Portfolio Construction and Analytics 读书笔记

1　资产管理的介绍

2　随机变量、概率分布和重要的统计概念

2.1　随机变量

随机变量：定义在样本空间ω 上的实值函数。

2.2　伯努利试验和概率分布函数

设伯努利试验一次成功的概率为p 那么一重伯努利试验有如下的概率分布函数：

$$\Pr(\widetilde{X}=x)=\begin{cases}1-p & \text{x=0}\\ p & \text{x=1}\end{cases}$$

2.3　n重伯努利试验

设伯努利试验一次成功的概率为p 那么n重伯努利试验成功x次的概率为：

$$\Pr(\widetilde{X}=x)=\frac{n!}{x!(n-x)!}p^x(1-p)^{n-x},x=0,\ldots,n$$

2.4　正态分布和概率分布函数

正态分布：

$$f(x|\mu,\sigma)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

概率分布函数(PDF): 表示随机变量在样本空间上的概率分布：

$$\Pr[a\leq X\leq b]=\int_a^b f_X(x)\,dx$$

2.5　累积分布函数

累积分布函数(CDF):

$$F(x)=\Pr[X\leq b]=\int_{-\infty}^b f_X(x)\,dx$$

2.6　描述分布

2.6.1集中趋势的度量

2.6.1.1　均值：

$$\mu=\mathrm{E}\left[X\right]=\int_{-\infty}^{\infty}xP(x)\,dx,$$

2.6.1.2 方差：

$$\operatorname{Var}(X) = \operatorname{E}\left[(X - \operatorname{E}[X])^2\right]$$

2.6.1.3 k阶中心矩:

$$\mu_k = \operatorname{E}\left[(X-\mu)^k\right] = \int_{-\infty}^{\infty} (x-\mu)^k P(x)\,dx,$$

4.矩量母函数

$$M_t(X) = \operatorname{E}\left[e^{tX}\right] = \int_{-\infty}^{\infty} e^{tx} P(x)\,dx,$$

2.6.2风险的度量

2.6.2.1 方差和标准差

在度量投资组合的风险时，首要考虑的就是投资组合的方差和标准差：

1.方差：

$$\begin{aligned}\operatorname{Var}(X) &= \operatorname{E}\left[(X - \operatorname{E}[X])^2\right] \\ &= \operatorname{E}\left[X^2 - 2X\operatorname{E}[X] + \operatorname{E}[X]^2\right] \\ &= \operatorname{E}\left[X^2\right] - 2\operatorname{E}[X]\operatorname{E}[X] + \operatorname{E}[X]^2 \\ &= \operatorname{E}\left[X^2\right] - \operatorname{E}[X]^2\end{aligned}$$

2.标准差：$\sigma_X = \sqrt{Var(X)}$

2.6.2.2 变异系数

当需要比较两组数据离散程度大小的时候，如果两组数据的测量尺度相差太大，或者数据量纲的不同，直接使用标准差来进行比较不合适，此时就应当消除测量尺度和量纲的影响，而变异系数可以做到这一点，它是原始数据标准差与原始数据平均数的比。CV没有量纲，这样就可以进行客观比较了。事实上，可以认为变异系数和极差、标准差和方差一样，都是反映数据离散程度的绝对值。其数据大小不仅受变量值离散程度的影响，而且还受变量值平均水平大小的影响。

变异系数的定义：

$$c_{\mathrm{v}} = \frac{\sigma}{\mu}$$

例如：投资组合A的变异系数为0.8,投资组合B的变异系数为0.5，那么我们可以认为投资组合A的风险比较大。

2.6.2.3 范围

随机变量的范围：即随机变量的取值范围。例如正态分布的取值范围是负无穷到正无穷。

2.6.2.4 百分位数

随机变量X或它的概率分布的分位数Z_a，是指满足条件$\Pr(X \leq Z_a) = \alpha$ 的实数

2.6.2.5　风险价值

风险价值VaR(Value at Risk)：在市场正常波动下，某一金融资产或证券组合的最大可能损失。更为确切的是指，在一定概率水平（置信度）下，某一金融资产或证券组合价值在未来特定时期内的最大可能损失。

给定置信度α:

$$\mathrm{VaR}_\alpha(X) = \inf\left\{x \in \mathbb{R} : F_X(x) > \alpha\right\} = F_Y^{-1}(1-\alpha).$$

2.6.2.6　条件风险价值

在投资组合超过某个给定VaR值的条件下，该投资组合的平均损失值。

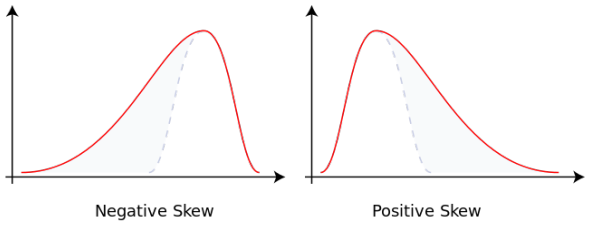
$$CVaR_\alpha(X) = E[-X \mid X \leq -VaR_\alpha(X)]$$

2.6.3偏度

偏度（skewness），是统计数据分布偏斜方向和程度的度量，是统计数据分布非对称程度的数字特征。我们可以从图片中（来自wiki百科）直观地看出正偏度和负偏度：

1.负偏度。密度函数左边的尾巴更加厚实，随机变量主要的取值分布在右边，通常我们也把他称为”右倾斜”。

2.正偏度。密度函数右边的尾巴更加厚实，随机变量主要的取值分布在左边，通常我们也把他称为”左倾斜”。



计算公式：

$$\gamma_1 = \mathrm{E}\left[\left(\frac{X-\mu}{\sigma}\right)^3\right] = \frac{\mu_3}{\sigma^3} = \frac{\mathrm{E}\left[(X-\mu)^3\right]}{\left(\mathrm{E}\left[(X-\mu)^2\right]\right)^{3/2}} = \frac{\kappa_3}{\kappa_2^{3/2}}$$

2.6.4峰度

峰度是描述总体中所有取值分布形态陡缓程度的统计量。这个统计量需要与正态分布相比较，峰度为3表示该总体数据分布与正态分布的陡缓程度相同；峰度大于3表示该总体数据分布与正态分布相比较为陡峭，为尖顶峰；峰度小于3表示该总体数据分

布与正态分布相比较为平坦，为平顶峰。峰度的绝对值数值越大表示其分布形态的陡缓程度与正态分布的差异程度越大。

计算公式：

$$\operatorname{Kurt}[X] = \mathrm{E}\left[\left(\frac{X-\mu}{\sigma}\right)^4\right] = \frac{\mu_4}{\sigma^4} = \frac{\mathrm{E}[(X-\mu)^4]}{(\mathrm{E}[(X-\mu)^2])^2}$$

2.7 协方差和相关系数

协方差:用于衡量两个变量的总体误差。而方差是协方差的一种特殊情况，即当两个变量是相同的情况。期望值分别为E[X]与E[Y]的两个实随机变量X与Y之间的协方差Cov(X,Y)定义为：

$$\operatorname{cov}(X,Y) = \mathrm{E}\left[(X - \mathrm{E}[X])(Y - \mathrm{E}[Y])\right]$$

相关系数：研究变量之间线性相关程度的量。随机变量X和Y 的相关系数定义为：

$$\rho_{X,Y} = \operatorname{corr}(X,Y) = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}$$

