Lecture 10
Principal Component Analysis and Factor Analysis

CHEN Ying
FE5209 Financial Econometrics

Outline

- Principal components analysis (PCA).
- Factor analysis.
- Factor models.

One of the primary goals of PCA and factor analysis is to identify a measurement model for a latent variable

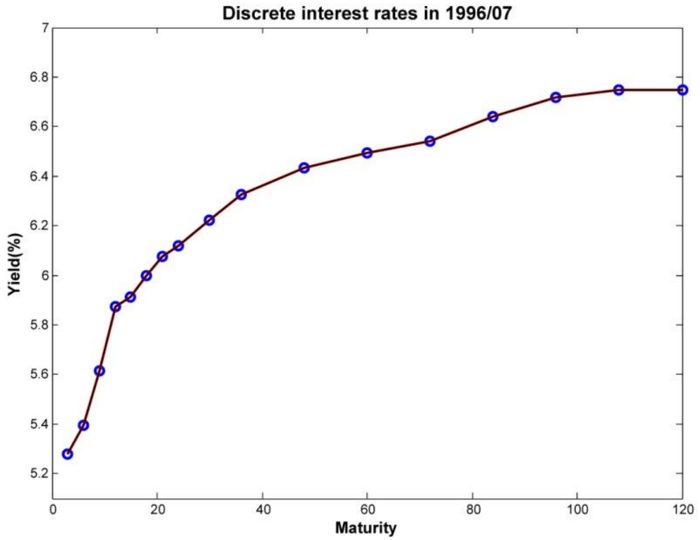
- identifying the items to include in the model
- identifying how many ‘factors’ there are in the latent variable
- identifying which items are “associated” with which factors

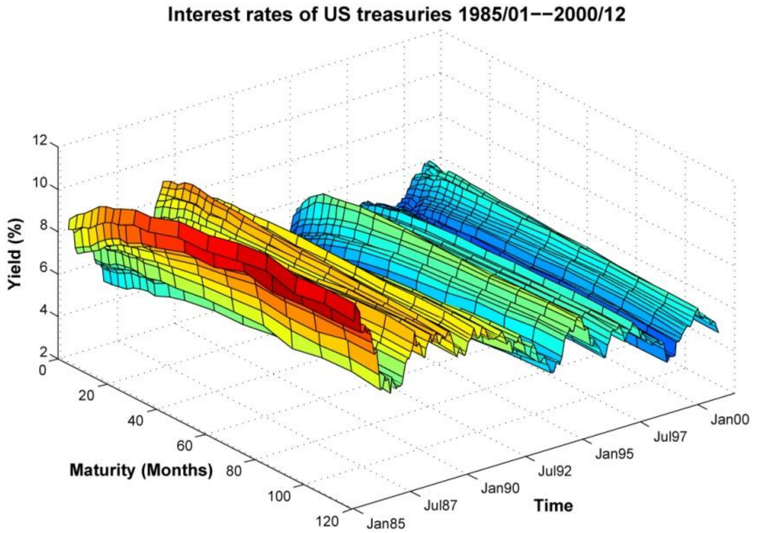
Readings
FE chapter 13
SDA chapter 17

Forecast yield curve

Interest rate is important in regulating and dealing with investment, inflation and unemployment. It is a primary input factor in economic and financial activities.

Yield Curve (or Term Structure) illustrates the relationship between interest rates (yields) and time to maturities.





The shape of yield curve reflects the market's expectation on monetary policy and economic conditions.

It is necessary to understand the dynamics of yield curve to price the interest rate derivatives.

But, we have too many variables (dimensions).

Why factor/component analysis?

Problem: we have too many variables (dimensions).

Solution: discover a new set of factors (combinations of variables) against which to represent, describe or evaluate the data

- Typically a smaller set of factors (*dimension reduction*). Can build more effective data analyses on the reduced-dimensional space: classification, clustering, pattern recognition.
- Better representation of data *without losing much information*. Observed data are described in terms of these factors rather than in terms of original variables/dimensions

Principal component analysis (PCA) convert a set of observations of *possibly correlated variables* into a set of values of linearly uncorrelated variables called **principal components.**

The new variables (PCs)

are linear combinations of the original ones

are uncorrelated with one another

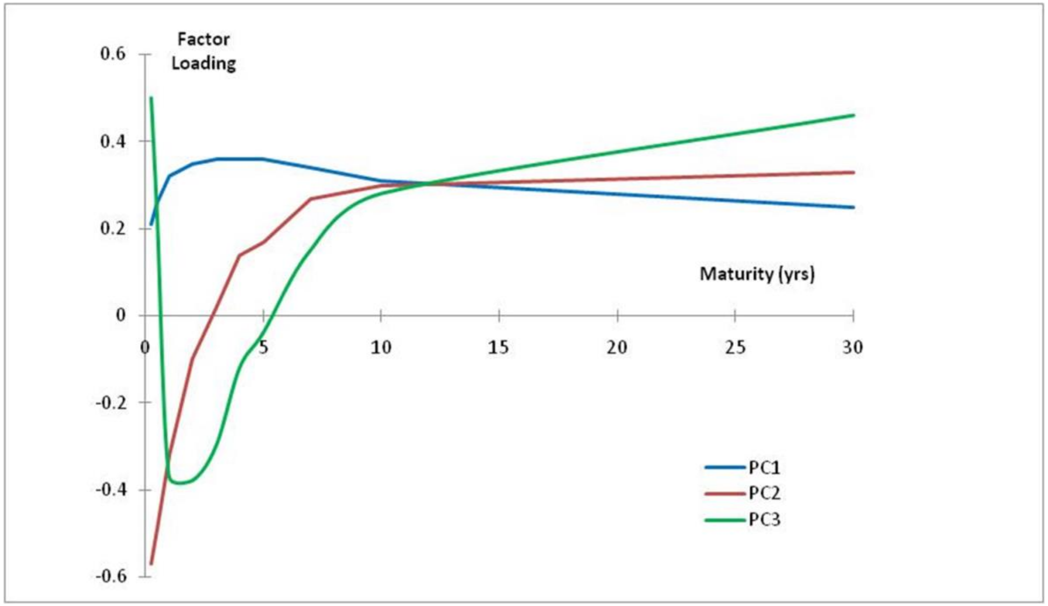
capture as much of the original variance in the data as possible

Principal components analysis

Interest rate movements are often highly correlated for different maturities. Principal components analysis attempts to define a set of factors that explains the movements in interest rates.

Example: 10 US Treasury rates with maturities between 3 months and 30 years. The period of analysis is 1989 to 1995 (1,543 daily observations).

Maturity	PC1	PC2	PC3	PC4	PC5
3m	0.21	-0.57	0.50	0.47	-0.39
6m	0.26	-0.49	0.23	-0.37	0.70
12m	0.32	-0.32	-0.37	-0.58	-0.52
2y	0.35	-0.10	-0.38	0.17	0.04
3y	0.36	0.02	-0.30	0.27	0.07
4y	0.36	0.14	-0.12	0.25	0.16
5y	0.36	0.17	-0.04	0.14	0.08
7y	0.34	0.27	0.15	0.01	0.00
10y	0.31	0.30	0.28	-0.10	-0.06
30y	0.25	0.33	0.46	-0.34	-0.18



*The first factor **PC1** represents a roughly parallel shift in the yield curve (**level**).*

*The second factor **PC2** corresponds to a "twist": Short-rates move in one direction and long-rates move in another (**slope**)*

*The third factor **PC3** corresponds to "bowing": Very short- and long-term rates move in one direction and mid-term rates move in another (**curvature**)*

Excursion: eigenvalues and eigenvectors

A square matrix A $(p \times p)$ has an eigenvector γ if $A\gamma = \lambda\gamma$ for some constant λ. We call λ the eigenvalue associated with the eigenvector γ.

For symmetric matrix A ($p \times p$ and $a_{ij} = a_{ji}$, e.g. covariance matrix) : Eigenvectors are *orthogonal and unit vectors*. For any eigenvectors γ_i and γ_j of matrix A, $\gamma_{\mathrm{i}}'\gamma_j = 0$ for $i \neq j$, and $\gamma_{\mathrm{j}}'\gamma_j = 1$ for all j.

Let $\Gamma = (\gamma_1, \gamma_2, \dots, \gamma_p)$ where γ_j are the eigenvectors of A. Since the eigenvectors are unit vectors and are orthogonal to each other, $\Gamma'\Gamma = \mathrm{I_p}$.

Let $\Gamma = (\gamma_1, \gamma_2, \dots, \gamma_p)$ and $\Lambda = diag(\lambda_1, \lambda_2, \dots, \lambda_p) = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & 0 & 0 & \lambda_p \end{pmatrix}$

where λ_j and γ_j be the j-th eigenvalue and its corresponding eigenvector of A (i.e. $A\gamma_j = \lambda_j\gamma_j$). Then $A\Gamma = \Gamma\Lambda$. Hence $A = \Gamma\Lambda\Gamma'$ and $\Gamma'A\Gamma = \Lambda$.

Remark: $A = \Gamma\Lambda\Gamma'$ ➔ $A^b = \Gamma\Lambda^{\mathrm{b}}\Gamma'$ where $\Lambda^{\mathrm{b}} = diag(\lambda_1^b, \lambda_2^b, \dots, \lambda_p^b)$.

Excursion: eigenvalues and eigenvectors

Theorem.
Let A be a square matrix of order p. If λ is an eigenvalue of A, then:

1. λ^m is an eigenvalue of A^m, for $m = 1,2, \ldots$
2. If A is invertible, then $\frac{1}{\lambda}$ is an eigenvalue of A^{-1}.
3. A is not invertible if and only if $\lambda = 0$ is an eigenvalue of A.
4. If α is any number, then $\lambda + \alpha$ is an eigenvalue of $A + \alpha I_p$.

Excursion: compute eigenvalues

$$A\gamma = \lambda\gamma \Rightarrow (A - \lambda I_p)\gamma = 0.$$

this can only have a non-zero solution for γ if $|A - \lambda I_p| = 0$, where $|A - \lambda I_p|$ is the determinant of $A - \lambda I_p$.

$|A - \lambda I_p|$ is called the characteristic equation of A.

Example: Consider the matrix $A = \begin{pmatrix} 4.6 & -1.2 \\ -1.2 & 1.4 \end{pmatrix}$

Its eigenvalues can be found as follows

$$\begin{pmatrix} 4.6 & -1.2 \\ -1.2 & 1.4 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} = \lambda\begin{pmatrix} x \\ y \end{pmatrix} \Rightarrow \begin{matrix} 4.6x - 1.2y = \lambda x \\ -1.2x + 1.4y = \lambda y \end{matrix} \quad \text{or } |A - \lambda I| = (4.6 - \lambda)(1.4 - \lambda) - 1.2^2 = (\lambda - 5)(\lambda - 1)$$

Solving the above system of equations, we have the eigenvalues for A are $\lambda = 1$, $y = 3x$ or $\lambda = 5$, $x = -3y$ with the corresponding eigenvectors:

$$\begin{pmatrix} 1/\sqrt{10} \\ 3/\sqrt{10} \end{pmatrix} \text{ and } \begin{pmatrix} -3/\sqrt{10} \\ 1/\sqrt{10} \end{pmatrix}.$$

(Unit vector: $x^2 + y^2 = 1$.)

