Modeling the Impact of Overnight Surprises on Intra-daily Stock Returns

Giampiero M. Gallo, Yongmiao Hong, and Tae-Hwy Lee
Università di Firenze, Dipartimento di Statistica,
Cornell University, Departments of Economics & Statistical Science, and
University of California, Riverside, Department of Economics

KEY WORDS functional-coefficient model, nonlinearity, predictive ability, volatility.

Abstract

In this paper we examine under what circumstances the information accumulated during market closing time and conveyed to the price formation at market opening may be exploited to predict where the stock price will be at the end of the trading day. In our sample of three financial time series, we find that, in spite of linear uncorrelatedness, there exists a strong nonlinear dependence structure in the conditional mean of the intra-daily returns. To model this structure we use the functional-coefficient (FC) model of Cai, Fan, and Yao (2000) where the coefficients are time-varying and dependent on the state of stock return volatility. Out-of-sample forecast performances of the FC models and linear models where the coefficients are constant are also compared using the criteria of mean square forecast errors, trading returns, and directional forecasts.

1. Introduction

Research on financial time series has long been based on the widely documented tenet that future asset prices returns are unpredictable whether one knows the past values of the series or even enlarged sets including other currently available public information. In statistical terms the assumption is that returns follow a martingale difference process. Many authors have attempted to show that if one breaks down the sample period, uses nonlinear models, introduces new explanatory variables, reproduces market behavior by chartists (Brock *et al.*, 1991) some degree of predictability is possible, although, as Granger (1992) argued, it may not lead to profitability of the outcome.

Several authors (e.g., Lo and MacKinlay, 1999; Sullivan, Timmerman and White 1999) have guarded against the so called data-snooping biases, i.e. the possibility that different analyses conducted on the same data sets may eventually uncover some sort of pattern that may be interpreted as evidence of predictability, and White (2000) has suggested a bootstrap–based *reality check* test to evaluate out–of–sample forecast performance.

More recently, also due to the increasing diffusion of ultra–high frequency data, the issue of market efficiency and market predictability has received considerable attention. Among others, Lo and MacKinlay (1999) claim that the larger quantity of information that is available to the markets and collected for analysis can reveal patterns which may cast some doubts about the correctness of the martingale difference hypothesis for asset returns even after taking data–snooping biases into consideration.

When working with daily data, it is customary that returns are measured on the basis of the closing prices. Amihud and Mendelson (1987) and Stoll and Whaley (1990) offered a rationale for this practice, arguing that returns measured as open-to-open are affected by specific trading mechanisms at work when markets open, resulting in a number of unappealing statistical features of the corresponding time series. In spite of this, opening prices are reckoned to be still of interest, since they convey the outcome of information accumulation during closing times as well and/or the feature of trading to convey a flow of information which was interrupted during closing times (Romer, 1993, Dow and Gorton, 1993). This information may be relevant when evaluating the behavior of intra-daily return, even without resorting to high frequency data.

In this paper, we will concentrate on one aspect of the price formation which can prove of interest for the study of market predictability, namely how one can use the information included in the price of an asset recorded when market opens, after many hours of potential information accumulation in the absence of active trading. In fact, we compute daily returns as usual as the log-difference of stock prices at market closing time but we decompose them

into two components, namely an overnight return (measured as the log-difference of the opening price and the price at the previous trading day closing time), and an intra-daily return measured as the log–difference of prices recorded the same trading day (at closing and opening time). The issue is then under what conditions the overnight return may contain useful information to predict the intra–daily return. Empirical evidence reported here shows that, depending on the overall daily return volatility, there is a correlation between overnight and intra-daily returns which can help in predicting the latter conditional on the value assumed by the former. This predictability is unlikely to be picked up by linear models which have constant coefficients. Rather, the model must be capable of reproducing the empirical regularity that coefficients depend on volatility. The candidate model which seems to include this needed flexibility is the functional coefficient model proposed by Cai *et al.* (2000, henceforth CFY), the coefficients of which are time-varying and can be made dependent on the degree of volatility prevailing that day.

The structure of the paper is as follows: we first discuss the nature of opening prices and review some of the evidence present in the literature (Section 2). In Section 3 we discuss the structure of the functional coefficient model and various strategies followed to capture some features present in the data. In Section 4 we compare the out–of–sample performance of the various estimated nonlinear models against a linear benchmark, using the methods of Diebold and Mariano (1995), West (1996) and White (2000). Section 5 contains some concluding remarks.

2. Overnight Surprises and Intra–daily Returns

To establish notation, let C_t be the closing price at time t, $t = 1, \ldots, n$, and O_t the opening price for the same day. Accordingly, the daily returns are approximated by the difference between the logarithms of closing prices, that is, $r_t = \ln C_t - \ln C_{t-1}$, . Clark (1973) considers r_t as the sum (over a random number of trades n_t) of independently and identically distributed price movements with mean 0 and constant variance σ^2. Accordingly, conditional on n_t, the variance of the daily returns is $n_t\sigma^2$. As noted by Gallo and Pacini (1998), though, one should keep into account that among these n_t trades, the first recorded price movement (occurring at market opening time) has different characteristics than the intra-day price movements. This different nature is a consequence of the price formation mechanisms at work around market opening time: next to the trading mechanisms specific to the exchange considered (cf. Cao, Ghysels, and Hathaway, 2000 for a discussion of pre-opening behavior at the NASDAQ), the overnight accumulation of information plays a special role. For example, cross-listings of the same company on other stock exchanges around the world convey some information available at opening time, and news released when markets are closed have not been translated into price movements.

