[**Points to cover**](#_ieqjt0joij1r) **1**

[About the dataset](#_j5yyxut0xki1) 2

[Yield Curve Data](#_jwtlsjyqghec) 2

[Goals](#_n58zji2jsp4v) 3

[Data Exploration](#_rzwk8rjmbp74) 3

[Data Visualization](#_hupchepavdvi) 3

[Data Preprocessing](#_wa9z3yvua8m2) 4

[Handling missing data](#_gukbt853xjyh) 4

[Smoothen data](#_j6sgcge0cdgm) 4

[Data Preprocessing for AR modeling:](#_wyhchzt2ffyx) 4

[Data Preprocessing for RF modelling:](#_hwc02picz7nc) 5

[Feature Extraction](#_rjey6ukinv64) 5

[Feature Extraction for RF model](#_kf7ek5yymtjq) 5

[Predicting Yields](#_5g3nsneg69cn) 6

[ARMA](#_80479pp5t3k8) 6

[Random Forest](#_tw2crlofgnnf) 8

[Literature Review](#_5cu3i3kt0c4x) 9

# Points to cover

a) A description of the problem and the data set, including details about data collection, if applicable

b) A description of initial data exploration

c) A statement of goals

d) A discussion of related literature

1. About the data - what does YTM mean, how in real world we arrive at that value, why is it important (1/2 page) - Zhirui,L

2. Goal - that you have historical yields and current yields for some time to maturity,predict next yield to maturity for next period - everyone

3. Data visualization - missing weekend,public holiday, data variation 18% yield in 62 and then drops to 3% in 90s. For some maturities, only data after/ before 1 year - Lakshya

3. Data preprocessing - savisky filter, analysing data stationarity (Vijay and Zhirui)

4. Feature Extraction - mean median,RMD , is weekend,public holiday, is weekday,

imputation(not needed as such) Vijay + Han

5. Random Forest Application to predict yields . L+Z

6. Literature review - Han

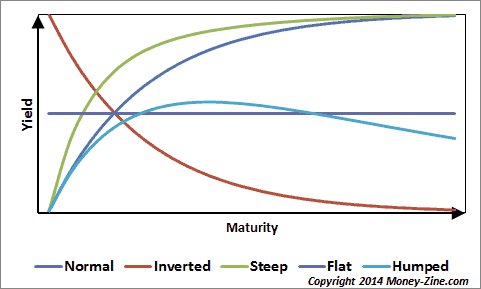
## About the dataset

### Yield Curve Data

A yield curve shows several yields or interest rates across different contract lengths (2 month, 2 year, 20 year) for a similar debt contract financial securities. The curve shows the relation between the interest rate (cost of borrowing) and the time to maturity, known as the "term", of the debt for a given borrower in a given currency. This IR (yield) can in-turn help gain insights into derive price deals, quantify risk, measure solvency, hedge, and more - thus foundational layer of quantitative financial analysis.

IR curve varies by asset class and this data may contain noise or have missing points depending upon trading frequency, liquidity levels or time of availability at the time of data snapshot. However, given their impact on financial analysis, it is quintessential to be able to accurately derive these while understanding the impact of this noise and missing values.

One very important YC is US Treasury yield curve - draws out a line chart to demonstrate a relationship between yields(price at which US treasury is willing to borrow money or investors are willing to lend) and maturities of on-the-run treasury fixed income debt securities. It illustrates the yields of Treasury securities at fixed maturities, viz. 1, 3 and 6 months and 1, 2, 3, 5, 7, 10, 20 and 30 years. These are zero coupon(no interest paid before maturity), risk-free( guaranteed return of principal and coupon promised at issue). Given this, these are often used as a benchmark to evaluate the relative worth of US Non-Treasury securities.



The spread between short term rates and long term rates that determines the slope of the yield curve, is a predictor of economic situation of the country. There are 4 major types of shapes:

1) Normal/upward slope: short term rates are lower than longer term rates, exhibiting an upward slope where investors are compensated for holding the longer term securities that possess greater investment risks. The higher yields on longer term maturity also means that the short term rates are likely to increase in the future as the growth in the economy would lead to higher inflation rates.

2)Inverted/Downward slope: This occurs when short term rates are greater than the long term rates. It would generally imply that both monetary and fiscal policies are currently restrictive in nature and the probability of the economy contracting in the future is high. According to empirical evidence, the Inverted Yield curve has been the best predictor of recessions in the economy.

