国信证券厦门营业部

R语言学习手册

版本：1.0

国信证券

方莲

2014年7月

国信证券厦门营业部

R 语言学习手册

版本：1.0

方　莲

2014 年 7 月

Guosen Security Company
厦 门 营 业 部

稿件版本： July 18, 2014
文版打印: XƎLaTeX © 方莲.

目录

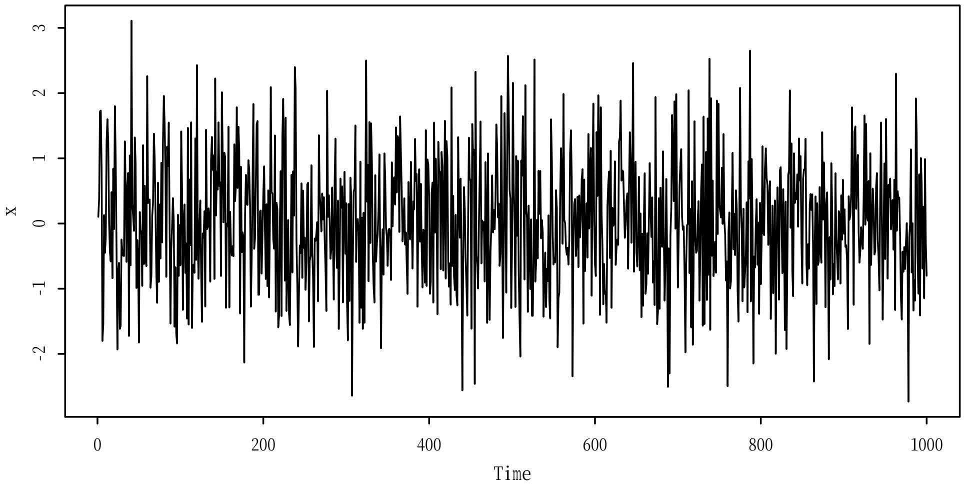
稿件版本： July 18, 2014
文版打印: XeLaTeX © 方莲.

第一章 引言

接触 R 也有一段时间了,大概是从研究生一年级的时候,有授课教师便开始鼓捣我们多多学习编程语言。可是以前也只是为了应付课程作业的需要,「幼稚」地把 R 当作一种简单的统计分析与画图之用,并没有深入编程语言的精髓部分。比如,对于 R 中对象特征的详细讨论,我并没有一个系统化的训练与学习过程。这不得不说是一种巨大的遗憾,因为现在 R 统计语言正发展的如火如荼,大有一统整个数据分析界的势头。而在这样的形势下,如果想要真正掌握这门语言的精髓与要领,也就不得不继续学习,希望通过对底层编程思想的透彻理解,为今后的软件使用、函数编程与数据分析提供扎实的基础。

```r
x <- rnorm(1000)

plot.ts(x)
```



1.1 R 是什么

1.2 R 能做什么

1.3 为什么使用 R

1.4 怎么安装

1.5 怎么用 R

第二章 统计与金融

现代金融的发展主要是借力统计技术的提高。从对风险的概率特征的描述开始，整个金融理论体系都是使用概率理论与统计方法来阐述的。Markowitz 的资产组合理论成功运用方差(variance)来刻画一项资产在时间范围内不确定性的量化指标，到最成功的运用统计随机思想的 Black-Scholes 期权定价公式，在整个金融领域都随处可见统计思想。因此，在深入理解整个金融市场之前，我们需要对描述其动态特征的概率统计理论有一定的掌握。对于金融市场建模，目前大概需要用到的数学理论有：

这里我特指滥觞于 Markowitz (1952)、大量使用数学尤其是概率统计理论与随机过程方法的金融理论与实证金融研究

个人强烈推荐 Taleb 写的 Fooled by Randomness。作者本人从科班出生，主要研究概率与统计思想。这是一个「神奇的书」，文学与数学兼谈，是一本从概率视角与金融交易员的角度看这个市场(与世界)的畅销书。

- 测度论框架下构建概率理论：与经典理论将概率解释为随机数出现的频率的极限值不同，现代概率论更多的从测度论的视角把概率当作一种可以衡量的 *Lebegue* 积分，即在样本空间上的一段「测量」。
- 随机过程论，尤其是布朗运动与 Ito 随机过程。
- 统计理论，需要掌握诸如回归分析、参数估计、统计推断、拟合分析等基本的分析技术，理解 Mento Carlo 方法做模拟的思路。

2.1 风险测量

风险指的是由于未来时间的不确定性导致金融变量在样本空间内的变动情况，即无法事先预知该变量出现的确切数值，而只知道其所属的整个样本概率空间。概率论使用实值函数映射来表示变量的不确定，即 $\{X \big| X(\omega) \in \mathcal{R}, \omega \in \Omega\}$。一项资产投资的风险可以理解为投资者对该资产未来预期收益的不确定性，即这笔投资可能远远超过初始投资额度，带来正值的投资回报；也可能由于不可预期的风险导致该资产回报小于原始投资成本，给投资者造成损失。因此，利用概率理论的「随机变量」这个重要的概念能够把投资者对未来不确定的担忧进行量化处理。这种对尚未发生事件的量化，可以是基于金融市场的历史数据而推断得到的客观证据，也可能是投资者自身的主管评价。通常的，前者被当作「经典」的概率理论，而后者则被贴上了「贝叶斯统计」，尽管这个标签其实也是「名不副、其不实」。

Bayesian Statistics

Risk Measure

■ [P1.] Risk means uncertainty in future returns from an investment, in particular, that the investment could earn less than the expted return and even results in a loss, that is, a negative return. Risk is often measured by the standart deviation of the return, which we also call the *volatility*. Recently there has been a trend toward measuring risk by value-at-risk (VaR) and expected shortfall

(ES). These focus on large losses and are more direct indications of financial risk than the standard deviatin of the return.