2.8 随机变量和的性质

1.随机变量和的期望。

$$\mathrm{E}[aX + bY] = a \cdot \mathrm{E}[X] + b \cdot \mathrm{E}[Y]$$

2.随机变量和的方差。

$$Var[aX + bY] = a^2 \cdot Var[X] + b^2 \cdot Var[Y] + 2 \cdot a \cdot b \cdot Cov(X,Y)$$

2.随机变量和的分布。对于独立的随机变量X 和Y, 随机变量Z = X + Y 的密度函数就是X的密度函数和Y的密度函数的卷积。

$$f_Z(z) = \int_{-\infty}^{\infty} f_Y(z-x) f_X(x)\, dx f_Z(z) = \int_{-\infty}^{\infty} f_Y(z-x) f_X(x)\, dx$$

2.9 联合概率分布和条件概率

对于离散型随机变量X,Y，当X=x时，Y 的条件分布为：

$$P_Y(y \mid X = x) = P(Y = y \mid X = x) = \frac{P(X = x \cap Y = y)}{P(X = x)}$$

同样对于连续型随机变量，当X=x时，Y 的条件分布为：

$$f_Y(y \mid X = x) = \frac{f_{X,Y}(x,y)}{f_X(x)}$$

一般的我们有如下结论：

1.重期望公式：$E(E(X \mid \mathcal{H})) = E(X)$

2.条件方差公式：$\operatorname{Var}(X) = \operatorname{E}(\operatorname{Var}(X \mid \mathcal{H})) + \operatorname{Var}(\operatorname{E}(X \mid \mathcal{H}))$

3.条件协方差公式：$\operatorname{cov}(X, Y) = \operatorname{E}(\operatorname{cov}(X, Y \mid Z)) + \operatorname{cov}(\operatorname{E}(X \mid Z), \operatorname{E}(Y \mid Z))$

边缘分布和密度函数

对于离散型随机变量：

$$\Pr(X = x) = \sum_y \Pr(X = x, Y = y) = \sum_y \Pr(X = x \mid Y = y) \Pr(Y = y)$$

对于连续型随机变量：

$$p_X(x) = \int_y p_{X,Y}(x, y)\, \mathrm{d}y = \int_y p_{X|Y}(x \mid y)\, p_Y(y)\, \mathrm{d}y$$

2.10　Copulas

copula函数描述的是变量间的相关性，实际上是一类将联合分布函数与它们各自的边缘分布函数连接在一起的函数，因此也有人将它称为连接函数。相关理论的提出可以追溯到1959年，Sklar通过定理形式将多元分布与Copula函数联系起来。

Copula函数的定义：$C : [0,1]^d \to [0,1]$ 称为是一个d维copula函数如果：

- $C(u_1, \ldots, u_{i-1}, 0, u_{i+1}, \ldots, u_d) = 0$, 如果某个分量为零，函数为零，
- $C(1, \ldots, 1, u, 1, \ldots, 1) = u$, 如果函数的d-1个分量为1，那么函数是把u映成u，
- C对于它的每一个分量是非减的。

Sklar定理

Sklar 定理（二元形式）：若$H(x, y)$是一个具有连续边缘分布的$F(x)$与$G(y)$的二元联合分布函数，那么存在唯一的copula函数C，使得$H(x, y) = C(F(x), G(y))$。反之，如果C是一个$copula$函数，而F和G是两个任意的概率分布函数，那么由上式定义的H函数一定是一个联合分布函数，且对应的边缘分布刚好就是F和G。

Copulas函数族

1.高斯Copula函数

高斯Copula函数是定义在单位立方体$[0,1]^d$.上面的函数。它是通过定义在$\mathbb{R}^d$上的多元正态函数构造的。

给定一个相关系数矩阵$R \in [-1,1]^{d\times d}$, 参数为R的高斯Copula函数可以写成：

$$C_R^{\text{Gauss}}(u) = \Phi_R\left(\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_d)\right),$$

其中Φ^{-1}标准正态分布函数的逆而Φ_R 是均值为零、协方差为R的多元正态分布。

2.阿基米德Copula函数

函数C 被称为阿基米德copula 如果它满足如下条件：

$$C(u_1, \ldots, u_d; \theta) = \psi^{[-1]}\left(\psi(u_1;\theta) + \cdots + \psi(u_d;\theta);\theta\right)$$

其中ψ: $[0,1]\times\Theta \to [0,\infty)$是连续、严格递减的凸函数，并且满足：$\psi(1;\theta) = 0$。$\theta$ 从属于某个参数空间Θ.

常用的阿基米德copula函数：

名称	$C_\theta(u,v)$	参数θ
Ali-Mikhail-Haq	$\frac{uv}{1-\theta(1-u)(1-v)}$	$\theta \in [-1,1)$
Clayton	$\left[\max\left\{u^{-\theta}+v^{-\theta}-1;0\right\}\right]^{-1/\theta}$	$\theta \in [-1,\infty)\backslash\{0\}$
Frank	$-\frac{1}{\theta}\log\left[1+\frac{(\exp(-\theta u)-1)(\exp(-\theta v)-1)}{\exp(-\theta)-1}\right]$	$\theta \in \mathbb{R}\backslash\{0\}$
Gumbel	$\exp\left[-\left((-\log(u))^\theta + (-\log(v))^\theta\right)^{1/\theta}\right]$	$\theta \in [1,\infty)$
Independence	uv	
Joe	$1-\left[(1-u)^\theta+(1-v)^\theta-(1-u)^\theta(1-v)^\theta\right]^{1/\theta}$	$\theta \in [1,\infty)$

2.11 概率分布和取样

在通常情况下，我们无法知道总体的分布情况。所以我们通常通过样本来估计总体。例如，我们独立地观测到n个样本数据：$X_1, \ldots, X_n$,通常采用如下的公式估计总体的均值、方差等参数：

样本均值：

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

样本方差：

$$s_2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_n - \bar{X})^2$$

样本标准差：

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_n - \bar{X})^2}$$

样本协方差：

$$sCov(\bar{X},\bar{Y}) = \frac{1}{n-1}\sum_{i=1}^{n}(X_n - \bar{X})(Y_n - \bar{Y})$$

样本相关系数：

$$r(\bar{X},\bar{Y}) = \frac{sCoV(\bar{X},\bar{Y})}{s_X s_Y}$$

2.11.1中心极限定理

独立同分布的中心极限定理:

设随机变量$X_1X_2\ldots X_n$独立同分布，并且具有有限的数学期望和方差：$E(X_i) = \mu$，$Var[X_i] = \sigma^2 < \infty$，

则对任意z:

$$\lim_{n\to\infty}\Pr\left[\sqrt{n}(S_n - \mu) \le z\right] = \Phi\left(\frac{z}{\sigma}\right),$$

其中

$$S_n := \frac{X_1 + \cdots + X_n}{n}$$

为样本的均值。

2.11.2置信区间

置信区间是一种常用的区间估计方法，所谓置信区间就是分别以统计量的置信上限和置信下限为上下界构成的区间。对于一组给定的样本数据，其平均值为μ，标准偏差为σ，则其整体数据的平均值的$100(1-\alpha)\,\%$ 置信区间为$(\mu - Z_{\frac{\alpha}{2}}\sigma, \mu + Z_{\frac{\alpha}{2}}\sigma)$，其中 α 为非置信水平在正态分布内的覆盖面积$Z_{\frac{\alpha}{2}}$ 即为对应的标准分数。

