第 9 章　主成分分析和因子模型

许多金融组合包含多个资产, 它们的收益率同时并动态地依赖于许多经济和金融变量. 因此利用合理的多元统计分析方法来研究组合收益率的行为和性质很重要. 然而, 如前几章所述, 对多元资产收益率的分析通常需要高维统计模型, 而这些模型很复杂并且很难应用. 为了使多元收益率的建模更加简单, 本章讨论一些降低维数的方法来寻找这些资产的内在结构. 一般来说, 降低维数最常用的统计方法是主成分分析 (PCA). 我们的讨论也从该方法开始. 实际中所观测到的收益率序列通常呈现出相似的特征, 这使得人们相信它们是由共同的因素驱动的. 这些共同的因素称为公共因子. 为了研究资产收益率的共同形式和简化组合分析, 许多文献给出了很多因子模型来分析多元资产收益率. 本章的第二个目的是引进一些有用的因子模型, 并说明它们在金融中的应用.

有三种类型的因子模型可用来研究资产收益率. 参见 Connor(1995) 与 Campbell, Lo 和 MacKinlay(1997). 第一种类型是宏观经济因子模型. 该模型利用宏观经济变量来描述资产收益率的共同的行为, 其中, 这些宏观经济变量包括 GDP 增长率、利率、通货膨胀率以及失业人数等. 由于该类模型的因子可以观测, 从而可以利用线性回归的方法来估计模型. 第二种类型是基本面因子模型. 该类模型用企业或资产的具体属性来构建公共因子. 例如企业规模、账面价值与市场价值以及产业分类. 第三种类型是统计因子模型. 该类模型把公共因子看成是需要用收益率序列估计的不可观测的变量或隐变量. 本章将讨论这三类因子模型以及它们在金融中的应用. Alexander (2001) 与 Zivot 和 Wang (2003) 也讨论了资产收益率的主成分分析和因子模型.

本章的结构安排如下: 9.1 节介绍资产收益率的一般因子模型; 9.2 节讨论宏观经济因子模型并给出一些简单的例子; 基本面因子模型及其应用在 9.3 节中给出; 9.4 节介绍统计因子分析最基本的方法 —— 主成分分析 (在多元分析中它是用来降低维数的); 9.5 节讨论正交因子模型, 包括因子旋转及其估计, 并给出了例子; 最后, 9.6 节介绍渐近主成分分析.

9.1　因 子 模 型

假定有 k 个资产和 T 个时间周期. r_{it} 表示资产 i 在第 t 个时间周期内的收益. 因子模型的一般形式为

$$r_{it} = \alpha_i + \beta_{i1} f_{1t} + \cdots + \beta_{im} f_{mt} + \varepsilon_{it}, \quad t = 1, \cdots, T; \quad i = 1, \cdots, k, \tag{9.1}$$

其中 α_i 是常数表示截距, $\{f_{jt}|j=1,\cdots,m\}$ 是 m 个公共因子, β_{ij} 是资产 i 在因子 j 上的负荷, ε_{it} 是资产 i 的个性因子.

对于资产收益率, 假定因子 $\boldsymbol{f}_t=(rf_{1t},\cdots,f_{mt})'$ 是 m 维平稳过程, 满足

$$\begin{aligned}\mathrm{E}(\boldsymbol{f}_t)&=\boldsymbol{\mu}_f,\\ \mathrm{Cov}(\boldsymbol{f}_t)&=\boldsymbol{\Sigma}_f,\quad m\times m\text{ 矩阵.}\end{aligned}$$

资产的个性因子 ε_{it} 是白噪声序列, 并且与公共因子 f_{jt} 和其他个性因子不相关. 具体地, 我们假定

$$\begin{aligned}E(\varepsilon_{it})&=0,\quad \text{所有的 } i \text{ 和 } t,\\ \mathrm{Cov}(f_{it},\varepsilon_{js})&=0,\quad \text{所有的 } j,i,t \text{ 和 } s,\\ \mathrm{Cov}(\varepsilon_{it},\varepsilon_{js})&=\begin{cases}\sigma_i^2, & \text{若 } i=j \text{ 且 } t=s,\\ 0, & \text{其他.}\end{cases}\end{aligned}$$

因此, 公共因子与个性因子不相关, 并且个性因子之间也是不相关的. 然而在一些因子模型中并不要求公共因子之间是不相关的.

在某些应用中, 资产的个数 k 可能比时间周期的个数 T 大. 我们将在 9.6 节分析这样的数据. 在因子分析中通常假定因子之间是序列不相关的, 从而 $\boldsymbol{r}_t$ 也是序列不相关的. 在应用中, 如果观测到的收益率序列是序列相关的, 则可以用第 8 章的模型消除序列相关性.

(9.1) 式的因子模型可以写成下述矩阵形式:

$$r_{it}=\alpha_i+\boldsymbol{\beta}_i\boldsymbol{f}_t+\boldsymbol{\varepsilon}_{it},$$

其中 $\boldsymbol{\beta}_i=(\boldsymbol{\beta}_{i1},\cdots,\boldsymbol{\beta}_{im})$, t 时刻 k 个资产的联合模型是

$$\boldsymbol{r}_t=\boldsymbol{\alpha}+\boldsymbol{\beta}\boldsymbol{f}_t+\boldsymbol{\varepsilon}_t,\quad t=1,\cdots,T, \tag{9.2}$$

其中 $\boldsymbol{r}_t=(r_{1t},\cdots,r_{kt})'$, $\boldsymbol{\alpha}=(\alpha_1,\cdots,\alpha_k)'$, $\boldsymbol{\beta}=[\boldsymbol{\beta}_{ij}]$ 是 $k\times m$ 因子负荷矩阵, $\boldsymbol{\varepsilon}_t=(\varepsilon_{1t},\cdots,\varepsilon_{kt})'$ 是误差向量且 $\mathrm{Cov}(\boldsymbol{\varepsilon}_t)=\boldsymbol{D}=\mathrm{diag}\{\sigma_1^2,\cdots,\sigma_k^2\}$ 是 $k\times k$ 对角矩阵. 从而, 收益率 $\boldsymbol{r}_t$ 的协方差矩阵为

$$\mathrm{Cov}(\boldsymbol{r}_t)=\boldsymbol{\beta}\boldsymbol{\Sigma}_f\boldsymbol{\beta}'+\boldsymbol{D}.$$

如果因子 f_{jt} 是可以观测的, 则 (9.2) 式的这种模型表示具有横截面回归的形式.

把 (9.1) 式的因子模型看做时间序列, 对第 i 个资产我们有

$$\boldsymbol{R}_i=\alpha_i\mathbf{1}_T+\boldsymbol{F}\boldsymbol{\beta}_i'+\boldsymbol{E}_i, \tag{9.3}$$

其中 $\boldsymbol{R}_i=(r_{i1},\cdots,r_{iT})'$, $i=1,\cdots,k$, $\mathbf{1}_T$ 是所有元素都为 1 的 T 维向量, $\boldsymbol{F}$ 是 $T\times m$ 矩阵且其第 t 行是 $\boldsymbol{f}_t'$, $\boldsymbol{E}_i=(\varepsilon_{i1},\cdots,\varepsilon_{iT})'$. $\boldsymbol{E}_i$ 的协方差矩阵 $\mathrm{Cov}(\boldsymbol{E}_i)=\sigma_i^2\boldsymbol{I}$ 是 $T\times T$ 对角阵.

最后, (9.2) 式可改写为

$$\boldsymbol{r}_t = \boldsymbol{\xi}\boldsymbol{g}_t + \boldsymbol{\xi}_t,$$

其中 $\boldsymbol{g}_t = (1, \boldsymbol{f}_t')'$, $\boldsymbol{\xi} = [\boldsymbol{\alpha}, \boldsymbol{\beta}]$ 是 $k \times (m+1)$ 矩阵. 对上式取转置并把所有的数据放在一块, 则可以得到

$$\boldsymbol{R} = \boldsymbol{G}\boldsymbol{\xi}' + \boldsymbol{E}, \tag{9.4}$$

其中 $\boldsymbol{R}$ 是 $T \times k$ 收益率矩阵, 其第 t 行是 $\boldsymbol{r}_t'$, 或等价地其第 i 列是由 (9.3) 式定义的 $\boldsymbol{R}_i$; $\boldsymbol{G}$ 是 $T \times (m+1)$ 矩阵, 其第 t 行是 $\boldsymbol{g}_t'$; $\boldsymbol{E}$ 是 $T \times k$ 个性因子矩阵, 其第 t 行是 $\boldsymbol{\varepsilon}_t'$. 如果公共因子 $\boldsymbol{f}_t$ 可以观测, 则 (9.4) 式是多元线性回归模型 (MLR) 的一种特殊形式. 参见 Johnson 和 Wichern (2002). 对于一般的 MLR 模型, 不要求 $\boldsymbol{\varepsilon}_t$ 的协方差矩阵是对角阵.

9.2 宏观经济因子模型

由于宏观经济因子模型中的因子是可以观测的, 从而可以利用最小二乘方法来估计 (9.4) 式的 MLR 模型. 估计为

$$\hat{\boldsymbol{\xi}}' = \begin{bmatrix} \hat{\boldsymbol{\alpha}}' \\ \hat{\boldsymbol{\beta}}' \end{bmatrix} = (\boldsymbol{G}'\boldsymbol{G})^{-1}(\boldsymbol{G}'\boldsymbol{R}),$$

从中可以很容易地得到 $\boldsymbol{\alpha}$ 和 $\boldsymbol{\beta}$ 的估计. (9.4) 式的残差为

$$\hat{\boldsymbol{E}} = \boldsymbol{R} - \boldsymbol{G}\hat{\boldsymbol{\xi}}'.$$

基于对模型的假定, $\boldsymbol{\varepsilon}_t$ 的协方差矩阵可以由下式估计:

$$\hat{\boldsymbol{D}} = \operatorname{diag}\{\hat{\sigma}_1^2, \cdots, \hat{\sigma}_k^2\},$$

其中 $\hat{\sigma}_i^2$ 是 $\hat{\boldsymbol{E}}'\hat{\boldsymbol{E}}/(T-m-1)$ 的第 (i,i) 个元素. 此外, 方程 (9.3) 中第 i 种资产的 R^2 为

$$R_i^2 = 1 - \frac{[\hat{\boldsymbol{E}}'\hat{\boldsymbol{E}}]_{i,i}}{[\boldsymbol{R}'\boldsymbol{R}]_{i,i}}, \quad i = 1, \cdots, k,$$

其中 $\boldsymbol{A}_{i,i}$ 表示矩阵 $\boldsymbol{A}$ 的第 (i,i) 元.

注意到先前的估计并没有要求个性因子 ε_{it} 彼此不相关. 因此一般来说所得到的估计不是有效的. 然而加上正交化限制经常需要大量的计算, 而且通常是可以忽略的. 我们可以检查 $\hat{\boldsymbol{E}}'\hat{\boldsymbol{E}}/(T-m-1)$ 的非对角线元素来验证所拟合模型的充分性. 这些元素应该接近于 0.

9.2.1 单因子模型

金融中最著名的宏观经济因子模型是市场模型. 参见 Sharpe (1970). 该市场模型就是下述单因子模型:

$$r_{it} = \alpha_i + \beta_i r_{mt} + \varepsilon_{it}, \quad i = 1, \cdots, k; \quad t = 1, \cdots, T, \tag{9.5}$$

其中 r_{it} 是第 i 个资产的超额收益率, r_{mt} 是市场的超额收益率. β_i 就是对股票收益率来说众所周知的 β. 为了进一步说明, 考虑 13 只股票的月收益率并且把标准普尔 500 指数的收益率作为市场收益率. 表 9-1 给出了所用到的股票及其代码. 样本区间是从 1990 年 1 月到 2003 年 12 月, 因此 $k = 13, T = 168$. 我们利用二级市场上的三个月期国库券的月收益作为无风险利率来计算股票和市场指数的超额收益. 这些收益率均以百分比的形式给出.

表 9-1　单因子模型分析中所用到股票及其代码 [a]

Tick	Company	$\bar{r}(\sigma_r)$	Tick	Company	$\bar{r}(\sigma_r)$
AA	Alcoa	1.09(9.49)	KMB	Kimberly-Clark	0.78(6.50)
AGE	A.G.Edwards	1.36(10.2)	MEL	Mellon Financial	1.36(7.80)
CAT	Caterpillar	1.23(8.71)	NYT	New York Times	0.81(7.37)
F	Ford Motor	0.97(9.77)	PG	Procter&Gamble	1.08(6.75)
FDX	FedEx	1.14(9.49)	TRB	Chicago Tribune	0.95(7.84)
GM	General Motors	0.64(9.28)	TXN	Texas Instrument	2.19(13.8)
HPQ	Hewlett-Packard	1.37(11.8)	SP5	S&P500 index	0.42(4.33)

a 表中还给出了超额收益率的样本均值和样本标准差. 样本区间是从 1990 年 1 月到 2003 年 12 月.

我们用 S-Plus 来执行上一小节所讨论的估计方法. 所用的大部分命令都能在免费软件 R 中应用.

```
> x=read.matrix(``m-fac9003.txt'',header=T)
> xmtx=cbind(rep(1,168),x[,14])
> rtn=x[,1:13]
> xit.hat=solve(xmtx,rtn)
> beta.hat=t(xit.hat[2,])
> E.hat=rtn-xmtx%*%xit.hat
> D.hat=diag(crossprod(E.hat)/(168-2))
> r.square=1-(168-2)*D.hat/diag(var(rtn,SumSquares=T))
```

下面给出了第 i 个资产收益率的 β_i, σ_i^2 和 R^2 的估计.

```
> t(rbind(beta.hat,sqrt(D.hat),r.square))
      beta.hat  sigma(i)  r.square
 AA      1.292     7.694     0.347
AGE      1.514     7.808     0.415
CAT      0.941     7.725     0.219
  F      1.219     8.241     0.292
FDX      0.805     8.854     0.135
 GM      1.046     8.130     0.238
HPQ      1.628     9.469     0.358
KMB      0.550     6.070     0.134
MEL      1.123     6.120     0.388
NYT      0.771     6.590     0.205
 PG      0.469     6.459     0.090
TRB      0.718     7.215     0.157
TXN      1.796    11.474     0.316
```

图 9-1 给出了 13 只股票 $\hat{\beta}_i$ 和 R^2 的条形图. 金融股票 AGE 和 MEL 以及高

科技股票 HPQ 和 TXN 似乎有较高的 β 和 R^2. 另一方面, KMB 和 PG 有较低的 β 和 R^2. R^2 的变化范围是从 0.09 到 0.41. 这表明市场收益对每只股票变化的解释少于 50%.

