

On Intraday Shanghai Stock Exchange Index

Hua Wang¹, Yan Yu² and Min Li³

¹*Alibaba Group*, ²*University of Cincinnati* and

³*California State University, Sacramento*

Abstract: This paper investigates the return, volatility, and trading on the Shanghai Stock Exchange with high-frequency intraday five-minute Shanghai Stock Exchange Composite Index (SHCI) data. The random walk hypothesis is rejected, indicating there are predictable components in the index. We adopt a time-inhomogeneous diffusion model using log penalized splines (log P -splines) to estimate the volatility. A GARCH volatility model is also fitted for comparison. A de-volatilized series are obtained by using the de-volatilization technique of Zhou (1991) that resample the data into different de-volatilized series with more desired properties for trading. A trading program based on local trends extracted with a State Space model is then implemented on the de-volatilized five-minute SHCI return series for profit. Volatility estimates from both models are found to be competitive for the purpose of trading.

Key words: GARCH, high-frequency data, intraday volatility, penalized splines, random walk, state space model, trading.

1. Introduction

There is a vast literature analyzing emerging stock markets. However, there is very limited research on the Shanghai Stock Exchange, especially using intraday (five-minute) data. It is well known that emerging equity markets like the Shanghai Stock Exchange have very different characteristics than developed equity markets. Compared with developed markets, emerging markets usually have higher average returns, low correlations with developed market returns, more predictable returns, and higher volatility.

Bekaert and Harvey (1997), Aggarwal, Inclan and Leal (1999), and Santis and Imrohorglu (1997) have analyzed emerging market volatility. However, they use daily (or lower frequency) data to study the relationship between the volatilities of emerging equity markets and other developed equity markets. The Chinese equity market is not included in their analysis. Mookerjee and Yu (1999) test

Shanghai and Shenzhen stock markets efficiency using daily stock price data. Lee, Chen and Rui (2001) examine time-series features of returns and volatility of daily stock indices from different Chinese stock markets. They apply GARCH-type models and find evidence that volatility is time-inhomogeneous, persistent, and predictable. Wei (2002) models China's weekly stock market volatility with the GARCH model and two of its nonlinear modifications. Zhou and Zhou (2005) examine stock returns, volatility, and cointegration among three Chinese stock markets before and after Hong Kong's return to China with the daily return data. Tian and Guo (2007) study the interaction between trading procedures and security price formation by examining the interday and intraday return volatility of the Shanghai Stock Exchange Composite Index.

In this paper, the intraday return volatility of the five-minute Shanghai Stock Exchange Composite Index (SHCI) is estimated using GARCH and log penalized splines (log P -splines). We (Yu, Yu, Wang and Li 2008, hereafter, YYWL) developed log P -splines estimation for a time-inhomogeneous volatility diffusion model framework as in FJZZ (Fan, Jiang, Zhang and Zhou 2003), which include many common diffusion models as a special case. With different volatility estimates, the de-volatilization technique (see Zhou 1991; Zhou 1996) is used to resample the data into different de-volatilized series for trading. The intraday high-frequency data like the five-minute SHCI are heteroscedastic time series with high volatility. The de-volatilization technique by Zhou (1991) removes heteroscedasticity from the original series by resampling it into a new homoscedastic and normalized time series that are more suitable for modeling trends and trading. We find that the volatility estimates from log P -splines can pick up fewer data points in a flat market but more data points in a volatile market. Instead of using statistical criteria to test which volatility model best captures the heteroscedasticity and time-inhomogeneous structure of five-minute SHCI in a certain period, a simple trading program based on local trends extracted with a State Space model is implemented on the de-volatilized series to compare the accuracy of the volatility estimates. The volatility estimates from GARCH and log P -splines are shown to be competitive for the purpose of trading.

The remainder of the paper is organized as follows. In Section 2, we find that there appear to be predictable components in the intraday return based on the rejection of the random walk hypothesis. In Section 3, intraday return volatility is estimated using GARCH and log P -splines. Log P -splines are discussed in detail. Section 4 describes the de-volatilization technique by Zhou (1991). A trading program based on de-volatilized series is provided in Section 5. Section 6 concludes the paper.

2. Intraday Return Analysis

The Shanghai stock market is an emerging equity market launched in 1990 and the largest stock exchange in China. As of December 2008, 864 companies are listed on the exchange with a market capitalization of over 10,000 billion RMB (US \$ 1,453 billion). The exchange is open for trading from Monday to Friday with two trading sessions each day, one in the morning from 9:30 AM to 11:30 AM and the other in the afternoon from 1:00 PM to 3:00 PM. Trades are executed through computerized trading systems without market makers.

