Multiple Imputation on Missing Values in Time Series Data

by

Sohae Oh

Program in Statistical and Economic Modeling
Duke University

Date:_____

Approved:

_____

Fan Li, Supervisor

_____

Jerry P. Reiter

_____

Charles M. Becker

Thesis submitted in partial fulfillment of
the requirements for the degree of
Master of Science in the Program of
Statistical and Economic Modeling in the Graduate School
of Duke University

2015

ABSTRACT

Multiple Imputation on Missing Values in Time Series Data

by

Sohae Oh

Program of Statistical and Economic Modeling
Duke University

Date:_____

Approved:

_____

Fan Li, Supervisor

_____

Jerry P. Reiter

_____

Charles M. Becker

An abstract of a thesis submitted in partial
fulfillment of the requirements for the degree
of Master of Science in the Program of
Statistical and Economic Modeling in the Graduate School of
Duke University

2015

Abstract

Financial stock market data, for various reasons, frequently contain missing values. One reason for this is that, because the markets close for holidays, daily stock prices are not always observed. This creates gaps in information, making it difficult to predict the following day's stock prices. In this situation, information during the holiday can be "borrowed" from other countries' stock market, since global stock prices tend to show similar movements and are in fact highly correlated. The main goal of this study is to combine stock index data from various markets around the world and develop an algorithm to impute the missing values in individual stock index using "information-sharing" between different time series. To develop imputation algorithm that accommodate time series-specific features, we take multiple imputation approach using dynamic linear model for time-series and panel data. This algorithm assumes ignorable missing data mechanism, as which missingness due to holiday. The posterior distributions of parameters, including missing values, is simulated using Monte Carlo Markov Chain (MCMC) methods and estimates from sets of draws are then combined using Rubin's combination rule, rendering final inference of the data set. Specifically, we use the Gibbs sampler and Forward Filtering and Backward Sampling (FFBS) to simulate joint posterior distribution and posterior predictive distribution of latent variables and other parameters. A simulation study is conducted to check the validity

and the performance of the algorithm using two error-based measurements: Root Mean Square Error (RMSE), and Normalized Root Mean Square Error (NRMSE). We compared the overall trend of imputed time series with complete data set, and inspected the in-sample predictability of the algorithm using Last Value Carried Forward (LVCF) method as a bench mark. The algorithm is applied to real stock price index data from US, Japan, Hong Kong, UK and Germany. From both of the simulation and the application, we concluded that the imputation algorithm performs well enough to achieve our original goal, predicting the stock price for the opening price after a holiday, outperforming the benchmark method. We believe this multiple imputation algorithm can be used in many applications that deal with time series with missing values such as financial and economic data and biomedical data.

Dedication

To my mom and dad for their unconditional support and love through all my walks of life

Contents

List of Tables

List of Figures

Acknowledgements

Many have contributed to my thesis in a various way. I would like to first express sincere gratitude to my academic advisor and committee chair, Prof. Fan Li. She taught me how to think, analyze and write in an academic way throughout my work on this Master's degree thesis. This will be my valuable asset that I will carry in my future.

Also I would like to thank Prof. Jerry Reiter for his comments and suggestions on the thesis. He advised me from formulating the research idea to final revision of this thesis and this helped me a lot. Research in this thesis was supported by a grant from the National Science Foundation (SES-11-31897).). I would also like to extend special thanks to the graduate school representative and Professor Charles Becker for serving as my general advisor, prof. Hyoung goo Kang and prof. Changmin Lee in Hanyang University for their support and guide. I thank all of my friends at Duke and Duraleigh Presbyterian Church for their love and support through all these tough but meaningful years.

Finally, I would like to thank God.

1. Introduction

Financial stock market data, for various reasons, frequently contain missing values. One reason for this is that, because the markets close for holidays, daily stock prices are not always observed. Such gaps in information make it difficult to predict future stock prices using the most up-to-date market information. For example, the US stock market is closed on Christmas, so the S&P 500 is not observed on December 25th. When the interest is in predicting an opening stock index price of December 26th, the common practice is to use market information from December 24th. Such a forecast would arguably be better if a Christmas day price were known. Imputing plausible values for such gaps could play a crucial role in predicting a stock's opening price on the following day.

In this situation, information during the holiday can be "borrowed" from other countries' stock market, since global stock prices tend to show similar movements and are in fact highly correlated. Holidays vary from country to country. So while stock prices are not observed in the US stock market on Christmas, for example, those from Japan are because Christmas is not a national holiday there. Therefore, the main goal of this study is to combine stock index data from various markets around the world and develop an algorithm to impute the missing values in individual stock index using "information-sharing" between different time series.

Missing values in data are often discarded for convenience. For example, a stock's price on December 25th is dropped from a data set. However, eliminating incomplete cases with missing values can result in the loss of key information relevant to the inference. Moreover, it ignores possible systematic differences between incomplete cases and complete cases in the data. Sometimes missing values are substituted with plausible single values. One of the common methods in panel data is Last Value Carried Forward (LVCF). LVCF uses the last observation to impute the missing values. In the example of a stock price, the price on Christmas Eve is substituted for the missing price for Christmas Day. Occasionally used in the imputation are the mean value, mode, or other summary statistics. Single imputation, however, is also problematic because it does not reflect the uncertainty that arises from the prediction of the missing value. Researchers can take the model-based single imputation approach such as Maximum-likelihood Estimation, which estimates parameters that produce the highest log-likelihood, resulting in using full information and unbiased parameters estimation, but this will shrink the standard error of the target estimand downward. The practice of using multiple imputations, which by now has become a standard strategy for handling missing values, was first developed by Rubin (1987). This practice takes the uncertainty of imputed missing values into account by replacing them with a set of plausible values.

Multiple imputation proceeds from a Bayesian perspective, where missing values are treated as unknown parameters and the information represented by these

missing value is expressed as the posterior predictive distribution, depending on observed values. Posterior predictive distribution of missing values, jointly with other parameters, can be obtained from the prior distribution of the missing values and unknown parameters and the likelihood distribution of observations based on Bayes theorem. The posterior distribution can be simulated using Monte Carlo Markov Chain (MCMC) methods and M sets of estimates of missing values; other parameters are drawn from this joint posterior distribution, making M-complete data sets. These multiple draws form the posterior distribution reflect the uncertainty in predicting missing values. Estimates from each M-complete data set is then combined using Rubin's combination rule, rendering final inference of the data set. Point estimates of the parameter of interest are the average of each point estimate from the M-complete data set, and the variance of the parameter is calculated using within-imputation variance and between-imputation variance.

