End-of-Day Stock Trading Volume Prediction with a Two-Component Hierarchical Model

Shuhao Chen, Rong Chen, Gary Ardell, and Biquan Lin

Shuhao Chen
is a Ph.D. candidate in statistics at Rutgers University in Piscataway, NJ.
bhmchen@stat.rutgers.edu

Rong Chen
is a professor of Rutgers University in Piscataway, NJ and Peking University, in Beijing, China.
rongchen@stat.rutgers.edu

Gary Ardell
is the head of Financial Engineering and Advanced Trading Solution at the ConvergEx Group in Boston, MA.
gardell@convergex.com

Biquan Lin
is in Financial Engineering and Advanced Trading Solution at the ConvergEx Group in Iselin, NJ.
blin@convergex.com

Both human traders and algorithmic trading engine designers have a profound interest in the high-quality prediction of the volume that will be traded in the remainder of the trading day. This volume represents the liquidity against which orders can be transacted while the available liquidity determines the market impact of working any order. Indeed, with inadequate liquidity, it may not be feasible or it may be too expensive to execute a large order in the remainder of the trading day. The commonly used guaranteed execution algorithms, such as the initiation price, face the challenge of working parent orders across the day. To minimize market impact, the algorithm must keep its volume participation rate as low as possible. On the other hand, the algorithm must also ensure that its volume participation rate is high enough in order to complete the order within the day. Thus, it is crucial to have an accurate prediction of the volume in the remainder of the day to effectively execute such algorithms.

Remainder day's volume prediction also plays an important role in analyzing transaction cost. Berkowitz [1988] introduced the concept of daily volume weighted average price (VWAP) and used the difference between the average execution price and the recorded VWAP to measure the cost of each trade. Minimizing such cost is among critical goals for institutional investors and having better knowledge of reminder day's volume would definitely help to make an efficient execution decision and, thus, favorable cost.

Recently, both academia and practitioners have been pursuing better models forecasting the trading volumes. Lo and Wang [2000] analyzed behavior of equity trading volume using the capital asset pricing model (CAPM). Hautsch [2002] modeled the intraday volume activity based on volume durations using autoregressive conditional duration (ACD) models, which were originally introduced by Engle and Russell [1998]. Darolles and Le Fol [2003] proposed a methodology of decomposing trading volume and Bialkowski, Darolles, and Le Fol [2008] extended the previous work into intraday data, decomposing intraday trading volume into two components: 1) reflects volume change associated with market evolutions, and 2) represents stock-specific volume pattern. It used the historical VWAP curve to estimate the market component and the autoregressive moving average (ARMA) and the self-exciting autoregressive (SETAR) models to estimate stock specific component.

Although our aim is to predict the volume to be traded in the reminder of the day, we instead investigate the total volume accrued throughout the day, or the end-of-day volume, for simplicity. Given the fact that the volume that has been accumulated

from the beginning of the day to the time of the prediction is known to us, the two volumes are equivalent.

Intuitively, there are two sources of useful information for projecting the end-of-day volume. Since we have observed the partial volume observed up to the time of prediction, and if the distribution of total volume throughout the day is relatively stable and can be estimated using historical data, then the total end-of-day volume prediction can be made with the partial volume and the estimated proportion it should assume for the total volume. We call this method *intraday* prediction. The second source of information is the dynamics of daily volume changing over time. If such dynamics is properly modeled, daily volume can be predicted as well. We denote this method as *daily* prediction.

Since the intraday method utilizes the volume accrued during the trading day while the daily method uses daily volume series, those two methods provide independent predictions. They can be improved by combining both sources of information. In this article, we propose a hierarchical model for such a combination. It extends to the stable seasonal pattern model of Chen and Fomby [1999], where the model was used to predict end-of-year total number of tourists. A similar idea was used by Oliver [1999] as well. This approach is different from that of Bialkowski, Darolles, and Le Fol [2008] who decomposed the trading volume into two components that reflect volume changes due to market evolutions and the stock-specific volume pattern.

This article is organized as follows: In the next section, the two-component hierarchical model is presented. Its prediction procedures and some extensions are shown in the third section. The fourth section shows an empirical study using Dow Jones Industrial Average component stocks, comparing out-of-sample prediction performance of different methods.

TWO-COMPONENT HIERARCHICAL MODEL

Consider a trading volume series …, x_{t1}, x_{t2}, … x_{td}, … where $t = 1, 2, \ldots, n$ corresponds to different trading days and d denotes the number of trading periods used each day. In this article, we use $d = 13$, corresponding to thirteen 30-minute trading intervals of the U.S. equity market. The end-of-day volume for day t is then given by $\gamma_t = x_{t1} + \ldots + x_{td}$.

