Volatility Predictions — SIG Data Exercise

Zheng Gao

30 Nov 2019

CONTENTS

1 VOLATILITY DEFINED, AND ESTIMATED

Among the many possible definitions, we shall refer to **volatility** exclusively as the standard deviation of the annualized percent return. Specifically, if we denote the (random) return over the next m trading days as $r^{(m)}$ (defined in Section 2 below), and denote its standard deviation as

$$\sigma_{(m)} = \sqrt{\mathrm{Var}(r^{(m)})},$$

then the volatility measured in annualized percent return is defined to be

$$\widetilde{\sigma}_{(m)} = \sigma_{(m)}\sqrt{252/m},$$

where we make the assumption that there are 252 trading days per year. In particular, the goal of this project is to estimate the volatility over the next month, i.e.,

$$\widetilde{\sigma}_{(21)} = \sigma_{(21)}\sqrt{252/21} = \sqrt{12\,\mathrm{Var}(r^{(21)})}, \tag{1}$$

where we assume on average $m = 21$ trading days per month. We summarize our estimates in Table 1.

Stock	Point estimate	95% confidence interval
a	27.12	(27.02, 27.22)
b	55.50	(51.15, 60.66)
c	77.72	(77.43, 78.02)
d	38.28	(35.28, 41.84)
e	32.73	(32.61, 32.86)
f	37.00	(36.86, 37.13)

Table 1: Summary of estimates of the volatility (1), measured in annualized percentage points.

2 DATA PREPROCESSING AND EXPLORATORY DATA ANALYSIS

Careful data preprocessing is the pre-requisite of valid inference and predictions downstream. We document the steps we have taken to clean the dataset in this section.

A preliminary inspection reveals the following issues in the dataset:

- **Missingness**. Four out of six stocks included in the dataset have missing data to various degrees; these missingness patterns are summarized in the 2$^{\text{nd}}$ column of Table 2 below.

- **Special values**. Two out of the six stocks have apparently erroneous records. In particular, stocks a and d have approximately 0.1% of their data replaced with 0's and 1's, respectively; see 3rd column of Table 2 below.

To facilitate downstream analysis, we replace all missing and special values (NA's, 0's and and 1's) with the last (i.e., closest prior) available data point that are neither missing nor taking the special values.

Stock	Missingness	Special values	Special features of the return series
a	0.07%	'0': 0.09%	ACF slightly negative at lags 1, 2, and 3.
b	-	-	ACF significantly negative at lag 1. "Bounce-back" behavior on 1-minute scale.
c	0.03%	-	A single large jump on day 149.
d	0.02% -	'1': 0.09%	ACF significantly negative at lag 1. "Bounce-back" behavior on 1-minute scale.
e	-	-	
f	1.39%	-	No movement over 99% of the time (on 1-minute scale)

Table 2: Summary of missingness, special values, and special features of the dataset.

We perform further inspections of the dataset, through the return series

$$r_t^{(m)} := \log x_t^{(m)} - \log x_{t-1}^{(m)} \approx x_t^{(m)} / x_{t-1}^{(m)} - 1,$$

where $x_t^{(m)}$ is the price series sampled at frequency m (see Section 3). Specifically, we inspect the autocorrelation functions (ACF) of $r_t^{(m)}$, and record any special features observed. These observations are tabulated in Table 2, column 4.

This process of exploratory data analysis (EDA) is illustrated with price data of stock d in Section 5 below. We also point out an interesting "bounce-back" behavior in the return series of stock b and d on the 1-minute scale, but not present in all other series.

3 (NON)INDEPENDENCE AND SAMPLING FREQUENCY

Returning to the problem of estimating annualized volatility, we consider a range of sampling frequencies for estimating the desired quantity (1).