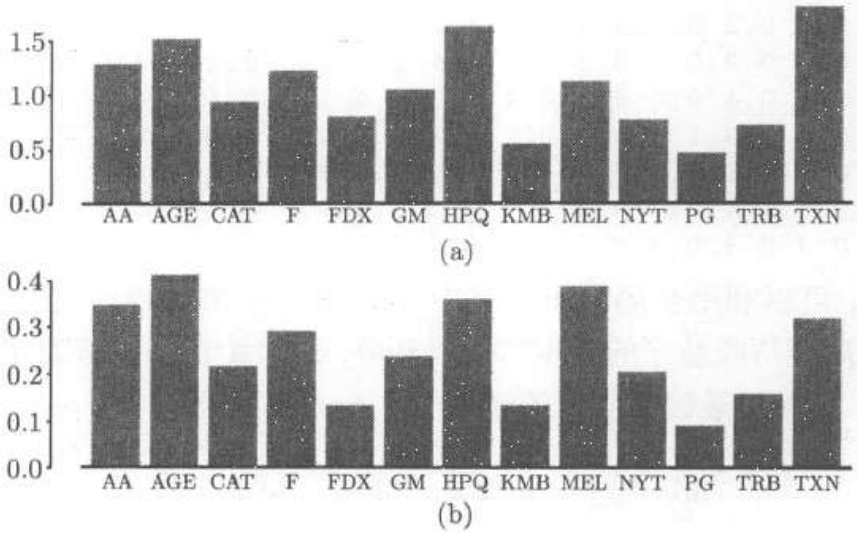

图 9-1 对 13 只股票的月超额收益拟合单因子模型时 β 和 R^2 的条形图: (a) β 的条形图; (b) R^2 的条形图. 标准普尔 500 指数的超额收益率作为市场指数. 样本区间是从 1990 年 1 月到 2003 年 12 月

在市场模型中, $\boldsymbol{r}_t$ 的协方差矩阵和相关矩阵可以如下估计:

```
> cov.r=var(x[,14])*(t(beta.hat)%*%beta.hat)+diag(D.hat)
> sd.r=sqrt(diag(cov.r))
> corr.r=cov.r/outer(sd.r,sd.r)
> print(corr.r,digits=1,width=2)
     AA  AGE CAT  F   FDX  GM HPQ KMB MEL NYT  PG TRB TXN
 AA 1.0 0.4 0.3 0.3 0.2 0.3 0.4 0.2 0.4 0.3 0.2 0.2 0.3
AGE 0.4 1.0 0.3 0.3 0.2 0.3 0.4 0.2 0.4 0.3 0.2 0.3 0.4
CAT 0.3 0.3 1.0 0.3 0.2 0.2 0.3 0.2 0.3 0.2 0.1 0.2 0.3
  F 0.3 0.3 0.3 1.0 0.2 0.3 0.3 0.2 0.3 0.2 0.2 0.2 0.3
FDX 0.2 0.2 0.2 0.2 1.0 0.2 0.2 0.1 0.2 0.2 0.1 0.1 0.2
 GM 0.3 0.3 0.2 0.3 0.2 1.0 0.3 0.2 0.3 0.2 0.1 0.2 0.3
HPQ 0.4 0.4 0.3 0.3 0.2 0.3 1.0 0.2 0.4 0.3 0.2 0.2 0.3
KMB 0.2 0.2 0.2 0.2 0.1 0.2 0.2 1.0 0.2 0.2 0.1 0.1 0.2
MEL 0.4 0.4 0.3 0.3 0.2 0.3 0.4 0.2 1.0 0.3 0.2 0.2 0.3
NYT 0.3 0.3 0.2 0.2 0.2 0.2 0.3 0.2 0.3 1.0 0.1 0.2 0.3
 PG 0.2 0.2 0.1 0.2 0.1 0.1 0.2 0.1 0.2 0.1 1.0 0.1 0.2
TRB 0.2 0.3 0.2 0.2 0.1 0.2 0.2 0.1 0.2 0.2 0.1 1.0 0.2
TXN 0.3 0.4 0.3 0.3 0.2 0.3 0.3 0.2 0.3 0.3 0.2 0.2 1.0
```

我们可以将所估计的超额收益率的协方差矩阵和相关矩阵与其样本协方差矩阵和样本相关矩阵进行比较.

```
> print(cor(rtn),digits=1,width=2)
     AA  AGE CAT  F   FDX  GM HPQ KMB MEL NYT  PG TRB TXN
 AA 1.0 0.3 0.6 0.5 0.2 0.4 0.5 0.3 0.4 0.4 0.1 0.3 0.5
```

```
AGE 0.3 1.0 0.3 0.3 0.3 0.3 0.3 0.3 0.4 0.4 0.2 0.2 0.3
CAT 0.6 0.3 1.0 0.4 0.2 0.3 0.2 0.3 0.4 0.3 0.1 0.4 0.3
  F 0.5 0.3 0.4 1.0 0.3 0.6 0.3 0.3 0.4 0.4 0.1 0.3 0.3
FDX 0.2 0.3 0.2 0.3 1.0 0.2 0.3 0.3 0.2 0.2 0.1 0.3 0.2
 GM 0.4 0.3 0.3 0.6 0.2 1.0 0.3 0.3 0.4 0.2 0.1 0.3 0.3
HPQ 0.5 0.3 0.2 0.3 0.3 0.3 1.0 0.1 0.3 0.3 0.1 0.2 0.6
KMB 0.3 0.3 0.3 0.2 0.3 0.3 0.1 1.0 0.3 0.2 0.3 0.3 0.1
MEL 0.4 0.4 0.4 0.4 0.2 0.4 0.3 0.4 1.0 0.3 0.4 0.3 0.3
NYT 0.4 0.4 0.3 0.4 0.3 0.2 0.3 0.2 0.3 1.0 0.2 0.5 0.2
 PG 0.1 0.2 0.1 0.1 0.1 0.1 0.1 0.3 0.4 0.2 1.0 0.3 0.1
TRB 0.3 0.2 0.4 0.3 0.3 0.3 0.2 0.3 0.3 0.5 0.3 1.0 0.2
TXN 0.5 0.3 0.3 0.3 0.2 0.3 0.6 0.1 0.3 0.2 0.1 0.2 1.0
```

在金融中, 可以利用全局最小方差组合 (GMVP) 来比较给收益率所拟合因子模型的协方差矩阵与收益率的样本协方差矩阵. 对于给定的协方差矩阵 $\boldsymbol{\Sigma}$, 全局最小方差组合 $\boldsymbol{\omega}$ 是下述最优化问题的解:

$$\min_{\boldsymbol{\omega}} \sigma^2_{p,\boldsymbol{\omega}} = \boldsymbol{\omega}'\boldsymbol{\Sigma}\boldsymbol{\omega}, \quad \text{满足} \quad \boldsymbol{\omega}'\mathbf{1} = 1.$$

其中 $\sigma^2_{p,w}$ 是投资组合的方差. 其解如下

$$\boldsymbol{\omega} = \frac{\boldsymbol{\Sigma}^{-1}\mathbf{1}}{\mathbf{1}'\boldsymbol{\Sigma}^{-1}\mathbf{1}},$$

其中 $\mathbf{1}$ 是元素全为 1 的 k 维向量.

对于所考虑的市场模型, 所拟合模型和数据的 GMVP 如下:

```
> w.gmin.model=solve(cov.r)%*%rep(1,nrow(cov.r))
> w.gmin.model=w.gmin.model/sum(w.gmin.model)
> t(w.gmin.model)
       AA      AGE     CAT     F       FDX     GM
[1,] 0.0117 -0.0306 0.0792 0.0225 0.0802 0.0533
        HPQ     KMB     MEL     NYT     PG      TRB      TXN
[1,] -0.0354 0.2503 0.0703 0.1539 0.2434 0.1400 -0.0388
> w.gmin.data=solve(var(rtn))%*%rep(1,nrow(cov.r))
> w.gmin.data=w.gmin.data/sum(w.gmin.data)
> t(w.gmin.data)
        AA       AGE     CAT      F       FDX     GM
[1,] -0.0073 -0.0085 0.0866 -0.0232 0.0943 0.0916
       HPQ     KMB     MEL     NYT     PG      TRB      TXN
[1,] 0.0345 0.2296 0.0495 0.1790 0.2651 0.0168  -0.0080
```

比较两个 GMVP, 给予 TRB 股票的权重变化很大. 然而, 这两个组合都给予 KMB, NYT 和 PG 股票较大的权重.

最后我们检查残差的协方差矩阵和相关矩阵以验证 13 只股票的个性因子不相关的假定. 下面给出了残差相关矩阵的前四列, 且在残差的交叉–相关矩阵中有取较大值的元素, 例如 Cor(CAT,AA) = 0.45 和 Cor(GM,F) = 0.48.

```
> resi.cov=t(E.hat)%*%E.hat/(168-2)
> resi.sd=sqrt(diag(resi.cov))
```

```
> resi.cor=resi.cov/outer(resi.sd,resi.sd)
> print(resi.cor,digits=1,width=2)
       AA    AGE   CAT    F
 AA  1.00 -0.13  0.45  0.22
AGE -0.13  1.00 -0.03 -0.01
CAT  0.45 -0.03  1.00  0.23
  F  0.22 -0.01  0.23  1.00
FDX  0.00  0.14  0.05  0.07
 GM  0.14 -0.09  0.15  0.48
HPQ  0.24 -0.13 -0.07 -0.00
KMB  0.16  0.06  0.18  0.05
MEL -0.02  0.06  0.09  0.10
NYT  0.13  0.10  0.07  0.19
 PG -0.15 -0.02 -0.01 -0.07
TRB  0.12 -0.02  0.25  0.16
TXN  0.19 -0.17  0.09 -0.02
```

9.2.2 多因子模型

Chen, Roll 和 Ross(1986) 考虑了股票收益率的多因子模型. 所用的因子包括宏观经济变量的不可预知的变化或意外. 这里不可预知的变化表示移除宏观经济变量动态依赖后所得到的残差. 得到不可预知的变化的一个简单方法是为宏观经济变量拟合一个第 8 章中的 VAR 模型. 为了进一步说明, 考虑下列两个月宏观经济变量.

(1) 城市居民的消费价格指数 (CPI): 包括所有项的指数, 且指数 $1982-1984=100$.

(2) 16 年及以上城市就业人数 (CE16): 以千记.

CPI 和 CE16 都已经进行了季节调整. 数据的时间区间是从 1975 年 1 月到 2003 年 12 月. 我们用更长的时间区间来得到变量的意外序列. 对于这两个序列, 我们通过取对数序列的一阶差分构造增长率序列. 增长率序列以百分比的形式给出.

为了得到意外序列, 我们用 BIC 准则来识别 VAR(3) 模型. 这样, 因子模型中所用的这两个宏观经济因子都是对数据拟合 VAR(3) 模型时从 1990 年到 2003 年的残差. 对于超额收益率序列, 我们仍然考虑前面所用到的 13 只股票. 下面给出了分析的细节:

```
> da=read.table('m-cpice16-dp7503.txt'),header=T)
> cpi=da[,1]
> cen=da[,2]
> x1=cbind(cpi,cen)
> y1=data.frame(x1)
> ord.choice=VAR(y1,max.ar=13)
> ord.choice$info
      ar(1)   ar(2)   ar(3)   ar(4)   ar(5)   ar(6)
BIC  36.992  38.093  28.234  46.241  60.677  75.810

      ar(7)  ar(8)   ar(9) ar(10)  ar(11)  ar(12)  ar(13)
BIC  86.23  99.294  111.27 125.46  138.01  146.71  166.92
```

```
> var3.fit=VAR(x1~ar(3))
> res=var3.fit$residuals[166:333,1:2]
> da=matrix(scan(file='m-fac9003.txt'),14)
> xmtx = cbind(rep(1,168),res)
> da=t(da)
> rtn=da[,1:13]
> xit.hat=solve(xmtx,rtn)
> beta.hat=t(xit.hat[2:3,])
> E.hat=rtn - xmtx%*%xit.hat
> D.hat=diag(crossprod(E.hat)/(168-3))
> r.square=1-(168-3)*D.hat/diag(var(rtn,SumSquares=T))
```

图 9-2 给出了 13 只股票的 β 和 R^2 估计的条形图. 有趣的是, 所有的超额收益率与 CPI 增长率的不可预知的变化都是负相关的. 这看起来是合理的. 然而, 所有超额收益率的 R^2 都很低, 这说明这两个宏观经济变量对这 13 只股票超额收益率的解释能力很低.

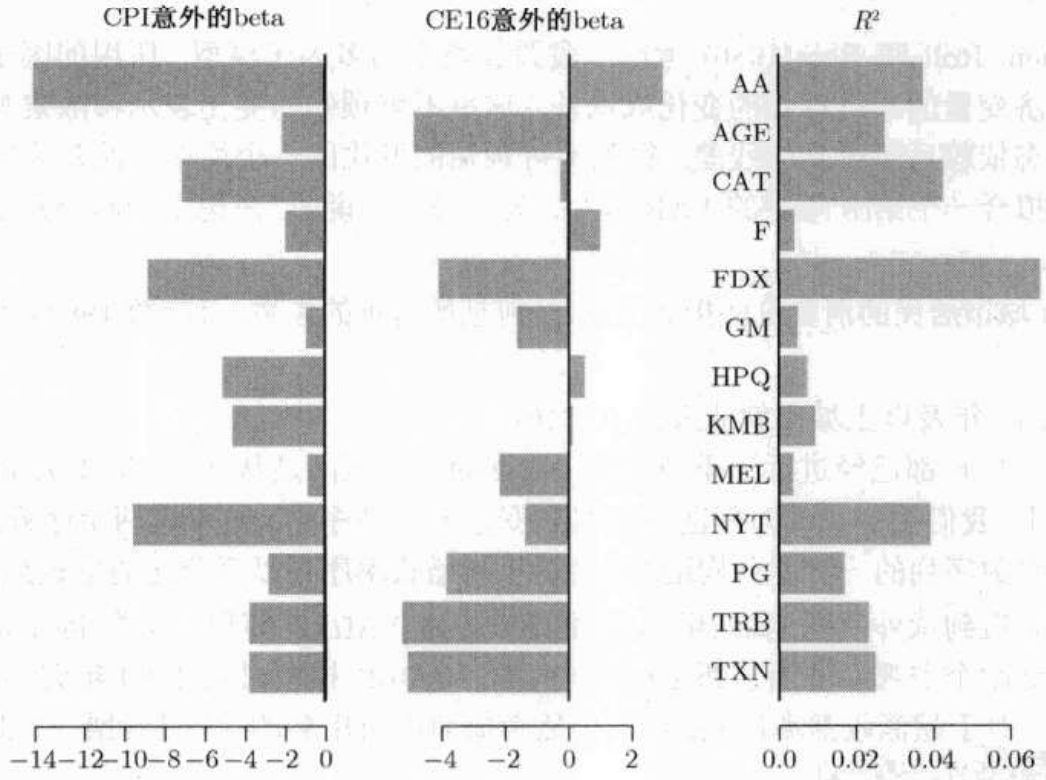

图 9-2 对 13 只股票的月超额收益率拟合二因子模型时 β 和 R^2 的条形图. 样本区间是从 1990 年 1 月到 2003 年 12 月

用下面的命令可以得到该二因子模型的协方差矩阵和相关矩阵的估计:

```
> cov.rtn=beta.hat%*%var(res)%*%t(beta.hat)+diag(D.hat)
> sd.rtn=sqrt(diag(cov.rtn))
> cor.rtn = cov.rtn/outer(sd.rtn,sd.rtn)
> print(cor.rtn,diits=1,width=2)
```

相关矩阵非常接近于单位矩阵, 表明所用的二因子模型并不能很好地拟合这些超额收益率. 最后, 下面给出二因子模型残差的相关矩阵.

```
> cov.resi=t(E.hat)%*%E.hat/(168-3)
> sd.resi=sqrt(diag(cov.resi))
> cor.resi=cov.resi/outer(sd.resi,sd.resi)
> print(cor.resi,digits=1,width=2)
```

如所料想, 该相关矩阵非常接近于前面由原始超额收益率序列给出的相关矩阵, 故此处略去.

