STAT 447C Project: Bora Guney and Damien Fung

2024-04-09

Introduction

Throughout the duration of this outgoing academic term, we have both been undergoing STAT 443: Time Series & Forecasting. Taking both STAT 443 and STAT 447C in tandem, we have grown a keen interesting in the utilization of Bayesian techniques within time series forecasting. Namely, this project aims to compare parameter estimates and prediction performance obtained using Bayesian techniques compared to those obtained using SARIMA and Box-Jenkins. To demonstrate this, we utilize the S&P 500's historic data from the previous 3 years. We develop our Bayesian model using MH MCMC to learn model parameters for the 3 year data set. Furthermore, we employ kernel mixtures to detect change-points for S&P 500's historic data from the previous 5 years; this encapsulates the effects of the COVID-19 pandemic where there exists a large sudden drop in the S&P 500 Index. Further, there are impacts of the war in Ukraine on the S&P 500 index. By completing this project, we aim to determine the better approach to predicting economic patterns despite their chaotic nature.

The main inspiration for our work in this project is the sentence put on the lecture slides by Prof. Alexandre Bouchard-Cote: "it is easier to describe how things change rather than how things are".

Literature Review

This project is heavily inspired by the STAT 447 and STAT 443 courses. The modelling and forecasting techniques used in the frequentist approach (Box-Jenkins SARIMA) are heavily inspired by the techniques and practices learned in STAT 443. As a result of this, we refer to the textbook used in STAT 443: C. Cha]ield and H. Xing 2019 for our techniques and theory. The code used to generate our results and models within the EDA and data preparation section are based off the lecture notes and various assignments provided in STAT 443.

The idea of using SARIMA for predicting the S&P 500 can be seen in Dun 2023. As such, we believe that an ARIMA model is an appropriate modelling approach that can be used as a baseline model that we may use as a comparator for our Bayesian model. This paper is relatively recent and contains data with a similar range as ours; they conclude that an $ARIMA(0,1,1)(0,0,0)[0]$ model is ideal for modelling the S&P500. Despite their promising results with a simple $ARIMA(0,1,1)(0,0,0)[0]$ model, some of their data preprocessing techniques are not outlined, and therefore it is hard to recreate their results accordingly. Using just the training data (details outlined in the upcoming section) and the same ARIMA model as Dun's paper, we obtain an AIC of 10693.23 compared to Dun's AIC of -34978.79. We think that it is possible to create a more flexible model using a Change Points Model utilizing kernel mixtures. A change point model should provide important implications on how to act in times of crisis and normality.

```
arima(x.ts.3.train, order=c(0,1,1))
```

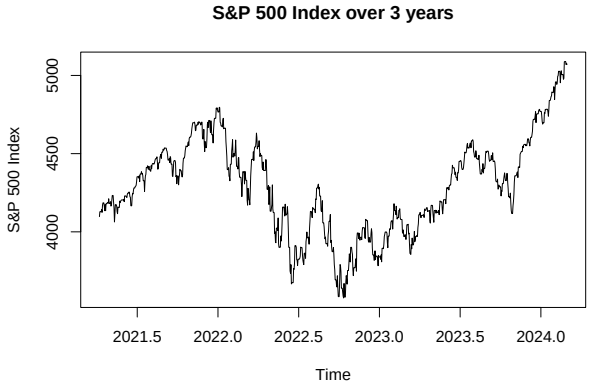
```
##
## Call:
## arima(x = x.ts.3.train, order = c(0, 1, 1))
```

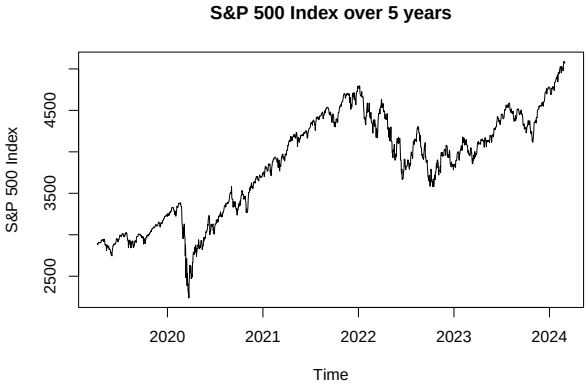
```
##
## Coefficients:
##           ma1
##       -0.0092
## s.e.   0.0326
##
## sigma^2 estimated as 1444:  log likelihood = -5344.62,  aic = 10693.23
```

