Beta Estimation Using High Frequency Data*

Angela Ryu
Duke University, Durham, NC 27708
April 2011

Faculty Advisor: Professor George Tauchen

Abstract

Using high frequency stock price data in estimating financial measures often causes serious distortion. It is due to the existence of the market microstructure noise, the lag of the observed price to the underlying value due to market friction. The adverse effect of the noise can be avoided by choosing an appropriate sampling frequency. In this study, using mean square error as the measure of accuracy in beta estimation, the optimal pair of sampling frequency and the trailing window was empirically found to be as short as 1 minute and 1 week, respectively. This surprising result may be due to the low market noise resulting from its high liquidity and the econometric properties of the errors-in-variables model. Moreover, the realized beta obtained from the optimal pair outperformed the constant beta from the CAPM when overnight returns were excluded. The comparison further strengthens the argument that the underlying beta is time-varying.

Keywords: Beta estimation, realized beta, high frequency data, market microstructure noise, optimal sampling interval, beta trailing window

JEL Classifications: C51; C58; G17.

*The Duke Community Standard was upheld in the completion of this thesis.
1 Introduction

The beta of a security represents the asset’s sensitivity to movements in the market. The beta is crucial in equity valuation and risk management. It represents the premium to induce risk-averse investors, which is proportional to the risk that cannot be diversified by adding more uncorrelated stocks in the portfolio - the systematic risk. More specifically, the beta represents the systematic risk by measuring the covariation of the portfolio return with the market return. For other things equal, the higher the risk, the more the investors should be compensated, and vice versa. For the beta being a powerful tool in measuring risk with respect to the market, the estimation of the underlying beta has been one of the most important concerns both in academic research and industry practice.

Dating back to Sharpe (1963) and Lintner (1965)’s work on the Capital Asset Pricing Model, the beta is often calculated as the ratio of the market reward-to-risk ratio to an individual stock or portfolio reward-to-risk ratio. For example, a stock with positive beta indicates that it generally moves in the same direction with the market, whereas negative beta indicates that it moves in the opposite direction. When the beta is less than one, the stock moves relatively less than the market whereas with beta greater than one means that the stock price movement surpasses the movement of the market. The assumptions of the Sharpe and Lintner model forces beta to be invariant with respect to the length of the period it is calculated. The constant beta obtained from one-factor CAPM has been widely used in asset pricing and risk management for its simplicity and handiness in measuring the systematic risk of a portfolio.

The CAPM has been under substantial criticism by the extensive number of studies in that the beta is actually time-varying in nature. The criticism originates from the fact that the one-factor CAPM beta does not take account of conditioning variables which may lead the beta to be time-varying. Moreover, in a statistical point of view, the beta - a ratio of time-varying covariance and variance - may be expected to fluctuate with persistency. A number of studies make application to different financial markets to show that the beta with conditioning variables captures the value of systematic risk better than the static CAPM (Faff, Hillier, and Hillier 2000; Choudhry 2002, 2004; Wang 2003) while earlier studies which attempted to use parametrized models in estimating beta (Ferson 1989; Ferson and Harvey 1991, 1993; and Ferson and Korajczyk 1995) have been outperformed by the constant beta.
model. A more recent studies provide mixed evidence on the applicability of the time-varying beta comparatively to the conventional CAPM (Faff and Brooks 2003; Galagedera 2007).

A non-parametric approach using high-frequency data is one of the recent methods utilized to estimate financial measures such as the market volatility. The method utilizes the price data with a very short return horizons, which now became widely available. By using observed variables for calculation, the approach is very handy in that the it trivializes calculation and avoids many distortive assumptions necessary for parametrized modeling. An intuitive approach has been made by Merton (1980), while others, including but not limited to Andersen, Bollerslev, Diebold, and Labys (2003) and Barndorff-Nielsen and Shephard (2004), have worked on the rigorous development of the measures. The realized measures, such as realized variance for instance, are known to be efficient estimators of underlying values like variance, covariance, etc. Within this context, the “realized beta,” defined as the ratio of the realized covariance of stock and market to the realized market variance, was suggested by Andersen, Bollerslev, Diebold and Wu (2006) as a possible estimator of the underlying beta. Unlike the constant period-by-period beta from the CAPM, the realized beta model allows continuous evolution in the beta estimation.

If the observed price data accurately presents the change of the underlying price of the stock, the realized measures are most accurate when the prices are sampled at the highest frequency and therefore all available data should be utilized. However, there exists a lag in adjusting to the price when incorporating new information in the market; the observed stock price does not instantly follow the underlying price. The lag imposes a serious distortion to return estimation particularly when the time interval between price sampling is very short. For example, for the data that record the snapshots of a fluctuating stock price minute-by-minute at 9:35am, at 9:36am, · · · , finally at 4:00pm, the time interval between each price sampling is high 1 minute, i.e. the sampling frequency is at 1 minute. Using data with extremely high frequency sampling incurs high level of noise in return calculation, namely the market microstructure noise.

