

## 苹果股票日收益率分析

一、下载一只股票 15-30 年日收益率数据 $r_t$ （附件中的沪深 300 或者 R 中的 quantmod 包），以收益率的负数 $X_t = -r_t$ 为研究对象，表示损失，首先拟合合适的 ARMA-GARCH 模型，得到新息序列 $Z_t$

利用 R 中的 quantmod 包，从雅虎获得了苹果公司（“AAPL”）从 2003 年 1 月 2 日到 2019 年 4 月 5 日的股票数据，利用调整后的价格 AAPL.Adjusted 可以得到苹果公司的日收益率数据 $r_t$ ，进而对日收益率数据取负后即可得到苹果公司的损失数据 $X_t$ 。为拟合合适的 ARMA-GARCH 模型，首先对损失序列 $\{X_t\}$ 进行分析，得到 $\{X_t\}$ 的时序图以及 QQ 图，并对 $\{X_t\}$ 进行 Jarque-Bera 正态性检验，结果如下：

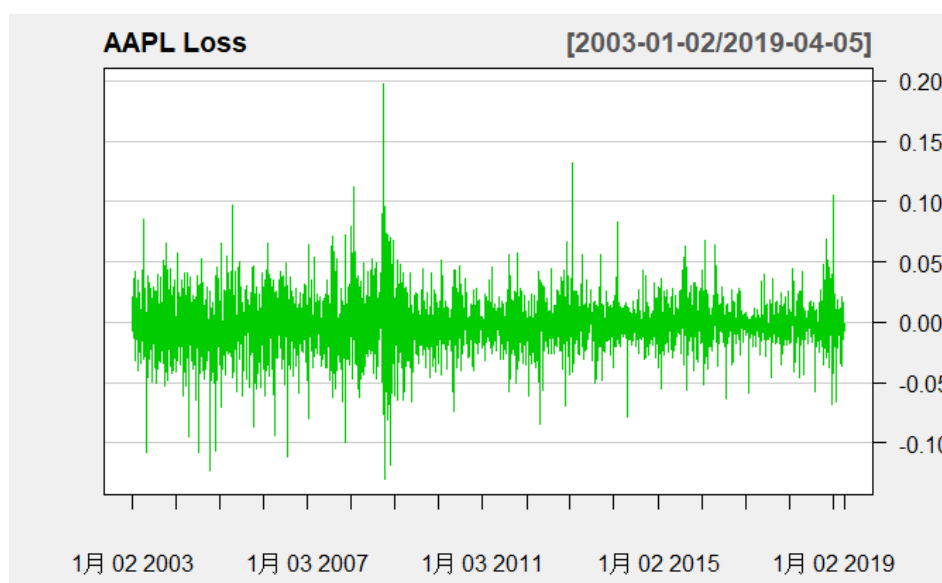


图 1：损失序列的时序图

由时序图可以看出苹果公司的损失序列 $\{X_t\}$ 在 0 附近波动，且存在一定的波动率集聚效应。可以发现，在 2009 年 1 月 3 日附近的时间里由较大的波动性，同时可以发现，2003 年至 2010 年的波动率比 2010 年之后的要更大一些。

