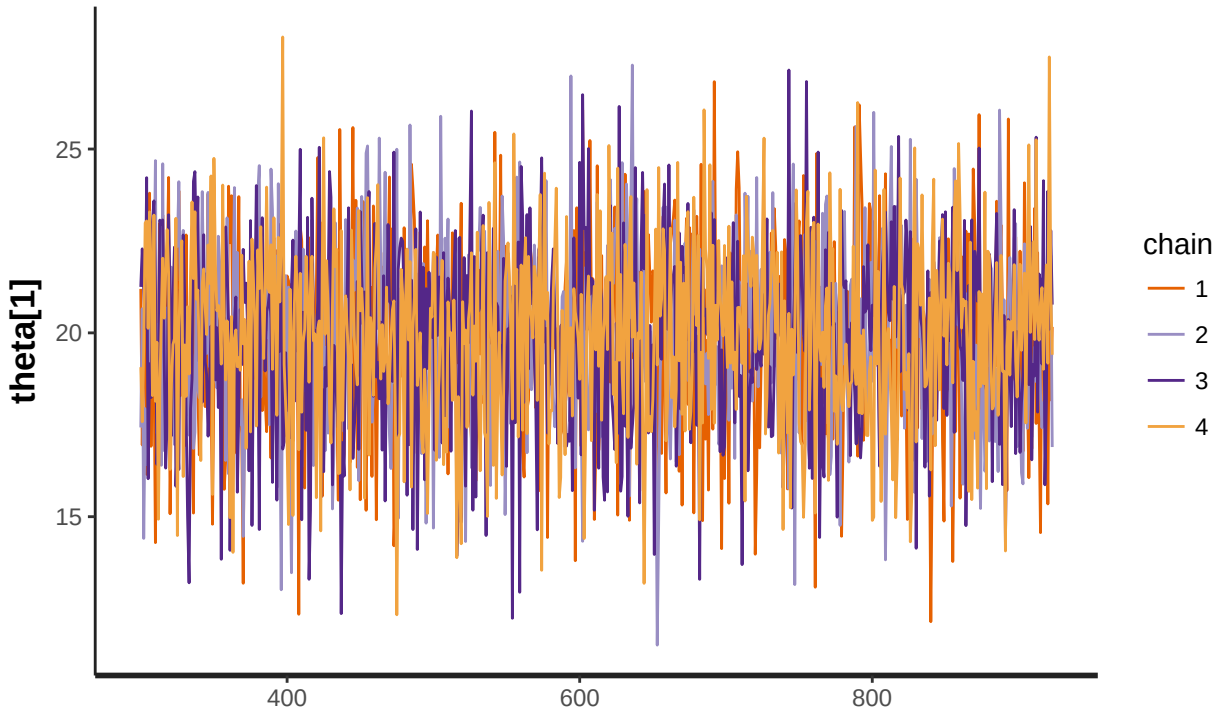
95% credible interval of θ can be shown as following.