Principal component analysis (PCA)

Given data $X = \begin{pmatrix} x_{11} & \cdots & x_{1n} \\ \vdots & \ddots & \vdots \\ x_{p1} & \cdots & x_{pn} \end{pmatrix}$ with p variables and n observations, the idea is to find PCs: $Y = \delta'(X - \mu_X)$ that maximize the variance and are uncorrelated from each other, where $\delta = (\delta_1, \dots, \delta_p)'$ is the assigned weights.

$$\max_{\delta:||\delta||=1} var(Y) = \max_{\delta:||\delta||=1} var(\delta' X) = \max_{\delta:||\delta||=1} \delta' var(X)\delta$$

Solution is the first eigenvector of $var(X)$.

Why the eigenvectors?

$$\max \delta' var(X)\delta = \max \delta'(XX')\delta \quad \text{s.t. } \delta'\delta = 1$$

Construct Lagrangian $\delta'(XX')\delta$-$\lambda\delta'\delta$.

FOC: $(XX')\delta$- $\lambda\, \delta = \left(XX' - \lambda I_p\right)\delta = 0$

As $\delta \neq 0$ then δ must be an eigenvector of $var(X) = \frac{1}{n} XX'$ with eigenvalue λ.

The PC transformation is defined as: $Y = \Gamma'(X - \mu_X)$

PCA: ordering

The PC transformation is defined as:

$$Y = \Gamma'(X - \mu_X)$$

Suppose $var(X) = \Sigma$, let Λ and Γ be the diagonal eigenvalue matrix and the corresponding eigenvector of Σ: $\Sigma\Gamma = \Gamma\Lambda$. We have $\Sigma = \Gamma\Lambda\Gamma'$.

$$var(Y) = var(\Gamma'X) = \Gamma' var(X)\Gamma = \Gamma'\Gamma\Lambda\Gamma'\Gamma = \Lambda = diag\left(\lambda_1, \lambda_2, \ldots, \lambda_p\right).$$

Let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$, maximizing the variance of $\delta' X$ leads to the choice $\delta_1 = \gamma_1$, the eigenvector corresponding to the largest eigenvalue λ_1 of Σ.

$$var(Y_1) = \lambda_1, var(Y_2) = \lambda_2, \cdots, var\left(Y_p\right) = \lambda_p.$$

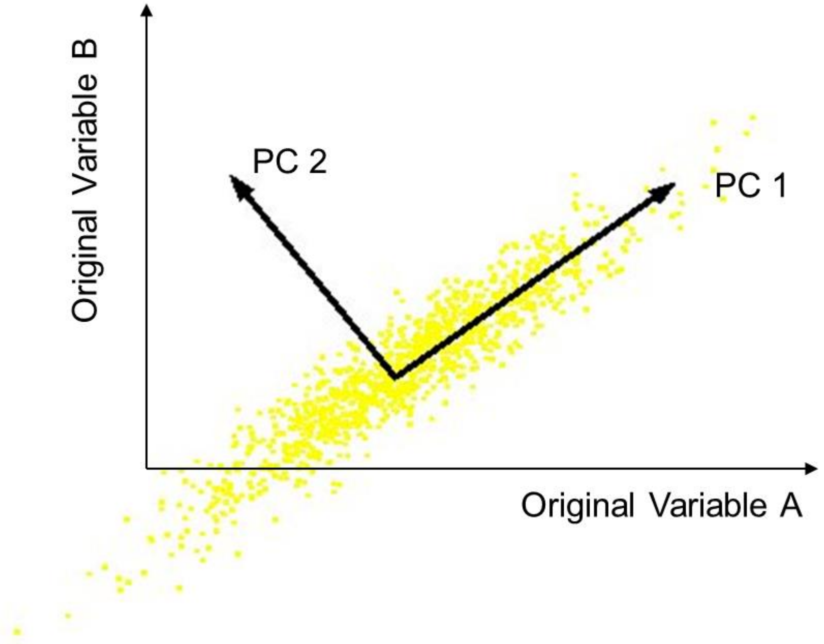
The PCs have zero means, variance $var\left(Y_j\right) = \lambda_j$ and zero correlations. Since $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$, we have $var(Y_1) \geq var(Y_2) \geq \cdots \geq var(Y_p)$.

Visualization

The PC transformation is defined as:

$$Y = \Gamma'(X - \mu_X)$$

This is a projection of X into the one-dimensional space, where the components of X are weighted by the elements of γ_1.
PC1 is the direction of greatest variability (covariance) in the data. PC2 is the next orthogonal (uncorrelated, indicated by the diagonal covariance matrix of Y) direction of greatest variability.



And so on…

How many PCs?

The PC transformation is defined as:

$$Y = \Gamma'(X - \mu_X)$$

Variance explained by the q-th PCs:

$$\Psi_q = \frac{var(Y_q)}{\sum_{j=1}^{p} var(Y_j)} = \frac{\lambda_q}{\sum_{j=1}^{p} \lambda_j}$$

Y_j	$var(Y_j)$	Ψ_j
1	305.90	83.1%
2	36.60	9.9%
3	9.61	2.6%
4	4.71	1.3%
5	3.88	1.1%
6	2.86	0.8%
7	1.61	0.4%
8	1.54	0.4%
9	0.64	0.2%
10	0.62	0.2%
Total variance =	367.97	1

Ignore the components of lesser significance.

You do lose some information, but if the eigenvalues are small, you don't lose much.

Applying PCA to Bond Portfolio Management

There are some applications in bond portfolio management where PCA has been employed.

- The first application is explaining the movement or dynamics in the yield curve and then applying the resulting principal components to measure and manage yield curve risk.
- The second application of PCA is to identify risk factors beyond changes in the term structure.

For example, given historical bond returns and factors that are believed to affect bond returns, PCA can be used to obtain principal components that are linear combinations of the variables that explain the variation in returns.

Using PCA to Control Interest Rate Risk

Using PCA, several studies have investigated the factors that have affected the historical returns on Treasury portfolios.

Robert Litterman and Jose Scheinkman found that three factors explained historical bond returns for U.S. Treasuries zero-coupon securities:

1. The changes in the *level* of rates;
2. The changes in the *slope of the yield curve*;
3. The changes in the *curvature of the yield curve*.

After identifying the factors, Litterman and Scheinkman use regression analysis to assess the relative contribution of these three factors in explaining the returns on zero-coupon Treasury securities of different maturities.

On average, the first principal component explained about 90% of the returns, the second principal component 8%, and the third principal component 2%.

Thus, only three principal components were needed to fully explain the dynamics of the yield curve.

Using PCA to Control Interest Rate Risk

There have been several studies that have examined the yield curve movement using PCA and reported similar results.

Once yield curve risk is described in terms of principal components, the factor loadings can be used to:

- Construct hedges that neutralize exposure to changes in the direction of interest rates.
- Construct hedges that neutralize exposure to changes in nonparallel shifts in the yield curve.
- Structure yield curve trades.

PCA of the dynamics of the yield curve have lead to the use of what is now referred to as principal component duration. Moreover, PCA can be used to estimate the probability associated with a given hypothetical interest rate shock so that a bond portfolio manager can better analyze the interest rate risk of a bond portfolio and traders can better understand the risk exposure of a bond trading strategy.

Example: Principal components analysis of yield curves

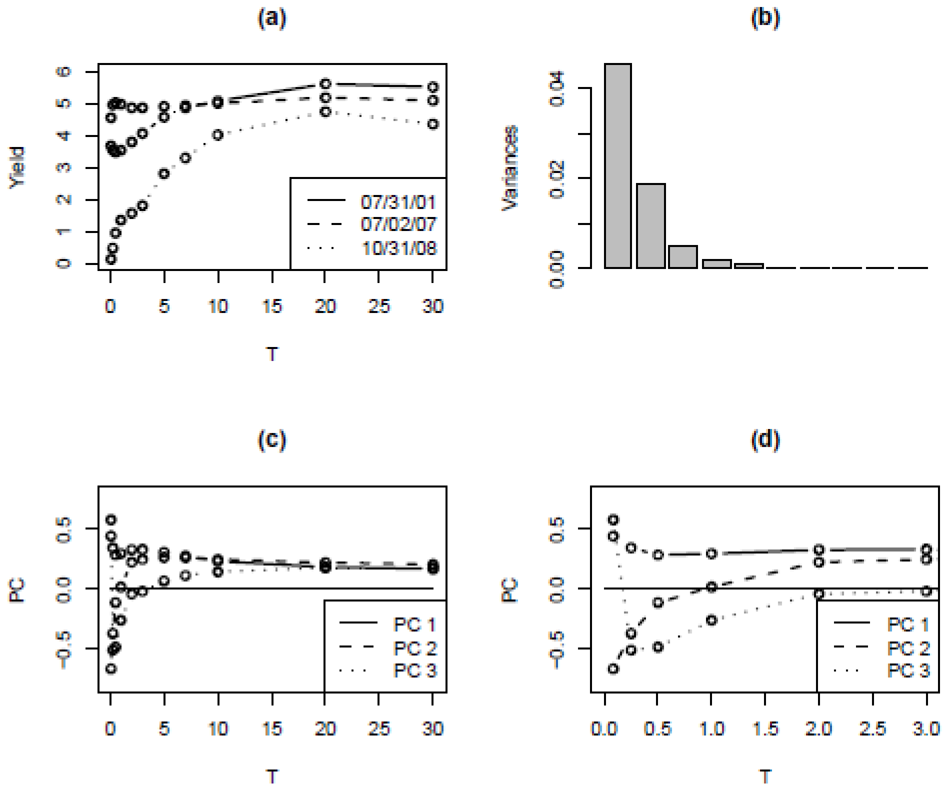


Fig. 17.1. *(a) Treasury yields on three dates. (b) Scree plot for the changes in Treasury yields. Note that the first three principal components have most of the variation, and the first five have virtually all of it. (c) The first three eigenvectors for changes in the Treasury yields. (d) The first three eigenvectors for changes in the Treasury yields in the range* $0 \le T \le 3$.

R: 9_YieldCurves.R

Example: Principal components analysis of yield curves

Importance of components:

	PC1	PC2	PC3	PC4	PC5	PC6
Standard deviation	0.21	0.14	0.071	0.045	0.033	0.0173
Proportion of Variance	0.62	0.25	0.070	0.028	0.015	0.0041
Cumulative Proportion	0.62	0.88	0.946	0.974	0.989	0.9932

PC7	PC8	PC9	PC10	PC11
0.0140	0.0108	0.0092	0.00789	0.00610
0.0027	0.0016	0.0012	0.00085	0.00051
0.9959	0.9975	0.9986	0.99949	1.00000