2.2 收益率

另外，我们也可以 Tsay 的 Financial Times Series Analysis 中发现类似的讨论。对于资产回报率的研究，能够逆向推断出资产价值。因此，我们看到绝大部分的文献都在讨论收益率的随机特征。

对于为何我们在处理金融实证研究时主要是针对资产 (股票、债券、衍生品等) 回报率, 可以参考经典的 Compbell,Lo and Makinley 的"The Econometrics of Financial Markets" 已经解释的十分到位。这是因为:

- 收益率衡量的是在单位投资数量上所获得的投资回报, 是一个相对的概念, 与投资规模大小没有关系。因而这有利于比较不同投资规模、投资期限的资产给予投资者的风险补偿。
- 收益率同股票价格相比在统计特征上更加方便。收益率反映了股票价格微小的变动情况, 比较服从随机过程对数据平稳性的要求。

2.2.1 简单净收益率

simple gross return: 简单毛收益率

一个单位时间内的简单毛收益率被定义为资产相对于初始投入所获得的增减的程度

$$1 + R_t = \frac{P_t}{P_{t-1}}. \tag{2.2.1}$$

simple net return: 简单净收益率

则简单净收益率 R_t 为

$$R_t = \frac{P_t}{P_{t-1}} - 1 = \frac{P_t - P_{t-1}}{P_{t-1}}. \tag{2.2.2}$$

由于期末价格 $P_t \geq 0$, 因此简单收益率的下限是 $R_t \geq -1$。

2.2.2 k 期收益率

同样的, 我们可以定义一个在 k 期内的资产收益率,

$$1 + R_t[k] = \frac{P_t}{P_{t-k}} = \frac{P_t}{P_{t-1}} \cdot \frac{P_{t-1}}{P_{t-2}} \cdots \frac{P_{t-(k-1)}}{P_{t-k}} = \prod_{j=0}^{k-1} (1 + R_{t-j}) \tag{2.2.3}$$

2.2.3 单期对数收益率

continuously compounded return

对数收益率 也称作「连续复合收益率」

$$r_t = \log(1 + R_t) = \log(\frac{P_t}{P_{t-1}}) = p_t - p_{t-1}. \tag{2.2.4}$$

式中 $p = \log(P)$。

2.2.4 k 期对数收益率

使用对数收益率的一个好处在于我们可以将一个 k 期的对数收益率化解为 k 个单期收益率的叠加，即

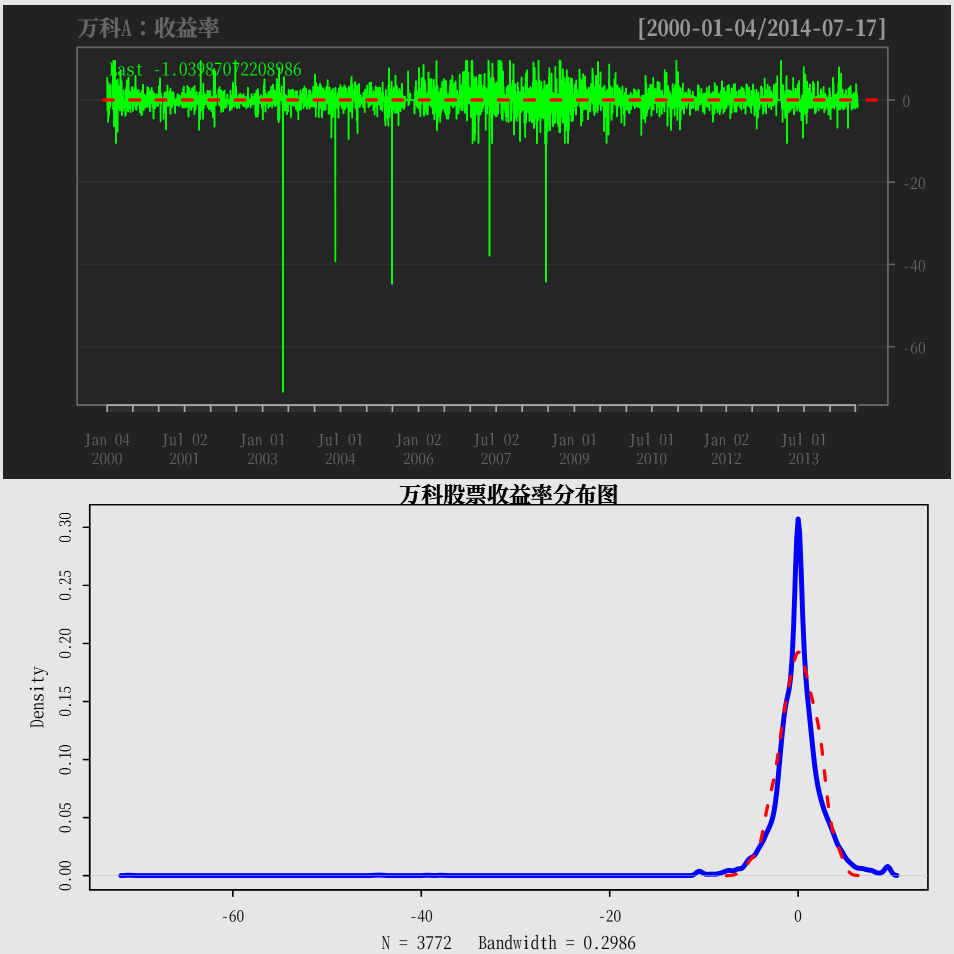
$$r_t[k] = \log(1 + R_t[k]) = \log\left(\prod_{j=0}^{k-1}(1 + R_{t-j})\right) = \sum_{j=0}^{k-1} r_{t-j} \tag{2.2.5}$$