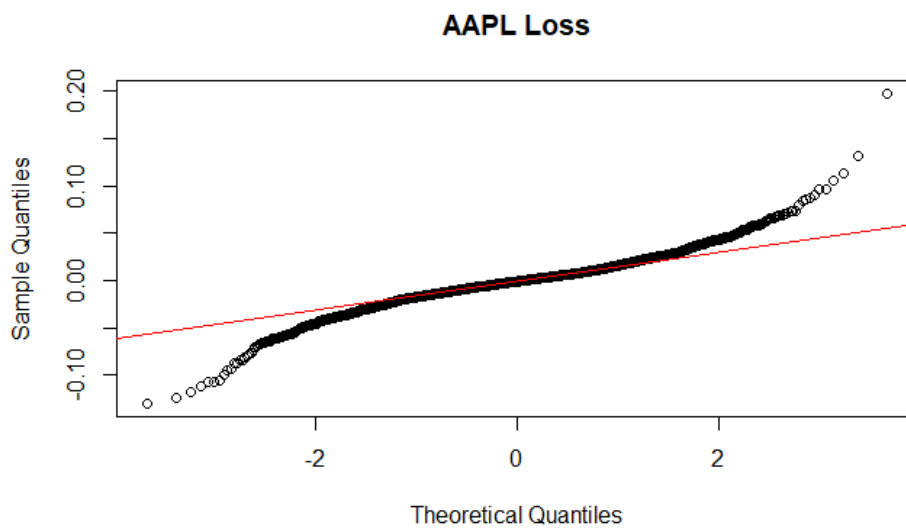


图 2：损失序列的 Q-Q 图

由 Q-Q 图可以看出，损失序列 $\{X_t\}$ 的分位数图与正态分布的分位数图并不重合，因此由 Q-Q 图可认为损失序列并不服从正态分布。

表 1：损失序列的相关统计量

Mean	Median	Stdev	Skewness	Kurtosis	J-B	Prob.
-0.001344	-0.001146	0.021097	0.077124	5.486251	5186.3703	<2.2e-16

表 1 列出了损失序列的均值、中位数、标准差、偏度、峰度、J-B 统计量及其 p 值。根据上述统计量，可以发现，数据均值左侧的离散程度比右侧强，且峰值高于正态分布。同时由 J-B 统计量及其 p 值，认为其不服从正态分布，与 Q-Q 图结果一致。

其次，对 $\{X_t\}$ 进行平稳性检验，以选择合适的 ARMA 模型。对 $\{X_t\}$ 画出 ACF 图和 PACF 图如下：

感谢中国人民大学统计学院徐慧超同学提供的数据、代码和报告

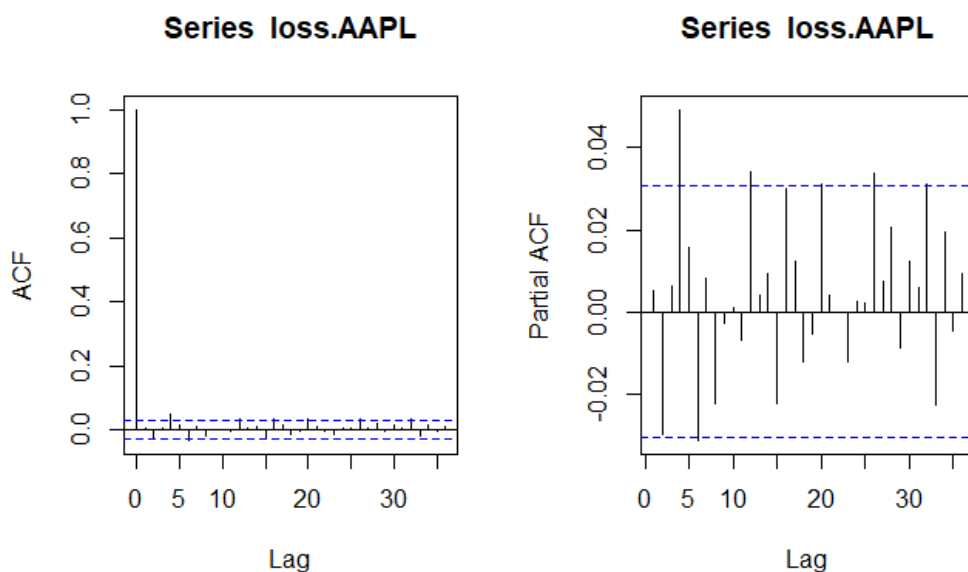


图 3：损失序列的 ACF 图及 PACF 图

由图 3 左侧的 ACF 图可以看出 $\{X_t\}$ 只在第 0 阶，自相关系数为 1，而之后各阶的自相关系数均很小，由图认为可看作是在两倍标准差以内，故认为损失序列 $\{X_t\}$ 是白噪声序列。同时观察图 3 右侧的 PACF 图，损失序列在第 3 阶的偏自相关系数明显超出两倍标准差，之后的第 5 阶、第 12 阶、第 25 阶均小幅度地超过了两倍标准差。因此利用 ACF 图和 PACF 图来给 ARMA 模型定阶在此处不是很合适。