Example: Principal components analysis of yield curves

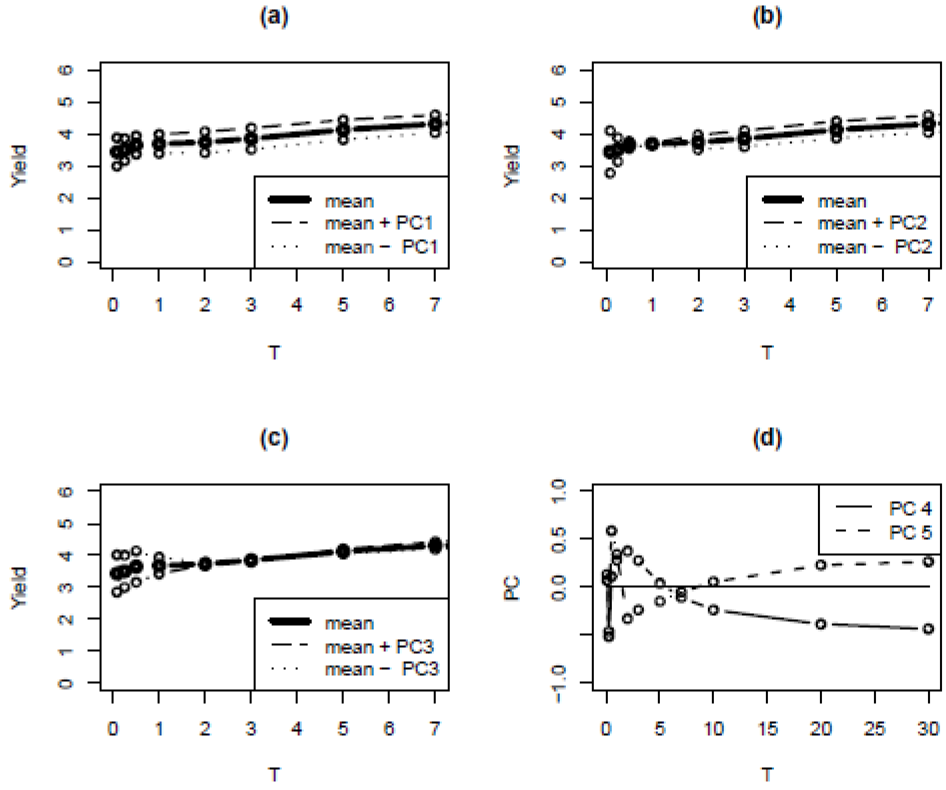


Fig. 17.2. *(a) The mean yield curve plus and minus the first eigenvector. (b) The mean yield curve plus and minus the second eigenvector. (c) The mean yield curve plus and minus the third eigenvector. (d) The fourth and fifth eigenvectors for changes in the Treasury yields.*

R: 9_YieldCurves.R

Example: Principal components analysis of yield curves

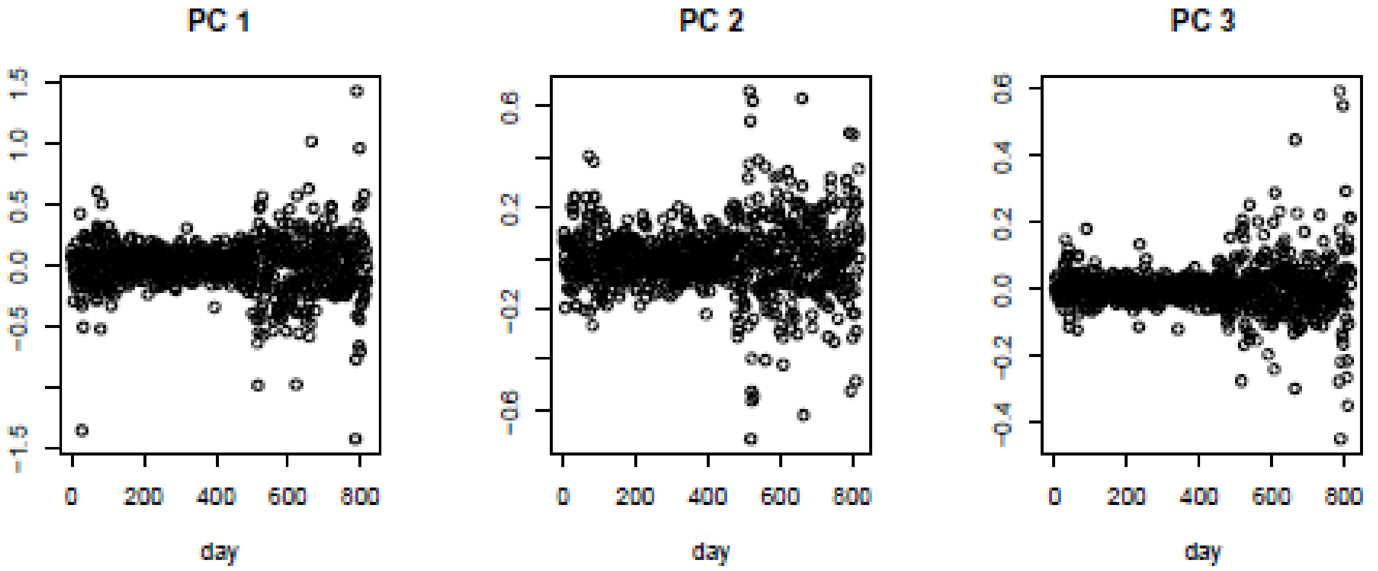


Fig. 17.3. *Time series plots of the first three principal components of the Treasury yields. There are 819 days of data, but they are not consecutive because of missing data; see text.*

R: 9_YieldCurves.R

Example: Principal components analysis of yield curves

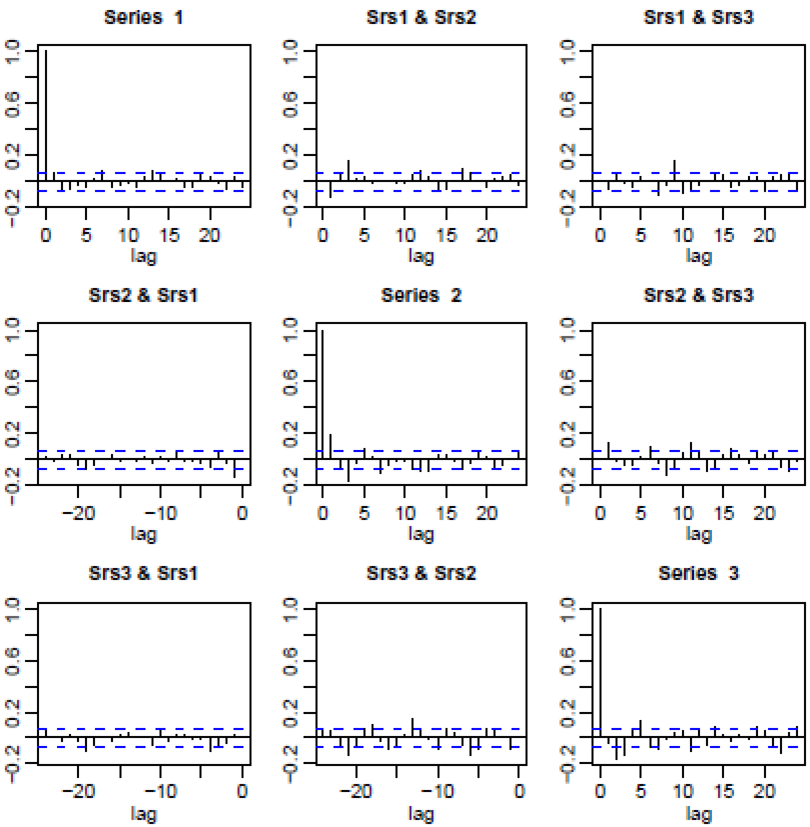


Fig. 17.4. *Sample auto- and cross-correlations of the first three principal components of the Treasury yields.*

R: 9_YieldCurves.R

Normalized PCA (NPCA)

When the data for separate variables exhibit different order of magnitude, or when the units of measurements differ between the variables, it is usual to standardize the data before applying the principal component analysis.

Standardization: $z_j = \frac{X_j - \bar{X}}{\sigma_X}$

The NPC transformation is defined as:

$$Y = \Gamma' Z$$

Note $var(Z) = \mathrm{corr}(\mathrm{X})$.

Example: PCs of a portfolio

Find PCs for an equally weighted portfolio of **four** asset classes, the €/$ currency exchange, the FTSE 100 index, the S&P 500 index, and the ST index. Use monthly returns from Jan 2000 to June 2009.

The (sample) covariance matrix is computed:

$$
\begin{array}{c}
 \\ \text{EUR/USD} \\ \text{FTSE100} \\ \text{S\&P500} \\ \text{STI}
\end{array}
\begin{array}{c}
\begin{array}{cccc}
\text{EUR/USD} & \text{FTSE100} & \text{S\&P500} & \text{STI}
\end{array} \\
\begin{pmatrix}
\color{red}{0.000644908} & -8.35658(10)^{-5} & -0.000282111 & -0.000469847 \\
 & 0.001787221 & 0.001711356 & 0.001830016 \\
 & & 0.002178172 & 0.002129165 \\
 & & & 0.004087288
\end{pmatrix}
\end{array}
$$

Example: PCs of a portfolio

The four eigenvalues and corresponding eigenvectors are:

Eigenvalues: $\lambda_1 = 0.0067552877$, $\lambda_2 = 0.0011182827$, $\lambda_3 = 0.0006033016$,

and $\lambda_4 = 0.0002207170$.

Eigenvectors:

$$\begin{pmatrix} 0.08580 \\ -0.44627 \\ -0.51142 \\ -0.72934 \end{pmatrix} \begin{pmatrix} 0.25073 \\ 0.53669 \\ 0.48798 \\ -0.64107 \end{pmatrix} \begin{pmatrix} 0.91408 \\ 0.13110 \\ -0.30085 \\ 0.23826 \end{pmatrix} \begin{pmatrix} -0.30698 \\ 0.70400 \\ -0.64017 \\ -0.01798 \end{pmatrix}$$

The first PC can be interpreted as a stock market indicator, which accounts for 77.66% of the total variance.

The second PC can be interpreted as a relative performance of STI against the rest assets. It accounts for 12.86% of the total variance.

The third PC is dominated by the FX rate, while the forth PC illustrates the difference between the Europe market and the US market.

Y_j	$var(Y_j)$	Ψ_j
1	0.0067552877	77.66%
2	0.0011182827	12.86%
3	0.0006033016	6.94%
4	0.0002207170	2.54%

The analysis depends on the magnitude of the covariance which is a function of the magnitude of data.

Example: NPCs of the portfolio

$$
\begin{array}{c|cccc}
 & \text{EUR/USD} & \text{FTSE100} & \text{P\&S500} & \text{ST} \\
\hline
\text{EUR/USD} & 1 & -0.07784 & -0.23803 & -0.28939 \\
\text{FTSE100} & & 1 & 0.86737 & 0.67709 \\
\text{P\&S500} & & & 1 & 0.71358 \\
\text{ST} & & & & 1
\end{array}
$$

The four eigenvalues and corresponding eigenvectors are:

Eigenvalues: $\lambda_1 = 2.5845461$, $\lambda_2 = 0.9627458$, $\lambda_3 = 0.3351569$, and $\lambda_4 = 0.1175484$.

Eigenvectors:

$$
\begin{pmatrix} 0.21471 \\ -0.56282 \\ -0.58493 \\ -0.54313 \end{pmatrix}
\begin{pmatrix} 0.94862 \\ 0.29361 \\ 0.10853 \\ -0.04613 \end{pmatrix}
\begin{pmatrix} -0.19339 \\ 0.35953 \\ 0.36144 \\ -0.83827 \end{pmatrix}
\begin{pmatrix} -0.12889 \\ 0.68394 \\ -0.71793 \\ 0.01351 \end{pmatrix}
$$

An example using standardized variables

The first eigenvector has a corresponding eigenvalue of 2.5845461 and accounts for 64.61% of the total variance.

The second eigenvector has a corresponding eigenvalue of 0.9627458 and accounts for 24.07% of the total variance.

The third and the fourth eigenvectors have eigenvalues of 0.3351569 and 0.1175484, respectively, with corresponding proportions of the total variance being 8.38% and 2.94%.

Y_j	$var(Y_j)$	Ψ_{j}
1	2.5845461	64.61%
2	0.9627458	24.07%
3	0.3351569	8.38%
4	0.1175484	2.94%

Factor Analysis (FA)

FA is a data reduction technique designed to represent a wide range of attributes on a smaller number of dimensions.

For example, suppose that a bank asked a large number of questions about a given branch. Consider how the following characteristics might be more parsimoniously represented by just a few constructs (factors).