3)Flat/Humped YC: This occurs when yields on medium term US Treasury Securities are higher than the yields on long term and the short term US Treasury Securities. This reflects that the current economic condition is unclear and the investors are uncertain about the economic scenario in the near future. It could also reflect that monetary policy is expansionary and fiscal policy is restrictive or vice-versa. It can also a predictor of an economic transition.

## Goals

* Create a ML model that can predict missing yield points by using historical data and partially available curve information provided as a time-series of yields for certain maturities.
* Analyze the impact of:
  + Random noise in yield data
  + Extent of missing information
* Once a stable model for US treasury yield curve is available, extend this to other asset classes with different liquidity levels and yield-maturity relationship.

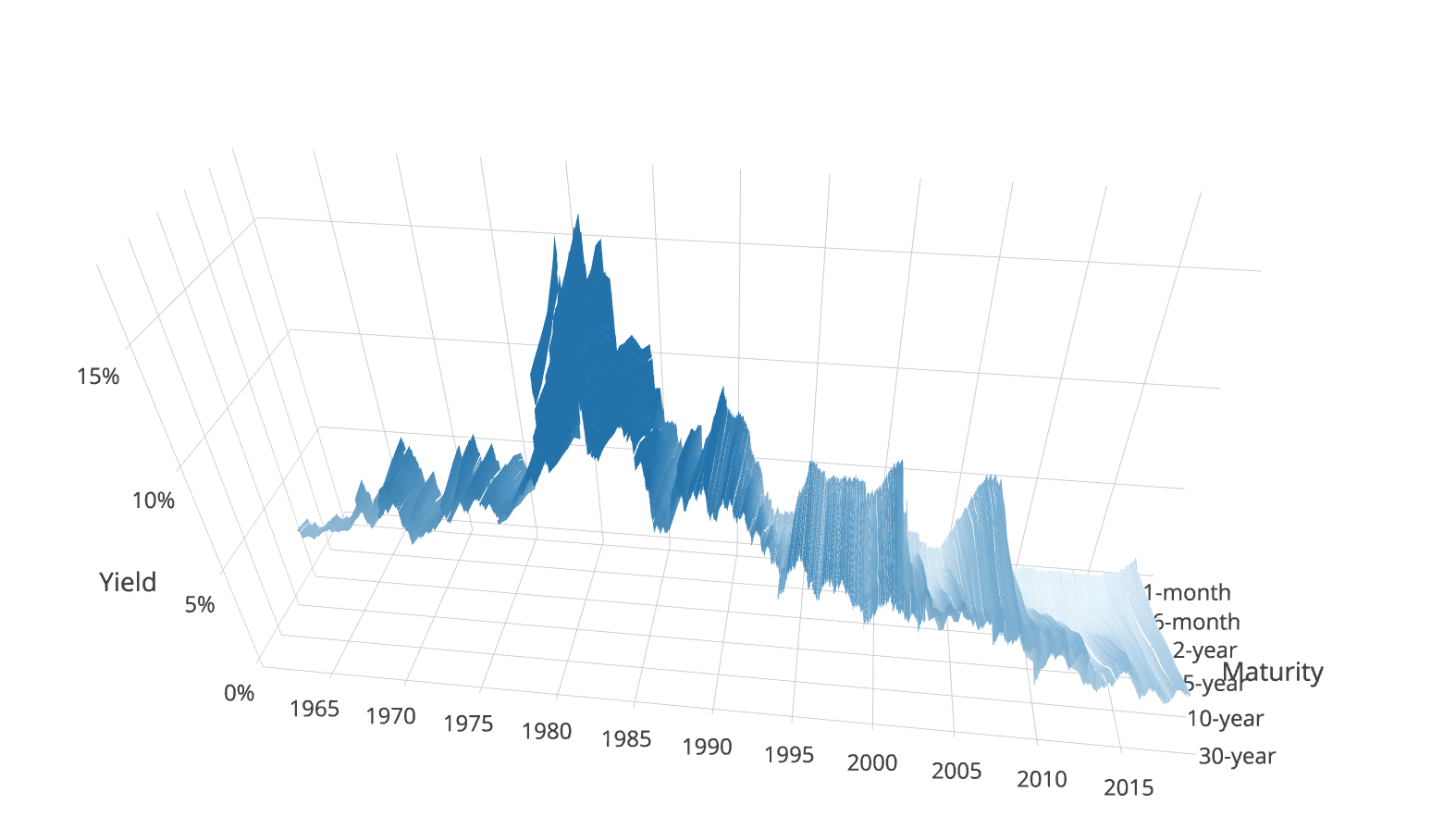
## Data Exploration

### Data Visualization

Below is the trend of last 47 years(1950 to 2017) of US treasury yields.

