BAYESIAN PREDICTIVE ANALYSIS OF S&P 500 INDEX

SHUYI TAN, ZHIPENG ZHU

Abstract. Forecasting the trend of the stock market has been a popular topic for researchers from various quantitative fields for years, with the aim to deliver informative messages to investors and invoke economic decisions. The accuracy of the prediction is the most critical part that matters to stock dealers. In this project, two methods: Bayesian structural time series models and Gaussian process will be utilized to forecast the stock market trend based on the S&P 500 index, after which the results show that both models demonstrate fairly good performance in the short-term forecasting, while the accuracy sharply declines in the long-term forecasting.

1. **Introduction**

The S&P 500 (Standard and Poor's 500) is a free-float, capitalization-weighted index of the top 500 publicly listed stocks in the US (top 500 by market cap), which is widely regarded as the best single benchmark of large-cap U.S. equities and therefore its performance is tracked by many funds.[1] The value of the index is calculated by summing the adjusted market caps of every company and dividing the result by a divider that is unreleased to the public as proprietary information to S&P. Since the S&P 500 index provides investors with a straightforward but informative proxy of the market trend, the analysis and prediction of which have been earning great interest within and outside of the financial and economics domains. For example, if the estimate of S&P indicates a pessimistic 12-month level or return, then cutting interest rates or lowering banking reserves will be considered.[2]

Multiple studies have developed methods for predicting the S&P 500 index from different perspectives, in which time series models and Bayesian methods are widely applied.[2] In this project, the main objective is to input the historical data of S&P 500 data into a Bayesian structural time series model as well as the Gaussian Process to estimate the S&P 500 index, with the motivation that the predicted results may help investors to choose the right stocks and timing to invest based on forecasting for the future. As a side activity, this project also compares the performances of those two methods.

The rest of the paper is organized as follows. Section 2 introduces the dataset used in this project. In Section 3, we apply the Bayesian structural time series model. In Section 4, the Gaussian process (GP) is explored. Section 5 summarizes and analyzes the obtained results. Section 6 presents the conclusion and discusses future work.

2. **Dataset Overview**

The data used in this project all come from the official website of Wall Street Journal (https://www.wsj.com). We collected 1581 daily S&P 500 (SPX) Index close price samples from the WSJ Markets over the January 2015 to April 2021 period, which forms the entire dataset. Since the stock market does not operate during weekends and holidays, data are

Date: April 01, 2021.

only available for weekdays with public holidays excluded. The S&P 500 index will be normalized with the starting price of the years. The normalization works as follows:

$$y_{new,ij} = ((y_{ij} - y_{i.})/\sigma_{i.}) - y_{new,i0} \tag{1}$$

where i is the year, j is the specific index of the trading day.

From Figure 1 below, plot[a] demonstrates the overall trend of the S&P 500 index over time, and plot[b] in the same figure displays the normalized S&P 500 index for each year respectively. Overall, it can be observed that the time series structure demonstrates an obvious trend as well as strong seasonal patterns. It is shown in the plots that there is obvious decline in the third quarter of 2015, which is because the Chinese market crash spilled over the global market in August 2015.[3] In the following modeling sections, data from 2015 to 2020 will be used as the training set, with the rest from 2021 will be used as the test set.

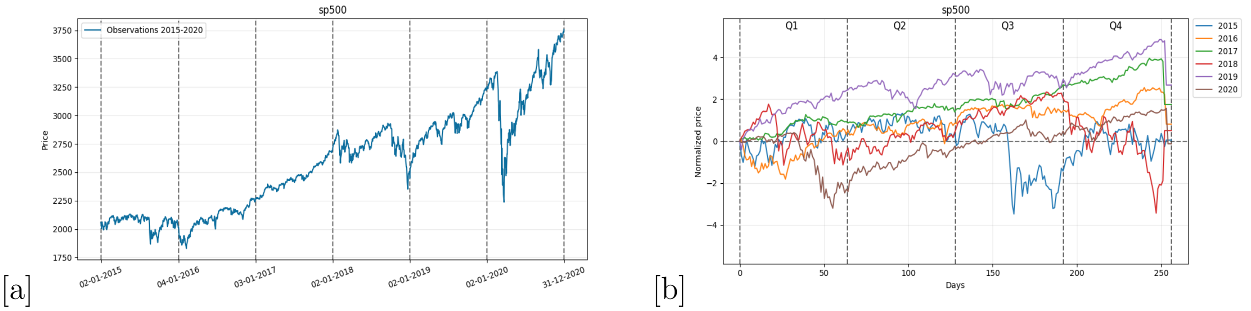


FIGURE 1. Left: Overall Trend of S&P 500 ; Right: Normalized S&P 500 Trend of Each Year