The market microstructure has been one of the central concerns in leading studies on high frequency data, specifically focusing in regards to the realized variance. In order to minimize the adverse effect of the noise, Ebens (1999), Andersen, Bollerslev, Diebold and Ebens (2001), and Bandi and Russell (2005a) have proposed adjustments by including
filtering whereas other researchers have suggested subsampling (Zhang, Mykland and At-Sahalia 2005: Andersen, Bollerslev and Meddahi 2011), correcting for overnight price changes (Hansen and Lunde 2004b), or using kernel estimators (Hansen and Lunde 2004a; B"anrdorff-Nielsen, Hansen, Lunde and Shephard 2004, 2008b). Andersen, Bollerslev, Diebold and Labys (2006) introduced using the volatility signature plot, a graph of average realized volatility against the sampling intervals, to select a frequency where the average volatility stabilizes. They specifically used 20 minutes for sampling interval as a viable bias-variance tradeoff. The relevant studies generally agree that it is not recommendable to use extremely high frequency sampling like 1 or 2 minutes are not recommended.

The realized beta has been studied relatively to a lesser extent than other realized measures, especially in finding its optimal sampling frequency. On the one hand, because realized beta has realized variance and realized covariance component, the microstructure noise may seriously distort the accuracy of beta estimation and using extremely high frequency should hence be avoided. On the other hand, the market as a whole has significantly greater liquidity than an individual stock, letting the former’s price to adjust more efficiently to the underlying value than the latter. That is, the impact of the noise is expected to be less in market returns, and the optimal sampling frequency for realized beta may not be as low as that of the realized variance. Consequently, the sampling interval that is optimal for beta calculation remains ambiguous. Another critical issue in beta estimation accompanied with the sampling frequency is the optimal span of time interval taken into account for calculation of realized beta, or the beta trailing window. Andersen, Bollerslev, Diebold and Wu (2003) discusses on the persistence of predictability of quarterly and monthly realized beta, but there has not been a study that examines which specific range of beta trailing window, paired with a sampling frequency, gives a relatively better estimation of true beta. In these two aspects, finding the optimal pair of sampling frequency and beta trailing window will add to the current literature on the non-parametrized beta estimation using high frequency data and present evidence on the time-varying nature of beta.

The objective of this study is to empirically determine the pair of sampling frequency and beta trailing window that yields the most precise estimation of underlying beta, and compare the results with the conventional constant beta model. The mean square error (MSE) will be defined and used as a measure of accuracy of beta estimation. Before getting into empirical approach, theoretical background on stochastic models of returns, microstructure noise and
realized beta will be briefly introduced in Section 2. In Section 3, a statistical overview on errors-in-variables will be given in order to justify the possibility of extremely high frequency price data being appropriate to use in beta estimation. In Section 4, realized betas and MSE for each sampling interval and beta trailing window days is calculated and then compared with the conventional model. In the final two sections, analysis and implication of the results are discussed with a brief conclusion.
2 Theoretical Framework

2.1 Stochastic Model of Returns

To begin, a model for stock price movement is defined using continuous-time stochastic model of stock prices. The stochastic differential equation introduced by Merton (1971) presents the evolution of the log-price $p(t)$:

$$dp(t) = \mu(t)dt + \sigma(t)dW(t),$$

where $\mu(t)dt$ represents the time-varying drift of prices and $\sigma(t)$ represents the instantaneous volatility where $W(t)$ is a standard Brownian motion. For the purpose of this study, we assume $\sigma(t)$ and $\mu(t)$ are strictly stationary. Model (1) produces a continuous sequence of prices with probability one; however, we empirically observe discontinuities in the prices, known as “jumps.” Merton (1976) has suggested a revised version of the model that incorporates discontinuities:

$$dp(t) = \mu(t)dt + \sigma(t)dW(t) + \kappa(t)dq(t)$$

where $\kappa(t)dq(t)$ is the jump component; $q(t)$ is a counting process that increments by one with each jump; and $\kappa(t)$ gives the magnitude of each jump. For the purpose of this study, however, we will not distinguish the jump component from the continuous variation term.

2.2 Beta: Capital Asset Pricing Model

The development of the Capital Asset Pricing Model led to the establishment of an appropriate measure of nondiversifiable risk and the market pricing for an individual stock (Sharpe 1964; Lintner 1965). The CAPM suggests that the required return of an asset, $r_i$, is equal to the risk-free rate of return, $r_f$, plus the expected excess return of the market multiplied by beta, $\beta_i$. That is,

$$r_i = r_f + \beta_i \cdot (r_m - r_f).$$

The CAPM displays the compensation to the investors in two terms: time value and risk. The risk free rate represents the time value of money, the rate at which the investors get
compensated as time passes. And the risk measure, beta, indicates how much they should be compensated for taking additional risk. The beta is calculated as the covariance between returns on the risky asset $i$ and the market divided by the market variance:

$$\beta_i = \frac{COV(r_i, r_m)}{Var(r_m)}.$$ (4)

Intuitively, a risk-free asset has a beta of zero since its covariance with the market is zero, whereas a market portfolio will have a beta of one since its covariance with itself equals the market variance. Stock market indices such as S&P 500 are often used as a proxy for the market portfolio.