- Friendliness of staff. - Time spent in line-up. - Assistance via telephone.	Service
- Distance of bank from home. - Hours of operation. - Availability of parking. - Proximity to other stores where you frequently shop.	Convenience
- Monthly account fee. - Charge for with-drawls and deposits. - Loan interest rate.	Cost

FA vs. PCA

FA and PCA are methods of data reduction:

explain many variables with a few “factors” or “components”;

correlated variables are grouped together and separated from other variables with low or no correlation.

But

Factors **cause** variables;

*Principal components **aggregates** the variables.*

FA analyzes only the variance shared among the variables (common variance **without error or unique variance**);

PCA analyzes all of the variance.

FA: “What are the underlying processes that could produce these correlations?”;

PCA: Just summarize empirical associations, very data driven

Factor analysis model

The data matrix $X_{(n \times p)}$ can be rewritten as a linear combination of the latent factors, plus an error term

$$\boxed{\mathbf{X} = \boldsymbol{\mu} + \boldsymbol{\Gamma}\mathbf{F} + \mathrm{E}} = \begin{cases} \mathbf{x_1} = \mu_1 + \gamma_{11}\mathbf{f_1} + \gamma_{12}\mathbf{f_2} + \ldots + \gamma_{1\mu}\mathbf{f_m} + \boldsymbol{\varepsilon}_1 \\ \mathbf{x_2} = \mu_2 + \gamma_{21}\mathbf{f_1} + \gamma_{22}\mathbf{f_2} + \ldots + \gamma_{2\mu}\mathbf{f_m} + \boldsymbol{\varepsilon}_2 \\ \ldots \\ \mathbf{x_p} = \mu_p + \gamma_{p1}\mathbf{f_1} + \gamma_{p2}\mathbf{f_2} + \ldots + \gamma_{pm}\mathbf{f_m} + \boldsymbol{\varepsilon}_p \end{cases}$$

f_i *(i=1,2,…,m) are* ***uncorrelated*** *random variables (common factors) $m \leq p$*

μ_i (i=1,2,…,p) are unique (specific) factors for each variable

e_i (i=1,2,…,p) are error random variables, ***uncorrelated with each other and with F*** *and represent* ***the residual error due to the use of common factors***

$x_j = \mu_j + f(f_k) + \varepsilon_j$

Total variability

Unique variability

Common variability

Error variability

Variability(X)=Variability(μ)+Variability(F)+Variability(ε)

Covariance (Correlation) matrix for {X}

The math of factor analysis

$$\boxed{X = \mu + \Gamma F + E} \longrightarrow X - \mu = \Gamma F + E$$

$$\downarrow$$

$$Var(X - \mu) = \Gamma Var(F)\Gamma' + Var(E)$$

This is the covariance matrix

Assumption: E(F) = 0 and **var(F)=I**

$$\Sigma = \Gamma\Gamma' + \Psi \longrightarrow Var(\mathbf{x_i}) = \sum_{j=1}^{m} \gamma_{ij}^2 + \psi_{ii}^2$$

By summarising the original data through *m* factors we commit an error measured by the residuals, whose **diagonal** variance-covariance matrix is:

$$\Psi = \begin{bmatrix} s_1^2 & 0 & \dots & 0 & \dots & 0 \\ & s_2^2 & \dots & 0 & \dots & 0 \\ & & \dots & \dots & \dots & \dots \\ & & & s_i^2 & \dots & 0 \\ & & & & \dots & \dots \\ & & & & & s_p^2 \end{bmatrix}$$

Some characteristics

Invariance of scale

Assume that we have the following FA model for X:

$$var(X) = \Gamma\Gamma' + \Psi.$$

What happens if we change the scale of X? $\mathrm{Y} = \mathrm{CX}, \mathrm{C} = \mathrm{diag}\,(c_1, \ldots, c_p)$

$var(Y) = \mathrm{C}\Sigma C' = C\Gamma\Gamma' C' + \mathrm{C}\Psi\mathrm{C}'.$

Hence the m-factor model is also true for Y with

$\Gamma_Y = C\Gamma$ and $\Psi_Y = \mathrm{C}\Psi\mathrm{C}'.$

The factors loadings are not unique!

For **orthogonal matrix G** we get:

$X = \mu + (\Gamma G)(G'F) + E.$

We get factor loadings ΓG and common factors $G'F$.

In practice, we will choose the **rotation which gives "desirable" interpretation.**

For the purpose of evaluation, the non-uniqueness can be solved by imposing additional constraints, e.g. $\Gamma'\Psi^{-1}\Gamma$ is diagonal.

Cars

Example: Data consist of the average marks (from 1= low to 7 = high) for 31 car types. Consider three variables **price, security and easy handling**. We look for **one factor**.

Price: $X_1 - \mu_1 = \gamma_1 F_1 + e_1$

Security: $X_2 - \mu_2 = \gamma_2 F_1 + e_2$

Easy handling: $X_3 - \mu_3 = \gamma_3 F_1 + e_3$

The correlation matrix:

$$R = \begin{pmatrix} 1 & 0.975 & 0.613 \\ & 1 & 0.620 \\ & & 1 \end{pmatrix} = \begin{pmatrix} \gamma_1^2 + s_1^2 & \gamma_1\gamma_2 & \gamma_1\gamma_3 \\ & \gamma_2^2 + s_2^2 & \gamma_2\gamma_3 \\ & & \gamma_3^2 + s_3^2 \end{pmatrix}.$$

$$var(X) = \Gamma\Gamma' + \Psi$$

We get the solution: $\gamma_1 = 0.982, \gamma_2 = 0.993, \gamma_3 = 0.624.$ $s_1^2 = 0.035$, $s_2^2 = 0.014$, $s_3^2 = 0.610.$

(Later you will understand, the first two variables are explained by the factor very well. This factor might be interpreted as a “Price+Security” factor)

In this example, we have a unique solution. But in most cases, factor scores cannot be computed exactly but be estimated.

Steps in exploratory factor analysis

(1) Collect and explore data: choose relevant variables.

(2) Determine the number of factors

(3) Estimate the model using predefined number of factors (there is no explicit solution as PCA)

(4) Rotate and interpret

(5) (a) Decide if changes need to be made (e.g. drop item(s), include item(s))

(b) repeat (3)-(4)

(6) Construct scales and use in further analysis

How many factors?

- Scree Plot (Cattell) - Not a test
  - Look for bend in plot
  - Include factor located right at bend point
- Kaiser (or Latent Root) criterion
  - Eigenvalues greater than 1
  - Also, 1 is the amount of variance accounted for by a single item ($r^2 = 1.00$). If eigenvalue < 1.00 then factor accounts for less variance than a single item.
  - Tinsley & Tinsley - Kaiser criterion can underestimate number of factors
- A priori hypothesized # of factors
- Percent of variance criterion
- Parallel analysis – eigenvalues higher than expect by chance

Percent of variance criterion

$$\boxed{\mathbf{X} = \mu + \Gamma\mathbf{F} + \mathbf{E}} \Rightarrow \quad Var(\mathbf{x_i}) = \sum_{j=1}^{m} \gamma_{ij}^2 + \psi_{ii}^2$$

The factor coefficient matrix Γ is sometime referred to as the *factor matrix*. Its elements γ are called *factor loadings* and link the latent factor to the original variables, like correlations.

$$\sum_{j=1}^{m} \gamma_{ij}{}^2$$

is defined as the communality of the i-th variable X_i. The communality is the part of variance of X_i explained by the m factors. *The specific variance* $\psi_{ii}{}^2$ *is the unexplained part.* **The goal of FA is to explain as much as possible.**

The relative communality measures the proportion of variance of the i-th variable explained by the common factors:

$$c_i = \frac{\sum_{j=1}^{m} \gamma_{ij}^2}{var(X_i)}$$

It is between zero and one and tells us – for each variable – what percentage of the original variability is explained by the extracted factors.

Estimation methods

Maximum likelihood: Estimates of factor loadings maximize a likelihood function conditional on the estimated sample covariance matrix. Multivariate normality is assumed for the data.

Least-squares : Estimates are those that minimize the sum of the squared differences between the actual and estimated correlation matrices.

Principal factoring: After the initial estimate on the specific variance (usually from regression analysis between the original variables), principal components are extracted from the difference between the sample covariance matrix and the estimated specific variances. New factor loadings and specific variances become available and the procedure is repeated iteratively.

Bartlett scores: This method considers the FA equation as a system of regression equations where the original variables are the dependent variables, the factor loadings are the explanatory variables and the factor scores are the unknown parameters.

Steps in estimation

Assuming a given number of factors:

1. Correlations among the original variables are exploited to obtain an initial guess for the factor loadings
2. On the basis of this initial guess specific variances are estimated
3. New estimates of the factor loadings are obtained according to some criterion (which varies according to the chosen method), conditional on estimates of the specific variances obtained in step 2
4. New estimates of the specific variances are obtained (again different methods can be used) conditional on estimates of the factor loadings obtained in step 3
5. Return to step (3) and proceed iteratively until estimates of factor loadings and specific variances are stable.

Covariance or correlation?

One may choose to use either the *correlation* or the *covariance matrix* of the original data.

Using the *correlation matrix* corresponds to standardizing the original variables and eliminates the effects of different measurement units or scales, but it also implies that all the original variables have the same variability (not always desirable)

If the measurement scale and unit are the same across variables, there is no reason to force variability to be equal across variables (better to use the *covariance matrix* and maintain the measurement scales)

Factor rotation

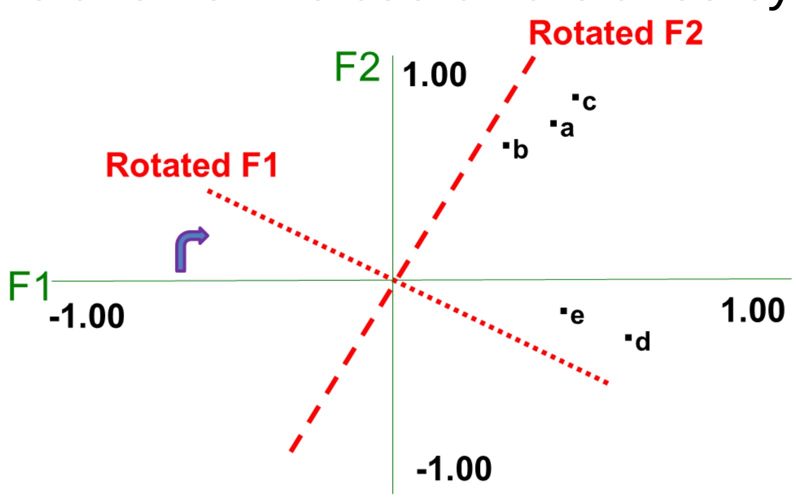
Factor loadings are not unique.
Estimation methods guarantee optimality according to some mathematical rule but not optimality with respect to interpretability.

Rotation may improve interpretability, without any loss of information. Usually, we rotate the factors in a way which provides reasonable interpretation which is consistent with the measured variables.

In the most simple case of m = 2 factors a rotation matrix is given by

$$G(\theta) = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix}$$

which represents a clockwise rotation of the coordinate axes by the angle θ (then $\Gamma^* = \Gamma G(\theta)$)



How to rotate?

Orthogonal methods assume that the factor remain uncorrelated. In some circumstances, the researchers may want to relax such an assumption to make interpretation even easier.

- *Varimax:* Maximizes variance of loadings on each factor (column of the factor matrix). Each factor is explained by few variables (but possibly the same variables can be relevant to explain two or more factors).
- *Quartimax*: Maximizes variance of loadings on each variable (row of the factor matrix). Each variable is expected to explain a reduced number of factors (but one factor may need to be interpreted by a large number of variables)

How to rotate?

Many factors are theoretically related, so rotation method should not “force” orthogonality.