3.1 *Sampling data on monthly intervals*

Since the goal is equivalent to estimating $\sigma^2_{(21)} = \mathrm{Var}(r^{(21)})$, a natural choice is to sample the data at monthly intervals. That is, we take the closing prices of every 21 trading days.

$$x_t^{(21)} = \text{closing price of the stock on the } (21 \times t)^{\text{th}} \text{ day},$$

and estimate (1) via the historical volatility

$$s^2_{(21)} = \widehat{\mathrm{Var}}\left(r_t^{(21)}\right). \tag{2}$$

If we assume *stationary monthly returns*, then this approach would yield an unbiased estimate of $\sigma^2_{(21)}$. The problem, however, is that given a year's worth of history, we can only calculate at most 12 monthly returns — and this leads to a large variance associate with the estimator (2). In particular, the $(1-\alpha)$-level confidence interval

$$\frac{(n-1)s^2_{(21)}}{\chi^2_{n-1,\alpha/2}} \le \sigma^2_{(21)} \le \frac{(n-1)s^2_{(21)}}{\chi^2_{n-1,1-\alpha/2}} \tag{3}$$

can be much wider than we would like it to be when n is small (i.e., $n = 12$).

We discuss a possible remedy to this problem next.

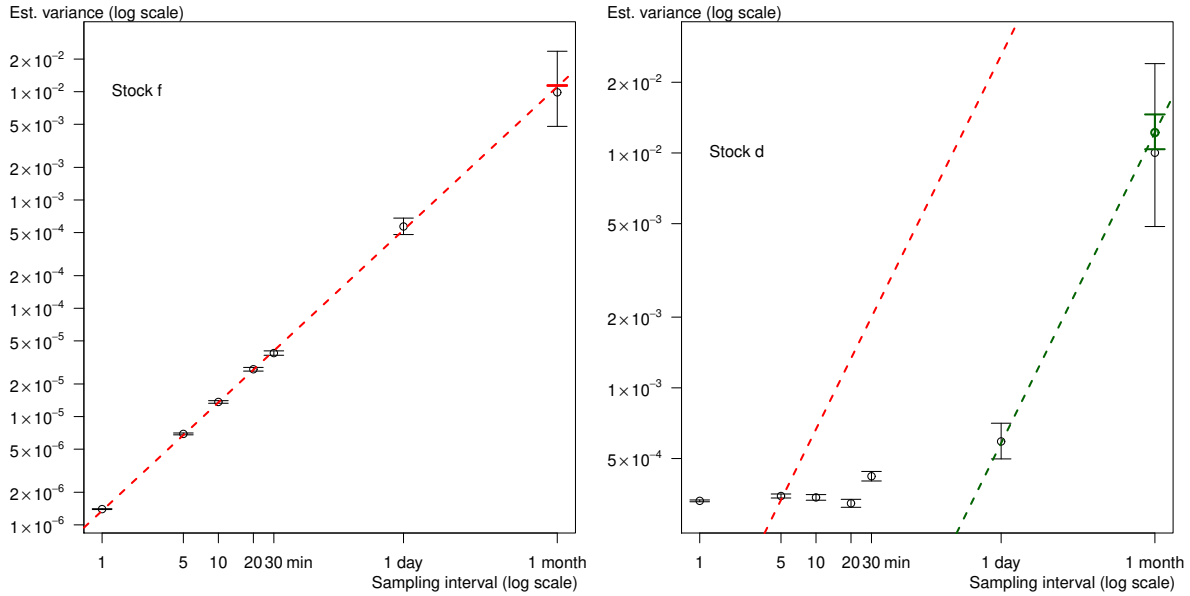

Figure 1: **Estimated variance of returns of Stock f (left) and Stock d (right) versus length of sampling intervals.** We also overlay the confidence intervals (3) (black solid bars), the confidence intervals of the pooled estimates (6) using all frequencies (red solid bars), and of the pooled estimates using only frequencies $m = 1$ and $m = 21$ (green solid bar, for Stock d). We add the "regression lines" through the origin and the pooled estimates for visual aid (dotted lines). Contrast Stock f, which demonstrates independence of its increments — and hence has variances scaling linearly — with Stock d, which is unlikely to have independent increments on the minute-scale. Further evidence (see Section 5) shows that returns in Stock d indeed display strong autocorrelations at 1-minute lag.