2.11.3Bootstrapping

Bootstrapping算法，指的就是利用有限的样本资料经由多次重复抽样，重新建立起足以代表母体样本分布的新样本。我们会在后续的章节中结合蒙特卡洛模拟给出详细的介绍。

2.11.4假设检验

假设检验(Hypothesis Testing)是数理统计学中根据一定假设条件由样本推断总体的一种方法。具体作法是：根据问题的需要对所研究的总体作某种假设，记作H_0；选取合适的统计量，这个统计量的选取要使得在假设H_0成立时，其分布为已知；由实测的样本，计算出统计量的值，并根据预先给定的显著性水平进行检验，作出拒绝或接受假设H_0的判断。常用的假设检验方法有u检验法、t检验法、χ^2检验法(卡方检验)、F 检验法，秩和检验等。我们会在后续的章节中给出详细的介绍。

3　常见的分布函数介绍

3.1　分布函数的样例

3.1.1记号说明

Gama 函数的定义如下：

$$\Gamma(z) = \int_0^{\infty} x^{z-1} e^{-x}\, dx$$

Beta函数的定义如下：

$$\mathrm{B}(x,y) = \int_0^1 t^{x-1}(1-t)^{y-1}\, dt$$

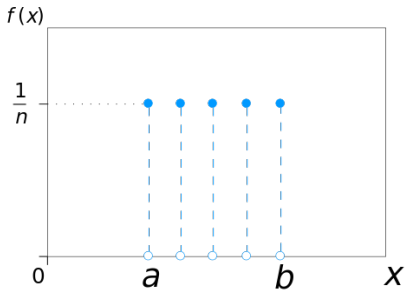
3.1.2离散型和连续型均匀分布

离散型均匀分布：

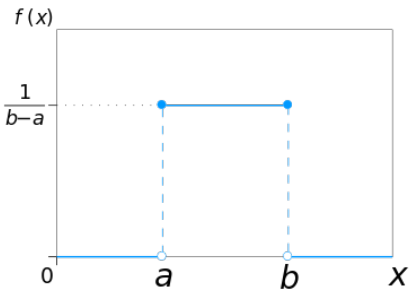
$$\Pr(X = x) = \frac{1}{N}$$

连续型均匀分布：

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{for } a \le x \le b, \\ 0 & \text{for } x < a \text{ or } x > b \end{cases}$$



3-1 离散型均匀分布



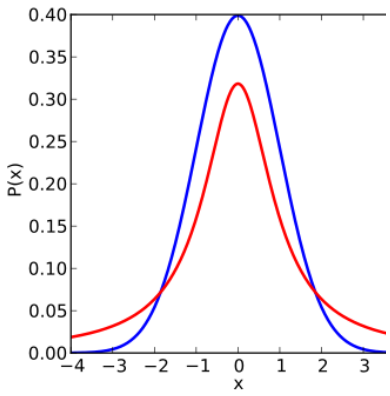
3-2 连续型均匀分布

3.1.3t分布

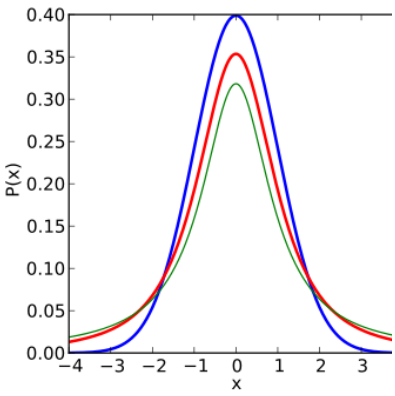
密度函数:

$$f(t) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\,\Gamma(\frac{\nu}{2})} \left(1 + \frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}},$$

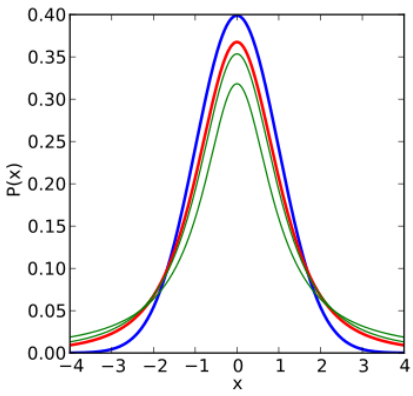
自由度分别为1，2，3的t分布的密度函数。其中蓝色的线条表示正态分布的密度函数。绿色的线条表示上一幅图中t分布的密度函数。



3-3 自由度为1



3-4 自由度为2



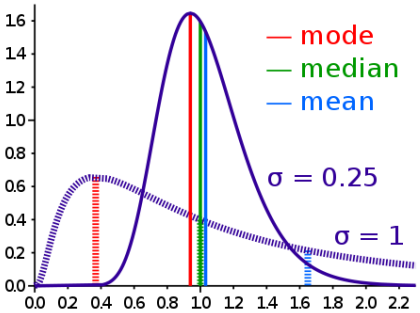
3-5 自由度为3

3.1.4对数正态分布

对数正态分布（logarithmic normal distribution）是指一个随机变量的对数服从正态分布，则该随机变量服从对数正态分布。对数正态分布从短期来看，与正态分布非常接近。但长期来看，对数正态分布向上分布的数值更多一些。

密度函数：

$$
\begin{aligned}
f_X(x) &= \frac{\mathrm{d}}{\mathrm{d}x}\Pr(X \le x) = \frac{\mathrm{d}}{\mathrm{d}x}\Pr(\ln X \le \ln x) \\
&= \frac{\mathrm{d}}{\mathrm{d}x}\Phi\left(\frac{\ln x - \mu}{\sigma}\right) \\
&= \varphi\left(\frac{\ln x - \mu}{\sigma}\right)\frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{\ln x - \mu}{\sigma}\right) \\
&= \varphi\left(\frac{\ln x - \mu}{\sigma}\right)\frac{1}{\sigma x} \\
&= \frac{1}{x}\cdot\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(\ln x - \mu)^2}{2\sigma^2}\right).
\end{aligned}
$$

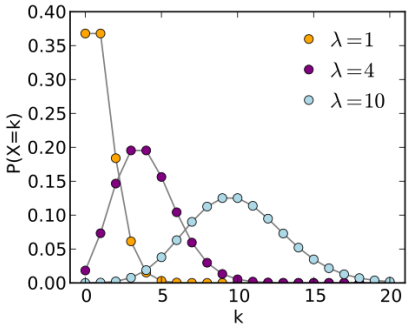


3-6 两种不同偏度的对数正态分布

3.1.5泊松分布

密度函数：

$$P(X = k) = e^{-\lambda}\frac{\lambda^k}{k!}$$



3-7 不同λ值的泊松分布

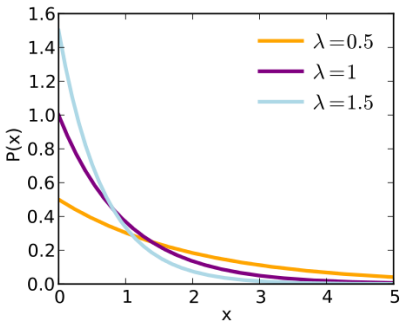
值得一提的是，随着λ的递增，泊松分布会不断逼近均值为λ，方差为$\sqrt{\lambda}$的正态分布，这一结论在金融模型中有重要的应用。