On the other hand, Christoffersen 2012 uses an MCMC approach that encorporates "jumps" to model the S&P 500 index; similarly to our case with COVID-19, their study included a relatively big fluctuation in their time series data: the 2008 financial crisis. Their model focuses on dealing with cases where sudden changes to the time series are observed, albeit, with a different approach to ours. The demonstration that MCMC is capable of handling sudden significant changes such as long as the model is provided with additional "tools" to account for them. With the limited time and resources of a course project, we will simply use kernel mixtures to detect change-points as outlined in the introduction.

Preliminary EDA and Data Preparation

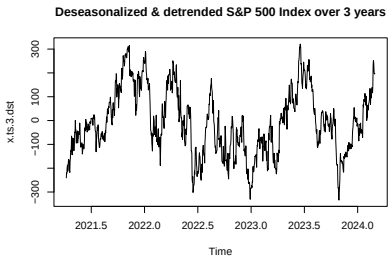
To determine model performance without data leakage, we reserve the data from Mar 2024 to Apr 2024 of the S&P 500 index data as an unseen test set. As such, we conduct our EDA and data preparation with the remaining data. This section also serves as a way to determine the appropriate SARIMA model using Box-Jenkins to give us our frequentist model comparator to our Bayesian model.

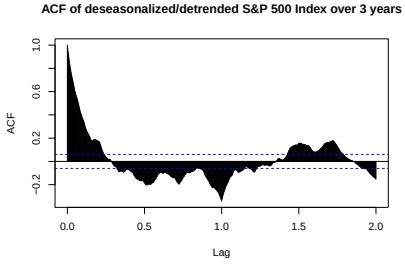


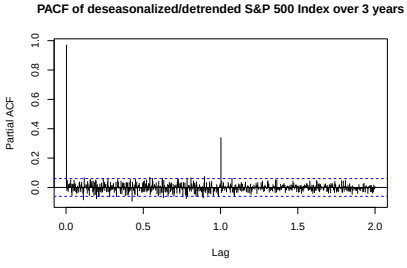


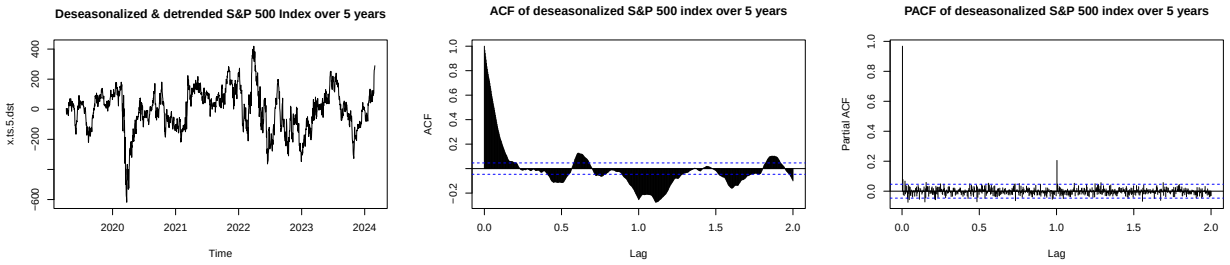
Detrending and deseasonalizing the data

The time series plots shown for both 3 year and 5 year data sets demonstrate that the data is not stationary (mean is not constant); namely some trend and seasonality may be observed. As a result of this, we remove trend and seasonal variation from both datasets.