3.2 *Sampling data at higher frequencies*

If the stock prices have *independent* — in addition to stationary — increments (so that $\{x_t\}$ is a so-called Lévy process), then we have

$$\sigma^2_{(m)} = \mathrm{Var}\left(r^{(m)}\right) = \frac{m}{m'}\mathrm{Var}\left(r^{(m')}\right) = \frac{m}{m'}\sigma^2_{(m')}. \tag{4}$$

In other words, the variances of the returns scales linearly as a function of time. Consequently, we may choose to estimate $\sigma^2_{(m')}$ instead of $\sigma^2_{(m)}$ where $m' < m$, so that effectively (m/m')-times more samples can be collected from the same length of historical data.

Let us denote the time-series sub-sampled daily as

$$x^{(1)}_t = \text{closing price of the stock on the } t^{\text{th}} \text{ day.}$$

Similarly, the time-series sampled at 1-, 5-, 10-, 20-, and 30-minute intervals are denoted

$$x^{(m)}_t, \quad \text{where } m \in \{1/391, 5/391, \ldots, 30/391\},$$

since there are 391 trading minutes per day. We calculate the variance estimates for the return series at each sampling frequency, and examine the linearity of the relationship (4) empirically. Figure 1 displays results for two of the stocks (f and d), and demonstrates two typical patterns: one in which we have approximately independent increments and one where the independent increments does not hold. In the interest of space, the plots for the other four stocks are deferred to Figure 5 attached to the end of this report.

When the assumption of independent increments holds (or in reality, approximately holds), we can pool the variance estimates across frequencies. Since the variance estimates $s^2_{(m)}$ are chi-square distributed and approximately independent across different frequencies, we have

$$\sum_{m=\frac{1}{391},\ldots,\frac{30}{391},1,21} \frac{(n_m - 1)s^2_{(m)}}{\sigma^2_{(m)}} \sim \chi^2_{\sum_m (n_m - 1)},$$

where n_m is the number of observations when we sample at frequency m. Now define

$$s^2_{\text{pooled}} = \frac{21 \sum_m (n_m - 1) s^2_{(m)}/m}{\sum_m (n_m - 1)} \sim \sigma^2_{(21)} \frac{\chi^2_{\sum_m (n_m - 1)}}{\sum_m (n_m - 1)}, \tag{5}$$

and we can construct confidence intervals for the the pooled estimates

$$\frac{\sum_m (n_m - 1) s^2_{\text{pooled}}}{\chi^2_{\sum_m (n_m - 1), \alpha/2}} \leq \sigma^2_{(21)} \leq \frac{\sum_m (n_m - 1) s^2_{\text{pooled}}}{\chi^2_{\sum_m (n_m - 1), 1-\alpha/2}} \tag{6}$$

For stocks with independent increments, the pooled estimates using all frequencies has significantly shorter confidence intervals; c.f., confidence intervals for $\sigma^2_{(21)}$ in Figure 1, left panel.

On the other hand, for stocks with non-independent increments, Relation (4) fails. In this case, we only pool data at frequencies that *do* demonstrate independence in increments. In particular, for stocks b and d, we pool the variance estimates only for data sampled daily ($m = 1$) and monthly ($m = 21$). The resulting estimates are still much more stable than using monthly data alone; c.f., confidence intervals for $\sigma^2_{(21)}$ in Figure 1, right panel.

4 (NON)STATIONARITY AND SAMPLING WEIGHTS

The two assumptions for the pooled esimate (5) must be borne in mind:

- **Independent increments**. This implies that the direction of the movement in stock prices cannot be predicted from past data. The ACF must display no significant deviation from zero at all lags.
- **Stationary increments**. The return series must not exhibit volatility clustering, i.e., the volatility during any two periods of the same length must be the same.

We discussed the role of independence in sampling frequency in the previous section. We now turn to stationarity and its role in determining the relative weights of samples to account for any changes in volatility over time.