Let us thus consider the decomposition of the daily returns r_t by adding and subtracting the log of opening prices:

$$\begin{aligned} r_t &= (\ln C_t - \ln O_t) + (\ln O_t - \ln C_{t-1}) \\ &\equiv \rho_t + \eta_t, \end{aligned}$$

so that ρ_t is the intra-daily return and η_t is the overnight return.

The series we use are the S&P500 index (1/3/1994 - 7/25/2001, 1909 daily observations) and two large caps stocks traded on the New York Stock Exchange which can be deemed representative of actively traded stocks: Citicorp and General Electric (both between 1/3/1994 and 8/4/2000, for a total of 1666 daily observations).

Like daily stock returns, also intra–daily returns exhibit volatility clustering, asymmetric response of volatility to the sign of returns and some moderate autocorrelation (not reported here). In addition, one should stress that when overall returns r_t are low (in absolute value) the "half-day" returns η_t and ρ_t are bound to be negatively correlated (since $\eta_t \approx -\rho_t$ when $|r_t| \approx 0$. Since $E|r_t|^2 \approx h_t$ (the conditional variance), absolute returns are connected to volatility, and hence a relationship between η_t and ρ_t should be detectable when daily volatility is low. Thus, enlarging the information set I_{t-1} available at closing time to include η_t should be relevant in modelling the conditional mean of ρ_t, at least for some states of volatility.

If the conjecture is correct, a linear constant parameter model,

$$\rho_t = a_0 + a_1\eta_t + \varepsilon_t, \tag{1}$$

should be incorrectly specified for the conditional mean of the intra-day return ρ_t; in particular, the coefficients a_0 and/or a_1 may happen to be statistically insignificant, whereas they may be time–varying. In view of what we argued above, the coefficient a_1 should capture a systematic pattern in the correlation between η_t and ρ_t and the volatility of the daily return process r_t. In synthesis, we want to investigate if the following statement holds for in-sample goodness-of-fit and for out-of-sample prediction:

Hypothesis: *The impact of overnight surprises η_t on the intra-daily return ρ_t depends on the conditional volatility of the daily return r_t.*

Under this hypothesis, the coefficient a_1 is not zero, not constant, and can be expressed as a function of some volatility measure of r_t, while the characteristics of a_0 are open to investigation. A suitable model which provides the needed flexibility of time-varying nonlinear response to some state variable (in our case daily volatility) is the functional coefficient model proposed by CFY.

3. The Functional-Coefficient Model

The linear model (1) exploits the information available at opening time, denoted as I^+_{t-1}, in a very restrictive fashion. We can think of a more general model in which the conditional mean of ρ_t, be it $E(\rho_t|I^+_{t-1})$, is a generic function of this information set, denoted as $g(I^+_{t-1})$. We can then write

$$\rho_t = g(I^+_{t-1}) + \varepsilon_t,$$

where $\{\varepsilon_t\}$ is a martingale difference sequence with respect to I^+_{t-1}. The key to forecasting is to manage and specify suitably the conditional mean $g(I^+_{t-1})$, which is generally a non-zero, time-varying function but is of complicated and unknown form. The functional-coefficient (FC) model of CFY, whose coefficients are time-varying and state-dependent, can be viewed as a linear model with time-varying and state-dependent coefficients, a special case of the more general state-dependent model of Priestley (1980), while retaining a good degree of flexibility, since it encompasses the models of Tong (1990) and Chen and Tsay (1993). It has the advantages of capturing a fine structure of the underlying dynamics and of giving good out-of-sample forecasts. A key feature of this model is that it makes use of a variable U_t which is function of the same information set I^+_{t-1} as a 'threshold variable' on which the functional-coefficients depend; that is, for the case at hand we have,

$$\rho_t = a_0(U_t) + a_1(U_t)\eta_t + \varepsilon_t \qquad (2)$$

where the $a_j(U_t)$'s, $j = 0, 1$, are the functional coefficients depending on U_t.

The coefficient functions $\{a_j(\cdot)\}$ are estimated by a locally linear regression method (e.g., Fan and Gijbels, 1996). For any given point U_0, we can approximate the functions $a_j(U_t)$'s locally by a linear function:

$$a_j(U_t) \approx \alpha_j + \beta_j(U_t - U_0),\ j = 0, 1,$$

for U_t in a neighborhood of U_0, where α_j and β_j are constants. The local linear estimator at point U_0 is given by $\hat{a}_j(U_0) = \hat{\alpha}_j$. The coefficients $\{(\hat{\alpha}_j, \hat{\beta}_j)\}^1_{j=0}$ are chosen as those values which minimize the sum of weighted squares

$$\sum_{t=1}^{n} \{\rho_t - a_0 - a_1\eta_t\}^2 K_h(U_t - U_0),$$

where $K_h(\cdot) = K(\cdot/h)/h$ for a given kernel function $K(\cdot)$ and bandwidth h. Note that here n denotes the number of observations used for in-sample estimation.