3. Bayesian Structural Time Series Model

Time series models have been widely used for the purpose of forecasting and explaining model structures in many files, in which the structural time series model is a powerful statistical tool that constructs the forecasting by structuring elements of time series data. Specifically, in a structural time series model, the observed equation of the observed data y_t is defined to be

$$y_t = Z_t^T \alpha_t + \varepsilon_t \tag{2}$$

where Z_t^T is a vector of model parameters, α_t is a vector of latent variables, and ε_t is an error term that follows Gaussian distribution with $\mu = 0$ and $\sigma^2 = H_t$. To further illustrate the latent variable, α_t is represented by the following transition equation[4]

$$\alpha_{t+1} = T_{t+1}\alpha_t + R_t\eta_t \tag{3}$$

where η_t follows Gaussian distribution with $\mu = 0$ and $\sigma^2 = Q_t$. The equation describes how the latent variable α_t evolves through time. T_t and R_t are the transition matrix and structural parameters respectively. R_t, T_t, and Z_t containing 0 and 1 indicates the relevance for the structural computation. The observation equation (2) and transition equation (3) together describe the state space of observed data. At this point, a time series model for forecasting can be constructed. Putting it into a Bayesian context, let us recall that the Bayesian statistics is computed as

$$\begin{aligned} P(\theta|x) &= \frac{P(x|\theta)P(\theta)}{\sum_\theta P(x,\theta)} \\ &= \frac{P(x|\theta)P(\theta)}{\sum_\theta P(x|\theta)P(\theta)} \\ &= \frac{P(x|\theta)P(\theta)}{P(x)} \end{aligned} \tag{4}$$

where x is the observed value and θ is the model parameter. In equation (4), $P(\theta)$ is the prior, $P(x|\theta)$ is the likelihood function, and $P(\theta|x)$ is the posterior, which will be updated by learning the observed data x given θ. In Bayesian analysis, different results will be obtained based on various selections of prior distributions. Based on the characteristic of our data, we will select a Gaussian distribution as the prior distribution of the BSTS model because the occured frequency values in $[0,\infty]$ is used. Basically, the BSTS model can be expressed with three state components: trend μ_t, seasonality τ_t and a regression component $\beta^T x_t$ as follows [5]

$$y_t = \mu_t + \tau_t + \beta^T \mathbf{x}_t + \varepsilon_t \tag{5}$$

where

$$\begin{aligned} \mu_{t+1} &= \mu_t + \delta_t + \eta_{0t} \\ \delta_{t+1} &= \delta_t + \eta_{1t} \\ \tau_t &= -\sum_{S=S}^{S-1} \tau_t + \eta_{2t} \text{ and } \tau_t \backsim Gaussian \end{aligned} \tag{6}$$

In our case, two models will be fitted to the data, one (M2) is of the form as equation (5), and the other (M1) is without the regression component. Cboe Volatility Index (VIX) is treated as the regression component in model M1, which is a real-time index that reflects the market's expectations for the relative strength of near-term price changes of the S&P 500 index.[11] It is not necessary to perform variable selection for our models because there is only one regressor in the regression component in our model. Results are presented in Section 5.