Hey insights:

* As we can see the yields around 1980 for all maturity lengths was significantly high(~16%) compared to those of present(~2-3%).
* Most of the public holiday and weekend data is missing.
* Some short terms yields are missing during most of the years.



### Data Preprocessing

#### Handling missing data

Some of the key features of this time series data includes missing values for particular days and one or more maturities.

* Missing weekend and public holiday data because markets close
* No 1/3/6-month yields available for pre 2000 data , because such extremely short yields were not offered at issue.

To circumvent above, we performed the following steps:

* yield rates are most correlated with the values from their immediate past e.g yield on monday = yield on friday
* For weekends instead of imputing values,we modified the time series such that there are no breaks after we ignore the weekend timestamps.

#### Smoothen data

The time series data needs to be statistically stationary for any model to provide valid results.

For RF we chose to use Savitzky–Golay filter and for ARMA rolling window functions which are described in detail in following sections.

#### Data Preprocessing for AR modeling:

* In order to smooth the data, we computed a rolling mean over a period of 100 days(approx. Business quarter)
* To detrend the data, we subtracted the smoothed values from the actual values.
* We then checked for stationarity using the Dickey Fuller test. This is applicable since the test is based on the intuition that if the series is integrated then the lagged level of the series (yt-1) will provide no relevant information in predicting the change in yt besides the one obtained in the lagged changes (yt-k).

#### Data Preprocessing for RF modelling:

* Smoothing peaks around end and start of week: In order to account for the extreme values on Monday or Friday, which is a common phenomenon in financial time series, we use Savitzky–Golay filter to smooth the series. It is a polynomial regressor that take in the previous rolling window of data points to fit a polynomial function, and predict the next period data point, so that there is seemingly continuity between each data point.
* We experimented with different window size and polynomial orders and chose 15 and 3 respectively, which gives us smooth enough series but still retains general trend in low and high values.

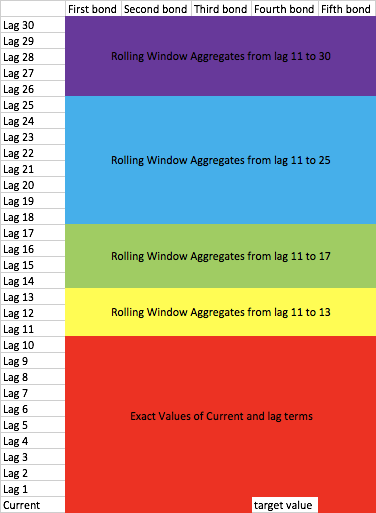
### Feature Extraction

Feature Extraction from time series data: The data we are dealing with only has datetimes and yield rates for different maturity periods. As a first logical step ,we computed following time based features derived from timestamps::

* is weekend or not
* is public holiday or not
* Day of week
* Month of year

#### Feature Extraction for RF model

Besides above features, we will also use a structural form VAR like setting to build up our additional features for RF models. For a certain target bond yield we want to predict, collect the current period exact value of yield of all other bonds; the lag 1 term to lag 10 terms of exact value of yield of all the bonds, including the target bond itself; and the rolling window aggregates of the lag 11 terms to 30 terms of yield of all bonds, including window size of 3, 7, 15 and 20. The whole setting can be visualized as following(note that target can be any of the 5 bonds, so there will be 5 RF models for this visualization):



For the rolling window aggregates, we will use:

1.Mean  
2.Standard Deviation  
3.Median  
4.Min  
5.Max  
6.Rooted Mean Square  
7.Crest Factor  
8.Zero-crossing Rate(here is Mean-crossing Rate)  
9.Trend(Slope of linear regression)

