Lecture 10: Factor Models and Principle Component Analysis (PCA)

Big Picture

- We want to explain the variation in the return of financial asset
- Instead of using the past history of return, we use other variables (factors) as the regressors
- For more discussion, read *Analysis of Financial Time Series* written by Ruey S. Tsay.

Factor Models

- Suppose there are k assets (most often stocks), and T periods. Let r_{it} be the (excess) return of asset i at time t. Throughout this note we assume r_{it} is stationary.
- A general form of the factor model is

$$r_{it} = \beta_{0i} + \beta_{1i} f_{1t} + \ldots + \beta_{mi} f_{mt} + e_{it} \tag{1}$$

  where we assume there are m factors, and f_{jt} is the j-th factor at time t.
- To distinguish various factor models, the key is to pay attention to the subscript.

Remarks

1. f_{jt} has the time subscript t. That means the factors are time-varying
2. f has no subscript i. That means the factors are common for all assets.
3. β has the subscript i. So the beta is asset-specific. That is, we need to run regression (1) separately for each asset return.
4. If we put k regressions together, it becomes a special case of seemingly unrelated regression (SUR). Because each regression has the same regressors (factors), the GLS estimator for this special SUR is the same as the equation-by-equation OLS estimator.

Example: one-factor model

- There are $k = 13$ monthly excess returns of stocks in the data file

  `672_2014_factor1.txt`

  from Jan 1990 to Dec 2003 (so T=168)

- The single factor (so $m = 1$) is the excess return of the S&P500 index, which approximates the market return.

- We regress each return series onto the S&P500 index. There are in total 13 regressions. The coefficients in each regressions are different.

- We want to find the association between the return of each individual asset and the market.

R Code

```
data = read.table("672_2014_factor1.txt", header=T)
n = 168
c = rep(1, n)   # intercept term
X = as.matrix(cbind(c, data[,14]))
Y = as.matrix(data[,1:13])
bet = matrix(0, 2, 13); ehat = matrix(0, n, 13)
for (i in 1:13) {
bet[,i] = solve(t(X)%*%X)%*%(t(X)%*%Y[,i])
ehat[,i] = Y[,i] - X%*%bet[,i]
}
rss = diag(crossprod(ehat))        # RSS
rsq = 1 - rss/diag(var(Y)*(n-1))  # R-Squared
```

Remarks

The key is the a loop of OLS estimation

```
for (i in 1:13) {
bet[,i] = solve(t(X)%*%X)%*%(t(X)%*%Y[,i])
ehat[,i] = Y[,i] - X%*%bet[,i]
}
```

1. For the return of the i-th stock, the coefficient estimate is

$$\beta_i = (X'X)^{-1}(X'Y_i)$$

   where the dimension of X is 168×2; Y_i is 168×1; and β_i is 2×1.

2. The residual is

$$E_i = Y_i - X\beta_i$$

Result

	Beta	Sigma	R-Squared
AA	1.292	7.694	0.347
AGE	1.514	7.807	0.415
CAT	0.941	7.724	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

Remarks

For example, for the stock of Alcoa (AA is its Tick or Ticker symbol), we find

1. its return is positively correlated with the market return because $\hat{\beta}_{1,AA} = 1.292 > 0$
2. the estimated standard error of regression (SER) σ is 7.694
3. the R^2 is 0.347. So the market return can explain around 34% of the variation of AA stock return.

BARRA Factor Model

BARRA Factor Model

1. The BARRA factor model is

$$r_t = \beta f_t + e_t \tag{2}$$

   where $r_t = (r_{1t}, \ldots, r_{kt})'$. β is given, and is a set of industry dummy variables.

2. f_t is called factor realization, and is unknown here, and needs to be estimated.

3. In short, we need to run the above regression repeatedly for each period. For a given period, the dependent variable is the returns of all assets at that period. The regressors are a set of time-invariant industry-dummies. The estimated coefficient is the factor realization at that period.

Example: BARRA Factor Model

- There are $k = 10$ monthly excess returns of assets in the data file

  `672_2014_factor2.txt`

  from Jan 1990 to Dec 2003 (so T=168)

- The regressors are three industry dummy variables:

```
d.fin = c(rep(1,4),rep(0,6))
d.it = c(rep(0,4),rep(1,3),rep(0,3))
d.ot = c(rep(0,7),rep(1,3))
```

  For example, d.fin equals one if a company is in the financial sector. In this case, AGE, C, MWD and MER are in financial sector, and other firms are not. So d.fin has four ones and six zeros.