4. **Gaussian Process**

Gaussian process (GP) is a popular Bayesian non-parametric model for time series data, for which various kernel functions have been developed for different modeling tasks.[6] Zhu summarized the realization of Bayesian non-parametric modeling with GP: a GP is the prior for the infinite set of random variables indexed by x, and another GP is the posterior upon observing some finite subset of the random variables.[7] To demonstrate the GP, we write:

$$f(x) \sim GP(m(\cdot), k(\cdot,\cdot)) \tag{7}$$

where x is some process $f(x)$, m is the mean, and $k(\cdot,\cdot)$ is the covariance. Specifically,

$$\begin{aligned} m(x) &= E[f(x)] \\ k(x_1,x_2) &= E([f(x_1) - m(x_1)][f(x_2) - m(x_2)]) \end{aligned} \tag{8}$$

In the regression setting, a training set $D = (x_i, y_i i = 1, 2, \ldots, N)$ is constructed where x_i, $y_i \in \mathbb{R}$. In most cases, the observed output is assumed to be noisy. However, since our response is the closing value of the S&P 500 index, with which the true prices of stocks are evaluated, the S&P 500 index we use will be assumed to be free of noise.[8] Namely, a key assumption we made is that $y_X = f(X), \forall X$, and $f \sim GP(0, k)$ for the covariance function k, which is the prior. Mathematically, the prior distribution of observed response (y) is given by

$$y \sim \mathcal{N}(0, K(X, X))$$

where $K(X, X)$ is the covariance matrix between pairs of the training points. In addition, the joint prior distribution of the training response (y) and the prediction of test response $(x_*)$is that

$$\left[\begin{array}{c} f(x) \\ f\left(x_*\right) \end{array}\right] \sim N\left(0,\left[\begin{array}{cc} K(X, X) & K\left(X, X_*\right) \\ K\left(X_*, X\right) & K\left(X_*, X_*\right) \end{array}\right]\right) \tag{9}$$

For simplicity, $f(x)$ and $f(x_*)$ will be denoted as f and f_*. Since the likelihood is conditionally independent of f_* given f, and $y|f \sim \mathcal{N}(\sigma_n^2 I_n)$ where I_n is a $n \times n$ identity matrix, we have $P(y|f, f_*) = P(y|f)$. Applying Bayes' Theorem, the joint posterior distribution given the training data can be calculated by

$$\begin{aligned} P\left(f, f_* \mid y\right) &= \frac{P\left(y \mid f, f_*\right) P\left(f, f_*\right)}{P(y)} \\ &= \frac{P(y \mid f) P\left(f, f_*\right)}{P(y)} \end{aligned} \tag{10}$$

Then the predictive distribution will be given by

$$P(y_*|y) = \int P(f, f_*|y) df$$

Knowing that these distributions and their marginals are normal, we can also obtain the mean of covariance of the predictive distribution:[9]

$$\begin{aligned} E\left[f_* \mid y\right] &= K\left(X_*, X\right)\left(K(X, X)+\sigma_n^2 I_n\right) y \\ \operatorname{Cov}\left[f_* \mid y\right] &= K\left(X_*, X\right)-K\left(X_*, X\right)\left(K(X, X)+\sigma_n^2 I_n\right) K\left(X, X_*\right) \end{aligned} \tag{11}$$

In our case, each time series will be treated as an independent predictor in the regression model.[10] Therefore, the historical data can provide knowledge to the prediction of a new series. Let i index the year M, the series is a sequence of the S&P 500 indexes during the period where it is traded. Excluding weekends and holidays for a stock market year, the length of M will be around 256. The lengths of all series do not have to be strictly identical if we assume the time periods spanned by the series follow an annual circle.

The forecasting problem is defined as given observations from a complete series $i = 1, 2, \ldots, N-1$ and optionally from a partial last series N, $\{y_\tau^N\}$ where $\tau = 1, 2, \ldots, M_N$. The last series will be extrapolated until it reaches a predetermined endpoint so that it characterize the joint distribution of $\{y_\tau^N\}$ where $\tau = M_{N+1}, \ldots, M_{N+H}$ for some H, and M_{N+H} is the last day of trading in H. Hereby We wish to find a proper representation of

$$P\left(\left\{y_\tau^N\right\}_{\tau=M_{N+1}, \ldots, M_{N+H}} \mid \left\{x_t^i, y_t^i\right\}_{i=1, \ldots, M_i}^{i=1, \ldots, N}\right) \tag{12}$$

given $\{x; i\}$ for each series, which is spanned over the forecasting horizon with i ranging over the available series, and t ranging over observations within a series. In other words, the goal is to predict the rest of training days in N, $\{y_\tau^N\}$.[10] As mentioned above, each time series is treated as an independent predictor in the regression model. Therefore, a representation of the trading dates is developed as independent predictors. The results are presented in Section 5.

