CHAPTER 9

Principal Component Analysis and Factor Models

Most financial portfolios consist of multiple assets, and their returns depend concurrently and dynamically on many economic and financial variables. Therefore, it is important to use proper multivariate statistical analyses to study the behavior and properties of portfolio returns. However, as demonstrated in the previous chapter, analysis of multiple asset returns often requires high-dimensional statistical models that are complicated and hard to apply. To simplify the task of modeling multiple returns, we discuss in this chapter some dimension reduction methods to search for the underlying structure of the assets. *Principal component analysis* (PCA) is perhaps the most commonly used statistical method in dimension reduction, and we start our discussion with the method. In practice, observed return series often exhibit similar characteristics leading to the belief that they might be driven by some common sources, often referred to as common factors. To study the common pattern in asset returns and to simplify portfolio analysis, various factor models have been proposed in the literature to analyze multiple asset returns. The second goal of this chapter is to introduce some useful factor models and demonstrate their applications in finance.

Three types of factor models are available for studying asset returns; see Connor (1995) and Campbell, Lo, and MacKinlay (1997). The first type is the *macroeconomic factor models* that use macroeconomic variables such as growth rate of GDP, interest rates, inflation rate, and unemployment rate to describe the common behavior of asset returns. Here the factors are observable and the model can be estimated via linear regression methods. The second type is the *fundamental factor models* that use firm or asset specific attributes such as firm size, book and market values, and industrial classification to construct common factors. The third type is the *statistical factor models* that treat the common factors as unobservable or latent variables to be estimated from the returns series. In this chapter, we discuss all

Analysis of Financial Time Series, Third Edition, By Ruey S. Tsay

three types of factor models and their applications in finance. Principal component analysis and factor models for asset returns are also discussed in Alexander (2001) and Zivot and Wang (2003).

The chapter is organized as follows. Section 9.1 introduces a general factor model for asset returns, and Section 9.2 discusses macroeconomic factor models with some simple examples. The fundamental factor model and its applications are given in Section 9.3. Section 9.4 introduces principal component analysis that serves as the basic method for statistical factor analysis. The PCA can also be used to reduce the dimension in multivariate analysis. Section 9.5 discusses the orthogonal factor models, including factor rotation and its estimation, and provides several examples. Finally, Section 9.6 introduces asymptotic principal component analysis.

9.1 A FACTOR MODEL

Suppose that there are k assets and T time periods. Let r_{it} be the return of asset i in the time period t. A general form for the factor model is

$$r_{it} = \alpha_i + \beta_{i1} f_{1t} + \cdots + \beta_{im} f_{mt} + \epsilon_{it}, \qquad t = 1, \ldots, T; \qquad i = 1, \ldots, k, \quad (9.1)$$

where α_i is a constant representing the intercept, $\{f_{jt} | j = 1, \ldots, m\}$ are m common factors, β_{ij} is the *factor loading* for asset i on the jth factor, and ϵ_{it} is the *specific factor* of asset i.

For asset returns, the factor $\boldsymbol{f}_t = (f_{1t}, \ldots, f_{mt})'$ is assumed to be an m-dimensional stationary process such that

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

and the asset specific factor ϵ_{it} is a white noise series and uncorrelated with the common factors f_{jt} and other specific factors. Specifically, we assume that

$$\begin{aligned} E(\epsilon_{it}) &= 0 \quad \text{for all } i \text{ and } t, \\ \text{Cov}(f_{jt}, \epsilon_{is}) &= 0 \quad \text{for all } j, i, t \text{ and } s, \\ \text{Cov}(\epsilon_{it}, \epsilon_{js}) &= \begin{cases} \sigma_i^2, & \text{if } i = j \text{ and } t = s, \\ 0, & \text{otherwise.} \end{cases} \end{aligned}$$

Thus, the common factors are uncorrelated with the specific factors, and the specific factors are uncorrelated among each other. The common factors, however, need not be uncorrelated with each other in some factor models.

In some applications, the number of assets k may be larger than the number of time periods T. We discuss an approach to analyze such data in Section 9.6. It is also common to assume that the factors, hence $\boldsymbol{r}_t$, are serially uncorrelated in factor analysis. In applications, if the observed returns are serially dependent, then the models in Chapter 8 can be used to remove the serial dependence.

In matrix form, the factor model in Eq. (9.1) can be written as

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

where $\boldsymbol{\beta}_i = (\beta_{i1}, \ldots, \beta_{im})$ is a row vector of loadings, and the joint model for the k assets at time t is

$$\boldsymbol{r}_t = \boldsymbol{\alpha} + \boldsymbol{\beta} \boldsymbol{f}_t + \boldsymbol{\epsilon}_t, \qquad t = 1, \ldots, T \tag{9.2}$$

where $\boldsymbol{r}_t = (r_{1t}, \ldots, r_{kt})'$, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_k)'$, $\boldsymbol{\beta} = [\beta_{ij}]$ is a $k \times m$ factor-loading matrix, and $\boldsymbol{\epsilon}_t = (\epsilon_{1t}, \ldots, \epsilon_{kt})'$ is the error vector with $\text{Cov}(\boldsymbol{\epsilon}_t) = \boldsymbol{D} = \text{diag}\{\sigma_1^2, \ldots, \sigma_k^2\}$, a $k \times k$ diagonal matrix. The covariance matrix of the return $\boldsymbol{r}_t$ is then

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

The model presentation in Eq. (9.2) is in a *cross-sectional* regression form if the factors f_{jt} are observed.

Treating the factor model in Eq. (9.1) as a time series, we have

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

for the ith asset ($i = 1, \ldots, k$), where $\boldsymbol{R}_i = (r_{i1}, \ldots, r_{iT})'$, $\boldsymbol{1}_T$ is a T-dimensional vector of ones, $\boldsymbol{F}$ is a $T \times m$ matrix whose tth row is $\boldsymbol{f}_t'$, and $\boldsymbol{E}_i = (\epsilon_{i1}, \ldots, \epsilon_{iT})'$. The covariance matrix of $\boldsymbol{E}_i$ is $\text{Cov}(\boldsymbol{E}_i) = \sigma_i^2 \boldsymbol{I}$, a $T \times T$ diagonal matrix.

Finally, we can rewrite Eq. (9.2) as

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

where $\boldsymbol{g}_t = (1, \boldsymbol{f}_t')'$ and $\boldsymbol{\xi} = [\boldsymbol{\alpha}, \boldsymbol{\beta}]$, which is a $k \times (m+1)$ matrix. Taking the transpose of the prior equation and stacking all data together, we have

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

where $\boldsymbol{R}$ is a $T \times k$ matrix of returns whose tth row is $\boldsymbol{r}_t'$ or, equivalently, whose ith column is $\boldsymbol{R}_i$ of Eq. (9.3), $\boldsymbol{G}$ is a $T \times (m+1)$ matrix whose tth row is $\boldsymbol{g}_t'$, and $\boldsymbol{E}$ is a $T \times k$ matrix of specific factors whose tth row is $\boldsymbol{\epsilon}_t'$. If the common factors $\boldsymbol{f}_t$ are observed, then Eq. (9.4) is a special form of the *multivariate linear regression* (MLR) model; see Johnson and Wichern (2007). For a general MLR model, the covariance matrix of $\boldsymbol{\epsilon}_t$ need not be diagonal.

9.2 MACROECONOMETRIC FACTOR MODELS

For macroeconomic factor models, the factors are observed and we can apply the least-squares method to the MLR model in Eq. (9.4) to perform estimation. The estimate is

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

from which the estimates of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are readily available. The residuals of Eq. (9.4) are

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

Based on the model assumption, the covariance matrix of $\boldsymbol{\epsilon}_t$ is estimated by

$$\widehat{\boldsymbol{D}} = \text{diag}\{\hat{\sigma}_1^2, \ldots, \hat{\sigma}_k^2\},$$

where $\hat{\sigma}_i^2$ is the (i, i)th element of $\widehat{\boldsymbol{E}}'\widehat{\boldsymbol{E}}/(T - m - 1)$. Furthermore, the R^2 of the ith asset of Eq. (9.3) is

$$R_i^2 = 1 - \frac{[\widehat{\boldsymbol{E}}'\widehat{\boldsymbol{E}}]_{i,i}}{[\boldsymbol{R}'\boldsymbol{R}]_{i,i}}, \qquad i = 1, \ldots, k,$$

where $\boldsymbol{A}_{i,i}$ denotes the (i, i)th element of the matrix $\boldsymbol{A}$.

Note that the aforementioned least-squares estimation does not impose the constraint that the specific factors $\boldsymbol{\epsilon}_{it}$ are uncorrelated with each other. Consequently, the estimates obtained are not efficient in general. However, imposing the orthogonalization constraint requires nontrivial computation and is often ignored. One can check the off-diagonal elements of the matrix $\widehat{\boldsymbol{E}}'\widehat{\boldsymbol{E}}/(T - m - 1)$ to verify the adequacy of the fitted model. These elements should be close to zero.

9.2.1 Single-Factor Model

The best known macroeconomic factor model in finance is the *market model*; see Sharpe (1970). This is a single-factor model and can be written as

$$r_{it} = \alpha_i + \beta_i r_{mt} + \epsilon_{it}, \qquad i = 1, \ldots, k; \qquad t = 1, \ldots, T, \tag{9.5}$$

where r_{it} is the excess return of the ith asset, r_{mt} is the excess return of the market, and β_i is the well-known β for stock returns. To illustrate, we consider monthly returns of 13 stocks and use the return of the S&P 500 index as the market return. The stocks used and their tick symbols are given in Table 9.1, and the sample period is from January 1990 to December 2003 so that $k = 13$ and $T = 168$. We use the monthly series of 3-month Treasury bill rates of the secondary market as

TABLE 9.1 Stocks Used and Their Tick Symbols in Analysis of Single-Factor Model[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&P 500 Index	0.42(4.33)

[a] Sample means (standard errors) of excess returns are also given. The sample period is from January 1990 to December 2003.

the risk-free interest rate to obtain simple excess returns of the stock and market index. The returns are in percentages.