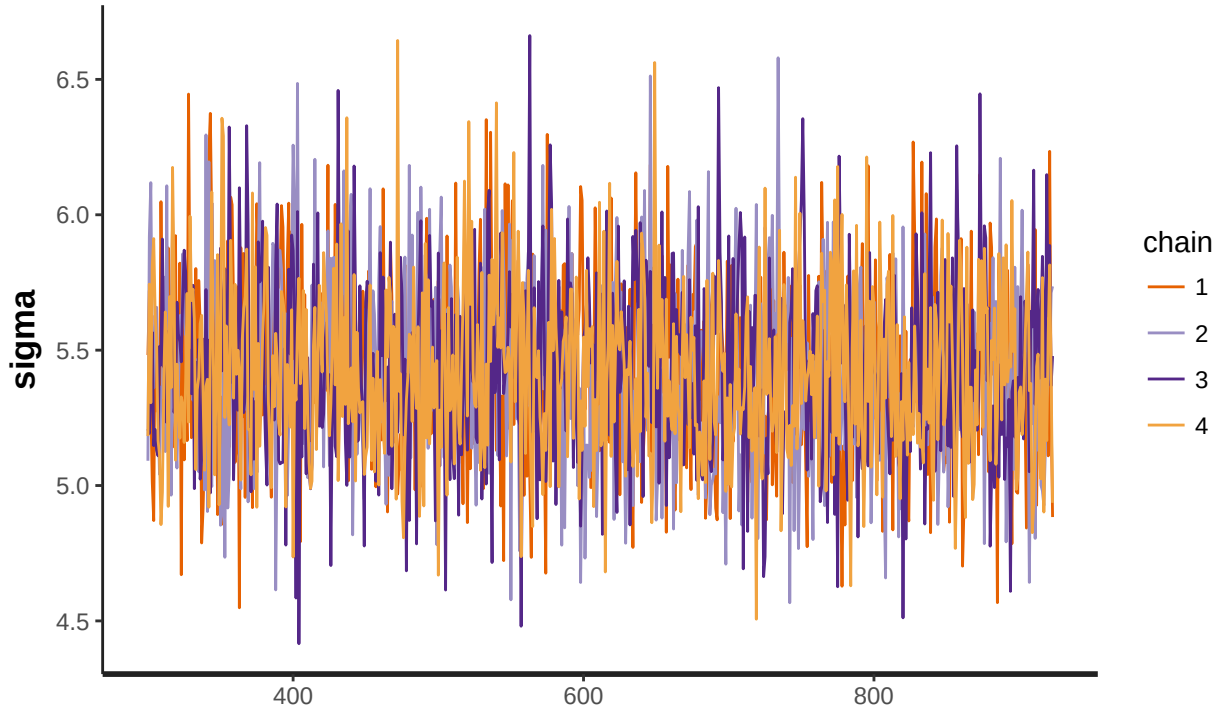


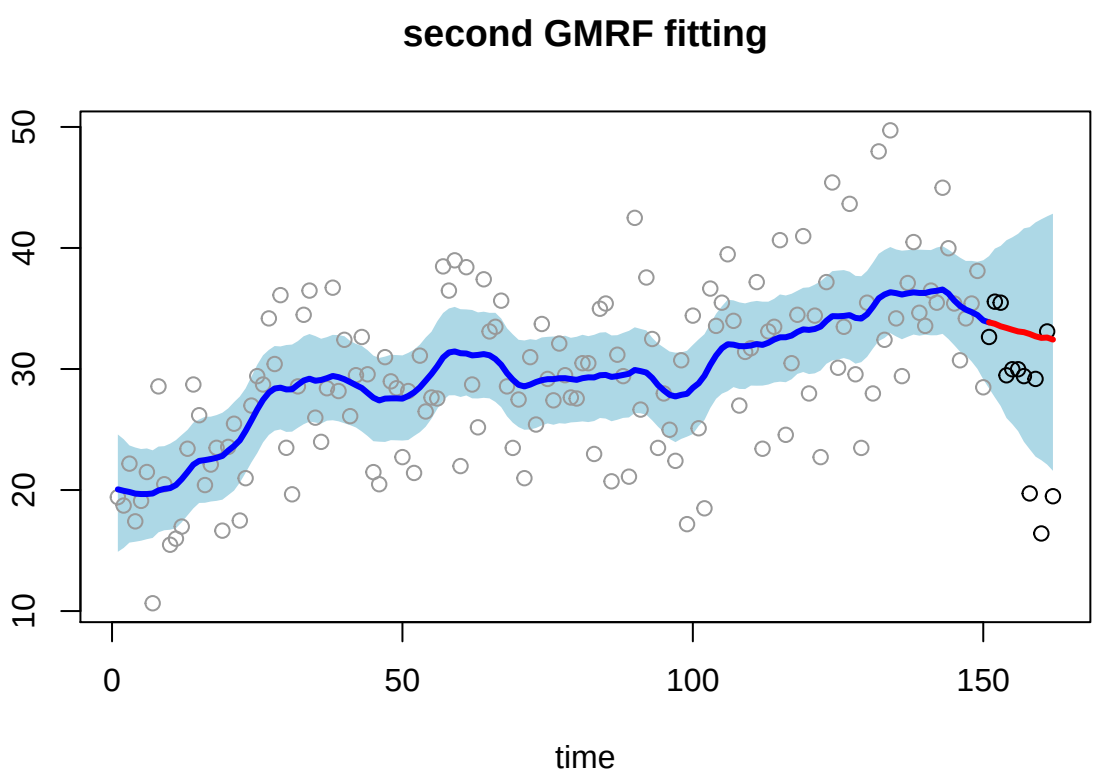
In the second fitting, there is no divergence issue either. Similarly, here are the results of second fitting.