9.3 基本面因子模型

基本面因子模型用资产的可观测的具体属性构建公共因子来解释超额收益率. 这些具体属性包括产业分类、企业规模、市场资本化、账面价值以及风格分类 (增长率或值). 对于基础因子模型, 文献中有两种方法. 第一种方法是由 BARRA 公司的创立者 Bar Rosenberg 提出的, 称为 BARRA 方法. 参见 Grinold 和 Kahn(2000). 与宏观经济因子模型相反, 该方法将观测到的资产的具体基本面作为因子 $\boldsymbol{\beta}_i$, 在每个时刻 t 通过回归的方法估计因子 $\boldsymbol{f}_t$. beta 不随时间改变, 但是 $\boldsymbol{f}_t$ 随时间演变. 第二种方法是由 Fama 和 French(1992) 提出的 Fama—French 方法. 在该方法中, 对于给定的具体基本面, 通过基于该具体基本面构造对冲组合来得到因子 f_{jt}. 下面两小节中, 我们简要讨论一下这两种方法.

9.3.1 BARRA 因子模型

假定超额收益率是均值修正的, 从而因子实现也是均值修正的, (9.2) 式的因子模型退化为

$$\tilde{\boldsymbol{r}}_t = \boldsymbol{\beta}\boldsymbol{f}_t + \boldsymbol{\varepsilon}_t, \tag{9.6}$$

其中 $\tilde{\boldsymbol{r}}_t$ 表示 (样本) 均值修正后的超额收益率序列, 为了符号上的简化, 这里继续用 $\boldsymbol{f}_t$ 作为因子实现. 由于 $\boldsymbol{\beta}$ 是给定的, (9.6) 式是有 k 个观测和 m 个未知量的多元线性回归. 由于公共因子的数目 m 应该小于资产的数目 k, 从而回归是可以估计的. 然而, 回归不是齐次的, 因为 $\boldsymbol{\varepsilon}_t$ 的协方差矩阵 $\boldsymbol{D} = \mathrm{diag}\{\sigma_1^2, \cdots, \sigma_k^2\}$ 依赖于第 i 个资产, 这里 $\sigma_i^2 = \mathrm{Var}(\varepsilon_{it})$. 因此, 时刻 t 的因子可以通过加权最小二乘 (WLS) 方法估计, 且权重为个性因子的标准误差. 这样得到的估计为

$$\hat{\boldsymbol{f}}_t = (\boldsymbol{\beta}'\boldsymbol{D}^{-1}\boldsymbol{\beta})^{-1}(\boldsymbol{\beta}'\boldsymbol{D}^{-1}\tilde{\boldsymbol{r}}_t). \tag{9.7}$$

在实际中协方差矩阵 $\boldsymbol{D}$ 是未知的, 从而估计时需要两个步骤.

第一步, 在每个时刻 t 利用普通最小二乘 (OLS) 方法得到 $\boldsymbol{f}_t$ 的一个初步估计如下

$$\hat{\boldsymbol{f}}_{t,o} = (\boldsymbol{\beta}'\boldsymbol{\beta})^{-1}(\boldsymbol{\beta}'\tilde{\boldsymbol{r}}_t),$$

其中第二个下标 o 表示 OLS 估计. 该因子实现的估计是相合的但不是有效的. OLS 回归的残差是

$$\boldsymbol{\varepsilon}_{t,o} = \tilde{\boldsymbol{r}}_t - \boldsymbol{\beta}\hat{\boldsymbol{f}}_{t,o}.$$

由于残差的协方差矩阵不随时间变化, 从而我们可以将所有的残差放在一起 (即对于 $t=1,\cdots,T$) 来得到 $\boldsymbol{D}$ 的估计

$$\hat{\boldsymbol{D}}_o = \operatorname{diag}\left\{\frac{1}{T-1}\sum_{t=1}^{T}(\boldsymbol{\varepsilon}_{t,o}\boldsymbol{\varepsilon}_{t,o}')\right\}.$$

第二步, 我们将估计 $\hat{\boldsymbol{D}}_o$ 嵌入以得到因子实现的修正估计

$$\hat{\boldsymbol{f}}_{t,g} = (\boldsymbol{\beta}'\hat{\boldsymbol{D}}_o^{-1}\boldsymbol{\beta})^{-1}(\boldsymbol{\beta}'\hat{\boldsymbol{D}}_o^{-1}\tilde{\boldsymbol{r}}_t), \tag{9.8}$$

其中第二个下标 g 表示广义最小二乘 (GLS) 估计, 是 WLS 估计的一个样本版本. 修正后回归的残差为

$$\boldsymbol{\varepsilon}_{t,g} = \tilde{\boldsymbol{r}}_t - \boldsymbol{\beta}\hat{\boldsymbol{f}}_{t,g},$$

由此我们可以估计残差的协方差矩阵

$$\hat{\boldsymbol{D}}_g = \operatorname{diag}\left\{\frac{1}{T-1}\sum_{t=1}^{T}(\boldsymbol{\varepsilon}_{t,g}\boldsymbol{\varepsilon}_{t,g}')\right\}.$$

最后被估因子实现的协方差矩阵为

$$\hat{\boldsymbol{\Sigma}}_f = \frac{1}{T-1}\sum_{t=1}^{T}(\hat{\boldsymbol{f}}_{t,g}-\bar{\boldsymbol{f}}_g)(\hat{\boldsymbol{f}}_{t,g}-\bar{\boldsymbol{f}}_g)',$$

其中

$$\bar{\boldsymbol{f}}_g = \frac{1}{T}\sum_{t=1}^{T}\hat{\boldsymbol{f}}_{t,g}.$$

由 (9.6) 式在 BARRA 方法下, 超额收益率的协方差矩阵为

$$\operatorname{Cov}(\boldsymbol{r}_t) = \boldsymbol{\beta}\hat{\boldsymbol{\Sigma}}_f\boldsymbol{\beta}' + \hat{\boldsymbol{D}}_g.$$

1. 产业因子模型

为了进一步说明, 考虑 10 只股票的超额收益率, 并用产业分类作为具体的资产基本面. 表 9-2 给出了所用的股票. 它们可以分为 3 个产业类别, 即, 金融服务、计算机和高科技以及其他类别. 样本区间仍然是从 1990 年 1 月到 2003 年 12 月. 在 BARRA 的框架下, 有三个公共因子表示这三个产业类别, 且 beta 是这三个产业类别的指示变量. 即

$$\tilde{r}_{it} = \beta_{i1}f_{1t} + \beta_{i2}f_{2t} + \beta_{i3}f_{3t} + \varepsilon_{it}, \quad i=1,\cdots,10, \tag{9.9}$$

beta 为

$$\beta_{ij} = \begin{cases} 1, & \text{若资产 } i \text{ 属于第 } j \text{ 个产业区} \\ 0, & \text{其他} \end{cases} \tag{9.10}$$

其中 $j=1,2,3$ 分别表示金融, 高科技和其他类别. 例如, IBM 股票收益率的 beta 向量为 $\boldsymbol{\beta}_i=(0,1,0)'$, Alcoa 股票收益的 beta 向量为 $\boldsymbol{\beta}_i=(0,0,1)'$.

表 9-2 产业因子模型分析中用到的股票及其代码 [a]

Tick	Company	$\bar{r}(\sigma_r)$	Tick	Company	$\bar{r}(\sigma_r)$
AGE	A.G.Edwards	1.36(10.2)	IBM	Int. Bus. Machines	1.06(9.47)
C	Citigroup	2.08(9.60)			
MWD	Morgan Stanley	1.87(11.2)	AA	Alcoa	1.09(9.49)
MER	Merrill Lynch	2.08(10.4)	CAT	Caterpillar	1.23(8.71)
DELL	Dell Inc.	4.82(16.4)	PG	Procter&Gamble	1.08(6.75)
HPQ	Hewlett-Packard	1.37(11.8)			

a 超额收益率的样本均值和样本标准差也在表中给出. 样本时间区间是从1990年1月到2003年12月.

(9.9) 式中, f_{1t} 是金融服务类的因子实现, f_{2t} 是计算机和高科技类的因子实现, f_{3t} 是其他类的因子实现. 因为 β_{ij} 是指示变量, 所以 $\boldsymbol{f}_t$ 的 OLS 估计非常简单. 事实上, $\boldsymbol{f}_t$ 是由 t 时刻每个类别的超额收益率的平均值构成的向量. 具体地,

$$\hat{\boldsymbol{f}}_{t,o} = \begin{bmatrix} \dfrac{\mathrm{AGE}_t + \mathrm{C}_t + \mathrm{MDW}_t + \mathrm{MER}_t}{4} \\ \dfrac{\mathrm{DELL}_t + \mathrm{HPQ}_t + \mathrm{IBM}_t}{3} \\ \dfrac{\mathrm{AA}_t + \mathrm{CAT}_t + \mathrm{PG}_t}{3} \end{bmatrix}.$$

第 i 个资产的个性因子仅仅是其超额收益率与其所属产业类样本均值的差. 于是可以得到残差协方差矩阵 $\boldsymbol{D}$ 的估计, 并由此得到广义最小二乘估计. 我们用 S-plus 进行分析. 首先, 将收益率加载到 S-plus 中, 移除掉样本均值, 创建产业类哑元并计算收益率的样本相关矩阵.

```
> da=read.table('m-barra-9003.txt'),header=T)
> rm = matrix(apply(da,2,mean),1)
> rtn = da - matrix(1,168,1)%*%rm
> fin = c(rep(1,4),rep(0,6))
> tech = c(rep(0,4),rep(1,3),rep(0,3)
> oth = c(rep(0,7),rep(1,3))
> ind.dum = cbind(fin,tech,oth)
> ind.dum
      fin tech oth
 [1,]   1    0   0
 [2,]   1    0   0
 [3,]   1    0   0
 [4,]   1    0   0
 [5,]   0    1   0
 [6,]   0    1   0
 [7,]   0    1   0
 [8,]   0    0   1
 [9,]   0    0   1
[10,]   0    0   1
> cov.rtn=var(rtn)
> sd.rtn=sqrt(diag(cov.rtn))
> corr.rtn=cov.rtn/outer(sd.rtn,sd.rtn)
> print(corr.rtn,digits=1,width=2)
```

```
      AGE   C   MWD MER DELL HPQ IBM AA   CAT PG
AGE   1.0 0.6 0.6 0.6 0.3   0.3 0.3 0.3 0.3 0.2
C     0.6 1.0 0.7 0.7 0.2   0.4 0.4 0.4 0.4 0.3
MWD   0.6 0.7 1.0 0.8 0.3   0.5 0.4 0.4 0.3 0.3
MER   0.6 0.7 0.8 1.0 0.2   0.5 0.3 0.4 0.3 0.3
DELL  0.3 0.2 0.3 0.2 1.0   0.5 0.4 0.3 0.1 0.1
HPQ   0.3 0.4 0.5 0.5 0.4   1.0 0.5 0.5 0.2 0.1
IBM   0.3 0.4 0.4 0.3 0.4   0.5 1.0 0.4 0.3-0.0
AA    0.3 0.4 0.4 0.4 0.3   0.5 0.4 1.0 0.6 0.1
CAT   0.3 0.4 0.3 0.3 0.1   0.2 0.3 0.6 1.0 0.1
PG    0.2 0.3 0.3 0.3 0.1   0.1-0.0 0.1 0.1 1.0
```

下面给出了 OLS 估计、残差和残差的方差估计.

```
> F.hat.o = solve(crossprod(ind.dum))%*%t(ind.dum)%*%rtn.rm
>  E.hat.o = rtn.rm - ind.dum%*%F.hat.o
> diagD.hat.o=rowVars(E.hat.o)
```

接下来便可以得到广义最小二乘估计.

```
> Dinv.hat = diag(diagD.hat.o^(-1))
> Hmtx=solve(t(ind.dum)%*%Dinv.hat%*%ind.dum)%*%t(ind.dum)
   %*%Dinv.hat
> F.hat.g = Hmtx%*%rtn.rm
> F.hat.gt=t(F.hat.g)
> E.hat.g = rtn.rm - ind.dum%*%F.hat.g
> diagD.hat.g = rowVars(E.hat.g)
> t(Hmtx)
             fin      tech       oth
 [1,]     0.1870    0.0000    0.0000
 [2,]     0.2548    0.0000    0.0000
 [3,]     0.2586    0.0000    0.0000
 [4,]     0.2995    0.0000    0.0000
 [5,]     0.0000    0.2272    0.0000
 [6,]     0.0000    0.4015    0.0000
 [7,]     0.0000    0.3713    0.0000
 [8,]     0.0000    0.0000    0.3319
 [9,]     0.0000    0.0000    0.4321
[10,]     0.0000    0.0000    0.2360
> cov.ind=ind.dum%*%var(F.hat.gt)%*%t(ind.dum)
   + diag(diagD.hat.g)
> sd.ind=sqrt(diag(cov.ind))
> corr.ind=cov.ind/outer(sd.ind,sd.ind)
> print(corr.ind,digits=1,width=2)
      AGE C   MWD MER DELL HPQ IBM AA   CAT PG
AGE   1.0 0.7 0.7 0.7 0.3   0.3 0.3 0.3 0.3 0.3
C     0.7 1.0 0.8 0.8 0.3   0.4 0.4 0.3 0.3 0.3
MWD   0.7 0.8 1.0 0.8 0.3   0.4 0.4 0.3 0.4 0.3
MER   0.7 0.8 0.8 1.0 0.3   0.4 0.4 0.3 0.4 0.3
DELL  0.3 0.3 0.3 0.3 1.0   0.5 0.5 0.2 0.2 0.2
HPQ   0.3 0.4 0.4 0.4 0.5   1.0 0.7 0.3 0.3 0.2
IBM   0.3 0.4 0.4 0.4 0.5   0.7 1.0 0.3 0.3 0.2
```

```
AA   0.3 0.3 0.3 0.3 0.2  0.3 0.3 1.0 0.7 0.5
CAT  0.3 0.3 0.4 0.4 0.2  0.3 0.3 0.7 1.0 0.6
PG   0.3 0.3 0.3 0.3 0.2  0.2 0.2 0.5 0.6 1.0
```

由模型得到的产业类内的股票的相关矩阵要比样本相关矩阵大. 例如, 股票 CAT 和 PG 的样本相关系数只有 0.1, 但是基于所拟合模型得到的相关系数是 0.6. 图 9-3 给出了基于广义最小二乘估计给出的因子实现的时间图.