We use S-Plus to implement the estimation method discussed in the previous section. Most of the commands used also apply to the software R.

```
> 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))
```

The estimates of β_i, σ_i^2, and R^2 for the ith asset return are given below:

```
> 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
```

Figure 9.1 shows the bar plots of $\hat{\beta}_i$ and R^2 of the 13 stocks. The financial stocks, AGE and MEL, and the high-tech stocks, HPQ and TXN, seem to have higher β and R^2. On the other hand, KMB and PG have lower β and R^2. The R^2

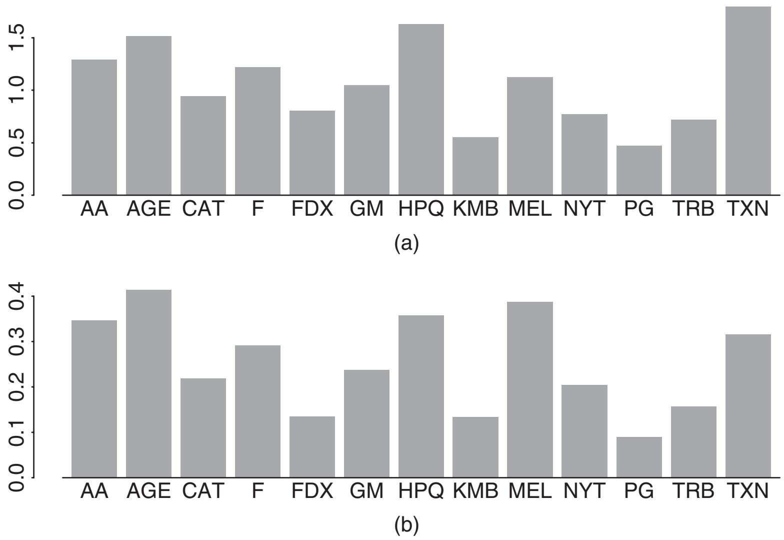


Figure 9.1 Bar plots of (a) beta and (b) R^2 for fitting single-factor market model to monthly excess returns of 13 stocks. S&P 500 index excess return is used as market index. Sample period is from January 1990 to December 2003.

ranges from 0.09 to 0.41, indicating that the market return explains less than 50% of the variabilities of the individual stocks used.

The covariance and correlation matrices of $\boldsymbol{r}_t$ under the market model can be estimated using the following:

```
> 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
```

We can compare these estimated correlations with the sample correlations of the excess returns.

```
> 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
```

In finance, one can use the concept of *global minimum variance portfolio* (GMVP) to compare the covariance matrix implied by a fitted factor model with the sample covariance matrix of the returns. For a given covariance matrix $\boldsymbol{\Sigma}$, the global minimum variance portfolio is the portfolio $\boldsymbol{\omega}$ that solves

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

where $\sigma^2_{p,\boldsymbol{\omega}}$ is the variance of the portfolio. The solution is given by

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

where $\mathbf{1}$ is the k-dimensional vector of ones.

For the market model considered, the GMVP for the fitted model and the data are as follows:

```
> 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
```

Comparing the two GMVPs, the weights assigned to TRB stock differ markedly. The two portfolios, however, have larger weights for KMB, NYT, and PG stocks.

Finally, we examine the residual covariance and correlation matrices to verify the assumption that the special factors are not correlated among the 13 stocks. The first four columns of the residual correlation matrix are given below and there exist some large values in the residual cross correlations, for example, Cor(CAT,AA) = 0.45 and 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 Multifactor Models

Chen, Roll, and Ross (1986) consider a multifactor model for stock returns. The factors used consist of *unexpected changes* or *surprises* of macroeconomic variables. Here unexpected changes denote the residuals of the macroeconomic variables after removing their dynamic dependence. A simple way to obtain unexpected changes is to fit a VAR model of Chapter 8 to the macroeconomic variables. For illustration, we consider the following two monthly macroeconomic variables:

1. Consumer price index (CPI) for all urban consumers: all items and with index 1982−1984 = 100.
2. Civilian employment numbers 16 years and over (CE16): measured in thousands.

Both CPI and CE16 series are seasonally adjusted, and the data span is from January 1975 to December 2003. We use a longer period to obtain the surprise series of the variables. For both series, we construct the growth rate series by taking the first difference of the logged data. The growth rates are in percentages.

To obtain the surprise series, we use the BIC criterion to identify a VAR(3) model. Thus, the two macroeconomic factors used in the factor model are the residuals of a VAR(3) model from 1990 to 2003. For the excess returns, we use the same 13 stocks as before. Details of the analysis follow:

```
> 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))
```

Figure 9.2 shows the bar plots of the beta estimates and R^2 for the 13 stocks. It is interesting to see that all excess returns are negatively related to the unexpected changes of CPI growth rate. This seems reasonable. However, the R^2 of all excess returns are low, indicating that the two macroeconomic variables used have very little explanatory power in understanding the excess returns of the 13 stocks.

The estimated covariance and correlation matrices of the two-factor model can be obtained using the following:

```
> 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)
```

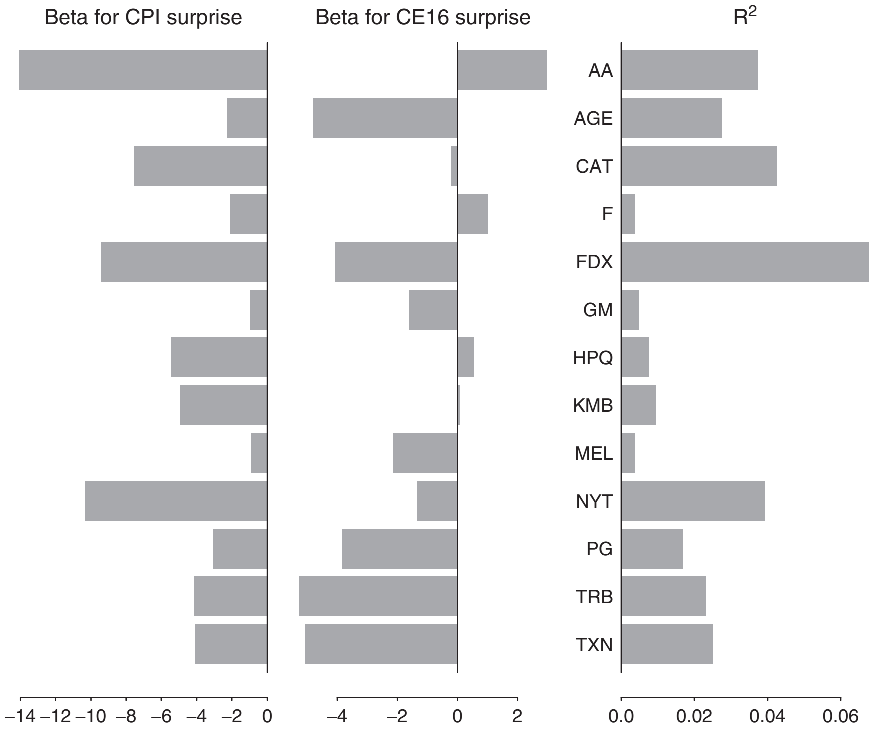


Figure 9.2 Bar plots of betas and R^2 for fitting two-factor model to monthly excess returns of 13 stocks. Sample period is from January 1990 to December 2003.

The correlation matrix is very close to the identity matrix, indicating that the two-factor model used does not fit the excess returns well. Finally, the correlation matrix of the residuals of the two-factor model is given by the following:

```
> 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)
```

As expected, this correlation matrix is close to that of the original excess returns given before and is omitted.

9.3 FUNDAMENTAL FACTOR MODELS

Fundamental factor models use observable asset specific fundamentals such as industrial classification, market capitalization, book value, and style classification (growth or value) to construct common factors that explain the excess returns.

There are two approaches to fundamental factor models available in the literature. The first approach is proposed by Bar Rosenberg, founder of BARRA Inc., and is referred to as the BARRA approach; see Grinold and Kahn (2000). In contrast to the macroeconomic factor models, this approach treats the observed asset specific fundamentals as the factor betas, $\boldsymbol{\beta}_i$, and estimates the factors $\boldsymbol{f}_t$ at each time index t via regression methods. The betas are time invariant, but the realizations $\boldsymbol{f}_t$ evolve over time. The second approach is the Fama–French approach proposed by Fama and French (1992). In this approach, the factor realization f_{jt} for a given specific fundamental is obtained by constructing some hedge portfolio based on the observed fundamental. We briefly discuss the two approaches in the next two sections.