利用 R 语言 TSA 包中的 `eacf` 函数可以直接确定  $ARMA(p, q)$  中的  $p$  和  $q$  值，在 `eacf` 地结果矩阵中，出现值为 0 的三角形的左上角的坐标即为可能  $p$ 、 $q$  值。对损失序列运用 `eacf` 函数，结果显示如下：

```
> eacf(loss.AAPL)
AR/MA
  0 1 2 3 4 5 6 7 8 9 10 11 12 13
0 0 0 0 x 0 x 0 0 0 0 0 x 0 0
1 x 0 0 x 0 0 0 0 0 0 0 x 0 0
2 x x 0 0 0 x 0 0 0 0 0 x 0 0
3 x x x 0 0 x 0 0 0 0 0 0 0 0
4 x x x 0 0 x 0 0 0 0 0 0 0 0
5 x x x x x x 0 0 0 0 0 0 0 0
6 x x x x 0 x 0 0 0 0 0 0 0 0
7 x x x x 0 x x 0 0 0 0 0 0 0
```

由结果，认为  $ARMA(1, 1)$ ,  $ARMA(2, 2)$ ,  $ARMA(3, 3)$  是可以考虑的模型，分别对其进行模拟，得到三个模型的 AIC 值分别为 -20127.48, -20137.51 和 -20133.7。根据 AIC 新息准则选择 AIC 值最小的模型，即为  $ARMA(2, 2)$ ，又对  $ARMA(2, 2)$  模型下的残差进行 Ljung-Box 白噪声检验：

### Box-Ljung test

```
data: resm2$residuals
x-squared = 0.00046269, df = 1, p-value = 0.9828
```

结果显示采用  $ARMA(2, 2)$  模型的残差是白噪声序列，检验结果显示支持该  $ARMA(2, 2)$  模型。

接下来，考虑序列  $\{X_t\}$  和  $\{X_t^2\}$ ，分别做作其 ACF 图，并分别对它们进行 Ljung-Box 白噪声检验：

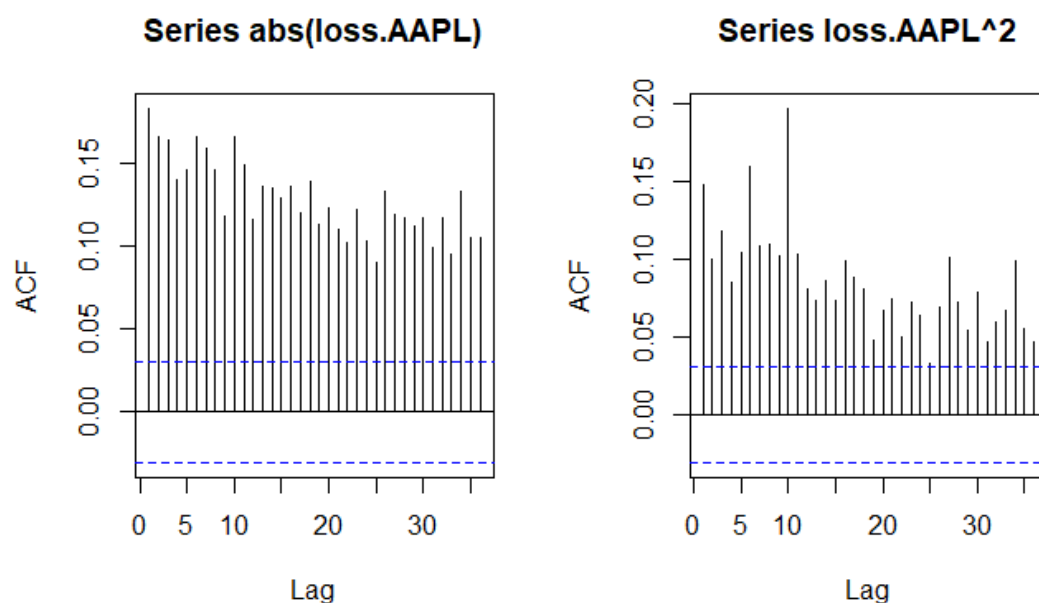


图 4：损失序列的绝对值及其平方的 ACF 图

由图 4 可以看出损失序列的绝对值及其平方的自相关系数都远远超过其两倍标准差，也就是说损失序列的绝对值序列及其平方序列都不是白噪声序列。

接下来对损失序列的绝对值序列和平方序列分别进行 Ljung-Box 检验，通过

感谢中国人民大学统计学院徐慧超同学提供的数据、代码和报告

统计量对两个序列的平稳性进行检验。

```
> Box.test(abs(loss.AAPL), type = 'Ljung-Box')
```

Box-Ljung test

```
data: abs(loss.AAPL)
x-squared = 137.87, df = 1, p-value < 2.2e-16
```

```
> Box.test(loss.AAPL^2, type = 'Ljung-Box')
```