How to handle missing values in surveys and cross-sectional data has been the subject of many studies (Rubin, 1987, Little, 1993, Raghunathan et al, 2003, Reiter, 2006). What has yet to be studied, however, are missing values in time series data. One of the first studies to conduct multiple imputation on time-series data was Hopke et al (2001). They adopted the integrated first-order moving average model and extended it to a more complex model that also captured the seasonal effect of the time series. Honaker et al (2010) suggested a multiple imputation model for a specific type of data, called time-

series cross-section (TSCS) data, which have T units for every N cross-sectional variable, often with T < N. The large number of N leads to computational difficulties, so Honaker and colleagues used bootstrapping to draw parameters from the posterior distribution. Multiple imputation models should be carefully constructed to incorporate unique features of time-series. The multiple imputation model used in this paper is designed to accommodate time series-specific features using state-space model for time-series and panel data.

Represented in a state-space model are many financial and economic time series that have unobserved components. Time series analyses using the state-space model have been done in numerous studies, including Durbin and Koopman (2002), West and Harrison (1997), Kunsch (2001), Migon et al (2005), among many others. The state-space model has a set of state vectors, which contain unobserved stochastic processes and the observation is dependent on these latent state variables. A general state-space model consists of two parts: state variables (θ_t) and observations (Y_t) for each time period, t. States, often in a vector form, are latent, autoregressive, and affect the observation of the current time period. This makes the relationship between Y_i and Y_j modelled only according to the dependence between θ_i and θ_j, ($i \neq j$), as can be seen in Figure 1.

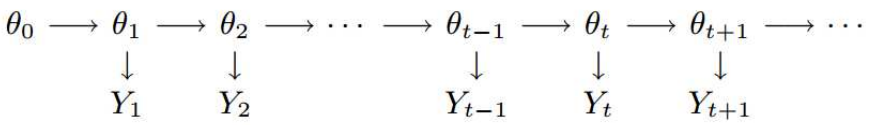

Figure 1: Dependence structure for a state-space model

A system of two equations, along with initial values, forms state-space models: an observation equation and a system equation. The observation equation describes the behavior of the observation in terms of state vectors and observational errors. The system equation depicts the evolution structure of state vectors over time. One benefit of using the state-space model is that it can be used for modeling multivariate time series in the presence of non-stationarity (West & Harrison, 1997). Typically, stock price data are non-stationary, especially when they follow a random walk, as in our model, described in Chapter 2. Results using non-stationary data are often unreliable in forecasting, so preliminary transformation is needed to achieve stationarity, which sometimes prevents us from investigating sudden movements and structural breaks in the time series. State-space models also allow terms to model trends, seasonality, and autoregressive features.

This thesis is organized as follows. Chapter 2 details the dynamic linear model, a special case of the state-space model being linear and Gaussian used in this model. Chapter 3 goes through the process of estimating unknown parameters and underlying latent state vectors in the model using the Gibbs sampler and Forward Filtering and

Backward Sampling (FFBS). This is followed in Chapter 4 with simulation studies. Chapter 5 applies the algorithm to daily stock index data from five markets. Chapter 6 wraps up this paper with a conclusion and further discussion.

2. Method

2.1 Notation

The multiple time series data used in this thesis, both in simulation studies in Chapter 4 and applications in Chapter 5, are in 2-dimensional matrix form: In rows are time periods and in columns are individual time series chains—indicating different stock indexes; time periods are denoted as t ($t = 1, \dots, T$), and time series chains as r ($r = 1, \dots, R$), resulting in a T X R matrix of multivariate time series data. Observation data matrix is denoted as Y, and the latent state matrix is denoted as θ. Both Y and θ have the size of T X R, and the [t, r] elements of Y and θ are notated as $Y_{t,r}$ and $\theta_{t,r}$. $Y_t = (Y_{t,1}, \dots, Y_{t,R})$ is the R-vector of observations at time t and $\theta_t = (\theta_{t,1}, \dots, \theta_{t,R})$ is the R-vector of state at time t, serving as rows in θ. Y can be partitioned to the observed part and the missing part: $Y = (Y^{obs}, Y^{mis})$, where Y_t^{obs} is a subset of Y_t that are observed, and Y_t^{mis} is a subset of Y_t that are missing.

2.2 Multiple Imputation

2.2.1 Missing Data Mechanism

Suppose Y is the data, partitioned to the observed part and the missing part, $Y = (Y^{obs}, Y^{mis})$, θ is a vector of unknown parameters. $Z = [z_{i,j}]$ is the binary missing data indicator variable such that

$$z_{i,j} = \begin{cases} 0 & y_{i,j} \; is \; observed \\ 1 & y_{i,j} \; is \; missing \end{cases}$$

Then, the joint model (likelihood) of full data is $f(\mathrm{Y},\mathrm{Z}|\theta) = f(Y^{obs}, Y^{mis}, Z|\theta)$. This likelihood for full data cannot be evaluated since it depends on $\mathrm{Y}^{\mathrm{mis}}$, which is also another unknown parameter. Hence, by integrating out $\mathrm{Y}^{\mathrm{mis}}$, we get $f\left(Y^{obs}, Z\middle|\theta\right) = \int f\left(Y^{obs}, Y^{mis}, Z\middle|\theta\right) \mathrm{dY}^{\mathrm{mis}} = \int f\left(Z\middle|Y^{obs}, Y^{mis}, \theta\right) * f\left(Y^{obs}, Y^{mis}\middle|\theta\right) \mathrm{dY}^{\mathrm{mis}}$. The missing data mechanism refers to the relationship between the probability of data missing and the observed values, i.e., $f\left(Z\middle|Y^{obs}, Y^{mis}, \theta\right)$ here. The missing data mechanism (Rubin, 1976) includes Missing Completely at Random (MCAR), Missing at Random (MAR), and Missing Not at Random (MNAR). MCAR is when the probability of missing values is dependent on neither the observed value nor the missing values, i.e., $f\left(Z\middle|Y^{obs}, Y^{mis}\right) = f(Z)$. MAR assumes that the probability of certain values are missing depend on observed values but does not depend on the value of missing observations, i.e., $f\left(Z\middle|Y^{obs}, Y^{mis}\right) = f(Z|Y^{obs})$. This is considerably a weaker assumption than MCAR. MNAR is a case where neither MCAR nor MAR hold. If MAR is assumed, the joint likelihood for full data becomes $f\left(Y^{obs}, Z\middle|\theta\right) = f\left(Z\middle|Y^{obs}, \theta\right) * f\left(Y^{obs}\middle|\theta\right)$ and we can ignore the missing data mechanism. Thus MAR is called ignorable. Priors for unknown model parameters and missing values $\pi\left(Y^{mis}, \theta\right)$ are imposed and $\mathrm{Y}^{\mathrm{mis}}$ are imputed, with M (>1) drawn from the posterior predictive distribution $f\left(Y^{mis}\middle|Y^{obs}, Z, \theta\right)$.