3.1.6指数分布

在概率理论和统计学中，指数分布（也称为负指数分布）是描述泊松过程中的事件之间的时间的概率分布，即事件以恒定平均速率连续且独立地发生的过程。这是伽马分布的一个特殊情况。

密度函数：

$$f(x;\lambda) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0, \\ 0 & x < 0. \end{cases}$$



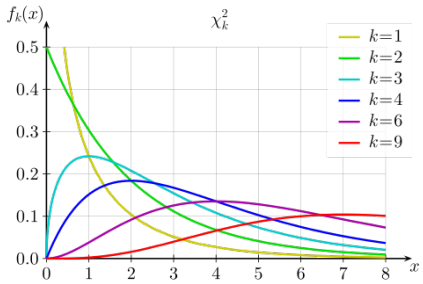
3-8 不同参数的指数分布

3.1.7卡方分布

若n个相互独立的随机变量$\xi_1, \xi_2, \ldots, \xi_n$，均服从标准正态分布（也称独立同分布于标准正态分布），则这n个服从标准正态分布的随机变量的平方和构成一新的随机变量，其分布规律称为卡方分布（chi-square distribution）。

密度函数：

$$f(x;\, k) = \begin{cases} \dfrac{x^{\frac{k}{2}-1} e^{-\frac{x}{2}}}{2^{\frac{k}{2}} \Gamma\left(\frac{k}{2}\right)}, & x > 0; \\ 0, & \text{otherwise.} \end{cases}$$



3-9 不同参数的卡方分布

3.1.8伽玛分布

伽玛分布（Gamma Distribution）是统计学的一种连续概率函数，是概率统计中一种非常重要的分布。”指数分布”和”χ^2分布”都是伽马分布的特例。

Gamma分布中的参数 α 称为形状参数（shape parameter），β 称为尺度参数（scale parameter）。

密度函数：

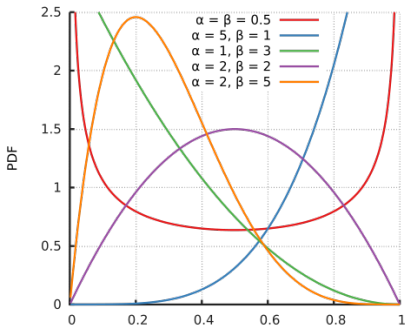
$$f(x;\alpha,\beta) = \frac{\beta^\alpha x^{\alpha-1} e^{-\beta x}}{\Gamma(\alpha)} \quad \text{for } x > 0 \text{ and } \alpha, \beta > 0,$$

3.1.9贝塔分布

在概率论中，贝塔分布，也称 B 分布，是指一组定义在(0,1) 区间的连续概率分布。

密度函数：

$$
\begin{aligned}
f(x;\alpha,\beta) &= \text{constant}\cdot x^{\alpha-1}(1-x)^{\beta-1} \\
&= \frac{x^{\alpha-1}(1-x)^{\beta-1}}{\displaystyle\int_0^1 u^{\alpha-1}(1-u)^{\beta-1}\,du} \\
&= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\,x^{\alpha-1}(1-x)^{\beta-1} \\
&= \frac{1}{\mathrm{B}(\alpha,\beta)}x^{\alpha-1}(1-x)^{\beta-1}
\end{aligned}
$$



3-10 不同参数的贝塔分布

3.2 金融回报率的分布模型

3.2.1椭圆分布族

椭圆分布的密度函数具有如下形式：

$$
f(x) = \frac{c}{\sqrt{|\Sigma|}}\cdot g((x-\mu)'\Sigma^{-1}(x-\mu))
$$

例如，我们所熟知的多元正态分布就属于椭圆分布族：

$$
f_{\mathbf{X}}(x_1,\ldots,x_k) = \frac{\exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)}{\sqrt{(2\pi)^k|\boldsymbol{\Sigma}|}}
$$

3.2.2稳定Paretian分布族

稳定Paretian分布族主要包括如下三个分布：正态分布，柯西分布，列维分布。例

如我们所熟知的柯西分布，其密度函数为:

$$f(x)=\frac{1}{\pi\gamma\left[1+\left(\frac{x-\mu}{\gamma}\right)^2\right]}=\frac{1}{\pi\gamma}\left[\frac{\gamma^2}{(x-\mu)^2+\gamma^2}\right],$$

以及列维分布，其密度函数为：

$$f(x;\mu,\gamma)=\sqrt{\frac{\gamma}{2\pi}}\ \frac{e^{-\frac{\gamma}{2(x-\mu)}}}{(x-\mu)^{3/2}}$$

我们把μ称为位置参数，而把γ称为尺度参数。

3.2.3广义λ分布族

Tukey λ分布：

$$F^{-1}(p)=Q\left(p;\lambda\right)=\begin{cases}\frac{1}{\lambda}\left[p^{\lambda}-(1-p)^{\lambda}\right], & \text{if } \lambda\neq 0\\ \log(\frac{p}{1-p}), & \text{if } \lambda=0,\end{cases}$$

Tukeyλ分布可以近似一些常见的分布：

λ = -1:　接近柯西分布$C(0,\Pi)$

λ = 0:　logistic分布

λ = 0.14:　接近正态分布$N(0,2.142)$

λ = 1:　均匀分布$U(-1,1)$

利用Tukey分布我们可以定义广义λ分布(GLD)：

$$F^{-1}(p)=Q\left(p;\lambda\right)=\lambda_1+\frac{1}{\lambda_2}\left[p_3^{\lambda}-(1-p)_4^{\lambda}\right]$$

上述分布的VaR和CVaR都是容易计算的。

3.3　金融回报率的尾部风险模型

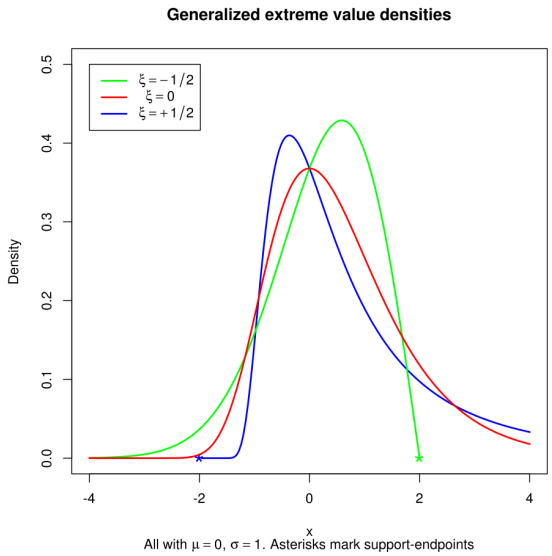
3.3.1广义极值分布

采用标准化的方法：

$$s=(x-\mu)/\sigma$$

可以得到：标准广义极值分布的密度函数：

$$f(s;\xi)=\begin{cases}(1+\xi s)^{(-1/\xi)-1}\exp(-(1+\xi s)^{-1/\xi}) & \xi\neq 0\\ \exp(-s)\exp(-\exp(-s)) & \xi=0\end{cases}$$