9.3.1 BARRA Factor Model

Assume that the excess returns and, hence, the factor realizations are mean corrected. At each time index t, the factor model in Eq. (9.2) reduces to

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

where $\tilde{\boldsymbol{r}}_t$ denotes the (sample) mean-corrected excess returns and, for simplicity in notation, we continue to use $\boldsymbol{f}_t$ as factor realizations. Since $\boldsymbol{\beta}$ is given, the model in Eq. (9.6) is a multiple linear regression with k observations and m unknowns. Because the number of common factors m should be less than the number of assets k, the regression is estimable. However, the regression is not homogeneous because the covariance matrix of $\boldsymbol{\epsilon}_t$ is $\boldsymbol{D} = \text{diag}\{\sigma_1^2, \ldots, \sigma_k^2\}$ with $\sigma_i^2 = \text{Var}(\epsilon_{it})$, which depends on the ith asset. Consequently, the factor realization at time index t can be estimated by the *weighted least-squares* (WLS) method using the standard errors of the specific factors as the weights. The resulting estimate is

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

In practice, the covariance matrix $\boldsymbol{D}$ is unknown so that we use a two-step procedure to perform the estimation.

In step one, the ordinary least-squares (OLS) method is used at each time index t to obtain a preliminary estimate of $\boldsymbol{f}_t$ as

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

where the second subscript o is used to denote the OLS estimate. This estimate of factor realization is consistent, but not efficient. The residual of the OLS regression is

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

Since the residual covariance matrix is time invariant, we can pool the residuals together (for $t = 1, \ldots, T$) to obtain an estimate of $\boldsymbol{D}$ as

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

In step two, we plug in the estimate $\widehat{\boldsymbol{D}}_o$ to obtain a refined estimate of the factor realization

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

where the second subscript g denotes the *generalized least-squares* (GLS) estimate, which is a sample version of the WLS estimate. The residual of the refined regression is

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

from which we estimate the residual variance matrix as

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

Finally, the covariance matrix of the estimated factor realizations is

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

where

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

From Eq. (9.6), the covariance matrix of the excess returns under the BARRA approach is

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

Industry Factor Model

For illustration, we consider monthly excess returns of 10 stocks and use industrial classification as the specific asset fundamental. The stocks used are given in Table 9.2 and can be classified into three industrial sectors—namely, financial services, computer and high-tech industry, and other. The sample period is again from January 1990 to December 2003. Under the BARRA framework,

TABLE 9.2 Stocks Used and Their Tick Symbols in Analysis of Industrial Factor Model[a]

Tick	Company	$\bar{r}(\sigma_r)$	Tick	Company	$\bar{r}(\sigma_r)$
AGE	A.G. Edwards	1.36(10.2)	HPQ	Hewlett-Packard	1.37(11.8)
C	Citigroup	2.08(9.60)	IBM	Int. Bus. Machines	1.06(9.47)
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)

[a] Sample mean and standard deviation of the excess returns are also given. The sample span is from January 1990 to December 2003.

there are three common factors representing the three industrial sectors and the betas are indicators for the three industrial sectors; that is,

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

with the betas being

$$\beta_{ij} = \left\{ \begin{array}{ll} 1 & \text{if asset } i \text{ belongs to the } j \text{ industrial sector,} \\ 0 & \text{otherwise,} \end{array} \right\} \tag{9.10}$$

where $j = 1, 2, 3$ representing the financial, high-tech, and other sectors, respectively. For instance, the beta vector for the IBM stock return is $\boldsymbol{\beta}_i = (0, 1, 0)'$ and that for Alcoa stock return is $\boldsymbol{\beta}_i = (0, 0, 1)'$.

In Eq. (9.9), f_{1t} is the factor realization of the *financial services* sector, f_{2t} is that of the *computer and high-tech* sector, and f_{3t} is for the other sector. Because the β_{ij} are indicator variables, the OLS estimate of $\boldsymbol{f}_t$ is extremely simple. Indeed, $\boldsymbol{f}_t$ is the vector consisting of the averages of sector excess returns at time t. Specifically,

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

The specific factor of the ith asset is simply the deviation of its excess return from its industrial sample average. One can then obtain an estimate of the residual variance matrix $\boldsymbol{D}$ to perform the generalized least-squares estimation. We use S-Plus to perform the analysis. The commands also apply to R. First, load the returns into S-Plus, remove the sample means, create the industrial dummies, and compute the sample correlation matrix of the returns.

```
> 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
```

The OLS estimates, their residuals, and residual variances are estimated as follows:

```
> 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)
```

One can then obtain the generalized least-squares estimates.

```
> 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
```

The model-based correlations of stocks within an industrial sector are larger than their sample counterparts. For instance, the sample correlation between CAT and PG stock returns is only 0.1, but the correlation based on the fitted model is 0.6. Finally, Figure 9.3 shows the time plots of the factor realizations based on the generalized least-squares estimation.

Factor Mimicking Portfolio

Consider the special case of BARRA factor models with a single factor. Here the WLS estimate of f_t in Eq. (9.7) has a nice interpretation. Consider a portfolio $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_k)'$ of the k assets that solves

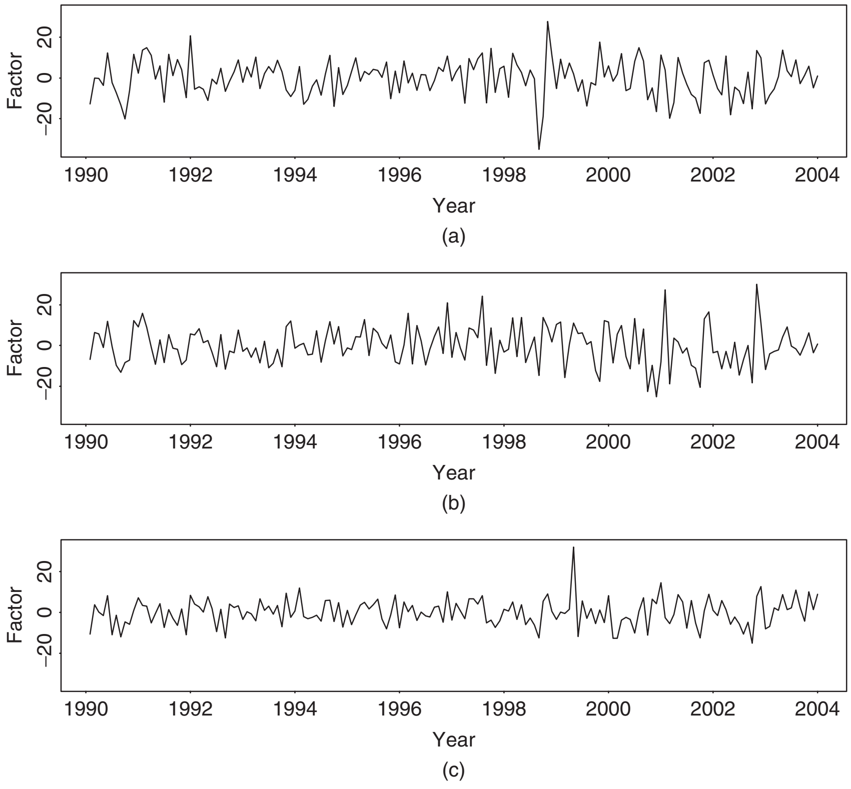
$$\min_{\boldsymbol{\omega}}(\tfrac{1}{2}\boldsymbol{\omega}'\boldsymbol{D}\boldsymbol{\omega}) \quad \text{such that} \quad \boldsymbol{\omega}'\boldsymbol{\beta} = 1.$$

It turns out that the solution to this portfolio problem is given by

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

Thus, the estimated factor realization is the portfolio return

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



(a)

(b)

(c)

Figure 9.3 Estimated factor realizations of BARRA industrial factor model for 10 monthly stock returns in 3 industrial sectors: (a) factor realizations: financial sector, (b) high-tech sector, and (c) other sector.

If the portfolio $\boldsymbol{\omega}$ is normalized such that $\sum_{i=1}^{k} \omega_i = 1$, it is referred to as a *factor mimicking portfolio*. For multiple factors, one can apply the idea to each factor individually.

Remark. In practice, the sample mean of an excess return is often not significantly different from zero. Thus, one may not need to remove the sample mean before fitting a BARRA factor model. □

9.3.2 Fama–French Approach

For a given asset fundamental (e.g., ratio of book-to-market value), Fama and French (1992) determined factor realizations using a two-step procedure. First, they sorted the assets based on the values of the observed fundamental. Then they formed a hedge portfolio, which is long in the top quintile ($\frac{1}{3}$) of the sorted assets and short in the bottom quintile of the sorted assets. The observed return on this hedge portfolio at time t is the observed factor realization for the given asset fundamental.

The procedure is repeated for each asset fundamental under consideration. Finally, given the observed factor realizations $\{\boldsymbol{f}_t | t = 1, \ldots, T\}$, the betas for each asset are estimated using a time series regression method. These authors identify three observed fundamentals that explain high percentages of variability in excess returns. The three fundamentals used by Fama and French are (a) the overall market return (market excess return), (b) the performance of small stocks relative to large stocks (SMB, small minus big), and (c) the performance of value stocks relative to growth stocks (HML, high minus low). The size sorted by market equity and the ratio of book equity to market equity is used to define value and growth stocks with value stocks having high book equity to market equity ratio.

Remark. The concepts of *factor* may differ between factor models. The *three factors* used in the Fama–French approach are three financial fundamentals. One can combine the fundamentals to create a new *attribute* of the stocks and refer to the resulting model as a single-factor model. This is particularly so because the model used is a linear statistical model. Thus, care must be exercised when one refers to the number of factors in a factor model. On the other hand, the number of factors is more well defined in statistical factor models, which we discuss next. □

9.4 PRINCIPAL COMPONENT ANALYSIS

An important topic in multivariate time series analysis is the study of the covariance (or correlation) structure of the series. For example, the covariance structure of a vector return series plays an important role in portfolio selection. In what follows, we discuss some statistical methods useful in studying the covariance structure of a vector time series.