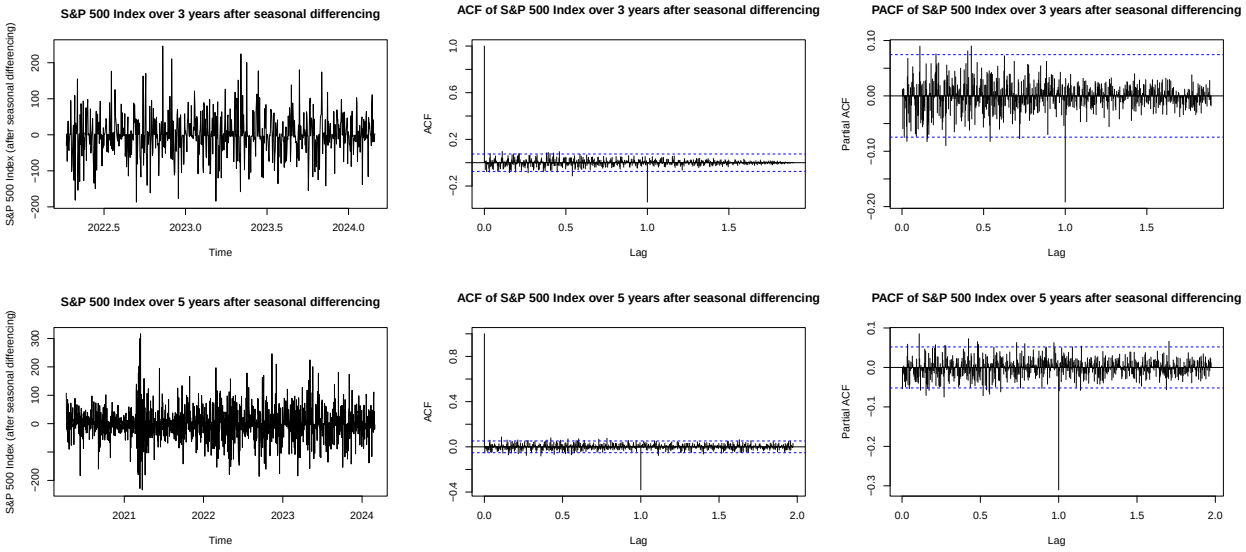






Applying differencing for SARIMA

Despite the deseasonalization and detrending of the data, the time series still exhibits some seasonality and trend. Therefore, we perform differencing at lag one and seasonal differencing at lag 365 (since we seem to have a seasonal period of 1 year. Note that the S&P 500 does not include data on days where the market is closed) in an attempt to obtain a stationary time series.



After the application of differencing on both time series, they may be considered stationary.

SARIMA Model

Using the EDA and prepared data from the previous section, we can determine the appropriate $SARIMA(p,d,q)(P,D,Q)_s$ model accordingly. From the application of differencing, we have determined that $d=1, D=1, s=365$ would be appropriate; but due to the limitations of the `stats` package in `R`, we are unable to use a period of $s=365$ for our model as its max period is $s=350$; instead, we adopt $s=182$ to represent biannual fluctuations instead. By the sample ACFs and PACFs, lag values before the 1.0 lag marker are above the significance line while lag values after are consistently under. This implies that $P=0$ and $Q=1$. To determine the ideal p and q values given that the remaining parameters in the $SARIMA$ model are set as described, we find the values of p,q that yield the lowest AIC using a simple grid search.

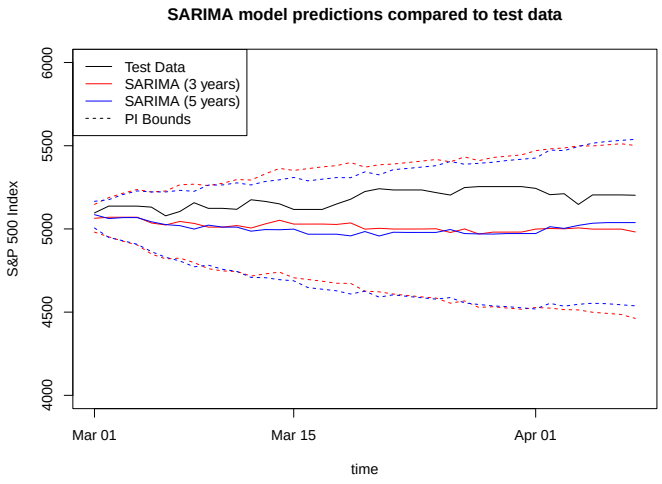
As a result of the heavy computations needed to generate each SARIMA model, we omit the results from this knit pdf. However, we conclude that $p=0, q=0$ are the most appropriate values for the SARIMA model as the decreases in the model's AIC are minimal (in the order of magnitude of 10^{-1} unit decrements). Therefore, we choose the simplest model through parsimony; we display the SARIMA models for both datasets in the below code cells.

```
sarima.model.3 <- arima(x.ts.3.train, order=c(0,1,0), seasonal=list(order=c(0,1,1), period = 182),inclu
sarima.model.3
```

```
##
## Call:
## arima(x = x.ts.3.train, order = c(0, 1, 0), seasonal = list(order = c(0, 1,
##     1), period = 182), include.mean = FALSE)
##
## Coefficients:
##          sma1
##       -0.9976
## s.e.   0.0711
##
## sigma^2 estimated as 1509:  log likelihood = -4606.22,  aic = 9216.44
```

```
sarima.model.5 <- arima(x.ts.5.train, order=c(0,1,0), seasonal=list(order=c(0,1,1), period = 182),inclu
sarima.model.5
```

```
##
## Call:
## arima(x = x.ts.5.train, order = c(0, 1, 0), seasonal = list(order = c(0, 1,
##     1), period = 182), include.mean = FALSE)
##
## Coefficients:
##          sma1
##       -1.0000
## s.e.   0.0554
##
## sigma^2 estimated as 1506:  log likelihood = -8361.1,  aic = 16726.19
```