Box-Ljung test

```
data: loss.AAPL^2
x-squared = 90.108, df = 1, p-value < 2.2e-16
```

检验结果表明 $\{X_t\}$ 和 $\{X_t^2\}$ 的 p 值均很小，也就是说两个序列均存在明显的自相关性，说明 $\{X_t\}$ 平稳、不相关，但并不前后独立。考虑用 GARCH 模型对损失序列 $\{X_t\}$ 进行拟合。通常用试错法尝试较低阶的 GARCH 模型即可，如 GARCH(1, 1)，GARCH(1, 2)，GARCH(2, 1)等。我们选择最简单的 GARCH (1, 1) 模型，分别利用正态新息和 t 新息去进行拟合。

下图分别是利用正态新息的 ARMA(2, 2)-GARCH(1, 1)模型拟合得到的 $|X_t|$ 的 ACF 图， $Z_t$ 的 Q-Q 图， $Z_t$ 的 ACF 图以及 $Z_t^2$ 的 ACF 图。

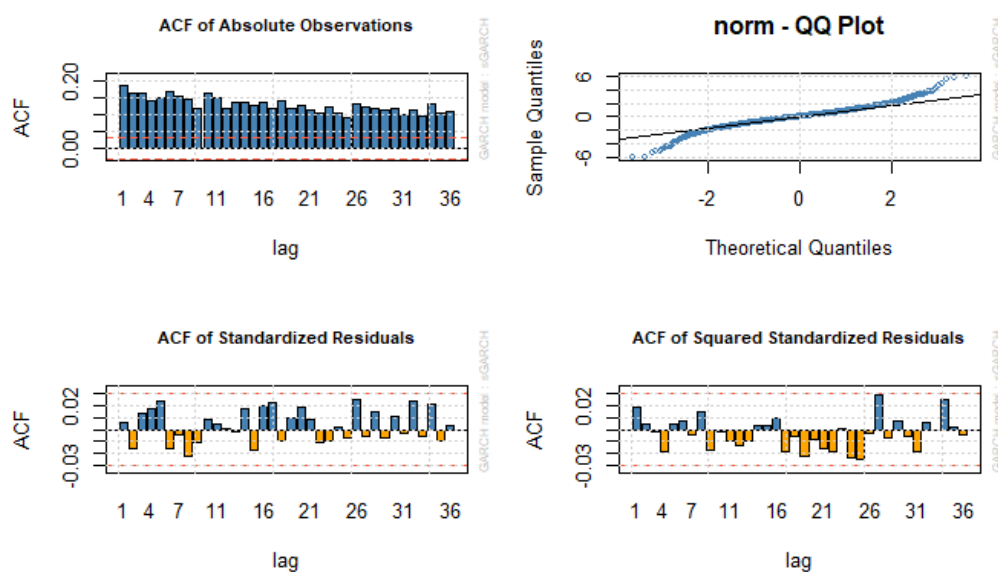


图 5：正态新息的相关图

感谢中国人民大学统计学院徐慧超同学提供的数据、代码和报告

由图 5 的左上图可以发现，利用正态新息的 ARMA(2, 2)-GARCH(1, 1) 模型拟合得到损失序列的绝对值序列的自相关系数远超过两倍标准差，即 $\{|X_t|\}$ 不是白噪声序列；由图 5 的右上图可发现拟合后得到的残差序列 $\{Z_t\}$ 的分位数曲线与正态分布的分位数曲线之间存在一定差异，故认为残差序列 $\{Z_t\}$ 不服从正态分布；由图 5 的左下图和右下图可发现残差序列及其平方序列的自相关系数都在两倍标准差之内，故认为 $\{Z_t\}$ 和 $\{Z_t^2\}$ 序列均是白噪声序列。