Given a k-dimensional random variable $\boldsymbol{r} = (r_1, \ldots, r_k)'$ with covariance matrix $\boldsymbol{\Sigma}_r$, a *principal component analysis* (PCA) is concerned with using a few linear combinations of r_i to explain the structure of $\boldsymbol{\Sigma}_r$. If $\boldsymbol{r}$ denotes the monthly log returns of k assets, then PCA can be used to study the main source of variations of these k asset returns. Here the keyword is *few* so that simplification can be achieved in multivariate analysis.

9.4.1 Theory of PCA

Principal component analysis applies to either the covariance matrix $\boldsymbol{\Sigma}_r$ or the correlation matrix $\boldsymbol{\rho}_r$ of $\boldsymbol{r}$. Since the correlation matrix is the covariance matrix of the standardized random vector $\boldsymbol{r}^* = \boldsymbol{S}^{-1}\boldsymbol{r}$, where $\boldsymbol{S}$ is the diagonal matrix of standard deviations of the components of $\boldsymbol{r}$, we use covariance matrix in our theoretical discussion. Let $\boldsymbol{w}_i = (w_{i1}, \ldots, w_{ik})'$ be a k-dimensional real-valued vector, where $i = 1, \ldots, k$. Then

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

is a linear combination of the random vector $\boldsymbol{r}$. If $\boldsymbol{r}$ consists of the simple returns of k stocks, then y_i is the return of a portfolio that assigns weight w_{ij} to the jth stock. Since multiplying a constant to $\boldsymbol{w}_i$ does not affect the proportion of allocation assigned to the jth stock, we standardize the vector $\boldsymbol{w}_i$ so that $\boldsymbol{w}_i'\boldsymbol{w}_i = \sum_{j=1}^k w_{ij}^2 = 1$.

Using properties of a linear combination of random variables, we have

$$\text{Var}(y_i) = \boldsymbol{w}_i'\boldsymbol{\Sigma}_r\boldsymbol{w}_i, \qquad i = 1, \ldots, k, \tag{9.11}$$

$$\text{Cov}(y_i, y_j) = \boldsymbol{w}_i'\boldsymbol{\Sigma}_r\boldsymbol{w}_j, \qquad i, j = 1, \ldots, k. \tag{9.12}$$

The idea of PCA is to find linear combinations $\boldsymbol{w}_i$ such that y_i and y_j are uncorrelated for $i \neq j$ and the variances of y_i are as large as possible. More specifically:

1. The first principal component of $\boldsymbol{r}$ is the linear combination $y_1 = \boldsymbol{w}_1'\boldsymbol{r}$ that maximizes Var(y_1) subject to the constraint $\boldsymbol{w}_1'\boldsymbol{w}_1 = 1$.
2. The second principal component of $\boldsymbol{r}$ is the linear combination $y_2 = \boldsymbol{w}_2'\boldsymbol{r}$ that maximizes Var(y_2) subject to the constraints $\boldsymbol{w}_2'\boldsymbol{w}_2 = 1$ and Cov$(y_2, y_1) = 0$.
3. The ith principal component of $\boldsymbol{r}$ is the linear combination $y_i = \boldsymbol{w}_i'\boldsymbol{r}$ that maximizes Var(y_i) subject to the constraints $\boldsymbol{w}_i'\boldsymbol{w}_i = 1$ and Cov$(y_i, y_j) = 0$ for $j = 1, \ldots, i - 1$.

Since the covariance matrix $\boldsymbol{\Sigma}_r$ is nonnegative definite, it has a spectral decomposition; see Appendix A of Chapter 8. Let $(\lambda_1, \boldsymbol{e}_1)$, ..., $(\lambda_k, \boldsymbol{e}_k)$ be the eigenvalue–eigenvector pairs of $\boldsymbol{\Sigma}_r$, where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k \geq 0$ and $\boldsymbol{e}_i = (e_{i1}, \ldots, e_{ik})'$, which is properly normalized. We have the following statistical result.

Result 9.1. The ith principal component of $\boldsymbol{r}$ is $y_i = \boldsymbol{e}_i'\boldsymbol{r} = \sum_{j=1}^k e_{ij}r_j$ for $i = 1, \ldots, k$. Moreover,

$$\text{Var}(y_i) = \boldsymbol{e}_i'\boldsymbol{\Sigma}_r\boldsymbol{e}_i = \lambda_i, \qquad i = 1, \ldots, k,$$

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

If some eigenvalues λ_i are equal, the choices of the corresponding eigenvectors $\boldsymbol{e}_i$ and hence y_i are not unique. In addition, we have

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

The result of Eq. (9.13) says that

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

Consequently, the proportion of total variance in $\boldsymbol{r}$ explained by the ith principal component is simply the ratio between the ith eigenvalue and the sum of all eigenvalues of $\boldsymbol{\Sigma}_r$. One can also compute the cumulative proportion of total variance explained by the first i principal components [i.e., $(\sum_{j=1}^{i}\lambda_j)/(\sum_{j=1}^{k}\lambda_j)$]. In practice, one selects a small i such that the resulting cumulative proportion is large.

Since $\text{tr}(\boldsymbol{\rho}_r) = k$, the proportion of variance explained by the ith principal component becomes λ_i/k when the correlation matrix is used to perform the PCA.

A by-product of the PCA is that a zero eigenvalue of $\boldsymbol{\Sigma}_r$, or $\boldsymbol{\rho}_r$, indicates the existence of an *exact* linear relationship between the components of $\boldsymbol{r}$. For instance, if the smallest eigenvalue $\lambda_k = 0$, then by Result 9.1 $\text{Var}(y_k) = 0$. Therefore, $y_k = \sum_{j=1}^{k} e_{kj} r_j$ is a constant and there are only $k-1$ random quantities in $\boldsymbol{r}$. In this case, the dimension of $\boldsymbol{r}$ can be reduced. For this reason, PCA has been used in the literature as a tool for dimension reduction.

9.4.2 Empirical PCA

In application, the covariance matrix $\boldsymbol{\Sigma}_r$ and the correlation matrix $\boldsymbol{\rho}_r$ of the return vector $\boldsymbol{r}$ are unknown, but they can be estimated consistently by the sample covariance and correlation matrices under some regularity conditions. Assuming that the returns are weakly stationary and the data consist of $\{\boldsymbol{r}_t | t = 1, \ldots, T\}$, we have the following estimates:

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

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

where $\widehat{\boldsymbol{S}} = \text{diag}\{\sqrt{\hat{\sigma}_{11,r}}, \ldots, \sqrt{\hat{\sigma}_{kk,r}}\}$ is the diagonal matrix of sample standard errors of $\boldsymbol{r}_t$. Methods to compute eigenvalues and eigenvectors of a symmetric matrix can then be used to perform the PCA. Most statistical packages now have the capability to perform principal component analysis. In R and S-Plus, the basic command of PCA is princomp, and in FinMetrics the command is mfactor.

Example 9.1. Consider the monthly log stock returns of International Business Machines, Hewlett-Packard, Intel Corporation, J.P. Morgan Chase, and Bank of America from January 1990 to December 2008. The returns are in percentages and include dividends. The data set has 228 observations. Figure 9.4 shows the time plots of these five monthly return series. As expected, returns of companies in the same industrial sector tend to exhibit similar patterns.

Denote the returns by $\boldsymbol{r}' = $ (IBM, HPQ, INTC, JPM, BAC). The sample mean vector of the returns is $(0.70, 0.99, 1.20, 0.82, 0.41)'$ and the sample covariance and correlation matrices are

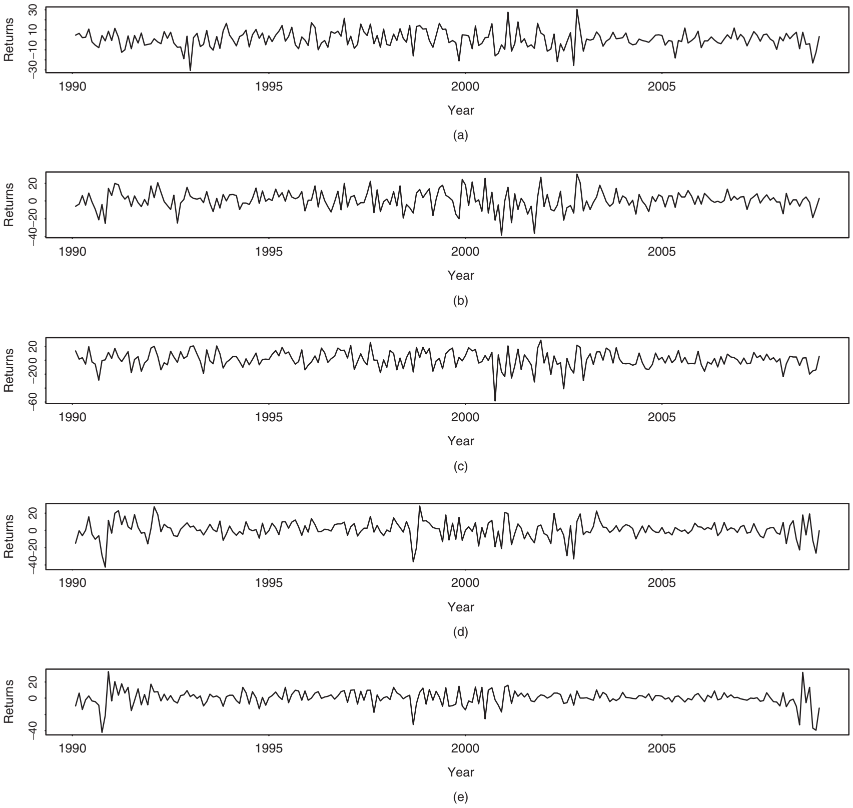


Figure 9.4 Time plots of monthly log stock returns in percentages and including dividends for (a) International Business Machines, (b) Hewlett-Packard, (c) Intel, (d) J.P. Morgan Chase, and (e) Bank of America from January 1990 to December 2008.