Figure 2.2.1: 万科 A 股

```
rtn <- function(stock.code, from = ”2000-01-01”) {
    y <- getSymbols(stock.code, from = from, auto.assign = FALSE)
    rtn <- diff(log(Cl(y)), 1) * 100
    rtn <- rtn[-1]
}

chartSeries(rtn(”000002.sz”), name = ”万科A: 收益率”)
abline(h = 0, col = ”red”, lwd = 2, lty = ”dashed”)

plot(density(rtn(”000002.sz”)), main = ”万科股票收益率分布图”, col = ”blue”,
    lwd = 3)
lines(density(rnorm(1000, 0, 2)), col = ”red”, lwd = 2, lty = ”dashed”)
```



■**说明** 万科 A 股数据来自 quantmod() 函数通过对 Yahoo Finance 股票市场的网页读取。

2.3 布朗运动

Brownian Motion

布朗运动,也称做「维纳过程」,最初指的是由生物学家布朗先生观测到花粉粒子在水中相互碰撞所形成的不规则的运动轨迹,后来经由 Bachelier 将其成功运用到股票市场,反映了股票价格变化的高度随机性特征。布朗运动是一个正态分布的独立增量连续随机过程。一个随机过程 $\{W(t)\}$ 是一个布朗运动,如果它具有如下特性:

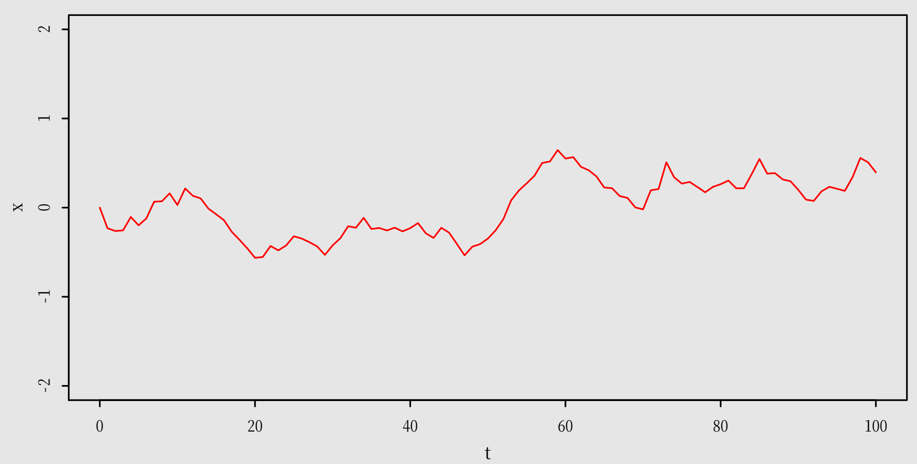
1. 初始状态 $W(0) = 0$。
2. 连续但不可微,即

$$\Delta W(t) = W(t + \Delta t) - W(t) = \varepsilon\sqrt{\Delta t}.$$

   其中,随机变量 $\varepsilon \sim \mathcal{N}(0, 1)$.
3. 独立增量过程:对于 $0 \leq t_1 \leq t_2 \leq \cdots \leq t_n$,增量 $W(t_2) - W(t_1)$、$W(t_3) - W(t_2)$、$\cdots$、$W(t_n) - W(t_{n-1})$ 是相互独立的。

Figure 2.3.2: 几何布朗运动模拟

```
t <- 0:100  # time
sig2 <- 0.01
## first, simulate a set of random deviates
x <- rnorm(n = length(t) - 1, sd = sqrt(sig2))
## now compute their cumulative sum
x <- c(0, cumsum(x))
plot(t, x, type = "l", ylim = c(-2, 2), col = "red")
```



2.3.1 随机游走模型

对于一个服从随机鞅过程的变量,无法预先确切知道其可能在下一期出现的实现值,我们有

$$\mathbb{E}[Y_{t+1}|Y_t] = Y_t. \tag{2.3.1}$$

一个随机游走模型是一个随机鞅过程,服从

Randow Walk Model:随机游走模型

$$Y_{t+1} = Y_t + \varepsilon_{t+1}. \tag{2.3.2}$$

随机游走最初指的是醉汉在深夜步行在街道所经历的轨迹,一般而言,这种步行路径往往飘忽不定,有可能向左,也有可能向右,无法准确判断其下一步可能迈向哪个具体的位置。在股票市场中,我们无法准确的预测下一期的股票价格会如何波动,因此,常常使用随机游走过程来比喻股票的不可预知特征。