### 2.3 Realized Beta

Andersen, Bollerslev and Diebold (2003) and Barndorff-Nielsen and Shephard (2002) have suggested using an empirical measure calculated from high frequency intraday data for volatility. As an extension, the realized beta, defined as the ratio of the realized covariance of individual stock and market to the realized market variance (Andersen, Bollerslev, Diebold and Wu 2006) can be used as an effective, non-parametric estimator of the underlying beta. First, we denote $r_{jt}$ as the intraday geometric return of price which is calculated as the difference of log price. For simplicity, consider time $t$ as in days. Then,

$$r_{jt} = \log p(t - 1 + \frac{j}{J}) - \log p(t - 1 + \frac{j - 1}{J})$$ (5)

where $J$ is the number of sampled data within time $t$ with respect to sampling interval $M$ minutes, and $j = 1, 2, 3, \ldots, J$. For example, data sampled at 1 minute from 9:35am to 4:00pm has 385 price data points per day and correspondingly 384 intraday returns; data sampled at 5 minutes - price data at 9:35am, 9:40am, 9:45am, $\cdots$ - has 77 price points per day and 76 intraday returns, and so on. Overnight returns from time $t$ to $t+1$ are calculated as the difference between the opening log price at time $t + 1$, $\log p^o(t + 1)$ and the closing log price of time $t$, $\log p^c(t)$:

$$r_t^{ON} = \log p^o(t + 1) - \log p^c(t).$$ (6)
Because the news during the interday period affects the overnight return, the calculation of financial measures that includes the overnight return may differ significantly from those that exclude the value.

The realized market variance excluding the overnight return at time $t$, $RV_{mt}$, with sampling interval $M$ in minutes is denoted and calculated as the sum of the squared log returns:

$$RV_{mt} = \sum_{j=1}^{J} r_{njt}^2$$  \hspace{1cm} (7)

where $r$ is the log return and $J$ is the number of sampled data in time $t$ with respect to $M$.

In order to obtain the realized variance including overnight returns, the overnight returns are squared and are added to $RV_{mt}$:

$$RV_{mt} = \sum_{j=1}^{J} r_{njt}^2 + (r_{t}^{ON})^2$$  \hspace{1cm} (8)

The realized variance converges to the integrated variance plus the jump component as the time between observations approaches zero. That is, as the sampling interval converges to zero,

$$RV_{mt} \rightarrow \int_{t-1}^{t} \sigma^2(s)ds + \sum_{j=1}^{N_t} \kappa_{jt}^2$$  \hspace{1cm} (9)

where $\int_{t-1}^{t} \sigma^2(s)ds$ is a consistent estimator of the integrated variance and $N_t$ is the number of jumps. Similarly, the realized covariance of stock $i$ and the market is calculated as

$$RCOV_{imt} = \sum_{j=1}^{J} r_{ijt}r_{mjt}$$  \hspace{1cm} (10)

and the calculation of the realized covariance including the overnight return is analogous to the corresponding realized variance. Finally, the realized beta is the ratio of the realized covariance of stock $i$ and market to the realized market variance:

$$\hat{\beta}_{it} = \frac{RCOV_{imt}}{RV_{mt}}.$$  \hspace{1cm} (11)

The realized beta estimation in the above process may be implemented by a simple
regression of individual high-frequency stock returns on that of the market return. Note that because the variance and covariance is treated as observed and not latent, this measure for beta is explanatory rather than predictive. The obtained OLS point estimate presents a consistent estimator of the realized regression coefficient calculated as the ratio of unbiased estimators of the covariance and the market variance. More detailed comments on how the realized beta measure can be a consistent estimator of the integrated beta are found in Andersen, Bollerslev, Diebold and Wu (2006).
3 Data

The calculations were implemented using the price data of 8 stocks: JP Morgan (JPM), Bank of America (BAC), Hewlett-Packard (HPQ), IBM (IBM), Johnson & Johnson (JNJ), Verizon (VZ), Wal-mart (WMT) and Exxon Mobil (XOM). They were specifically chosen for their high market capitalization and the range of representation across industries. For market reference, the price data of S&P 500 (SPY) was used. Each stock data was 1 minute price data from 9:35am to 4:00pm, obtained from an online vendor, price-data.com. Data from August 23, 2004 to January 7, 2009, a total of 1093 days was observed. Stock splits and any other inconsistencies with price data were adjusted.

3.1 Market Microstructure Noise

In a perfectly efficient market where stock prices are adjusted instantaneously to new information, the price of a stock at any point in time should be equal to the sum of all discounted future profits. By the Gordon’s growth model, assuming the required rate of return according to the CAPM is $k$, the constant growth rate, $g$, and the expected profit, $\Pi^e$, the present value of the stock $P$ is calculated as

$$P = \frac{\Pi^e}{(k - g)}.$$  \hspace{1cm} (12)

Assuming that the market determines the price of an individual stock using the Gordon’s growth model, small changes in variables, $\Pi^e$, $k$ and $g$ can significantly impact the resulting price. For example, consider a company with $1 profit per share, with discount rate $k = 10\%$ and constant growth rate $g = 5\%$. According to the growth model, the true value of the stock issued by the company should be

$$P = \frac{1}{10\% - 5\%} = \$20.00.$$  \hspace{1cm} (13)

If the market views the discount rate to be $13\%$ instead of $10\%$, the underlying value expected from the model is $37.5\%$ less than the result from the calculation in (13):

$$P = \frac{1}{13\% - 5\%} = \$12.50.$$  \hspace{1cm} (14)
In reality, an individual stock price do not instantly adjust to the effective price due to market friction. Because small difference in values of variables may seriously divert the observed price from the true price, frequent sampling of price data may not correctly reflect the true value of the stock and is susceptible to noise. This requires particular attention and appropriate selection of data for studies using high frequency stock price data.