5. Model Evaluation

5.1. **BSTS Models.** The first model (M1) fitted by BSTS will only include the components of trend(μ_t) and seasonality (τ_t) and leave the Regression as NULL. The second model (M2) will use the VIX index as the Regression component. Figure 2 displays the trend and seasonality components for both models.

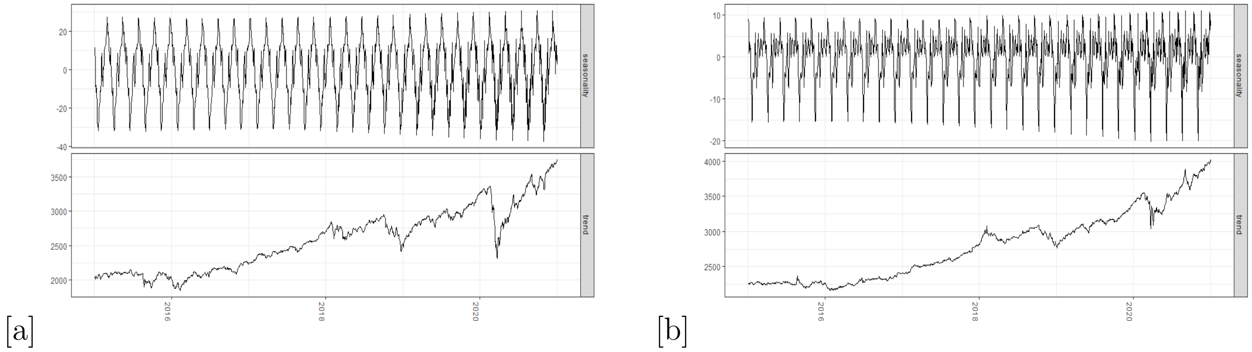


Figure 2. Trend and Seasonality of Model 1; Right: Trend and Seasonality of Model 2

In both models, Trend and Seasonality have apparent effects, while the volatility of Trend in M2 seems to be minor than that in M1. Therefore, it suggests that the Regression component in M2 may have substantial effects on the S&P 500 index. Predictions are made for the period of the first four months of 2021. The modeling results are summarized in Figure 3.

