

Intraday Volatility Modelling Introduction

- 第一日之内的 Volatility (用的是 Intraday Returns)
也可以说是 realized variance
(用 Returns 算出来的 variance)
- 在这节课之前，我们用的是 daily returns.
但现代金融市场的数据几乎是连续的 → 这启发了我们使用更高频的数据

首先要注意的一点是：用高频数据算 mean return 是没有意义的

$$\hat{\mu} = \frac{1}{T} \sum_{t=0}^T (\ln(S_t) - \ln(S_{t-1})) = \frac{1}{T} (\ln(S_T) - \ln(S_0))$$

只有第 0 & 最后一个数据起作用，中间没什么贡献

但对于方差来说：

$$\delta^2 = \frac{1}{T} \sum_{t=0}^T (\ln(S_t) - \ln(S_{t-1}) - \hat{\mu})^2 \text{ 可以非零，虽然通常我们认为是零}$$

- 用高频数据算方差是非常非常略的

tremendous benefits !

类似于用日数据算月方差，我们可以用小时数据算日方差。用分钟数据算日方差

这样会很精确

假设有 1 24 小时交易的东西 (e.g. currency pairs USD/EUR)

现在我们令 m 为观测数。 j 是第 j 观测。If $m=1440 \rightarrow$ 按分钟观测

$$R_{t+\frac{j}{m}} = \ln(S_{t+\frac{j}{m}}) - \ln(S_{t+\frac{j-1}{m}})$$

Realized variance

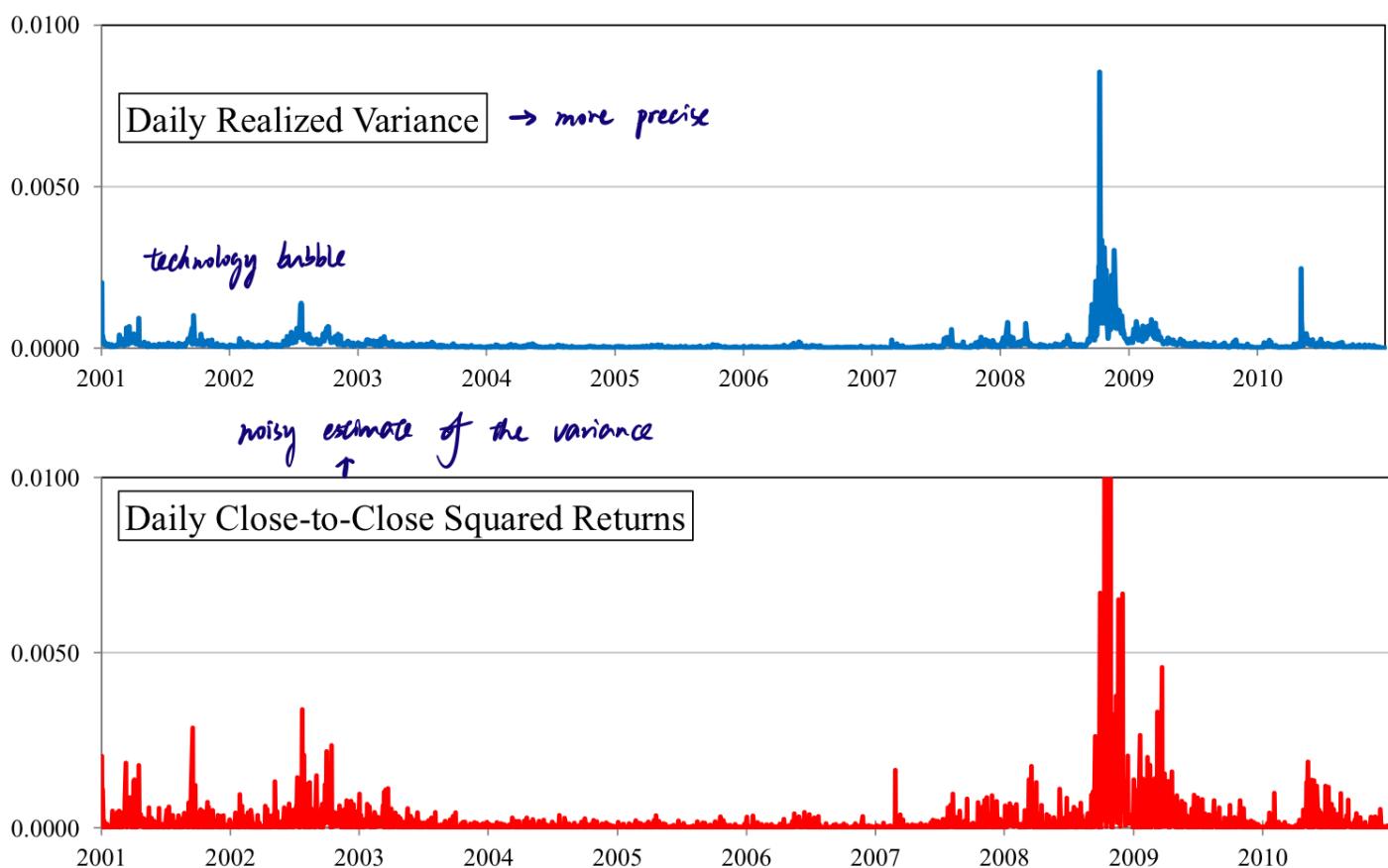
$$RV_{t+1}^m = \sum_{j=1}^m R_{t+\frac{j}{m}}^2$$