2.1 Data and descriptive statistics

SHCI is a weighted-average market-capitalization index. Its base date is December 19, 1990 and its base value is 100. The intraday data we used for analysis are provided by Guangdong Min An Securities Company Ltd. in Guangzhou, China. The available five-minute data start on April 18, 2001 and end on October 15, 2001, representing a total of 5,616 observations. There are 48 data points on each day: 24 data points are recorded in the two-hour morning trading session with the first data point taken at 9:35 AM; the other 24 data points are recorded in the afternoon trading session with the last data point taken at 3:00 PM. In order to make it easier to derive the time-series properties of additive processes than of multiplicative processes, we follow Campbell, Lo, and MacKinlay (1997) to calculate the continuously compounded percentage returns:

$$dS_t = 100 \times (S_t - S_{t-1}) = 100 \times [\log(x_t) - \log(x_{t-1})], \quad (2.1)$$

where x_t is the five-minute SHCI at time t , $S_t = \log(x_t)$, and dS_t is the continuously compounded every-five-minute logarithmic percentage return at time t .

The five-minute SHCI over the sample period are plotted in Figure 1. Some summary statistics on the return data are obtained. The five-minute SHCI has a small negative average return of about one in two hundredth of a percent per five minutes in the particular period. The skewness coefficient 0.60643 indicates that the return distribution is substantially positively skewed. Furthermore, the kurtosis, a measure of the thickness of the tails of the distribution, is very high with a value of about 41, much more than a Gaussian distribution kurtosis of 3. The Kolmogorov-Smirnov test for normality of the return distribution provides a statistic of 0.13298, thus rejecting the null hypothesis of normal distribution of the five-minute SHCI at 1% significant level. The Jarque-Bera test also rejects the null hypothesis of normality with a p-value of 0.

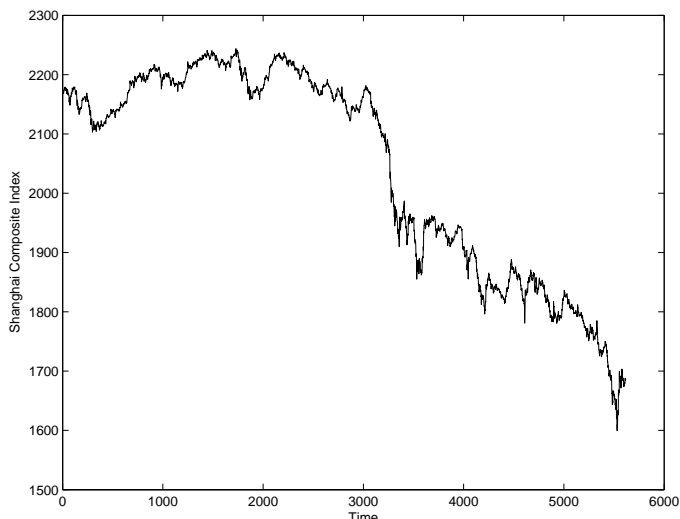


Figure 1: The Five-minute Shanghai Stock Exchange Composite Index, April 18, 2001 to October 15, 2001. The Shanghai Composite Index is the original market index level. Time is the sequence number of data points by time with five minutes between every two data points.

2.2 Random walk

We apply the variance-ratio test of Lo and MacKinlay (1988) to investigate whether SHCI follows a random walk at a market microstructure level. The results in Table 1 show that the random-walk null hypothesis on the five-minute SHCI is rejected at all usual significance levels with homoscedastic-consistent test statistics. With heteroscedastic-consistent test statistics $Z_k(q)$, which is considered to be more appropriate for the five-minute SHCI data, the random-walk null hypothesis is rejected at 5% significance level at most lags except at lag 2.

Table 1: Variance ratios for five-minute Shanghai stock exchange composite continuously compounded returns.

	Number of lags				
	2	4	8	16	32
Variance ratio	1.21	1.14	1.20	1.28	1.42
$z(q)$	15.73***	2.99***	5.67***	4.77***	5.31***
$Z_h(q)$	1.52	4.11***	2.21**	2.39***	2.69***

*** 1% significant level; ** 5% significant level; * 10% significant level.

$z(q)$ is the asymptotic normal test statistic under homoscedasticity; $Z_h(q)$ is the asymptotic normal test statistic under heteroscedasticity.