The deviation of the observed price from the underlying price is called the market microstructure noise. The noise can be modeled as the difference between the efficient log price, \( p^*(t) \) and the observed log price, \( p(t) \) at time \( t \). That is,

\[
p(t) = p^*(t) + \epsilon(t).
\]

(15)

where \( \epsilon(t) \) is the microstructure noise.

The market microstructure noise is crucial in studying high frequency data. Theoretically, it is possible to obtain arbitrarily accurate estimates of the integrated volatility as the sampling interval converges to 0. However, the realized variance estimators will be severely biased by the market microstructure noise with extremely short sampling interval. Consider \( p(t) \). Since \( p(t) \) is continuous, \( p(t + \Delta t) - p(t) \rightarrow 0 \) as \( \Delta t \rightarrow 0 \). However, because \( \epsilon(t) \) is nonzero and is not continuous, \( \epsilon(t + \Delta t) - \epsilon(t) \) does not converge to 0. That is, the change in the observed price, \( p^*(t) \), is increasingly dominated by the differences in the noise term as the sampling interval decreases. For example, we observe negative serial correlation of the returns due to high level of microstructure noise in jump tests, the tests that are designed to detect and distinguish the jump component from the continuous variance. The existence of the market microstructure noise leaves a serious task to the researchers to select an appropriate sampling frequency when they use high frequency data to estimate financial measures. In this respect, we will now carefully present how high frequency data may be used for better beta estimation.
4 Errors-in-variables

As reviewed in the previous section, the market microstructure noise may seriously distort the return estimation if the price data is sampled at high frequency. For this reason, it is suggested by many studies in high frequency literature to avoid using sampling interval shorter than 5, 10, or even 20 minutes for realized variance calculation. Even when 1 minute data is available, using the whole data set to calculate realized variables is generally uncommon.

Nonetheless, in case of the beta estimation, the line for acceptable sampling frequency remains uncertain due to the properties of errors-in-variables modeling. Consider a simple linear regression model of underlying stock return, $y_i^*$, and the underlying market return, $x_i^*$, with coefficient $\beta$, as the following:

$$y_i^* = x_i^* \beta + u_i^*$$  \hspace{1cm} (16)

where $i = 1, 2, ..., m$. The observed individual stock price, $y_i$, is the underlying stock price plus the noise, $\epsilon_{yi}$:

$$y_i = y_i^* + \epsilon_{yi}.$$ \hspace{1cm} (17)

In other words, the underlying stock price can be expressed as the observed price less the noise:

$$y_i^* = y_i - \epsilon_{yi}.$$ \hspace{1cm} (18)

By the initial model in equation (16) on true prices of the market and the stock price, the observed price of a stock can be expressed in terms of the underlying market price $x_i^*$, the beta, and two noises: the underlying noise $u_i^*$ plus the microstructure noise of the stock, $\epsilon_{yi}$. That is,

$$y_i = x_i^* \beta + u_i^* + \epsilon_{yi}.$$ \hspace{1cm} (19)

Note that the equation (19) is another Ordinary Linear Regression between the underlying market price and the observed individual stock return. The R-square value may drop due
to the additional noise $\epsilon_{yi}$, but the beta coefficient does not change. For simplicity, let $u_i$ denote the noise in the regression between the underlying market price and the observed individual stock price. That is, let

$$u_i = u_i^* + \epsilon_{yi}, \quad (20)$$

then the model becomes,

$$y_i = x_i^* \beta + u_i. \quad (21)$$

While the impact of the noise in the dependent variable does not affect the regression, having noise in the independent variable can potentially incur a serious distortion in beta estimation. To see this problem, the observed market return, $x_i$, is expressed as the sum of the underlying market price and the market noise, $\epsilon_{xi}$:

$$x_i = x_i^* + \epsilon_{xi}, \quad (22)$$

and the underlying market price can be expressed as the following:

$$x_i^* = x_i - \epsilon_{xi}. \quad (23)$$

Plugging in for $x_i^*$,

$$y_i = (x_i - \epsilon_{xi}) \beta + u_i \quad (24)$$

and letting $v_i$ be the sum of noises in the new regression, the regression between the observed stock price and the observed market price is finally obtained:

$$y_i = x_i \beta + v_i \quad (25)$$

where

$$v_i = u_i - \epsilon_{xi} \beta. \quad (26)$$
The latter regression in equation (25) is different from the former in equation (21) in that
now the noise term, \( v_i \), and the independent variable, \( x_i \), are not independent because \( \epsilon_{xi} \) is present in both terms. That is, their covariance is nonzero:

\[
\text{COV}(x_i, v_i) = -\sigma^2_{\epsilon_i} \beta.
\]  

(27)

Because of the negative correlation between the error and the observed stock price, the
estimated coefficient from the regression, or \( \hat{\beta} \), does not converge to the underlying value \( \beta \),
but instead

\[
\hat{\beta} \rightarrow \frac{\sigma^2_{\epsilon_i}}{\left(\sigma^2_x + \sigma^2_{\epsilon_i}\right)} \beta.
\]  