前面没有 $\frac{1}{m}$ → 如果有就是在算分钟方差了。
但我们是想算日方差的

这里我们没有减 mean → 对于分钟而言，这实在是太小了

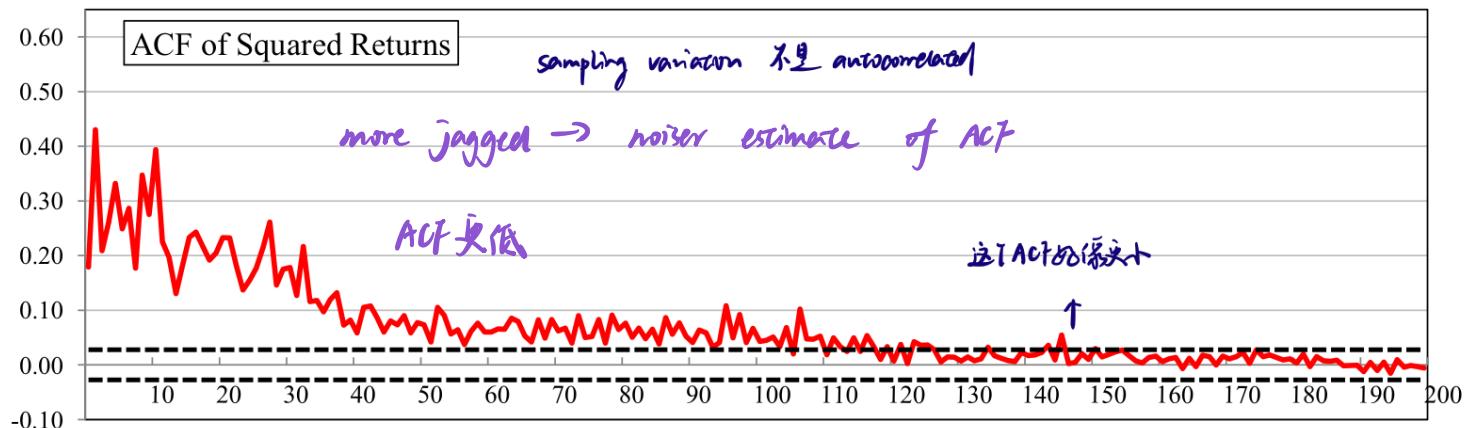
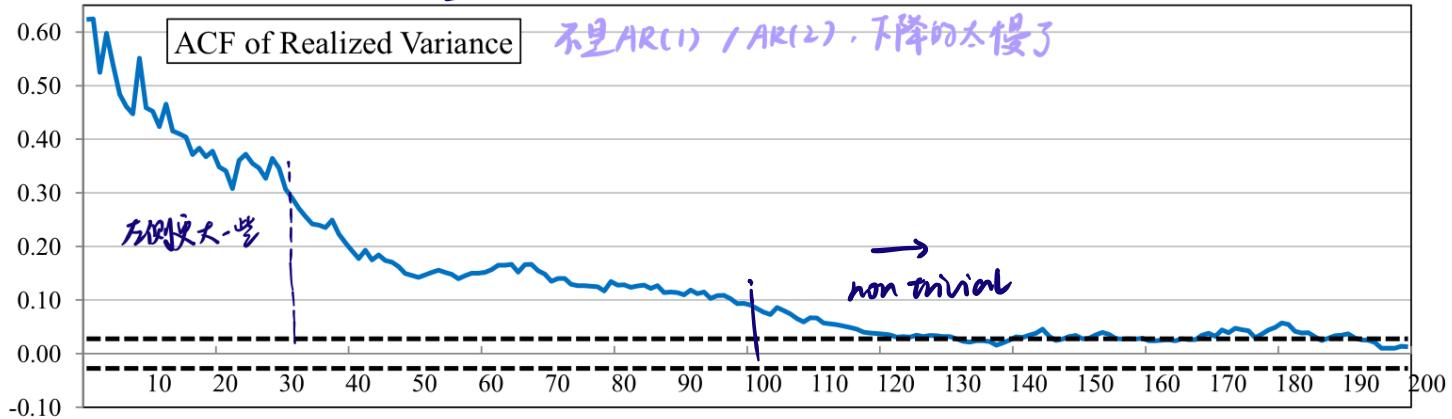
接下来介绍一下 Realized Variance 的四个特征：

- a. RVs are MUCH MORE PRECISE indicators of daily variance than are daily squared returns → 为什么要预测