In Equation (1), we assume that the daily pattern is stable across different days, *which is to say that* given the end-of-day volume total γ_t, $(x_{t1}, x_{t2}, \ldots x_{td})$ are conditionally independent for $t = 1, 2, \ldots, n$. Namely

$$p[x_{11}, \ldots, x_{11d}, \ldots, x_{n1}, \ldots, x_{nd} \mid \gamma_1, \ldots, \gamma_n] = \prod_{t=1}^{n} p[x_{t1}, \ldots, x_{td} \mid \gamma_t] \quad (1)$$

where $p[\cdot|\cdot]$ denotes the conditional density.

This assumption allows us to model the series …, x_{t1}, x_{t2}, …, x_{td}, … with a two-component hierarchical structure, shown in Equation (2).

$$p[x_{t1}, \ldots, x_{td} \mid \gamma_t], \sim M_1(\gamma_t, \theta) \quad \text{and} \quad \{\gamma_t\} \sim M_2(\gamma_{t-1}, \ldots, \beta), \quad (2)$$

where the intraday model M_1 is a $(d - 1)$-dimensional distribution and the daily model M_2 is a time series model for the daily volume series $\{\gamma_t\}$. θ and β are parameters to be estimated.

There are several important characteristics associated with this hierarchical model. First, it is assumed that the dependency between different days is only through the total volume $\{\gamma_t\}$, not the individual observations within the days. Second, given the total volume, the intraday distribution of the total volume throughout the day is the same. Third, the two models, one for intraday distribution and the other for the daily volume series, can be modeled separately. Those three properties are ideal in terms of making intraday predictions of the end-of-day volume.

In the following steps, we discuss the possible models for M_1 and M_2. Since $\{x_{tk}\}$ is integer-valued, one immediate choice for model M_1 is the multinomial distribution, with the total being the daily total volume. However, trading volume is often large enough to be treated as a continuous variable. Treating it as a continuous variable also provides the advantage of more flexibility in the modeling of total volume series γ_t. Nevertheless, we still want its distribution to possess the proportional interpretation of the multinomial distribution. Chen and Fomby [1999] proposed a continuous analogue of the multinomial distribution called the Gaussian multinomial distribution. Specifically, as shown in Equation (3), a d-dimensional random variable $(x_{t1}, x_{t2}, \ldots, x_{td})$ is said to follow Gaussian-multinomial(G-MN) distribution $(\gamma_t, \theta_1, \ldots, \theta_d, \sigma^2)$ if

$$p[x_{t1}, \ldots, x_t(d-1)] \sim N(\mu, \sigma^2 \Sigma) \quad \text{and} \quad x_{td} = \gamma_t - \sum_{i=1}^{d-1} x_{ti}, \tag{3}$$

where $0 < \theta_i < 1, \sum_{i=1}^{d-1} \theta_i = 1$ and

$$\mu = \begin{pmatrix} \theta_1 \gamma_t \\ \theta_2 \gamma_t \\ \vdots \\ \theta_{d-1} \gamma_t \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \theta_1(1-\theta_1) & -\theta_1\theta_2 & \ldots & -\theta_1\theta_{d-1} \\ -\theta_1\theta_2 & \theta_2(1-\theta_2) & \ldots & \theta_2\theta_{d-1} \\ \vdots & \vdots & \ddots & \vdots \\ -\theta_1\theta_{d-1} & -\theta_2\theta_{d-1} & \ldots & \theta_{d-1}(1-\theta_{d-1}) \end{pmatrix}$$

Note that μ and Σ do not depend on t, due to our assumption of a stable daily pattern.

This distribution can be viewed as a continuous version of the multinomial distribution. The mean and correlation coefficient matrix are the same as the multinomial distribution, except for the extra variance parameter σ^2. In the multinomial distribution, the variance varies with the total volume in a relatively restricted way. In the Gaussian multinomial distribution, the variance is constant. It is possible to allow for varying variance (depending on observable variables) similar to the weighted regression setting or a GARCH type of heteroscedasticity. Here we choose to assume constant variance.

It can be shown that the Gaussian multinomial distribution also has the combination property of the multinomial distribution. For example, if $(x_{t1}, \ldots, x_{td}) \sim G-MN(\gamma, \theta_1, \ldots, \theta_d, \sigma^2)$, then $(x_{t1} + x_{t2}, x_{t3}, \ldots, x_{td}) \sim G-MN(\gamma, \theta_1 + \theta_2, \ldots, \theta_d, \sigma^2)$. More critically, this property implies that for $1 \le k < d$

$$\sum_{i=1}^{k} x_{ti} \sim N(\gamma_k \gamma_t, \sigma^2 \gamma_k (1-\gamma_k)), \quad \text{where } \gamma_k = \sum_{i=1}^{k} \theta_i$$