The Shanghai stock market returns are predictable to some degree on daily or weekly frequency (see Darrat and Zhong 2000). The variance-ratio tests above also confirm that the Shanghai stock market returns are also predictable to some degree on intraday frequency. This finding supports our motivation to develop some models to identify this predictability and try to outperform the market index in the Shanghai stock market.

3. Time-inhomogeneous Volatility Estimation

3.1 GARCH

To estimate time-varying return volatility, the parametric GARCH (Generalized Autoregressive Conditional Heteroscedasticity) model of Bollerslev (1986) is first employed. As what was shown by Engle and Pattern (2001) with Dow Jones Industrial Index, GARCH is successful on modeling the conditional volatilities and capturing some important stylized facts such as persistence and mean-reversion on time series of asset returns. Given the continuously compounded percentage return $dS_t = 100 \times [\log(x_t) - \log(x_{t-1})]$ where x_t is the five-minute SHCI at time t and $S_t = \log(x_t)$, the GARCH(p, q) model has the following form:

$$\begin{aligned} dS_t &= \beta_0 + \beta_1 + \epsilon_t \\ \epsilon_t &= \sqrt{h_t} e_t \\ h_t &= \phi + \sum_{j=1}^q \alpha_j \epsilon_{t-j}^2 + \sum_{j=1}^p \gamma_j h_{t-j}, \end{aligned} \quad (3.1)$$

where e_t is independent and identically-distributed with mean 0 and variance 1 (following either the standard normal or standardized Student t-distribution), $\alpha_i, \gamma_j, \phi, \beta_0$ and β_1 are the parameters. The model can be estimated by maximum likelihood and Schwarz's Bayesian Information Criterion (BIC) is chosen to determine p and q . It is found that GARCH(1,2) is relatively efficient on fitting the five-minute SHCI returns and thus the setting of $p = 1$ and $q = 2$ is chosen. The conditional volatility estimates from this GARCH(1,2) model are used in Section 4 to obtain de-volatilized series.

3.2 Log P -splines

Financial market conditions change as time passes by and it is reasonable to expect the return and volatility of SHCI depend on both time and the underlying asset price level. We apply the semiparametric log P -splines time-inhomogeneous diffusion model we developed recently (see YYWL 2008) to estimate volatility. Log P -splines maximize penalized likelihood with an iterative algorithm and give

very competitive volatility estimates compared with GARCH.

The log P -splines time-inhomogeneous diffusion model maximizes penalized likelihood with the smoothing parameter selected by generalized cross validation (GCV) and the EBBS of Ruppert (1997). The time-inhomogeneous diffusion model takes the following form:

$$dS_t = [\alpha(t) + \beta(t)]dt + \sigma(t)S'_\gamma dW_t \quad (3.2)$$

where $S_t = \log(x_t)$, γ is a scalar parameter independent of time t , $\alpha(t)$ and $\beta(t)$ are time-dependent coefficients of the drift $\alpha(t) + \beta(t)$, $\sigma(t)$ is a time-dependent coefficient of volatility, and W_t is the standard Brownian motion. This model encompasses most well-known diffusion models, such as the CKLS (Chan, Karolyi, Longstaff and Sanders 1992) model where $\alpha(t)$, $\beta(t)$ and $\sigma(t)$ are constants or time-independent. When $\gamma = 1$, the CKLS model corresponds to the famous Black-Scholes Geometric Brownian Motion option pricing model (Black and Scholes 1973). When $\gamma = 0.5$, the CKLS model corresponds to the CIR (Cox, Ingersoll and Ross 1985) model. When $\gamma = 0$, the CKLS model corresponds to the Vasicek (1977) model. FJZZ (Fan, Jiang, Zhang and Zhou 2003) consider a more general model with depends on time t but they note possible over-parameterization and unreliable estimates due to high collinearity. Our log P -splines time-inhomogeneous diffusion model adopted here treats γ as a parameter (time-independent). It has the advantage of allowing testing parametric restrictions corresponding to the model that fit the data adequately using formal tests such as the Wald test. YYWL also present the local log-linear method maximizing kernel-weighted likelihood with bandwidth selected by the Rule-of-thumb. Estimation results from log P -splines and the local log-linear method were found to be similar. Below we describe the estimation algorithm of log P -splines to estimate volatility.