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

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

Table 9.3 gives the results of PCA using both the covariance and correlation matrices. Also given are eigenvalues, eigenvectors, and proportions of variabilities

TABLE 9.3 Results of Principal Component Analysis for Monthly Log Returns, Including Dividends of Stocks of IBM, Hewlett-Packard, Intel, J.P. Morgan Chase, and Bank of America from January 1990 to December 2008[a]

Using Sample Covariance Matrix					
Eigenvalue	284.17	112.93	57.43	46.81	29.87
Proportion	0.535	0.213	0.108	0.088	0.056
Cumulative	0.535	0.748	0.856	0.944	1.000
Eigenvector	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
Using Sample Correlation Matrix					
Eigenvalue	2.607	1.072	0.569	0.451	0.301
Proportion	0.522	0.214	0.114	0.090	0.060
Cumulative	0.522	0.736	0.850	0.940	1.000
Eigenvector	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] The eigenvectors are in columns.

explained by the principal components. Consider the correlation matrix and denote the sample eigenvalues and eigenvectors by $\hat{\lambda}_i$ and $\widehat{\boldsymbol{e}}_i$. We have

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

for the first two principal components. These two components explain about 74% of the total variability of the data, and they have interesting interpretations. The first component is a roughly equally weighted linear combination of the stock returns. This component might represent the general movement of the stock market and hence is a *market component*. The second component represents the difference between the two industrial sectors—namely, technologies versus financial services. It might be an *industrial component*. Similar interpretations of principal components can also be found by using the covariance matrix of $\boldsymbol{r}$.

An informal but useful procedure to determine the number of principal components needed in an application is to examine the *scree plot*, which is the time plot of the eigenvalues $\hat{\lambda}_i$ ordered from the largest to the smallest (i.e., a plot of $\hat{\lambda}_i$ versus i). Figure 9.5(a) shows the scree plot for the five stock returns of Example 9.1. By looking for an elbow in the scree plot, indicating that the remaining eigenvalues

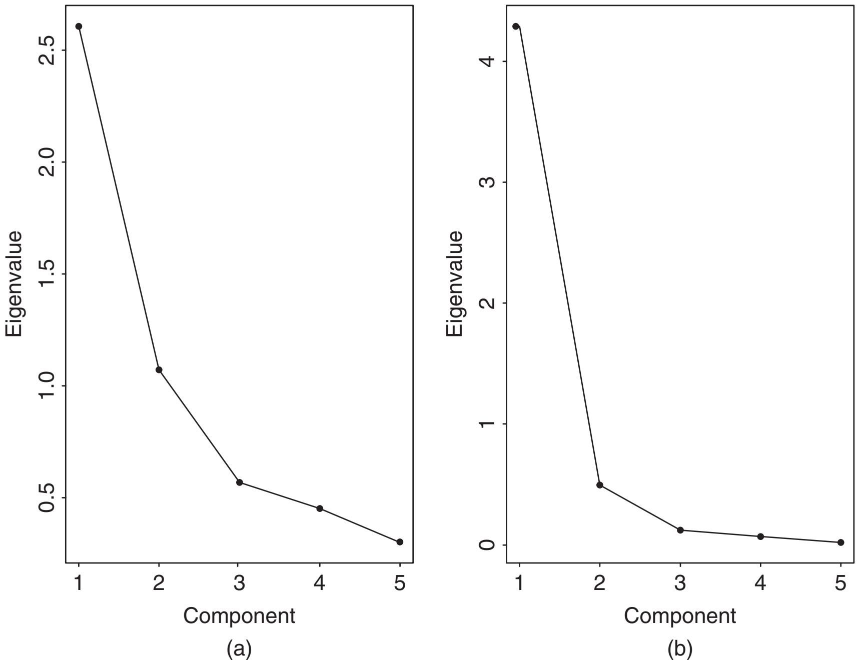


Figure 9.5 Scree plots for two 5-dimensional asset returns: (a) series of Example 9.1 and (b) bond index returns of Example 9.3.

are relatively small and all about the same size, one can determine the appropriate number of components. For both plots in Figure 9.5, two components appear to be appropriate. Finally, except for the case in which $\lambda_j = 0$ for $j > i$, selecting the first i principal components only provides an approximation to the total variance of the data. If a small i can provide a good approximation, then the simplification becomes valuable.