This feature allows us to estimate γ_k much more easily than estimating $\theta_1, \ldots, \theta_k$ in the high dimensional space. The parameter γ_k is of critical interest as our prediction procedure involves only γ_k, instead of the individual parameters $\theta_1, \ldots, \theta_k$. In Equation (4), the maximum likelihood estimator of γ_k in our setting is

$$\hat{\gamma}_k = \frac{\sum_{t=1}^{n} \left(\sum_{i=1}^{k} x_{ti} \right) \gamma_t}{\sum_{t=1}^{n} \gamma_t^2} \tag{4}$$

with standard error $\hat{\sigma} = [\hat{\gamma}_k (1 - \hat{\gamma}_k) / \sum \gamma_t^2]^{1/2}$. The maximum likelihood estimator for σ^2 is

$$\hat{\sigma}^2 = \frac{\sum_{t=1}^{n} (x_t - \hat{\theta}\gamma_t)' \hat{\Sigma}^{-1} (x_t - \hat{\theta}\gamma_t)}{n}$$

In Equation (5), to model the daily volume dynamics, we use a Gaussian model where

$$p[\gamma_t \mid \gamma_{t-1}] \sim N(\mu_t, \sigma_t^2) \tag{5}$$

where $\gamma_{t-1} = (\gamma_{t-1}, \ldots, \gamma_1)$, $\mu_t = f(\gamma_{t-1}, \theta)$, and $\sigma_t = g(\gamma_{t-1}, \theta)$. This model includes all the standard autoregressive integrated moving average (ARIMA) models of Box and Jenkins [1976] and Brockwell and Davis [1986], with or without GARCH errors (Bollerslev [1986]).

PREDICTION BASED ON THE TWO-COMPONENT HIERARCHICAL MODELS

At a given time k of the trading day t, we are interested in estimating the end-of-day volume γ_t using volume accumulated up to time k, namely $(x_{t1}, x_{t2}, \ldots, x_{tk})$ and the historical daily volume series $\gamma_{t-1}, \ldots, \gamma_1$. Specifically, let $x_t^{(k)} = (x_{t1}, x_{t2}, \ldots, x_{tk})$ and $\Sigma_k = (V_{tj})_{k \times k}$, where $V_{ij} = -\theta_i \theta_j$ for $i \ne j$ and $v_{ii} = \theta_i(1-\theta_i)$. Then,

$$p[\gamma_t \mid x_{t1}, \ldots, x_{tk}, \gamma_{t-1}] \propto p[x_{t1}, \ldots, x_{tk} \mid \gamma_t] \times p[\gamma_t \mid \gamma_{t-1}] \sim N(\mu_{t,k}, \sigma_{t,k}^2)$$

where

$$\sigma_{t,k}^2 = \left(\frac{\theta_k' \Sigma_k^{-1} \theta_k}{\sigma^2} + \frac{1}{\sigma_t^2} \right)^{-1} = \left(\frac{\gamma_k}{(1-\gamma_k)} \sigma^2 + \frac{1}{\sigma_t^2} \right)^{-1}$$

and

$$\mu_{t,k} = \left(\frac{x_k^{(k)} \Sigma_k^{-1} \theta_k}{\sigma^2} + \frac{\mu_t}{\sigma_t^2} \right) \sigma_{t,k}^2 = \left(\frac{\sum_{i=1}^{k} x_{ti}}{(1-\gamma_k)\sigma^2} + \frac{\mu_t}{\sigma_t^2} \right) \sigma_{t,k}^2 \tag{6}$$

Hence, the least squares prediction of γ_t is $\mu_{t,k}$. In fact, letting $c_t = \sigma^2 / \sigma_t^2$, Equation (6) can be written as Equation (7)

$$\mu_{t,k} = \frac{1}{\gamma_k + 1(1-\gamma_k)c_t}\sum_{i=1}^{k} x_{ti} + \frac{c_t}{\gamma_k + (1-\gamma_k)c_t}(1-\gamma_k)\mu_t$$

$$\triangleq \quad w_{tk}\frac{\sum_{i=1}^{k} x_{ti}}{\gamma_k} + (1-w_{tk})\mu_t \qquad (7)$$

Here, $\frac{\Sigma_{i=1}^{k} x_{ti}}{\gamma_k}$ and μ_t are the predictions of the end-of-day volume based on M_1 model alone and M_2 model alone, respectively. The weights ω_{tk} and $1-\omega_{tk}$ dictate the contribution of models M_1 and M_2 to the prediction of the total volume.