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



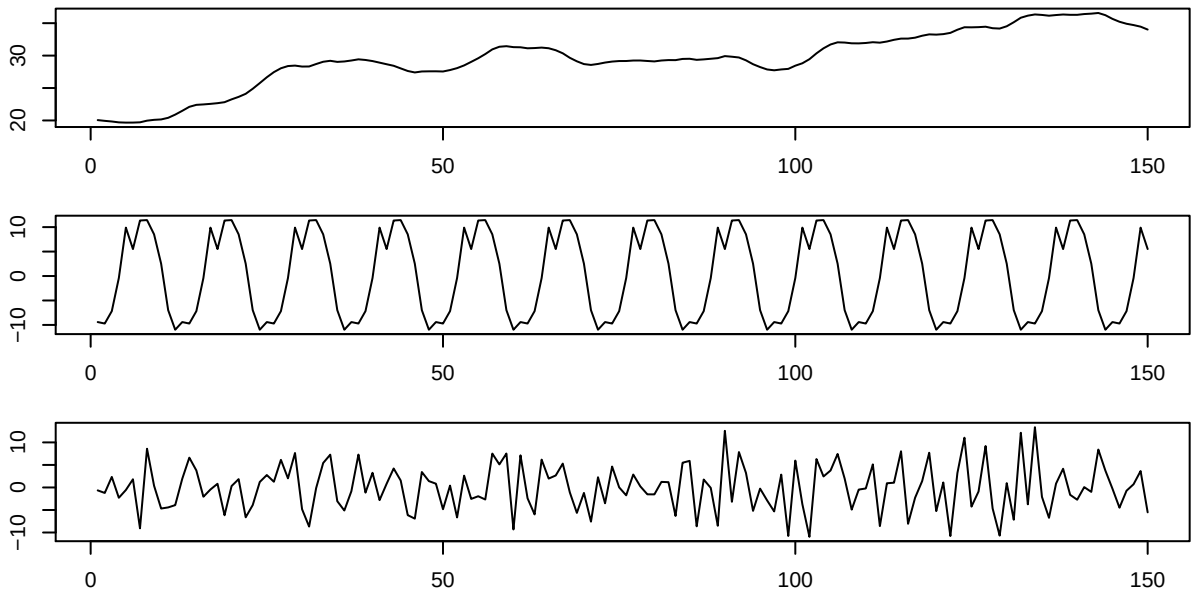

theta[1]
25
20
15
400
600
800
chain
1
2
3
4


sigma
6.5
6.0
5.5
5.0
4.5
400
600
800
chain
1
2
3
4

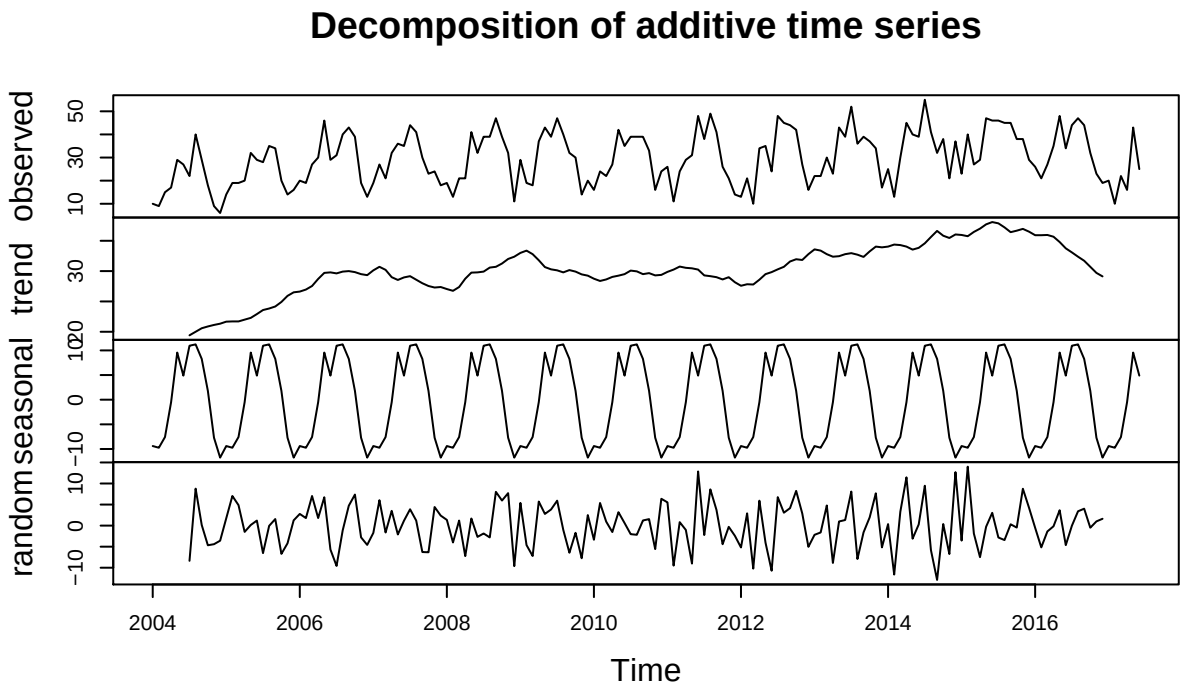


where the last 12 θs are simulated and will be used in forecasting.