Oblique rotation methods allow for correlated factors. Oblique or correlated components (less or more than 90 degrees) - Account for same % var, but factors correlated.

- *Promax:* Starting from an orthogonal solution, further oblique rotation aims to drive to zero those factor loadings that are smaller. This is achieved by raising to powers the orthogonal loadings.

Although they can lead to further simplification of the factor matrix, it becomes less straightforward to evaluate the contribution of each factor in explaining the original variability.

Rotating or not?

If it is assumed that the latent factor exist but cannot be directly measured, then meaningfulness of the results is a priority.

A good factor should

- Makes sense
- will be easy to interpret
- simple structure
- Lacks complex loadings

After rotation, variance accounted for by a factor is spread out. First factor no longer accounts for max variance possible; others get more variance. Total variance accounted for is the same.

If the objective is only data reduction, regardless of the factor meaning, there is no need for rotation.

How to interpret?

<u>Loadings</u>: represent correlations between item and factor

High loadings: define a factor

Low loadings: item does not “load” on factor

This example:

factor 1 is defined by Distance of bank from home, Hours of operation, Parking availability, Proximity to other stores and frequently shop. (Convenience)

factor 2 is defined by Friendliness of staff, Time spent in line-up, Assistance via phone, Effectiveness of staff, Monthly fee of account, Charge for withdraw. (Service)

Variable	Factor1	Factor2
friendli	-0.3118	**0.5870**
timespe	-0.3498	**0.6155**
assistencete	-0.1919	**0.8381**
staeffective	-0.2269	**0.7345**
distancehome	**0.5682**	-0.1748
hoursoperat	**0.8184**	-0.1212
parkingavail	**0.9233**	-0.1968
proximit	**0.6238**	-0.2227
frequentlysh	**0.8817**	-0.2060
monaccount	-0.0308	**0.4165**
chargewith	-0.1872	**0.5647**

High loadings are highlighted in red.

Example: FA for the portfolio

Back to the portfolio example in PCA: for an equally weighted portfolio of four asset classes, the €/$ currency exchange, the FTSE 100 index, the S&P 500 index, and the ST index, we look for three factors.

$$X_1 = \gamma_{11}f_1 + \gamma_{12}f_2 + \gamma_{13}f_3 + \epsilon_1, \qquad X_2 = \gamma_{21}f_1 + \gamma_{22}f_2 + \gamma_{23}f_3 + \epsilon_2,$$
$$X_3 = \gamma_{31}f_1 + \gamma_{32}f_2 + \gamma_{33}f_3 + \epsilon_3, \qquad X_4 = \gamma_{41}f_1 + \gamma_{42}f_2 + \gamma_{43}f_3 + \epsilon_4.$$

The (sample) correlation matrix is computed:

$$\begin{array}{c} \\ \text{EUR/USD} \\ \text{FTSE100} \\ \text{S\&P500} \\ \text{STI} \end{array}
\begin{array}{c} \begin{array}{cccc} \text{EUR/USD} & \text{FTSE100} & \text{S\&P500} & \text{STI} \end{array} \\ \begin{pmatrix} 1 & -0.07784 & -0.23803 & -0.28939 \\ & 1 & 0.86737 & 0.67709 \\ & & 1 & 0.71358 \\ & & & 1 \end{pmatrix} \end{array}$$

The matrix $\Gamma\Gamma'$ is $$\Gamma\Gamma' = \begin{pmatrix} \boxed{0.998054} & -0.067478 & -0.248905 & -0.289191 \\ & \boxed{0.945017} & 0.925098 & 0.676003 \\ & & \boxed{0.939421} & 0.714725 \\ & & & \boxed{0.99973} \end{pmatrix}$$

The communality have been highlighted in the matrix.

Using the varimax method to rotate. The orthogonal transformation G is

$$\begin{array}{c|c} & \begin{array}{ccc} 1 & 2 & 3 \end{array} \\ \begin{array}{c} 1 \\ 2 \\ 3 \end{array} & \begin{pmatrix} 0.81021 & -0.23619 & 0.53644 \\ 0.29708 & 0.95443 & -0.02848 \\ -0.50527 & 0.18244 & 0.84346 \end{pmatrix} \end{array}$$

Example: FA for the portfolio

The new λ matrix resulting from the rotation is

$$
\begin{array}{c c}
 & \begin{array}{ccc} \text{Factor 1} & \text{Factor 2} & \text{Factor 3} \end{array} \\
\begin{array}{c} \text{EUR/USD} \\ \text{FTSE100} \\ \text{P\&S500} \\ \text{ST} \end{array} &
\begin{pmatrix}
-0.05971 & 0.99032 & -0.11724 \\
0.92385 & 0.02328 & 0.30162 \\
0.89926 & -0.15865 & 0.32493 \\
0.44880 & -0.16089 & 0.87902
\end{pmatrix}
\end{array}
$$

Interpretation:

- Factor 1 has virtually no EUR/US, but large combinations of equity indexes (mainly FTSE & SP500).
- Factor 2 has its largest element the currency with virtually no equity index.
- Factor 3 has a large elements of ST index.

Example: Factor analysis of equity funds

```
> factanal(equityFunds[,2:9],4,rotation="none")

Call:
factanal(x = equityFunds[, 2:9], factors = 4,
      rotation = "none")

Uniquenesses:
EASTEU  LATAM  CHINA  INDIA ENERGY MINING   GOLD  WATER
 0.735  0.368  0.683  0.015  0.005  0.129  0.005  0.778

Loadings:
       Factor1 Factor2 Factor3 Factor4
EASTEU  0.387   0.169   0.293
LATAM   0.511   0.167   0.579
CHINA   0.310   0.298   0.362
INDIA   0.281   0.951
ENERGY  0.784                   0.614
MINING  0.786           0.425  -0.258
GOLD    0.798                  -0.596
WATER   0.340           0.298   0.109

               Factor1 Factor2 Factor3 Factor4
SS loadings       2.57    1.07    0.82    0.82
Proportion Var    0.32    0.13    0.10    0.10
```

Data: 9_equityFunds.csv

R: 9_FactorAnalysisOfEquityFunds.R

Example: Factor analysis of equity funds

```
Cumulative Var    0.32    0.46    0.56    0.66

Test of the hypothesis that 4 factors are sufficient.
The chi square statistic is 17 on 2 degrees of freedom.
The p-value is 2e-04
```

- factanal standardizes the variables, the factor model estimate of the correlation matrix is the estimate of the covariance matrix, that is,

$$\widehat{\boldsymbol{\beta}}^T \widehat{\boldsymbol{\beta}} + \widehat{\boldsymbol{\Sigma}}_{\epsilon}$$

Example: Factor analysis of equity funds: Varimax rotation

```
Call:
factanal(x = equityFunds[, 2:9], factors = 4,
      rotation = "varimax")

Uniquenesses:
EASTEU  LATAM  CHINA  INDIA ENERGY MINING   GOLD  WATER
 0.735  0.368  0.683  0.015  0.005  0.129  0.005  0.778

Loadings:
       Factor1 Factor2 Factor3 Factor4
EASTEU 0.436   0.175   0.148   0.148
LATAM  0.748   0.174           0.180
CHINA  0.494           0.247
INDIA  0.243           0.959
ENERGY 0.327   0.118           0.934
MINING 0.655   0.637           0.168
GOLD   0.202   0.971
WATER  0.418                   0.188

               Factor1 Factor2 Factor3 Factor4
SS loadings       1.80    1.45    1.03    1.00
Proportion Var    0.23    0.18    0.13    0.12
Cumulative Var    0.23    0.41    0.54    0.66

Test of the hypothesis that 4 factors are sufficient.
The chi square statistic is 17 on 2 degrees of freedom.
The p-value is 2e-04
```

Data: 9_equityFunds.csv
R: 9_FactorAnalysisOfEquityFunds.R

Criticisms of factor analysis

Labels of factors can be arbitrary or lack scientific basis.

Derived factors often very obvious.

defense: but we get a quantification

Too many steps that could affect results.

Correlation matrix is often poor measure of association of input variables.

PCA vs FA at a glance

Factor analysis	Principal component analysis
Number of factors predetermined	Number of components evaluated ex post
Many potential solutions	Unique mathematical solution
Factor matrix is estimated	Component matrix is computed
Factor scores are estimated	Component scores are computed
More appropriate when searching for an underlying structure	More appropriate for data reduction (no prior underlying structure assumed).
Factors are not necessarily sorted	Factors are sorted according to the amount of explained variability
Only common variability is taken into account	Total variability is taken into account
Estimated factor scores may be correlated	Component scores are always uncorrelated
A distinction is made between common and specific variance	No distinction between specific and common variability
Preferred when there is substantial measurement error in variables	Preferred as a preliminary method to cluster analysis or to avoid multicollinearity in regression
Rotation is often desirable as there are many equivalent solutions	Rotation is less desirable, unless components are difficult to interpret and explained variance is spread evenly across components

Factor Models

Factor models are statistical models that try to explain complex phenomena through a small number of basic causes or factors.

Factor models serve two main purposes:

- They reduce the dimensionality of models to make estimation possible;
- They find the true causes that drive data.

Factor models were introduced by Charles Spearman in 1904.

The Spearman model explains intellectual abilities through one common factor, the famous ”general intelligence” g factor, plus another factor s which is specific to each distinct ability.

Louis Leon Thurstone developed the first true multifactor model of intelligence, where were identified the following seven primary mental abilities:

Verbal Comprehension	Word Fluency
Number Facility	Spatial Visualization
Associative Memory	Perceptual Speed
Reasoning.	

Factor Models

In the early applications of factor models to psychometrics, the statistical model was essentially a conditional multivariate distribution. The objective was to explain psychometric tests as probability distributions conditional on the value of one or more factors. In this way, one can make predictions of, for example, the future success of young individuals in different activities.

In economics, factor models are typically applied to time series. The objective is to explain the behavior of a large number of stochastic processes, typically price, returns, or rate processes, in terms of a small number of factors. These factors are themselves stochastic processes.

.

Factor Models

In order to simplify both modeling and estimation, most factor models employed in financial econometrics are static models. This means that time series are assumed to be sequences of temporally independent and identically distributed (IID) random variables so that the series can be thought as independent samples extracted from one common distribution.

In financial econometrics, factor models are needed not only to explain data but to make estimation feasible. Factor models are able to explain all pairwise correlations in terms of a much smaller number of correlations between factors.

Linear Factor Models Equations

Linear factor models are regression models of the following type:

$$X_i = \alpha_i + \sum_{j=1}^{K} \beta_{ij} f_j + \epsilon_i$$

where

X_i = a set of N random variables

f_j = a set of K common factors

ϵ_i = the noise terms associated with each variable X_i

β_{ij}'s are the ***factor loadings*** or ***factor sensitivities***, which express

the influence of the j-th factor on the i-th variable.

Note: In this formulation, factor models are essentially static models, but it is possible to add a dynamics to both the variables and the factors.

Factor Models: Types of Factors and Their Estimation

In financial econometrics, the factors used in factor models can belong to three different categories:

- Macroeconomic factors
- Fundamental factors
- Statistical factors

Macroeconomic factors are macroeconomic variables that are believed to determine asset returns (Example: GNP, the inflation rate, the unemployment rate, or the steepness of the yield curve).

Fundamental factors are variables that derive from financial analysis.

Statistical factors are factors that derive from a mathematical process.

Macroeconomic factors

Macroeconomic factors are exogenous factors that must be estimated as variables exogenous to the factor model. They influence the model variables but are not influenced by them.

A factor model is estimated as a linear regression model, means that there is indeed a linear relationship between the factors and the model variables.