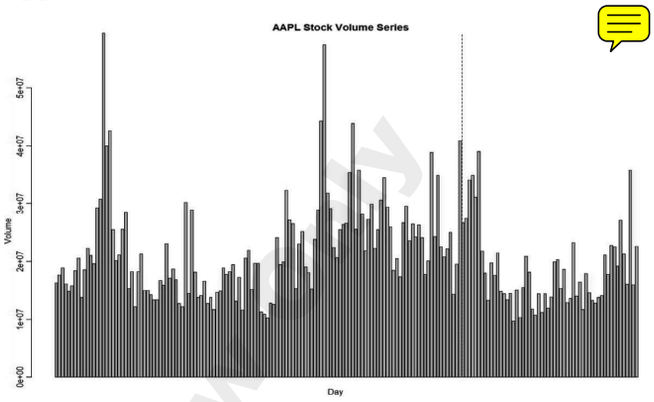
Note that, when the model for the total volume series and the model for volumes within the day have equal precision, that is, $\sigma^2 = \sigma_t^2$, then $\mu_{t,k} = \Sigma_{i=1}^{k} x_{ti} + (1-\gamma_k)\mu_t = \Sigma_{i=1}^{k} x_{ti} + \Sigma_{i=k-1}^{d} \theta_i \mu_t$. Hence, the adjustment procedure is simply to replace the predicted contribution of the individual observations $\theta_i\mu_t$ by their observed values x_{ti}. When $c < 1$, $\Sigma_{i=1}^{k} x_{ti}$, bears more weight in prediction, because the individual volume observation has less variance (more accuracy) than the daily volume model.

An alternative is to treat the combination weights ω_{tk} and $1-\omega_{tk}$ in Equation (7) as unknown parameters. If one assumes homoscedasticity in the daily volume series, the weights are constant over time. They can be estimated with least squares as if the daily total volume follows a regression model with $\Sigma_{i=1}^{k} x_{ti}$ and μ_t as explanatory variables. This model has a nonparametric modeling flavor as it bypasses the assumptions of M_1 and M_2 and optimizes the linear combination directly. We call this model the regression approach. One can further extend this model by including other informative variables in the combination, such as market trading volume up to period k.

EMPIRICAL STUDY

In this section, we apply the two-component hierarchical model to the historical volume profile of the 30 selected stocks in the Dow Jones Industrial Average from January 2010 to September 2010. Detailed results are shown using the volume profile of Apple, Inc.[1] There are 185 trading days in the study. We use the first 155 days for modeling and parameter estimation and the last 30 days for out-of-sample prediction performance comparison. Exhibit 1 shows the volume series of Apple, Inc. The volumes to the left of vertical dash line are used in fitting, while the right-side volumes are used for prediction and testing purposes.

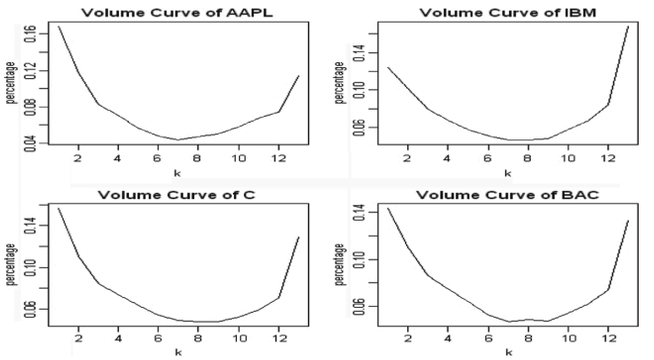
EXHIBIT 1
Apple, Inc., Volume Series



To avoid dealing with micro-structure noise and for the ease of computation, we aggregate the minute-by-minute volume data to 30-minute intervals with 13 periods per day. Here we use 1:00 p.m. as our time for the end-of-day volume prediction, that is, we have observed the trading volume up to 1:00 p.m. (thus, $k = 7$) and want to estimate the end-of-day volume using the accumulated volume that day and the historical daily volumes before that day.

We use the Gaussian multinomial distribution for Model M_1. Exhibit 2 depicts the estimated θ_i, $1 \le i \le 13$ for Apple, Inc., for the intraday distribution of the total volume into the 13 intraday periods. Regularity of the

EXHIBIT 2
Intraday Volume Distribution of Different Stocks



Note: (x-axis corresponds to the 13 trading periods and y-axis represents the volume percentage θ. Volumes would accumulate to 100%.)

pattern for different stocks is also demonstrated and verified. Furthermore, Exhibit 3 shows the estimates and standard errors of $\gamma_k \Sigma_{i=1}^{k} \theta_i$.

We fit the historical daily volume series using Gaussian ARMA models and ARMA-GARCH models. The autocorrelation function (ACF) and partial autocorrelation function (PACF) of Apple, Inc.'s daily volume series are shown in Exhibit 4. They appear to be quite stationary. Further study using the extended autocorrelation function of Tsay and Tiao [1984] and the adjusted extended autocorrelation function of Chen, Min, and Chen [2010] and the AIC criterion of Akaike [1974] identifies an AR(2) model as an appropriate model for the daily volume series of Apple, Inc. To include heteroscedasticity, we also model the error term of the AR(2) model as a GARCH(2, 0) process, after carrying out a model selection procedure. Exhibit 5 shows the estimated parameters for AR(2) and AR(2)–GARCH(2, 0) model.

The GARCH component is marginally significant. We also fit other volume series individually and identify the best ARMA and ARMA–GARCH models. They are used as Model M_2 in the prediction exercises.