For SHCI five-minute data, a discretized version of the semiparametric time-inhomogeneous model (3.2) is based on the Euler scheme for approximation:

$$\Delta S_t = [\alpha(t) + \beta(t)S_t]\Delta t + \sigma(t)S'_\gamma \sqrt{\Delta t}\epsilon_t, \quad (3.3)$$

where $\Delta S_t = 100 \times (S_t - S_{t-1})$, $\Delta t = 1/(252 \times 48)$ (252 trading days per year and 48 five-minute data points per day), and ϵ_t are independent and identically distributed as standard normal $N(0, 1)$. The time-dependent components are estimated with P -splines, and $\sigma(t)$ is estimated with log P -splines so that positive volatility is naturally embedded: $\alpha(t) = B_\alpha(t)\delta_\alpha$, $\beta(t) = B_\beta(t)\delta_\beta$, $\log \sigma^2(t) = 2B_\sigma(t)\delta_\sigma$, where $B(t)$ is a vector of spline basis functions (either truncated power basis or B-spline basis) and are vectors of spline coefficients. The log likelihood function, excluding constants, is negative

$$\sum (\Delta t^{-1} \{ \Delta S_t - (B_\alpha(t)\delta_\alpha + B_\beta(t)\delta_\beta S_t)\Delta t \}^2 \exp\{-(2B_\sigma(t)\delta_\sigma + \gamma \log S_t^2)\})$$

For notational consistency, we reserve the subscript 1 for drift $(\alpha(t) + \beta(t)S_t)$ and 2 for volatility $\sigma(t)$: parameter vectors $\delta_1 = (\delta_\alpha^T, \delta_\beta^T)^T$ for drift and $\delta_2 = (\delta_\sigma^T, \gamma)^T$ for volatility. Write the extended design matrix for drift as $B_1(t) = [B_\alpha(t), B_\beta S_t]$ and the extended design matrix for volatility as $B_2(t) = [B_\sigma(t), \log S(t)]$. Further denote the parameter vector by $\theta = (\delta_1^T, \delta_2^T)^T = (\delta_\alpha^T, \delta_\beta^T, \delta_\sigma^T, \gamma)^T$.

The smoothing parameter vectors are $\lambda = (\lambda_\alpha, \lambda_\beta, \lambda_2)^2$ and $\lambda_1 = (\lambda_\alpha, \lambda_\beta)^T$, where $\lambda_\alpha, \lambda_\beta$ and λ_2 are smoothing parameters for $\alpha(t), \beta(t)$ and $\log \sigma^2(t)$, respectively. The penalized likelihood estimator of θ maximizes the penalized log likelihood function $Q_{N,\lambda}(\theta) = L_N(\theta) - (N/2)\lambda\theta^T D\theta$, where $N = 5,615$ and

$$\begin{aligned} L_N(\theta) &:= \sum \ell_N(\theta, t) \\ &= - \sum (\Delta t^{-1} \{\Delta S_t - B_1(t)\delta_1 \Delta t\}^2 \exp\{-2B_2(t)\delta_2\} + 2B_2(t)\delta_2). \end{aligned}$$

Here D is an appropriate positive semi-definite matrix that penalizes jumps at the knots in the p th derivative of the spline, if truncated power basis is adopted. The two-step estimation algorithm as in YYWL is implemented to obtain volatility estimates for SHCI:

Step 1: Drift Estimation.

The time-inhomogeneous drift $\mu(t, S_t) = \alpha(t) + \beta(t)S_t$ is estimated by minimizing

$$\sum \{(\Delta S_t / \Delta t) - B_1(t)\delta_1\}^2 + (N/2)\lambda_1 \delta_1^T D_1 \delta_1$$

This can be achieved by a simple ridge regression $\hat{\delta}_1 = (B_1^T B_1 + N\lambda_1 D_1)^{-1} B_1^T (\Delta S_t / \Delta t)$, where D_1 is a diagonal matrix, and is a spline basis function. The smoothing parameter vector can be chosen by criteria such as GCV or EBBS.

Step 2: Log P -splines Volatility Estimation.

Denote the residual from the previous drift estimation by $e_t = \Delta t^{-1/2}(\Delta S_t - \hat{\mu}(t, S_t)\Delta t)$. Then we have $e_t \approx \sigma)t^\gamma \epsilon_t$, and estimate the parameter for volatility by minimizing the negative penalized likelihood

$$\sum (E_t^2 \exp\{-2B_2(t)\delta_2\} + 2B_2(t)\delta_2) + (N/2)\lambda_2 \delta_2^T D_2 \delta_2$$

The nonlinear optimization routine *lsqnonlin*() from Matlab's optimization toolbox is used. A preliminary parameter estimate $\hat{\delta}_{2,pre}$ for volatility can be

obtained by a simple ridge regression $\hat{\delta}_{2,pre} = (B_2^T B_2 + N\lambda_2 D_2)^{-1} B_2^T E$, where vector is an element in $\log |e_t|$. The volatility estimate is given by $\hat{\sigma}(t, S_t) = \sqrt{\sigma^2(t) S_t^{2\hat{\gamma}}}$.