2.3.2 几何布朗运动

这里,我们使用具体的 S_t 来表示股票价格,并且假定期初始状态价格 S_0 是给定的,是在建模时已经实现的随机变量值。那么,在持有期内的股票累计收益率,$r_t[t]$,可以表示为

$$r_t[t] = \log\left(\frac{S_t}{S_0}\right) = \sum_{j=0}^{t-1} r_{t-j}. \tag{2.3.3}$$

则可以推出

$$S_t = S_0 \cdot \exp r_t[t] = S_0 \cdot \exp \sum_{j=0}^{t-1} r_{t-j} \tag{2.3.4}$$

我们知道,正态分布具有线性叠加的特征,即

$$\begin{aligned} X &= \sum_{i=1}^{k} X_i \sim \mathcal{N}(\sum_{i=1}^{k} \mu_i, \sum_{i=1}^{k} \sigma_i^2), \\ X_i &\sim \mathcal{N}(\mu_i, \sigma_i^2) \end{aligned} \tag{2.3.5}$$

因此,如果我们假设股票的对数收益率服从正态分布,$r_t \sim \mathcal{N}(\mu, \sigma^2)$,则有

$$\log\left(\frac{S_t}{S_0}\right) = r_t[t] = \sum_{j=0}^{t-1} r_{t-j} \sim \mathcal{N}(t\mu, t\sigma^2). \tag{2.3.6}$$

我们称 S_t 服从一个对数正态分布,又名「几何布朗运动」,其参数集为 $\boldsymbol{\theta} = \{\mu, \sigma\}$。

2.4 随机微分方程

微分方程描述了关于自变量函数变化轨迹的动态特征，如物理学中的加速度与能量的关系。由于早期的微分方程主要是用于解运动方程的问题，是处于牛顿力学里面通常假定一个完美、没有摩擦的理想世界。这导致其无法捕捉到客观世界的随机性特征。顾名思义，随机微分方程就是在微分方程中加入随机项，使得其能够更好解释随机项如何影响因变量的动态变化。在金融研究中，我们通常假设股票价格的变化服从以下运动方程（随机微分方程，SDE）

$$dS_t = \mu S_t dt + \sigma S_t dW_t,$$

其中 dW_t 是一个（连续的）*Brownian motion*，$W_t \sim \mathcal{N}(0,t)$。等价的，我们可以得到股票价格变化率：

$$d\ln S_t = \frac{dS_t}{S_t} = \mu dt + \sigma dW_t. \tag{2.4.1}$$

下面我们来推导股票价格 S_t 的动态方程，这主要是使用著名的 Ito 公式。过于理论的东西我在这里不推导了，主要是利用一个简单的例子来说明如何使用 Ito 公式解 SDE 类问题。

2.4.1 Ito-Doeblin 公式

首先，令 $Y_t = \ln S_t$，对其分别做简单的偏微分求导得到

$$\frac{\partial Y}{\partial t} = 0$$
$$\frac{\partial Y}{\partial S} = \frac{1}{S}$$
$$\frac{\partial^2 Y}{\partial S^2} = -\frac{1}{S^2}$$

早期的文献叫做 Ito 公式，是以日本著名的数学家伊藤清在开创随机过程、尤其是随机积分方面的伟大成就。Ito 据说是在华尔街最为响亮的日本名字了，在很长一段时间内，Ito 一直受到追捧。直到近几年人们在巴黎艺术学院清理文件时发现一封近乎于 60 年前寄给数学家 Poincare 的通讯信封，里面详细记录了另一位伟大而爱国的年轻数学家 Wolfgang Doeblin 独立创造的随机积分理论。这个足足比 Ito 公式早了将近 30 年。后来有一部电影纪录片讲述了年轻数学家英年早逝的悲壮故事，"Mathematician Rediscovered"。为了共同纪念两位数学家的研究，国际数学会将原先的 Ito 公式改名为 Ito-Doeblin 公式。

则根据 Ito-Doeblin 公式，我们可以推出以下方程

$$\begin{aligned}
d\ln S_t = dY_t &= \frac{\partial Y}{\partial t}dt + \frac{\partial Y}{\partial S}dS_t + \frac{1}{2}\frac{\partial^2 Y}{\partial S^2}dS_t dS_t \\
&= 0\cdot dt + \frac{1}{S_t}dS_t - \frac{1}{2}\cdot\frac{1}{S_t^2}dS_t dS_t \\
&= \frac{1}{S_t}\cdot S_t\cdot(\mu dt + \sigma dW_t) - \frac{1}{2}\cdot\frac{1}{S_t^2}\cdot\sigma^2 S_t^2 dt \\
&= (\mu - \frac{1}{2}\sigma^2)dt + \sigma dW_t.
\end{aligned}$$

在区间 $[0,t]$ 内对上式两边求积分，得到如下式子

$$\int_0^t d\ln S_u = \int_0^t dY_u = \int_0^t (\mu - \frac{1}{2}\sigma^2)du + \int_0^t \sigma dW_u$$
$$\Rightarrow \ln S_t - \ln S_0 = Y_t - Y_0 = (\mu - \frac{1}{2}\sigma^2)t + \int_0^t \sigma dW_u.$$

