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).

Specifically, we treat the missing value x_h as an unknown parameter, and we model the time series as AR of order p as $x_t = \varphi_1 x_{t-1} + \cdots + \varphi_p x_{t-p} + a_t$. Then the parameters we want to estimate are $\boldsymbol{\theta} = (\boldsymbol{\varphi}, x_h, \sigma^2)$, where $\boldsymbol{\varphi} = (\varphi_1 \cdots \varphi_p)'$ and σ^2 is the variance of the innovation term a_t. Assume the prior distributions of these parameters as:

$$\boldsymbol{\varphi} \sim N(\boldsymbol{\varphi}_0, \boldsymbol{\Sigma}_0) \qquad x_h \sim N(\mu_0, \sigma^2_0) \qquad \frac{v\lambda}{\sigma^2} \sim \chi^2_v$$

Here, $\boldsymbol{\varphi}_0$, $\boldsymbol{\Sigma}_0$, μ_0, σ^2_0, 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 $x_t = \varphi x_{t-1} + a_t$, with prior of $\varphi \sim N(\varphi_0, \Sigma_0)$ as univariate normal. The posterior distributions are described below:

1. $f(\varphi \mid \mathbf{X}, x_h, \sigma^2)$ - The posterior distribution of φ is normal with $\sigma^2_* = \left(\frac{\sum_{t=2}^{n} x_{t-1}^2}{\sigma^2} + \Sigma_0^{-2}\right)^{-1}$,

   $\mu_* = \sigma^2_* \left[\frac{\sum_{t=2}^{n} x_{t-1}^2}{\sigma^2} \left(\sum_{t=2}^{n} x_{t-1}^2\right)^{-1} \left(\sum_{t=2}^{n} x_{t-1} x_t\right) + \Sigma_0^{-2} \varphi_0\right]$

2. $f(\sigma^2 \mid \mathbf{X}, x_h, \varphi)$ - The posterior distribution of σ^2 is an inverted chi-squared distribution:

   $$\frac{v\lambda + \sum_{t=2}^{n} (x_t - \varphi x_{t-1})^2}{\sigma^2} \sim \chi^2_{v+(n-1)}$$

3. $f(x_h \mid \mathbf{X}, \sigma^2, \varphi)$ - The posterior distribution of x_h is normal with $\mu_* = \frac{\sigma^2 \mu_0 + \sigma_0^2 (1+\varphi^2) \widehat{x_h}}{\sigma^2 + \sigma_0^2 (1+\varphi^2)}$,

   $\sigma^2_* = \frac{\sigma^2 \sigma_0^2}{\sigma^2 + \sigma_0^2 (1+\varphi^2)}$, where $\widehat{x_h} = \frac{\varphi}{1+\varphi^2} (x_{h-1} + x_{h+1})$

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

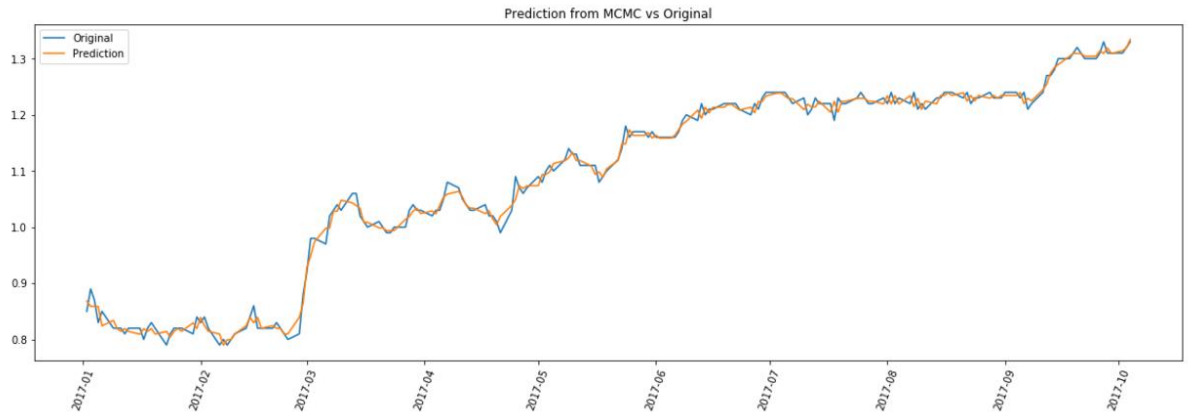
Using the above conditional posterior distributions, we can estimate φ, σ^2 and x_h using Gibbs sampling with thousands of iterations as follows:

1. Specify arbitrary starting values for φ, σ^2 and x_h.
2. Use the normal distribution $f(\varphi \mid \mathbf{X}, x_h, \sigma^2)$ to draw a random realization for φ
3. Use the chi-squared distribution $f(\sigma^2 \mid \mathbf{X}, x_h, \varphi)$ to draw a random realization for σ^2
4. Use the normal distribution $f(x_h \mid \mathbf{X}, \sigma^2, \varphi)$ to draw a random realization for x_h

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.

```
RMSE:  0.0127455492948
R sq:  0.993953804648
```



I also record the Gibbs sampling estimation of model parameters φ and σ^2 for each day. The mean of φ for 198 test days is 0.99265, and the mean of σ^2 is 0.00164. These estimations are close to maximum likelihood estimation using actual data without missing values, where φ_{MLE} is 0.99246 and σ^2_{MLE} 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.