However, such a model will have no explanatory power. The variance of each variable that is not explained by common factors appears as noise.

Adding factors might improve the explanatory power of the model but, in general, worsens the ability to estimate the model because there are more parameters to estimate. There is a trade-off between adding explanatory factors and the ability to estimate them.

Example: A macroeconomic factor model

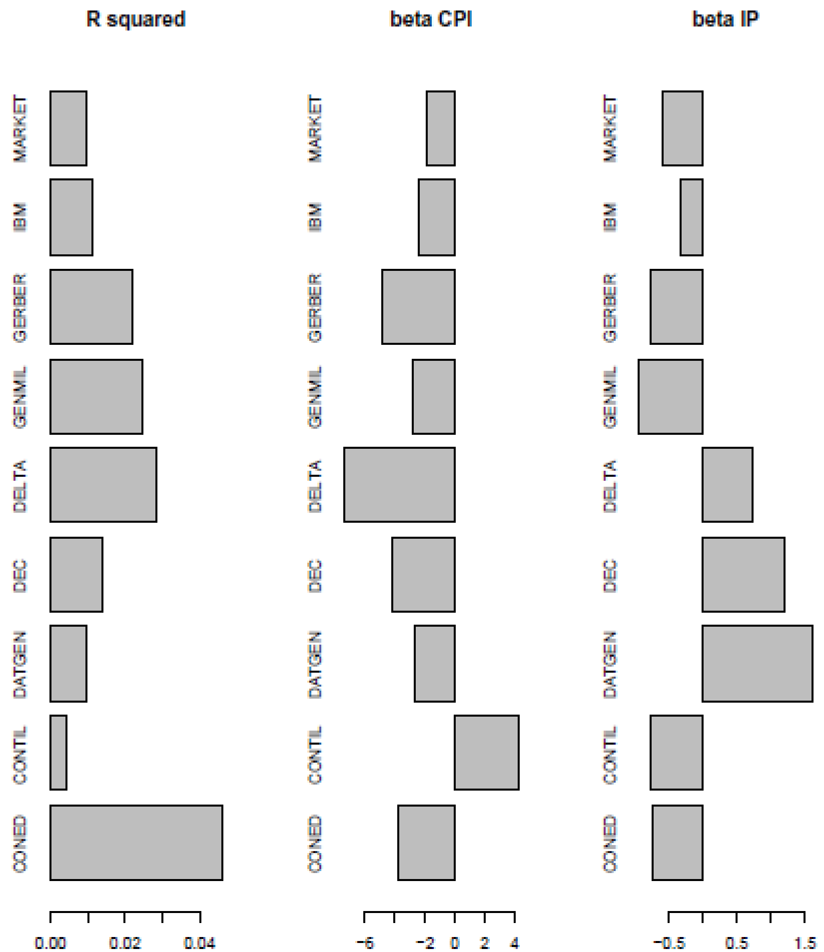


Fig. 17.6. *R^2 and slopes of regressions of stock returns on CPI residuals and IP residuals.*

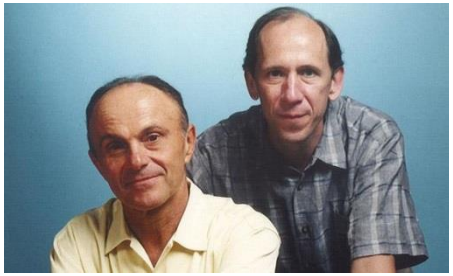
Data: 7_CPI.dat.csv, 7_IP.dat.csv, 9_berndtInvest.csv
R: 9_MacroeconomicFactorModel.R

Financial factors

Factor models generalize the CAPM by allowing more factors than simply the market risk and the unique risk of each asset. A *factor* can be any variable thought to affect asset returns. Examples of factors include:

1. returns on the market portfolio;
2. growth rate of the GDP;
3. interest rate on short term Treasury bills or changes in this rate;
4. inflation rate or changes in this rate;
5. interest rate spreads, for example, the difference between long-term Treasury bonds and long-term corporate bonds;
6. return on some portfolio of stocks, for example, all U.S. stocks or all stocks with a high ratio of book equity to market equity --- this ratio is called BE/ME in Fama and French (1992, 1995, 1996);
7. the difference between the returns on two portfolios, for example, the difference between returns on stocks with high BE/ME values and stocks with low BE/ME values.

Fama-French model



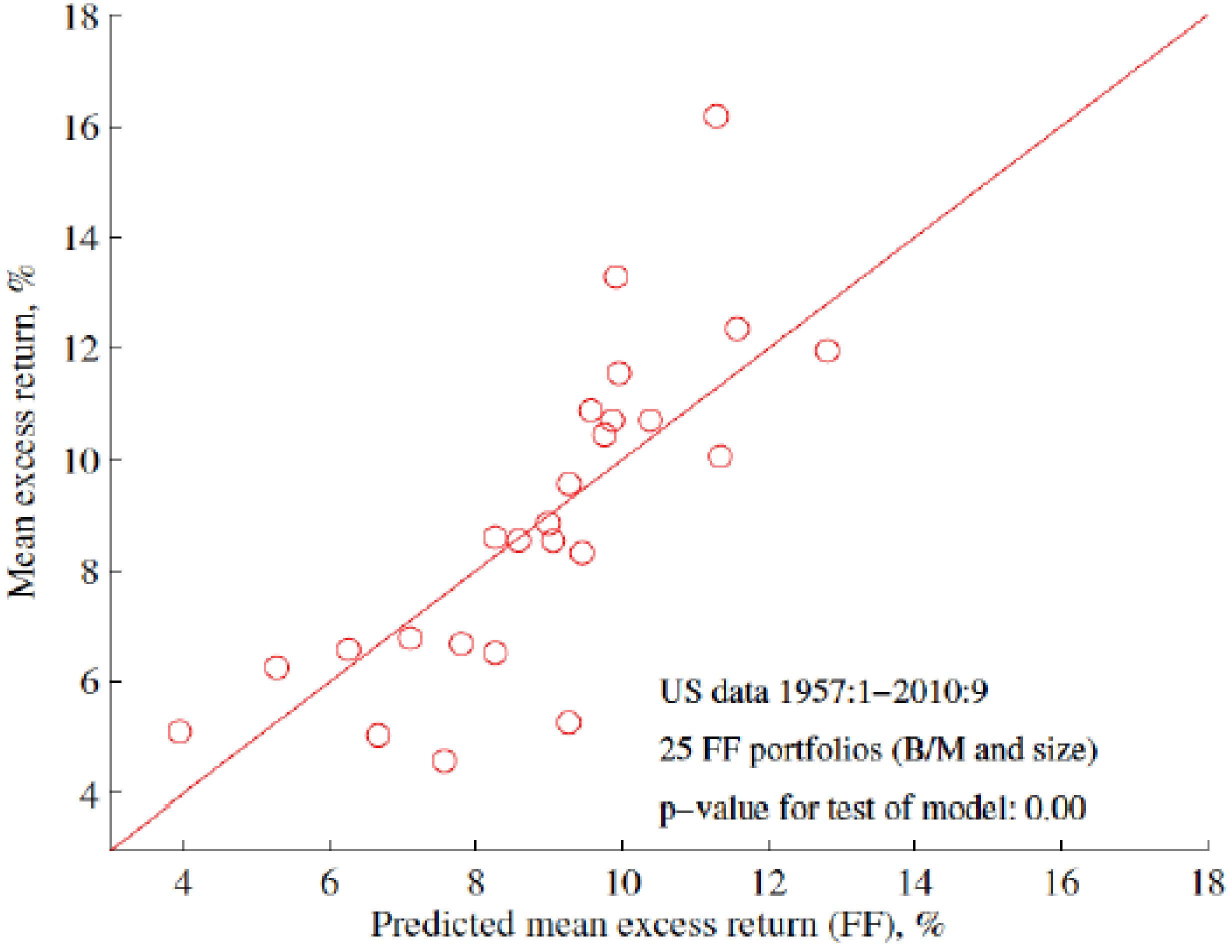
Fama and French (1993, 1996).

Construct 25 stock portfolios according to two characteristics of the firm:
- ✓ the size
- ✓ the book-value-to-market-value ratio (B/M)

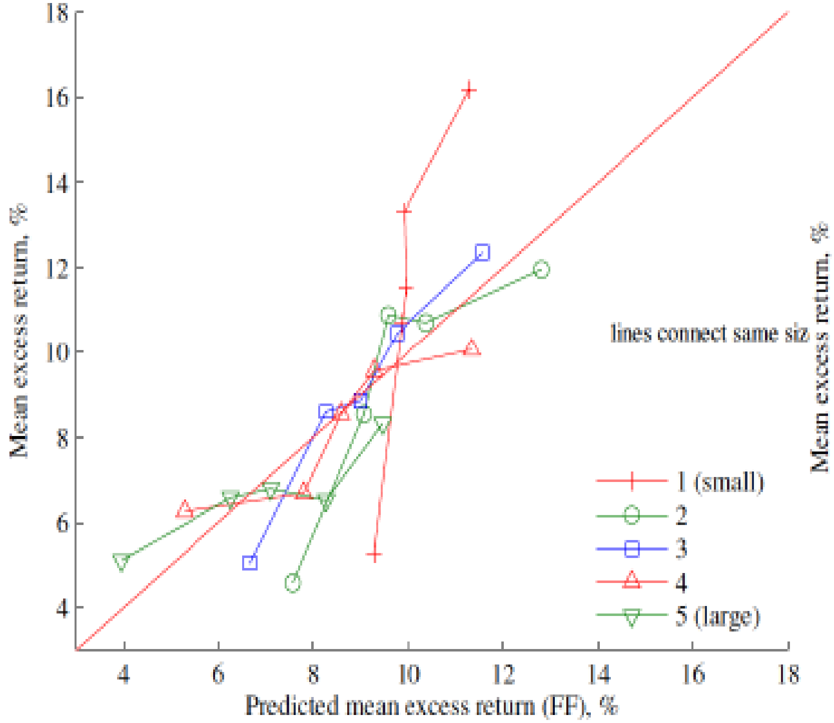
Fama and French find that a three-factor model fits the stock portfolios fairly well.
- ❑ the market return
- ❑ the return on a portfolio of small stocks minus the return on a portfolio of big stocks
- ❑ the return on a portfolio with high B/M minus the return on a portfolio with low B/M

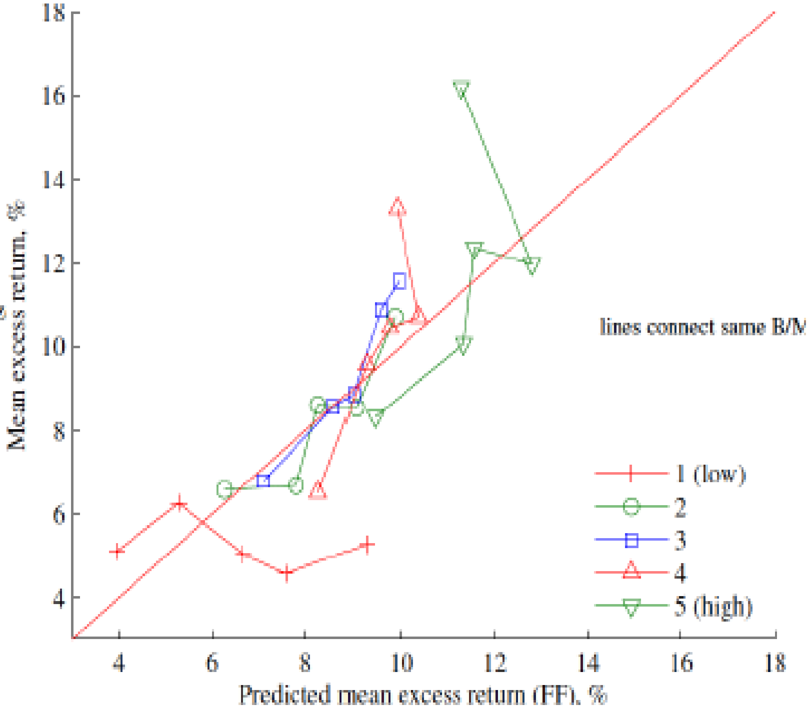
The three-factor model is rejected at traditional significance levels, but it can still capture a fair amount of the variation of expected returns.