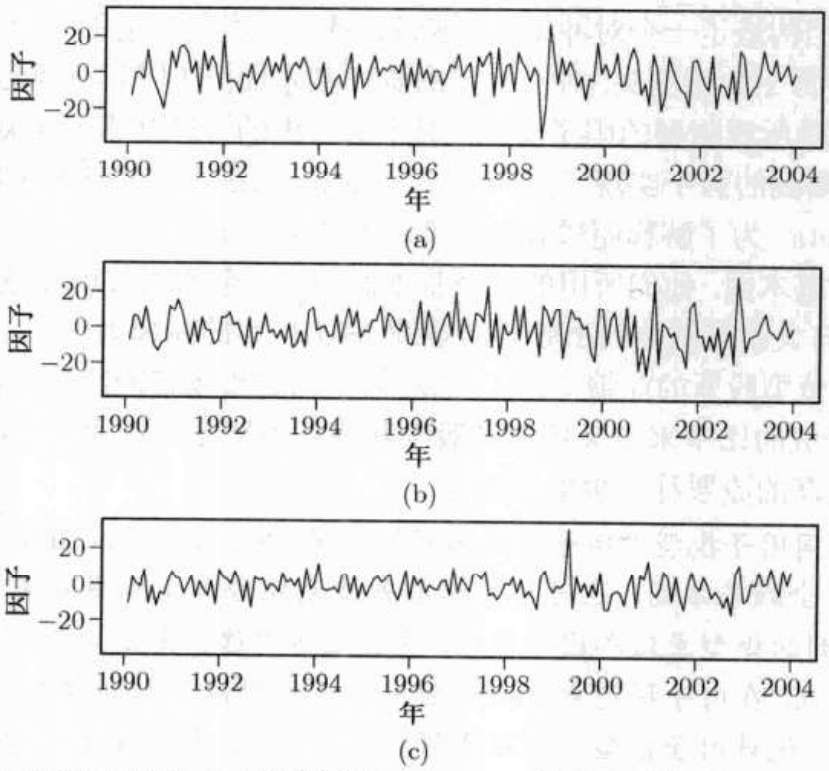

图 9-3 对三个产业类的 10 只股票拟合 BARRA 产业因子模型所估计的因子实现. 其中 (a) 因子实现: 金融类, (b) 高科技类及 (c) 其他类

2. 因子模拟组合

考虑带单因子的 BARRA 因子模型这种特殊情况. 这里 (9.7) 式给出的 f_t 的 WLS 估计提供了很好的解释. 考虑 k 个资产的组合 $\boldsymbol{\omega} = (\omega_1, \cdots, \omega_k)'$. 该组合是下述最小化问题的解:

$$\min_{\boldsymbol{\omega}} \left(\frac{1}{2} \boldsymbol{\omega}' \boldsymbol{D} \boldsymbol{\omega} \right), \quad \text{满足} \quad \boldsymbol{\omega}' \boldsymbol{\beta} = 1.$$

该组合问题的解由下式给出:

$$\boldsymbol{\omega}' = (\boldsymbol{\beta}' \boldsymbol{D}^{-1} \boldsymbol{\beta})^{-1} (\boldsymbol{\beta}' \boldsymbol{D}^{-1}).$$

因此, 被估因子实现是如下组合的收益率:

$$\hat{f}_t = \boldsymbol{\omega}' \boldsymbol{r}_t.$$

如果将组合 $\boldsymbol{\omega}$ 标准化, 即满足 $\sum_{i=1}^{k} \omega_i = 1$, 则称之为因子模拟组合. 对多个因子的情况, 可以对每个因子单独应用该思想.

注释　在实际中, 超额收益率的样本均值经常与 0 没有显著的区别. 因此, 在拟合一个 BARRA 因子模型之前, 通常不需要移除样本均值. □

9.3.2 Fama-French 方法

对一个给定的资产基本面 (例如账面价值与市场价值的比率), Fama 和 French (1992) 使用两个步骤来决定因子实现. 首先, 他们基于观测到的基本面的值将资产分类. 然后他们构造了一个对冲组合. 该组合持有分类资产前面 1/3 的多头, 且持有分类资产后面 2/3 的空头. 对于给定的资产基本面, t 时刻所观测到的该对冲组合的收益率就是所观测到的因子实现. 对于所考虑的资产基本面重复上面的步骤. 最后, 给定观测到的因子实现 $\{\boldsymbol{f}_t|t=1,\cdots,T\}$, 并用时间序列的回归方法来估计每个资产的 beta. 为了解释超额收益率变动性的高百分比, Fama 和 French 确认了三个观测到的基本面. 他们所用的三个基本面是 (a) 全部的市场收益率 (市场超额收益率); (b) 与大股票相关的小股票的业绩 (SMB, 小的减掉大的); (c) 与成长型股票相联系的价值型股票的业绩 (HML, 高对低). 通过市场资产净值和市场资产净值对账面资产净值的比率来定义价值型股票和成长型股票. 账面资产净值对市场资产净值的比率高的股票称为价值型股票.

注释　不同因子模型中因子的概念可能不同. 在 Fama-French 方法中所用的三个因子是三个金融基本面. 也可以将这些基本面组合起来构成股票的一个新的属性, 并将所得到的模型看做单因子模型. 这里之所以这样是因为所用的模型是线性统计模型. 因此, 在因子模型中当提到因子的个数时应该特别注意. 另一方面, 对于因子的个数, 统计因子模型中有相当好的定义. 下面我们将对此进行讨论. □

9.4　主成分分析

在多元时间序列分析中, 一个重要的问题是对序列的协方差 (或相关系数) 结构的研究. 例如, 向量收益率序列的协方差结构在组合选择中起着很重要的作用. 下面, 我们讨论一些统计方法. 它们在研究时间序列的协方差结构时非常有用.

给定一个 k 维随机向量 $\boldsymbol{r}=(r_1,\cdots,r_k)'$, 其协方差矩阵为 $\boldsymbol{\Sigma}_r$, 则主成分分析 (principal component analysis, 简记为 PCA) 关心的是利用 r_i 很少的线性组合来解释 $\boldsymbol{\Sigma}_r$ 的结构. 如果 $\boldsymbol{r}$ 表示 k 个资产的月对数收益率, 则可用 PCA 来研究这 k 个资产收益率变化的原因. 这里关键词是很少, 从而使得多元分析可以获得简化.

9.4.1 PCA 理论

PCA 对 $\boldsymbol{r}$ 的协方差矩阵 $\boldsymbol{\Sigma}_r$ 或相关矩阵 $\boldsymbol{\rho}_r$ 都适用. 因为相关矩阵是标准化随机变量 $\boldsymbol{r}^*=\boldsymbol{S}^{-1}\boldsymbol{r}$ 的协方差矩阵, 此处 $\boldsymbol{S}$ 是 $\boldsymbol{r}$ 的分量的标准差组成的对角矩阵, 所以在我们的理论分析中使用协方差矩阵. 令 $\boldsymbol{\omega}_i=(\omega_{i1},\cdots,\omega_{ik})'$ 表示 k 维向量,

这里 $i=1,\cdots,k$. 那么

$$y_i=\boldsymbol{\omega}_i'\boldsymbol{r}=\sum_{j=1}^{k}\omega_{ij}r_j$$

是随机向量 $\boldsymbol{r}$ 的线性组合. 若 $\boldsymbol{r}$ 由 k 只股票的简单收益率组成, 则 y_i 是对第 j 只股票赋予权重 ω_{ij} 之后所形成的组合的收益率. 因为将 $\boldsymbol{\omega}_i$ 乘上一个常数并不影响分配到第 j 支股票上的权重, 所以我们将向量 $\boldsymbol{\omega}_i$ 标准化, 使得 $\boldsymbol{\omega}_i'\boldsymbol{\omega}_i=\sum\limits_{j=1}^{k}\omega_{ij}^2=1$. 利用随机变量线性组合的性质, 我们有

$$\mathrm{Var}(y_i)=\boldsymbol{\omega}_i'\boldsymbol{\Sigma}_r\boldsymbol{\omega}_i,\quad i=1,\cdots,k, \tag{9.11}$$

$$\mathrm{Cov}(y_i,y_j)=\boldsymbol{\omega}_i'\boldsymbol{\Sigma}_r\boldsymbol{\omega}_j,\quad i,j=1,\cdots,k. \tag{9.12}$$

PCA 的思想就是找到线性组合 $\boldsymbol{\omega}_i$ 使得对 $i\neq j$ 有 y_i 与 y_j 是不相关的, 并且 y_i 的方差尽可能大. 更具体地:

(1) $\boldsymbol{r}$ 的第一个主成分是在 $\boldsymbol{\omega}_1'\boldsymbol{\omega}_1=1$ 的限制下, 使得 $\mathrm{Var}(y_1)$ 最大的线性组合 $y_1=\boldsymbol{\omega}_1'\boldsymbol{r}$;

(2) $\boldsymbol{r}$ 的第二个主成分是在 $\boldsymbol{\omega}_2'\boldsymbol{\omega}_2=1$ 与 $\mathrm{Cov}(y_2,y_1)=0$ 的限制下, 使得 $\mathrm{Var}(y_2)$ 最大的线性组合 $y_2=\boldsymbol{\omega}_2'\boldsymbol{r}$;

(3) $\boldsymbol{r}$ 的第 i 个主成分是在 $\boldsymbol{\omega}_i'\boldsymbol{\omega}_i=1$ 与 $\mathrm{Cov}(y_i,y_j)=0,\ j=1,\cdots,i-1$ 的限制下, 最大化 $\mathrm{Var}(y_i)$ 的线性组合 $y_i=\boldsymbol{\omega}_i'\boldsymbol{r}$.

因为 $\boldsymbol{\Sigma}_r$ 的协方差矩阵是非负定的, 所以它具有谱分解 (见第 8 章附录 A). 令 $(\lambda_1,\boldsymbol{e}_1)\cdots(\lambda_k,\boldsymbol{e}_k)$ 为 $\boldsymbol{\Sigma}_r$ 的特征值 (特征向量组), 其中 $\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_k\geqslant 0$, 则我们有下面的统计结果.

结果 9.1 $\boldsymbol{r}$ 的第 i 个主成分是 $y_i=\boldsymbol{e}_i'\boldsymbol{r}=\sum\limits_{j=1}^{k}e_{ij}r_j, i=1,\cdots,k$. 而且

$$\mathrm{Var}(y_i)=\boldsymbol{e}_i'\boldsymbol{\Sigma}_r\boldsymbol{e}_i=\lambda_i,\quad i=1,\cdots,k,$$

$$\mathrm{Cov}(y_i,y_j)=\boldsymbol{e}_i'\boldsymbol{\Sigma}_r\boldsymbol{e}_j=0,\quad i\neq j.$$

如果某些特征值 λ_i 是相等的, 则对应特征向量 $\boldsymbol{e}_i$ 的选择不是唯一的, 从而 y_i 也不是唯一的. 另外, 我们有

$$\sum_{i=1}^{k}\mathrm{Var}(r_i)=\mathrm{tr}(\boldsymbol{\Sigma}_r)=\sum_{i=1}^{k}\lambda_i=\sum_{i=1}^{k}\mathrm{Var}(y_i). \tag{9.13}$$

等式 (9.13) 说明

$$\frac{\mathrm{Var}(y_i)}{\sum\limits_{i=1}^{k}\mathrm{Var}(r_i)}=\frac{\lambda_i}{\lambda_1+\cdots+\lambda_k}.$$

因此, $\boldsymbol{r}$ 的总方差中由第 i 个主成分解释的比例是 $\boldsymbol{\Sigma}_r$ 的第 i 个特征值占 $\boldsymbol{\Sigma}_r$ 的所有特征值总和的比率. 也可以计算由前 i 个主成分所能解释的总方差的累积比例 (即 $\sum_{j=1}^{i}\lambda_j)/(\sum_{j=1}^{k}\lambda_j)$, 实际中, 可以选择一个较小的 i 使得前面的累积比例很大.

因为 $\mathrm{tr}(\boldsymbol{\rho}_r)=k$, 所以当采用相关阵来进行主成分分析时, 由第 i 个主成分解释的方差比例变为 λ_i/k.

PCA 的一个副产品是 $\boldsymbol{\Sigma}_r$ 或 $\boldsymbol{\rho}_r$ 的 0 特征值表明 $\boldsymbol{r}$ 的分量之间存在精确的线性关系. 例如, 如果最小特征值 $\lambda_k=0$, 则由结果 9.1 知 $\mathrm{Var}(y_k)=0$. 因此, $y_k=\sum_{j=1}^{k}e_{kj}r_j$ 是个常数, 从而在 $\boldsymbol{r}$ 中只有 $k-1$ 个随机量. 在这种情形下, $\boldsymbol{r}$ 的维数可以降低. 由于这个原因, 文献中常用 PCA 作为降低维数的工具.