We perform this two-step estimation algorithm on the five-minute SHCI data. Both $\hat{\sigma}(t)$ and the volatility estimates $\hat{\sigma}(t, S_t)$ are obtained. The volatility estimates from both GARCH and log P -splines are used below for de-volatilization.

4. De-volatilization

The financial markets are known to be very volatile. The volatilities in the emerging equity markets are even higher. Fundamental economic facts and market trends are often buried in strong noise and are very difficult to be detected. Heteroscedasticity makes things even worse. To utilize the traditional models to efficiently model volatile high-frequency heteroscedastic time series, such as foreign-exchange rates, Zhou (1991) proposed a technique, de-volatilization, to remove heteroscedasticity from the time series of interest based on his naive and Bayesian volatility estimators. His de-volatilization technique was proved to be capable of resampling the heteroscedastic time series (high-frequency foreign-exchange rates) into a new homoscedastic and normalized time series that can be used more efficiently and accurately for forecasting trends and trading.

Based on the volatility estimates from GARCH and log P -splines, the de-volatilization technique described below is implemented to obtain two de-volatilized series S_d from the five-minute logarithmic SHCI S_{n_i} where i represents one data point at time i on a day n :

1. Let initial value $S_0 = S_{n_i}$.
2. Suppose we have obtained ν -th data S_d in d -series.
3. Estimate the cumulative volatility process $V(\Delta S_k)$ from GARCH or log P -splines:

$$V(\Delta S_k) = \sum_{i=1}^k \hat{\sigma}_i^2 \quad (4.1)$$

where $\hat{\sigma}_i^2$ is the estimated variance of the five-minute SHCI returns and ΔS_k is the logarithmic five-minute SHCI return at time k .

1. Include next S_{n_k} into the d -series as soon as cumulative volatility $V(\Delta S_k)$ cross next τ level, i.e., let $S_{d+1} = S_{n_i}$ such that $k = \min\{i : V(\Delta S_i) > (d+1)\tau\}$.

2. Repeat step 2 until the end of series S_{n_i} and a d -series S_d is found.

We choose $\tau = 1$ and provide two d -series. They have similar sizes to that of daily series so that those d -series may be compared later. Two d -series are plotted in Figure 2, based on GARCH and log P -splines volatility estimates. The d -series have more data points in the volatile period but have fewer data points in the smooth period. The de-volatilized series based on log P -splines volatility estimates have nice property of obviously picking up fewer data points when the market is flat while obviously including more data points when the market is volatile, indicating that we can avoid unnecessary trading if we use the time-dependent log P -spline d -series to trade when the market does not change dramatically.

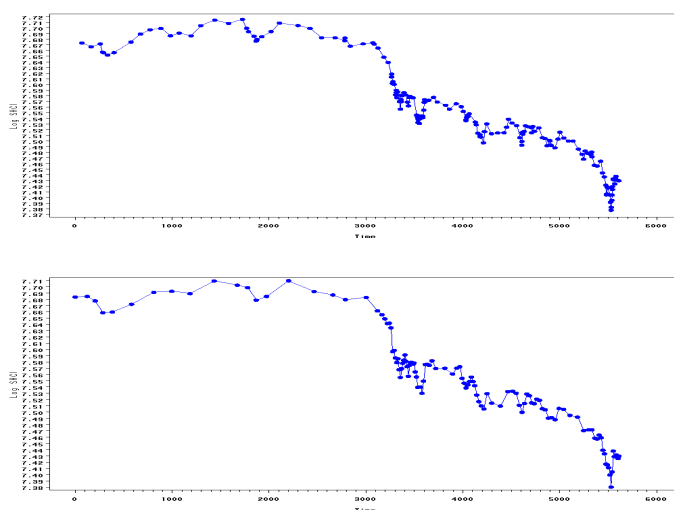


Figure 2: D -series of Five-minute SHCI Based on Volatility Estimates from: GARCH with 173 Data Points (upper), log P -splines with 118 Data Points (lower).