Specifically, if there are volatility clustering (as in the case of Stocks b, and to lesser extent, in the other series), then more recent samples should be placed relatively higher weights. I do not have time to go into this, but I see at least two ways to approach the problem:

1. **Models/methods based on stationarity assumptions, e.g., GARCH, which accounts for volatility clustering.** This may not be suitable for series such Stock b, where returns r_t are not stationary; we may be better off modeling the difference series Δx_t which looks roughly stationary.
2. **Models/methods that ditch stationarity assumptions.** This could entail anything, including ad-hoc weighting of samples by recent-ness.

A formal criterion for comparing — and tuning, if necessary — these methods can be pseudo-out-of-sample predictive performances, with some loss function measuring closeness of predicted values to observed.

5 EDA ILLUSTRATED WITH STOCK D

The data pre-processing steps were briefly summarized in Section 2. We provide some further details in the case of Stock d; the steps are illustrated in Figure 2.

Notice that the returns series is roughly stationary. Also, notice that the autocorrelation on the 1-minute scale is significantly negative at lag 1. We investigated this autocorrelation further in Figure 3. The returns display an interesting "bounce-back" behavior, i.e., a large movement is usually followed by a large (and almost equal magnitude) movement in the opposite direction the following minute — and hence the off-axis line in the r_{t+1}-r_t plane. This explains the negative lag-1 auto correlation seen in the ACF. Similar phenomenon takes place in Stock b (but not in the others).

An intra-day volalitity pattern is also observed: volatility tends to be higher in opening and closing minutes,with the first minute showing much larger price movement than any other minute in trading. See Figure 4.