9.4.2　经验的 PCA

应用中, 收益率向量 $\boldsymbol{r}$ 的协方差矩阵 $\boldsymbol{\Sigma}_r$ 和相关矩阵 $\boldsymbol{\rho}_r$ 是未知的, 但在一些正则性条件下, 它们可以通过样本协方差矩阵和样本相关矩阵相合地估计. 假定收益率是弱平稳的, 且数据为 $\{r_t|t=1,\cdots,T\}$, 则我们有如下估计:

$$\hat{\boldsymbol{\Sigma}}_r\equiv[\hat{\sigma}_{ij,r}]=\frac{1}{T-1}\sum_{t=1}^{T}(\boldsymbol{r}_t-\bar{\boldsymbol{r}})(\boldsymbol{r}_t-\bar{\boldsymbol{r}})',\quad \bar{\boldsymbol{r}}=\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{r}_t, \tag{9.14}$$

$$\hat{\boldsymbol{\rho}}_r=\hat{\boldsymbol{S}}^{-1}\hat{\boldsymbol{\Sigma}}_r\hat{\boldsymbol{S}}^{-1}, \tag{9.15}$$

其中 $\hat{\boldsymbol{S}}=\mathrm{diag}\{\sqrt{\hat{\sigma}_{11,r}},\cdots,\sqrt{\hat{\sigma}_{kk,r}}\}$ 是由 $\boldsymbol{r}_t$ 的样本标准差构成的对角矩阵. 从而可以利用计算对称矩阵的特征值和特征向量的方法来进行主成分分析. 现在大多数统计包都能进行主成分分析. 在 S-Plus 中, 进行主成分分析的基本命令是`princomp`, 在 FinMetrics 中则为`mfactor`.

例 9.1　考虑 IBM、Hewlett-Packard、Intel Corporation、Merrill Lynch 与 Morgan Stanley Dean Witter 从 1990 年 1 月至 2008 年 12 月的月对数收益率. 此收益率以百分比表示, 且包括红利. 数据集共有 228 个观测值. 图 9-4 给出了这 5 种月收益率序列的时间图. 如所料想, 同一工业部门的公司收益率倾向于展现出相似的模式.

用 $\boldsymbol{r}'$=(IBM, HPQ, INTC, JPM, BAC) 表示这些收益率, 其样本均值向量为 $(0.70, 0.99, 1.20, 0.82, 0.41)'$, 样本协方差矩阵和样本相关矩阵为

$$\hat{\boldsymbol{\Sigma}}_r=\begin{bmatrix}74.64 & & & & \\ 42.28 & 112.22 & & & \\ 48.03 & 70.45 & 146.50 & & \\ 30.10 & 42.42 & 44.59 & 106.04 & \\ 21.07 & 26.30 & 29.24 & 67.45 & 91.83\end{bmatrix},$$

$$\hat{\boldsymbol{\rho}}_r = \begin{bmatrix} 1.00 & & & & \\ 0.46 & 1.00 & & & \\ 0.46 & 0.55 & 1.00 & & \\ 0.34 & 0.39 & 0.36 & 1.00 & \\ 0.25 & 0.26 & 0.25 & 0.68 & 1.00 \end{bmatrix}.$$

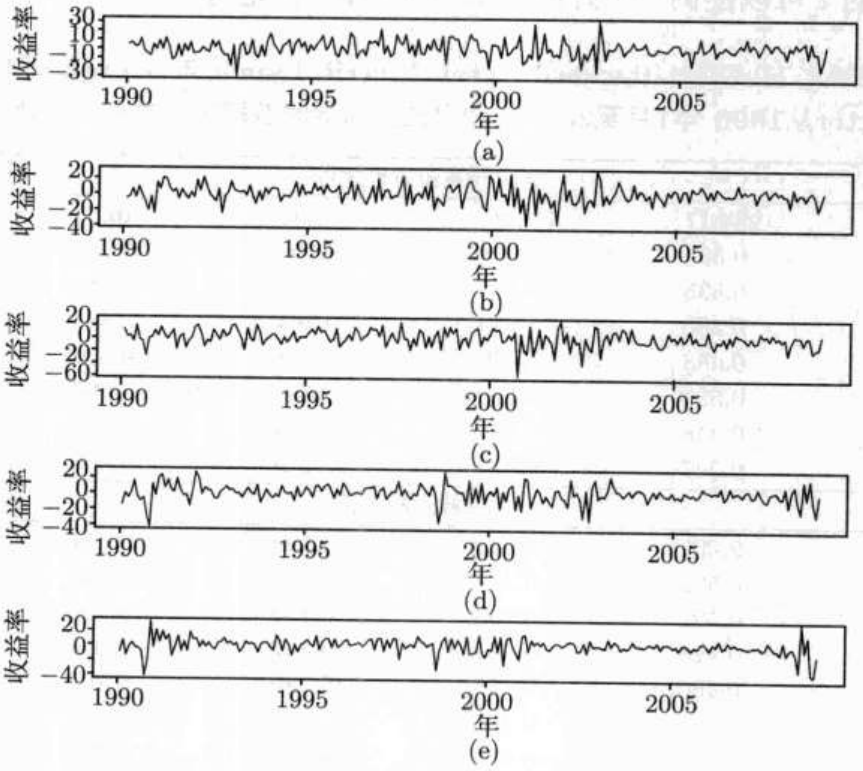

图 9-4 以下 5 个公司月对数收益率的时间图, 其中对数收益率以百分比表示且包含分红, 时间区间是从 1990 年 1 月至 2008 年 12 月: (a) IBM; (b) Hewlett-Packard; (c) Intel; (d) J.P. Morgan Chase; (e) Bank of America

表 9-3 给出了利用协方差矩阵和相关矩阵进行主成分分析的结果. 还给出了特征值、特征向量以及由主成分解释的变化比. 考虑相关矩阵, 并用 $\hat{\lambda}_i$ 和 $\hat{e}_i$ 来表示样本特征值与特征向量. 对前两个主成分, 我们有

$$\hat{\lambda}_1 = 2.608, \quad \hat{e}_1 = (0.428, 0.460, 0.451, 0.479, 0.416)',$$
$$\hat{\lambda}_2 = 1.072, \quad \hat{e}_2 = (0.341, 0.356, 0.385, -0.469, -0.623)'$$

这两个成分大约解释了数据全部变化的 72%, 且它们具有有趣的解释. 第一个成分是股票收益率的一个大致为等权重的线性组合. 这个成分可能代表股票市场的一般运动, 从而是一个市场成分. 第二个成分代表两个工业部门 (即技术和金融服务) 的差. 它可能是一个工业成分. 利用 $\boldsymbol{r}$ 的协方差矩阵也可以发现主成分的类似解释.

应用中确定主成分个数的一个非正式但是很有用的方法是检查斜坡图 (scree plot). 它是特征值 $\hat{\lambda}_i$ 按由大到小次序排列之后的时间图 (即 $\hat{\lambda}_i$ 对 i 的图). 图 9-5a

给出了例 9.1 中的 5 种股票收益率的斜坡图. 通过在斜坡图中寻找转弯处, 这表明余下的特征值相对较小并大致看上去是相同的, 所以我们可以选择一个恰当的主成分的个数. 对图 9-5 的两个图来说, 两个主成分看起来是合适的. 最后, 除了对 $j > i$ 有 $\lambda_j = 0$ 的情形, 选择前 i 个主成分仅仅提供了数据总方差的一个近似. 如果一个很小的 i 可以提供一个好的近似, 则这种简化是有价值的.

表 9-3　对 IBM, Hewlett-Packard, Intel, Merrill Lynch 与 Morgan Stanley Dean Witter从1990 年1月至2008年12月的月对数收益率进行主成分分析的结果 [a]

			利用样本协方差矩阵		
特征值	284.17	112.93	57.43	46.81	29.87
比例	0.535	0.213	0.108	0.088	0.056
累积	0.535	0.748	0.856	0.944	1.000
特征向量	0.330	0.139	−0.264	0.895	−0.014
	0.483	0.279	−0.701	−0.430	−0.116
	0.581	0.478	0.652	−0.096	−0.016
	0.448	−0.550	0.013	−0.064	0.702
	0.347	−0.610	0.119	−0.009	−0.702
			利用样本相关阵		
特征值	2.607	1.072	0.569	0.451	0.301
比例	0.522	0.214	0.114	0.090	0.060
累积	0.522	0.736	0.850	0.940	1.000
特征向量	0.428	0.341	0.837	−0.002	0.008
	0.460	0.356	−0.380	0.704	0.145
	0.451	0.385	−0.389	−0.704	0.022
	0.479	−0.469	−0.046	0.052	−0.739
	0.416	−0.623	0.035	−0.073	0.658

a 特征向量是以列向量形式给出的.

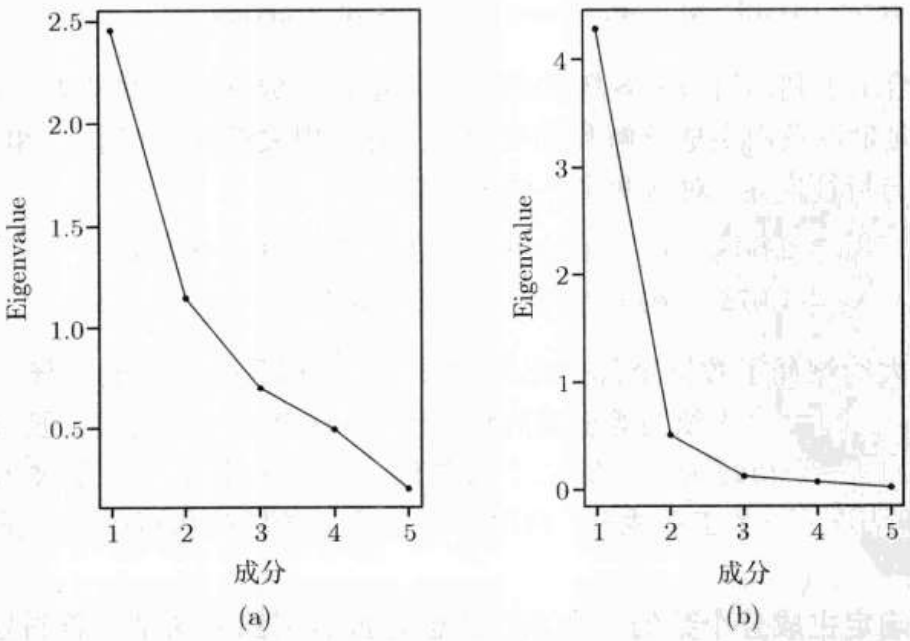

图 9-5　两个 5 维资产收益率的斜坡图: (a) 例 9.1 中的序列; (b) 例 9.3 中的债券指数收益率

注释 下面给出使用 R 和 S-Plus 进行 PCA 的命令. 命令 princomp 求出了特征值的平方根, 将其作为标准差.

```
> rtn=read.table(''m-5clog-9008.txt''),header=T)
> pca.cov = princomp(rtn)
> names(pca.cov)
> summary(pca.cov)
> pca.cov$loadings
> screeplot(pca.cov)
> pca.corr=princomp(rtn,cor=T)
> summary(pac.corr)
```

□

9.5 统计因子分析

我们现在转向统计因子分析. 多元统计分析中的一个主要困难是 "维数的祸害". 特别地, 当模型的阶或时间序列的维数增加时, 参数模型的参数数量也经常陡增. 通常要寻找简化方法来克服维数所带来的祸害. 从实证的观点出发, 多元数据经常表现出一些相似的模式, 这表明数据中存在潜藏的共同结构. 统计因子分析是文献中可以利用的简化方法之一. 它的目的是识别几个因子, 使得它们能够解释数据的协方差矩阵与相关矩阵中的绝大部分变化.

传统的统计因子分析假定数据没有序列相关性. 由于金融数据经常是以小于或等于一周的频率抽取的, 所以经常违反这个假定. 然而, 这个假定对低频数据的资产收益率 (如股票或市场指数的月收益率) 看上去是合理的. 如果违背了这个假定, 则可以利用本书中讨论的参数模型消除数据的线性动态依赖, 并对残差序列应用因子分析.

下面, 我们讨论基于正交因子模型 (orthogonal factor model) 的因子分析. 考虑第 t 期 k 个资产的收益率 $\boldsymbol{r}_t = (r_{1t}, \cdots, r_{kt})'$, 并假定 $\boldsymbol{r}_t$ 是弱平稳的, 其均值为 $\boldsymbol{\mu}$, 协方差矩阵为 $\boldsymbol{\Sigma}_r$. 因子模型假定 $\boldsymbol{r}_t$ 线性地依赖于少数不可观测的随机变量 $\boldsymbol{f}_t = (f_{1t}, \cdots, f_{mt})'$ 与 k 维附加噪声 $\boldsymbol{\varepsilon}_t = (\varepsilon_{1t}, \cdots, \varepsilon_{kt})'$. 这里 $m < k$, f_{it} 是公共因子, ε_{it} 是误差. 数学上, 统计因子模型具有 (9.1) 式的形式, 只是用均值收益 $\boldsymbol{\mu}$ 代替截距 $\boldsymbol{\alpha}$. 因此因子模型有如下形式

$$\boldsymbol{r}_t - \boldsymbol{\mu} = \boldsymbol{\beta}\boldsymbol{f}_t + \boldsymbol{\varepsilon}_t, \tag{9.16}$$

其中 $\boldsymbol{\beta} = [\beta_{ij}]_{k\times m}$ 是因子负荷矩阵, β_{ij} 是第 i 个变量在 j 个因子上的负荷, ε_{it} 是 r_{it} 的个性误差. 统计因子模型的一个关键特征是 m 个因子 f_{it} 和因子负荷 β_{ij} 都是不可观测的. 正因为如此, (9.16) 式不是一个多元线性回归模型, 尽管它看上去与多元线性回归模型相似.

称 (9.16) 式的因子模型是一个正交因子模型, 如果它满足下面的假定:

(1) $\mathrm{E}(\boldsymbol{f}_t) = \boldsymbol{0}$, $\mathrm{Cov}(\boldsymbol{f}_t) = \boldsymbol{I}_m$ 为 $m \times m$ 的单位矩阵;

(2) $\mathrm{E}(\boldsymbol{\varepsilon}_t)=\mathbf{0}$, $\mathrm{Cov}(\boldsymbol{\varepsilon}_t)=\boldsymbol{D}=\mathrm{diag}\{\sigma_1^2,\cdots,\sigma_k^2\}$(即 $\boldsymbol{D}$ 是 $k\times k$ 对角矩阵);

(3) $\boldsymbol{f}_t$ 与 $\boldsymbol{\varepsilon}_t$ 是独立的, 从而 $\mathrm{Cov}(f_t,\varepsilon_t)=\mathrm{E}(\boldsymbol{f}_t\boldsymbol{\varepsilon}_t')=\mathbf{0}_{m\times k}$.