Since we are interested in out-of-sample predictive ability of the FC model, we can select h using an out-of-sample cross-validation procedure, as suggested by CFY. Let m and Q be two positive integers such that $n > mQ$. The basic idea is first to use Q sub-series of lengths $n - qm$ $(q = 1, \ldots, Q)$ to estimate the coefficient functions and then to compute the one-step forecast errors of the next segment of the time series of length m based on the estimated models. That is, we choose h to minimize the average of the mean square forecast errors

$$AMS(h) = \frac{1}{Q}\sum_{q=1}^{Q} AMS_q(h)$$

where

$$AMS_q(h) = \frac{1}{m}\sum_{t=n-qm+1}^{n-qm+m} \{\rho_t - \hat{a}_{0,q}(U_t) - \hat{a}_{1,q}(U_t)\eta_t\}^2$$

and $\{\hat{a}_{j,q}(\cdot)\}^1_{j=0}$ are computed from the sample $\{\eta_t, \rho_t, U_t\}^{n-qm}_{t=1}$. Following CFY, we use

$m = [0.1n]$, $Q = 4$, and the Epanechnikov kernel $K(u) = \frac{3}{4}(1-u^2)\mathbf{1}(|u| < 1)$, where $\mathbf{1}(\cdot)$ is the indicator function. As often occurs, the particular choice of the kernel function is not crucial for the results.

3.1 The choice of U_t

It is obviously important to choose an appropriate variable U_t when estimating the FC model. Knowledge of the data structure or of some economic theory may be helpful, but when no prior information is available, U_t may be chosen as a function of given explanatory variables or may be chosen using some data-driven methods as AIC-based selection and cross-validation.

In the absence of a specific theory, here we adopt an heuristic approach by choosing the following variables for U_t:

- The daily return, as level, square and absolute value; the level preserves the sign of the returns and would signal a dependence of the coefficients on the size and sign of the returns. Squared returns and returns in absolute value are more closely related to volatility.

- The spread between daily absolute returns and their moving average of length N. A moving average of absolute returns can be seen as a rough measure of local volatility and the spread from it represents whether the most recent return is above or below this "average" volatility. For the problem analyzed here the spread may signal an incoming increase or decrease in volatility. Various lengths can be specified: here we chose $N = 5, 10$ and 20 trading days. Absolute returns are considered in forming the moving average rules instead of the squared returns, since the former has very interesting statistical properties (e.g. stronger evidence of long memory), as emphasized in Granger (1998, p. 269) and Granger and Ding (1995).

- The daily high-low spread which can be seen as an alternative measure of volatility (cf. Garman and Klass, 1980, and Parkinson, 1980, for estimating the variance of returns from high–low range data; and, more recently, Gallant *et al.*, 1999);

- The estimated conditional variance of r_t, $\sigma_t^2 \equiv Var(r_t|I_{t-1})$ modelled according to three specifications of the GARCH family: these are special cases of the threshold GARCH (TGARCH) model of Glosten *et al.* (1993),

  $$\sigma_t^2 = \omega+\beta\sigma_{t-1}^2+\alpha e_{t-1}^2+\gamma e_{t-1}^2\mathbf{1}(e_{t-1} \geq 0),$$

  where $e_t \equiv r_t - a_0 - a_1 r_{t-1}$. The GARCH model of Bollerslev (1986) is the case with $\gamma = 0$. J.P. Morgan's Riskmetrics (1996) model, which suggests an exponentially weighted moving average (EWMA) of past squared innovations as an estimate of the variance, can also be seen as another special case where we do not resort to estimation but we choose parameters as $\omega = 0$, $\beta = 0.94$, $\alpha = 1 - \beta$, $\gamma = 0$.

- The trading volume: the relationship between variability of returns and trading volumes has been analyzed by several authors (Epps and Epps, 1976, Cornell, 1981, Tauchen and Pitts, 1983, Cooper, 1999) and often modelled jointly.

Summarizing, in the empirical implementation, we include the following eleven choices for U_t, all included in the information set I_{t-1}:

$$\begin{aligned}
&U_{1,t} = r_{t-1},\\
&U_{2,t} = r_{t-1}^2,\\
&U_{3,t} = |r_{t-1}|,\\
&U_{4,t} = |r_{t-1}| - N^{-1}\textstyle\sum_{j=1}^{N}|r_{t-j}|,\ N = 5,\\
&U_{5,t} = |r_{t-1}| - N^{-1}\textstyle\sum_{j=1}^{N}|r_{t-j}|,\ N = 10,\\
&U_{6,t} = |r_{t-1}| - N^{-1}\textstyle\sum_{j=1}^{N}|r_{t-j}|,\ N = 20,\\
&U_{7,t} = (\text{High-Low Spread})_{t-1},\\
&U_{8,t} = \text{Riskmetrics EWMA},\\
&U_{9,t} = \text{GARCH},\\
&U_{10,t} = \text{TGARCH, and}\\
&U_{11,t} = (\text{Trading Volume})_{t-1}.
\end{aligned}$$

In Tables 1-3, the FC model using U_k will be denoted as FC_k ($k = 1, \ldots, 11$). For S&P500 index, the trading volume data is not available and thus U_{11} will not be used.