3-11 不同参数下的极值分布

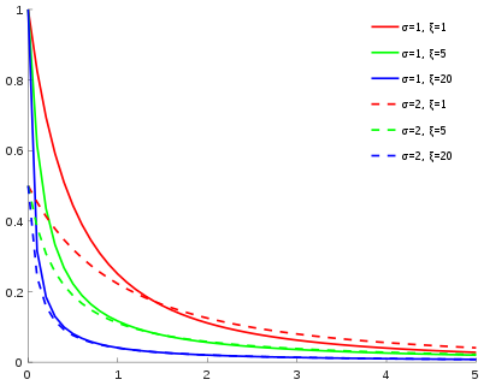
3.3.2广义帕累托分布

同样的采用标准化的方法：

$$s = (x - \mu)/\sigma$$

可以得到：标准广义帕累托分布的密度函数：

$$f_{(\xi,\mu,\sigma)}(x) = \frac{1}{\sigma}\left(1 + \frac{\xi(x-\mu)}{\sigma}\right)^{\left(-\frac{1}{\xi}-1\right)},$$



3-12 不同参数下的帕累托分布

3.3.3极值模型

为了拟合GDP模型，我们通常采用极大似然法估计参数（MLE）一旦GDP的参数

估计完成以后，我们就可以计算模型的VaR和CVaR了：

$$(100-\epsilon)\%VaR = u + \frac{\theta}{\xi}((\frac{n}{N_u}(1-\frac{\epsilon}{100}))^{-\xi}-1)$$

其中u是门限，n是观测数量，N_u是观测值大于门限值的数量。

4　统计学模型

4.1　经典收益模型

最为经典的线性回归模型可以表示为：

$$r_i = \alpha_i + f_1\beta_{i1} + \cdots + f_p\beta_{ip} + \varepsilon_i \qquad i = 1, \ldots, n,$$

r_i:第i份资产第收益率

f_k：影响因子

β:回归系数（灵敏度）

α:常数

ε_i随机扰动

我们之所以给出上述的模型，是假设收益率可以通过一些可观测的因子线性表出。上述的模型的衍生是多种多样的。例如：上述的模型是静态的，我们可以在不同的时间下观测数据，从而建立模型：

$$r_{i,t+1} = \alpha_i + f_{1t}\beta_{i1} + \cdots + f_{pt}\beta_{ip} + \varepsilon_{it} \qquad i = 1, \ldots, n,$$

4.2　回归分析

4.2.1一个简单的例子

假设我们想研究保洁公司股票的收益率和标普500指数的关系，我们可以建立如下的回归模型：

$$r_{P\&G} = \alpha + \beta r_{S\&P500} + \varepsilon$$

其中:

$r_{P\&G}$：保洁公司股票的收益率。

$r_{P\&G500}$:标普500指数收益率。

我们选取了过去78个月份的数据进行回归分析，得到如下的回归方程：

$$r_{P\&G} = 0.0021 + 0.4617 r_{S\&P500}$$

经过P值检验，β系数显著不为0，经过F检验回归方程显著。为了确保回归模型的有效性，我们还需要对残差进行如下的检验：

1.对ε进行正态性检验

2.对ε进行方差齐性检验

3.对ε进行自相关检验

4.2.2回归分析在投资中的应用

回归分析在投资中的应用主要在于以下四个方面：

1.建立投资策略

2.选择投资策略

3.选择标的资产

4.评估策略表现

4.3 因子分析

一个因子分析模型和回归分析模型十分相似，例如资产回报率模型可以写成：

$$r = \alpha + B \cdot f + \varepsilon$$

其中α是N维列向量，表示资产的平均回报率。f是K维因子向量，B是N×K维因子载荷矩阵。

因子分析存在的问题：尽管我们可以计算出具体的因子，但是有时候我们难以对因子给出合理的解释。

4.4 主成分分析

主成分分析（Principal Component Analysis，PCA），是一种统计方法。通过正交变换将一组可能存在相关性的变量转换为一组线性不相关的变量，转换后的这组变量叫主成分。

设是数据的协方差矩阵，PCA的主要步骤:

求解矩阵$\mathbf{W^T}$ 满足如下的条件。

$$Max\ \mathbf{W^T \Sigma W}$$

$$s.t. \mathbf{W^T W} = 1$$

作为一个例子，我们考虑10支股票（AXP, T, BA, CAT, CVX, CSCO, KO, DD, XOM, GE）构成的投资组合在过去78个月内的收益率。我们计算得到各个主成分如下：

EXHIBIT 4.4 Principal components.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
AXP	−0.3125	−0.3965	−0.0053	−0.1065	0.2329	−0.3591	0.7197	−0.0855	−0.0244	0.1563
T	−0.262	0.2583	−0.5274	−0.5126	0.5049	0.2171	−0.0773	0.0167	0.0072	−0.1233
BA	−0.319	−0.1623	0.3003	0.4218	0.51	0.3489	−0.2128	−0.2478	−0.3287	0.0829
CAT	−0.3592	−0.0749	0.0022	−0.3208	−0.3152	−0.3924	−0.4024	−0.092	−0.5604	0.1562
CVX	−0.316	0.4582	0.2953	−0.0732	0.0018	−0.0931	−0.0679	−0.1833	0.4842	0.5622
CSCO	−0.3206	−0.1258	0.1474	−0.2682	−0.4821	0.6971	0.2597	0.0245	−0.0325	−0.0057
KO	−0.2625	0.2633	−0.5851	0.5335	−0.3049	−0.0072	0.1629	−0.3324	−0.0654	−0.0178
DD	−0.3654	−0.3085	0.081	−0.0091	−0.0691	−0.162	−0.2745	−0.2502	0.5104	−0.5788
XOM	−0.2559	0.5627	0.3717	0.0788	0.0302	−0.1638	0.256	0.2926	−0.2266	−0.4931
GE	−0.3634	−0.1923	−0.1883	0.2768	−0.0081	−0.0166	−0.1704	0.7934	0.1595	0.1854

4-13 主成分

EXHIBIT 4.5 Standard deviations of the 10 principal components.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
Standard deviation	2.3003	1.1724	0.9020	0.7931	0.7036	0.6943	0.5649	0.4933	0.4500	0.3865
Proportion of variance	0.5291	0.1374	0.0814	0.0629	0.0495	0.0482	0.0319	0.0243	0.0203	0.0149
Cumulative proportion	0.5291	0.6666	0.7480	0.8109	0.8604	0.9086	0.9405	0.9648	0.9851	1.0000

4-14 各个主成分占比

从中我们可以看到，前三个主成分解释了总体74 %的方差。通常我们选取占比较高的前几个成分作为我们的主成分，一旦我们确定了主成分我们就可以采用如下的公式计算得分了：

$$x_k = \sum_{i=1}^{N} \beta_{ik} r_i$$

4.5　自回归条件方差模型

传统的回归模型假设残差的方差是相同的，然而这样的假设存在如下的问题：

1.金融资产的回报率的振幅随时间变化，并且在某一时期，巨大的资产振幅往往预示着未来一段时间的振幅也会比较大。这也是我们所说的波动集群效应。

2.高频金融数据往往具有厚尾性，这与传统的数据服从正态分布的假设相矛盾。

基于上述的问题，我们引入了ARCH和GARCH模型。

ARCH(q):

以ε_t 表示收益或者收益残差，假设$\varepsilon_t = \sigma_t z_t$ ，此处$z_t \sim i.i.d\ N(0,1)$ （即独立同分布，均符合期望为0，方差为1 的正态分布）。在此条件下σ_t^2可以写成：

$$\sigma_t^2 = \alpha_0 + \alpha_1 \epsilon_{t-1}^2 + \cdots + \alpha_q \epsilon_{t-q}^2 = \alpha_0 + \sum_{i=1}^{q} \alpha_i \epsilon_{t-i}^2,$$