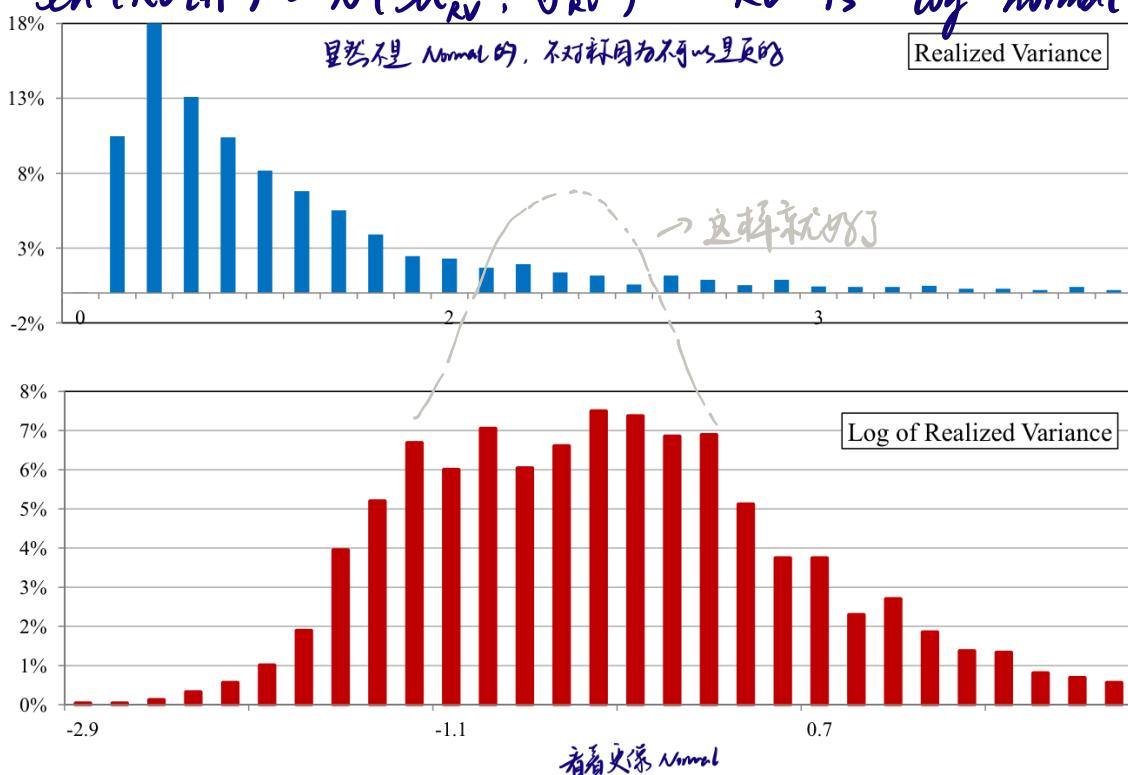
这两下图看起来趋势差不多 close-to-close is noisy estimate. 看起来 more jagged. RVs are less noisy
其实就是说 RV 是 pretty good estimator

2. RV is extremely persistent \rightarrow easy to forecast \rightarrow 为什么能预测



很多 squared return 只是 sampling variation. 而 sampling variation is not autocorrelated \rightarrow 下面的 ACF 更低

3. $\ln(RV_{t+1}^m) \sim N(\mu_{RV}, \sigma_{RV}^2) \rightarrow RV$ is log normal \rightarrow 如何预测



-T问题是：我们用 GARCH model 算出来的 RV 会像上面的图一样么？

→ Remain to be tested

4. $\frac{R_{t+1}}{\sqrt{RV_{t+1}^m}} \stackrel{i.i.d.}{\sim} N(0, 1)$

算是 scaled by s.d. RV_{t+1}^m 是 good estimate of variance

和 GARCH 很像 ($Z_t \sim N(0, 1)$)

不是一下子就能用的 因为 RV_{t+1}^m 在一天结束之后才能知道
但如果能够预测的话，那是极好的

$RV_{t+1|t}^m$ 是我们的预测，那么

$$R_{t+1} / \sqrt{RV_{t+1|t}^m} \stackrel{i.i.d.}{\sim} N(0, 1)$$

接下来我们利用四 T stylized facts 来进行一波预测：

(主要是通过前三 T 来预测)

而对于第四 T fact，这意味着我们可以预测 $R_{t+1|t}^m + \hat{\text{Var}}$

具体如何预测呢？

AR(1) model :

$$RV_{t+1}^m = \phi_0 + \phi_1 RV_t^m + \varepsilon_{t+1}$$

ϕ_0, ϕ_1 用 OLS 来估计，那么 $RV_{t+1|t}^m = \phi_0 + \phi_1 RV_t^m$

还能再给点么？

呆

→ RV 其实不是 true variance, 会有 measurement error

ARMA(1,1)

$$\ln(RV_{t+1}^m) = \phi_0 + \phi_1 \ln(RV_t^m) + \theta_1 \varepsilon_t + \varepsilon_{t+1}, \text{ with } \varepsilon_{t+1} \stackrel{i.i.d.}{\sim} N(0, \sigma_\varepsilon^2)$$