An MAR assumption is reasonable here, as the data used in this study are stock price index data with missing values due to holidays that occur in each country. This study focuses on stock price indexes from July to September of 2009, reflecting the

markets in the U.S., Japan, Hong Kong, the UK, and Germany. Missingness due to holidays qualifies as missing at random: The probability of a specific day being designated as a holiday in any country is unrelated to the stock price on that particular day if the price had been observed. It may be argued, however, that some anomalies and seasonal trends exist in the stock market, like the January effect (i.e., stock prices tend to rise at the beginning of each calendar year due to optimism and the selling and subsequent buying of stocks to avoid tax) or pre-holiday effect (i.e., stock prices tend to rise before major holidays like Christmas and Thanksgiving, due to optimism and increased spending for retail firms), meaning stock price data is MNAR (Missing Not at Random). However, these anomalies are not observed consistently and no research result exists about these seasonal trends since the stock market is, generally speaking, efficient.

2.2.2 Steps of Multiple Imputation

The multiple imputation procedure follows three steps. First, for every missing entry in the data, imputation is conducted by drawing samples from the posterior predictive distribution M times, creating M sets of complete data, denoted as $\mathrm{Y}^{\mathrm{inc},(\mathrm{m})} = \left(Y^{obs}, Y^{mis,(m)}\right), m = 1, \dots, M$. Let Q be quantities of interest, $\hat{Q}$ be its estimator, and U be the associated variance. Analysis is made on each M-complete data set individually: Q is

estimated using each complete data, denoted as $\hat{Q}_{*m}$ and the associated variances as U_{*m}. Valid inference can be made by simply combining inferences from M-complete data set, using Rubin's combination rule (Rubin, 1987). A repeated-imputation estimator is the average of the M-complete data estimates, $\bar{Q}_M = \sum_{m=1}^{M} \hat{Q}_{*m}$ /M and the variance associated with $\bar{Q}_M$ is $T_M = \bar{U}_M + \left(1 + \frac{1}{M}\right) B_M$, where $\bar{U}_M = \frac{1}{M}\sum_{m=1}^{M} U_{*m}$ is the within-imputation variability and $B_M = \frac{1}{M-1}\sum_{m=1}^{M}\left(\hat{Q}_{*m} - \bar{Q}_M\right)^T\left(\hat{Q}_{*m} - \bar{Q}_M\right)/(M-1)$ is the between-imputation variability. $\left(Q - \hat{Q}\right) \sim T_\nu(0, U)$ with moderate size of M, where $\nu = (M-1)(1 + r_M^{-1})B_M/\bar{U}_M$ is a degree of freedom of the t-distribution and $r_M = (1 + M^{-1})B_M/\bar{U}_M$. A $(1-\alpha)$% confidence interval for Q is $\bar{Q}_M \pm t_\nu\left(\frac{\alpha}{2}\right)\sqrt{T_M}$.

2.3 Multivariate Dynamic Linear Model

A dynamic linear model (DLM) is a special case of state-space models—linear and Gaussian (i.e., the joint distribution of $\{Y_t\}$ and $\{\theta_t\}$ are multivariate normal). The general class of DLM models are represented with a quadruple $\{F_t, G_t, V_t, W_t\}$, called system matrices, and defined (West and Harrison, 1997) as follows:

$$\text{Observation equation:} \quad Y_t = F_t\theta_t + \nu_t, \qquad \nu_t \sim N(0, V_t)$$

$$\text{System equation:} \quad \theta_t = G_t\theta_{t-1} + \omega_t, \qquad \omega_t \sim N(0, W_t)$$

$$\text{Initial information:} \quad (\mu_0|D_0) \sim N(m_0, C_0)$$

where ν_t is independent of ν_s, ω_t is independent of ω_s for $t \neq s$. System matrices may contain non-random components.

The basic form of the model considered in this thesis follows that of Prado and West (2010)

$$Y_t = \theta_t + \nu_t, \nu_t \sim N(0, V_t * \Sigma_t) \tag{1}$$

$$\theta_t = \theta_{t-1} + \Omega_t, \Omega_t \sim MN(0, W_t, \Sigma_t) \tag{2}$$

For each time period, there is a latent state θ_t for each time series and the observed value Y_t is assumed to be affected by this state with the observational error. The state value θ_t follows random walk (i.e., $G_t = I$) with the evolution noise Ω_t. $\nu_t = [\nu_{t,1}, \ldots, \nu_{t,R}]$ is a R-vector of the observation errors at time t and distributed according to

multivariate normal distribution with mean vector 0_R and variance-covariance matrix equals to $V_t * \Sigma_t$. V_t is assumed to be a known, multiplicative scalar to Σ_t. $\Omega_t = [\omega_{t,1}, \dots, \omega_{t,R}]$ is a R-vector of the evolution errors with matrix-variate normal distribution, with mean matrix 0_R, row variance matrix W_t and column variance matrix Σ_t. W_t is a known, row variance matrix, and, in this model, a scalar. Finally, Σ_t is a time-varying, unknown covariance matrix between R numbers of time series. In detail, for two different time series chain i and j, $\mathrm{cov}(\nu_{t,i}, \nu_{t,j}) = V_t * \sigma_{t,i,j}$ and $\mathrm{cov}(\omega_{t,i}, \omega_{t,j}) = \sigma_{t,i,j} * W_t$ $(i, j = 1, \dots, r, i \neq j)$

The covariance matrix between time series Σ_t is assumed to be changed slowly and unpredictably over time. Second, in this thesis, this covariance matrix is "discounted" as time progresses at a rate of β. Information available at time t, is denoted as I_t. The posterior distribution of covariance matrix at time t-1 follows Inverse-Wishart distribution, $(\Sigma_{t-1}|I_{t-1}) \sim IW(n_{t-1}, D_{t-1})$, where n and D are the degree of freedom and the scale matrix of the inverse-Wishart distribution, respectively. At the beginning of time t, the prior of the covariance matrix becomes $(\Sigma_t|I_{t-1}) \sim IW(\beta * n_{t-1} - R + 1, \beta * D_{t-1})$. After Y_t is observed, the posterior distribution is finally derived as $(\Sigma_t|I_t) \sim IW(n_t, D_t)$, where $n_t = n_{t-1} + 1$ and $D_t = D_{t-1} + e_t * \frac{e_t^T}{q_t}$. This is described in the estimation section in detail.

This study assumes multiple chains of time series to be highly correlated. The multiple time series are thought to have the same state evolution structure over time,

thus enabling us to extend the univariate dynamic linear model into a multivariate dynamic linear model. This type of time series chains is referred to, in Prado and West (2010), as exchangeable time series. According to Hamao et al (1990), there exists a high degree of correlation among the stock price changes in major international stock indexes, which serves as the ground of assuming our data a multivariate exchangeable time series.