其中$\alpha_i \geq 0,\ i > 0.$

GARCH(p,q):

$y_t = x_t' b + \epsilon_t$

$\epsilon_t | \psi_{t-1} \sim \mathcal{N}(0, \sigma_t^2)$

$\sigma_t^2 = \omega + \alpha_1 \epsilon_{t-1}^2 + \cdots + \alpha_q \epsilon_{t-q}^2 + \beta_1 \sigma_{t-1}^2 + \cdots + \beta_p \sigma_{t-p}^2 = \omega + \sum_{i=1}^{q} \alpha_i \epsilon_{t-i}^2 + \sum_{i=1}^{p} \beta_i \sigma_{t-i}^2$

对于ARCH和GARCH模型，我们容易计算他们的VaR和CVaR：

$$VaR_{1-\varepsilon} = V_t(-u_r + q_{1-\varepsilon}\sigma_r)$$

其中$q_{1-\varepsilon}$为标准正态分布$100(1-\varepsilon)$分位数。同时：

$$CVaR_(1-\varepsilon) = V_t(-u_r + \frac{\phi(q_{1-\varepsilon})}{\varepsilon}\sigma_r)$$

5 模型模拟

5.1 蒙特卡罗模拟

假设你现在有1000美元。你计划用这笔资金投资标普500指数。C_0, C_1分别表示期初和期末（一年以后）的资金。$r_{0,1}$表示期间资金的回报率。则有：

$$C_1 = (1 + r_{0,1})C_0$$

在$[t, t+1]$区间内资金的回报率可以表示成：

$$r_t = \frac{P_{t+1} - P_t + D_t}{P_t}$$

P_t:标普500指数在t时刻的价格

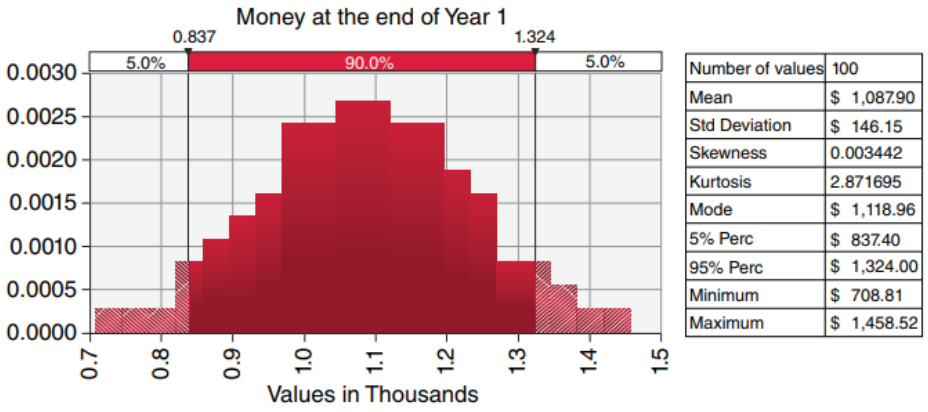
D_t:在区间$[t, t+1]$内的标普500指数的股息

5.1.1选择分布函数

为了采用蒙特卡罗模拟我们的模型，我们首先要确定模拟数据的分布函数。这里有两种方法确定数据的分布函数。第一种是采用历史数据，第二种是随机产生一种给定分布的数据序列。例如，我们可以生成一组均值为μ，方差为σ的正态分布的数据。

5.1.2理解蒙特卡罗模拟的输出

作为蒙特卡罗模拟的一个例子，我们模拟了标普指数一年以后的收益情况。我们生成了100个服从正态分布的随机数（均值为8.79%,方差为14.65%）图表中的直方图展示了未来一年的收益率分布：



5-15 未来一年标普500收益率分布

根据统计学的知识我们可以知道，我们有95%的置信度认为一年后的收益率分布的区间：

$$\left(\bar{x} - z^* \frac{\sigma}{\sqrt{n}}, \bar{x} + z^* \frac{\sigma}{\sqrt{n}}\right) = (1058.9, 1116.9)$$

5.2　为什么采用蒙特卡罗模拟

上面的例子给出了蒙特卡罗模拟的基本方法，我们采用上述方法，给出了一年以后收益率的分布情况。值得注意的是，如果标普500收益率不满足正态分布，那么我们的模拟就不准确了。下面的例子是一个更加复杂的模拟。

5.2.1多个输入变量和混合分布

假设你为未来养老而计划做一笔投资，投资的本金为1000美元，投资的周期为30年。假设标普500指数服从均值为μ，方差为σ的正态分布。记期初的资金为C_0,期末的资金为C_{30}。容易知道，期末的收益可以写成：

$$r_{0,t} = (1 + r_{0,1})(1 + r_{1,2}) \dots (1 + r_{t-1,t}) - 1$$

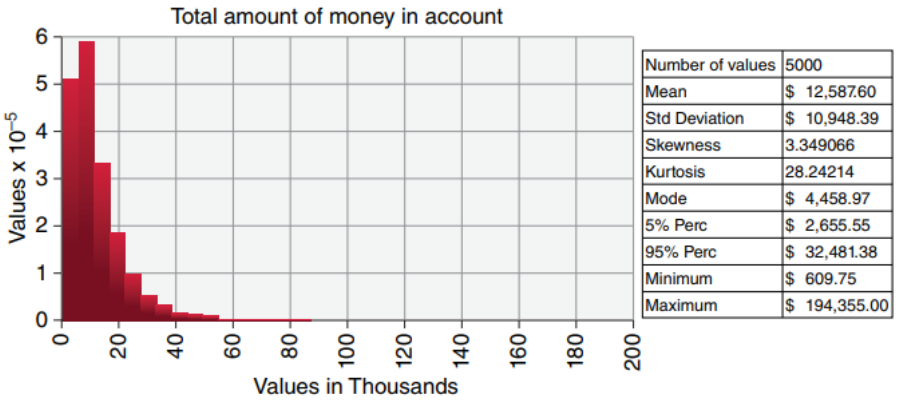
期末的收益取决于这30个正态分布。

5.2.2合并相关

如果我们投资的是两类标的：国债和股票。这二者的收益率存在着负相关。那么我们又该如何模拟呢？不妨假设股票和国债收益率的相关系数为-0.2。国债收益率服从均值为4%，方差为7%的正态分布。我们采用蒙特卡罗进行了500次模拟，下图展示了30年后收益率的分布情况：

5.2.3模型评估

如何评估模型的好坏呢？我们考虑下面两种不同的策略：



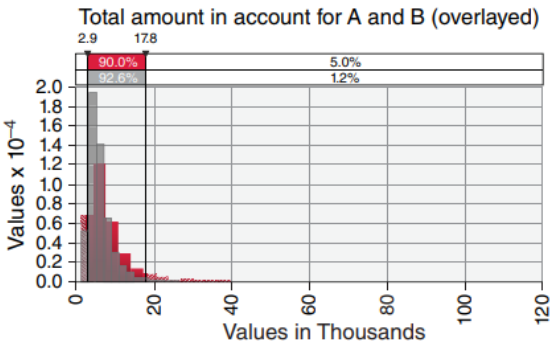
Number of values	5000
Mean	$ 12,587.60
Std Deviation	$ 10,948.39
Skewness	3.349066
Kurtosis	28.24214
Mode	$ 4,458.97
5% Perc	$ 2,655.55
95% Perc	$ 32,481.38
Minimum	$ 609.75
Maximum	$ 194,355.00

Exhibit 5.3 Output distribution for amount of capital after 30 years.