## Predicting Yields

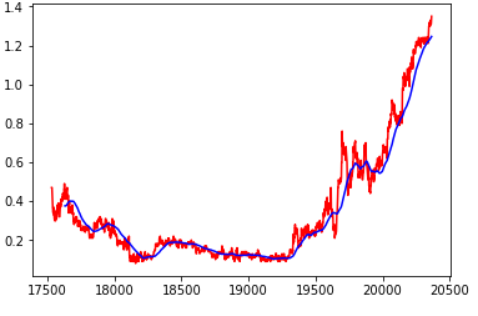
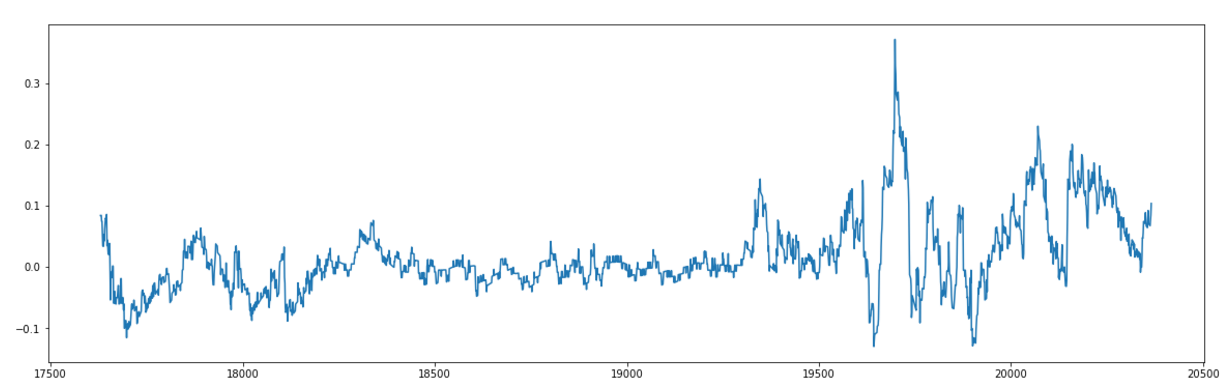
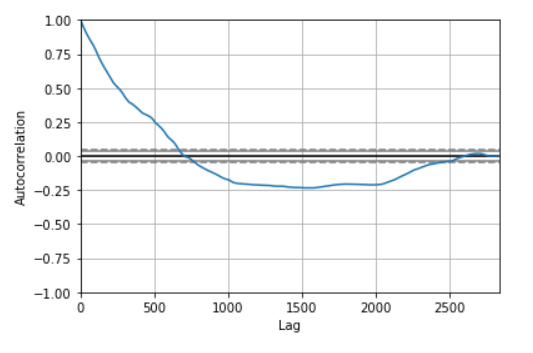
The following models are used to create a baseline and use US treasury Yield data.

The two which are described below:

### ARMA

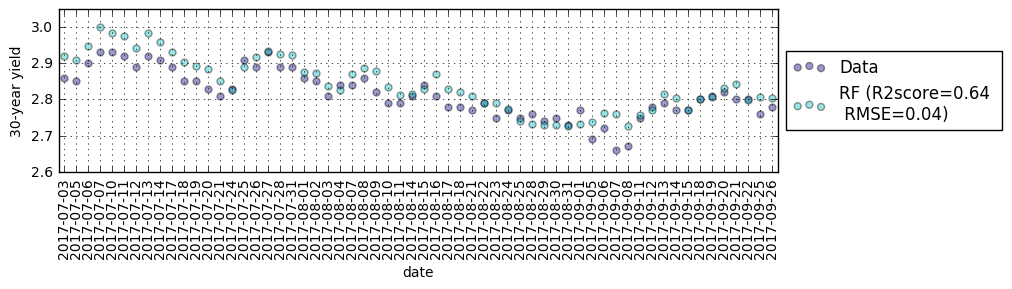
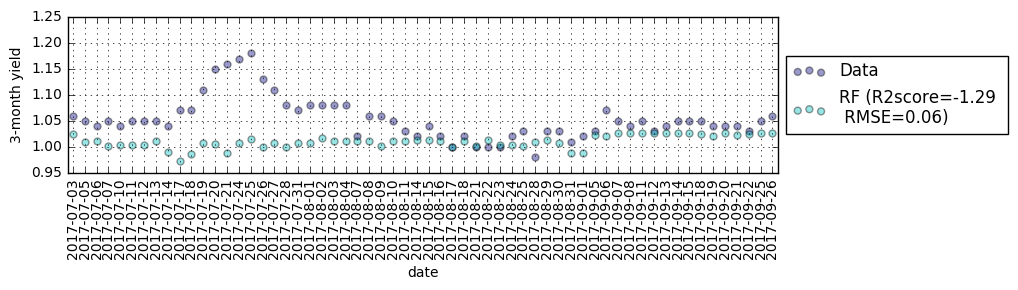
For AR/MA models to work, the data has to be made statistically stationary. This means that the data must have the same statistical properties such as mean,variance and autocorrelation over time. The data must also be present in equally spaced intervals.