我们知道，Brownian Motion 表示在一定时间段内随机游走走过的路径，并且如果我们假定在初始阶段为 $W(0)=0$，那么，上面等式的最后一项是

$$\int_0^t \sigma dW_u = \sigma(W_t - W_0) = \sigma W_t.$$

因此，可以得到

$$\begin{aligned}\ln S_t - \ln S_0 = Y_t - Y_0 = (\mu - \frac{1}{2}\sigma^2)t + \sigma W_t \\ \Rightarrow Y_t = Y_0 + (\mu - \frac{1}{2}\sigma^2)t + \sigma W_t \\ \Rightarrow S_t = S_0 \cdot \exp\{(\mu - \frac{1}{2}\sigma^2)t + \sigma W_t\}.\end{aligned}$$

2.4.2 分布特征

我们知道，对于任何一个正态分布做线性转换后依然服从正态分布。由于布朗运动 $W_t \sim \mathcal{N}(0,t)$，则

$$\begin{aligned}Y_t &= Y_0 + (\mu - \frac{1}{2}\sigma^2)t + \sigma W_t \\ \Rightarrow \mathbb{E}[Y_t|Y_0] &= \mathbb{E}[Y_0 + (\mu - \frac{1}{2}\sigma^2)t + \sigma W_t] \\ &= Y_0 + (\mu - \frac{1}{2}\sigma^2)t + \underbrace{\mathbb{E}[\sigma W_t]}_{0} = Y_0 + (\mu - \frac{1}{2}\sigma^2)t \qquad (2.4.2) \\ \Rightarrow \mathrm{Var}[Y_t|Y_0] &= \mathrm{Var}[\sigma W_t] = \sigma^2\,\mathrm{Var}(Wt) = \sigma^2 \cdot t \\ \Rightarrow Y_t &\sim \mathcal{N}(Y_0 + (\mu - \frac{1}{2}\sigma^2)t, \sigma^2 t) \qquad (2.4.3)\end{aligned}$$

那么，$\ln S_t = Y_t$ 则服从 ** 正态分布 **，则 $S_t = e^{Y_t}$ 服从 *log-normal disctribution*，即 $\ln S_t \sim \mathcal{N}(\mathbb{E}[S_t], \mathrm{Var}[S_t])$.

下面我们来推导 S_t 的分布特征。

2.4.3 Moment Generation Function

我们知道，对于对任何一个随机变量建模，往往需要假设该变量服从某一类随机过程，而这个随机过程由分布函数（distribution function）给定。可是，有些时候，我们并不一定需要知道整个分布函数的具体形式，而只是关注该随机变量的几个「统计特征」，如一阶矩、二阶矩等。下面要介绍的「矩条件生成函数」就针对这种情况提出的。随机变量的矩条

件可以在 **Moment Generation Function** (MGF) 十分方便的推导出来。比如，对于正态分布，我们只需要知道一阶矩和二阶矩条件就可以对变量做统计推断(method of moment, MM，还有更一般的 GMM)。对于一个可测空间 $(\Omega, \mathcal{F}, P)$，随机变量 $X \in \sigma(\mathcal{F})$ 的 MGF 定义为

$$M_X(\tau) = \mathbb{E}[e^{x\tau}] = \int_\Omega e^{x(\omega)\tau} dP(\omega)$$

其任一 $m-$ 阶的矩公式可以对 τ 在 $\tau = 0$ 处求 m 次导数得到

$$\begin{aligned}
\frac{\partial M_X(\tau)}{\partial \tau}|_{\tau=0} &= \frac{\partial}{\partial \tau}\mathbb{E}[e^{x\tau}]|_{\tau=0} = \mathbb{E}[x \cdot e^{x\tau}]|_{\tau=0} \\
&= \mathbb{E}[x] \\
\frac{\partial^2 M_X(\tau)}{\partial \tau^2}|_{\tau=0} &= \mathbb{E}[x^2] \\
\frac{\partial^m M_X(\tau)}{\partial \tau^m}|_{\tau=0} &= \mathbb{E}[x^m]
\end{aligned}$$

特别的，对于一个正态分布，$X \sim \mathcal{N}(\mu, \sigma^2)$，有

$$M_X(\tau) = \mathbb{E}[e^{x\tau}] = \exp\{\mu\tau + \frac{1}{2}\sigma^2\tau^2\}$$

2.4.4 期望与方差

这个特征对于求一个「对数正态分布」十分有用。由 (2.4.3)

$$Y_t \sim \mathcal{N}(Y_0 + (\mu - \frac{1}{2}\sigma^2)t, \sigma^2 t)$$

则其 MGF 为

$$M_Y(\tau) = \mathbb{E}[e^{y\tau}] = \exp\{\tilde{\mu}\tau + \frac{1}{2}\tilde{\sigma}^2\tau^2\},$$

其中，$\tilde{\mu} = Y_0 + (\mu - \frac{1}{2}\sigma^2)t$，$\tilde{\sigma}^2 = \sigma^2 t$. 因此，我们可以得到如下公式

$$\mathbb{E}[S^\tau] = \mathbb{E}[e^{y\tau}] = \exp\{\tilde{\mu}\tau + \frac{1}{2}\tilde{\sigma}^2\tau^2\}$$