A：50%资金投资股票，50%资金投资债券

B：30%资金投资股票，70%资金投资债券

我们对投资组合A和投资组合B都进行了400次模拟。从模拟的结果我们可以看到，30年后，投资组合A 的平均收益大于投资组合B。但是投资组合A的变异系数也大于投资组合B。这说明投资组合A的风险比B大。课本上的图标罗列了二者的均值、方差、峰度、偏度。从中我们可以直观地看到两个投资组合的差异。



	Strategy A	Strategy B
Number of values	5000	5000
Mean	$ 7,905.30	$6,040.17
Std Deviation	$ 5,341.57	$3,219.06
Skewness	3.698306	3.514493
Kurtosis	43.44567	41.09948
Mode	$ 4,856.33	$4,389.39
5% Perc	$ 2,930.51	$2,809.56
95% Perc	$ 17,834.32	$11,940.22
Minimum	$ 1,131.24	$1,287.31
Maximum	$ 116,550.20	$70,578.09

5.2.4模拟多少次？

模拟的数据至少要大于等于30才具有统计学意义。在上述的例子中我们模拟了400个样本用于分析。一般来说提高样本的数量有助于我们提高模拟的精度。

5.2.5随机数的生成

Excel、R、MATLAB 都可以很方便地生成随机数，可参考各软件的文档、也可百度。

6　模型优化

本章我们主要讨论的是优化问题－－在一系列限制条件下如何使得我们的模型是最优的。

6.1　优化公式

一个优化问题的数学表达式主要由以下三个部分组成：

1.一系列的决策变量（一般由一个$N \times 1$维的向量组成）

2.一个目标函数

3.一系列的限制条件$(g_i(x), h_j(x))$满足：$g_i(x) \leq 0, h_j(x) = 0$

一般来说我们的目标函数总是可以写成：

Maximize：资产预期回报率

6.1.1最大化和最小化

通常来说，最优化问题可以写成如下的表达式：

$$\begin{aligned} &\text{minimize } f(x) \\ &\text{subject to } g_i(x) \leq 0 \;\; i \in \{1, \ldots, I\} \\ &\qquad\qquad h_j(x) = 0 \;\; j \in \{1, \ldots, J\}. \end{aligned}$$

同时最大化问题和最小化问题也是可以相互转化的：

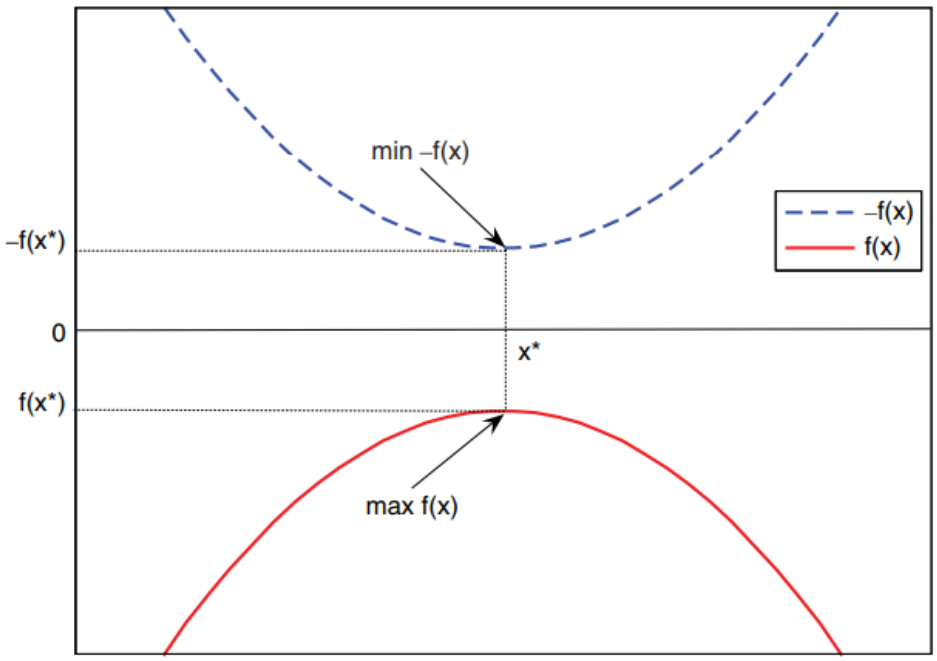
$$\max_x f(x) = -\min_x -f(x)$$

6.1.2局部最优和全局最优

在求解最优解的时候，部分算法（例如梯度下降法）容易陷入局部最优解而无法得到全局最优解。

6.1.3多目标优化

多目标规划是数学规划的一个分支。研究多于一个的目标函数在给定区域上的最优化。又称多目标最优化。通常记为MOP(multi-objective programming)。求解多目标线性规划的基本思想是将多目标转化为单目标，常见的方法有理想点法、线性加权法、最大最小法、目标规划法、模糊数学解法等。



6-16 最大化f(x)相当于最小化-f(x)

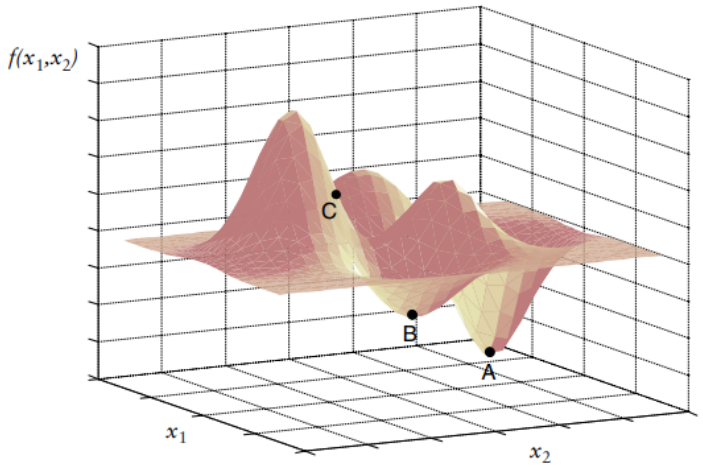


Exhibit 6.2 Global (point A) versus local (point B) minimum for a function of two variables x_1 and x_2.

6.2　重要的优化问题

6.2.1凸优化

所谓的凸优化问题是指我们的目标函数和约束条件都是凸函数：

$$\begin{aligned} \min_{x} \quad & f(x) \\ \text{subject to} \quad & g_i(x) \leq 0 \; i = 1, \ldots, I \\ & A\mathbf{x} \leq \mathbf{b} \end{aligned}$$

其中$f(x)$和$g_i(x)$都是凸函数。

6.2.2线性规划

线性规划（Linear programming,简称LP）是运筹学中研究较早、发展较快、应用广泛、方法较成熟的一个重要分支，它是辅助人们进行科学管理的一种数学方法。研究线性约束条件下线性目标函数的极值问题的数学理论和方法。线性规划问题的标准形式：

$$\begin{aligned} \min_{x} \quad & \mathbf{c}^{\mathrm{T}}\mathbf{x} \\ \text{subject to} \quad & A\mathbf{x} \leq \mathbf{b} \\ & \mathbf{x} \geq \mathbf{0} \end{aligned}$$