Making the data stationary: In order to detrend the data, we first smoothed it by computing the rolling mean over a period of 100 days. We chose this time period as it comes close to the economic quarter. Once we have the smoothed values, we compute the difference between the actual and the smoothed values. This results in a series which looks visibly stationary. To confirm whether our series is stationary, we performed the Dicky Fuller test for stationarity and scored a significantly low p-value. Now our data is ready to be used for any ARIMA model.

* Aim of experiment : Forecast yield for single maturity such as 3-month or 30-years.
* Data scope: Data from 2000 to 2016 was considered as training data, as data before this point should have little to no effect on predicting values for 2017.
* Data preprocessing:
  + Impute missing values which do not fall on weekends, by using most recently available past value.
  + Smooth the data by computing rolling mean for a period of 100 days
  + 
  + Detrend the data by subtracting smoothed value from original value
  + 
  + Check stationarity of resulting values using Dickey Fuller Test
  + Check autocorrelation plots to identify p and q values for AR and MA
  + 
* Results and Conclusion: Our model achieved an RMSE of 0.017 and R squared score of 0.98. But this model cannot be implemented on it’s own as we have not considered same period features of other maturity rates. We plan to ensemble this model with some other models.

### Random Forest

* Aim of experiment : Given historical yields, predict yield for a single maturity on a given day
* Assumption: All other maturity yields for that day are present i.e with historical yield data from 01/01/2010 to 08/31/2017, we predict 1-year maturity yield for 09/01/2017 assuming all other maturity yields are available.
* Data scope: We consider Jan’2010 to Aug’2017 data as train and predict yield for every day of Sept for each maturity.
* Data preprocessing and feature extraction:
  + Remove missing NA data which is for public holidays as described in data pre processing section
  + Smoothen data to remove Monday/Friday peaks as described in data pre processing section
  + Add rolling window and lag features such that information from one week ago is more significant than that a month ago as described in feature extraction section
* Model
  + RF model with max depth 30. Pending other hyper parameter tuning.
  + Train and test shapes: Y\_train: (1869, 1), X\_train: (1869, 342), Y\_test: (60, 1), X\_test: (60, 342)
* Results:
  + The model does a decent job while predicting longer maturity yields but performs miserably for shorter yields. This can also be attributed to volatility in short term yields.
  + We achieve a R square as high as .92 for 10-year yield. Infact, for maturities greater than 3 year, R square metric stays above 0.6. However, we also report negative figures for this metric for shorter maturity such as -7 for 1 month yield where predictions are void.
  + With this as baseline, we will start thinking of an ensemble model that uses different model for different yields and perhaps different features.



## Literature Review

In the book “Analysis of Financial Time Series,” Ruey S. Tsay mentions a method to fill missing values in a time series as an application of MCMC. Markov Chain Monte Carlo (MCMC) methods are widely used in time series analysis as the computing facilities advance in recent years. In the literature, missing values in a time series can be handled by Bayesian inference via Gibbs sampling (Tsay 613).

In our project, this method is useful for data imputation: we can impute missing data in our training set more accurately than some simple imputation methods such as filling the previous day value. In this way, we can potentially improve the performance of the trained model.

Specifically, we treat the missing value xh as an unknown parameter, and we model the time series as AR of order p as xt = φ1xt-1 + … + φpxt-p + at . Then the parameters we want to estimate are ***θ*** = （***φ***，xh ，σ²）, where ***φ***  = (φ1 … φp)’ and σ² is the variance of the innovation term at. Assume the prior distributions of these parameters as:

***φ ~*** N(***φ***0 , ***Σ***0 ) xh ~ N（μ0, σ20） ~

Here, ***φ***0 , ***Σ***0 , μ0, σ20 , v and λ are all hyperparameters from our choice.

With the three prior distributions defined above and the AR (p) model, we can derive three conditional posterior normal distributions of the parameters. For simplicity, I only try the AR(1) model xt = φxt-1 + at , with prior of φ ***~*** N(φ0 , Σ0 ) as univariate normal. The posterior distributions are described below:

1. f(φ |**X**, xh , σ²) - The posterior distribution of φ is normal with ,
2. f(σ² |**X**, xh ,φ) - The posterior distribution of is an inverted chi-squared distribution:

~

1. f(xh |**X**, σ², φ) - The posterior distribution of xh is normal with , ，where