3.2 Testing for functional-coefficients

To provide evidence for the usefulness of the FC model, we apply CFY's goodness-of-fit test for a specific parametric model against an FC alternative, based on bootstrap. We consider the linear constant parameter model (1) as the null hypothesis and the functional coefficient model (2) as the alternative, i.e.,

$$H_0 : a_j(U_t) = a_j,\ j = 0, 1, \qquad (3)$$

to test for parameter constancy. Under H_0, the process $\{\rho_t\}$ is linear in conditional mean. When H_0 in (3) does not hold, the coefficients are functionals of U_t and the linear model suffers from 'neglected nonlinearity'. CFY test consists of comparing the residual sum of squares (RSS) under the null hypothesis

$$RSS_0 \equiv \sum_{t=1}^{n} \hat{\varepsilon}_t^2 = \sum_{t=1}^{n} (\rho_t - \hat{a}_0 - \hat{a}_1 \eta_t)^2$$

with RSS under the alternative

$$RSS_1 \equiv \sum_{t=1}^{n} \tilde{\varepsilon}_t^2 = \sum_{t=1}^{n} (\rho_t - \hat{a}_0(U_t) - \hat{a}_1(U_t)\eta_t)^2.$$

The statistic is $T_n = (RSS_0 - RSS_1)/RSS_1$. We reject the null hypothesis for large values of T_n.

Fan, Zhang, and Zhang (2001, Theorem 5) show the asymptotic normality of T_n with a suitable normalization. An important consequence of this result is that we do not have to derive theoretically the normalizing factors in order to be able to use the test, but we can directly simulate the distribution of the test statistic T_n under the null hypothesis via bootstrap:

1. Generate the bootstrap residuals $\{\varepsilon_t^*\}$ from the centered residuals $(\tilde{\varepsilon}_t - \bar{\tilde{\varepsilon}})$ where $\bar{\tilde{\varepsilon}} = n^{-1} \sum \tilde{\varepsilon}_t$ and define $\rho_t^* \equiv \hat{a}_0 + \hat{a}_1 \eta_t + \varepsilon_t^*$.

2. Construct the bootstrap sample $\{\rho_t^*, U_t\}_{t=1}^n$ and calculate the bootstrap statistic T_n^*. This step is repeated over the number of desired replications.

3. Reject the null hypothesis H_0 in (3) when the test statistic T_n computed over the original data is greater than the $100 \times (1 - \alpha)$ percentile of the conditional distribution of T_n^* given $\{\rho_t, U_t\}_{t=1}^n$. The bootstrap p-value of T_n is approximately the relative frequency of the event $\{T_n^* \geq T_n\}$ in the bootstrap resamples.

3.3 Results on the FC models

The bootstrap p-values of T_n are reported in the tables. Both the naive-bootstrap (Efron, 1979) and the wild-bootstrap (Wu 1986, Liu 1988) procedures are used, whose p-values are denoted as P_B and P_W, respectively. The bandwidth h for a FC model is chosen such that AMS(h) is minimized, among the 11 values of $h = 2^j n^{-1/5}$ where $j = -5, -4, -3, -2, -1, 0, 1, 2, 3, 4,$ and 5. All U_t's are standardized by dividing them by their unconditional standard deviations. In the second columns of Tables 1-3, reported are the values of j chosen. The p-values, computed from 100 bootstrap resamples and reported in brackets, indicate strong rejection of H_0 in (3) in favor of the FC models in most cases. Indeed, the estimated statistics T_n are positive for all cases indicating $RSS_0 > RSS_1$. Thus, the result that we get in estimation is that neglected nonlinearity in the linear model may be explored using the FC model. Most choice of U_t delivers significant improvement in goodness-of-fit with many p-values are close to zero. Comparing the choices of U_t in terms of T_n and its p-values, it may be interesting to note that FC_9 with U_t being estimated from GARCH is the worst model for S&P500 and GE. FC_9 is also the second worst for Citicorp. In general, asymmetric TGARCH seems to work better. Whether this nonlinearity can be exploited even in an out-of-sample prediction exercise is an issue explored in the next section.

4. Out-of-Sample Predictive Ability of FC Models

In addition to specification testing and estimation, out-of-sample forecast evaluation is also important to make the analysis robust to the possible consequences of structural changes and data snooping. To evaluate the nonlinear models in terms of out-of-sample predictive ability, in Section 4.1, we first discuss three forecast evaluation criteria – mean squared forecast error, mean trading returns, and mean correct directional forecasts. Our primary objective is to compare the FC model with the linear constant parameter model in (1). When several models using the same data are compared for predictive ability, it is crucial to take into account the dependence among the forecasts from various models because of the data-snooping problem, which occurs when a model is searched extensively until a match with the given data is found. Conducting inference without taking into account specification search is commonly referred to as 'data-mining' and can be extremely misleading (cf. Lo and MacKinlay 1999, Ch.8). White (2000) develops a test to compare multiple models in predictive ability accounting for specification search, built on West (1996) and Diebold and Mariano

(1995). Section 4.2 provides a short discussion of the method.