ARMA(1,1)有什么好处？它反映了 measurement error

model for "true" variance 是 $\ln V_{t+1} = \phi_0 + \phi_1 \ln V_t + \varepsilon_{t+1}$

但是我们既没 V_t 也没 V_{t+1} ，只有 RV_t 和 RV_{t+1}

$RV_t = V_t + \eta_t$ measurement error due to sampling variation in RV_t

$$\begin{aligned} \ln RV_{t+1} &= \phi_0 + \phi_1 (\ln V_t - \eta_t) + \varepsilon_{t+1} + \eta_{t+1} \\ &\quad \uparrow \qquad \qquad \uparrow \\ &\quad \text{measurement error} \qquad \text{measurement error} \\ &= \phi_0 + \phi_1 (\ln V_t) + \varepsilon_{t+1} + \eta_{t+1} - \phi_1 \eta_t \end{aligned}$$

$$\varepsilon_{t+1} + \eta_{t+1} - \phi_1 \eta_t = \varepsilon_{t+1} - \theta_1 \varepsilon_t$$

下面我们来预测。注意

$$RV_{t+1|t}^m = E_t [RV_{t+1}] = \bar{E}_t [\exp(\ln(RV_{t+1}^m))] \neq \exp(E_t[\ln(RV_{t+1}^m)])$$

$$\varepsilon_{t+1} \sim N(0, \sigma_\varepsilon^2) \Rightarrow E[\exp(\varepsilon_{t+1})] = \exp\left(\frac{\sigma_\varepsilon^2}{2}\right)$$

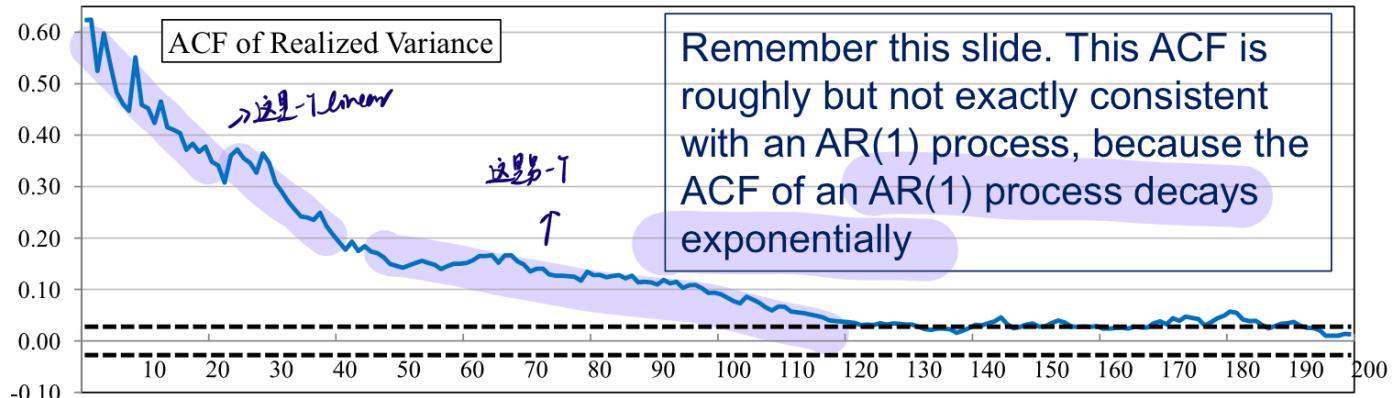
在 AR(1) model 中

$$\begin{aligned} RV_{t+1|t}^m &= E_t[\exp(\phi_0 + \phi_1 \ln RV_t^m + \varepsilon_{t+1})] \\ &= \exp(\phi_0 + \phi_1 \ln RV_t^m) E_t[\exp(\varepsilon_{t+1})] \\ &= (RV_t^m)^{\phi_1} \exp(\phi_0 + \frac{\sigma_\varepsilon^2}{2}) \end{aligned}$$