接下来利用  $t$  新息的 ARMA(2, 2)-GARCH(1, 1) 模型进行拟合，得到的 $|X_t|$ 的 ACF 图， $Z_t$ 的 Q-Q 图， $Z_t$ 的 ACF 图以及 $Z_t^2$ 的 ACF 图如下。

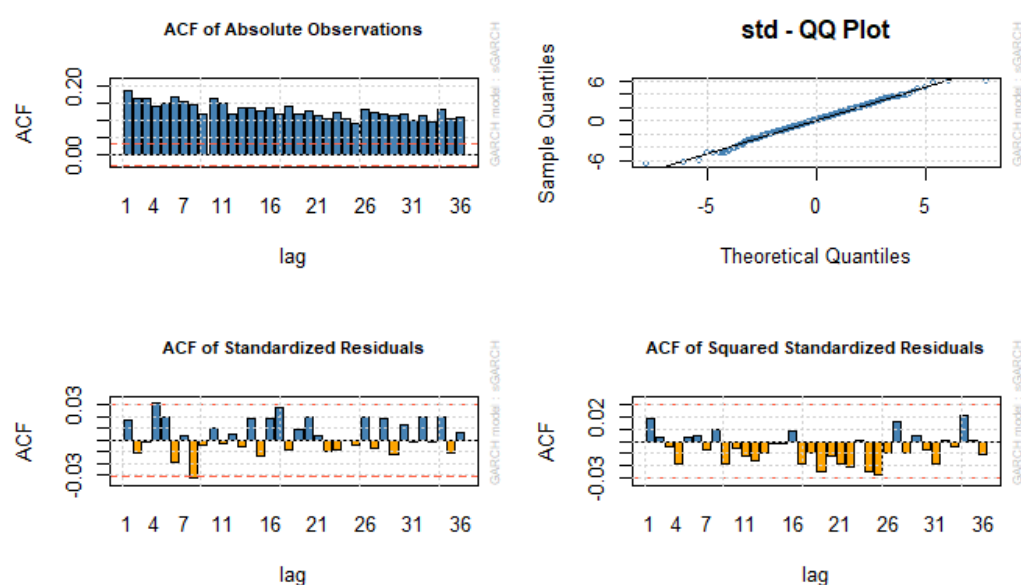


图 6:  $t$  新息的相关图

由图 6 的左上图可以发现，利用  $t$  新息的 ARMA(2, 2)-GARCH(1, 1) 模型拟合得到损失序列的绝对值序列的自相关系数远超过两倍标准差，即 $\{|X_t|\}$ 不是白噪声序列；由图 6 的右上图可发现拟合后得到的残差序列 $\{Z_t\}$ 的分位数曲线与正态分布的分位数曲线之间十分相似，故认为残差序列 $\{Z_t\}$ 服从正态分布；由图 6 的左下图和右下图可发现残差序列及其平方序列的自相关系数都可认为在两倍标准差之内，故认为 $\{Z_t\}$ 和 $\{Z_t^2\}$ 序列均是白噪声序列。

根据图 5 和图 6 之间的比较，认为采用  $t$  新息的模型进行拟合更可取。接下来继续对正态新息和  $t$  新息的残差进行 Ljung-Box 白噪声检验，同时计算两种新

感谢中国人民大学统计学院徐慧超同学提供的数据、代码和报告

息情况下的 AIC，结果如下表：

表 2：正态新息和  $t$  新息的比较

	L-B 统计量	P 值	AIC
正态新息	0.30819	0.5788	-20828.9174
$t$ 新息	0.25395	0.6143	-21260.4628

由表 2 可以看出，正态新息和  $t$  新息的 L-B 统计量值及其  $p$  值的差异不大，但  $t$  新息下的 AIC 值明显小于正态新息，故认为  $t$  新息的模型是更合理的。

综合以上情况，选择带有  $t$  新息的 ARMA(2, 2)-GARCH(1, 1) 模型。确定了模型以后，就可以得到新息序列  $\{z_t\}$ ，画出  $\{z_t\}$  的时序图如下：