4.1 Forecast evaluation criteria

Our evaluation of out-of-sample forecasts proceeds as follows. There are P predictions in all for each model. Suppose one-step predictions are to be made for P prediction periods, indexed from R through n, so that $n = R+P-1$. Here, P and R may increase as the sample size n increases. The first forecast is based on the model parameter estimator $\hat{\beta}_R$, formed using observations 1 though R, the next based on the model parameter estimator $\hat{\beta}_{R+1}$, formed using observations 2 though $R+1$, and so forth, with the final forecast based on the model parameter estimator $\hat{\beta}_n$. Based on the estimated models using a series of rolling samples, each of size R, one-step ahead forecasts are generated for P post-samples, resulting in P forecasts to evaluate each model. Let $\{\hat{\rho}_{t+1}\}_{t=R}^{n}$ be an estimated forecast of $\{\rho_{t+1}\}_{t=R}^{n}$ using information $\{I_t^+\}_{t=R}^{n}$. We compare the forecasts in terms of mean squared forecast errors (MSE)

$$MSE_P \equiv P^{-1}\sum_{t=R}^{n}(\rho_{t+1} - \hat{\rho}_{t+1})^2.$$

However, our main aim is to investigate profitability of using a FC model relative to that of using a benchmark linear model. Because the investors are ultimately trying to maximize profits rather than minimize forecast errors, MSE may not be the most appropriate evaluation criterion. We consider two additional forecast evaluation criteria.

Our second criterion is the mean trading return (MTR) of a strategy defined as

$$MTR_P \equiv P^{-1}\sum_{t=R}^{n} S_{t+1}\rho_{t+1},$$

where S_{t+1} is a signal function at time t for the next period $t+1$ representing the recommended trading position. The estimation of S_{t+1} will be carried out based on the linear models and the FC models. The signal function is

$$S_{t+1} = \mathbf{1}(\hat{\rho}_{t+1} > 0) - \mathbf{1}(\hat{\rho}_{t+1} < 0),$$

which takes a value of $+1$ (for a buy signal), -1 (for a sell signal), or 0. If ρ_{t+1} is predicted to be positive, then $S_{t+1} = 1$. Four interesting cases may worth mentioning. First, for the martingale model, we have $\rho_{t+1} = 0$ and $S_{t+1} = 0$ for all t. Hence, $MTR_P = 0$. Second, the Buy-and-Hold strategy, which is defined with $S_{t+1} = 1$ for all t, has the mean trading return $MTR_P^{\text{Buy-Hold}} = P^{-1}\sum_{t=R}^{n}\rho_{t+1}$. Third, if ρ_{t+1} and its forecast $\hat{\rho}_{t+1}$ have the same signs for all t, i.e., if we could make the perfect directional forecasts for all t, then $S_{t+1}\rho_{t+1} = |\rho_{t+1}|$ for all t, and $MTR_P^{\text{Perfect}} = P^{-1}\sum_{t=R}^{n}|\rho_{t+1}|$.

The third forecast evaluation criterion is about the directional forecasts. The forecast $\hat{\rho}_{t+1}$ of ρ_{t+1} is correct in direction (sign) if $\hat{\rho}_{t+1}\rho_{t+1} > 0$. The probability that a model generates a correct directional prediction of ρ_{t+1} is $\Pr(\hat{\rho}_{t+1}\rho_{t+1} > 0)$, which can be estimated by mean correct directional prediction (MCD)

$$MCD_P \equiv P^{-1}\sum_{t=R}^{n}\mathbf{1}(\hat{\rho}_{t+1}\rho_{t+1} > 0).$$

4.2 Comparing forecasting models

Model comparison via forecast criteria can be conveniently formulated as hypothesis testing of some suitable moment conditions. Consider an $l \times 1$ vector of moments, $E(\psi^*)$, where $\psi^* = \psi(Z, \beta^*)$ is an $l \times 1$ vector with elements $\psi_k^* \equiv \psi_k(Z, \beta^*)$ for a random vector $Z = (\rho, \eta, r, U)'$ and $\beta^* \equiv \text{plim}\,\hat{\beta}_n$. The appropriate null hypothesis is that the best model is no better than a benchmark, expressed formally as

$$H_0 : \max_{1\le k\le l} E(\psi_k^*) \le 0. \tag{4}$$

This is a multiple hypothesis, the intersection of the one-sided individual hypotheses $E(\psi_k^*) \le 0,\ k = 1, \ldots, l$. The alternative is that H_0 is false, that is, that the best model is superior to the benchmark. White's (2000) results for testing H_0 in (4) hold whenever the $l \times 1$ sample moment vector

$$\bar{\psi} = P^{-1}\sum_{t=R}^{n}\psi(Z_{t+1}, \hat{\beta}_t)$$

has a continuous limiting distribution.

West (1996, Theorem 4.1) shows that under proper regularity conditions,

$$\sqrt{P}(\bar{\psi} - E(\psi^*)) \to N(0, \Omega) \text{ in distribution}$$

as $P \equiv P(n) \to \infty$ when $n \to \infty$, where Ω is a $l \times l$ matrix

$$\Omega = \lim_{n\to\infty} \text{var}[P^{-\frac{1}{2}}\sum_{t=R}^{n}\psi(Z_{t+1}, \hat{\beta}_t)],$$

which is a complicated expression as Ω depends on the estimated parameter $\hat{\beta}_t$.