即，期望可以表示为

$$\begin{aligned}
\mathbb{E}[S_t] &= \mathbb{E}[e^{y\tau}]|_{\tau=1} = \exp\{\tilde{\mu} + \frac{1}{2}\tilde{\sigma}^2\} = \exp\{Y_0 + (\mu - \frac{1}{2}\sigma^2)t + \frac{1}{2}\sigma^2 t\} = S_0 \cdot e^{\mu t} \\
\mathbb{E}[S^2] &= \mathbb{E}[e^{y\tau}]|_{\tau=2} = \exp\{2\tilde{\mu} + 2\tilde{\sigma}^2\} = \exp\{2[Y_0 + (\mu - \frac{1}{2}\sigma^2)t] + 2\sigma^2 t\} = S_0^2 \cdot \exp\{2\mu t + \sigma^2 t\}
\end{aligned}$$

因此，S_t 的方差可以表示为

$$\text{Var}[S_t] = \mathbb{E}[S^2] - (\mathbb{E}[S])^2 = S_0^2 \cdot \exp\{2\mu t + \sigma^2 t\} - S_0^2 \cdot \exp\{2\mu t\} = S_0^2 \cdot e^{2\mu t} \cdot (e^{\sigma^2 t - 1}) \tag{2.4.4}$$

2.4.5 O-U Process

其实,我们同样可是使用这种方法来解一类更加广义的随机过程——O-U 过程。这个过程为

$$dX(t) = -\kappa(X(t) - \theta)dt + \sigma dW(t). \tag{2.4.5}$$

在式子 (2.4.5) 中,我们注意到参数 (κ, θ, σ) 决定了整个随机过程的特征:

- κ:是随机过程的「变化率」,即控制了整个随机过程向长期均值回归的快慢程度;
- θ:代表了随机过程的「长期均值水平」,O-U 过程最显著的特征是其具有了「均值回复」,即在变化率 κ 的控制下变量趋于稳定的状态。这个在利率期限结构建模中经常使用,如最早的 Vasicek 利率模型就是一个典型的 O-U 过程。
- σ:表示随机过程的「瞬时方差」。

下面我们来求解显示解。

首先对方程 (2.4.5) 在区域 $[0, t]$ 进行积分得到

$$\int_0^t dX(u) = -\int_0^t \kappa X(u)du + \int_0^t \kappa\theta du + \int_0^t \sigma dW(u),$$

然后两边同时乘以 $e^{\kappa u}$ 得到

$$\int_0^t e^{\kappa u}dX(u) = -\int_0^t \kappa X(u)e^{\kappa u}du + \int_0^t \kappa\theta e^{\kappa u}du + \int_0^t \sigma e^{\kappa u}dW(u). \tag{2.4.6}$$

我们分开求解等式两边。

1. 先是对左边进行分步积分得到

$$LHS = e^{\kappa u}X(u)|_{u=0}^t - \int_0^t \kappa X(u)e^{\kappa u}du = e^{\kappa t}X(t) - X(0) - \int_0^t \kappa X(u)e^{\kappa u}du. \tag{2.4.7}$$

2. 同样的,我们也可以求出右边式子

$$\begin{aligned} RHS &= -\int_0^t \kappa X(u)e^{\kappa u}du + \int_0^t \theta de^{\kappa u} + \int_0^t \sigma e^{\kappa u}dW(u) \\ &= -\int_0^t \kappa X(u)e^{\kappa u}du + \theta(e^{\kappa t} - 1) + \int_0^t \sigma e^{\kappa u}dW(u) \end{aligned} \tag{2.4.8}$$

3. 对比 (2.4.7) 与 (2.4.8),

$$e^{\kappa t}X(t) - X(0) - \int_0^t \kappa X(u)e^{\kappa u}du = -\int_0^t \kappa X(u)e^{\kappa u}du + \theta(e^{\kappa t} - 1) + \int_0^t \sigma e^{\kappa u}dW(u)$$

$$e^{\kappa t}X(t) - X(0) = \theta(e^{\kappa t} - 1) + \int_0^t \sigma e^{\kappa u}dW(u)$$

$$\Rightarrow X(t) = e^{-\kappa t}(X(0) - \theta) + \theta + \int_0^t \sigma e^{-\kappa(t-u)} dW(u) \tag{2.4.9}$$

2.4.6 期望

由 (2.4.9) 得到

$$X(t) = e^{-\kappa t}(X(0) - \theta) + \theta + \int_0^t \sigma e^{-\kappa(t-u)} dW(u).$$

则其期望可以表示为

$$\begin{aligned} \mathbb{E}[X(t)|X(0)] &= \mathbb{E}[e^{-\kappa t}(X(0) - \theta) + \theta + \int_0^t \sigma e^{-\kappa(t-u)} dW(u)] \\ &= e^{-\kappa t}(X(0) - \theta) + \theta. \end{aligned} \tag{2.4.10}$$

最后一项由 $W_t \sim \mathcal{N}(0, t)$ 得到。

利用 (2.4.10), 我们可以求出 O-U 过程在长期的一个均值回复项, 即

$$\lim_{t \to +\infty} e^{-\kappa t}(X(0) - \theta) + \theta = \theta.$$

2.4.7 协方差

任一区间内 $[s, t]$ 协方差可以由以下求出。

$$\begin{aligned} \text{Cov}(X_s, X_t) &= \mathbb{E}[(X_s - \mathbb{E}[X_s]) \cdot (X_t - \mathbb{E}[X_t])] \\ &= \sigma^2 \mathbb{E}\Big[\int_0^s e^{-\kappa(s-u)} dW(u) \cdot \int_0^t e^{-\kappa(t-v)} dW(v)\Big] \end{aligned} \tag{2.4.11}$$