Given the estimated parameters of M_1 and M_2, we compare following six different prediction methods.

EXHIBIT 3
Maximum Likelihood Estimates (standard errors) of the Gamma

	γ_1	γ_2	γ_3	γ_4	γ_5	γ_6	γ_7	γ_8	γ_9	γ_{10}	γ_{11}	γ_{12}
Estimate	0.165	0.280	0.363	0.434	0.490	0.537	0.580	0.630	0.681	0.741	0.813	0.888
Std. err.	0.011	0.013	0.014	0.015	0.015	0.015	0.015	0.014	0.014	0.013	0.012	0.010

EXHIBIT 4
ACF and PACF for Apple, Inc.'s Daily Volume Series

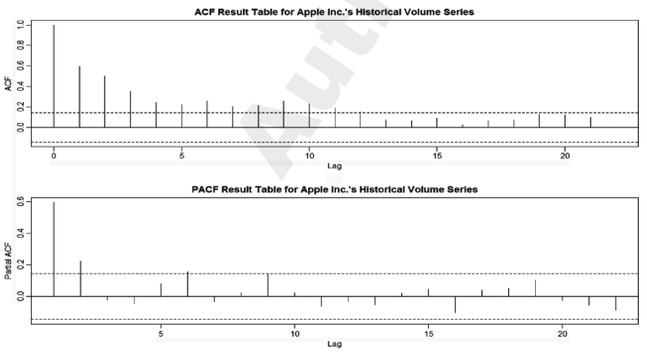


EXHIBIT 5
Fitting Results for Candidate ARMA Models

AR(2)			AR(2)–GARCH(2, 0)		
	Value	***Sd***		***Value***	***Sd***
Intercept	0.9966	0.0703	Intercept	0.3122	0.0680
AR1	0.4599	0.0714	AR1	0.4863	0.1038
AR2	0.2239	0.0714	AR2	0.1844	0.0834
σ^2		0.0943	GAR0	0.0515	0.0100
			GAR1	0.1711	0.1081
			GAR2	0.4069	0.1886
Log Likelihood		–44.35	Log Likelihood		–36.54

1. **ARMA:** Prediction using the daily volume series only, under the ARMA model;
2. **GARCH:** Prediction using the daily volume series only under ARMA–GARCH model;
3. **Intraday:** Prediction using intraday model M_1 only;
4. **New-ARMA:** Prediction using the new hierarchical model with ARMA as model M_2;
5. **New-GARCH:** Prediction using the new hierarchical model with ARMA–GARCH as model M_2;
6. **Reg:** Prediction using combination (7) with least squares optimized weights.

We first show the detailed results of Apple, Inc. Exhibit 6 numerates the estimated variance σ_2 of the intraday Gaussian–Multinomial distribution (M_1) and σ_t^2 for the ARMA model (M_2) and the corresponding weights ω_{tk} and $1-\omega_{tk}$ in Equation (7). Note that we assume that the ARMA models for M_2, c_t and ω_{tk} are indeed independent of time t. In this case, the prediction provided by the daily series is relatively more accurate (smaller variance), hence, having a large weight in the combined prediction. The estimated weights under least square criterion is also presented and it tends to put relatively more weight on intraday prediction, which is not optimal in this case.

EXHIBIT 6
Variance Comparison I

Intraday (σ^2)	ARMA (σ_t^2)	w_{tk}	$1 - w_{tk}$	Reg w_1	Reg w_2
6.41E + 13	4.22E + 13	0.48	0.52	0.77	0.23

Exhibit 7 shows the estimated variance σ_t^2 under the ARMA–GARCH model for the 30 days in our prediction period and its corresponding ratio c_t, comparing to the estimated σ^2 listed in Exhibit 6. It does change quite significantly, from 0. 3 to almost 3.

Exhibit 8 shows the predictions of ARMA, GARCH, intraday, reg, new-ARMA and new-GARCH, with the true observations marked as dots. Day 28 is an unusual observation. The actual volume is significantly larger than normal. The new prediction method is able to capture such a large movement, while the daily model underpredicts and the intraday model overpredicts.

Exhibit 9 shows the actual prediction of the volume (in millions of shares) of the 30 day prediction period, under different methods. The true observation, labelled as *real* is shown at the bottom line for comparison.

Exhibit 10 summarizes the prediction performance by showing the root-mean-squared prediction error $RMSE = \sqrt{\Sigma_{i=1}^{n}(Pred_i - True_i)^2}$.