FF three-factor model



FF three-factor model (B/M and size)





Example: Fitting the Fama-French model to GE, IBM, and Mobil

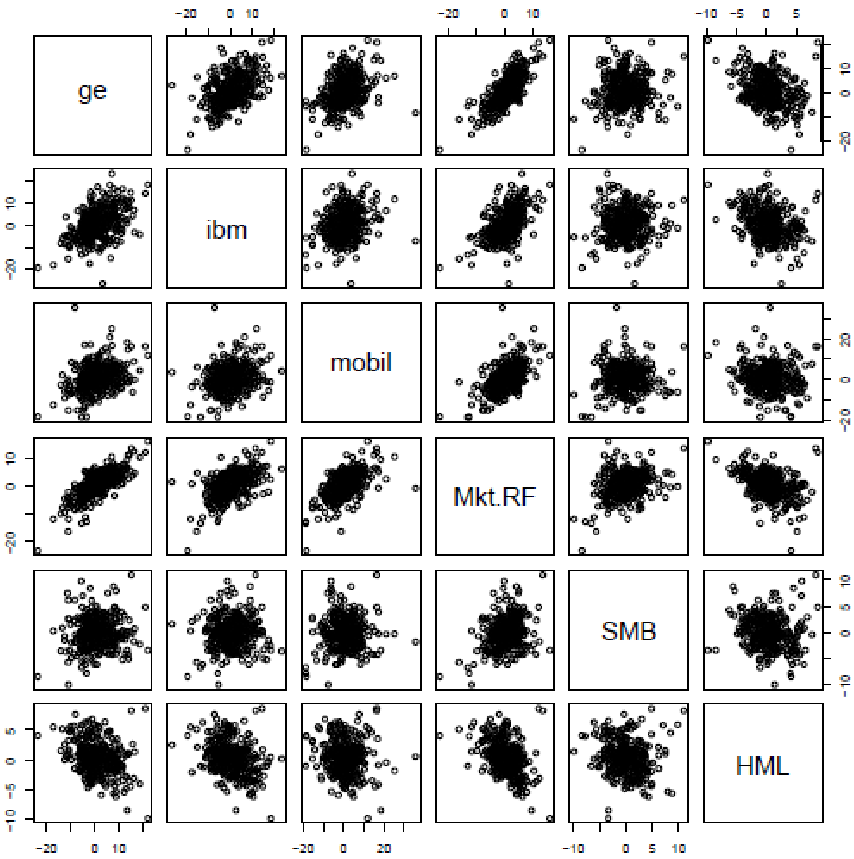


Fig. 17.7. *Scatterplot matrix of the excess returns on GE, IBM, and Mobil and the three factors in the Fama–French model.*

Data: 9_FamaFrench_mon_69_98.txt, 9_CRSPmon.txt
R: 9_Fama-FrenchModel.R

Example: Fitting the Fama-French model to GE, IBM, and Mobil

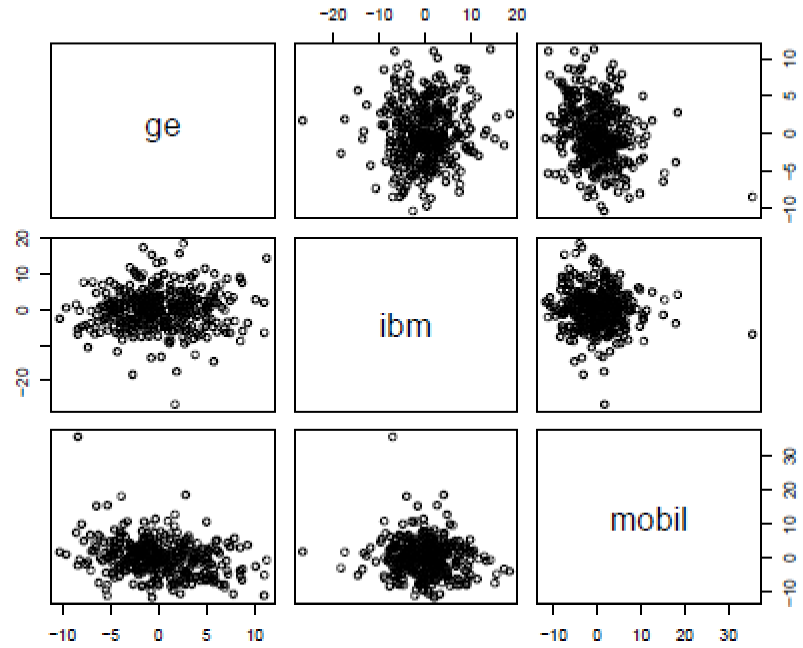


Fig. 17.8. *Scatterplot matrix of the residuals for GE, IBM, and Mobil from the Fama–French model.*

Data: 9_FamaFrench_mon_69_98.txt, 9_CRSPmon.txt
R: 9_Fama-FrenchModel.R

Statistical factors

Statistical factors are obtained through a logical process of analysis of the given variables.

Statistical factors are factors that are endogenous to the system. They are typically determined with one of two statistical processes; namely, principal component analysis or factor analysis.

Note that factors defined through statistical analysis are linear combinations of the variables.

R lab

In this section, we will start with the one-factor CAPM model of Chapter 16 and then extend this model to the three-factor Fama-French model. We will use the data set Stock_FX_Bond_2004_to_2005.csv on the book's website, which contains stock prices and other financial time series for the years 2004 and 2005. Data on the Fama-French factors are available at Prof. Kenneth French's website

http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/

data_library.html#Research

where RF is the risk-free rate and Mkt.RF, SMB, and HML are the Fama-French factors.

Go to Prof. French's website and get the daily values of RF, Mkt.RF, SMB, and HML for the years 2004-2005. It is assumed here that you've put the data in a text file FamaFrenchDaily.txt. Returns on this website are expressed as percentages.

R lab

Now fit the CAPM to the four stocks using the lm command. This code fits a linear regression model separately to the four responses. In each case, the independent variable is Mkt.RF.

```
# Uses daily data 2004-2005
stocks = read.csv("Stock_FX_Bond_2004_to_2005.csv",header=T)
stocks_subset=as.data.frame(cbind(GM_AC,F_AC,UTX_AC,MRK_AC))
stocks_diff = as.data.frame(100*apply(log(stocks_subset),
2,diff) - FF_data$RF)
names(stocks_diff) = c("GM","Ford","UTX","Merck")
FF_data = read.table("FamaFrenchDaily.txt",header=T)
FF_data = FF_data[-1,] # delete first row since stocks_diff
# lost a row due to differencing
fit1 = lm(as.matrix(stocks_diff)~FF_data$Mkt.RF)
summary(fit1)
```

R lab

Problem 1 The CAPM predicts that all four intercepts will be zero. For each stock, using $\alpha = 0.025$, can you accept the null hypothesis that its intercept is zero? Why or why not? Include the p-values with your work.

Problem 2 The CAPM also predicts that the four sets of residuals will be uncorrelated. What is the correlation matrix of the residuals? Give a 95% confidence interval for each of the six correlations. Can you accept the hypothesis that all six correlations are zero?

Problem 3 Regardless of your answer to Problem 6, assume for now that the residuals are uncorrelated. Then use the CAPM to estimate the covariance matrix of the excess returns on the four stocks. Compare this estimate with the sample covariance matrix of the excess returns. Do you see any large discrepancies between the two estimates of the covariance matrix?

R lab

Next, you will fit the Fama-French three-factor model. Run the following R code, which is much like the previous code except that the regression model has two additional predictor variables, SMB and HML.

```
fit2 = lm(as.matrix(stocks_diff)~FF_data$Mkt.RF +
FF_data$SMB + FF_data$HML)
summary(fit2)
```

Problem 4 The CAPM predicts that for each stock, the slope (beta) for SMB and HML will be zero. Explain why the CAPM makes this prediction. Do you accept this null hypothesis? Why or why not?

Problem 5 If the Fama-French model explains all covariance between the returns, then the correlation matrix of the residuals should be diagonal. What is the estimated correlations matrix? Would you accept the hypothesis that the correlations are all zero?

Problem 6 Which model, CAPM or Fama-French, has the smaller value of AIC? Which has the smaller value of BIC? What do you conclude from this?

R lab

Problem 7 What is the covariance matrix of the three Fama-French factors?

Problem 8 In this problem, Stocks 1 and 2 are two stocks, not necessarily in the Stock_FX_Bond_2004_to_2005.csv data set. Suppose that Stock 1 has betas of $0.5, 0.4$, and -0.1 with respect to the three factors in the Fama-French model and a residual variance of 23.0. Suppose also that Stock 2 has betas of 0.6, 0.15, and 0.7 with respect to the three factors and a residual variance of 37.0. Regardless of your answer to Problem 9, when doing this problem, assume that the three factors do account for all covariances.

a) Use the Fama-French model to estimate the variance of the excess return on Stock 1.

b) Use the Fama-French model to estimate the variance of the excess return on Stock 2.

c) Use the Fama-French model to estimate the covariance between the excess returns on Stock 1 and Stock 2.

R lab

This section applies statistical factor analysis to the log returns of 10 stocks in the data set Stock_FX_Bond.csv. The data set contains adjusted costing (AC) prices of the stocks, as well as daily volumes and other information that we will not use here.

The following R code will read the data, compute the log returns, and fit a two-factor model. Note that factanal works with the correlation matrix or, equivalently, with standardized variables.

```
dat = read.csv("Stock_FX_Bond.csv")
stocks_ac = dat[,c(3,5,7,9,11,13,15,17)]
n = length(stocks_ac[,1])
stocks_returns = log(stocks_ac[-1,] / stocks_ac[-n,])
fact = factanal(stocks_returns,factors=2,,rotation="none")
print(fact)
```

Loadings less than the parameter cutoff are not printed. The default value of cutoff is 0.1, but you can change it as in “print(fact,cutoff=.01)" or “print(fact,cutoff=0)".

R lab

Problem 9 What are the factor loadings? What are the variances of the unique risks for Ford and General Motors?

Problem 10 Does the likelihood ratio test suggest that two factors are enough? If not, what is the minimum number of factors that seems sufficient?

The following code will extract the loadings and uniquenesses.

```
loadings = matrix(as.numeric(loadings(fact)),ncol=2)
unique = as.numeric(fact$unique)
```

Problem 11 Regardless of your answer to Problem 2, use the two-factor model to estimate the correlation of the log returns for Ford and IBM.

R lab

Problem 1.