这里我们需要使用到两个基本的概念

- 布朗运动的「独立增量」(indepdent increasement), 即任何一个布朗运动在不同期间内的增量是相互独立的, 即对于区间 $[0, t], 0 \leq t_0 \leq t_1 \leq \cdots \leq t_n \leq t$, $W(t_1) - W(t_0), W(t_2) - W(t_1), \cdots, W(t_n) - W(t_{n-1})$ 是相互独立的增量过程。
- Ito Isometry 性质: 即对于一个布朗运动的独立增量, dW_t, 有关其多项式有如下性质

$$\mathbb{E}\Big[\Big(\int_0^t F(u) dW(u)\Big)^2\Big] = \int_0^t \mathbb{E}[F^2(u)] du$$

利用这两个性质, 对于 (2.4.11), 如果 $s \geq t$ (反之, $s \leq t$, 则令二者调换, 即 $t = \min\{s, t\}$), 我们有

$$\text{Cov}(X_s, X_t) = \sigma^2 \mathbb{E}\Big[\Big(\int_0^t e^{-\kappa(s-u)} dW(u) + \int_t^s e^{-\kappa(s-u)} dW(u)\Big) \cdot \int_0^t e^{-\kappa(t-v)} dW(v)\Big] \tag{2.4.12}$$

$$
\begin{aligned}
&= \sigma^2 \mathbb{E}\left[\int_0^t e^{-\kappa(s-u)} dW(u) \cdot \int_0^t e^{-\kappa(t-v)} dW(v) \right. \\
&\qquad \left. + \int_t^s e^{-\kappa(s-u)} dW(u) \cdot \int_0^t e^{-\kappa(t-v)} dW(v)\right] && (2.4.13) \\
&= \sigma^2 \left\{ \mathbb{E}\left[\int_0^t e^{-\kappa(s-u)} dW(u) \cdot \int_0^t e^{-\kappa(t-v)} dW(v)\right] \right. \\
&\qquad \left. + \mathbb{E}\left[\int_t^s e^{-\kappa(s-u)} dW(u) \cdot \int_0^t e^{-\kappa(t-v)} dW(v)\right] \right\}, && (2.4.14)
\end{aligned}
$$

- 对于第一项我们需要使用 Ito Isometry 性质,

$$
\mathbb{E}\left[\int_0^t e^{-\kappa(s-u)} dW(u) \cdot \int_0^t e^{-\kappa(t-v)} dW(v)\right] = \mathbb{E}\left[\int_0^t e^{-\kappa(s-u)} \cdot e^{-\kappa(t-u)} du\right]
$$

- 而第二向可以由布朗运动的独立增量性质消除,

$$
\begin{aligned}
&\mathbb{E}\left[\int_t^s e^{-\kappa(s-u)} dW(u) \cdot \int_0^t e^{-\kappa(t-v)} dW(v)\right] \\
=&\mathbb{E}\left[\int_t^s e^{-\kappa(s-u)} dW(u)\right] \cdot \mathbb{E}\left[\int_0^t e^{-\kappa(t-v)} dW(v)\right] = 0
\end{aligned}
$$

因此,方程 (2.4.14) 变为

$$
\begin{aligned}
\text{Cov}(X_s, X_t) &= \sigma^2 \left\{ \mathbb{E}\left[\int_0^t e^{-\kappa(s-u)} \cdot e^{-\kappa(t-u)} du\right] \right\} \\
&= \sigma^2 \left\{ e^{-\kappa(s+t)} \mathbb{E}\left[\int_0^t e^{2\kappa u} du\right] \right\} = \sigma^2 \cdot e^{-\kappa(s+t)} \cdot \frac{1}{2\kappa} \cdot e^{2\kappa u} \Big|_{u=0}^{t} \\
&= \frac{\sigma^2}{2\kappa} e^{-\kappa(s+t)} \cdot \left(e^{2\kappa t} - 1 \right).
\end{aligned}
$$

也就是说,对于任何一个协方差,我们都有

$$
\text{Cov}(X_s, X_t) = \frac{\sigma^2}{2\kappa} e^{-\kappa(s+t)} \cdot \left(e^{2\kappa \cdot \min\{s,t\}} - 1 \right).
$$

2.4.8 方差

方差也就是 0 阶协方差,

$$
\begin{aligned}
\text{Var}(X_t) = \text{Cov}(X_t, X_t) &= \frac{\sigma^2}{2\kappa} e^{-\kappa(t+t)} \cdot \left(e^{2\kappa \cdot \min\{t,t\}} - 1 \right) \\
&= \frac{\sigma^2}{2\kappa} \left(1 - e^{-2\kappa t} \right)
\end{aligned}
$$