(28)

Stock and Watson (2010) presents detailed derivation process for equation (28). Note that
the ratio \( \frac{\sigma^2_{\epsilon_i}}{\left(\sigma^2_x + \sigma^2_{\epsilon_i}\right)} \) is less than 1, since the variance of the individual stock noise, \( \sigma^2_{\epsilon_i} \), and
the variance of underlying stock price, \( \sigma^2_x \), are both positive. Thus, the estimated beta, \( \hat{\beta} \),
converges to a smaller value than the underlying beta. Especially in case where the noise
level in \( x_i \) is high, the coefficient \( \frac{\sigma^2_{\epsilon_i}}{\left(\sigma^2_x + \sigma^2_{\epsilon_i}\right)} \) is highly biased towards 0 and the estimation may
not accurately capture the true value of the beta. The negative correlation on the regression
erroneously “flats-out” the slope as visualized in Figure 1, where the regressions with and
without the measurement error on 200 sample data are compared. Figure 1-(b) has a lower
slope than that of Figure 1-(a) due to the existence of error.

For the realized beta, however, the distortion due to the bias is expected to be small
because the noise for the observed market price is significantly lower than that of the indi-
vidual stock. This is due to the fact that the market price adjusts more efficiently than
any individual stocks do due to the former’s significantly greater liquidity. That is, the
coefficient of \( \beta \) in the convergence, or \( \frac{\sigma^2_{\epsilon_i}}{\left(\sigma^2_x + \sigma^2_{\epsilon_i}\right)} \), may be close to 1 because \( \sigma^2_{\epsilon_i} \) is very low
relatively to \( \sigma^2_x \). Depending on how much the coefficient is biased towards zero, the beta
estimation using high frequency data, or specifically those at extremely short sampling in-
tervals like 1 minute, may or may not be accurate enough to be used as a viable measure
for the true beta. Hence, by the property of errors-in-variables model, the level of appro-
priate sampling frequency for beta estimation remains ambiguous and should be viewed in
a different context from the realized variance.

In this perspective, it remains uncertain up to which frequency would give a viable
estimation of beta, since the “optimal” sampling frequency paired with the beta trailing window days that gives a relatively better estimation of beta has yet been identified. If the “optimal” pair exists, then the comparison between the beta calculated from the pair and the constant beta from the CAPM will give us a valuable insight in how beneficial the relatively low market noise is on high frequency beta estimation. Following the recent high frequency studies, this study will utilize the non-parametric, empirical approach in finding and examining the “optimal” pair as described in the next section.
5 Methods

The range of the sampling frequency - length of the time interval in between each sampled price data - and the beta trailing window days - the number of days of which data is used in beta calculation - were allowed to vary from 1 to 20 minutes and from 1 to 200 days, denoted as $M$ and $T$, respectively. The longest sampling interval was set at 20 minutes following Andersen, Bollerslev, Diebold and Labys (2003). They suggest that 20 minutes is a “reasonable tradeoff (point) between microstructural bias and minimizing sampling error” (2003) and more scarce sampling with fewer data points may not be sufficient in calculating realized measures. Moreover, the 200-day range was considered to be sufficient in length to cover enough range of data to estimate beta.

5.1 Calculating Realized Beta

Given $M$ sampling minutes and $T$ beta trailing window days, the corresponding realized betas were calculated. Consider calculation excluding overnight returns, and $M = 5$, $T = 30$ for simplicity. First, the market index and a stock price data was sampled at $M$ minutes. Denote the log price data of the individual stock price for day $t$ as $\tilde{X}_t = (x_{t,1}, x_{t,2}, \cdots, x_{t,385})$ where $x_{t,1}$ indicates the first minute data at day $t$ or the data at 9:35am, $x_{t,2}$ the second minute data at 9:36am, $\cdots$, $x_{t,385}$ the last minute data at 4:00pm. Suppose $\tilde{X}_t$ is sampled at $M = 5$ minutes. Then, $\tilde{X}_t = (x_{t,1}, x_{t,6}, x_{t,11}, \cdots, x_{t,381})$ is of length 77. By taking the difference between the log price of consecutive sampling intervals, we obtain the stock return vector for day $t$, $\tilde{r}_{X_t}$, of length 76:

$$\tilde{r}_{X_t} = \left( x_{t,6}, x_{t,11}, x_{t,16}, \cdots, x_{t,381} \right) - \left( x_{t,1}, x_{t,6}, x_{t,11}, \cdots, x_{t,376} \right)$$  \hspace{1cm} (29)

Similarly, the log price data of the market price for day $t$ sampled at 5 minutes is denoted $\tilde{S}_t = (s_{t,1}, s_{t,6}, s_{t,11}, \cdots, s_{t,381})$, and the market return, $\tilde{r}_{S_t}$, is obtained analogously to the stock return:

$$\tilde{r}_{S_t} = \left( s_{t,6}, s_{t,11}, s_{t,16}, \cdots, s_{t,381} \right) - \left( s_{t,1}, s_{t,6}, s_{t,11}, \cdots, s_{t,376} \right)$$  \hspace{1cm} (30)

Then beta was obtained from the regression of the market return and the stock return for each day. Consider day 1, with stock return $\tilde{r}_{X_1}$ and market return $\tilde{r}_{S_1}$. Market return
vector is regressed against the stock return vector using Ordinary Least Squares to find the beta value for day 1:

\[ \beta_1 = (\bar{r}_{S1}' \bar{r}_{S1})^{-1} \bar{r}_{S1}' \bar{r}_{X1} \]  

(31)

where \( \beta_1 \) denotes the value of beta at day 1. By repeating the process for day 2, 3, \( \cdots \), 1093, the realized beta for the whole range is obtained.