When we compare a single model ($l = 1$) with a benchmark we can use Diebold and Mariano's (1995) test and West's (1996) test, with an appropriate estimator of Ω. When we compare multiple forecasting models ($l > 1$) against a given benchmark model, however, sequential use of Diebold and Mariano (1995) and West (1996) tests may result in a data-snooping bias since the test statistics are mutually dependent due to the use of the same data. To account for possible bias due to data snooping, we use White's (2000) procedure. White's (2000) test statistic for H_0 in (4) is formed as follows:

$$\bar{V} \equiv \max_{1 \le k \le l} \sqrt{P}\bar{\psi}_k,$$

which converges in distribution to $\max_{1 \le k \le l} \mathcal{Z}_k$ under H_0, where the limit random vector $\mathcal{Z} = (\mathcal{Z}_1, \ldots, \mathcal{Z}_l)'$ is $N(0, \Omega)$. White (2000) suggests to use the stationary bootstrap of Politis and Romano (1994, PR) to obtain the null distribution of $\bar{V}$. This gives appropriate p-values for testing the null hypothesis that the best model has no predictive superiority relative to the benchmark (White, 2000, Corollary 2.4). The p-value is called the 'Reality Check p-value' for data snooping. White (2000, Proposition 2.5) also shows that the test's level can be driven to zero at the same time the power approaches to one as $\bar{V}$ diverges at rate $P^{\frac{1}{2}}$ under the alternative.

In our application, we will evaluate the predictive ability of $l = 10$ or $l = 11$ FC models using the three criteria. For example, if we compare the FC_k model with the benchmark linear model using mean trading returns, then we set

$$\bar{\psi}_k = MTR_P^k - MTR_P^{\text{Benchmark}}, \quad k = 1, ..., l.$$

4.3 Results of predictive ability tests

We compare the FC models of (2) with a benchmark linear models in (1). P_{RC} denotes the p-values of White (2000) test computed using PR's stationary bootstrap. The Bootstrap Reality Check p-values are computed with 1,000 bootstrap resamples and the bootstrap smoothing parameter $q = 0.5$. See PR or White (2000) for the details. The other values of q (say, $q = 0.25$, 0.75) give similar p-values (not reported). The bootstrap p-value of White's (2000) test with $l = 1$ are reported next to the values of criterion functions, and the 'bootstrap reality check p-values' to compare the ten or eleven FC models ($l = 10$ or 11) with the benchmark are reported in the last rows of Tables 1-3, which is to test for the null hypothesis that the best of the ten FC models has no predictive superiority over the benchmark linear model. The difference between the p-values with $l = 1$ and $l = 10$ (or 11) gives an estimate of the data-snooping bias, which may be substantial, and enables to quantify the effects of blind specification search and eliminate our illusions to confuse the spurious with the salient.

The statistically significant nonlinearities in conditional mean found in in-sample analysis are not generally carried over to significant out-of-sample forecasts, after accounting for data-snooping. As expected the choice of the loss function directly affects the forecast evaluation results. Some significant out-of-sample forecast improvement of the FC models is found in terms of MCD, for Citicorp and GE. The predictive performance of the FC models in terms of MSE is generally dominated by a linear model.

5. Conclusions

This article has demonstrated the relation between the impact of overnight returns on the intra-daily returns and volatility in daily stock returns on the S&P500 index and on two large U.S. firms (Citicorp and GE). In terms of in-sample goodness of fit, we do find some significant evidence that the impact of the overnight surprises on the intra-daily returns may depend on the state of daily volatility. However, the statistically significant nonlinear responses of intra-daily return to the overnight surprises found in the in-sample analysis are not generally carried over to significant out-of-sample forecasts, after accounting for data-snooping. As expected the choice of the loss function directly affects the forecast evaluation results. There are many possible reasons for the rather disappointing results. One is that the nonlinear models used are not the most suitable ones. Another possible reason is that nonlinearities may be exogenous, arising from outliers, structural shifts, and government intervention, which may render various nonlinearity tests to reject while not being useful for out-of-sample forecasts. It is also possible that the nonlinearity in conditional mean of these

series may not be strong enough to be exploited for forecasting. It is important to explore these possible reasons, but this is beyond the scope of this paper, and has to be left for further research.

References

Amihud, Y. and H. Mendelson (1987), "Trading Mechanisms and Stock Returns: An Empirical Investigation", *Journal of Finance*, 62, 533-553.

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

Brock, W., D. Hsieh and B. LeBaron (1991), *Nonlinear Dynamics, Chaos, and Instability: Statistical Theory and Economic Evidence*. MIT Press: Cambridge, MA.

Cai, Z., J. Fan, and Q. Yao (2000), "Functional-coefficient Regression Models for Nonlinear Time Series", *Journal of American Statistical Association,* 95(451), 941-956.

Cao C., E.Ghysels, and F. Hathaway (2000), "Price Discovery without Trading: Evidence from the Nasdaq Pre-Opening, *Journal of Finance*, 55, 1339-1365.

Chen, R. and R. S. Tsay (1993), "Functional-coefficient Autoregressive Models", *Journal of American Statistical Association*, 88, 298-308.

Clark, P. (1973), "A Subordinated Stochastic Process Model with Finite Variances for Speculative Prices", *Econometrica*, 55, 987–1008.