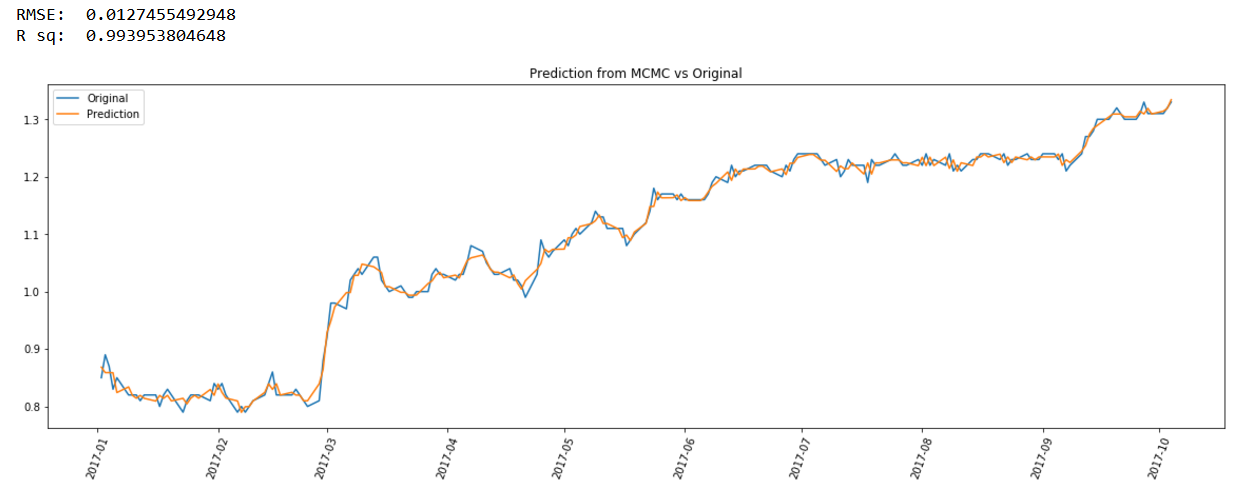
Note: in the above formula, n is the sample size, and 1<h<n. And **X** represents the available data points.

Using the above conditional posterior distributions, we can estimate φ, σ² and xh using Gibbs sampling with thousands of iterations as follows:

1. Specify arbitrary starting values for φ, σ² and xh.
2. Use the normal distribution f(φ |**X**, xh , σ²) to draw a random realization for φ
3. Use the chi-squared distribution f(σ² |**X**, xh ,φ) to draw a random realization for σ²
4. Use the normal distribution f(xh |**X**, σ², φ) to draw a random realization for xh

Repeat steps 2-4 for many iterations to obtain a Gibbs sample. Then use the sample means to get a point estimates of the parameters (Tsay 626).

I implement the above method using the one-year maturity bond YTM data after 2000, from ‘FRB\_H15’ （US Fed treasury curves）dataset. To test the method, I assume one day’s value in 2017 is missing, but the remaining values are all available. I repeat this for 198 days in 2017, and then compute the RMSE of the method. The RMSE is 0.01275, and the R square score is 0.99395. This result is better than using AR(1) model to predict the missing day value with traditional MLE method and treat the problem as a simple prediction problem. The figure below shows the predicted missing value from MCMC method versus the actual value.



I also record the Gibbs sampling estimation of model parameters φ and for each day. The mean of φ for 198 test days is 0.99265, and the mean of is 0.00164. These estimations are close to maximum likelihood estimation using actual data without missing values, where φMLE is 0.99246 and is 0.00153.

Problems of the method described above:

1. The values of hyperparameters, and the starting values of Gibbs Sampling are from our arbitrary choice, and different values might lead to different estimates.
2. Estimate one missing value needs thousands (perhaps more) iterations. If we use a more complex model than AR(1) and we have a large number of missing values, the method might be time-consuming.
3. This method might not perform as good as the machine learning models that we have tried, as it only focuses on a single time series and ignores many other features in our machine learning models.

Possible Future Research:

1. We should check the convergence of a Gibbs sample. We might repeat the Gibbs sampling several times with different starting values to ensure the algorithms has converged.
2. We should try more complex models with more lag terms than a simple AR(1) model.
3. In order to include some interactive features like our random forest model, we can consider adding explanatory variables to the model such as the YTM of other bonds with different maturity time. For example, we can use a regression model with serially correlated errors.
4. We can modify the Gibbs sampling process for consecutive missing values.

Works Cited

Tsay, Ruey S. *Analysis of financial time series*. 3rd ed., John Wiley & Sons, 2010.

Works Cited

Tsay, Ruey S. *Analysis of financial time series*. 3rd ed., John Wiley & Sons, 2010.