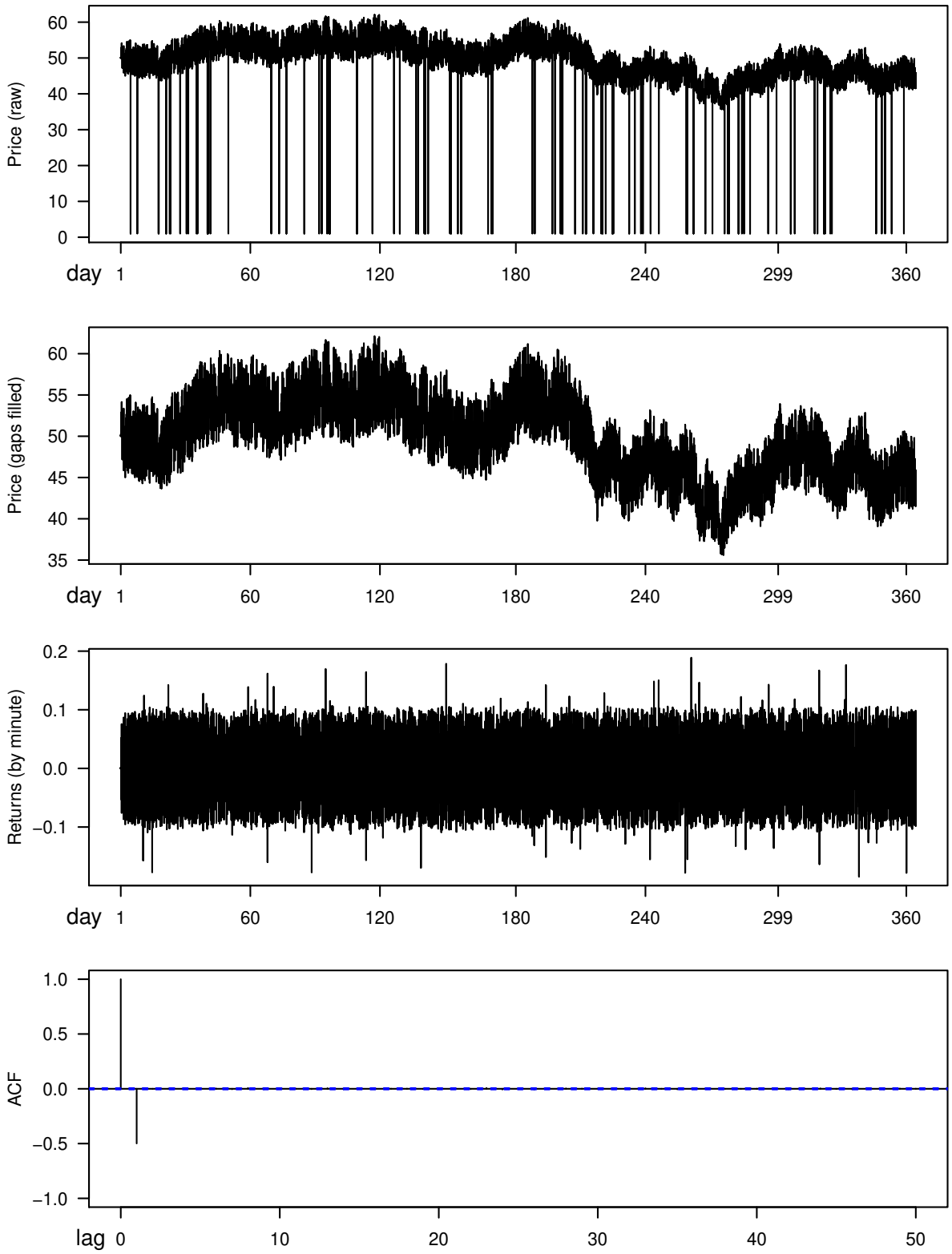

Figure 2: **Exploratory data analysis illustrated with Stock d.** From top to bottom: raw price series, price series with missing and special values replaced, return series, and the ACF of return series on 1-minute scale. The returns are roughly stationary, but not independent, as the ACF clearly indicates. This is corroborated by Figure 1, where we see sub-linear scaling of variances in returns as a function of sampling intervals.