3. Estimation

Imputing missing values can be done by treating missing objects as unknown parameters. In this model, state vectors, covariance matrices, and missing values, respectively $\{\theta_t\}, \{\Sigma_t\}$ and $\{Y_t^{mis}\}$ for $t = 1, \dots, T$ are the unknown parameters. In general, it is hard to obtain the analytic form of the joint posterior distribution and posterior predictive distribution of those parameters. Consequently, for the posterior simulation, the study uses the Monte Carlo Markov Chain (MCMC; Gelfand & Smith, 1990) scheme. In this model, full conditional distribution for each parameter, given other parameters, is available. We use the Gibbs sampler to simulate joint posterior distribution and posterior predictive distribution.

3.1 Gibbs Sampler

The Gibbs Sampler is an iterative MCMC algorithm that enables generating joint distribution of multiple parameters from each of their "full conditional distribution." For an R-dimensional parameter vector X, full conditional distribution is defined as $p(X_r|X_{(-r)})$, where $X_{(-r)} = \{X_1, X_2, \dots, X_{r-1}, X_{r+1}, \dots X_R\}$. For each iteration s, Gibbs sampler cycles through R-number of full conditional distributions of parameters and draws values from them, conditioning values from the previous iteration:

$$X_r^{(s)} \sim p\left(X_r \middle| X_{-r}^{(s-1)}\right) for\ r = 1, \dots, R, s = 1, \dots, S$$

As the number of iterations S increases, $p(X^{(S)})$ converges to $P(X)$.

In this thesis, unknown parameters are divided into $\{(\theta_t, \Sigma_t)\}$ and $\{Y_t^{mis}\}$ for all t and their full conditional distribution is denoted as $p(\theta_t, \Sigma_t | Y_t, Y_t^{mis})$ and $p(Y_t^{mis} | Y_t, \theta_t, \Sigma_t)$. At each iteration s, a random sample of the parameter is drawn from their full conditional distributions:

$$\left(\theta_t^{(s)}, \Sigma_t^{(s)}\right) \sim p\left(\theta_t, \Sigma_t \middle| Y_t, Y_t^{mis,(s-1)}\right)$$

$$\left(Y_t^{mis,(s)}\right) \sim p\left(Y_t^{mis} \middle| Y_t, \theta_t^{(s)}, \Sigma_t^{(s)}\right)$$

As shown in Dawid (1981) and Press (1982), there is a conjugacy between the matrix-variate normal distribution for θ_t and the inverse-Wishart distribution for Σ_t. As θ_t follows a matrix-variate normal distribution, which is dependent on Σ_t, and Σ_t and follows an inverse-Wishart distribution, (θ_t, Σ_t) jointly follows a MNIW (Matrix-variate Normal, Inverse-Wishart) distribution, that is

$$(\Sigma_t) \sim IW(n_t, D_t)$$

$$(\theta_t | \Sigma_t) \sim \mathrm{MN}(M_t, C_t, \Sigma_t)$$

$$\Rightarrow \ p(\theta_t, \Sigma_t) \sim MNIW(M_t, C_t, n_t, D_t)$$

Derivation of the full conditional distribution of Y_t^{mis} can be found in Hoff (2009). Let a be the subset of the index that Y_t is observed and b be the subset of the index that

Y_t is missing. $\theta_t^{[b]}$ is an element of θ_t corresponding to the index in b, and $\Sigma_t^{[b,a]}$ is the submatrix made up corresponding to Row a and Column b of Σ_t. Then the full conditional distribution of missing values is

$$P\left(Y_t^{mis} \middle| \theta_t, \Sigma_t, Y_t^{obs}\right) \propto N\left(\theta_t^{b|a}, \Sigma_t^{b|a}\right)$$

$$\text{where } \theta_t^{b|a} = \theta_t^{[b]} + \Sigma_t^{[b,a]}\left(\Sigma_t^{[a,a]}\right)^{-1}\left(Y_t^{obs} - \theta_t^{[a]}\right) \text{ and } \Sigma_t^{b|a} = \Sigma_t^{[b,b]} - \Sigma_t^{[b,a]}\left(\Sigma_t^{[a,a]}\right)^{-1}\Sigma_t^{[a,b]}$$

3.2 Forward Filtering Backward Sampling

To generate the latent state vector $\{\theta_t\}$ and covariance matrices $\{\Sigma_t\}$, Carlin et al (1992) suggested "state-by-state" simulation, which sequentially drew state variables one at a time from t = 1 to t = T. This technique can accommodate non-Gaussian, non-linear class of state-space model but the convergence to the posterior distribution can be very low. For efficiency, this thesis adopts the approach developed by Carter and Kohn (1994) and Fruhwirth-Schnatter (1994), which generates all of the state variables at once. This technique is referred to as "Forward Filtering Backward Sampling" in Fruhwirth-Schnatter (1994). Carter and Kohn (1994) showed that FFBS converges to the posterior distribution faster than the state-by-state approach.

The FFBS algorithm consists of two parts: forward filtering and backward sampling. For each time t, let I_t denote all the information available up to time t. Starting from the initial time period (t = 0) to the end period (t = T), information is updated from

I_{t-1} to $I_t = \{ I_{t-1}, Y_t\}$ every time Y_t is observed. The forward filtering part sequentially updates components in evolution equations using the Kalman filter, obtaining $p(\theta_t, \Sigma_t | Y_t, Y_t^{mis}, I_t)$ for all t = 1, ..., T. After it reaches the last period, we do the backward sampling, moving successively from time t = T to t = 1, simulating sets of states vectors and covariance matrix $(\theta_{T:1}, \Sigma_{T:1})$.

Forward Filtering We now describe the updating process for $\{\theta_t\}$ and $\{\Sigma_t\}$. At time t-1, the posterior distribution of θ_{t-1} and Σ_{t-1} are respectively $(\theta_{t-1}|\Sigma_{t-1}, I_{t-1}) \sim MN(M_{t-1}, C_{t-1}, \Sigma_{t-1})$ and $(\Sigma_{t-1}|I_{t-1}) \sim IW(n_{t-1}, D_{t-1})$, which leads to $p(\theta_{t-1}, \Sigma_{t-1}|I_{t-1}) \propto MNIW(M_{t-1}, C_{t-1}, n_{t-1}, D_{t-1})$, the posterior distribution of (θ_{t-1}, Σ_t) based on the all the information available up to t-1.