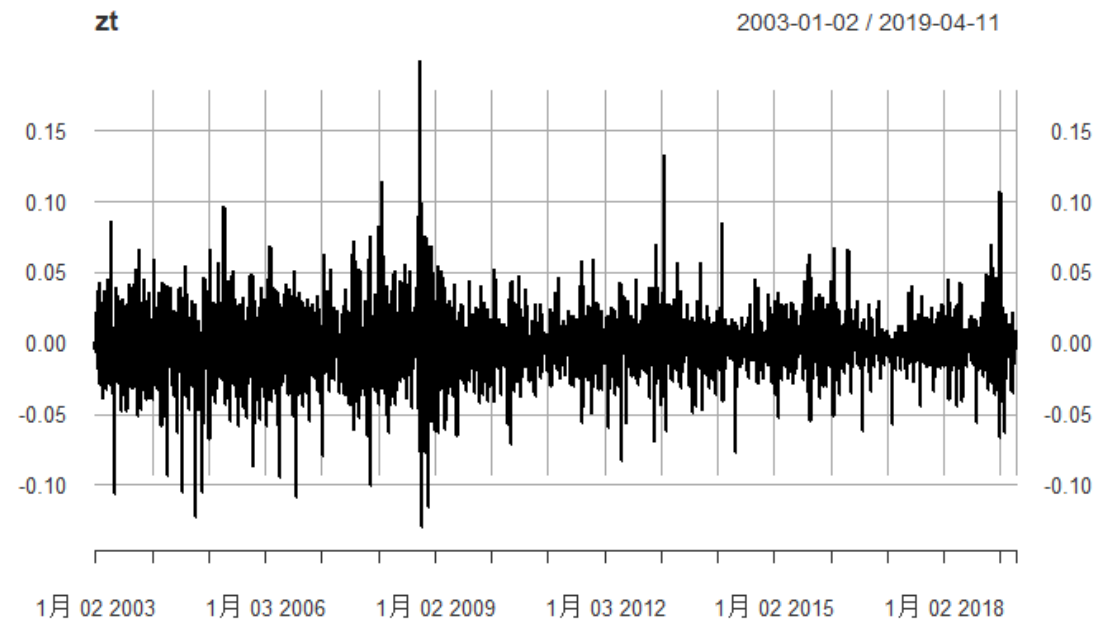


图 7：新息序列的时序图

这里列出  $\{z_t\}$  的前六项：

```
> head(zt)
```

```
          zt
2003-01-02 0.001475746
2003-01-03 -0.005257426
2003-01-06 0.001508416
2003-01-07 0.004873880
2003-01-08 0.021724788
2003-01-09 -0.007584006
```

感谢中国人民大学统计学院徐慧超同学提供的数据、代码和报告

## 2、Block maxima

### 2.1 对 $Z_t$ 分块 $n = 60, 120, 240$ ，拟合分块极值的 GEV 分布参数 $\xi, \mu, \sigma$

将残差序列以“天”为单位，根据时间进行分块。对 $Z_t$ 分块  $n=60, 120, 240$ ，也就是按 60 天、120 天和 240 天对残差序列进行分块。将分块后的序列分别记为 M. 60、M. 120 和 M. 240，对它们分别利用 R 语言 qrmtools 包中的 fit\_GEV\_MLE 进行 GEV 分布的拟合，通过可利用 fit 结果中的 `par[["shape"]]`、`par[["loc"]]` 和 `par[["scale"]]` 分别得到拟合的 GEV 分布中的参数 $\xi, \mu, \sigma$ ，其中 $\xi$ 是形状参数， $\mu$ 是位置参数， $\sigma$ 是规模参数。拟合后求得参数的结果如下：

表 3：分块后 GEV 分布的参数

	$\xi$	$\mu$	$\sigma$
M. 60	0.10609	0.04195	0.01786
M. 120	0.22174	0.05048	0.01823
M. 240	0.19569	0.05996	0.02378

### 2.2 计算 $Z_t$ 的 return level $r_{240, 10}$ 和 return period $r_{240, 0.05}$

由上题，已知 $M_{240} \sim H_{0.11025, 0.05771, 0.02807}$ ，残差序列 $\{Z_t\}$ 的 return level  $r_{240, 10}$ 即为 GEV 分布的  $(1-1/10)$  分位数，return period  $r_{240, 0.05}$ 即为 1 减去 GEV 分布在 0.05 处的函数值后的倒数，计算结果如下表：