在上述假定下, 很容易看出

$$\begin{aligned}\boldsymbol{\Sigma}_r&=\mathrm{Cov}(\boldsymbol{r}_t)=\mathrm{E}[(\boldsymbol{r}_t-\boldsymbol{\mu})(\boldsymbol{r}_t-\boldsymbol{\mu})']\\&=\mathrm{E}[(\boldsymbol{\beta}\boldsymbol{f}_t+\boldsymbol{\varepsilon}_t)(\boldsymbol{\beta}\boldsymbol{f}_t+\boldsymbol{\varepsilon}_t)']\\&=\boldsymbol{\beta}\boldsymbol{\beta}'+\boldsymbol{D}\end{aligned}\tag{9.17}$$

且

$$\mathrm{Cov}(\boldsymbol{r}_t,\boldsymbol{f}_t)=\mathrm{E}[(\boldsymbol{r}_t-\boldsymbol{\mu})\boldsymbol{f}_t']=\boldsymbol{\beta}\mathrm{E}(\boldsymbol{f}_t\boldsymbol{f}_t')+\mathrm{E}(\boldsymbol{\varepsilon}_t\boldsymbol{f}_t')=\boldsymbol{\beta}.\tag{9.18}$$

利用 (9.17) 式和 (9.18) 式, 我们看出, 对 (9.16) 式中的正交因子模型

$$\begin{aligned}\mathrm{Var}(r_{it})&=\beta_{i1}^2+\cdots+\beta_{im}^2+\sigma_i^2,\\\mathrm{Cov}(r_{it},r_{jt})&=\beta_{i1}\beta_{j1}+\cdots+\beta_{im}\beta_{jm},\\\mathrm{Cov}(r_{it},f_{jt})&=\beta_{ij}.\end{aligned}$$

由 m 个公共因子贡献的 r_{it} 的方差部分 $\beta_{i1}^2+\cdots+\beta_{im}^2$ 称为共性方差 (Communality). r_{it} 方差的剩余部分 σ_i^2 称为唯一性方差或个性方差. 令 $c_i^2=\beta_{i1}^2+\cdots+\beta_{im}^2$ 为共性方差, 它是第 i 个变量对 m 个公共因子的负荷的平方和. 分量 r_{it} 的方差变为 $\mathrm{Var}(r_{it})=c_i^2+\sigma_i^2$.

实际中, 并非每个协方差矩阵都具有正交因子表示. 换句话说, 一个不具有任何正交因子表示的随机变量 $\boldsymbol{r}_t$ 是存在的. 而且, 随机变量的正交因子表示并不唯一. 事实上, 对任何满足 $\boldsymbol{P}\boldsymbol{P}'=\boldsymbol{P}'\boldsymbol{P}=\boldsymbol{I}$ 的 $m\times m$ 正交矩阵 $\boldsymbol{P}$, 令 $\boldsymbol{\beta}^*=\boldsymbol{\beta}\boldsymbol{P}$, $\boldsymbol{f}_t^*=\boldsymbol{P}'\boldsymbol{f}_t$, 则

$$\boldsymbol{r}_t-\boldsymbol{\mu}=\boldsymbol{\beta}\boldsymbol{f}_t+\boldsymbol{\varepsilon}_t=\boldsymbol{\beta}\boldsymbol{P}\boldsymbol{P}'\boldsymbol{f}_t+\boldsymbol{\varepsilon}_t=\boldsymbol{\beta}^*\boldsymbol{f}_t^*+\boldsymbol{\varepsilon}_t.$$

另外 $\mathrm{E}(\boldsymbol{f}_t^*)=\mathbf{0}$, $\mathrm{Cov}(\boldsymbol{f}_t^*)=\boldsymbol{P}'\mathrm{Cov}(\boldsymbol{f}_t)\boldsymbol{P}=\boldsymbol{P}'\boldsymbol{P}=\boldsymbol{I}$. 这样, $\boldsymbol{\beta}^*$ 和 $\boldsymbol{f}_t^*$ 对 $\boldsymbol{r}_t$ 建立了另一个正交因子模型. 正交因子表示的这种不唯一性既是缺点, 又是因子分析中的一个优点. 说它是缺点是因为它使得因子负荷的意义不确定了. 它也是一个优点, 因为它允许我们进行旋转来寻找具有良好解释的公共因子. 因为 $\boldsymbol{P}$ 是一个正交矩阵, 所以变换 $\boldsymbol{f}_t^*=\boldsymbol{P}'\boldsymbol{f}_t$ 是 m 维空间中的一个旋转.

9.5.1 估计

(9.16) 式中的正交因子模型可以通过两种方法估计. 第一种方法利用前一节中的主成分分析. 这个方法不要求数据的正态性假定, 也不要求预先指定公共因子的个数. 它对协方差矩阵和相关矩阵都是适用的. 但是同 PCA 中所提到的一样, 这个解通常只是一个近似. 第二种估计方法是最大似然方法. 它利用正态密度, 并要求预先指定公共因子的个数.

主成分方法

再次令 $(\hat{\lambda}_1, \hat{\boldsymbol{e}}_1), \cdots, (\hat{\lambda}_k, \hat{\boldsymbol{e}}_k)$ 是样本协方差矩阵 $\hat{\boldsymbol{\Sigma}}_r$ 的特征值和特征向量对, 其中 $\hat{\lambda}_1 \geqslant \hat{\lambda}_2 \geqslant \cdots \geqslant \hat{\lambda}_k$. 令 $m < k$ 为公共因子的个数, 则因子负荷矩阵由下式给出

$$\hat{\boldsymbol{\beta}} \equiv [\hat{\beta}_{ij}] = \left[\sqrt{\hat{\lambda}_1}\hat{\boldsymbol{e}}_1 \middle| \sqrt{\hat{\lambda}_2}\hat{\boldsymbol{e}}_2 \middle| \cdots \middle| \sqrt{\hat{\lambda}_m}\hat{\boldsymbol{e}}_m\right]. \tag{9.19}$$

个性方差的估计是矩阵 $\hat{\boldsymbol{\Sigma}}_r - \hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}'$ 的主对角线上的元素. 即 $\hat{\boldsymbol{D}} = \operatorname{diag}\{\hat{\sigma}_1^2, \cdots, \hat{\sigma}_k^2\}$, 其中 $\hat{\sigma}_i^2 = \hat{\sigma}_{ii,r} - \sum_{j=1}^{m} \hat{\beta}_{ij}^2$, $\hat{\sigma}_{ii,r}$ 是 $\hat{\boldsymbol{\Sigma}}_r$ 的第 (i,i) 个元素. 共性方差的估计为

$$\hat{c}_i^2 = \hat{\beta}_{i1}^2 + \cdots + \hat{\beta}_{im}^2.$$

由近似产生的误差矩阵为

$$\hat{\boldsymbol{\Sigma}}_r - (\hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}' + \hat{\boldsymbol{D}}).$$

我们当然希望这个矩阵接近于 0. 可以证明 $\hat{\boldsymbol{\Sigma}}_r - (\hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}' + \hat{\boldsymbol{D}})$ 的元素的平方和小于或等于 $\hat{\lambda}_{m+1}^2 + \cdots + \hat{\lambda}_k^2$. 因此, 近似误差的上界为所忽略的特征值的平方和.

由 (9.19) 式的解, 基于主成分方法的因子负荷估计并不随着公共因子 m 的增加而改变.

最大似然方法

如果公共因子 $\boldsymbol{f}_t$ 和个性因子 $\boldsymbol{\varepsilon}_t$ 是联合正态的, 那么 $\boldsymbol{r}_t$ 是多元正态的, 且其均值为 $\boldsymbol{\mu}$、协方差矩阵为 $\boldsymbol{\Sigma}_r = \boldsymbol{\beta}\boldsymbol{\beta}' + \boldsymbol{D}$. 在 $\boldsymbol{\beta}'\boldsymbol{D}^{-1}\boldsymbol{\beta} = \boldsymbol{\Delta}$(它是一个对角矩阵) 的限制下, 可以利用最大似然方法得到 $\boldsymbol{\beta}$ 和 $\boldsymbol{D}$ 的估计. 这里 $\boldsymbol{\mu}$ 是由样本均值估计的. 对这个方法的细节, 读者可以参考 Johnson 和 Wichern (2007).

在利用最大似然方法时, 公共因子的个数必须事先给定. 在实际中, 可以用修正的似然比检验来检查所拟合的 m- 因子模型的充分性. 检验统计量是

$$\mathrm{LR}(m) = -\left[T - 1 - \frac{1}{6}(2k+5) - \frac{2}{3}m\right](\ln|\hat{\boldsymbol{\Sigma}}_r| - \ln|\hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}' + \hat{\boldsymbol{D}}|). \tag{9.20}$$

在 m 个因子的零假设下, 上述检验统计量的渐近分布是自由度为 $\frac{1}{2}[(k-m)^2 - k - m]$ 的卡方分布. 我们将在 9.6.1 小节讨论选择 m 的一些方法.

9.5.2 因子旋转

正如前面提到的, 对任何 $m \times m$ 正交矩阵 $\boldsymbol{P}$,

$$\boldsymbol{r}_t - \boldsymbol{\mu} = \boldsymbol{\beta}\boldsymbol{f}_t + \boldsymbol{\varepsilon}_t = \boldsymbol{\beta}^*\boldsymbol{f}_t^* + \boldsymbol{\varepsilon}_t,$$

其中 $\boldsymbol{\beta}^* = \boldsymbol{\beta}\boldsymbol{P}$, $\boldsymbol{f}_t^* = \boldsymbol{P}'\boldsymbol{f}_t$. 另外,

$$\boldsymbol{\beta}\boldsymbol{\beta}' + \boldsymbol{D} = \boldsymbol{\beta}\boldsymbol{P}\boldsymbol{P}'\boldsymbol{\beta}' + \boldsymbol{D} = \boldsymbol{\beta}^*(\boldsymbol{\beta}^*)' + \boldsymbol{D}.$$

这个结果说明：共性方差与个性方差在正交变换下保持不变. 因此寻找一个正交矩阵 $\boldsymbol{P}$ 来变换因子模型使得公共因子具有良好的性质就是合理的. 这样一个变换等价于将公共因子在 m 维空间中旋转. 事实上, 有无限个可以利用的因子旋转. Kaiser(1958) 提出了一个方差最大化准则 (Varimax criterion) 来选择旋转. 这在许多应用中都运作地很好. 记 $\boldsymbol{\beta}^* = [\boldsymbol{\beta}^*_{ij}]$ 是旋转后的因子负荷矩阵, c_i^2 表示第 i 个共性方差. 定义 $\widetilde{\beta}^*_{ij} = \beta^*_{ij}/c_i$ 为经过共性方差 (正) 平方根的尺度变换之后的旋转系数. 方差最大化方法是选择对角矩阵 $\boldsymbol{P}$, 使得下式最大化

$$V = \frac{1}{k}\sum_{j=1}^{m}\left[\sum_{i=1}^{k}(\widetilde{\beta}^*_{ij})^4 - \frac{1}{k}\left(\sum_{i=1}^{k}\widetilde{\beta}^{*2}_{ij}\right)^2\right].$$

这个复杂的表示有一个简单的解释. 最大化 V, 对应于尽可能多地分散每个因子负荷的平方. 因此, 此方法是为了在因子负荷的旋转矩阵的任何列中寻找大的但可以忽略联合功效的组. 在实际应用中, 使用因子旋转来帮助解释公共因子. 这在一些应用中可能有益, 但在其他应用中未必有用. 对于因子旋转有许多可用的准则.

9.5.3　应用

给定资产收益率的数据 $\{\boldsymbol{r}_t\}$, 因子分析使得我们能够找到一些公共因子来解释收益率变化. 由于因子分析假定数据没有序列相关性, 所以在使用因子分析前应该检验这个假定的正确性. 为此, 我们可以使用多元混成统计量. 如果发现有序列相关性, 则可以构造一个 VARMA 模型来消除数据中的动态相依性, 并且对残差序列运用因子分析. 对许多收益率序列, 线性模型残差的相关矩阵经常非常接近于原始数据的相关矩阵. 在这种情形下, 动态依赖对因子分析的影响是可以忽略的.

本小节考虑 3 个例子. 前 2 个例子用 Minitab 软件进行分析, 第 3 个例子用 R 或 S-Plus 分析. 也可以用其他程序包.

例 9.2　再次考虑例 9.1 中使用的 IBM, HewLett-Parkard, Intel, J.P. Morgan Chase 和 Bank of America 的月对数股票收益率. 为了检验序列不相关的假设, 我们计算混成统计量得到 $Q_5(1) = 39.99$, $Q_5(5) = 160.60$, $Q_5(10) = 293.04$. 与自由度为 25, 125 和 250 的 χ^2 分布比较, 这些检验统计量的 p 值分别是 0.029, 0.017 和 0.032. 因此, 收益率存在某种微弱的序列相依, 但是, 这种相依性在 1%的水平上是不显著的. 为了简便, 我们在因子分析中忽略了序列相依性.

表 9-4 给出了基于相关矩阵运用主成分方法和最大似然方法的因子分析结果. 我们假定共同因子的个数是 2. 根据例 9.1 中的主成分分析, 这种取法是合理的. 从表中可见, 因子分析揭示了几个有趣的发现.

- 用最大似然方法确定的两个因子能够解释股票收益率变异性的 60%.
- 根据旋转后的因子载荷, 两个公共因子有一些有意义的解释. 技术类股票 (IBM、Hewlett-Packard 及 Intel) 对第一个因子的负荷很大, 而金融类股票

(J.P.Morgan Chase 和 Bank of America) 对第二个因子的负荷很大. 这两个旋转后的因子共同区分了产业部门.

- 在这个特例中, 方差最大化旋转似乎改变了两个公共因子的顺序.
- IBM 股票收益率的特殊方差相对较大, 表明该股票有自己的特性, 这值得进一步研究.