2.5 有关统计模型的思考

Ruppert 提到的几个在处理量化分析时需要注意的问题,我个人觉得都十分的在理。

这里我简单说说对运用统计模型在金融分析中的理解。此处纯属个人观点,与公司无关。。。。。。

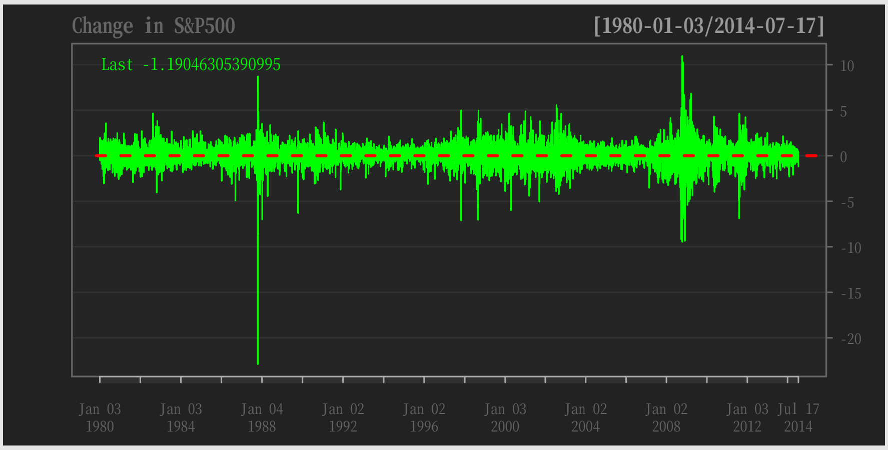
模型是对现实的简要、精辟而不失去其关键信息的数学描述。有关如何将统计模型运用到数量金融分析当中,我们应该多加思考,而不是一味的盲目追随模型。

- 永远都要把数据放在首要位置,只有对数据做出合理的解释才能进一步去拟合各种理论。这里我们可以使用 R 中非常多的优秀做图软件包, 如 ggplot2()。Ruppert 提到说我们可以使用绘图来检测出「好的数据」和「坏的数据」以及「异常值」(outlier)。对于「bad data」, 要么是尽量去修复使其符合需要, 要么就是果断的将其从样本中删除,因为这样的「坏数据」不仅无益于结果分析,反而会提供干扰信息。而对那些看起来是「异常」的数据, 我们需要格外提高警惕。比如, 从金融危机当中产生的「outlier」, 反映了由于市场的结构调整与重新构建而出现的与先前模式完全不同的特征,这些新的数据集要求我们为此提供具有针对性的统计模型。比如,现在比较前沿的统计极值理论,可以运用于对极端风险的建模,这些在金融理论文献当中通常被归类为「Rare Disaster」。从一定意义上说,正是由于我们通过图形发现的「异常特征」数据,才启发学术界与实务界不要止步于现有的数学模型,而是应该激流勇进、迎刃而上,提出更加可信的理论模型。
- 时刻警惕模型的适用性。Box 有句很出名的话:『All models are false but some are usefull』。意思是说,没有任何一个模型能够保证其完全正确性,但是一个科学「有用」的模型必须能够为解释某个现象、某个问题而提出针锋相对的思路,为我们理解问题找到合理而可行的切入点。
- 偏差 -方差的权衡。我们知道,在统计理论里讲究对变量估计的无偏性、相合性。所谓的无偏,指的是变量的估计值等于其期望值,是对真是数值的完美拟合。可是真正做到无偏估计十分的困难,有时是难以找到这样的参数估计,有时是因为纯粹意义上的无偏估计并没有合理的解释。因此,我们面临这相合性的束缚,即估计值对期望值收敛的逼近程度。许多的情况是,无偏性与相合性是一对矛盾的产物,必须在两者之间做出一定的取舍。一个总的原则是采用「简约」(principle of parsimony)原则,即用尽量少的参数来取得尽可能小的误差范围。
- 模型的不确定性。我们知道,任何一个模型对只是对问题的无限逼近的选择,由经济体自发生成系统均衡时的客观数据无法道出其产生的内在机制。我们经济计量学家也只是通过这些显现的测量数据来「推断」其内在的随机过程。因此,任何一个模型都具有不完美的内在缺陷。模型的这样不确定性分析要求我们通过比较不同模型的分析结果来尽可能的选择一个合理的模型。现在,我们已经可以借助 Baysian 方法来比较多个复杂模型的分析结果。
- 金融数据并不一定(甚至完全不是)正态分布。我们原来在课本里学习到的大多数金融理论都把资产回报率假设为一个服从正态高斯分布的随机过程,从而可以利用高斯分布计算风险概率及相应的方差。可是真实的金融数据并非是一个服从常态的高斯分布。比如,股票的回报率往往具有「偏峰」(skewness)、「后尾」(fat tail)的特点,左右两边的分布是不对称的,而且表现出波动率聚集(volatility clustering)。很明显的,这些特征有别于正态高斯分布。往往的,一个可能的备选方案是 t-distribution。
- 金融数据的方差往往不是固定的,而是时变的。我们知道,正态高斯分布有两个参数,$\mathcal{N}(\mu, \sigma^2)$。如果假定收益率服从高斯分布,则暗含着假定其方差是固定的。可

是我们往往看到金融数据的方差是随着时间变化的，尤其在一些特定的时点上，股票市场的波动往往高的惊人。比如，1987 年的黑色星期四，整个 NSE 股票下跌了近 30%。为了处理时变的方差，我们需要引入如 ARCH、GARCH、Voatility Model 等。

Figure 2.5.1 : S&P500 指数变动率情况

```r
chartSeries(rtn("^GSPC", from = "1980-01-01"), name = "Change in S&P500")  ## 画图
abline(h = 0, col = "red", lwd = 2, lty = "dashed")
```



■说明 从这个图形可以看出，金融数据的变动率往往存在着「聚集出现」的现象，即一个较高的波动率往往会导致下一个时期更加可能的出现较高的波动情况。金融市场的这种时变的波动率特征要求我们需要特别处理。