At the beginning of time t, we assume the prior distribution of θ_t given (Σ_t, I_{t-1}) to be $MN(a_t, R_t, \Sigma_t)$ and the prior distribution of Σ_t to be $IW(\beta * n_{t-1} - R + 1, \beta * D_{t-1})$. This leads to a joint prior

$$p(\theta_t, \Sigma_t | I_{t-1}) \sim MNIW(a_t, R_t, \beta * n_{t-1} - R + 1, \beta * D_{t-1})$$

where $a_t = M_{t-1}$, is a mean matrix of θ_t,

$R_t = C_{t-1} + W_t$, is a column variance matrix of θ_t,

β is the discounting factor for Σ_t

As the likelihood distribution of time t is assumed to be $(\mathrm{Y_t}|\theta_t, \Sigma_t) \sim N(\theta_t, V_t * \Sigma_t)$, the joint posterior of $(\theta_\mathrm{t}, \Sigma_t)$ at time t follows $\mathrm{p}(\theta_\mathrm{t}, \Sigma_t|I_t) \sim MNIW(M_t, C_t, n_t, D_t)$ since $(\theta_\mathrm{t}|\Sigma_t, I_t) \sim \mathrm{MN}(\mathrm{M_t}, C_t, \Sigma_t)$ and $(\Sigma_\mathrm{t}|I_t) \sim IW(n_t, D_t)$. Each component can be calculated based on the following:

Table 1: Detailed Calculation of Components in the Updating Equations for parameters

Updating Equation for $\theta_\mathrm{t}|\Sigma_\mathrm{t}$	Updating Equations for Σ_t
$\mathrm{M_t} = a_t + A_t * e_t^\mathrm{T}$ $\mathrm{A_t} = \frac{R_t}{q_t}$ $\mathrm{e_t} = Y_t - f_t$ $\mathrm{f_t} = a_t$ $\mathrm{q_t} = R_t + V_t$ $\mathrm{C_t} = R_t - A_t * Q_t * A_t^T$	$\mathrm{n_t} = n_{t-1} + 1$ $\mathrm{D_t} = D_{t-1} + e_t * \frac{e_t^T}{q_t}$

Backward Sampling After obtaining the posterior distribution for the parameters at each time period from 1 to T, we generate $\theta_{\mathrm{T:1}}$ and $\Sigma_{\mathrm{T:1}}$. At time t = T, $(\theta_\mathrm{T}|\Sigma_T, I_T) \sim MNIW(s_T, S_T, n_T, D_T)$, where $\mathrm{s_T} = M_T, S_T = C_T$. Then, for each t from T-1 to 1, sample $(\theta_\mathrm{t}, \Sigma_t)$ from $(\theta_\mathrm{t}, \Sigma_t|I_T) \sim MNIW(s_t, S_t, n_t, D_t)$, where $\mathrm{s_t} = M_t + C_{t+1} * (R_{t+1})^{-1} * C_{t+1} * (\theta_{t+1} - a_{t+1})$ and $S_t = C_t - C_t * (R_t)^{-1} * C_t$.

3.3 Joint Posterior Simulation

Given the previous state of the parameters, $(\theta_t^{(s-1)}, \Sigma_t^{(s-1)}, Y_t^{mis,(s-1)})$, we generate a new state as follows: first, sample $(\theta_t, \Sigma_t)^{(s)} \sim MNIW\left(M_t^{(s-1)}, C_t^{(s-1)}, n_t^{(s-1)}, D_t^{(s-1)}\right)$, where $M_t^{(s-1)}, C_t^{(s-1)}, n_t^{(s-1)}, D_t^{(s-1)}$ are all from the previous state intermediate components from the FFBS algorithm. Then sample $Y_t^{mis,(s)} \sim N\left(\theta_t^{b|a,(s)}, \Sigma_t^{b|a,(s)}\right)$, where $\theta_t^{b|a,(s)}, \Sigma_t^{b|a,(s)}$ are calculated based on the current state of $(\theta_t, \Sigma_t)^{(s)}$. This sequence of state converge to the joint posterior distribution of $(\theta_t, \Sigma_t, Y_t^{mis})$ after a large number of iterations.

4. Simulation

In this chapter, a simulation study is conducted to check the validity and the performance of the algorithm. To mimic as closely as possible the application of the algorithm to the actual stock price data, all of the data used in this chapter was generated based on the stock price index data.

4.1 Simulation Design

The actual stock data with missing values will be notated as ZY, and the simulated data which mimics the actual data is denoted as CY. This section details how CY was simulated.

Using ZY, run the algorithm preliminarily and get $\{\theta_t\}, \{\Sigma_t\}$ and $\{Y_t^{mis}\}$ estimated. Averaged values from the preliminary estimation of parameters, $\{\widetilde{\theta_t}\}$ and $\{\widetilde{\Sigma_t}\}$, are used as the "true" parameters:

$$\theta_{ZY,t} = E\left[E\left[\widetilde{\theta_t}\right]_S\right]_M$$

$$\Sigma_{ZY,t} = E\left[E\left[\widetilde{\Sigma_t}\right]_S\right]_M$$

Using $\theta_{ZY,t}$ and $\Sigma_{ZY,t}$, generate complete data set CY according to the observational Equation (2) in the multivariate dynamic linear model:

$$CY_t = \theta_{ZY,t} + \nu_t, \nu_t \sim N\left(0, V_t * \Sigma_{ZY,t}\right)$$

Initial vague prior was set according to Prado and West (2010) as $m_0 = 0, C_0 = 100, n_0 = 20, D_0 = 20 * diag(R)$. The known multiplicative scalar V_t and column

covariance matrix (in this code, a scalar) W_t are both the absolute values of what are sampled from normal distribution with mean 0 and standard deviation 1. The discounting factor was set as $\beta = 0.99$. S, the number of simulation is 10000 and initial 1000 iteration was taken out as a Burn-in.

ZY and the CY of the first time series chain (S&P 500) are graphically presented in Figure 2. It can be seen that the simulated, complete data CY successfully mimicked the actual data ZY with no missing values in it. We will use this CY as a "complete data" to test how well the imputation algorithm works.

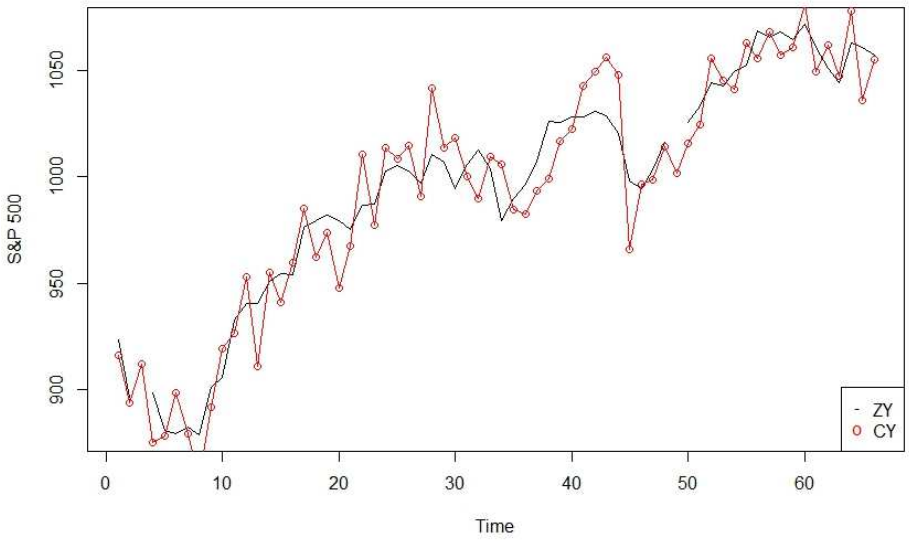

Figure 2: Time series plot of ZY and CY of S&P 500

4.2 Measurement of a performance of the imputation

For an imputation algorithm, the performance of it is assessed based on the closeness of imputed values to true values. In this sense, two error-based measurements were chosen to evaluate the quality of MI: Root Mean Square Error (RMSE), and Normalized Root Mean Square Error (NRMSE).