表 9-4 IBM, HewLett-Packard, Intel, J.P.Morgan Chase 和 Bank of America 的月对数股票收益率的因子分析 [a]

变量	因子负荷的估计		旋转因子负荷		共性方差
	f_1	f_2	f_1^*	f_2^*	$1-\sigma_i^2$
			最大似然法		
IBM	0.327	0.530	0.593	0.189	0.387
HPQ	0.348	0.669	0.733	0.177	0.568
INTC	0.337	0.647	0.709	0.171	0.531
JPM	0.734	0.186	0.358	0.667	0.573
BAC	0.960	−0.111	0.124	0.958	0.934
变量	1.801	1.193	1.535	1.459	2.994
比例	0.360	0.239	0.307	0.292	0.599

a 收益率包含分红, 时间是从 1990 年 1 月到 2008 年 12 月. 分析基于样本交叉-相关阵并假定有两个公共因子.

例 9.3 在这个例子中, 我们考虑期限为 30 年、20 年、10 年、5 年和 1 年的美国债券指数的月对数收益率. 例 8.2 中描述过这个数据, 但被转换成了对数收益率. 总共有 696 个观测值. 正如例 8.2 中显示的, 数据具有序列依赖性. 然而, 通过拟合一个 VARMA(2,1) 模型来消除序列依赖几乎不对同步相关矩阵具有任何影响. 事实上, 拟合一个 VARMA(2,1) 模型之前和之后的相关矩阵分别为

$$\hat{\boldsymbol{\rho}}_o = \begin{bmatrix} 1.0 & & & & \\ 0.98 & 1.0 & & & \\ 0.92 & 0.91 & 1.0 & & \\ 0.85 & 0.86 & 0.90 & 1.0 & \\ 0.63 & 0.64 & 0.67 & 0.81 & 1.0 \end{bmatrix}, \quad \hat{\boldsymbol{\rho}} = \begin{bmatrix} 1.0 & & & & \\ 0.98 & 1.0 & & & \\ 0.92 & 0.92 & 1.0 & & \\ 0.85 & 0.86 & 0.90 & 1.0 & \\ 0.66 & 0.67 & 0.71 & 0.84 & 1.0 \end{bmatrix},$$

其中 $\hat{\boldsymbol{\rho}}_o$ 是原始对数收益率的相关矩阵. 因此, 我们直接对收益率序列应用因子分析.

表 9-5 中给出了数据因子分析的结果. 对两种估计方法, 前两个公共因子对数据总方差的解释都超过了 90%. 事实上, 高的共性方差说明对五种债券指数收益率而言, 其个性方差都非常小. 因为两种方法的结果接近, 故我们只讨论主成分方法. 非旋转因子负荷说明: (a) 所有 5 种收益序列对第一个因子的负荷粗略地相等; (b) 对第二个因子的负荷与期限长短是正相关的. 因此, 第一个公共因子代表了一般的美国债券收益率, 第二个因子体现了 "期限" 效应. 而且, 第二个因子负荷的和接近于 0. 因此, 这个公共因子也可以解释为长期债券与短期债券的比较. 这里一

个长期债券指的是期限为 10 年或更长的债券. 对旋转后的因子, 其负荷也是有趣的. 对第一个旋转因子的负荷与期限成正比例, 而对第二个因子的负荷与期限成反比例.

表 9-5 期限为 30 年、20 年、10 年、5 年和 1 年的美国债券指数的月对数收益率的因子分析 [a]

变量	因子负荷的估计		旋转因子负荷		共性方差
	f_1	f_2	f_1^*	f_2^*	$1-\sigma_i^2$
			主成分法		
30 年	0.952	0.253	0.927	0.333	0.970
20 年	0.954	0.240	0.922	0.345	0.968
10 年	0.956	0.140	0.866	0.429	0.934
5 年	0.955	−0.142	0.704	0.660	0.931
1 年	0.800	−0.585	0.325	0.936	0.982
变量	4.281	0.504	3.059	1.726	4.785
比例	0.856	0.101	0.612	0.345	0.957
			最大似然法		
30 年	0.849	−0.513	0.895	0.430	0.985
20 年	0.857	−0.486	0.876	0.451	0.970
10 年	0.896	−0.303	0.744	0.584	0.895
5 年	1.000	0.000	0.547	0.837	1.000
1 年	0.813	0.123	0.342	0.747	0.675
变量	3.918	0.607	2.538	1.987	4.525
比例	0.784	0.121	0.508	0.397	0.905

a 时间是从 1942 年 1 月到 1999 年 12 月. 分析基于样本交叉相关阵, 假定有两个公共因子.

例 9.4 再一次考虑表 9-2 中的 10 只股票的月超额收益率. 时间区间是从 1990 年 1 月到 2003 年 12 月, 收益率以百分比的形式给出. 我们的目的是用 S-Plus 命令`factanal`演示一下统计因子模型的应用. 我们从二因子模型开始, 但是 (9.20) 式的似然比检验拒绝了二因子模型的假设. 检验统计量是 LR(2)=72.96. 基于自由度 26 的渐近卡方分布, 检验统计量的 p 值接近为零.

```
> rtn=read.table(''m-barra-9003.txt'',header=T)
> stat.fac=factanal(rtn,factors=2,method='mle')
> stat.fac
Sums of squares of loadings:
  Factor1 Factor2
 2.696479 2.19149

Component names:
 "loadings" "uniquenesses" "correlation" "criteria"
 "factors" "dof" "method" "center" "scale" "n.obs"
 "scores" "call"
```

接着应用一个三因子模型, 在 5%的显著性水平下该模型似乎是合理的. LR(3) 统计量的 p 值是 0.089 2.

```
> stat.fac=factanal(rtn,factor=3,method='mle')
> stat.fac
Test of the hypothesis that 3 factors are sufficient
```

```
versus the alternative that more are required:
The chi square statistic is 26.48 on 18 degrees of freedom.
The p-value is 0.0892
> summary(stat.fac)
Importance of factors:

                  Factor1   Factor2   Factor3
   SS loadings      2.635     1.825     1.326
Proportion Var      0.264     0.183     0.133
Cumulative Var      0.264     0.446     0.579

Uniquenesses:
   AGE    C      MWD    MER   DELL    HPQ    IBM
 0.479 0.341 0.201 0.216 0.690 0.346 0.638
   AA     CAT    PG
 0.417 0.000 0.885

Loadings:
    Factor1 Factor2 Factor3
AGE  0.678   0.217   0.121
C    0.739   0.259   0.213
MWD  0.817   0.356
MER  0.819   0.329
DELL 0.102   0.547
HPQ  0.230   0.771
IBM  0.200   0.515   0.238
AA   0.194   0.546   0.497
CAT  0.198   0.138   0.970
PG   0.331
```

因子负荷也可以用下述命令>plot(loadings(stat.fac))在图上表示出来, 即图9-6. 从图中可以看出因子 1 本质上代表金融服务类, 而因子 2 主要由高科技类和 Alcoa 股票的超额收益率构成, 因子 3 很大程度上依赖于 CAT 股票和 AA 股票的超额收益率, 因此代表剩余的产业股.

用命令rotate可以进行因子旋转. 该命令允许许多种旋转方法. 由命令predict可以得到因子实现.

```
> stat.fac2 = rotate(stat.fac,rotation='quartimax')
> loadings(stat.fac2)
    Factor1 Factor2 Factor3
AGE  0.700   0.171
C    0.772   0.216   0.124
MWD  0.844   0.291
MER  0.844   0.264
DELL 0.144   0.536
HPQ  0.294   0.753
IBM  0.258   0.518   0.164
AA   0.278   0.575   0.418
CAT  0.293   0.219   0.931
PG   0.334
> factor.real=predict(stat.fac,type='weighted.ls')
```

最后基于所拟合的三因子统计因子模型可以得到这 10 只股票超额收益率的相关矩阵. 如所料想, 与 9.3.1 节的产业因子模型所得到的相关矩阵相比, 此处得到的相关矩阵与样本相关矩阵的对应部分更接近. 可以用 GMVP 来比较收益率和统计因子模型的协方差矩阵.

```
> corr.fit=fitted(stat.fac)
> print(corr.fit,digits=1,width=2)
      AGE  C  MWD MER DELL HPQ IBM AA  CAT PG
AGE   1.0 0.6 0.6 0.6 0.19 0.3 0.3 0.3 0.3 0.2
C     0.6 1.0 0.7 0.7 0.22 0.4 0.3 0.4 0.4 0.3
MWD   0.6 0.7 1.0 0.8 0.28 0.5 0.4 0.4 0.3 0.3
MER   0.6 0.7 0.8 1.0 0.26 0.5 0.4 0.4 0.3 0.3
DELL  0.2 0.2 0.3 0.3 1.00 0.5 0.3 0.3 0.1 0.0
HPQ   0.3 0.4 0.5 0.4 0.45 1.0 0.5 0.5 0.2 0.1
IBM   0.3 0.3 0.4 0.3 0.31 0.5 1.0 0.4 0.3 0.1
AA    0.3 0.4 0.4 0.4 0.33 0.5 0.4 1.0 0.6 0.1
CAT   0.3 0.4 0.3 0.3 0.11 0.2 0.3 0.6 1.0 0.1
PG    0.2 0.3 0.3 0.3 0.03 0.1 0.1 0.1 0.1 1.0
```

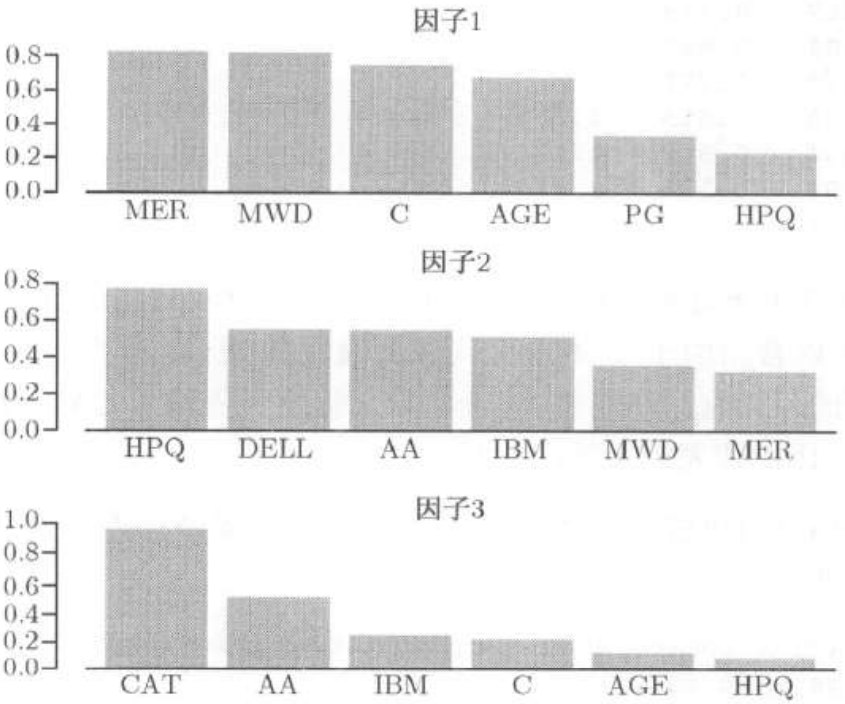

图 9-6　对表 9-2 中 10 只股票的月超额收益率拟合一个三因子模型时的因子负荷图

9.6　渐近主成分分析

到现在为止, 关于 PCA 的讨论都是假定资产的个数小于所考虑时期的个数, 即 $k < T$. 为了处理 T 较小 k 较大的情形, Conner 和 Korajczyk(1986,1988) 提出了渐近主成分分析 (APCA) 的概念. 该方法与传统的 PCA 相似, 但是依赖于资产

数目趋于无穷时的渐近结果. 因此 APCA 是基于下述 $T \times T$ 矩阵的特征值 — 特征向量分析:

$$\hat{\boldsymbol{\Omega}}_T = \frac{1}{k}(\boldsymbol{R} - \mathbf{1}_T\bar{\boldsymbol{r}}')(\boldsymbol{R} - \mathbf{1}_T\bar{\boldsymbol{r}}')',$$

其中 $\mathbf{1}_T$ 是元素全为 1 的 T 维向量, $\bar{\boldsymbol{r}} = (\bar{r}_1, \cdots, \bar{r}_k)$, $\bar{r}_i = (\mathbf{1}_T'\boldsymbol{R}_i)/T$ 为第 i 个收益率序列的样本均值. Conner 和 Korajczyk(1988) 证明了当 $k \to \infty$ 时, $\hat{\boldsymbol{\Omega}}_T$ 的特征值–特征向量分析等价于传统的统计因子分析. 换言之, 因子 $\boldsymbol{f}_t$ 的 APCA 估计是 $\hat{\boldsymbol{\Omega}}_T$ 的前 m 个特征向量. 令 $\hat{\boldsymbol{F}}_t$ 是由 $\hat{\boldsymbol{\Omega}}_T$ 的前 m 个特征向量组成的 $m \times T$ 矩阵, 则 $\hat{\boldsymbol{f}}_t$ 是 $\hat{\boldsymbol{F}}_t$ 的第 t 列. 应用类似于 BARRA 因子模型的估计的思想, Conner 和 Korajczyk(1988) 建议按如下步骤修正 $\hat{\boldsymbol{f}}_t$ 的估计:

(1) 对 $t = 1, \cdots, T$, 利用样本协方差矩阵 $\hat{\boldsymbol{\Omega}}_T$ 得到初始估计 $\hat{\boldsymbol{f}}_t$;

(2) 对每个资产, 给出模型

$$r_{it} = \alpha_i + \boldsymbol{\beta}_i\hat{\boldsymbol{f}}_t + \varepsilon_{it}, \quad t = 1, \cdots, T,$$

的 OLS 估计, 其中 $\boldsymbol{\beta}_i = (\beta_{i1}, \cdots, \rho_{im})$, 并计算残差的方差 $\hat{\sigma}_i^2$.

(3) 构造对角矩阵 $\hat{\boldsymbol{D}} = \mathrm{diag}\{\hat{\sigma}_1^2, \cdots, \hat{\sigma}_k^2\}$, 并且将收益率进行如下刻度变换

$$\boldsymbol{R}_* = \boldsymbol{R}\hat{\boldsymbol{D}}^{-1/2}.$$

(4) 利用 $\boldsymbol{R}_*$ 计算下述 $T \times T$ 协方差矩阵

$$\hat{\boldsymbol{\Omega}}_* = \frac{1}{k}(\boldsymbol{R}_* - \mathbf{1}_T\bar{\boldsymbol{r}}_*')(\boldsymbol{R}_* - \mathbf{1}_T\bar{\boldsymbol{r}}_*')',$$

其中 $\bar{\boldsymbol{r}}_*$ 是由 $\boldsymbol{R}_*$ 的列中值所构成的 k 维向量, 然后对 $\hat{\boldsymbol{\Omega}}_*$ 进行特征值–特征向量分析来得到 $\boldsymbol{f}_t$ 的修正估计.