Zhou (1991) illustrates that the d -series of high-frequency foreign-exchange rates have attractive properties such as normality, homoscedasticity, and independence. The new d -series will enhance the capability of traditional time series models to produce more accurate forecasts. If the local trends can be accurately found as the deterministic part of the market index, naturally they can be used to forecast market index and then trade.

5. Local Trend Extraction and Trading

The State Space model or the Dynamic Linear model, which was originally proposed as a method primarily for use in aerospace-related research, was intro-

duced by Kalman (1960) and Kalman and Bucy (1961). The State Space model has been applied to many economic data by various researchers, such as Shumway and Stoffer (1982, 2000) and Harvey and Pierse (1984). The State Space form has been shown to be a powerful tool that opens the way to handling a wide range of time series models. Once a model is put in a State Space form, the Kalman filter may be applied and this in turn leads to algorithms for smoothing and prediction.

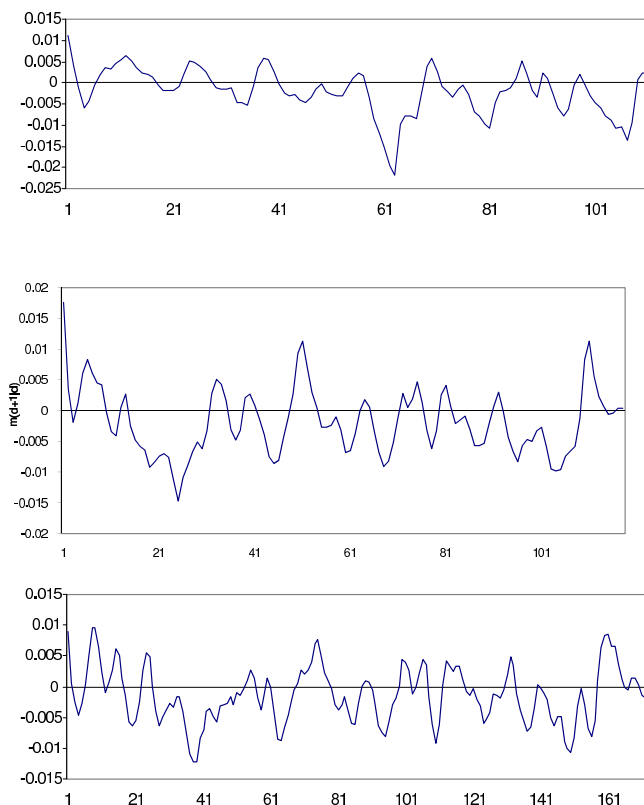


Figure 3: Extracted local trend $m_{d+1|d}$: Daily Series, GARCH D -series, and log P -splines D -series (from top to bottom).

Based on the framework of State Space model of Harvey (1991) and the application of Zhou (1991) to high-frequency foreign-exchange rates, a State Space model is set up to extract the unobservable local trend, say from the five-minute SHCI d -series as follows:

$$S_d = a_d + \phi_d \quad (5.1)$$

$$a_d = a_{d-1} + m_{d-1} + \xi_d, \quad (5.2)$$

$$m_d = m_{d-1} + \Psi_d, \quad (5.3)$$

where S_d is the observed logarithmic five-minute SHCI d -series and $[a_d, m_d]^T$ is the state vector. The disturbance ω_d is essentially set as a constant, zero, in this State Space form model. Furthermore, ξ_d and ψ_d are white-noise disturbances with zero means and variances σ_ξ^2 and σ_ψ^2 respectively. Both disturbances are independent to each other.

The State Space form above is applied to the actual daily SHCI series (April 18, 2001 - October 15, 2001, 111 trading days), GARCH d -series and log P -spline d -series of S_d . The daily series are low-frequency data with less noise than the high-frequency five-minute SHCI series. The predicted local trends $m_{d+1|d}$ in those three series extracted from the State Space models are plotted in Figure 3.

In Figure 3, even with the same settings in the State Space model, we can observe that the forecasts of GARCH and log P -splines d -series fit better than those of the daily series. Considering the assumptions of the State Space model, we can conclude that d -series based on GARCH and log P -splines volatility estimates from the intraday SHCI series have more goodness of normality and independence.