Next, denote \( \bar{\beta}_{1:T} \) as the list of betas for day 1, 2, \( \cdots \), \( T \). Letting \( T = 30 \), \( \bar{\beta}_{1:30} = (\beta_1, \beta_2, \cdots, \beta_{30}) \). By shifting a day while keeping the number of days in trailing window constant, \( \bar{\beta}_{2:31} = (\beta_2, \beta_2, \cdots, \beta_{31}) \) is obtained. By repeating the process trailing windows of beta for the window day \( (1094 - T) \) to day 1093, denoted as \( \bar{\beta}_{1094-T:1093} \), is calculated. In total, there are \( 1094 - 30 = 1064 \) trailing windows for \( T = 30 \). For comparison purposes, only the windows from \( \bar{\beta}_{201-T:200} \) to \( \bar{\beta}_{1064:1093} \) is used for each \( T \). It is because the maximum size of trailing window is 200 days and thus, the data for day 201 is the first set that has preceding windows for all \( T \) ranging from 1 day to 200 days.

A series of beta with a specific \((M, T)\) later chosen in the study is compared with the conventional constant beta in Figure 5. We then proceed to determine which pair of \((M, T)\) yields the beta with the lowest mean square error, a measure defined and used in order to evaluate the accuracy of estimation in the following section. Moreover, it is later shown that the set of high frequency betas calculated from the pair where MSE is the lowest is a better estimation of the underlying beta than that of the CAPM.

5.2 Mean Square Error (MSE)

The next step is to calculate square error (SE) and mean square error (MSE) the measures of accuracy in beta estimation of each sampling interval \( M \) and trailing window \( T \). The SE for each \((M, T)\) is calculated as the squared difference between the observed market price and individual stock price multiplied by the corresponding beta. Note that in order to have betas with different \((M, T)\) comparable, it is necessary to fix a specific sampling interval \( M \) for the market index and individual stock price data. In this way, only the values of beta each obtained from the corresponding pair of the sampling frequency and trailing window will vary, and all other variables are kept the same for each SE calculation. Because SE will vary only with beta values from different \((M, T)\), accuracy of the beta estimation for each
pair of \((M, T)\) may be determined by comparing SE values.

The volatility signature plot, a plot of average realized volatility against the sampling interval suggested by Andersen, Bollerslev, Diebold and Labys (2006), is used in order to minimize the additional distortion of beta estimation results due to the microstructure noise in price data. While higher frequency sampling provides more data points to be utilized in estimation and thus gives more precision in general, high level of noise may distort the estimation. The volatility signature plot visualizes the level of volatility in each sampling interval. By looking at the tendency of the volatility, the balancing point with appropriate bias-variance tradeoff may be chosen at the sampling interval at which the level of average volatility sharply drops and stabilizes. For this study, by observing the volatility signature plots of the 8 stock data, the tradeoff points were approximately at 20 minutes and thus \(M\) was set as 20. The volatility signature plot for 5 stocks, BAC, IBM, JPM, WMT and XOM, are presented in Figure 2. That is, given the individual stock, all variables in SE except for the beta use the market index price and stock price all equally sampled at \(M = 20\), denoted as \(S_t\) and \(X_t\) respectively.

After setting up observed market index price \(S_t\) and individual stock price \(X_t\), SE was calculated using the beta from the preceding window for each day starting from day 201, given \((M, T)\). That is, SE at time \(t\) is calculated as the following:

\[
SE_t = (S_t - \hat{\beta}_{t-T:t-1} \cdot X_t)^2.
\] (32)

Recall that \(\hat{\beta}_{t-T:t-1}\) denotes the beta calculated using \(T\) days of data right before day \(t\). \(SE_t\) is a vector of length 19 because \(S_t\) and \(X_t\) are sampled at 20 minutes, and \(201 \leq t \leq 1093\). Then, the average of all SE vectors was taken to get \((1093 - 200) = 893\) scalar values, and taken again to finally get a scalar value, the mean square error:

\[
MSE_{(M,T)} = \frac{1}{(893)} \sum_{t=1}^{893} \left( \frac{1}{19} \sum_{j=1}^{19} SE_{jt} \right).
\] (33)

The MSE will be used as a measure of accuracy for beta estimation obtained from each pair of sampling interval and trailing window, and later for comparison to the constant beta model. The lower the MSE is for an estimated beta, the better we assume that the estimation captures the true value of beta. Note that the MSE is a measure for evaluating
the estimated beta and not for the prediction of the stock returns.