Throughout this thesis, simulated outputs are denoted with hat, i.e. $\{\widehat{\theta_t}\}$, whereas corresponding true values are denoted without hat. In the model, the sizes of $\mathrm{Y} = \{Y_{1:T}\}$ is $(\mathrm{T} \times \mathrm{R})$ and the size of $\hat{Y}$ is $(\mathrm{T} \times \mathrm{R} \times \mathrm{S} \times \mathrm{M})$, each dimension stands for time (T), the number of time series chain (R), the number of iteration (S), and the number of imputations (M).

First, the error of simulated values is defined as such:

$$\epsilon_Y^{t,r,s,m} = \hat{Y}_{t,r,s,m} - Y_{t,r}$$

$\hat{Y}_{t,r,s,m}$ is $[\mathrm{t}, \mathrm{r}, \mathrm{s}, \mathrm{m}]^{\mathrm{th}}$ element of $\hat{Y}$. Scales of each stock index are all different, so all the measurements are calculated for specific time series chains.

RMSE (Root Mean Square Error)

$$\mathrm{RMSE}_{\mathrm{Y}}^{r} = \sqrt{E\left[E\left[E\left[\left(\epsilon_Y^{t,r,s,m}\right)^2\right]_T\right]_S\right]_M}$$

$\mathrm{RMSE}_Y^{\mathrm{r}}$ is squared-rooted MSE, which is squared errors, averaged across times, iterations and imputations in time series r. Initial 1000 draws are discarded as a burn-in

and according to Rubin's combination rule, averaging across M imputation was done to draw a combined inference

NRMSE (Normalized Root Mean Square Error)

NRMSE is RMSE divided by the range of the values. Unlike MSE and RMSE, it is scale-free and often expressed as a percentage. The lower the NRMSE are, the smaller the degree of deviation from the true value.

$$\mathrm{NRMSE}_Y^{\mathrm{r}} = \frac{RMSE_Y^r}{\max Y_r - \min Y_r}$$

NRMSE_Y^r is $\mathrm{RMSE}_Y^{\mathrm{r}}$ divided by the range of Y_r for each time series.

4.3 Simulation study Result

To test the imputation algorithm, the study checked the overall performance of imputation and the point-wise examination of the in-sample prediction performance of the algorithm.

Overall Trend Trends of CY and simulated data set from the imputation algorithm were compared. Of CY, 20% were made to be randomly missing values, denoted as $\mathrm{CY}^{\mathrm{mis}}$. Using this data set with missing data, the study ran the imputation algorithm M = 5

times, each m^{th} imputation contained S = 10000 iterations. The simulated, complete data set is notated as SY and its size is ($T \times R \times S \times M$). SY is evaluated based on the difference between CY and SY, and the difference between θ and simulated $\hat{\theta}$ from the algorithm.

Presented below are the time series plot of CY and the averaged SY of the first time series chain (S&P 500). The results of the rest of the four time series chains can be found in Appendix A. Averaging in SY is done in element-wise across the iteration after burn-in for each m^{th} imputation of one time series chain, and then averaged again across the M = 5 number of imputations. In Figure 3, it can be seen that SY clearly shows a high degree of co-movement with CY with small errors.

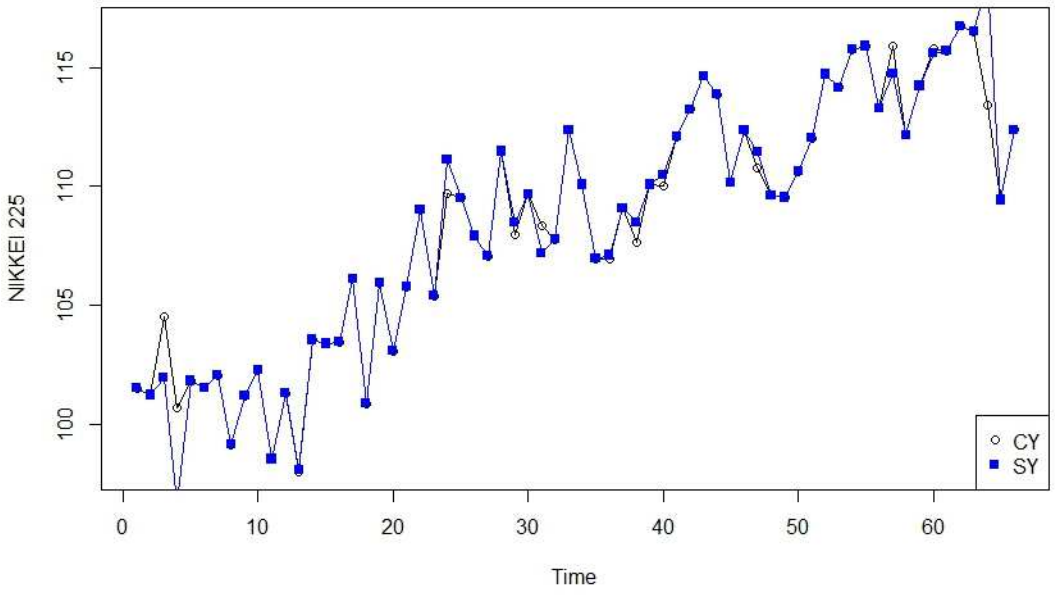

Figure 3: Time series plot of CY and SY of NIKKEI 225

To see the exact error, three error-based measurements were used and are provided in Tables 3 and 4. The following table is the result from a simulation study for each M = 5 imputation of theta and Y. As it can be seen in Table 3 and Table 4, RMSE and Normalized RMSE are both pretty low, indicating that the imputation algorithm works well.