Cooper, M. (1999), "Filter Rules Based on Price and Volume in Individual Security Overreaction", *Review of Financial Studies*, 12(4), 901-935.

Cornell, B. (1981), "The Relationship between Volume and Price Variability in Futures Markets", *Journal of Futures Markets*, 1, 303-316.

Diebold, F.X. and R.S. Mariano (1995), "Comparing Predictive Accuracy", *Journal of Business and Economic Statistics*, 13, 253-263.

Dow, J. and G. Gorton (1993),"Trading, Communication and the Response of Asset Prices to News", *Economic Journal*, 103, 639–46.

Efron, B. (1979), "Bootstrap Methods: Another Look at the Jackknife", *Annals of Statistics*, 7, 1-26.

Epps, T.W., and M.L. Epps (1976), "The Stochastic Dependence of Security Price Changes and Transaction Volumes: Implications for the Mixture-of-Distribution Hypothesis", *Econometrica*, 44, 305-321.

Fan, J., and I. Gijbels (1996), *Local Polynomial and Its Applications.* Chapman and Hall: London.

Fan, J., C.M. Zhang, and J. Zhang (2001), "Generalized Likelihood Ratio Statistics and Wilks Phenomenon", *Annals of Statistics,* 29(1), 153-193.

Gallant A.R., C.-T. Hsu, and G. Tauchen (1999), "Using Daily Range Data to Calibrate Volatility Diffusions and Extract the Forward Integrated Variance", *Review of Economics and Statistics*, 81, 617-631.

Gallo, G.M. and B. Pacini (1998), "Early News is Good News: The Effects of Market Opening on Volatility", *Studies in Nonlinear Dynamics and Econometrics*, 2, 115-132.

Garman M.B. and M.J Klass (1980), "On the Estimation of Security Price Volatilities from Historical Data", *Journal of Business*, 53, 67-78.

Glosten, L. R., R. Jaganathan and D. Runkle (1993), "On the Relationship Between the Expected Value and the Volatility of the Nominal Excess Return on Stocks", *Journal of Finance*, 48, 1779-1801.

Granger, C.W.J. (1992), "Forecasting Stock Market Prices: Lessons for Forecasters", *International Journal of Forecasting*, 8, 3-13.

Granger, C.W.J. (1998), "Real and Spurious Long-Memory Properties of Stock-Market Data: Comment", *Journal of Business and Economic Statistics*, 16(3), 268-269.

Granger, C.W.J. and Z. Ding (1995), "Some Properties of Absolute Return: An Alternative Measure of Risk," *Annales d'Economie et de Statistique*, 40, 67-92.

J.P. Morgan (1996), *Riskmetrics Technical Manual*, 4rd ed.

Liu, R.Y. (1988), "Bootstrap Procedures under Some Non-iid Models", *Annals of Statistics*, 16, 1697-1708.

Lo, A.W. and A.C. MacKinlay (1999), *A Non-Random Walk Down Wall Street*, Princeton University Press: Princeton.

Parkinson, M. (1980), "The Extreme Value Method for Estimating the Variance of the Rate of Return", *Journal of Business*, 53, 61-66.

Politis, D. N. and J. P. Romano (1994), "The Stationary Bootstrap", *Journal of American Statistical Association,* 89, 1303-1313.

Priestley, M.B. (1980), "State-dependent Models: A General Approach Nonlinear Time Series Analysis", *Journal of Time Series Analysis,* 1, 47-71.

Romer, D. (1993), "Rational Asset-Price Movements without News", *American Economic Review*, 83, 1112–30.

Stoll, H.R. and R.E. Whaley (1990), "Stock Market Structure and Volatility", *Review of Financial Studies*, 3, 37–71.

Sullivan,R., A. Timmermann, and H. White (1999), "Data Snooping, Technical Trading Rule Performance, and the Bootstrap," *Journal of Finance*, 54, 1647-1692.

Tauchen, G.E. and M. Pitts (1983), "The Price Variability-Volume Relationships on Speculative Markets", *Econometrica*, 51(2), 485-505.

Tong, H. (1990), *Nonlinear Time Series: A Dynamic System Approach*. Clarendon Press: Oxford.

West, K.D. (1996), "Asymptotic Inference about Predictive Ability", *Econometrica,* 64 1067-1084.

White, H. (2000), "A Reality Check for Data Snooping", *Econometrica*, 68(5), 1097-1126.

Wu, C.F.J. (1986), "Jackknife, Bootstrap, and Other Resampling Methods in Regression Analysis", *Annals of Statistics*, 14, 1261-1350.