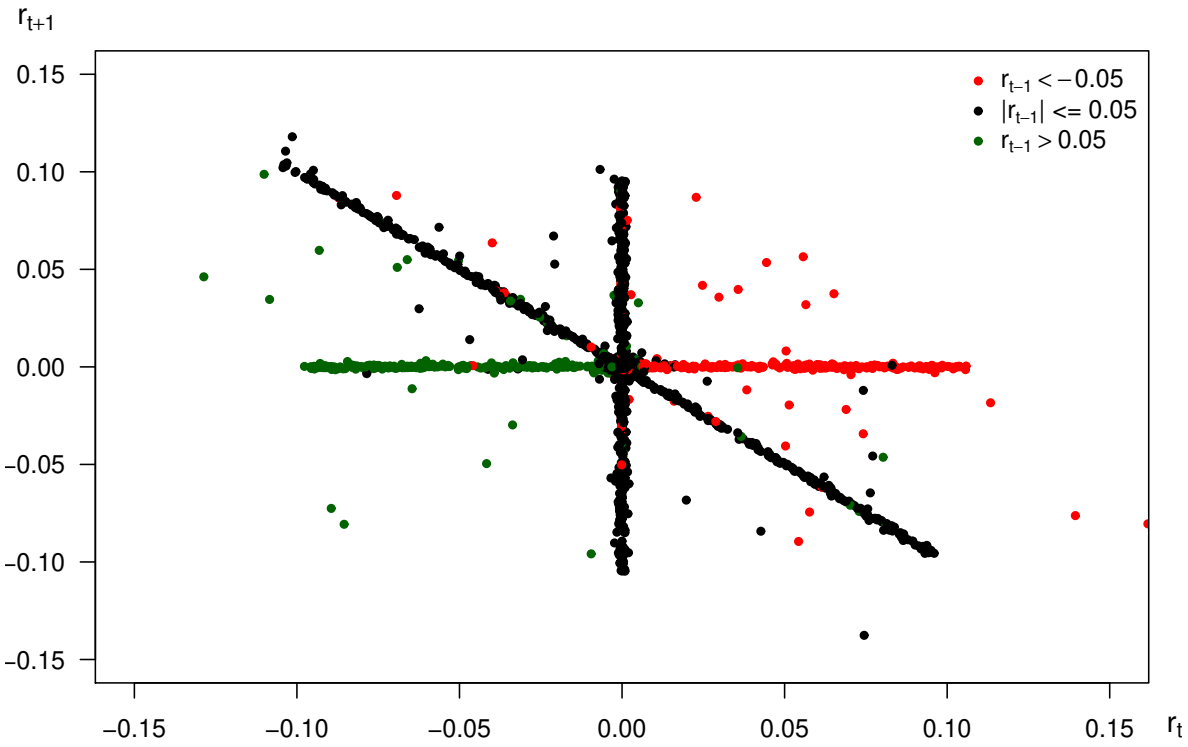

Figure 3: **Lagged returns on Stock d, r_{t+1} versus r_t, colored by r_{t-1}.** The returns display an interesting "bounce-back" behavior, i.e., a large movement is usually followed by a large (and almost equal magnitude) movement in the opposite direction the following minute — and hence the off-axis line in the r_{t+1}-r_t plane. This explains the negative lag-1 auto correlation seen in the ACF. Similar phenomenon takes place in Stock b (but not in the others).

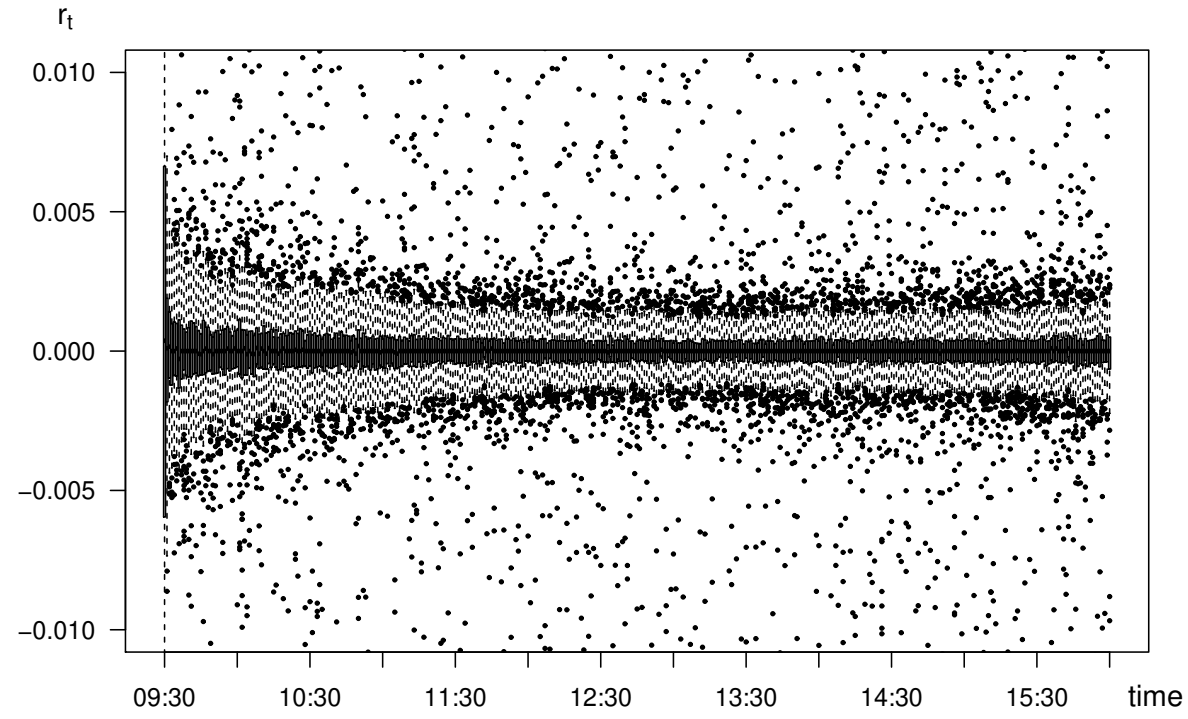

Figure 4: **Intra-day volatility pattern of Stock d.** Volatility tends to be higher in opening and closing minutes, with the first minute showing much larger price movement than any other minute in trading.