Table 2: Deviation of SY to CY

	S&P 500	**NIKKEI 225**	**Hang Seng**	**FTSE**	**DAG**
$\mathbf{RMSE}_{\mathbf{Y}}^{r}$	2.492	0.486	6.199	11.175	32.207
$\mathbf{NRMSE}_{\mathbf{Y}}^{r}$	0.011	0.025	0.010	0.005	0.014

Table 3: Deviation of LVCF to CY

	S&P 500	**NIKKEI 225**	**Hang Seng**	**FTSE**	**DAG**
$\mathbf{RMSE}_{\mathbf{Y}}^{r}$	9.306	0.927	18.445	92.730	92.101
$\mathbf{NRMSE}_{\mathbf{Y}}^{r}$	0.041	0.049	0.030	0.047	0.040

In-sample predictability In this section, we check the in-sample predictability of our algorithm. The motivation of this study is to cope with the missing values in the stock markets that are closed on holidays. By imputing these missing values, we can more closely guess the opening price for the following day than we can by predicting based on data from the day before holiday. Values from the following day of missing data were inspected as to test the in-sample predictability. For example, the U.S. stock market closes on December 25th for Christmas, so there exists missing values for that day. Our

purpose is to see how closely the imputation algorithm simulated the data on December 26th, assuming that we don't have any information after December 25th. The day of missing values is denoted as time t, which is the holiday: the stock price on Christmas Eve is Y_{t-1}; the stock price on Christmas day is Y_t, which is missing in our original data ZY; finally the stock price on the December 26th is Y_{t+1}, which is assumed to be unknown.

According to our model, we can predict Y_{t+1} based on the observational equation (1) and the evolutional equation (2). If Y_t is missing, Y_{t+1} is predicted based on Y_{t-1}:

$$\begin{aligned} Y_{t+1} &= \theta_{t+1} + \nu_{t+1} \\ &= Y_{t-1} - \nu_{t-1} + \Omega_t + \Omega_{t+1} + \nu_{t+1} \end{aligned} \quad (3)$$

However, if we have Y_t imputed, denoted as Y_t^{im}, Y_{t+1} is predicted as such:

$$\begin{aligned} Y_{t+1} &= \theta_{t+1} + \nu_{t+1} \\ &= \theta_t + \Omega_{t+1} + \nu_{t+1} \\ &= Y_t^{im} - \nu_t + \Omega_{t+1} + \nu_{t+1} \end{aligned} \quad (4)$$

Values were compared in three cases: the "Oracle Truth (OT)," the actual stock price Y_{t+1}, the "Imputed by the Algorithm (IA)," which is the prediction made based on the Y_t^{im} by the algorithm, denoted as Y_{t+1}^{IA} and "Last Value Carried Forward (LVCF)" case, which is the prediction made based on the Y_{t-1}, denoted as Y_{t+1}^{LVCF}. For this simulation study, CY was used again as the complete data set. $CY_{t,i}$ and $CY_{t,i}^{mis}$ are (t, i) elements of CY and CY^{mis}, respectively. For each missing value in CY, three values are assessed through error-based measurements (RMSE and NRMSE) for each time series chain.

Using CY^{mis}, we ran the imputation algorithm M = 5 times, each m^{th} imputation containing S = 10000 iterations and generate SY. Note that $Y_{t,i,s,m}^{im}$ is the $[t, i, s, m]^{th}$ element of SY. If $CY_{t,i}^{mis}$is missing, save $CY_{t+1,i}$ as the "Oracle Truth". If $CY_{t+1,i}^{mis}$ is missing as well, save $CY_{t+2,i}$. For "Imputed by the Algorithm," we calculated the value according to Equation (4):

$$Y_{t+1,i,s,m}^{IA} = Y_{t,i,s,m}^{im} - \nu_{t,i,s,m} + \Omega_{t+1,i,s,m} + \nu_{t+1,i,s,m}$$

Observational errors ν_t and ν_{t+1} are distributed as $N(0, V_t * \Sigma_t)$ and $N(0, V_{t+1} * \Sigma_{t+1})$, respectively in each iteration and each imputation. "Last Value Carried Forward" values can be calculated with Equation (3):

$$Y^{LVCF}_{t+1,i,s,m} = CY_{t-1,i,s,m} - \nu_{t-1,i,s,m} + \Omega_{t,i,s,m} + \Omega_{t+1,i,s,m} + \nu_{t+1,i,s,m}$$

Here, evolution errors Ω_t and Ω_{t+1} are sampled from $MN(0, W_t, \Sigma_t)$ and $MN(0, W_{t+1}, \Sigma_{t+1})$, respectively. For each iteration, and for each imputation, the deviation of "Imputed by Algorithm" and "Last Value Carried Forward" to the "Oracle Truth" is calculated, and then averaged across iterations and imputations.

Table 4: RMSE of "Imputed by Algorithm" and "Last Value Carried Forward" to "Oracle truth"

RMSE	**S&P 500**	**NIKKEI 225**	**Hang Seng**	**FTSE**	**DAG**
IA	10.649	2.582	70.751	192.308	163.764
LVCF	19.321	3.660	82.463	225.009	273.854

Table 5: NRMSE of "Imputed by Algorithm" and "Last Value Carried Forward" to "Oracle truth"

NRMSE	**S&P 500**	**NIKKEI 225**	**Hang Seng**	**FTSE**	**DAG**
IA	0.130	0.290	0.355	0.257	0.135
LVCF	0.236	0.411	0.414	0.301	0.226

As we can see in Tables 4 and 5, IA is smaller than LVCF in both measurements for most of the time series. From this we can see that the imputation algorithm performs

well enough to achieve our original goal, predicting the stock price for the opening price after a holiday.

5. Application

As it is mentioned in the introduction, the motivation of this thesis is to improve prediction on opening stock price index of the day after holiday. We applied the imputation algorithm to the actual stock price index data, which is same as that used in the simulation study.

Stock price data used in this thesis is daily stock index from different stock markets, which are highly correlated to one another. The study chose the S&P 500 (United States, "SPX"), Nikkei 225(Japan, "NEIKKEI"), Hang Seng Index (Hong Kong, "HKHS"), FTSE Index (United Kingdom, "FTSE"), and Deutsche Boerse AG German Stock Index (Germany, "DAG"). These stock indexes are chosen for their size and only one index is selected per each country.

Daily stock index data during July through September of 2009 were collected from Bloomberg, resulting in T = 65 and R = 5. Note that market closings on weekends were not considered as missing in data. The real data, which contains missing values, will be notated as ZY. Approximately 2% of the observation points are missing. In the data we use, not many points are missing due to limited number of holidays, the algorithm shows good performance of imputation up to 20% as we can see in the simulation study result. Correlation between each stock index is provided in Table 6.

Table 6: Correlation between Stock Indexes

	SPX	NEIKKEI	HKHS	FTSE	DAG
SPX	1.000	0.906	0.926	0.985	0.989
NEIKKEI	0.906	1.000	0.875	0.914	0.888
HKHS	0.926	0.875	1.000	0.928	0.920
FTSE	0.985	0.914	0.928	1.000	0.985
DAG	0.989	0.888	0.920	0.985	1.000

Most of settings are same as in the simulation studies in Chapter 4. Initial prior was set according to Prado and West(2010) as $\mathrm{m}_0 = 0, C_0 = 100, n_0 = 20, \mathrm{D}_0 = 20 * diag(R)$. The known multiplicative scalar V_t and column covariance matrix (in this code, a scalar) W_t are both the absolute values of what are sampled from normal distribution with mean 0 and standard deviation 1. The discounting factor was set as $\beta = 0.99$. S, the number of simulation is 10000 and initial 1000 iteration was taken out as a Burn-in and M=5 times of imputations are taken.