Remark. The R and S-Plus commands used to perform the PCA are given below. The command `princomp` gives the square root of the eigenvalue and denotes it as standard deviation.

```
> 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 STATISTICAL FACTOR ANALYSIS

We now turn to statistical factor analysis. One of the main difficulties in multivariate statistical analysis is the "curse of dimensionality." For serially correlated data, the number of parameters of a parametric model often increases dramatically when the order of the model or the dimension of the time series is increased. Simplifying methods are often sought to overcome the curse of dimensionality. From an empirical viewpoint, multivariate data often exhibit similar patterns indicating the existence of common structure hidden in the data. Statistical factor analysis is one of those simplifying methods available in the literature. The aim of statistical factor analysis is to identify, from the observed data, a few factors that can account for most of the variations in the covariance or correlation matrix of the data.

Traditional statistical factor analysis assumes that the data have no serial correlations. This assumption is often violated by financial data taken with frequency less than or equal to a week. However, the assumption appears to be reasonable for asset returns with lower frequencies (e.g., monthly returns of stocks or market indexes). If the assumption is violated, then one can use the parametric models discussed in this book to remove the linear dynamic dependence of the data and apply factor analysis to the residual series.

In what follows, we discuss statistical factor analysis based on the *orthogonal factor model*. Consider the return $\boldsymbol{r}_t = (r_{1t}, \ldots, r_{kt})'$ of k assets at time period t and assume that the return series $\boldsymbol{r}_t$ is weakly stationary with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}_r$. The statistical factor model postulates that $\boldsymbol{r}_t$ is linearly dependent on a few *unobservable* random variables $\boldsymbol{f}_t = (f_{1t}, \ldots, f_{mt})'$ and k additional noises $\boldsymbol{\epsilon}_t = (\epsilon_{1t}, \ldots, \epsilon_{kt})'$. Here $m < k$, f_{it} are the common factors, and ϵ_{it} are the errors. Mathematically, the statistical factor model is also in the form of Eq. (9.1) except that the intercept $\boldsymbol{\alpha}$ is replaced by the mean return $\boldsymbol{\mu}$. Thus, a statistical factor model is in the form

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

where $\boldsymbol{\beta} = [\beta_{ij}]_{k\times m}$ is the *matrix of factor loadings*, β_{ij} is the loading of the ith variable on the jth factor, and ϵ_{it} is the *specific error* of r_{it}. A key feature of the statistical factor model is that the m factors f_{it} and the factor loadings β_{ij} are *unobservable*. As such, Eq. (9.16) is not a multivariate linear regression model, even though it has a similar appearance. This special feature also distinguishes a statistical factor model from other factor models discussed earlier.

The factor model in Eq. (9.16) is an orthogonal factor model if it satisfies the following assumptions:

1. $E(\boldsymbol{f}_t) = \boldsymbol{0}$ and $\text{Cov}(\boldsymbol{f}_t) = \boldsymbol{I}_m$, the $m \times m$ identity matrix.
2. $E(\boldsymbol{\epsilon}_t) = \boldsymbol{0}$ and $\text{Cov}(\boldsymbol{\epsilon}_t) = \boldsymbol{D} = \text{diag}\{\sigma_1^2, \ldots, \sigma_k^2\}$ (i.e., $\boldsymbol{D}$ is a $k \times k$ diagonal matrix).
3. $\boldsymbol{f}_t$ and $\boldsymbol{\epsilon}_t$ are independent so that $\text{Cov}(\boldsymbol{f}_t, \boldsymbol{\epsilon}_t) = E(\boldsymbol{f}_t \boldsymbol{\epsilon}_t') = \boldsymbol{0}_{m\times k}$.

Under the previous assumptions, it is easy to see that

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

and

$$\text{Cov}(\boldsymbol{r}_t, \boldsymbol{f}_t) = E[(\boldsymbol{r}_t - \boldsymbol{\mu})\boldsymbol{f}_t'] = \boldsymbol{\beta}E(\boldsymbol{f}_t\boldsymbol{f}_t') + E(\boldsymbol{\epsilon}_t\boldsymbol{f}_t') = \boldsymbol{\beta}. \qquad (9.18)$$

Using Eqs. (9.17) and (9.18), we see that for the orthogonal factor model in Eq. (9.16)

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

The quantity $\beta_{i1}^2 + \cdots + \beta_{im}^2$, which is the portion of the variance of r_{it} contributed by the m common factors, is called the *communality*. The remaining portion σ_i^2 of the variance of r_{it} is called the *uniqueness* or *specific variance*. Let $c_i^2 = \beta_{i1}^2 + \cdots + \beta_{im}^2$ be the communality, which is the sum of squares of the loadings of the ith variable on the m common factors. The variance of component r_{it} becomes $\text{Var}(r_{it}) = c_i^2 + \sigma_i^2$.

In practice, not every covariance matrix has an orthogonal factor representation. In other words, there exists a random variable $\boldsymbol{r}_t$ that does not have any orthogonal factor representation. Furthermore, the orthogonal factor representation of a random variable is not unique. In fact, for any $m \times m$ orthogonal matrix $\boldsymbol{P}$ satisfying $\boldsymbol{P}\boldsymbol{P}' = \boldsymbol{P}'\boldsymbol{P} = \boldsymbol{I}$, let $\boldsymbol{\beta}^* = \boldsymbol{\beta}\boldsymbol{P}$ and $\boldsymbol{f}_t^* = \boldsymbol{P}'\boldsymbol{f}_t$. Then

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

In addition, $E(\boldsymbol{f}_t^*) = \boldsymbol{0}$ and $\text{Cov}(\boldsymbol{f}_t^*) = \boldsymbol{P}'\text{Cov}(\boldsymbol{f}_t)\boldsymbol{P} = \boldsymbol{P}'\boldsymbol{P} = \boldsymbol{I}$. Thus, $\boldsymbol{\beta}^*$ and $\boldsymbol{f}_t^*$ form another orthogonal factor model for $\boldsymbol{r}_t$. This nonuniqueness of orthogonal factor representation is a weakness as well as an advantage for factor analysis. It is a weakness because it makes the meaning of factor loading arbitrary. It is an advantage because it allows us to perform rotations to find common factors that have nice interpretations. Because $\boldsymbol{P}$ is an orthogonal matrix, the transformation $\boldsymbol{f}_t^* = \boldsymbol{P}'\boldsymbol{f}_t$ is a rotation in the m-dimensional space.

9.5.1 Estimation

The orthogonal factor model in Eq. (9.16) can be estimated by two methods. The first estimation method uses the principal component analysis of the previous section. This method does not require the normality assumption of the data nor the prespecification of the number of common factors. It applies to both the covariance and correlation matrices. But as mentioned in PCA, the solution is often an approximation. The second estimation method is the maximum-likelihood method that uses normal density and requires a prespecification for the number of common factors.

Principal Component Method

Again let $(\hat{\lambda}_1, \widehat{\boldsymbol{e}}_1), \ldots, (\hat{\lambda}_k, \widehat{\boldsymbol{e}}_k)$ be pairs of the eigenvalues and eigenvectors of the sample covariance matrix $\widehat{\boldsymbol{\Sigma}}_r$, where $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_k$. Let $m < k$ be the number of common factors. Then the matrix of factor loadings is given by

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

The estimated specific variances are the diagonal elements of the matrix $\widehat{\boldsymbol{\Sigma}}_r - \widehat{\boldsymbol{\beta}}\widehat{\boldsymbol{\beta}}'$. That is, $\widehat{\boldsymbol{D}} = \text{diag}\{\hat{\sigma}_1^2, \ldots, \hat{\sigma}_k^2\}$, where $\hat{\sigma}_i^2 = \hat{\sigma}_{ii,r} - \sum_{j=1}^m \hat{\beta}_{ij}^2$, where $\hat{\sigma}_{ii,r}$ is the (i, i)th element of $\widehat{\boldsymbol{\Sigma}}_r$. The communalities are estimated by

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

The error matrix caused by approximation is

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

Ideally, we would like this matrix to be close to zero. It can be shown that the sum of squared elements of $\widehat{\boldsymbol{\Sigma}}_r - (\widehat{\boldsymbol{\beta}}\widehat{\boldsymbol{\beta}}' + \widehat{\boldsymbol{D}})$ is less than or equal to $\hat{\lambda}_{m+1}^2 + \cdots + \hat{\lambda}_k^2$. Therefore, the approximation error is bounded by the sum of squares of the neglected eigenvalues.

From the solution in Eq. (9.19), the estimated factor loadings based on the principal component method do not change as the number of common factors m is increased.

Maximum-Likelihood Method

If the common factors $\boldsymbol{f}_t$ and the specific factors $\boldsymbol{\epsilon}_t$ are jointly normal, then $\boldsymbol{r}_t$ is multivariate normal with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}_r = \boldsymbol{\beta}\boldsymbol{\beta}' + \boldsymbol{D}$. The maximum-likelihood method can then be used to obtain estimates of $\boldsymbol{\beta}$ and $\boldsymbol{D}$ under the constraint $\boldsymbol{\beta}'\boldsymbol{D}^{-1}\boldsymbol{\beta} = \boldsymbol{\Delta}$, which is a diagonal matrix. Here $\boldsymbol{\mu}$ is estimated by the sample mean. For more details of this method, readers are referred to Johnson and Wichern (2007).

In using the maximum-likelihood method, the number of common factors must be given a priori. In practice, one can use a modified likelihood ratio test to check the adequacy of a fitted m-factor model. The test statistic is

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

which, under the null hypothesis of m factors, is asymptotically distributed as a chi-squared distribution with $\frac{1}{2}[(k-m)^2 - k - m]$ degrees of freedom. We discuss some methods for selecting m in Section 9.6.1.

9.5.2 Factor Rotation

As mentioned before, for any $m \times m$ orthogonal matrix $\boldsymbol{P}$,

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

where $\boldsymbol{\beta}^* = \boldsymbol{\beta}\boldsymbol{P}$ and $\boldsymbol{f}_t^* = \boldsymbol{P}'\boldsymbol{f}_t$. In addition,

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

This result indicates that the communalities and the specific variances remain unchanged under an orthogonal transformation. It is then reasonable to find an orthogonal matrix $\boldsymbol{P}$ to transform the factor model so that the common factors have nice interpretations. Such a transformation is equivalent to rotating the common factors in the m-dimensional space. In fact, there are infinite possible factor rotations available. Kaiser (1958) proposes a *varimax* criterion to select the rotation that works well in many applications. Denote the rotated matrix of factor loadings by $\boldsymbol{\beta}^* = [\beta_{ij}^*]$ and the ith communality by c_i^2. Define $\tilde{\beta}_{ij}^* = \beta_{ij}^*/c_i$ to be the rotated coefficients scaled by the (positive) square root of communalities. The varimax procedure selects the orthogonal matrix $\boldsymbol{P}$ that maximizes the quantity

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

This complicated expression has a simple interpretation. Maximizing V corresponds to spreading out the squares of the loadings on each factor as much as possible. Consequently, the procedure is to find groups of large and negligible coefficients in any column of the rotated matrix of factor loadings. In a real application, factor rotation is used to aid the interpretations of common factors. It may be helpful in some applications, but not informative in others. There are many criteria available for factor rotation.

9.5.3 Applications

Given the data $\{\boldsymbol{r}_t\}$ of asset returns, the statistical factor analysis enables us to search for common factors that explain the variabilities of the returns. Since factor analysis assumes no serial correlations in the data, one should check the validity of this assumption before using factor analysis. The multivariate portmanteau statistics can be used for this purpose. If serial correlations are found, one can build a VARMA model to remove the dynamic dependence in the data and apply the factor analysis to the residual series. For many returns series, the correlation matrix of the residuals of a linear model is often very close to the correlation matrix of the original data. In this case, the effect of dynamic dependence on factor analysis is negligible.

We consider three examples in this section. The first and third examples use the R or S-Plus to perform the analysis and the second example uses Minitab. Other packages can also be used.

Example 9.2. Consider again the monthly log stock returns of IBM, Hewlett-Parkard, Intel, J.P. Morgan Chase, and Bank of America used in Example 9.1. To check the assumption of no serial correlations, we compute the portmanteau statistics and obtain $Q_5(1) = 39.99$, $Q_5(5) = 160.60$, and $Q_5(10) = 293.04$. Compared with chi-squared distributions with 25, 125, and 250 degrees of freedom, the p values of these test statistics are 0.029, 0.017, and 0.032, respectively. Therefore, there exists some minor serial dependence in the returns, but the dependence is not significant at the 1% level. For simplicity, we ignore the serial dependence in factor analysis.

Table 9.4 shows the results of factor analysis based on the correlation matrix using the maximum-likelihood method. We assume that the number of common factors is 2, which is reasonable according to the principal component analysis of Example 9.1. From the table, the factor analysis reveals several interesting findings:

- The two factors identified by the maximum-likelihood method explain about 60% of the variability of the stock returns.
- Based on the rotated factor loadings, the two common factors have some meaningful interpretations. The technology stocks (IBM, Hewlett-Packard,

TABLE 9.4 Factor Analysis of Monthly Log Stock Returns of IBM, Hewlett-Packard, Intel, J.P. Morgan Chase, and Bank of America[a]

	Estimates of Factor Loadings		Rotated Factor Loadings		Communalities
Variable	f_1	f_2	f_1^*	f_2^*	$1 - \sigma_i^2$
			Maximum-Likelihood Method		
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
Variance	1.801	1.193	1.535	1.459	2.994
Proportion	0.360	0.239	0.307	0.292	0.599

[a] The returns include dividends and are from January 1990 to December 2008. The analysis is based on the sample cross-correlation matrix and assumes two common factors.

and Intel) load heavily on the first factor, whereas the financial stocks (J.P. Morgan Chase and Bank of America) load highly on the second factor. These two rotated factors jointly differentiate the industrial sectors.

- In this particular instance, the varimax rotation seems to alter the ordering of the two common factors.
- The specific variance of IBM stock returns is relatively large, indicating that the stock has its own features that are worth further investigation.

Example 9.3. In this example, we consider the monthly log returns of U.S. bond indexes with maturities in 30 years, 20 years, 10 years, 5 years, and 1 year. The data are described in Example 8.2 but have been transformed into log returns. There are 696 observations. As shown in Example 8.2, there is serial dependence in the data. However, removing serial dependence by fitting a VARMA(2,1) model has hardly any effects on the concurrent correlation matrix. As a matter of fact, the correlation matrices before and after fitting a VARMA(2,1) model are

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

$$\widehat{\boldsymbol{\rho}} = \begin{bmatrix} 1.0 & & & & \\ 0.98 & 1.0 & & & \\ 0.92 & 0.92 & 1.0 & & \\ 0.85 & 0.86 & 0.90 & 1.0 & \\ 0.66 & 0.67 & 0.71 & 0.84 & 1.0 \end{bmatrix},$$

where $\widehat{\boldsymbol{\rho}}_o$ is the correlation matrix of the original log returns. Therefore, we apply factor analysis directly to the return series.

Table 9.5 shows the results of statistical factor analysis of the data. For both estimation methods, the first two common factors explain more than 90% of the total variability of the data. Indeed, the high communalities indicate that the specific variances are very small for the five bond index returns. Because the results of the two methods are close, we only discuss that of the principal component method. The unrotated factor loadings indicate that (a) all five return series load roughly equally on the first factor, and (b) the loadings on the second factor are positively correlated with the time to maturity. Therefore, the first common factor represents the general U.S. bond returns, and the second factor shows the "time-to-maturity" effect. Furthermore, the loadings of the second factor sum approximately to zero. Therefore, this common factor can also be interpreted as the contrast between long-term and short-term bonds. Here a long-term bond means one with maturity 10 years or longer. For the rotated factors, the loadings are also interesting. The loadings for the first rotated factor are proportional to the time to maturity, whereas the loadings of the second factor are inversely proportional to the time to maturity.

TABLE 9.5 Factor Analysis of Monthly Log Returns of U.S. Bond Indexes with Maturities in 30 Years, 20 Years, 10 Years, 5 Years, and 1 Year[a]

	Estimates of Factor Loadings		Rotated Factor Loadings		Communalities
Variable	f_1	f_2	f_1^*	f_2^*	$1 - \sigma_i^2$
		Principal Component Method			
30 years	0.952	0.253	0.927	0.333	0.970
20 years	0.954	0.240	0.922	0.345	0.968
10 years	0.956	0.140	0.866	0.429	0.934
5 years	0.955	−0.142	0.704	0.660	0.931
1 year	0.800	−0.585	0.325	0.936	0.982
Variance	4.281	0.504	3.059	1.726	4.785
Proportion	0.856	0.101	0.612	0.345	0.957
		Maximum-Likelihood Method			
30 years	0.849	−0.513	0.895	0.430	0.985
20 years	0.857	−0.486	0.876	0.451	0.970
10 years	0.896	−0.303	0.744	0.584	0.895
5 years	1.000	0.000	0.547	0.837	1.000
1 year	0.813	0.123	0.342	0.747	0.675
Variance	3.918	0.607	2.538	1.987	4.525
Proportion	0.784	0.121	0.508	0.397	0.905

[a] The data are from January 1942 to December 1999. The analysis is based on the sample cross-correlation matrix and assumes two common factors.

Example 9.4. Again, consider the monthly excess returns of the 10 stocks in Table 9.2. The sample span is from January 1990 to December 2003 and the returns are in percentages. Our goal here is to demonstrate the use of statistical factor models using the R or S-Plus command `factanal`. We started with a two-factor model, but it is rejected by the likelihood ratio test of Eq. (9.20). The test statistic is LR(2) $=$ 72.96. Based on the asymptotic χ^2_{26} distribution, p value of the test statistic is close to zero.