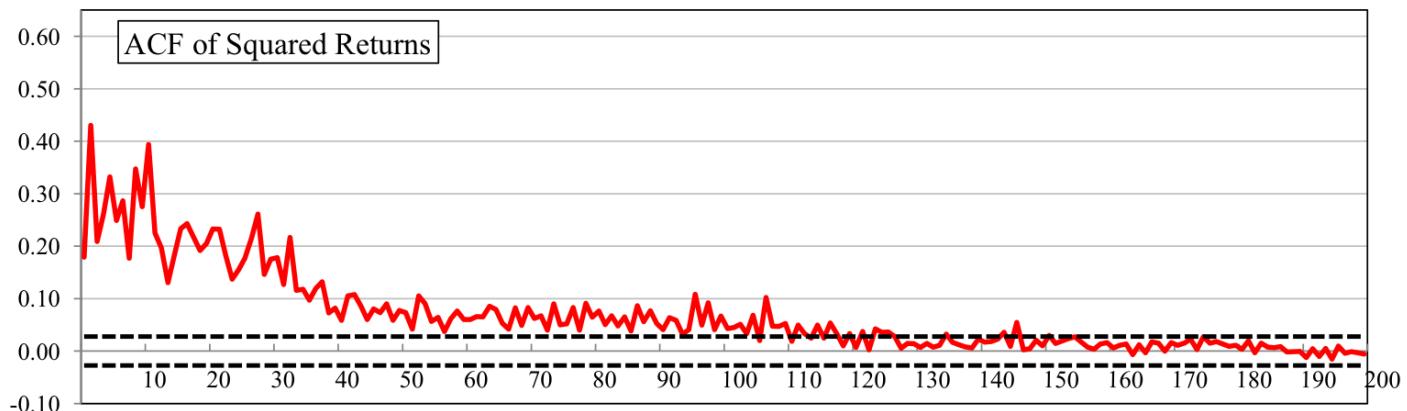
如果我们用 ARMA(1,1) model

$$RV_{t+1|t}^m = (RV_t^m)^{\phi_1} \exp(\phi_0 + \theta_1 \varepsilon_t + \frac{\sigma_\varepsilon^2}{2})$$

接下来我们回顾一下上面用过的图



Remember this slide. This ACF is roughly but not exactly consistent with an AR(1) process, because the ACF of an AR(1) process decays exponentially



我们考虑一下 AR(p) model :

因为 P 要选的很大 \rightarrow AR(21) (1月有21天)

$$\ln(RV_{t+1}^m) = \phi_0 + \phi_1 \ln(RV_t^m) + \phi_2 \ln(RV_{t-1}^m) + \dots + \phi_{21} \ln(RV_{t-20}^m) + \varepsilon_{t+1}$$

这些应该都是正确的. 为啥呢? future realized variance 和今天的 RV 有因果逻辑

注意: 我们是在估 22 个系数 ($\phi_0 \sim \phi_{21}$). 有的可能小于 0. 有的且大了

那我们怎么做呢? 下面引入 HAR (Heterogeneous Autoregressions)

也叫 mixed-frequency. 它是怎么做的呢?

将 21 个交易日简化为 3 个: $RV_{D,t}$, $RV_{W,t}$, $RV_{M,t}$ 分别是日, 周, 月的 RV

$$RV_{D,t} = RV_t$$

$RV_{W,t}$ 是从 RV_{t-4} 到 RV_t 的均值: $\frac{1}{5} \times (RV_{t-4} + RV_{t-3} + RV_{t-2} + RV_{t-1} + RV_t)$

$RV_{M,t}$ 是从 RV_{t-20} 到 RV_t 的均值: $\frac{1}{21} \times (RV_{t-20} + \dots + RV_t)$

因为一周有五个交易日, 一月有二十个交易日

HAR Model :

$$RV_{t+1} = \phi_0 + \phi_D RV_{D,t} + \phi_W RV_{W,t} + \phi_M RV_{M,t} + \varepsilon_{t+1}$$

在数学上有一点： RV_t 在 $RV_{D,t}$, $RV_{W,t}$, $RV_{M,t}$ 中都出现了

$$RV_{t-1} \sim RV_{t-4} \text{ 在 } RV_{W,t}, RV_{M,t} \text{ 中出现}$$

$$RV_{t-5} \sim RV_{t-10} \text{ 只在 } RV_{M,t} \text{ 中}$$

所以还是有 20 个系数

$$RV_t = \phi_D + \frac{1}{5}\phi_W + \frac{1}{21}\phi_M$$

$$RV_{t-1} \sim RV_{t-4} = \frac{1}{5}\phi_W + \frac{1}{21}\phi_M$$

$$RV_{t-5} \sim RV_{t-10} = \frac{1}{21}\phi_M$$

虽然很多系数是一样的，但是避免了奇怪的系数

下面介绍更复杂一点的模型

$$\ln(RV_{t+1}) = \phi_0 + \phi_D \ln(RV_{D,t}) + \phi_W \ln(RV_{W,t}) + \phi_M \ln(RV_{M,t}) + \varepsilon_{t+1}$$

这样系数会精确一些，此时

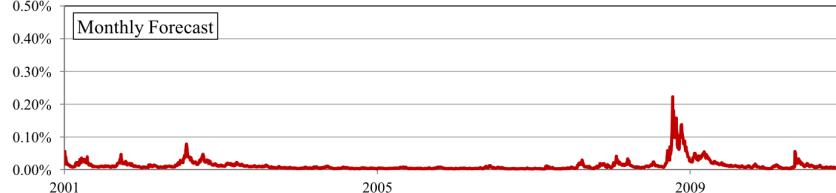
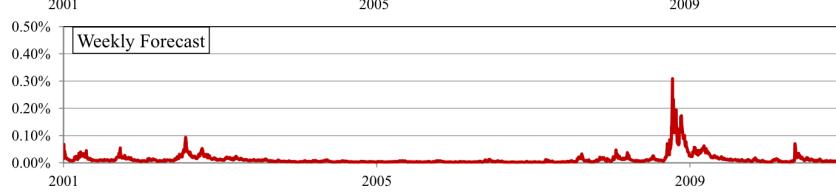
$$RV_{t+1|t}^m = (RV_t)^{\phi_D} (RV_{W,t})^{\phi_W} (RV_{M,t})^{\phi_M} e^{\phi_0 + \frac{1}{2}\phi_\varepsilon^2}$$