The above plot shows the SARIMA model's predicted values compared to the test data. We can see that the model seems to underestimate the S&P 500 Index instead of overestimate. However, the prediction intervals seem to still capture the test data with 95% confidence. Therefore, the SARIMA model is reasonably appropriate for modelling the S&P 500 for both the 3 year dataset and the 5 year dataset.

To compare the SARIMA model's performance with the bayesian approach our project demonstrates, we compute the mean squared prediction error (MSPE) of both SARIMA models to be 32688.32 for the model

trained on the 3 year dataset, and 36738.45 for the model trained on the 5 year dataset. We will compute the MSPE for the MCMC model later in this report.

Model Development

MCMC Model

We start by constructing a Bayesian model for the three year data, then we will make the model more complex. ACF and PACF analysis of the data showed that an $AR(1)$ process is suitable for the data too. Therefore, we incorporate this knowledge into our Bayesian model. Let i be the time index and y be the deseasonalized time series.

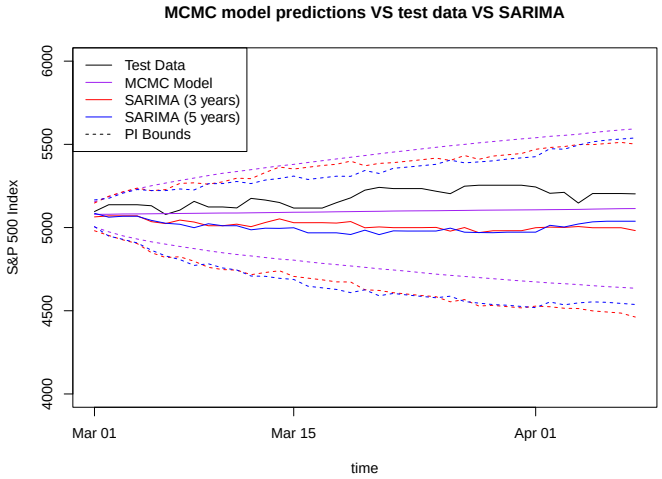
$$\sigma \sim Exp\left(\frac{1}{100}\right)$$

$$\alpha \sim N(0, 10)$$
$$y_i \sim N\left(\alpha \cdot y_{i-1}, \sigma\right)$$

```
data {
int N; // number of samples
real initial;
real final;
vector [N] y; // SP 500 index value
}

parameters {
  real <lower = 0> sigma;
  real alpha;
}

model {
  sigma ~ exponential(1.0/100);
  alpha ~ normal(0, 10);
  for (i in 1:N){
    if (i == 1){
      y[i] ~ normal(alpha*initial, sigma);
    } else{
      y[i] ~ normal(alpha*y[i-1], sigma);
}}}

generated quantities {
  vector [39] yhat;
  for (i in 1:39){
    if (i == 1){
      yhat[i] = normal_rng(alpha*final, sigma);
    } else{
      yhat[i] =  normal_rng(alpha*yhat[i-1], sigma);
}}}
```

We use the data from April 2021 to Feburary 2024 for training. Further, the data from March 2024 to April 2024 will be used for testing.