```
> 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"
```

We then applied a three-factor model that appears to be reasonable at the 5% significance level. The p value of the LR(3) statistic is 0.0892.

```
> 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
```

The factor loadings can also be shown graphically using

```
> plot(loadings(stat.fac))
```

and the plots are in Figure 9.6. From the plots, factor 1 represents essentially the financial service sector, and factor 2 mainly consists of the excess returns from the high-tech stocks and the Alcoa stock. Factor 3 depends heavily on excess returns of CAT and AA stocks and, hence, represents the remaining industrial sector.

Factor rotation can be obtained using the command rotate, which allows for many rotation methods, and factor realizations are available from the command predict.

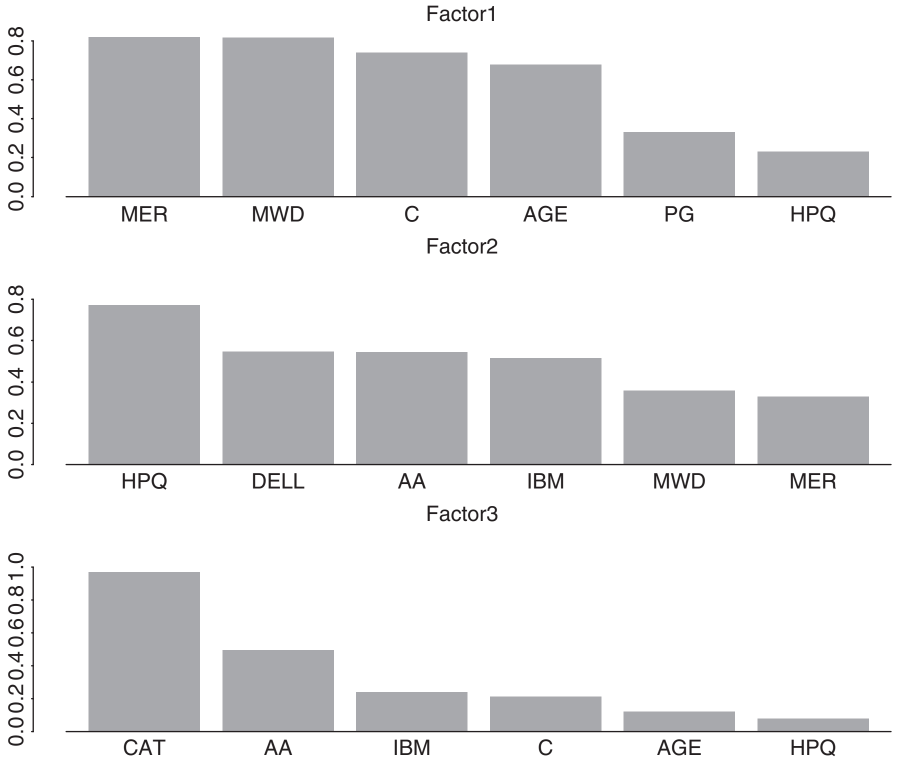


Figure 9.6 Plots of factor loadings when a 3-factor statistical factor model is fitted to 10 monthly excess stock returns in Table 9.2.

```
> 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')
```

Finally, we obtained the correlation matrix of the 10 excess returns based on the fitted three-factor statistical factor model. As expected, the correlations are closer to their sample counterparts than those of the industrial factor model in Section 9.3.1.

One can also use GMVP to compare the covariance matrices of the returns and the statistical factor model.

```
> 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
```

9.6 ASYMPTOTIC PRINCIPAL COMPONENT ANALYSIS

So far, our discussion of PCA assumes that the number of assets is smaller than the number of time periods, that is, $k < T$. To deal with situations of a small T and large k, Conner and Korajczyk (1986, 1988) developed the concept of *asymptotic principal component analysis* (APCA), which is similar to the traditional PCA but relies on the asymptotic results as the number of assets k increases to infinity. Thus, the APCA is based on eigenvalue–eigenvector analysis of the $T \times T$ matrix

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

where $\boldsymbol{1}_T$ is the T-dimensional vector of ones and $\bar{\boldsymbol{r}} = (\bar{r}_1, \ldots, \bar{r}_k)'$ with $\bar{r}_i = (\boldsymbol{1}_T'\boldsymbol{R}_i)/T$ being the sample mean of the ith return series. Conner and Korajczyk (1988) showed that as $k \to \infty$ eigenvalue–eigenvector analysis of $\widehat{\boldsymbol{\Omega}}_T$ is equivalent to the traditional statistical factor analysis. In other words, the APCA estimates of the factors $\boldsymbol{f}_t$ are the first m eigenvectors of $\widehat{\boldsymbol{\Omega}}_T$. Let $\widehat{\boldsymbol{F}}_t$ be the $m \times T$ matrix consisting of the first m eigenvectors of $\widehat{\boldsymbol{\Omega}}_T$. Then $\widehat{\boldsymbol{f}}_t$ is the tth column of $\widehat{\boldsymbol{F}}_t$. Using an idea similar to the estimation of BARRA factor models, Connor and Korajczyk (1988) propose refining the estimation of $\widehat{\boldsymbol{f}}_t$ as follows:

1. Use the sample covariance matrix $\widehat{\boldsymbol{\Omega}}_T$ to obtain an initial estimate of $\widehat{\boldsymbol{f}}_t$ for $t = 1, \ldots, T$.
2. For each asset, perform the OLS estimation of the model

$$r_{it} = \alpha_i + \boldsymbol{\beta}_i\widehat{\boldsymbol{f}}_t + \epsilon_{it}, \qquad t = 1, \ldots, T,$$

where $\boldsymbol{\beta}_i = (\beta_{i1}, \ldots, \beta_{im})$ and compute the residual variance $\hat{\sigma}_i^2$.

3. Form the diagonal matrix $\widehat{\boldsymbol{D}} = \text{diag}\{\hat{\sigma}_1^2, \ldots, \hat{\sigma}_k^2\}$ and rescale the returns as

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

4. Compute the $T \times T$ covariance matrix using $\boldsymbol{R}_*$ as

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

   where $\bar{\boldsymbol{r}}_*$ is the k-dimensional vector of the column means of $\boldsymbol{R}_*$, and perform eigenvalue–eigenvector analysis of $\widehat{\boldsymbol{\Omega}}_*$ to obtain a refined estimate of $\boldsymbol{f}_t$.

9.6.1 Selecting the Number of Factors

Two methods are available in the literature to help select the number of factors in factor analysis. The first method proposed by Connor and Korajczyk (1993) makes use of the idea that if m is the proper number of common factors, then there should be no significant decrease in the cross-sectional variance of the asset specific error ϵ_{it} when the number of factors moves from m to $m + 1$. The second method proposed by Bai and Ng (2002) adopts some information criteria to select the number of factors. This latter method is based on the observation that the eigenvalue–eigenvector analysis of $\widehat{\boldsymbol{\Omega}}_T$ solves the least-squares problem

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

Assume that there are m factors so that $\boldsymbol{f}_t$ is m-dimensional. Let $\hat{\sigma}_i^2(m)$ be the residual variance of the inner regression of the prior least-squares problem for asset i. This is done by using $\widehat{\boldsymbol{f}}_t$ obtained from the APCA analysis. Define the cross-sectional average of the residual variances as

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

The criteria proposed by Bai and Ng (2002) are

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

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

where M is a prespecified positive integer denoting the maximum number of factors and $P_{kT} = \min(\sqrt{k}, \sqrt{T})$. One selects m that minimizes either $C_{p1}(m)$ or $C_{p2}(m)$ for $0 \leq m \leq M$. In practice, the two criteria may select different numbers of factors.

9.6.2 An Example

To demonstrate asymptotic principal component analysis, we consider monthly simple returns of 40 stocks from January 2001 to December 2003 for 36 observations. Thus, we have $k = 40$ and $T = 36$. The tick symbols of stocks used are given in Table 9.6. These stocks are among those heavily traded on NASDAQ and the NYSE on a particular day of September 2004. The main S-Plus command used is mfactor.

To select the number of factors, we used the two methods discussed earlier. The Connor–Korajczyk method selects $m = 1$, whereas the Bai–Ng method uses $m = 6$. For the latter method, the two criteria provide different results.