The above blue line is my trend component. After second fitting, my decomposition will look like the following.

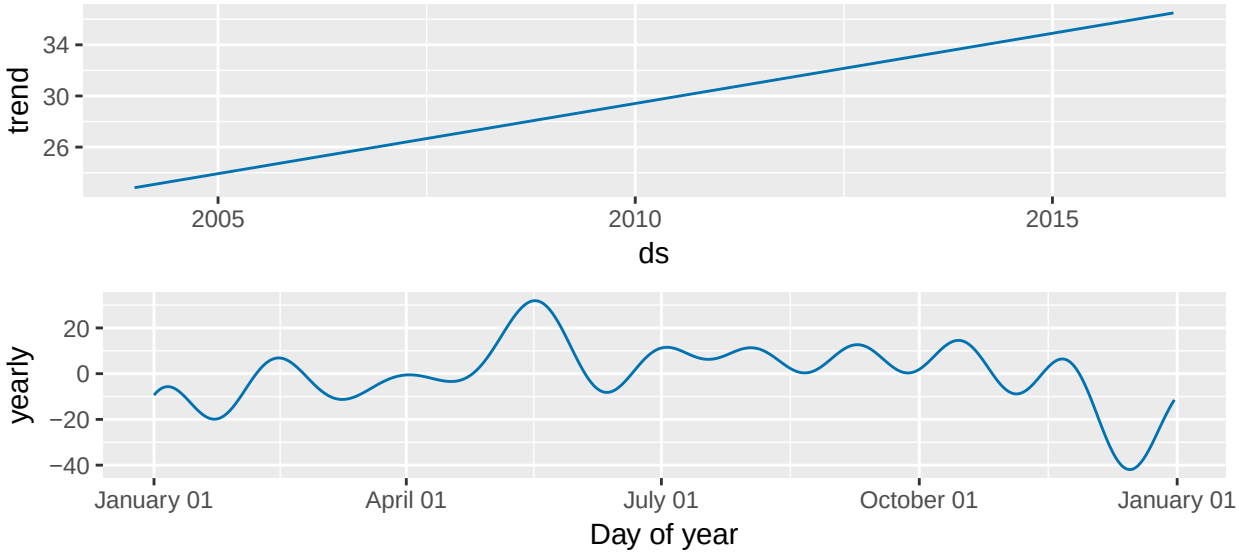


As a comparison, decomposition based on ARIMA looks like the following. Generally speaking, the seasonal part seems to be exactly the same, but my trend line is more smoothing.



And decomposition based on `Prophet` looks like the following.

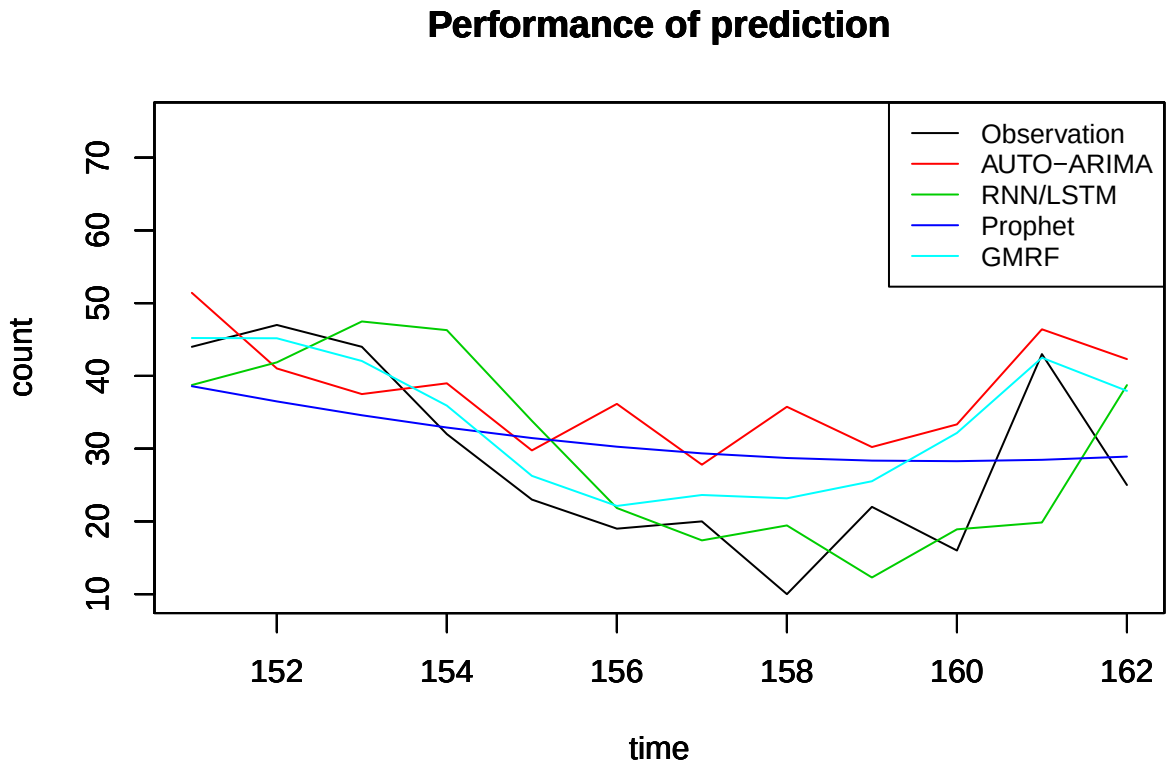
```
## Initial log joint probability = -6.35307
## Optimization terminated normally:
##   Convergence detected: absolute parameter change was below tolerance
```