Table 1. S&P500

	Test for FC (in-sample)				Comparing Predictive Ability (out-of-sample)					
Model	j	T_n	P_B	P_W	MSE_P	P_{RC}	MTR_P	P_{RC}	MCD_P	P_{RC}
L					1.752		0.022		0.507	
FC_1	4	0.030	(0.00)	(0.00)	1.709	(0.198)	0.002	(0.655)	0.497	(0.707)
FC_2	5	0.024	(0.00)	(0.00)	1.675	(0.118)	0.058	(0.083)	0.507	(0.467)
FC_3	3	0.027	(0.00)	(0.00)	1.719	(0.245)	0.027	(0.416)	0.508	(0.421)
FC_4	4	0.024	(0.00)	(0.00)	1.693	(0.173)	0.040	(0.216)	0.509	(0.367)
FC_5	4	0.025	(0.00)	(0.00)	1.727	(0.300)	0.005	(0.887)	0.499	(0.871)
FC_6	4	0.024	(0.00)	(0.00)	1.716	(0.225)	0.029	(0.373)	0.503	(0.654)
FC_7	5	0.010	(0.02)	(0.08)	1.699	(0.189)	0.067	(0.028)	0.524	(0.016)
FC_8	5	0.023	(0.00)	(0.00)	1.673	(0.128)	0.046	(0.159)	0.504	(0.627)
FC_9	3	0.007	(0.08)	(0.24)	1.855	(0.808)	0.027	(0.436)	0.514	(0.185)
FC_{10}	5	0.011	(0.00)	(0.03)	1.743	(0.424)	0.046	(0.184)	0.511	(0.297)
						(0.344)		(0.356)		(0.247)

TABLE 2. Citicorp

	Test for FC (in-sample)				Comparing Predictive Ability (out-of-sample)					
Model	j	T_n	P_B	P_W	MSE_P	P_{RC}	MTR_P	P_{RC}	MCD_P	P_{RC}
L					5.616		-0.134		0.462	
FC_1	4	0.021	(0.00)	(0.07)	13.620	(1.000)	0.042	(0.088)	0.509	(0.046)
FC_2	4	0.027	(0.00)	(0.00)	20.400	(0.984)	0.015	(0.131)	0.505	(0.055)
FC_3	4	0.016	(0.00)	(0.14)	13.077	(1.000)	0.072	(0.064)	0.509	(0.043)
FC_4	4	0.012	(0.02)	(0.34)	13.205	(1.000)	0.035	(0.118)	0.498	(0.094)
FC_5	4	0.014	(0.03)	(0.12)	13.484	(1.000)	0.022	(0.113)	0.498	(0.095)
FC_6	3	0.026	(0.00)	(0.03)	13.480	(1.000)	-0.014	(0.191)	0.494	(0.125)
FC_7	5	0.007	(0.03)	(0.21)	12.587	(1.000)	0.054	(0.088)	0.502	(0.076)
FC_8	2	0.054	(0.00)	(0.00)	17.677	(0.999)	0.004	(0.154)	0.503	(0.070)
FC_9	5	0.005	(0.15)	(0.55)	12.504	(1.000)	0.046	(0.095)	0.503	(0.069)
FC_{10}	5	0.009	(0.00)	(0.24)	17.292	(0.980)	0.055	(0.069)	0.506	(0.042)
FC_{11}	5	0.002	(0.76)	(0.82)	12.976	(1.000)	0.072	(0.063)	0.506	(0.048)
						(1.000)		(0.106)		(0.068)

TABLE 3. GE

	Test for FC (in-sample)				Comparing Predictive Ability (out-of-sample)					
Model	j	T_n	P_B	P_W	MSE_P	P_{RC}	MTR_P	P_{RC}	MCD_P	P_{RC}
L					3.293		0.064		0.480	
FC_1	4	0.034	(0.00)	(0.00)	6.111	(1.000)	0.076	(0.443)	0.523	(0.042)
FC_2	5	0.020	(0.00)	(0.04)	6.572	(1.000)	0.074	(0.475)	0.523	(0.042)
FC_3	5	0.017	(0.00)	(0.07)	5.723	(1.000)	0.086	(0.392)	0.526	(0.036)
FC_4	5	0.014	(0.00)	(0.20)	13.875	(0.921)	0.070	(0.449)	0.523	(0.036)
FC_5	5	0.019	(0.00)	(0.05)	5.711	(1.000)	0.070	(0.472)	0.521	(0.040)
FC_6	4	0.027	(0.00)	(0.00)	7.468	(0.990)	0.058	(0.502)	0.520	(0.053)
FC_7	4	0.010	(0.02)	(0.37)	6.462	(1.000)	0.077	(0.423)	0.518	(0.052)
FC_8	5	0.020	(0.00)	(0.07)	6.600	(1.000)	0.073	(0.455)	0.521	(0.040)
FC_9	4	0.003	(0.56)	(0.72)	6.662	(1.000)	0.122	(0.264)	0.531	(0.023)
FC_{10}	0	0.085	(0.00)	(0.00)	5.745	(1.000)	0.082	(0.396)	0.520	(0.059)
FC_{11}	0	0.080	(0.00)	(0.00)	5.749	(1.000)	0.074	(0.488)	0.523	(0.041)
						(1.000)		(0.363)		(0.038)

Notes: (1) Both naive and wild bootstraps are used, whose p-values are denoted as P_B and P_W, respectively. The bandwidth h is chosen to minimize $\text{AMS}(h)$, among the 11 values of $h = 2^j n^{-1/5}$ where $j = -5, -4, \ldots, 4$, and 5. (2) P_{RC} denotes denote the p-values of White (2000) test. The P_{RC} values in all rows (except in the last row) are to compare each of FC_k with the benchmark linear model L. The P_{RC} values in the last row are to compare the best of the eleven FC model with the benchmark model L.