Overall Trend In figure 4, the time series plot of Imputed values and original data are presented. During the period, three points are missing and all of them are imputed following the overall trend of original data.

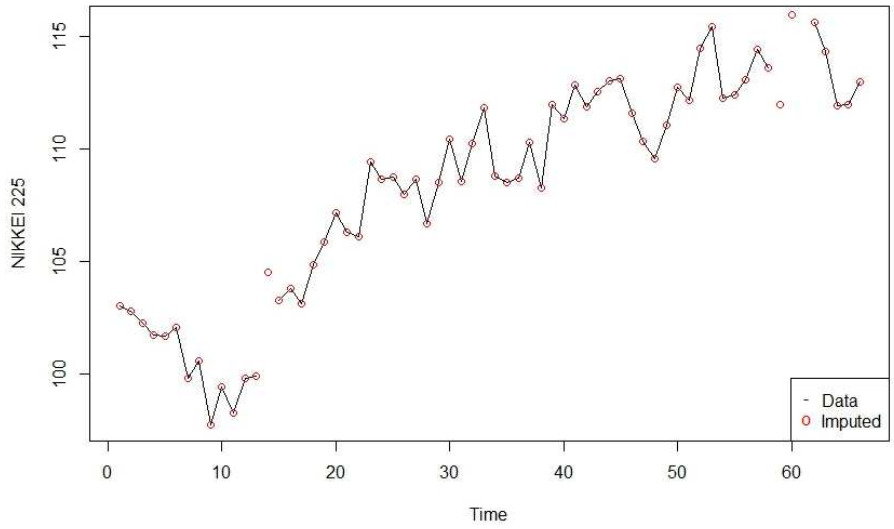

Figure 4: Time series plot of Imputed Values and the Original data

In-sample Predictability In this section, in-sample predictability is presented, as it is done in the simulation studies in section 4.3. Instead of using CY, the real stock price ZY is used. The process is identical to that of simulation studies. Values are compared in three cases as in the simulation: the 'Oracle Truth (OT)', the 'Imputed by the Algorithm (IA)', and 'Last Value Carried Forward (LVCF)' case. For each missing value in ZY, 'IA' and 'LVCF' are assessed through error-based measurements (RMSE, and NRMSE) to the 'OT' for each time series chain. Some of indices, Hang Seng and FTSE have very high errors compared to others. DAG do not have missing values during the period and Hang Seng and FTSE have only one missing values so NRMSE is not available.

Table 7: Errors of "Imputed by Algorithm" and "Last Value Carried Forward" to "Oracle truth"

		Imputed by Algorithm	Last Value Carried Forward
RMSE	S&P 500	1.324	26.388
	NEKKEI	3.838	5.443
	Hang Seng	141.594	157.605
	FTSE	86.355	289.748
	DAG	NA	NA
NRMSE	S&P 500	0.010	0.208
	NEKKEI	0.310	0.440
	Hang Seng	0.495	0.678
	FTSE	0.332	0.498
	DAG	NA	NA

6. Conclusion and Discussion

Stock price changes very fast and continuously reflecting various information on each stock, firm and market environment. Therefore, even a day close of market results in information reflection gap to the opening stock index price. We believe this algorithm help this problem by taking autoregressive structure on prices and correlated information borrowed from other markets into account. As we can see from the simulation study, imputing missing values using the algorithm outperforms to using other methods such as Last Value Carried Forward.

As we discussed previously, this multiple imputation algorithm can be used in many applications that deal with time series with missing values. Financial and economic data and biomedical data seem to be a good candidates since such data are usually highly correlated and exchangeable.

The imputation algorithm is based on the assumption of MAR so the algorithm is not applicable to data that contains systematic missing values. Missing values that exist in financial time series were generated due to a variety of reasons and in many cases, they are not missing at random. For example, for hourly stock data around the world, due to different time zones, some stock data exist as missing when a market is closed whereas on the other side of the world, the stock is traded and the price is observed. This type of missing data is not missing at random. The algorithm studied

here is not applicable to such types of missing data, making this a limitation of this study.

Another issue to overcome is the length of the time series. In this study, the length of the time series was T = 20, which is a quite short for time series, especially for daily data. With longer time series, due to evolution error, the preciseness will become lower. Future research should address these limitations.

Reference

Honaker, J. and King, G. (2010), "What to Do about Missing Values in Time-Series Cross-Section Data", American Journal of Political Science, 54: 561–581. doi: 10.1111/j.1540-5907.2010.00447.x

Hopke, P. K., Liu, C., and Rubin, D. B (2001) Multiple Imputation for Multivariate Data with Missing and Below-Threshold Measurements: Time-Series Concentrations of Pollutants in the Arctic.

Tusell, F. (2012) Multiple imputations of time series with an application to the construction of historical price indices. MS

Baraldi, A.N. and Enders, C.K. (2010), "An Introduction to Modern Missing Data Analyses", Journal of School Psychology, 48: 5-37. doi:10.1016/j.jsp.2009.10.001

Raghunathan T. E., Lepkowski, J. M., Van Hoewyk, J. & Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a series of regression models. Survey Methodol. 27, 85–96.

Raghunathan, T. E. (2006). Combining information from multiple surveys for assessing health disparities. Allgemeines Statist. Archiv. 90, 515–26.

Reiter, J. P. (2004). Simultaneous use of multiple imputation for missing data and disclosure limitation. Survey Methodol. 30, 235–42.

Reiter, J. P. & Raghunathan, T. E. (2007). The multiple adaptations of multiple imputation. J. Am. Statist. Assoc. 102, 1462–71.

Rubin, D. B. (1987). Multiple Imputation for Nonresponse in Surveys. New York: John Wiley & Sons.

Rubin, D. B. & Schenker, N. (1987). Interval estimation from multiply-imputed data: A case study using census agriculture industry codes. J. Offic. Statist. 3, 375–87.

Harrison, J., & West, M. (1999). Bayesian Forecasting & Dynamic Models. Springer.

Cargnoni, C., Müller, P., & West, M. (1997). Bayesian forecasting of multinomial time series through conditionally Gaussian dynamic models. Journal of the American Statistical Association, 92(438), 640-647.

Sims, C. A., & Zha, T. (1998). Bayesian methods for dynamic multivariate models. International Economic Review, 949-968.

Prado, R., & West, M. (2010). Time series: modeling, computation, and inference. CRC Press.