表 4：return level 与 return period

$r_{240, 10}$	$r_{240, 0.05}$
12.72%	1.27 $\approx$ 305 天

### 2.3 基于 GEV 分布计算 $Z_t$ 和 $X_{t+1}$ 的 $\text{VaR}(\alpha = 0.95)$ 指标

感谢中国人民大学统计学院徐慧超同学提供的数据、代码和报告



基于 GEV 分布计算分块后的残差序列 $\{Z_t\}$ 的 VaR 等价于计算分块后的 GEV 分布的 $\alpha^n$ 分位点。基于 GEV 分布计算 $X_{t+1}$ 的 VaR 则等价于预测出的 $X_{t+1}$ 的均值  $\mu$  加上协方差  $\sigma$  乘以 $Z_t$ 的相应的 VaR 值。在本题中得到的  $\mu$  和  $\sigma$  分别为  $-0.000841$  和  $0.02099$ ，计算得到的结果如下表所示。

表 5：基于 GEV 分布的 VaR

	n=60		n=120		n=240	
	$Z_t$	$X_{t+1}$	$Z_t$	$X_{t+1}$	$Z_t$	$X_{t+1}$
VaR	0.02303	-0.00036	0.02322	-0.00036	0.01279	-0.00058

### 3、Threshold Exceedances

3.1 针对 $Z_t$ 实现书中 Figure 5.4：绘制 Sample mean excess 图像，选取合适的阈值水平  $u$ ，并拟合 GPD 分布参数 $\xi$ ， $\beta$

针对 $Z_t$ 画出的 Sample mean excess 图像如下图所示，选择的阈值水平是在其之后图像不再有巨大转折点的横坐标所在直线。由图 8，选取阈值水平  $u=0.035$ ，同时，对残差序列 $\{Z_t\}$ 画出 GPD shape 图，并画出其置信带如图 9。

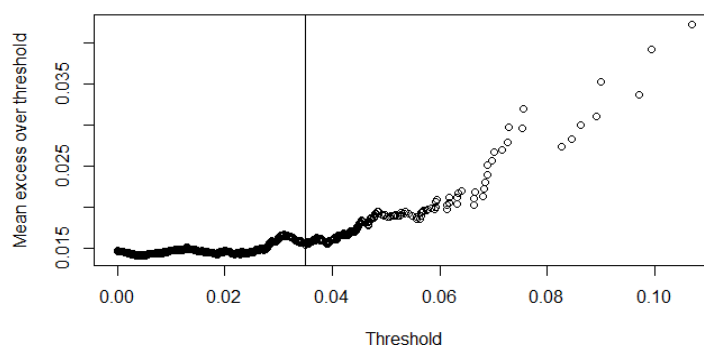


图 8：sample mean excess 图

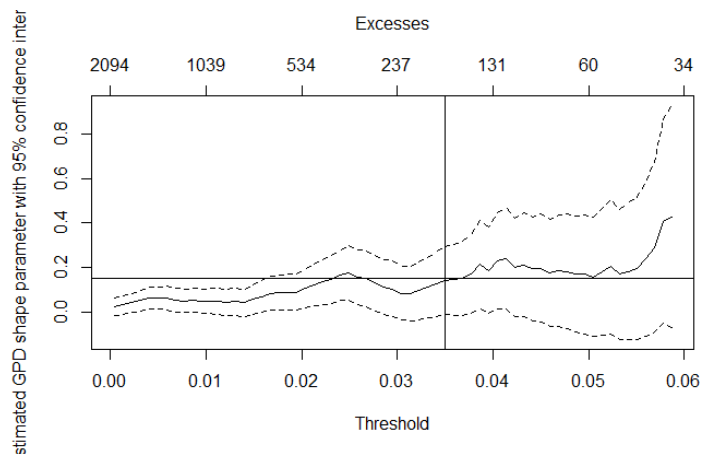


图 9: GPD shape 图

接下来利用 `fit_GPD_MLE` 函数对残差序列进行 GPD 模型的拟合，拟合得到 GPD 分布的形状参数 $\xi$ 和规模参数 $\beta$ 分别为 0.14375 和 0.01344