We plot the predictions, actual values, and 95% confidence intervals as follows :



This plot suggests that the simple MCMC model with 3 years of data matches the performance of the SARIMA models. It achieves $MSPE = 8591.045$ which is significantly better than the MSPE for the SARIMA models. However, it it can be observed that the prediction interval seems to be larger. The error of the model is not improved compared to the frequentist approach.

Change point model implementation

Now, we will use the five year data, and will try to find the change points within the data. From the time series plot of the five year data, it seems that there are multiple change point candidates, one of them being around COVID times. We assume that there are four change points, and therefore five α values corresponding the divided regions. Further, we assume that each region has the same standard deviation for simplicity. To find the change point we use the following alternation of Kernels model :

$$K = K_1 \circ K_2 \circ K_3$$

, where K_1 only modifies σ, K_2 only modifies the α's, and K_3 only modifies the change points. We know that if K_1, K_2, and K_3 are π-invariant, then K is too. For this particular target distribution, we can say that K_1 is irreducible if the proposal for K_1 can reach all of R_+ from any starting point. Further, K_2 is irreducible if the the proposal for K_2 can reach all of R_+^5 from any starting point. Lastly, K_3 is irreducible if the proposal for K_3 can reach all of

$$\{1, 2, ..., 447\} \times \{448, 315, ..., 894\} \times \{895, 628, ..., 1341\} \times \{1342, 941, ..., 1788\}$$

from any starting point. For the sigma, we use a normal distribution centered at the current σ as the proposal distribution:

$$\sigma' \sim N(\sigma, 5)$$

For α's we will use a multivariate normal distribution centered at the current α's as the proposal distribution :

$$\alpha' \sim N(\alpha, 0.001)$$

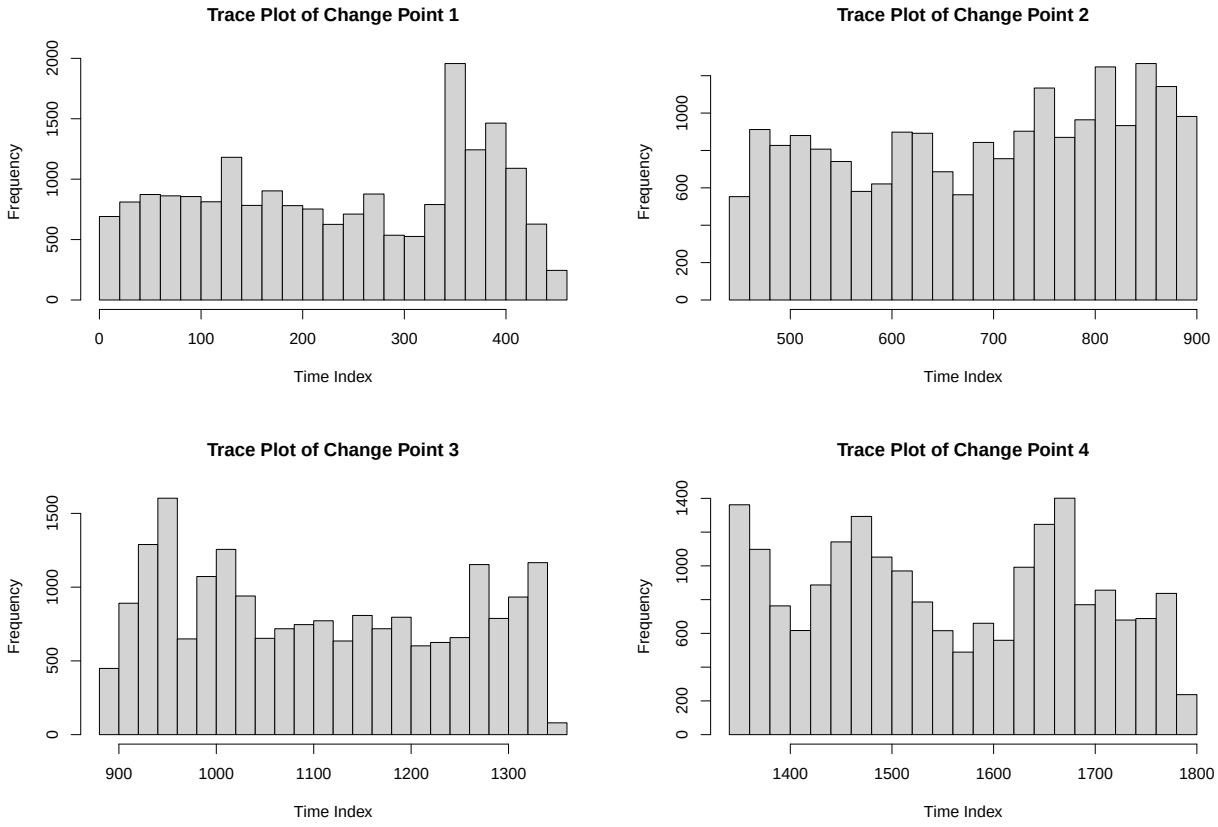
Lastly, for each change point, we use a discrete uniform distribution as the proposal