Since $m_{d+1|d}$ is considered to be the predicted slope of the logarithmic five-minute SHCI series, the simple trading program proposed by Zhou (1991), in which foreign-exchange rates were shown to be successfully traded with excess returns, can be applied to trade in Shanghai stock market. The trading program is set up as follows:

1. At time d
 - a. Long or keep long in a SHCI portfolio, if $m_{d+1|d} > 0$.
 - b. Short or keep short in a SHCI portfolio, if $m_{d+1|d} < 0$
2. Always on one position with fixed amount of fund. Transaction cost c is set to be 0.75%.
3. If in a transaction round buy at time $d + i$ and sell at time $d + i + j$, the profit is:

$$\begin{aligned} \text{profit} &= \exp(S_{d+i+j} - S_{d+i}) - 1 - c \quad \text{if buy long} \\ &= \exp(S_{d+i} - S_{d+i+j}) - 1 - c \quad \text{if sell short} \end{aligned} \quad (5.4)$$

Note: Even though selling short was not allowed in 2001, recently (2008) China lifted shorting restriction.

The profit/loss based on GARCH or log P -splines d -series is much better. The profit/loss based on GARCH model is 21.27% higher than the profit/loss based on the daily series. While the profit/loss based on log P -splines is 17.41% higher than the profit/loss based on the daily series. Even when the transaction

costs are included, the trading results with the trading program applied on all three series are still very favorable. The results are listed in Table 2 and Table 3. With the log P -splines d -series, the profit/loss before transaction fees is 4% less than that with the GARCH d -series. At the same time, we only need to trade 23 times with the log P -splines d -series but need to trade 28 times with the GARCH d -series. The profit/loss after transaction fees with the log P -splines d -series (72.61%) is essentially as good as that with the GARCH d -series (72.72%). As we mentioned before, this is due to the nice property of time-dependent log P -splines volatility estimates that allows us to pick up only a few observations through time when applying the de-volatilization scheme in a flat market. This is meaningful when trading in the emerging equity markets because by trading fewer times the high transaction costs in those markets can be avoided and the profit in a long-term prospect can be improved.

Table 2: Profit/loss not including transaction costs

Time Series	Daily Series	GARCH D-series	PS D-series
Profit/Loss without Transaction Cost (%)	72.45	93.72	89.86
Difference (%) with the Profit/Loss of $\exp(\text{first}_{S_d} - \text{last}_{S_d}) - 1 = 28.90\%$	43.55	64.82	60.96
Difference (%) with the Profit/Loss of $\exp(\text{max}_{S_d} - \text{min}_{S_d}) - 1 = 40.29\%$	32.16	53.42	49.57

Note: PS stands for log P -splines.

Table 3: Profit/Loss Including Transaction Costs

Time Series	Daily Series	GARCH D-series	PS D-series
Profit/Loss without Transaction Cost (%)	72.45	93.72	89.96
Transaction Times	17	28	23
Transaction Cost (%)	12.75	21.00	17.25
Profit/Loss with Transaction Cost (%)	59.70	72.72	72.61

6. Conclusion

With the support of fast-growing computing power, collecting and analyzing intraday or high-frequency data is a feasible task with low costs. The enormous intraday or high-frequency data steadily available on the emerging stock market provide a great opportunity of extracting more information from those data at market microstructure level that is considered to be closer to the continuous-time model setting.

In this paper, the random-walk hypothesis for SHCI based on five-minute data is rejected, supporting the intention of utilizing some structural model to seize the predictability in Shanghai stock market and gain excess returns. Also, re-sampling the high-frequency SHCI data into low-frequency data with the de-volatilization technique is applied. To further capture the dynamics of the five-minute SHCI, a time-dependent coefficient diffusion model with log P -splines and GARCH are applied to obtain volatility estimates. When applying the de-volatilization scheme, volatility estimates from log P -splines are shown to have nice property of picking up fewer data points in a flat market while more data points in a volatile market when applying the de-volatilization scheme.

The trading results show that accurate volatility estimation is the key to improving trading profit. It is found that to use traditional time series model, such as the State Space model, to extract signals for trading on the Shanghai Stock Exchange, the characteristics such as normality and homoscedasticity of the input time series are important for producing accurate forecasts as trading signals. With the log P -splines d -series, we can gain almost as much return as that with the GARCH d -series while trading fewer times.