### 3.2 基于 GPD 分布计算 $Z_t$ 和 $X_{t+1}$ 的 VaR 和 ES 指标( $\alpha = 0.95$ )

基于 GPD 分布计算 $Z_t$ 的 VaR 和 ES 指标可以直接分别利用 `VaR_GPDtail` 函数和 `ES_GPDtail` 函数进行计算。对 $X_{t+1}$ 的 VaR 和 ES 指标的计算如 GEV 分布，用 $Z_t$ 的 VaR 值和 ES 值分别乘以预测的 $X_{t+1}$ 的协方差 `sigma` 加上它的均值 `mu` 得到。结果如下表：

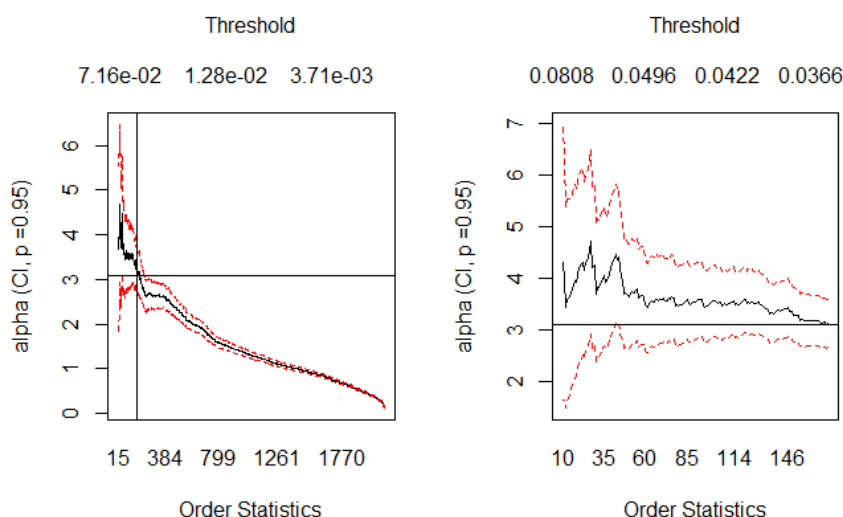
表 6: 基于 GPD 分布的 VaR 和 ES

	$Z_t$	$X_{t+1}$
VaR	0.0350	-0.000109
ES	0.0507	0.00022

## 4、Hill Method

### 4.1 针对 $Z_t$ 实现书中 Figure 5.8: 选取合适的 $k$ , 得到尾部指数 $\hat{\alpha}_{k,n}^{(H)}$

针对 $Z_t$ ，可以利用 `hillplot` 命令，画出 hill plot 如下：



选取  $k=150$ ，尾部指数  $\hat{\alpha}_{k,n}^{(H)}$  通过公式  $\hat{\alpha}_{k,n}^{(H)} = \left( \frac{1}{k} \sum_{i=1}^k \log Z_{i,n} - \log Z_{k,n} \right)^{-1}$  计算得到  $\hat{\alpha}_{k,n}^{(H)} = 3.19194$ ，其中  $Z_{i,n}$  是从大到小排序后的  $Z_t$  序列的第  $i$  个元素。

#### 4.2 基于尾部指数 $\hat{\alpha}_{k,n}^{(H)}$ 计算 $Z_t$ 和 $X_{t+1}$ 的 VaR 和 ES 指标 ( $\alpha = 0.95$ )

基于尾部指数  $\hat{\alpha}_{k,n}^{(H)}$  计算  $Z_t$  的 VaR 和 ES 指标是分别通过公式

$$\widehat{VaR}_{\alpha}(Z) = \left( \frac{n}{k} (1 - \alpha) \right)^{-\frac{1}{\hat{\alpha}_{k,n}^{(H)}}} Z_{k,n}$$

$$\widehat{ES}_{\alpha}(Z) = \frac{\hat{\alpha}_{k,n}^{(H)}}{\hat{\alpha}_{k,n}^{(H)} - 1} \widehat{VaR}_{\alpha}(Z)$$

计算结果如下表所示：

表 7：hill estimation 的 VaR 和 ES

	$Z_t$	$X_{t+1}$
VaR	0.042205	0.00004
ES	0.06146	0.00045