For GM the p-value is 0.008 so we reject that the intercept is zero. For Ford the p-value is 0.007 so again we reject that the intercept is zero. For the other two stocks, the p-values are 0.955 and 0.335 and we accept that the intercepts are zero.

```
> summary(fit1)
Response GM :
Call:
lm(formula = GM ~ FF_data$Mkt.RF)

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    -0.22902    0.08646  -2.649  0.00833 **
FF_data$Mkt.RF  1.25000    0.12730   9.819  < 2e-16 ***
---

Response Ford :
Call:
lm(formula = Ford ~ FF_data$Mkt.RF)

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    -0.18347    0.06757  -2.715  0.00685 **
FF_data$Mkt.RF  1.31952    0.09950  13.262  < 2e-16 ***
---

Response UTX :
Call:
lm(formula = UTX ~ FF_data$Mkt.RF)

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.002193   0.038835   0.056    0.955
FF_data$Mkt.RF 0.919322   0.057183  16.077   <2e-16 ***
---

Response Merck :
Call:
lm(formula = Merck ~ FF_data$Mkt.RF)

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    -0.08883    0.09203  -0.965    0.335
FF_data$Mkt.RF  0.62545    0.13552   4.615 4.99e-06 ***
---
```

R lab

Problem 2.

The following code performs a correlation test, and gives confidence intervals, on all six pairs of residual. Clearly, the correlation between GM and Ford residuals in positive, which is not unexpected since they are in the same industry. The correlation between GM and Merck residuals is borderline significant -- the p-value is 0.049.

```
> for (i in 1:3)
+ {
+ for (j in (i+1):4)
+ {
+ print(c(i,j))
+ print(cor.test(fit1$resid[,i],fit1$resid[,j]))
+ }
+ }
[1] 1 2

        Pearson's product-moment correlation

data:  fit1$resid[, i] and fit1$resid[, j]
t = 13.6323, df = 501, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.4533535 0.5811635
sample estimates:
      cor
0.5201647

[1] 1 3

        Pearson's product-moment correlation

data:  fit1$resid[, i] and fit1$resid[, j]
t = -0.2249, df = 501, p-value = 0.8222
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.09738817  0.07745125
sample estimates:
        cor
-0.01004523

[1] 1 4

        Pearson's product-moment correlation

data:  fit1$resid[, i] and fit1$resid[, j]
t = -1.9719, df = 501, p-value = 0.04918
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.173850629 -0.000330075
sample estimates:
        cor
-0.08775601

[1] 2 3

        Pearson's product-moment correlation

data:  fit1$resid[, i] and fit1$resid[, j]
t = -0.532, df = 501, p-value = 0.595
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.11095967  0.06379929
sample estimates:
        cor
-0.02376172

        Pearson's product-moment correlation

data:  fit1$resid[, i] and fit1$resid[, j]
t = -0.2144, df = 501, p-value = 0.8303
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.09692651  0.07791450
sample estimates:
         cor
-0.009579221

[1] 3 4

        Pearson's product-moment correlation

data:  fit1$resid[, i] and fit1$resid[, j]
t = -0.1232, df = 501, p-value = 0.902
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.09288645  0.08196523
sample estimates:
         cor
-0.005502672
```

R lab

Problem 3.

There are some large differences between the two estimates of the covariance matrix. The largest discrepancy is between GM and Ford where the estimated based on the CAPM model is 0.761 and the sample covariance is 2.282

```
betas1 = as.vector(coef(fit1)[2,])
> covModelBased = var(FF_data$Mkt.RF)* betas1 %*% t(betas1) +
+       diag(diag(cov(fit1$residuals)))
> options(digits=3)
> covModelBased
      [,1]  [,2]  [,3]  [,4]
[1,] 4.464 0.761 0.530 0.360
[2,] 0.761 3.090 0.559 0.381
[3,] 0.530 0.559 1.145 0.265
[4,] 0.360 0.381 0.265 4.423
> covModelFree = cov(stocks_diff)
> covModelFree
          GM  Ford   UTX  Merck
GM    4.4641 2.282 0.513 0.0108
Ford  2.2825 3.090 0.528 0.3507
UTX   0.5130 0.528 1.145 0.2553
Merck 0.0108 0.351 0.255 4.4228
```

R lab

Problem 4.

CAPM predicts that the market returns explain all correlations between equity returns, so returns on other portfolios will have slopes of zero. We see in the results below that the p-values for SMB and HML arc occasionally quite small, e.g., for Merck and HML. Therefore, we reject this hypothesis.

```
> summary(fit2)
Response GM :

Call:
lm(formula = GM ~ FF_data$Mkt.RF + FF_data$SMB + FF_data$HML)

Residuals:
     Min       1Q   Median       3Q      Max
-13.7591  -0.7271  -0.0326   0.7751  15.0074

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     -0.2511     0.0865   -2.90   0.0038 **
FF_data$Mkt.RF   1.3889     0.1517    9.16   <2e-16 ***
FF_data$SMB     -0.2504     0.2236   -1.12   0.2633
FF_data$HML      0.6006     0.2643    2.27   0.0235 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 1.93 on 499 degrees of freedom
Multiple R-squared: 0.173,      Adjusted R-squared: 0.168
F-statistic: 34.9 on 3 and 499 DF,  p-value: <2e-16
```

```
Response Merck :

Call:
lm(formula = Merck ~ FF_data$Mkt.RF + FF_data$SMB + FF_data$HML)

Residuals:
    Min      1Q  Median      3Q     Max
-30.627  -0.454   0.106   0.691  12.310

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     -0.0598     0.0915   -0.65  0.51337
FF_data$Mkt.RF   0.7093     0.1605    4.42  1.2e-05 ***
FF_data$SMB     -0.4174     0.2366   -1.76  0.07832 .
FF_data$HML     -0.9559     0.2796   -3.42  0.00068 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 2.04 on 499 degrees of freedom
Multiple R-squared: 0.066,      Adjusted R-squared: 0.0604
F-statistic: 11.8 on 3 and 499 DF,  p-value: 1.86e-07
```

R lab

Problem 5.

The correlation matrix is clearly not diagonal. In particular, there is still a very high correlation between the two auto companies. The residual correlation matrix is, in fact, similar to that of the CAPM model.

```
> cor(fit2$resid)
             GM      Ford      UTX     Merck
GM       1.0000  0.517015 -0.0185 -0.077819
Ford     0.5170  1.000000 -0.0258  0.000729
UTX     -0.0185 -0.025771  1.0000 -0.01359
Merck   -0.0778  0.000729 -0.0136  1.00000
```

```
> for (i in 1:3)
+ {
+ for (j in (i+1):4)
+ {
+ print(c(i,j))
+ print(cor.test(fit2$resid[,i],fit2$resid[,j]))
+ }
+ }
[1] 1 2

        Pearson's product-moment correlation

data:  fit2$resid[, i] and fit2$resid[, j]
t = 13.5, df = 501, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.450 0.578
sample estimates:
  cor
0.517

[1] 1 3

        Pearson's product-moment correlation

data:  fit2$resid[, i] and fit2$resid[, j]
t = -0.415, df = 501, p-value = 0.6786
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.106  0.069
sample estimates:
    cor
-0.0185

[1] 1 4
```

R lab

Problem 6.

The Fama-French model has the smallest value of AIC, but the CAPM has the smallest value of BIC. Therefore, it is not clear which is the better of the two models. Certainly, there is not clear evidence that the Fama-French model is superior.

```
> AIC(fit1)
[1]7752
> AIC(fit2)
[1]7721
> n = dim(stocks_diff)[1]
> BIC1 = AIC(fit1)+(log(n)-2)*4
> BIC2= AIC(fit2)+(log(n)-2)*12
> BIC1 [1]7769
> BIC2
[1] 7772
```

R lab

Problem 7.

```
> cov_FF_factors = cov(FF_data[,2:4])
> cov_FF_factors
         Mkt.RF     SMB     HML
Mkt.RF  0.4611  0.1723 -0.0348
SMB     0.1723  0.2146 -0.0290
HML    -0.0348 -0.0290  0.1102
```

R lab

Problem 8.

(a) Use the Fama-French model to estimate the variance of the excess return on Stock 1: 23.2

(b) Use the Fama-French model to estimate the variance of the excess return on Stock 2: 37.2

(c) Use the Fama-French model to estimate the covariance between the excess returns on Stock 1 and Stock 2: 0.18

```
> betaStock1 = c(0.5, 0.4, -0.1)
> residVar1=23.0
> betaStock2 = c(0.6, 0.15, 0.7)
> residVar2 = 37
> t(betaStock1) %*% cov_FF_factors %*% betaStock1 + residVar1[,1]
[1,] 23.2
> t(betaStock2) %*% cov_FF_factors %*% betaStock2 + residVar2[,1]
[1,] 37.2
> t(betaStock1) %*% cov_FF_factors %*% betaStock2[, 1]
[1,]  0.18
```

R lab

Problem 9. The loadings are the 8 by 2 matrix shown above. The variances of the unique risks for Ford and General Motors are 0.423 and 0.399, respectively.

```
> print(fact)

Call:
factanal(x = stocks_returns, factors = 2, rotation = "none")

Uniquenesses:
  GM_AC    F_AC  UTX_AC  CAT_AC  MRK_AC  PFE_AC  IBM_AC MSFT_AC
  0.399   0.423   0.718   0.714   0.519   0.410   0.760   0.749

Loadings:
        Factor1 Factor2
GM_AC    0.693  -0.348
F_AC     0.692  -0.313
UTX_AC   0.531
CAT_AC   0.529
MRK_AC   0.551   0.421
PFE_AC   0.574   0.511
IBM_AC   0.490
MSFT_AC  0.499

               Factor1 Factor2
SS loadings      2.643   0.666
Proportion Var   0.330   0.083
Cumulative Var   0.330   0.414

Test of the hypothesis that 2 factors are sufficient.
The chi square statistic is 564.7 on 13 degrees of freedom.
The p-value is 2.6e-112
```

R lab

Problem 10. The p-value for testing that 2 factors are sufficient is very small, so we conclude that more than 2 factors is needed. Re-running the code with 3 and then 4 factors, we see that 3 factors are also not sufficient but 4 factors are enough.

Problem 11. The code below will compute the estimate of the correlation matrix using the factor model.

```
loadings= matrix(as.numeric(loadings(fact)),ncol=2)
unique = as.numeric(fact$unique)
loadings %*% t(loadings) + diag(unique)
```

This code uses the fact that the factors are uncorrelated with unit variances as implicit in the following from R's help:

The factor analysis model is x = Lambda f + e

for a p-element row-vector x, a p x k matrix of loadings, a k-element vector of scores and a p-element vector of errors. None of the components other than x is observed, but the major restriction is that the scores be uncorrelated and of unit variance, and that the errors be independent with variances Phi, the uniquenesses. Thus factor analysis is in essence a model for the covariance matrix of x,

Sigma = Lambda'Lambda + Psi

R lab

Problem 11. The correlation between the log-returns of Ford and IBM (stocks 2 and 7) is 0.3379.

```
> loadings = matrix(as.numeric(loadings(fact)),ncol=2)
> unique = as.numeric(fact$unique)
> loadings %*% t(loadings) + diag(unique)
       [,1]   [,2]   [,3]   [,4]   [,5]   [,6]   [,7]   [,8]
[1,] 1.0000 0.5883 0.3774 0.3939 0.2357 0.2197 0.3383 0.3305
[2,] 0.5883 1.0000 0.3759 0.3906 0.2495 0.2366 0.3379 0.3314
[3,] 0.3774 0.3759 1.0000 0.2829 0.2807 0.2899 0.2600 0.2636
[4,] 0.3939 0.3906 0.2829 1.0000 0.2587 0.2633 0.2591 0.2606
[5,] 0.2357 0.2495 0.2807 0.2587 1.0000 0.5314 0.2722 0.2937
[6,] 0.2197 0.2366 0.2899 0.2633 0.5314 1.0000 0.2834 0.3088
[7,] 0.3383 0.3379 0.2600 0.2591 0.2722 0.2834 1.0000 0.2449
[8,] 0.3305 0.3314 0.2636 0.2606 0.2937 0.3088 0.2449 1.0000
```