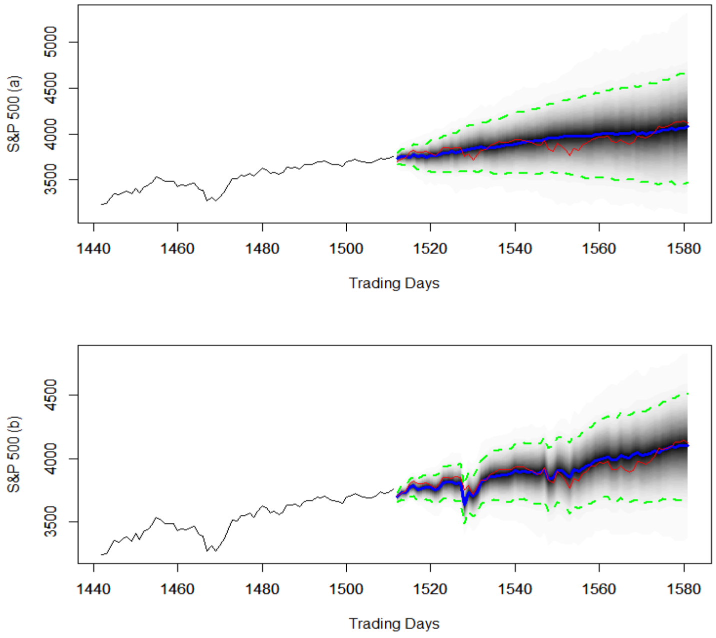


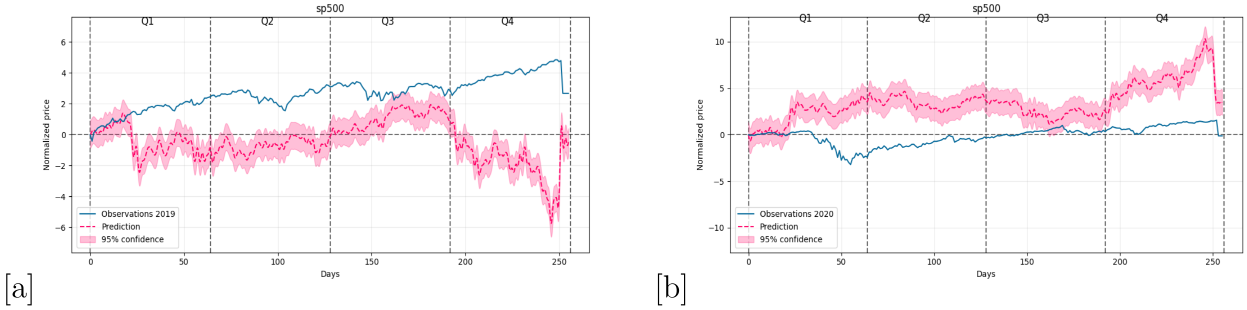
Figure 3. Predictions of S&P 500 in 2021 Q1

Compared to M1, the confidence interval of the predicted values are slimmer in M2, and M2 also catches the minor volatility of several spots as shown in Figure 3. According to the MSE and R-squared scores in Table 1, M2 has better fit to the data than M1. This confirms the hypothesis that the VIX index may have significant improvement in predicting the S&P 500 index.

Model	MSE	R-Square
M1	4742.573	0.61
M2	2251.439	0.84

Table 1. Performance Comparison between M1 & M2

5.2. **GP Results.** With the Gaussian Process Regression model, three years, 2019, 2020, and 2021 will be set as the prediction targets. For each year, the training data will include the S&P 500 from 2015 to the previous year of the target year. For example, to predict the S&P 500 index in 2019, a set of data from 2015 to 2018 will be used as the training data in the GP model. Figure 4 presents the predictions for 2019 and 2020.



[a] [b]

Figure 4. Predictions of S&P 500 in 2019 (a) and 2020 (b)

Both predictions for 2019 and 2020 show fairly good fittings at the beginning of the year. After that, dramatic deviations from the true observations appear and such deviations become worse in the later quarters. The prediction of 2019 in Q4 even has a contradicted direction compared to the observed trend. Figure 5 displays the prediction for 2021, with partially observed data.

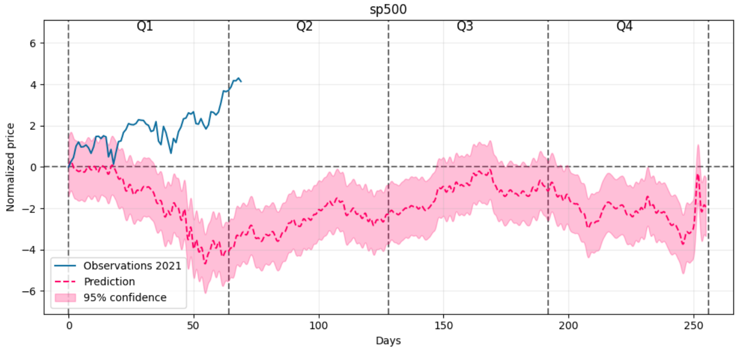


Figure 5. Predictions of S&P 500 in 2021

The predicted curve in Figure 5 seems to draft a lot based on the observed Q1 data, though it has some good values in the first 20 days. The prediction for the rest of the quarters in

2021 still needs verification in the future. Generally, the predictions of the GP model show good performance at the beginning of the years, while they tend to become inaccurate in the following quarters.