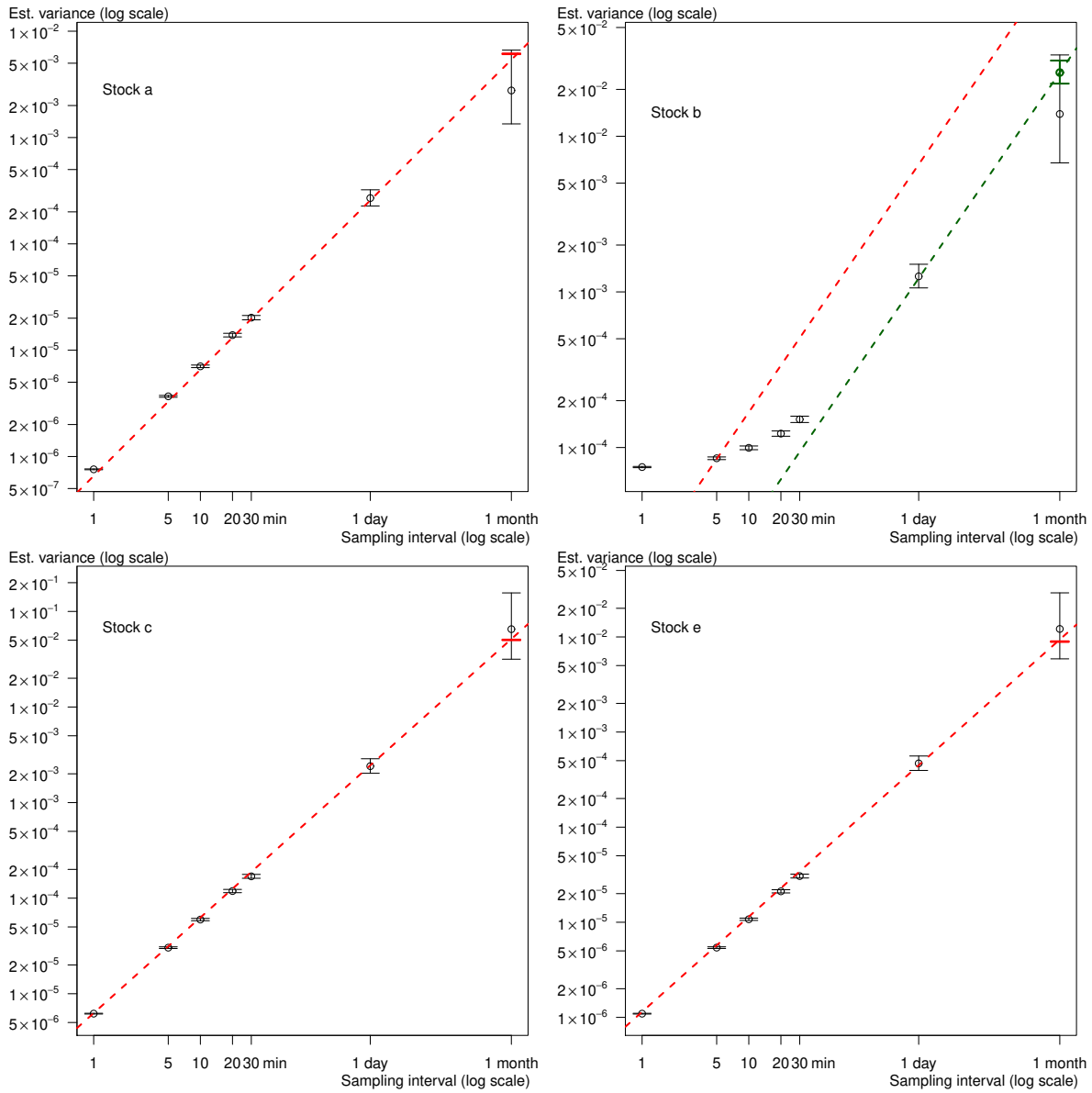

Figure 5: **Estimated variance of returns of Stock a (upper-left), Stock b (upper-right), Stock c (lower-left), and Stock e (lower-right), versus length of sampling intervals.** We also overlay the confidence intervals (3) (black solid bars), the confidence intervals of the pooled estimates (6) using all frequencies (red solid bars), and of the pooled estimates using only frequencies $m = 1$ and $m = 21$ (green solid bar, for Stock d). We add the "regression lines" through the origin and the pooled estimates for visual aid (dotted lines).