9.6.1 因子个数的选择

文献中有两种方法来选择因子分析中因子的个数. 第一种方法是由 Conner 和 Korajczyk(1993) 提出的. 该方法所用的思想是如果 m 是正确的因子个数, 则当因子个数从 m 变到 $m+1$ 时资产个性误差 ε_{it} 的横截面方差应该不会显著下降. 第二种方法是 Bai 和 Ng(2002) 提出来的. 该方法采用一些信息准则来选择因子个数. 后一种方法基于这样一个观测到的事实: $\hat{\boldsymbol{\Omega}}_T$ 的特征值 — 特征向量分析求解出了下述最小二乘问题

$$\min_{\alpha,\boldsymbol{\beta},\boldsymbol{f}_t} \frac{1}{kT}\sum_{i=1}^{k}\sum_{t=1}^{T}(r_{it} - \alpha_i - \boldsymbol{\beta}_i\boldsymbol{f}_t)^2.$$

假定存在 m 个因子, 则 $\boldsymbol{f}_t$ 是 m 维的. 令 $\hat{\sigma}_i^2(m)$ 表示对资产 i 进行前面的最小二乘问题的组内回归的残差的方差, 这里要利用 APCA 分析中得到的 $\hat{\boldsymbol{f}}_t$. 定义残差的横截面方差如下:

$$\hat{\sigma}^2(m) = \frac{1}{k}\sum_{i=1}^{k}\hat{\sigma}_i^2(m).$$

Bai 和 Ng(2002) 给出的准则是

$$C_{p1}(m) = \hat{\sigma}^2(m) + m\hat{\sigma}^2(M)\left(\frac{k+T}{kT}\right)\ln\left(\frac{kT}{k+T}\right),$$

$$C_{p2}(m) = \hat{\sigma}^2(m) + m\hat{\sigma}^2(M)\left(\frac{k+T}{kT}\right)\ln\left(P_{kT}^2\right),$$

其中 M 是事先指定的正整数, 它表示因子的最大个数; $P_{kT} = \min(\sqrt{k}, \sqrt{T})$. 使得 $C_{p1}(m)$ 或 $C_{p2}(m)$ 最小的 m 便是我们所要选择的因子个数, 这里 $0 \leqslant m \leqslant M$. 实际中, 这两个准则可能会选择不同的因子个数.

9.6.2 例子

为了进一步说明渐近主成分分析, 考虑 40 只股票的月简单收益率, 时间区间是从 2001 年 1 月到 2003 年 12 月, 共 36 个观测. 于是我们有 $k = 40, T = 36$. 表 9-6 给出了这些股票的代码. 这些股票是在 2004 年 9 月份的某一天在 NASDAQ 和 NYSE 中交易频繁的股票. 主要用到的 S-Plus 命令是`mfactor`.

表 9-6 渐近主成分分析中所用到的股票的代码, 样本时间区间是从 2001 年 1 月到 2003 年 12 月

市 场	Tick Symbol				
NASDAQ	INTC	MSFT	SUNW	CSCO	AMAT
	ORCL	SIRI	COCO	CORV	SUPG
	YHOO	JDSU	QCOM	CIEN	DELL
	ERTS	EBAY	ADCT	AAPL	JNPR
NYSE	LU	PFE	NT	BAC	BSX
	GE	TXN	XOM	FRX	Q
	F	TWX	C	MOT	JPM
	TYC	HPQ	NOK	WMT	AMD

我们用前面所讨论的两种方法来选择因子个数. Conner 和 Korajczyk 提出的方法选择了 $m = 1$, 而 Bai 和 Ng 提出的方法选择了 $m = 6$. 对于后一种方法, 两个准则给出了不同的选择.

```
> dim(rtn)  % rtn is the return data.
[1] 36 40
> nf.ck=mfactor(rtn,k='ck',max.k=10,sig=0.05)
> nf.ck
Call:
mfactor(x = rtn, k = "ck", max.k = 10, sig = 0.05)

Factor Model:
 Factors Variables Periods
       1        40      36
Factor Loadings:
     Min. 1st Qu. Median   Mean 3rd Qu.  Max.
F.1 0.069  0.432  0.629  0.688  1.071 1.612
Regression R-squared:
```

```
   Min. 1st Qu. Median   Mean 3rd Qu.   Max.
 0.090  0.287   0.487   0.456  0.574   0.831
> nf.bn=mfactor(rtn,k='bn',max.k=10,sig=0.05)
Warning messages:
Cp1andCp2 did not yield same result.The smaller one is used.
> nf.bn$k
[1] 6
```

取 $m = 6$, 我们对收益率序列应用 APCA 可以得到斜坡图和被估收益率因子.

```
> apca = mfactor(rtn,k=6)
> apca
Call:
mfactor(x = rtn, k = 6)
Factor Model:
 Factors Variables Periods
       6        40      36
Factor Loadings:
         Min  1st Qu. Median    Mean  3rd Qu.  Max.
F.1    0.048   0.349   0.561   0.643   0.952  2.222
F.2   -1.737   0.084   0.216   0.214   0.323  1.046
F.3   -1.512   0.002   0.076   0.102   0.255  1.093
F.4   -0.965  -0.035   0.078   0.048   0.202  0.585
F.5   -0.722  -0.008   0.056   0.066   0.214  0.729
F.6   -0.840  -0.088   0.003   0.003   0.071  0.635
Regression R-squared:
   Min. 1st Qu. Median  Mean 3rd Qu.  Max.
 0.219  0.480   0.695  0.651  0.801  0.999

> screeplot.mfactor(apca)
> fplot(factors(apca))
```

图 9-7 给出了 40 支股票收益率的 APCA 的斜坡图. 6 个公共因子大约解释了变化的 89.4%. 图 9-8 给出了 6 个被估因子的时间图.

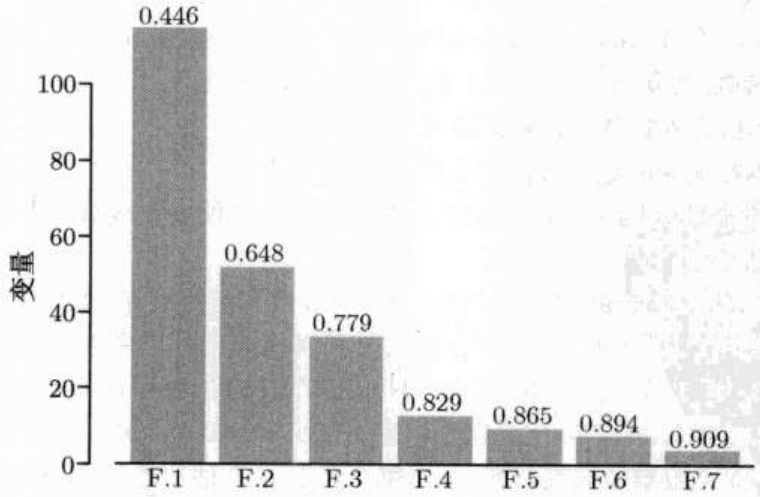

图 9-7 对 40 只股票的月简单收益率进行 APCA 的斜坡图. 样本时间区间是从 2001 年 1 月到 2003 年 12 月

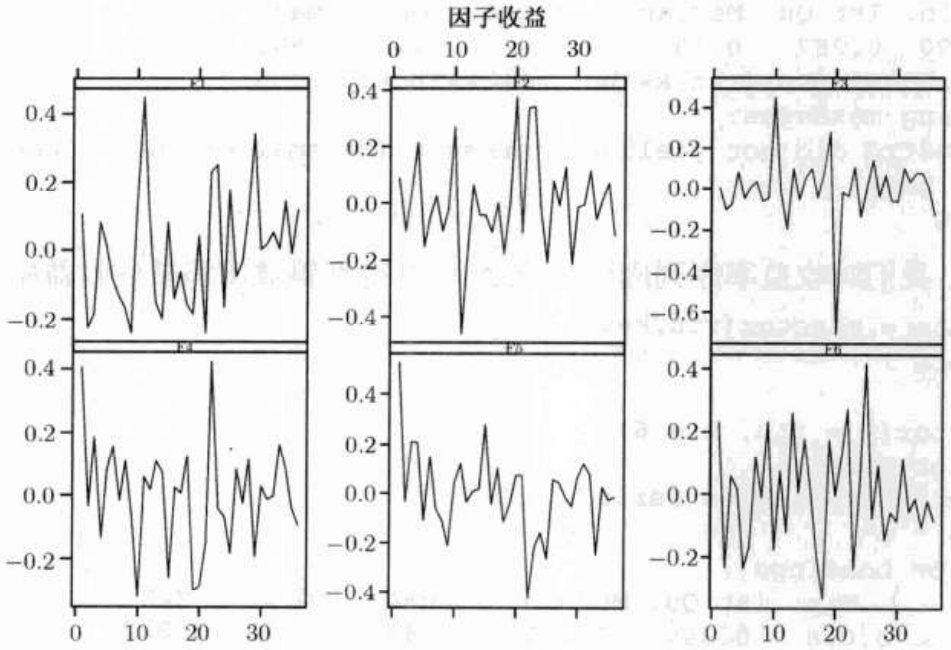

图 9-8　对 40 只股票的月简单收益率进行 APCA 所得到的因子收益的时间图. 样本时间区间是从 2001 年 1 月到 2003 年 12 月

练　习　题

9.1 考虑用百分数表示并且包括红利的从 1990 年 1 月到 2008 年 12 月的 13 只股票和标准普尔 500 综合指数的简单超额月收益率. 使用二级市场上 3 个月国债月利率作为无风险利率计算超额收益率. 股票的小记号分别为 AA、AXP、CAT、DE、F、FDX、HPQ、IBM、JNJ、KMB、MMM、PG 和 WFC. 数据包含在文件 `m-fac-ex-9008.txt` 中. 对 13 只股票收益率进行 9.2.1 节中的市场模型分析, 对每个股票收益率序列求 β_i、σ_i^2 和 R^2 的估计值.

9.2 考虑用百分数表示并且包括红利的从 1960 年 1 月到 2008 年 12 月的 Merck & Company、Johnson & Johnson、General Electric、General Motors、Ford Motor Company 和价值加权指数; 见第 8 章练习 8.1 的文件 `m-mrk2vw.txt`.

(a) 使用样本协方差矩阵进行数据的主成分分析.

(b) 使用样本相关矩阵进行数据的主成分分析.

(c) 对数据进行统计因子分析. 确定公共因子数量. 使用主成分和最大似然方法求因子载荷的估计值.

9.3 文件 `m-excess-c10sp-9003.txt` 包含 10 只股票和标准普尔 500 综合指数的简单超额月收益率. 使用二级市场上 3 个月国债月利率作为无风险利率计算超额收益率. 样本期为从 1990 年 1 月到 2003 年 12 月, 共有 168 个观察值. 文件的 11 列分别包含 ABT、LLY、MRK、PFE、F、GM、BP、CVX、RD、XOM 和 SP5. 使用单因子市场模型分析 10 个股票超额收益率, 画出每个股票的 β 估计值和 R^2, 使用全局最小方差投资组合比较所拟合模型和数据的协方差矩阵.

9.4 再次考虑文件 `m-excess-c10sp-9003.txt` 中的 10 只股票的收益率, 股票来自 3 个产业的公司, ABT、LLY、MRK 和 PFE 是主要的医药公司, F 和 GM 是汽车公司, 其他是

大的石油公司. 使用 BARRA 产业因子模型分析超额收益率. 画出三因子实现, 并评价所拟合模型的充分性.

9.5 再次考虑文件 `m-excess-c10sp-9003.txt` 中的 10 只股票的超额收益率, 对收益率进行主成分分析, 并获取碎石图, 共有几个公共因子? 为什么? 解释公共因子.

9.6 再次考虑文件 `m-excess-c10sp-9003.txt` 中 10 只股票的超额收益率, 对其进行统计因子分析, 在 5%的显著性水平下要用到多少个公共因子? 画出所拟合模型的被估因子负荷图. 这些公共因子有意义吗?

9.7 文件 `m-fedip.txt` 包含从 1954 年 7 月到 2003 年 12 月的联邦基金有效年利率和月利率以及工业生产指数. 工业生产指数已经经过了季节调整. 用联邦基金利率和工业生产指数作为宏观经济变量为文件 `m-excess-c10sp-9003.txt` 中 10 只股票的超额收益率拟合一个宏观经济因子模型. 可以用 VAR 模型来得到宏观经济变量的意外序列. 解释所拟合的因子模型.

参考文献

Alexander, C. (2001). *Market Models: A Guide to Financial Data Analysis*. Wiley, Hoboken, NJ.

Bai, J. and Ng, S. (2002). Determining the number of factors in approximate factor models. *Econometrica* **70**: 191–221.

Campbell, J. Y., Lo, A. W., and MacKinlay, A. C. (1997). *The Econometrics of Financial Markets*. Princeton University Press, Princeton, NJ.

Chen, N. F., Roll, R., and Ross, S. A. (1986). Economic forces and the stock market. *Journal of Business* **59**: 383–404.

Connor, G. (1995). The three types of factor models: A comparison of their explanatory power. *Financial Analysts Journal* **51**: 42–46.

Connor, G. and Korajczyk, R. A. (1986). Performance measurement with the arbitrage pricing theory: A new framework for analysis. *Journal of Financial Economics* **15**: 373–394.

Connor, G. and Korajczyk, R. A. (1988). Risk and return in an equilibrium APT: Application of a new test methodology. *Journal of Financial Economics* **21**: 255–289.

Connor, G. and Korajczyk, R. A. (1993). A test for the number of factors in an approximate factor model. *Journal of Finance* **48**: 1263–1292.

Fama, E. and French, K. R. (1992). The cross-section of expected stock returns. *Journal of Finance* **47**: 427–465.

Grinold, R. C. and Kahn, R. N. (2000). *Active Portfolio Management: A Quantitative Approach for Producing Superior Returns and Controlling Risk*, 2nd ed. McGraw-Hill, New York.

Johnson, R. A. and Wichern, D. W. (2007). *Applied Multivariate Statistical Analysis*, 6th ed. Prentice Hall, Upper Saddle River, NJ.

Kaiser, H. F. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika* **23**: 187–200.

Sharpe, W. (1970). *Portfolio Theory and Capital Markets*. McGraw-Hill, New York.

Zivot, E. and Wang, J. (2003). *Modeling Financial Time Series with S-Plus*. Springer New York.