6.2.3二次规划

二次规划的一般形式可以表示为：

$$\begin{aligned} \min_{x} \quad & \tfrac{1}{2}\mathbf{x}^{\mathrm{T}}Q\mathbf{x} + \mathbf{c}^{\mathrm{T}}\mathbf{x} \\ \text{subject to} \quad & A\mathbf{x} \leq \mathbf{b}, \\ & \mathbf{x} \geq \mathbf{0} \end{aligned}$$

其中:

x是一个N维决策变量

Q是一个$N \times N$维矩阵

c是一个N维向量

A是一个$J \times N$维矩阵

b是一个J维矩阵

6.2.4二阶锥规划

一个二阶锥规划问题(SOCP)是指具有如下形式的图规划问题：

$$\begin{aligned} \min_{x} \quad & c^T x \\ \text{subject to} \quad & Ax = b \\ & \|A_i x + d_i\|_2 \leq c_i^T x + e_i, \quad i = 1, \ldots, I \end{aligned}$$

其中:

c是一个N维变量

A是一个$J \times N$的矩阵

b是一个J维向量

C_i是一个$I_i \times N$维矩阵

d_i是一个I_i维矩阵

e_i是标量

6.2.5整数规划

整数规划是指规划中的变量（全部或部分）限制为整数，若在线性模型中，变量限制为整数，则称为整数线性规划。

整数规划又分为：

1、纯整数规划：所有决策变量均要求为整数的整数规划

2、混合整数规划：部分决策变量均要求为整数的整数规划

3、纯0－1整数规划：所有决策变量均要求为0－1的整数规划

4、混合0－1规划：部分决策变量均要求为0－1的整数规划

6.3 一个简单的优化例子:资产分配

现在有一个资产管理者计划投资1千万美元于下面四种基金：

EXHIBIT 6.4 Data for the portfolio manager's problem.

Fund Type	Growth	Index	Bond	Money Market
Fund #	1	2	3	4
Expected return	20.69%	5.87%	10.52%	2.43%
Risk level	4	2	2	1
Max investment	40%	40%	40%	40%

他对自己的投资分配有如下的限定：投资于任何一种基金的比例不超过40%、投资于基金1和基金3的金额之和不超过总投资的60%、平均的投资风险水平不能超过2。

我们假定$x=(x_1,x_2,x_3,x_4)$表示投资于四种基金的金额。那么我们的目标函数可以写成：

$$f(x)=\mu^T x=(20.69\%)x_1+(5.87\%)x_2+(10.52\%)x_3+(2.43\%)x_4$$

我们的约束条件可以写成：

1.总的投资金额应该等于1千万美元：

$$x_1+x_2+x_3+x_4=10,000,000$$

2.投资于基金1和基金3的金额之和不超过总投资的60%：

$$x_1+x_3\leq 6,000,000$$

3.平均的投资风险水平不能超过2:

$$\frac{4x_1+2x_2+2x_3+x_4}{x_1+x_2+x_3+x_4}\leq 2$$

由于$x_1+x_2+x_3+x_4=10,000,000$,所以我们化简得到：

$$4x_1+2x_2+2x_3+x_4\leq 20,000,000$$

4.投资于各个基金的比例不超过40%:

$$x_1\leq 4,000,000, x_2\leq 4,000,000, x_3\leq 4,000,000, x_4\leq 4,000,000$$

5.当然，投资金额是非负的：

$$x_1\geq 0, x_2\geq 0, x_3\geq 0, x_4\geq 0$$

最后我们的问题可以写成如下的形式：

$$\max_{x_1,x_2,x_3,x_4}\begin{bmatrix}0.2069 & 0.0587 & 0.1052 & 0.0243\end{bmatrix}\cdot\begin{bmatrix}x_1\\x_2\\x_3\\x_4\end{bmatrix}$$

$$\begin{bmatrix}1 & 1 & 1 & 1\end{bmatrix}\cdot\begin{bmatrix}x_1\\x_2\\x_3\\x_4\end{bmatrix}=10,000,000$$

$$\begin{bmatrix} 1 & 0 & 1 & 0 \\ 4 & 2 & 2 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} \leq \begin{bmatrix} 6,000,000 \\ 20,000,000 \\ 4,000,000 \\ 4,000,000 \\ 4,000,000 \\ 4,000,000 \end{bmatrix}$$

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} \geq \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

6.4 优化算法

求解线性规划问题的基本方法是单纯形法。单纯形算法利用多面体的顶点构造一个可能的解，然后沿着多面体的边走到目标函数值更高的另一个顶点，直至到达最优解为止。虽然这个算法在实际上很有效率，在小心处理可能出现的“循环”的情况下，可以保证找到最优解，但它的最坏情况可以很坏：可以构筑一个线性规划问题，单纯形算法需要问题大小的指数倍的运行时间才能将之解出。事实上，有一段时期内人们曾不能确定线性规划问题是NP完全问题还是可以在多项式时间里解出的问题。

第一个在最坏情况具有多项式时间复杂度的线性规划算法在1979年由前苏联数学家Leonid Khachiyan提出。这个算法建基于非线性规划中Naum Shor发明的椭球法（ellip-soid method），该法又是Arkadi Nemirovski（2003年冯诺伊曼运筹学理论奖得主）和D. Yudin的凸集最优化椭球法的一般化。

理论上，“椭球法”在最恶劣的情况下所需要的计算量要比“单形法”增长的缓慢，有希望用之解决超大型线性规划问题。但在实际应用上，Khachiyan的算法令人失望：一般来说，单纯形算法比它更有效率。它的重要性在于鼓励了对内点算法的研究。内点算法是针对单形法的“边界趋近”观念而改采“内部逼近”的路线，相对于只沿着可行域的边沿进行移动的单纯形算法，内点算法能够在可行域内移动。

1984年，贝尔实验室印度裔数学家卡马卡（Narendra Karmarkar）提出了投影尺度法（又名Karmarkar’s algorithm）。这是第一个在理论上和实际上都表现良好的算法：它的最坏情况仅为多项式时间，且在实际问题中它比单纯形算法有显著的效率提升。自此之后，很多内点算法被提出来并进行分析。一个常见的内点算法为Mehrotra predictor-corrector method。尽管在理论上对它所知甚少，在实际应用中它却表现出色。

6.5 优化软件

matlab、python、c++、java 理论上都可以用于求解。

6.6 一个求解的例子

6.6.1Excel求解

这里我采用的是matlab求解。代码如下：

```
f=[-0.2069  -0.0587  -0.1052  -0.0243];
Aineq=[1 0 1 0;4 2 2 1;1 0 0 0 ;0 1 0 0 ;0 0 1 0 ;0  0 0
    1];
bineq=[6 20 4 4 4 4];
Aeq=[1,1,1,1];
beq=10;
lb=[0 0 0 0 ];
%% Start with the default options
options = optimoptions('linprog');
%% Modify options setting
options = optimoptions(options,'Display', 'off');
options = optimoptions(options,'Algorithm', 'interior-
   point');
[x,fval,exitflag,output,lambda] = ...
linprog(f,Aineq,bineq,Aeq,beq,lb,[],[],options);
```

6.6.2求解结果

运行上述代码得到求解的结果：

$$x = \begin{bmatrix} 2,000,000 \\ 0 \\ 4,000,000 \\ 4,000,000 \end{bmatrix}$$

$$fmax = 931,800$$

所以我们投资于四种基金的金额分别为：2,000,000美元，0美元，4,000,000美元，4,000,000美元。一年后我们的预期收益为931,800美元。