$c_1' \sim Unif\{1, 447\}$, $,c_2' \sim Unif\{447, 894\}$, $c_3' \sim Unif\{895, 1341\}$, $c_4' \sim Unif\{1342, 1788\}$. By our choice of proposals, the MCMC can reach any point in $\{1, 2, ..., 447\} \times \{448, 315, ..., 894\} \times \{895, 628, ..., 1341\} \times \{1342, 941, ..., 1788\}$

Let ρ be the parameter we used to determine which kernel to use, we define it as :

$$\rho \sim Unif\{1, 3\}$$

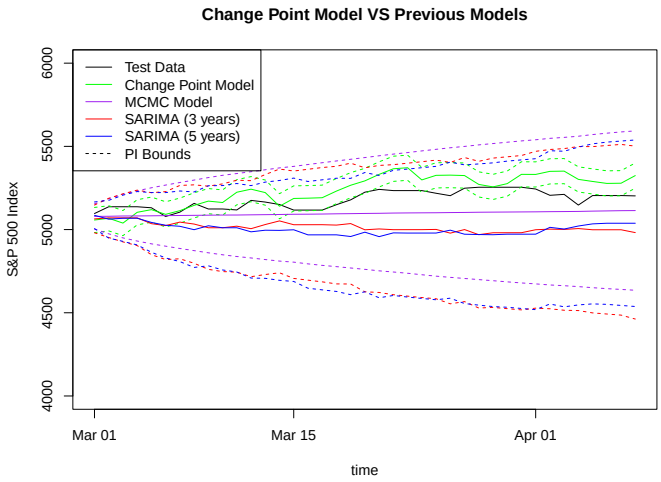
if $\rho = 1$, we use K_1, if $\rho = 2$ we use K_2 and if $\rho = 3$, we use K_3. Let $\alpha = (\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5)$ and $c = (c_1, c_2, c_3, c_4)$. In each K_1 we define the ratio as $\frac{\gamma(\sigma', \alpha, c)}{\gamma(\sigma, \alpha, c)}$, in K_2 as : $\frac{\gamma(\sigma, \alpha', c)}{\gamma(\sigma, \alpha, c)}$, in K_3 as: $\frac{\gamma(\sigma, \alpha, c')}{\gamma(\sigma, \alpha, c)}$ Then let $R^{(m)}$ be the ratio, we sample $A^{(m)} \sim Bern(\min\{1, R^{(m)}\})$ and then accept the proposal if $A = 1$, and set $X^{(m)} = \tilde{X}^{(m)}$. If $A^{(m)} = 0$, we stay at the current point $X^{(m)} = X^{(m-1)}$.



The spike in the plot for the first change point indicates that the change change point is around March 2023, when COVID restriction started, which is aligned with our expectations. The second change point seems to be between 800th and 900th time indices, which corresponds to around August 2021, . Further, the third change point seem to be around 950th or 1350th time index, which is also aligned with the plot. There seems to be three potential candidates for the fourth change point, the most notable one being around $1600th$ and 1700th time indices, which corresponds to around October 2023. Therefore, the fourth change point is also aligned with the plot of the five year time series.

It seems that the standard deviation 38.56 is less than the Bayesian standard deviation estimate for the three year data 45.46. Further, by using a change point model using kernel mixtures, we had more flexibility in terms of having Bayesian estimates of the α values for each region determined by the change points. The highest α value was chosen for the fifth region, which is the one that is on the most right hand side. The first and the fourth regions has the smallest α values.