如果我们想预测很多天的情况：

$$\ln(RV_{t+1,t+k}) = \phi_{0,k} + \phi_{D,k} \ln(RV_{D,t}) + \phi_{W,k} \ln(RV_{W,t}) + \phi_{M,k} \ln(RV_{M,t})$$

$$\text{其中 } RV_{t+1,t+k} = \frac{1}{k} (RV_{t+1} + \dots + RV_{t+k})$$

所以也是不变的



更平滑，因为 average RV。
不会有太大的极端值

HAR 会 capture leverage effect

我们可以在右侧加 R_t + 右侧可以加 dummy 变量 Monday

$$\ln(RV_{t+1}) = \phi_0 + \phi_D \ln(RV_{D,t}) + \phi_W \ln(RV_{W,t}) + \phi_M \ln(RV_{M,t}) + \phi_R R_t + \varepsilon_{t+1}$$

$RV_{t+1|t} = E_t[\exp(\ln(RV_{t+1}^m))]$ 表示为止，所以系数没有限制条件

由 Stylized fact 4 知

$$R_{t+1} / \sqrt{RV_{t+1|t}} \stackrel{iid}{\sim} N(0, 1)$$

\Rightarrow 可以算 Monte Carlo VaR

下面结合 GARCH 和 RV

$$R_{t+1} = \delta_{t+1} Z_{t+1} \text{, where}$$

$$\delta_{t+1}^2 = w + \alpha R_t^2 + \beta \delta_t^2$$

↓

$$\delta_{t+1}^2 = w + \alpha R_t^2 + \beta \delta_t^2 + \gamma RV_t^m \text{, 用 MLE 即可估计系数}$$

It's not useful for forecasting multiple days horizon

想估计 δ_{t+2} , 需要 RV_{t+1}

如何解决呢：

$$R_{t+1} = \delta_{t+1} Z_{t+1}.$$

$$\delta_{t+1}^2 = w + \alpha R_t^2 + \beta \delta_t^2 + \gamma RV_t^m$$

$$RV_t^m = w_{RV} + p_{RV} \delta_t^2 + \varepsilon_t$$

Intraday Volatility Modelling Introduction II

我们具体如何计算 RV%?

之前讲的有什么问题? → 有 Bid-Ask Spread, 不是我們想要的 fundamental value

- 現实世界中是用 Bid- Ask Price 而不是 fundamental value 来交易的

在前一节课中, 我们认为资产^I一天都有24小时在交易.^{II} bid-ask spread 可以忽略不计

$$RV_{t+1}^m = \sum_{j=1}^m R_{t+j/m}^2 = \sum_{j=1}^m (\ln(S_{t+j/m}) - \ln(S_{t+(j-1)/m}))^2$$

$$\ln S_{t+j/m}^{Fund} = \ln S_{t+(j-1)/m}^{Fund} + e_{t+j/m}, \text{ with } e_{t+j/m} \stackrel{i.i.d.}{\sim} N(0, \sigma_e^2)$$

S^{Fund} 是 fundamental asset price

但是我们观测到的是 bid & ask price

We observe (我们进行一些假设, 搞-丁假设模型出来) → illustrate the question

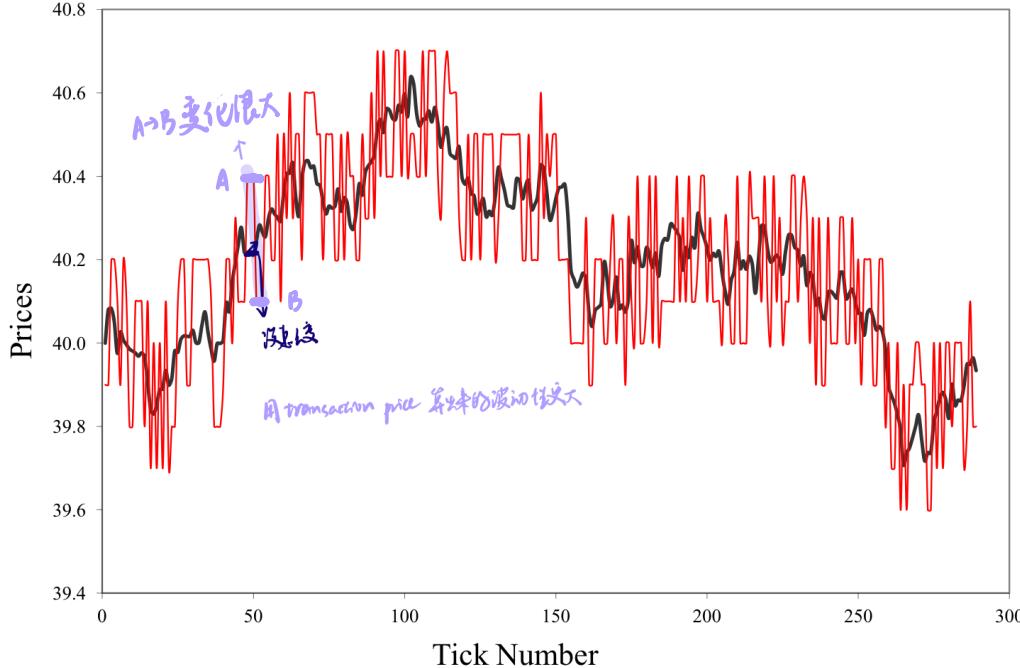
$$S_{t+j/m}^{Obs} = B_{t+j/m} I_{t+j/m} + A_{t+j/m} (1 - I_{t+j/m})$$