5.3 Comparing MSE: High-Frequency Beta and Constant Beta

Having MSE as the measure of beta estimation accuracy, the pair of sampling interval $M$ and beta trailing window $T$ that yields the lowest MSE was found. Firstly, for a given stock, the pair $(M, T)$ that yields the lowest MSE was found for a given sampling interval, denoted as $\min(M, T)$, a vector of length 20. $\min(M, T)$ is a set of local minimum for each sampling interval ranging from 1 to 20 minutes. Among all pairs of $\min(M, T)$, the pair with the lowest MSE was denoted as $g\min(M, T)$, $g\min$ for “global minimum.” That is, $g\min(M, T)$ is the pair of sampling interval and trailing window that yields the lowest MSE level for a stock.

For comparison purposes with the CAPM beta, the “median pair” of $g\min(M, T)$ of 8 stocks - median value for each $M$ and $T$ - was denoted as $med\min(M, T)$. Further, the level of MSE using constant beta and the level of MSE using beta from $med\min(M, T)$ were compared for each individual stock. For $med\min(M, T)$ beta, MSE was calculated in two ways, one including the overnight return and the other excluding. The daily returns in the first quarter was used to calculate the constant beta. For all three cases, 1) constant, 2) time-varying with overnight returns, and 3) time-varying without overnight returns, MSE was calculated and was compared. The comparison with the constant beta is particularly important in that 1) it provides a valid criterion on whether the empirically found “optimal” realized beta gives a viable estimation of underlying beta and 2) the time-varying nature of beta may be investigated.
6 Results

6.1 Finding the “Optimal” pair of \((M, T)\)

Figure 3 presents a plot of mean square error vs. the trailing window days \(T\), calculated with JPM price data at different sampling intervals. MSE level was very high at extremely short trailing window, as expected, for it had too few data points to estimate beta. For all sampling intervals, MSE was the lowest when \(T\) ranged from 5 to 60 for all sampling intervals. After a short rise starting at roughly \(T = 40\), MSE started to drop again, but it did not recover to the lowest level. After a certain threshold number of days ranging within relatively short period, from 1 week to 6 weeks, increased \(T\) did not enhance the estimation of beta.

Figure 4 presents a comprehensive view in the tendency of MSE levels for all pairs of sampling interval and trailing window days. It shows a 3-dimensional graph of MSE levels of 5 stocks, BAC, IBM, JPM, WMT and XOM, with sampling interval \(M\) on the \(x\)-axis, trailing window days \(T\) on \(y\), and MSE on \(z\). The outlier \(T = 1\) was excluded because as seen in Figure 3 its MSE levels were too high to be considered as an appropriate estimation of beta. The 3-D plot visualizes the resulting \(g_{min}(M, T)\), or the pair of sampling interval and trailing window that yields the global minimum value of MSE for a given stock, listed in Table 1. The values of \(g_{min}(M, T)\) shows that extremely high frequency of 1 or 2 minutes paired with relatively short trailing window as short as one week yielded the lowest MSE. \(med_{min}(M, T)\), the pair of median value for \(M\) and \(T\) of 8 stocks, was found to be (1, 20).

Note that all stocks except for JNJ had their global minimum MSE at 1 minute sampling interval. JNJ had its global minimum MSE at 2 minutes sampling. And the number of days in the trailing window \(T\) that yielded the global minimum ranged from 7 to 36 for the observed stocks, as expected in the tendancy of MSE for given sampling intervals (Figure 3). Again, the “optimal” trailing window days is relatively short, considering that the practitioner like Merrill Lynch uses 60 monthly returns, or five years of price data, for a single beta.

More notably, the most accurate beta estimation using the realized method was found at the highest frequency, which diverts from the general suggestions on “optimal” sampling interval for realized measures made by the current high frequency literature. As carefully analyzed with the errors-in-variables model in Section 4, it is likely that the microstructure
noise distorts the realized beta estimation relatively less than it does other realized measures because of the high liquidity and relatively low level of microstructure noise in the market portfolio. Because the noise factor was less present in the regression, the estimation of beta might have benefited from more data points obtained from high frequency sampling.

6.2 MSE Comparison with Constant Beta Model

Figure 5 presents the “optimal” realized beta at $medmin(M, T) = (1, 20)$ in straight lines and the quarterly beta with daily returns in dotted lines. Overnight returns were not included. Unlike in Andersen, Bollerslev, Diebold and Wu (2006), the betas were plotted with each trailing window shifted for each day instead of for the size of the trailing window $T$, because 1) it gives a more continuous view of time-varying beta estimation and 2) it better represents the series of betas used for the MSE calculation used in this study. The variation of the realized beta was observed to be smaller than that of the quarterly beta. Comparing the overall level of two betas, a downward bias in the realized beta of $medmin(M, T)$ was observed. With overnight returns, the level of variance of “optimal” beta was closer to the quarterly beta, but the systematic bias was not reduced.

Despite the possible existence of bias, the level of MSE calculated with $medmin(M, T)$ beta estimation was lower than with that of the constant beta when the overnight returns were excluded except for BAC (Table 2). When the overnight returns were included, the comparison was inconclusive. Clearly, at empirically found “optimal” sampling frequency along with appropriate trailing window, the realized beta captured the underlying beta with more accuracy than the conventional constant beta model did. The findings that the constant beta was outperformed by the realized beta at appropriate range of sampling interval and trailing window reassure the claim that the underlying beta is time-varying. Nonetheless, our results also indirectly extend Ghysels (1998)’s argument in that the constant beta model may outperform misspecified time-varying beta models, in that the realized beta with overnight returns does not strictly dominate the conventional model in beta estimation. We may cautiously conclude that as long as the range of the trailing window is properly set and the high frequency minute-by-minute data is available, the realized beta excluding the overnight returns gives a good estimation of underlying beta that may outperform the constant beta.
7 Conclusion

The beta has been one of the key financial measures, and various attempts to accurately estimate its underlying value have been made extensively both in academics and among practitioners. Recently, there have been non-parametric approaches in estimating the time-varying beta with high frequency data, denoted as the realized beta. The market microstructure noise, the lag of price adjustment to the true asset price, is known to cause high level of distortion when price data is sampled at extremely high frequencies. For the realized volatility, using 5 minutes or shorter sampling interval is generally discouraged.