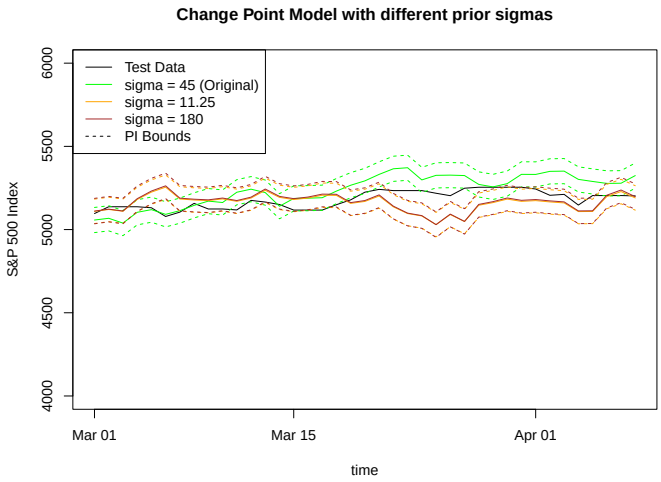
We make predictions with the change point model as follows. We use the α value corresponding to the rightmost region.



From the results shown in the above, the prediction interval of the change point model does not always capture the true values in the test data, however the model is able to obtain $MSPE = 6859.014$ compared to the previous MCMC model without the change point kernel which obtained $MSPE = 8591.045$. The difference in $MSPE$ within the MCMC models is much smaller compared to the difference in $MSPE$ between the MCMC models and the SARIMA models. This implies that the change point model is capable of obtaining good predictions. Notably, the confidence intervals for the Change Point Model are not large; this is because we did not make posterior predictions. Instead, we used the posterior estimates of the parameters to make predictions, which was done to save computation time. The main benefit of the change point kernel is its flexibility and ability to deal with sudden changes in the time series; it can be assumed that this model performs better in cases such as COVID-19 or the 2008 economic crisis compared to the simple MCMC model.

Sensitivity analysis for change point model

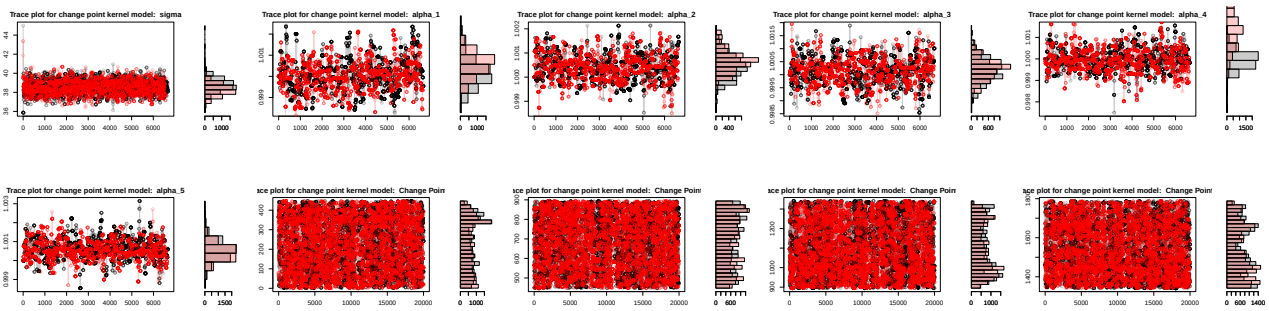
To demonstrate our that our prior choice was appropriate, we choose different sets of priors for and compute their posteriors. Namely, we change the value of the σ prior to a quarter of its original value $(45/4 = 11.25)$ in addition to four times its original value $(45 \cdot 4 = 180)$. We keep the remaining prior choices constant, in order to properly evaluate the sensitivity of the σ parameter. In the interest of computational power and complexity, we do not perform a grid search or further exploration on the remaining prior choices as each model requires several minutes to run. This sensitivity analysis simply demonstrates the importance of the *sigma* prior



The MSPE of the models with different *sigma* priors (11.25 and 180) are relatively close to that of the original prior choice of $\sigma = 45$. $\sigma = 11.25$ obtained $MSPE = 7293.212$ and $\sigma = 180$ obtained $MSPE = 7334.161$. Although both are close to that of $\sigma = 45$ at $MSPE = 6859.014$, we can observe on the graph that the predictions from $\sigma = 11.25$ are quite rigid, and do not account for the variation in the time series quite as well; this is naturally expected as the chosen standard deviation is lower. Conversely, $\sigma = 180$ does not have a prediction interval that is significantly larger than that of $\sigma = 45$, but, when the point estimates in $\sigma = 180$ deviate away from the true values of the data, they seems to deviate much more than the predictions seen in $\sigma = 45$. This suggests that the prediction error for values further in-the-future would be significantly higher. Overall, the model's point estimate predictions are not very sensitive to prior changes as demonstrated here: only the prediction intervals are significantly sensitive as expected by definition.