6. Conclusion and Future Work

This project undertakes the task of predicting the stock market trend based on the S&P 500 index. The BSTS model shows decent accuracy in predicting the S&P 500 in short term, and additional variables may improve the accuracy. Besides, the GP model has good accuracy in short-term prediction as well, while the long-term forecasting is also unstable. It reveals that either the BSTS or the GP models are insufficient to predict the S&P 500 index as an actual trading strategy. It is known that the S&P 500 is affected by various political and economic factors, and single events may result in vigorous fluctuations. For example, the coronavirus pandemic has been known to have a devastating effect on the stock market performance in 2020. [12] It is essential to incorporate as many as appropriate covariates as a supplement of market information to approach more robust and detailed results. In this case, Bayesian variable selection with Spike and Slab will be a useful tool especially for the structural time series model.

In addition, as we have noticed that the predictive accuracy of both models sharply degrades in long-term forecasting, which arises our consideration of the further possible methods to improve the long-term performance. Specifically, the BSTS model provides an option for long-term prediction, which adds more constraint on the uncertainty. It assumes a stationary distribution where the uncertainty will grow to a finite asymptote instead of an infinity. [4] For the Gaussian process regression, there are two possible approaches for long-term forecasting. The first is to construct a complex kernel from a few base kernels. The second is to make the time series a multi-dimensional input, and hence, an autoregressive correction can be used over the time series. [13]

References

[1] Kenton, Will *S&P 500 Index - Standard Poor's 500 Index.* Investopedia. March 23, 2021. Received from http: https://www.investopedia.com/terms/s/sp500.asp.

[2] Chan, Eric Glenn. *Forecasting the S&P 500 Index Using Time Series Analysis and Simulation Methods.* Massachusetts Institute of Technology. June 23, 2009.

[3] Clinch, Matt. *How the S&P 500 experts got it wrong in 2015.* January 1, 2016. CNBC. Received from https://www.cnbc.com/2016/01/01/how-the-sp-500-experts-got-it-wrong-in-2015.html

[4] Scott, Steven L. *Fitting Bayesian structural time series with the bsts R package.* The Unofficial Data Science Blog. July 11, 2017. Received from https://www.unofficialgoogledatascience.com/2017/07/fitting-bayesian-structural-time-series.html

[5] Steven L. Scott and Hal Varian. *Predicting the Present with Bayesian Structural Time Series.* June 28, 2013. International Journal of Mathematical Modelling and Numerical Optimisation. vol. 5. pp. 4-23.

[6] Cheng, L., Ramchandran, S., Vatanen, T. et al. *An additive Gaussian process regression model for interpretable non-parametric analysis of longitudinal data.* Nat Commun 10, 1798 (2019). Received from https://doi.org/10.1038/s41467-019-09785-8.

[7] Zhu, Xiaojin.*Bayesian Nonparametrics.* CS 731 Spring 2011 Artificial Intelligence. University of Wisconsin-Madison. Received from http://pages.cs.wisc.edu/ jerryzhu/cs731/bnp.pdf.

[8] Farrell, Todd and Correa, Andrew. *Gaussian Process Regression Models for Predicting Stock Trends.* January 2007.

[9] Ou, P., Wang, H. *Modeling and Forecasting Stock Market Volatility by Gaussian Processes based on GARCH , EGARCH and GJR Models.* Proceedings of the World Congress on Engineering 2011 Vol I WCE 2011, July 6 - 8, 2011, London, U.K.

[10] Chapados, N., Bengio, Y. *Forecasting and Trading Commodity Contract Spreads with Gaussian Processes.* University of Montreal. June 12th, 2007

[11] Kupper, Justin. *Cboe Volatility Index (VIX).* Guideline to Volatility. Investopedia. Received form https://www.investopedia.com/terms/v/vix.asp

[12] Takyi, Paul and Bentum-Ennin, Isaac. *The impact of COVID-19 on stock market performance in Africa: A Bayesian structural time series approach* J Econ Bus. December 8, 2020.105968. doi: 10.1016/j.jeconbus.2020.10596

[13] Swastanto, Bagas Abisena. *Gaussian Process Regression for Long-Term Time Series Forecasting.* TUDelft. Received from https://repository.tudelft.nl/islandora/object/uuid%3A7f3916c9-795c-403f-bec8-5eb6caa3e398

Department of Statistics, University of British Columbia