However, the results remain ambiguous for the realized beta. Firstly, the realized beta has not been studied extensively especially in the context of finding the “optimal” sampling frequency and the size of the trailing window, the number of days included in beta calculation. Secondly and more importantly, relatively low level of the microstructure noise in the market index leaves the result of regression inconclusive. The econometric property of the beta’s errors-in-variables model shows that depending on the level of the impact on beta estimation by the low market noise, the cutoff for the appropriate sampling frequency may vary. Hence, the realized beta should be studied in a different context from that of the realized volatility when the “optimal” sampling frequency is at hand.

The purpose of this study was to empirically determine the level of “optimal” sampling frequency paired with the corresponding “optimal” trailing window. Having the mean square error as the measure of accuracy, the “optimal” pair of sampling interval and trailing window, \( (M, T) \), was found for 8 stocks. The levels of MSE of 1) the beta estimation from the “optimal” pair and 2) constant beta from the CAPM were compared to check the viability of the result. Surprisingly, extremely short sampling intervals of 1 or 2 minutes and relatively short trailing window from 1 to 5 weeks were found to yield the lowest MSE, that is, found to be “optimal” in beta estimation. For the comparison with the CAPM, the pair of sampling interval and trailing window that yields the lowest MSE, or \( gmin(M, T) \), was found to give a better estimation than the constant beta model if calculated without the overnight returns. With the overnight returns, the realized beta did not have strict advantage over the constant beta.

This study opens up a new point of view towards the microstructure noise and time-varying beta estimation. On the one hand, it reassured that constant beta model, of all
odds it may have, is a reasonable and practical assumption. On the other hand, in case high
frequency price data are available for 5 weeks or more days, an estimation that captures well
the underlying value of beta can be made without any parametrized modeling if sampled at
the highest frequency. The use of extremely short sampling interval is potentially enabled by
the low level of market noise due to its high liquidity in the beta’s errors-in-variables model.
Our results present that as long as the beta and its realized estimation are concerned, we
can fully take advantage of the price data available by minutes without microstructure noise
distorting the result.
References


### Tables

#### Table 1: $gmin(M, T)$

<table>
<thead>
<tr>
<th>Stocks</th>
<th>$gmin(M, T)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>BAC</td>
<td>(1, 7)</td>
</tr>
<tr>
<td>HPQ</td>
<td>(1, 29)</td>
</tr>
<tr>
<td>IBM</td>
<td>(1, 16)</td>
</tr>
<tr>
<td>JNJ</td>
<td>(2, 20)</td>
</tr>
<tr>
<td>JPM</td>
<td>(1, 16)</td>
</tr>
<tr>
<td>VZ</td>
<td>(1, 17)</td>
</tr>
<tr>
<td>WMT</td>
<td>(1, 13)</td>
</tr>
<tr>
<td>XOM</td>
<td>(1, 23)</td>
</tr>
</tbody>
</table>

#### Table 2: MSE Comparison (All MSE levels: $\times 10^{-4}$)

<table>
<thead>
<tr>
<th>Stocks</th>
<th>MSE (constant)</th>
<th>MSE (varying, w/ overnight)</th>
<th>MSE (varying, w/o overnight)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BAC</td>
<td>0.0369</td>
<td>0.2032</td>
<td>0.0491</td>
</tr>
<tr>
<td>HPQ</td>
<td>0.1005</td>
<td>0.0573</td>
<td>0.0529</td>
</tr>
<tr>
<td>IBM</td>
<td>0.0360</td>
<td>0.0497</td>
<td>0.0229</td>
</tr>
<tr>
<td>JNJ</td>
<td>0.0443</td>
<td>0.0643</td>
<td>0.0179</td>
</tr>
<tr>
<td>JPM</td>
<td>0.2579</td>
<td>0.0882</td>
<td>0.0456</td>
</tr>
<tr>
<td>VZ</td>
<td>0.0555</td>
<td>0.0475</td>
<td>0.0410</td>
</tr>
<tr>
<td>WMT</td>
<td>0.0783</td>
<td>0.0597</td>
<td>0.0555</td>
</tr>
<tr>
<td>XOM</td>
<td>0.0379</td>
<td>0.0461</td>
<td>0.0342</td>
</tr>
</tbody>
</table>
B Figures

Figure 1: Impact of the Errors-in-Variables on OLS
Figure 2: Volatility Signature Plot
Figure 3: MSE vs. T for JPM
Figure 4: 3-D Plot of MSE
Figure 5: “Optimal” Realized Beta at $medmin(M, T) = (1, 20)$ vs. Quarterly Beta