trade occurs randomly at bid/ask price

where $B_{t+j/m}$ is the bid price, which is the fundamental price rounded down to the nearest \$1/10 其实有点大, 但这是我们的一个假设, 便于我们有些什么

- $A_{t+j/m}$ is the ask price, which is the fundamental price rounded up to the nearest \$1/10. trade occurs at ask p
- $I_{t+j/m}$ is an i.i.d. random variable, which takes the values 1 and 0 each with probability 1/2. ↑
↓
trade occurs at bid price
- $I_{t+j/m}$ is thus an indicator variable of whether the observed price is a bid or an ask price

那么结果是什么呢？



Returns computed using trading price 会变化很大 (more volatile)
用高频数据会导致 $RV \uparrow$

我们为什么想要的是 fundamental value 呢？

→ 这样我们可以在不同时间尺度上应用不同的数据
要求数据 uncorrelated. 但 bid-ask spread is correlated

如何解决这个问题： The Sparse RV estimator

下面介绍 Sparse RV estimator

核心原理是避开老大难问题：

一 股决定不用这么高频的数据了，又留一个 bid-ask spread

一 可是皇上，那我们这些数据不是白收集了？

$$RV_{t+1}^s = \sum_{j=1}^{m/s} R_{t+j s/m}^2$$

一 但这问题是有办法解决的！

一 fundamental value 的变化远大于 bid-ask 的误差

那么 s 如何选择呢？

对于 liquid stock (AAPL, GOOGL) : $s \rightarrow 1$ illiquid: $s = 15 / 20 / 30$

有根据的计算 s 的方式

Volatility signature plots provide a convenient graphical tool for choosing s :

First compute RV_{t+1}^s for values of s going from 1 to 120 minutes.

Second, scatter plot the average RV across days on the vertical axis against s on the horizontal axis.

Third, look for the smallest s such that the average RV does not change much for values of s larger than this number

* 对于非常 illiquid 的 stock : 用 spread 会低估 return
不管如何，只要 RV 等于一个值，选最大的 s 就行了

一 下面我们来讲这丁办法：



① 我们取什么值就可以了： 建立 RV estimator

在 $0 \sim 15$ 有 157 RV

$$\text{RV}_{t+1}^{\text{Av}} = \frac{1}{s} \sum_{i=1}^s \text{RV}_{t+1}^{s,i}$$

② 第二丁办法是把 error 体现在模型中：

$$\ln(S_{t+j/m}^{\text{Obs}}) = \ln(S_{t+j/m}^{\text{Fund}}) + u_{t+j/m}, \text{ with } u_{t+j/m} \stackrel{i.i.d.}{\sim} N(0, \sigma_u^2)$$

这里我们为了方便作出的假设

那么 log return 会是 fundamental returns plus an MA(1) error

$$\begin{aligned} R_{t+j/m}^{\text{Obs}} &= \ln(S_{t+j/m}^{\text{Obs}}) - \ln(S_{t+(j-1)/m}^{\text{Obs}}) \\ &= \ln(S_{t+j/m}^{\text{Fund}}) + u_{t+j/m} - (\ln(S_{t+(j-1)/m}^{\text{Fund}}) + u_{t+(j-1)/m}) \\ &= R_{t+j/m}^{\text{Fund}} + \underbrace{u_{t+j/m} - u_{t+(j-1)/m}}_{\text{MA}(1) \text{ error process}} \end{aligned}$$

$$RV_{t+1}^m = \sum_{j=1}^m \left(R_{t+j/m}^{\text{Obs}} \right)^2 = \sum_{j=1}^m \left(R_{t+j/m}^{\text{Fund}} + u_{t+j/m} - u_{t+(j-1)/m} \right)^2$$

那么误差是什么呢？把括号打开，bias 是 $\text{var}(u_{t+j/m}^2) + \text{var}(u_{t+j-1/m})$

If we carry out the multiplication, we will get terms $(R_{t+j/m}^{\text{Fund}})^2$,

$u_{t+j/m}^2, u_{t+(j-1)/m}^2, 2R_{t+j/m}^{\text{Fund}}u_{t+j/m}$, and $2R_{t+j/m}^{\text{Fund}}u_{t+(j-1)/m} = 2u_{t+j/m} \cdot u_{t+(j-1)/m}$