```
> 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:
Cp1 and Cp2 did not yield same result. The smaller one
   is used.
> nf.bn$k
[1] 6
```

TABLE 9.6 Tick Symbols of Stocks Used in Asymptotic Principal Component Analysis for Sample Period from January 2001 to December 2003

Market	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

Using $m = 6$, we apply APCA to the returns. The scree plot and estimated factor returns can also be obtained.

```
> 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))
```

Figure 9.7 shows the scree plot of the APCA for the 40 stock returns. The 6 common factors used explain about 89.4% of the variability. Figure 9.8 gives the time plots of the returns of the 6 estimated factors.

EXERCISES

9.1. Consider the monthly simple excess returns, in percentages and including dividends, of 13 stocks and the S&P 500 composite index from January 1990 to December 2008. The monthly 3-month Treasury bill rate in the secondary market is used as the risk-free interest rate to compute the excess returns. The tick symbols for the stocks are AA, AXP, CAT, DE, F, FDX, HPQ, IBM, JNJ, KMB, MMM, PG, and WFC. The data are in the file `m-fac-ex-9008.txt`. Perform the *market model* analysis of Section 9.2.1 for the 13 stock returns to obtain the estimates of β_i, σ_i^2, and R^2 for each stock return series.

9.2. Consider the monthly log stock returns, in percentages and including dividends, of Merck & Company, Johnson & Johnson, General Electric, General Motors, Ford Motor Company, and value-weighted index from January 1960 to December 2008; see the file `m-mrk2vw.txt` of Exercise 8.1 of Chapter 8.

(a) Perform a principal component analysis of the data using the sample covariance matrix.

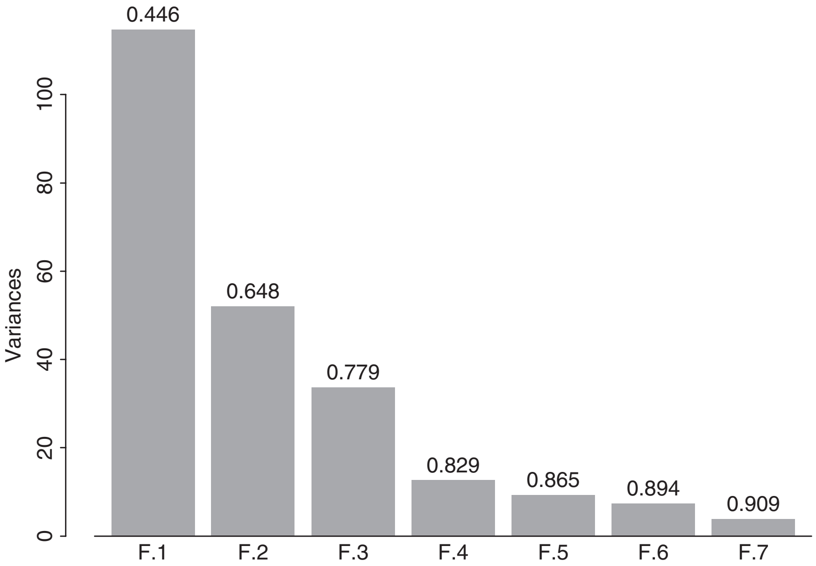


Figure 9.7 Scree plot of asymptotic principal component analysis applied to monthly simple returns of 40 stocks. Sample period is from January 2001 to December 2003.

(b) Perform a principal component analysis of the data using the sample correlation matrix.

(c) Perform a statistical factor analysis on the data. Identify the number of common factors. Obtain estimates of factor loadings using both the principal component and maximum-likelihood methods.

9.3. The file `m-excess-c10sp-9003.txt` contains the monthly simple excess returns of 10 stocks and the S&P 500 index. The 3-month Treasury bill rate on the secondary market is used to compute the excess returns. The sample period is from January 1990 to December 2003 for 168 observations. The 11 columns in the file contain the returns for ABT, LLY, MRK, PFE, F, GM, BP, CVX, RD, XOM, and SP5, respectively. Analyze the 10 stock excess returns using the single-factor market model. Plot the beta estimate and R^2 for each stock, and use the global minimum variance portfolio to compare the covariance matrices of the fitted model and the data.

9.4. Again, consider the 10 stock returns in `m-excess-c10sp-9003.txt`. The stocks are from companies in 3 industrial sectors. ABT, LLY, MRK, and PFE are major drug companies, F and GM are automobile companies, and the rest are big oil companies. Analyze the excess returns using the BARRA industrial factor model. Plot the 3-factor realizations and comment on the adequacy of the fitted model.

9.5. Again, consider the 10 excess stock returns in the file `m-excess-c10sp-9003.txt`. Perform a principal component analysis on the returns and obtain

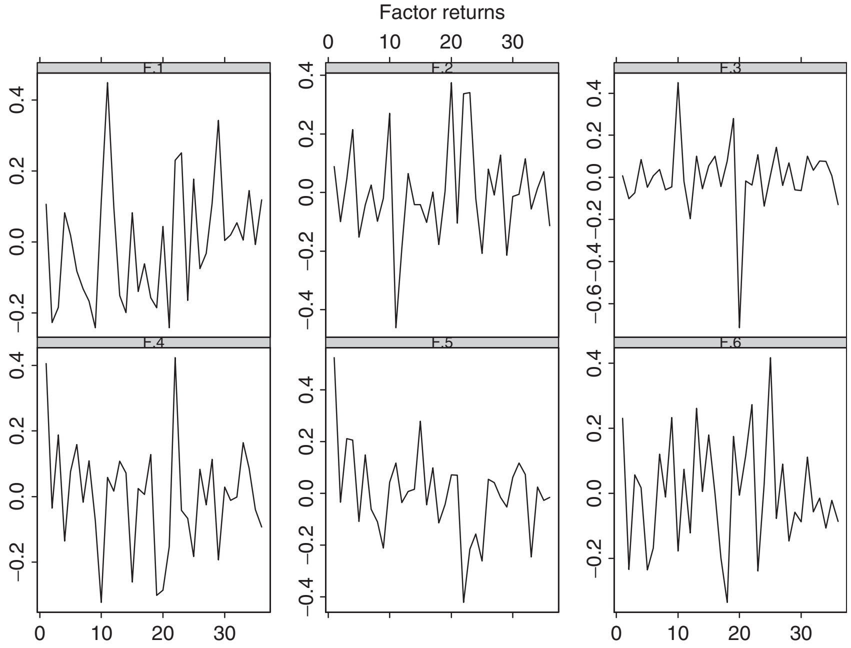


Figure 9.8 Time plots of factor returns derived from applying asymptotic principal component analysis to monthly simple returns of 40 stocks. Sample period is from January 2001 to December 2003.

the scree plot. How many common factors are there? Why? Interpret the common factors.

9.6. Again, consider the 10 excess stock returns in the file `m-excess-c10sp-9003.txt`. Perform a statistical factor analysis. How many common factors are there if the 5% significance level is used? Plot the estimated factor loadings of the fitted model. Are the common factors meaningful?

9.7. The file `m-fedip.txt` contains year, month, effective federal funds rate, and the industrial production index from July 1954 to December 2003. The industrial production index is seasonally adjusted. Use the federal funds rate and the industrial production index as the macroeconomic variables. Fit a macroeconomic factor model to the 10 excess returns in `m-excess-c10sp-9003.txt`. You can use a VAR model to obtain the surprise series of the macroeconomic variables. Comment on the fitted factor model.

REFERENCES

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.