Model Diagnostics

We would like to ensure that the MCMC model is fast mixing. Therefore, we present trace plots obtained by using different chains. To create the second chain, we utilize a different random seed.



The MCMC model seems to be mixing well as the trace plots obtained from different chains are hard to distinguish and the histograms are approximately uniform.

Conclusion

In conclusion, both the Frequentist and the Bayesian approaches we have taken to modelling and forecasting the S&P 500 Index achieve a similar level of performance. However, when Kernel Mixtures are implemented

into the MCMC algorithm, we are more likely to be able to handle sudden changes and fluctuations in the time series; in other words, the model which incorporates Kernel Mixtures is significantly more robust despite the fact that it obtains a slightly lower MSPE compared to the simple MCMC model. Since it is easier to describe how things change rather than how things are, the change point model we have developed should help with future crisis in terms how much the index will deviate from its performance in normality. To make the Change Point Model even better, we could have incorporated the α values corresponding to other regions as well. In particular, while making predictions, we could have assigned a probability α values corresponding to each region. For example, smaller probabilities for times of crisis etc. Then while making predictions, we could have used the probability distribution to change the α value as time passes. This could result in a reasonably dynamic model for long term predictions. Yet, since we made predictions just for about a month, we didn't use the mentioned methodology to make predictions. Instead, we used the α value corresponding to the rightmost region.

Limitations

Our project is largely limited to the short amount of time that we are able to spend developing these models and lack of available computation resources at our disposal. With enough hyperparameter tuning, we expect that our models would be capable of performing much better and providing more accurate results. Further, incorporating further tools for the model to use would be useful, such as recent news or events that can significantly impact values including, but not limited to FED interest rates, new pandemics, unforeseen global events, and more. By incorporating these tools and information, the model could be more dynamic and flexible to sudden changes.

Another limitation of the model is the utilization of several statistical assumptions with regard to the time series. We made the strong assumption that the time series was stationary after data preparation; however all models may benefit significantly more if we had the techniques and tools that could further assist in handling non-stationary time series. This would require additional knowledge in time series forecasting and modelling techniques.

Further, while making predictions for the Change Point Model we used posterior estimates of the parameters to make predictions instead of directly having posterior predictions. The limitation that behind this was the computation cost of having posterior predictions would be much larger given the complexity of the Change Point Model. The consequence of our choice to avoid high computation cost was narrow confidence intervals for the predictions of the Change Point Model.

Finally, our data is also limited to just 3 and 5 years of the S&P 500 Index. This is primarily because we wanted to observe the difference in model performance when the impact of the COVID-19 pandemic is incorporated into the data. Under the assumption that a minimal number of events impact the S&P 500 Index, it would be interesting to see how much better the model could perform with additional post-pandemic data and crisis events exclusively.

Contributions

In this project, we split the workload into relatively even parts; Bora took charge of implementing the two MCMCs that were demonstrated, while Damien completed the data preparation, SARIMA models. We collaborated on main bodies of writing. The data analysis across the project was also distributed evenly; both of us creating visualizations for the project in general, and providing insight to the statistics behind the analyses using the tools acquired throughout STAT 447.

References

Chatfield, Christopher, and Haipeng Xing. The Analysis of Time Series: An Introduction. CRC Press, 2019.

Christoffersen, Peter, Kris Jacobs, and Chayawat Ornthanalai. "Dynamic Jump Intensities and Risk Premiums: Evidence from S&P500 Returns and Options." Journal of Financial Economics, vol. 106, no. 3, 2012, pp. 447-472.

Dun, Vadym. "Analysis and Forecasting the Price of the S&P 500 Index using the Arima Model." Financial Markets, Institutions and Risks, vol. 7, no. 4, 2023, pp. 113-134.