EXHIBIT 7
Variance Comparison II

T	σ_t^2	σ^2/σ_t^2	t	σ_t^2	σ^2/σ_t^2	t	σ_t^2	σ^2/σ_t^2
1	3.32E + 13	1.93	11	6.03E + 13	1.06	21	3.57E + 13	1.80
2	2.46E + 13	2.60	12	4.71E + 13	1.36	22	2.43E + 13	2.64
3	2.99E + 13	2.14	13	2.55E + 13	2.51	23	2.86E + 13	2.24
4	2.51E + 13	2.56	14	3.84E + 13	1.67	24	2.18E + 13	2.94
5	3.22E + 13	1.99	15	2.72E + 13	2.36	25	3.04E + 13	2.11
6	2.54E + 13	2.53	16	2.75E + 13	2.34	26	4.86E + 13	1.32
7	3.43E + 13	1.87	17	2.79E + 13	2.30	27	2.70E + 13	2.38
8	2.65E + 13	2.42	18	2.45E + 13	2.62	28	8.58E + 13	0.75
9	3.95E + 13	1.62	19	2.15E + 13	2.99	29	1.77E + 14	0.36
10	3.10E + 13	2.07	20	2.52E + 13	2.54	30	7.23E + 13	0.89

EXHIBIT 8
Comparison of Different Prediction Methods

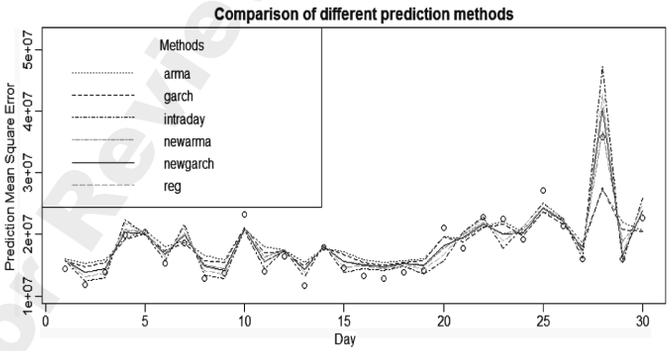


EXHIBIT 9
Prediction Results from Various Methods (in millions of shares)

	Day 1	Day 2	Day 3	Day 4	Day 5	Day 6	Day 7	Day 8	Day 9	Day 10
ARMA	16.05	15.30	15.83	19.32	20.62	18.10	18.88	16.55	15.84	20.93
GARCH	15.78	14.67	15.48	19.22	20.15	17.30	18.64	15.74	15.47	21.10
Intraday	15.68	12.39	13.00	22.41	20.13	15.70	21.65	13.38	12.79	20.65
New-ARMA	15.87	13.89	14.46	20.82	20.38	16.94	20.22	15.01	14.36	20.80
New-GARCH	15.74	13.86	14.49	20.36	20.14	16.72	19.95	14.86	14.22	20.92
Reg	15.85	13.14	13.73	21.82	20.35	16.34	21.13	14.19	13.57	20.83
Real	14.38	11.88	13.87	19.94	20.31	15.27	18.62	12.88	13.64	23.27
	Day 11	**Day 12**	**Day 13**	**Day 14**	**Day 15**	**Day 16**	**Day 17**	**Day 18**	**Day 19**	**Day 20**
ARMA	18.01	17.46	15.49	17.72	17.25	15.89	15.38	15.76	16.09	19.70
GARCH	17.00	17.20	14.88	17.70	16.69	15.49	15.03	15.47	15.75	19.69
Intraday	14.48	17.49	13.20	18.38	13.90	14.41	14.11	14.83	13.60	15.72
New-ARMA	16.30	17.47	14.38	18.04	15.62	15.17	14.77	15.31	14.89	17.77
New-GARCH	15.56	17.35	14.27	18.01	15.64	15.08	14.68	15.25	15.05	18.26
Reg	15.38	17.58	13.81	18.33	14.75	14.84	14.49	15.13	14.26	16.73
Real	13.98	16.42	11.68	17.92	14.61	13.25	12.83	13.80	14.12	21.09
	Day 21	**Day 22**	**Day 23**	**Day 24**	**Day 25**	**Day 26**	**Day 27**	**Day 28**	**Day 29**	**Day 30**
ARMA	19.38	21.23	22.11	20.41	23.73	22.40	18.56	27.22	21.99	20.63
GARCH	18.77	21.06	21.62	19.79	23.64	21.63	17.94	27.65	20.83	20.54
Intraday	20.36	22.81	17.79	21.29	25.20	22.55	16.19	47.18	15.47	25.91
New-ARMA	19.86	22.00	20.02	20.83	24.44	22.47	17.41	36.89	18.83	23.19
New-GARCH	19.48	21.67	20.13	20.28	24.27	22.11	17.28	40.48	16.55	23.86
Reg	20.25	22.57	18.89	21.20	25.00	22.64	16.83	42.82	17.07	24.83
Real	17.76	22.74	22.51	19.22	27.14	21.30	16.00	35.75	15.98	22.60

In the following, we present the prediction performance comparison for the 30 stocks in the Dow Jones Industrial Average Index. The first two columns of Exhibit 11 provide the company information; the following two columns store the ARMA and GARCH orders of the daily volume series. Mean-squared prediction errors of the six prediction methods are then listed, followed by the ratio comparison statistics.