The expected values of these terms are

$\text{var}(R_{t+j/m}^{\text{Fund}}), \text{var}(u_{t+j/m}^2), \text{var}(u_{t+(j-1)/m}^2), 0$, and 0

那怎么解决误差呢？

$$R_{t+j/m}^{\text{Obs}} R_{t+(j-1)/m}^{\text{Obs}} = (R_{t+j/m}^{\text{Fund}} + u_{t+j/m} - u_{t+(j-1)/m}) \\ \times (R_{t+(j-1)/m}^{\text{Fund}} + u_{t+(j-1)/m} - u_{t+(j-2)/m})$$

We can estimate the other variance $\text{var}(u_{t+j/m})$ by considering the cross-products of $R_{t+(j+1)/m}^{\text{Obs}} R_{t+j/m}^{\text{Obs}}$

Combining the ideas from the previous two slides, we can correct for the bias in the RV estimator as follows:

$$RV_{t+1}^{\text{AR}(1)} = \sum_{j=1}^m (R_{t+j/m}^{\text{Obs}})^2 + \sum_{j=2}^m R_{t+j/m}^{\text{Obs}} R_{t+(j-1)/m}^{\text{Obs}} + \sum_{j=1}^{m-1} R_{t+j/m}^{\text{Obs}} R_{t+(j+1)/m}^{\text{Obs}}$$

但是这方法有些过于简单了，它依赖于我们上面的模型是正确的，并不通用

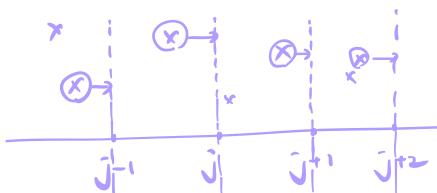
接下来我们试图解决 data issue:

- Market Closure — 市场并不 24 小时开着的
- prices & quotes 分布是随机的，但我们希望它分布在 near, evenly spaced price grid 上

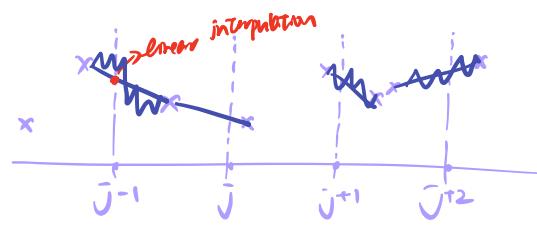
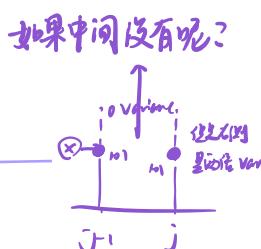
有两种方法可以解决，但只有一种方法是正确的

1. 直接往右挪 ✓

Actual trade 是发的很随机的

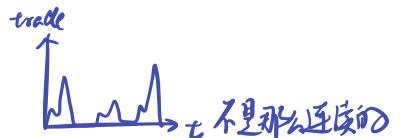


2. 线性插值法 X → price 是不可预测的！



一些其他的问题：

我们一直讨论的是 $S=1$ 的情形 \rightarrow 对于 S&P 500 很适用 但是很多股票的情况这样



这时候我们需要谨慎地处理区间的长度

另一个问题：市场不是 24 小时开着的，那 RV 怎么算呢？(解决)

其实我们只能算出市场开着的时候的 $RV - RV^{open}$

但市场关着的时候 price 仍然是变化的

直观上我们可能认为 U.S. $\frac{24}{6.5} RV^{open}$ CHN : $\frac{24}{4} RV^{open}$
 ↑
 U.S. open for 6.5 hrs

但我们假设了 open 和 close 时 volatility 是一样的

而 open 时 volatility 更高 — 我们需要 scale，但不是 $\frac{24}{6.5}$

下面找 ratio :

该一：24-hour variance \rightarrow usual close-to-close squared return (这是 unbiased 的)
 $RV^{open} \rightarrow$ 我们是 ignorant 的

$$RV_{t+1}^{24H} = \left(\frac{\sum_{t=1}^T R_t^2}{\sum_{t=1}^T RV_t^{open}} \right) RV_{t+1}^{open}$$

Or, we can add to RV_{t+1}^{open} the squared return constructed from the close on day t to the open on day $t+1$:

$$RV_{t+1}^{24H} = \ln(S_{t+1}^{Open}/S_t^{Close})^2 + RV_{t+1}^{open}$$

A third approach is to find optimal weights for the two terms.

This can be done by minimizing the variance of the RV_{t+1}^{24H} estimator subject to having a bias of zero

Specifically, to solve

$$\min_{a,b} (a \ln(S_{t+j}^{Open}/S_{t+j-1}^{Close}) + b RV_{t+j}^{open})$$

When computing optimal weights a and b , a much larger weight is found for RV_{t+1}^{open} than for $\ln(S_{t+1}^{Open}/S_t^{Close})$

This occurs because $\ln(S_{t+1}^{Open}/S_t^{Close})$ is a noisy estimate of the overnight return variance