OLS Estimator

```
Y = as.matrix(data[,1:10])
X = cbind(d.fin, d.it, d.ot)
f.o = solve(crossprod(X))%*%(t(X)%*%t(Y))  # OLS
f.o[,1:2]
```

Note that

1. you get transpose of the data matrix using $t(Y)$.
2. so each row of Y (or each column of $t(Y)$) is used as the dependent variable
3. in the first period $t = 1$, for instance, the coefficient of d.fin is -10.80. In the second period, the coefficient of d.fin is 2.212500. There are in total 168 coefficients of d.fin. That coefficient series is the factor realization for the financial dummy.

OLS and Industrial Average

Because X includes three industry dummies (and no intercept term). The estimated f_t is

$$f_t = \begin{pmatrix} \frac{AGE_t + C_t + MWD_t + MER_t}{4} \\ \frac{DELL_t + HPQ_t + IBM_t}{3} \\ \frac{AA_t + CAT_t + PG_t}{3} \end{pmatrix}$$

So each component of f_t is the average of returns of firms in that specific industry in the t-th period.

WLS Estimator

1. The OLS estimator is inefficient because it ignores the heteroskedasticity

$$\Omega = Ee_t e_t' = \texttt{diag}(\sigma_1^2, \ldots, \sigma_k^2) \neq \sigma^2 I$$

   where σ_k^2 is the variance of the k-th asset return.

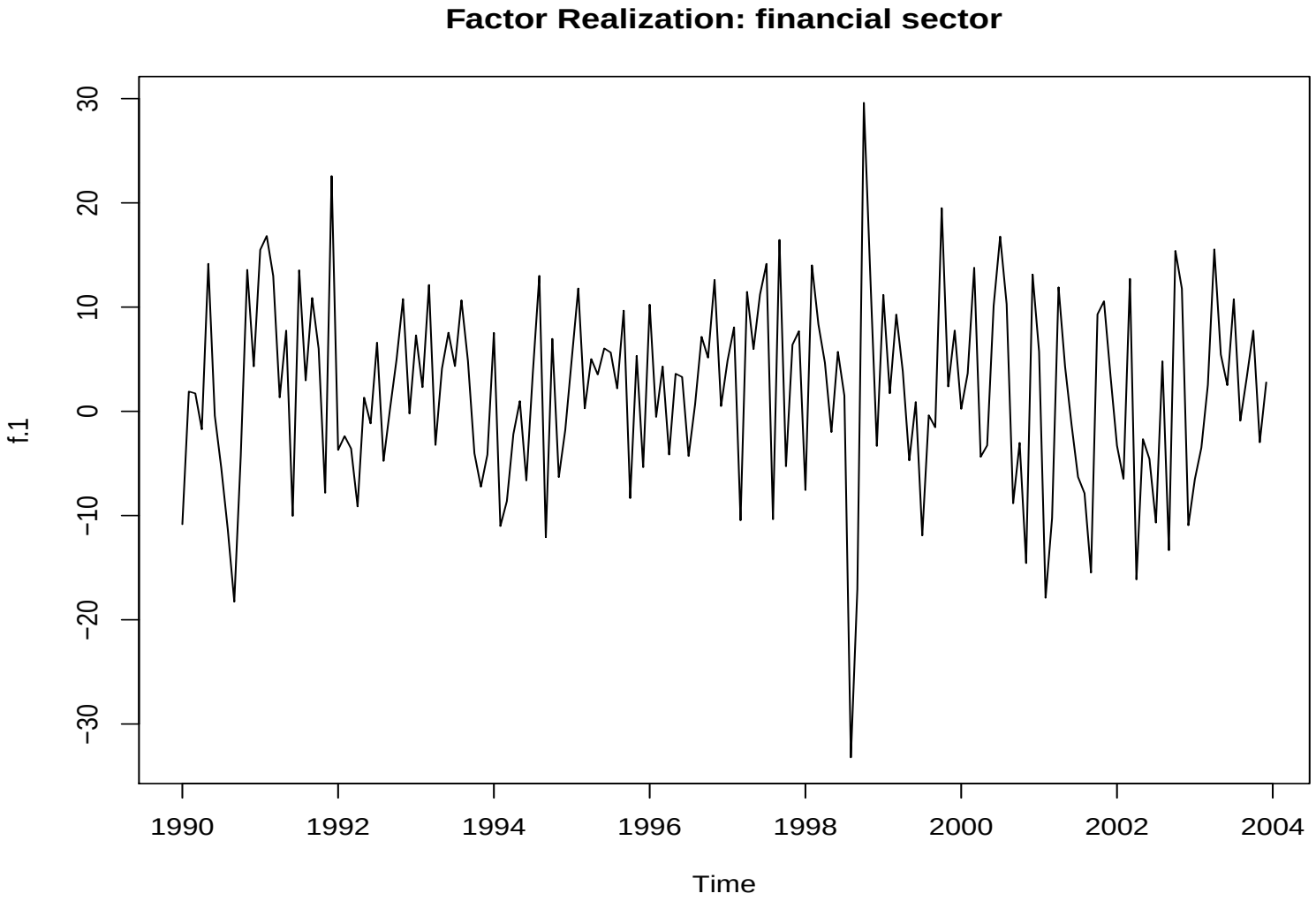
2. We can estimate Ω using the OLS residual. Then the more efficient weighted least squares (WLS) estimator is