We make the following observations:

1. In most cases, the new two-component hierarchical model performs much better than using the intraday model alone and using the daily series dynamics alone. In more than half of the cases, the improvement is over 20%, some significantly higher.
2. Although the ARMA–GARCH model may be marginally better than ARMA model for modeling the daily volume series dynamics (as in the case of Apple, Inc.), the ARMA model works almost as well as ARMA–GARCH model in terms of prediction.
3. The alternative combination method with the least-squares-optimized weights outperforms the intraday model alone and the daily series alone. Its overall performance is not as good as the one based on the hierarchical model.

EXHIBIT 10

Root Mean Square Error from Various Methods

ARMA	GARCH	Intraday	New-ARMA	New-GARCH	Reg	GARCH / ARMA	New-GARCH / GARCH	New-GARCH / Intraday
15.85	13.76	15.62	9.72	9.72	11.95	86.84%	70.69%	62.24%

EXHIBIT 11

Prediction Comparison on Dow Jones Industrial Average Components Based on Prediction Mean Square Error (in millions of shares)

Company	Symbol	ARMA Orders	GARCH Orders	ARMA	GARCH	Intraday	New-ARMA	New-GARCH	Reg	GARCH / ARMA	New-GARCH / New-ARMA	New-GARCH / GARCH	New-GARCH / Intraday
3M	MMM	1,0	1,0	3.4	3.23	1.62	1.73	1.64	1.43	94.95%	95.05%	50.98%	101.80%
Alcoa	AA	1,0	2,0	28.33	21.58	12.45	11.75	10.55	10.9	76.18%	89.74%	48.87%	84.71%
American Express	AXP	1,0	1,1	5.93	4.65	5.57	3.96	3.73	4.96	78.42%	94.28%	80.19%	66.99%
ATT	T	1,1	1,0	16.24	16.13	12.94	11.29	11.1	11.56	99.31%	98.27%	68.78%	85.74%
Boeing	BA	1,1	1,1	6.34	6.08	3.64	3.07	3.27	2.8	95.93%	106.67%	53.81%	89.98%
Caterpillar	CAT	1,1	1,1	7.36	6.49	5.4	4.43	4.36	4.74	88.07%	98.55%	67.26%	80.79%
Chevron Corporation	CVX	1,1	1,1	5.08	4.76	3.84	3.14	2.92	3.32	93.84%	93.12%	61.32%	76.02%
Cisco Systems	CSCO	1,0	1,0	40.18	31.66	34.9	23.55	22.9	28.57	78.79%	97.21%	72.33%	65.61%
Coca-Cola	KO	1,0	0,0	7.51	7.51	4.88	3.81	3.81	4.01	100.00%	100.00%	50.78%	78.09%
DuPont	DD	1,1	1,0	4.87	4.91	3.63	3.12	3.1	3.09	100.83%	99.40%	63.05%	85.23%
Exxon Mobil	XOM	1,1	1,1	13.2	11.53	7.95	7.14	7.14	7.08	87.32%	100.04%	61.94%	89.84%
General Electric	GE	2,0	0,0	53.57	53.57	31.28	24.82	24.82	23.53	100.00%	100.00%	46.33%	79.34%
Hewlett-Packard	HPQ	1,1	1,1	27.03	26.97	20.82	16.28	17.95	16.5	99.79%	110.26%	66.53%	86.21%
The Home Depot	HD	1,1	0,0	9.05	9.05	7.8	6.28	6.28	6.94	100.00%	100.00%	69.37%	80.48%
Intel	INTC	1,0	0,0	53.88	53.88	51.88	26.28	26.28	37.96	100.00%	100.00%	48.78%	50.66%
IBM	IBM	1,1	1,1	4.05	3.32	3.73	2.79	2.53	3.25	81.90%	90.56%	76.24%	67.77%
Johnson & Johnson	JNJ	1,1	1,0	6.33	6.43	5.29	3.9	3.77	4.94	101.53%	96.64%	58.69%	71.38%
J.P. Morgan Chase	JPM	1,1	2,0	25.91	17.45	17.37	12.92	11.41	14.45	67.34%	88.32%	65.42%	65.70%
Kraft Foods	KFT	1,0	2,0	8.95	6.7	5.48	4.29	4.36	4.56	74.91%	101.42%	65.01%	79.52%
McDonald's	MCD	1,0	1,1	4.77	4.21	5.62	2.39	3.55	5.08	88.19%	148.64%	84.36%	63.13%
Merck	MRK	1,1	1,1	8.31	6.58	6.46	5.53	5.37	5.31	79.15%	97.19%	81.65%	83.11%
Microsoft	MSFT	1,0	1,1	40.50	37.43	51.02	35.37	37.1	46.32	92.41%	104.90%	99.11%	72.71%
Pfizer	PFE	1,1	1,1	35.56	35.36	21.79	20.07	19.46	18.7	99.41%	96.97%	55.05%	89.30%
Procter & Gamble	PG	1,1	1,1	6.48	5.76	5.27	3.96	3.67	4.79	88.85%	92.77%	63.75%	69.65%
Travelers	TRV	1,1	1,0	4.01	4.03	3.22	2.45	2.44	2.94	100.43%	99.40%	60.48%	75.67%
United Technologies Corporation	UTX	1,0	1,1	3.63	3.31	2.63	2.32	2.18	2.44	91.14%	94.13%	65.98%	82.95%
Verizon	VZ	1,1	2,0	14.15	14.09	9.14	8.79	8.38	7.74	99.54%	95.36%	59.48%	91.71%
Wal-Mart	WMT	1,1	1,1	9.13	8.35	6.21	5.15	4.98	5.25	91.40%	96.59%	59.62%	80.17%
Walt Disney	DIS	1,1	2,0	6.92	4.86	5.66	4.30	3.27	4.83	70.21%	76.01%	67.27%	57.72%
Average										90.34%	98.67%	64.57%	77.65%