References

- Aggarwal, R., Inclan, C., and Leal, R. (1999). Volatility in emerging stock markets. *Journal of Financial and Quantitative Analysis* **34**, 33-55.
- Bekaert, G., and Harvey, C. R. (1997). Emerging equity market volatility. *Journal of Financial Economics* **43**, 29-77.
- Black, F., and Scholes, M. (1973). The Pricing of options and corporate liabilities. —it *Journal of Political Economy* **81**, 637-654.
- Bollerslev, T. (1986). Generalized autoregressive conditional heteroscedasticity. *Journal of Econometrics* **31**, 307-327.
- Campbell, J. Y., Lo, A. W., and MacKinlay, A. C. (1997). *The Econometrics of Financial Markets*. Princeton University Press.
- Chan, K. C., Kayolyi, G. A., Longstaff, F. A., and Sanders, A. B. (1992). An empirical comparison of alternative models of the short-term interest rate. *Journal of Finance* **47**, 1209-1227.
- Cox, J. C., Ingersoll J. E., and Ross, S. A. (1985). A theory of the term structure of interest rates. *Econometrica* **61**, 929-952.
- Darrat, A. F., and Zhong, M. (2000). On testing the random-walk hypothesis: a model-comparison approach. *The Financial Review* **35**, 105-124.
- Engle, R. F. and Patton, A. J. (2001). What good is a volatility model? *Quantitative Finance* **1**, 237-245.

- Fan, J., Jiang, J., Zhang, C., and Zhou, Z. (2003). Time-dependent diffusion models for term structure dynamics and the stock price volatility. *Statistica Sinica* **13**, 965-992.
- Harvey, A. C., and Pierse, R. G. (1984). Forecasting economic time series with structural and box-jenkins models: a case study. *Journal of Business Economic and Statistics* **1**, 299-307.
- Harvey, A. C. (1991). *Time Series Analysis*, second Edition. MIT Press.
- Kalman, R. E. (1960). A New approach to linear filtering and prediction problems. *Trans. ASME Journal of Basic Engineering* **82**, 35-45.
- Kalman, R. E., and Bucy, R. S. (1961). New results in filtering and prediction theory. *Trans. ASME Journal of Basic Engineering* **83**, 95-108.
- Lee, C. F., Chen, G. M., and Rui, O. M. (2001). Stock returns and volatility on china's stock markets. *Journal of Financial Research* **26**, 523-543.
- Lo, A. W. and MacKinlay, A. C. (1988). Stock market prices do not follow random walks: evidence from a simple specification test. *Review of Financial Studies* **1**, 41-66.
- Mookerjee, R. and Yu, Q. (1999). An empirical analysis of the equity markets in China. *Review of Financial Economics* **8**, 41-60.
- Ruppert, D. (1997). Empirical-bias bandwidths for local polynomial nonparametric regression and density estimation. *Journal of the American Statistical Association* **92**, 1049-62.
- Santis, G. D., and Imrohorglu, S. (1997). Stock returns and volatility in emerging financial market. *Journal of International Money and Finance* **16**, 561-879.
- Shumway, R. H., and Stoffer, D. S. (1982). An approach to time series smoothing and forecasting using the EM algorithm. *Journal of Time Series Analysis* **3**, 253-264.
- Shumway, R. H., and Stoffer, D. S. (2000). *Time Series Analysis and Its Applications*. Springer- Verlag.
- Tian, G. G. and Guo, M. (2007). Interday and intraday volatility: additional evidence from the Shanghai stock exchange. *Review of Quantitative Finance & Accounting* **28**, 287-306.
- Vasicek, O. A. (1977). An equilibrium characterization of the term structure. *Journal of Financial Economics* **5**, 177-188.
- Wei, W. (2002). Forecasting stock market volatility with non-linear GARCH Models: A case for China. *Applied Economics Letters* **9**, 163-166.
- Yu, Y., Yu, K., Wang, H., and Li, M. (2008). Semiparametric estimation for a class of time-inhomogeneous diffusion processes. *Statistica Sinica*, to appear.
- Zhou, B. (1991). Analysis of volatile time series, with reference to foreign exchange rates. Ph.D Dissertation, University of California, Berkeley.

Zhou, B. (1996). High-frequency data and volatility in foreign-exchange rates. *Journal of Business & Economic Statistics* **14**, 45-52.

Zhou, Q. and Zhou, Z. (2005). Stock returns, volatility, and cointegration among Chinese stock markets. *China and World Economy* **13**, 106-22.

Received December 24, 2008; accepted May 6, 2009.

Hua Wang
P4P Product & Operations
2-3/F Xi Hu International Technology Building
391 Wen Er Road
Hangzhou, China, 310099
derekwang@alibaba-inc.com

Yan Yu
Department of Quantitative Analysis and Operations Management
College of Business
PO Box 210130
Cincinnati, OH 45221-0130, USA
Yan.Yu@uc.edu

Min Li
Decision Sciences Area
College of Business Administration
California State University, Sacramento
6000 J St, Sacramento, CA 95819-6088, USA
limin@csus.edu