Then, I will go further and compare them with regard to forecasting the last 12 points. Seasonal part seems to be straightforward to use in forecasting. For trend part, I use posterior median of *theta*. And then I sum up the seasonal component with a linear extension of trend component as my prediction.

The results of the above forecasting strategy along with other model's prediction, like ARIMA and RNN/LSTM are shown as following. I also show the result of `Prophet`.

Method	MSE
AUTO-ARIMS	160.4
RNN/LSTM	109.9
Prophet	106.9
GMRF	56.0



As is shown above, GMRF obviously outperforms the other methods. RNN/LSTM and `Prophet` make similar performances. They are well-used models in parctice, for example, `Prophet` is used in many applications across Facebook for producing reliable forecasts for planning and goal setting. The main reason that they are not competitive here could be the amount of data is really small, or that, the issue of overfitting.

Discussion

Conclusion

As is mentioned above, GMRF is a powerful nonparametric regression method which can also be applied into time series. It has a good chance to catch various characteristic features of interests, like autocorrelation structure and periodic component.

In fact, priors have a shrinkage effect, which can also be seen as a kind of regularization. when it comes to MLE, ridge regression is equivalent to Gaussian prior, and Lasso regression is equivalent to Laplace prior. Under Bayesian framework, we are more flexible to choose from different kinds of priors.

When it comes to model structure, GMRF is a realization of partial-pooling model. The connections between θ_i is limited by distance, which balances local adaptation and global control. A partial-pooling model is always a good idea, bacause it is more flexible than a fully-pooling model and more controllable than a non-pooling model. It is a tradeoff of bias and variance.

Limitation

The main issue here is computational efficience. Bayesian method is time-consuming. In simulation study, for example, it takes totally about an hour to sample. And Horseshoe prior has a more extreme problem of it than Gaussian prior and Laplace prior. Moreover, if we try funtions like $y = sin(100x)$, divergence issues raise up.

The idea is to never let a Bayesian model to deal with a too complicated function fitting. Alternatively, we could firstly remove some component to make the model simpler. Another idea is to leave a mangeable amount of data for Bayesian model and fit a freqentist model with the rest data. And then include the results of frequentist model as priors. Besides, A general idea of online learning and ensemble learning is also appealing to me.

Reference

[1]Faulkner, James R.; Minin, Vladimir N. Locally Adaptive Smoothing with Markov Random Fields and Shrinkage Priors. Bayesian Anal. 13 (2018), no. 1, 225–252. doi:10.1214/17-BA1050. https://projecteuclid.org/euclid.ba/1487905413

[2]Vehtari, A., Gelman, A., and Gabry, J. (2017a). Practical Bayesian model evaluation using leaveone-out cross-validation and WAIC. Statistics and Computing. 27(5), 1413â “1432. doi:10.1007/s11222- 016-9696-4. (published version, arXiv preprint)

[3]Gelman, A. and D. B. Rubin (1992) Inference from iterative simulation using multiple sequences (with discussion). Statistical Science, 7:457-511