4. The phenomenon observed on the 28th day of the Apple, Inc., volume series is quite important. It shows that the hierarchical model can indeed effectively combine two independent sources of information and produce a more accurate prediction. This phenomenon is also observed in other stocks as well.
5. The idea of including other factors, such as market volume, in the combination has been tried, without success. More research needs to be done to find more appropriate factors.

In conclusion, our empirical study shows that the proposed two-component hierarchical model and its associated prediction method are effective in making predictions of the end-of-day volume, and is more accurate compared to that using the intraday model alone and daily volume series alone.

ENDNOTES

[1]Arbitrarily chosen, not included in DJIA though.

[2]Apple, Inc., stock trades under AAPL on the NASDAQ.

REFERENCES

Akaike, H. "A New Look at the Statistical Model Identification." *IEEE Transactions on Automatic Control*, Vol. 19, No. 6 (1974), pp. 716-723.

Berkowitz, S.A., D.E. Logue, and E.A.J. Noser. "The Total Cost of Transaction at the NYSE, " *Journal of Finance*, Vol. 43, No. 1 (1988), pp. 97-112.

Bertsimas, D., and A.W. Lo. Optimal Control of Execution Costs, *Journal of Financial Markets*, 1 (1998), pp. 1-50.

Bialkowski, J., S. Darolles, and G. Le Fol. "Improving VWAP Strategies: A Dynamic Volume Approach." *Journal of Banking and Finance*, Vol. 32, No. 9 (September 2008), pp. 1709-1722.

Bollerslev, T. "Generalized Autoregressive Conditional Heteroskedasticity." *Journal of Econometrics*, 31 (1986), pp. 307-327.

Box, G.E.P., and G.M. Jenkins. *Time Series Analysis: Forecasting and Control,* San Francisco, CA: Holden-Day, 1976.

Brockwell, P.J., and R.A. Davis. *Time Series: Theory and Methods*, New York: Springer-Verlag, 1986.

Chen, R., and T.B. Fomby. "Forcasting with Satable Seasonal Pattern Models with an Application to Hawaiian Tourism Data." *Journal of Business and Economic Statistics*, Vol. 17, No. 4 (October 1999).

Chen, S., W. Min, and R. Chen. *Model Indentification for Time Series with Dependent Innovations*, Technical report, Department of Statistics, Rutgers University, (2010).

Darolles, S., and G. Le Fol. "Trading Volume and Arbitrage." Working paper, *CREST*, 2003.

Engle, R.F. "Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation." *Econometrica*, Vol. 50, No. 4 (1982), pp. 987-1008.

Engle, R.F., and J. Russell. "Autoregreesive Conditonal Duration: A New Model for Irregularly Spaced Transaction Data." *Econometrica*, 66 (1998), pp. 1127-1162.

Hautsch, N. "Modelling Intraday Trading Activity Using Box-Cox ACD Models." Working paper, Center of Finance and Econometrics, University of Konstanz, February 2005.

Lo, A., and J. Wang. "Trading Volume: Definition, Data Analysis, and Implication of Portfolio Theory." *Review of Financial Studies*, 13 (2000), pp. 257-300.

Madhavan, A. *Vwap Strategies. Transaction Performance: The Changing Face of Trading Investment Guides Series*, New York: Institutional Investor, Inc., 2002.

Oliver, R.M. "Bayesian Forecasting with Stable Seasonal Patterns." *Journal of Business and Economic Statistics*, Vol. 5 (October 1999), pp. 77-85.

Tsay, R.S., and G.C. Tiao. "Consistent Estimates of Autoregressive Parameters and Extended Sample Autocorrelation Function for Stationary and Nonstationary ARMA Models." *Journal of American Statistical Association*, Vol. 79, No. 385 (March 1984).