$$f_t^{\texttt{WLS}} = (\beta' \widehat{\Omega}^{-1} \beta)^{-1} (\beta' \widehat{\Omega}^{-1} r_t)$$

R Code

```
ehat.o = t(Y) - X%*%f.o
omega = var(t(ehat.o))
weight = diag(omega^(-1))
w.m = Diagonal(10, x = weight)
x.wls = t(X)%*%w.m%*%X
y.wls = t(X)%*%w.m%*%t(Y)
f.wls = solve(x.wls)%*%y.wls                    # WLS
```

Plot of Factor Realization



Statistical Factor Model

Statistical Factor Model

1. In the statistical factor model, the factors are the principle components (PC) of the return series.
2. Consider a linear combination (portfolio) of k returns at the t-th period

$$\sum_{i=1}^{k} w_i r_{it}$$

   where the weight for the i-th asset is w_i.
3. PC is a special linear combination so that
   (a) each PC is uncorrelated with each other
   (b) each PC will explain the maximum amount of remaining variance-covariance of r_t after the previous PC.

Eigenvalue and Eigenvector

1. For a square matrix A, its eigenvalue solves the equation

$$A\lambda = \lambda c,$$

   where c is a (nonzero) column vector, called the eigenvector

2. Mathematically, there are k (possibly duplicate or complex) eigenvalues if the dimension of A is $k \times k$. They are obtained as the roots of

$$\texttt{determinant}(A - \lambda I) = 0$$

Principle Component (PC)

1. Denote the variance-covariance matrix of k returns by Ω.
2. Let $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_k$ be the eigenvalues of Ω in descending order
3. Let $c_1, c_2, \ldots, c_k$ be the corresponding eigenvectors
4. We can prove that the j-th PC is the j-th eigenvector multiplied by the return series, and the variance of the j-th PC is λ_j.

Example: Statistical Factor Model

- There are $k = 5$ monthly excess returns of assets in the data file

  `672_2014_factor3.txt`

  from Jan 1990 to Dec 1999 (so T=120)

- The eigenvalues and eigenvectors are

```
data = read.table("672_2014_factor3.txt", header=T)
data = data[1:120,]
v.m = var(data)
e.value = eigen(v.m)$values
e.vector = eigen(v.m)$vector
```

Principle Components

```
pc1 = as.matrix(data)%*%e.vector[,1]
pc2 = as.matrix(data)%*%e.vector[,2]
pc3 = as.matrix(data)%*%e.vector[,3]
var(pc1)
var(pc2)
var(pc3)
cor(cbind(pc1, pc2, pc3))
```

We can verify the three PCs are mutually uncorrelated, and the variance of each PC is the corresponding eigenvalues.

Statistical Factor Model

```
summary(lm(data[,1]~pc1+pc2+pc3))
summary(lm(data[,5]~pc1+pc2+pc3))
lm(formula = data[, 5] ~ pc1 + pc2 + pc3)

             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.32290    0.40212  -0.803    0.424
pc1         -0.46613    0.02456 -18.981   <2e-16 ***
pc2         -0.48495    0.03759 -12.902   <2e-16 ***
pc3          0.03609    0.04771   0.756    0.451
```

For instance, we find all three PCs are significant in explaining the first asset. However, for the 5th asset, only the first two PCs are significant.