16 Statistical Factor Analysis with Principal Components

Text: pp. 175-179.

- This is about modeling the cross-sectional correlation across asset returns without an economic theory.
- Useful for determining the number of common factors
- Market only **rewards** for bearing **systematic risk**

16.1 Factor models versus factor analysis

- A one-factor model: A common latent (unobserved) factor drives all returns

$$r_{t,i}^e = \alpha_i + \delta_i f_t + \epsilon_{it}$$

  The δ_i are factor loadings, and can differ across assets. f_t is the common latent factor ϵ_{it} are idiosyncratic components.

- A two-factor model.

$$r_{t,i}^e = \alpha_i + \delta_{1i} f_{t,1} + \delta_{2i} f_{t,2} + \epsilon_{t,i}$$

- **Factor analysis** is a statistical method to describe variability among observed, correlated variables in terms of a smaller number of *unobserved* (latent) variables called factors. The *observed* variables are modelled as linear combinations of the potential factors, plus "error" terms.
- Statistical **Factor** analysis is related to **principal component** analysis (PCA). They are almost the same thing, but not exactly
- **Statistical factor analysis** represents a observations $r_1, r_2, ..., r_n$ in terms of a small number of common factors plus an **idiosyncratic** component. The common factors are unobserved and sometimes referred to as **latent** factors.
- This analysis uses concepts of eigenvalues and eigenvectors from linear algebra. The text by Brooks has a nice review of these concepts.
- If successful in statistically modeling the factor structure, then try to identify those factors in the data.
- **Economic factors:** Build a theory about why asset returns on all sorts of assets $r_1, r_2, ..., r_n$ are **driven** (dependent) on a small set of common factors. These can be quantities (economics) or asset returns (finance).

- Consumption growth, GDP growth
- The market return (CAPM)
- Portfolios of returns sorted from small to big on firm size and sorted on book-value to market-value. (Fama-French factors).

- Statistical factor analysis useful in studying the **term-structure of interest rates**.
- Was popular for study of stock returns, forms the basis of **Arbitrage Pricing Theory**. After discovery of the **Fama-French factors**, it fell out of favor for stock returns. Recently, became useful for studying **exchange rates**.

16.2 The method of Principal Components

- To estimate the factors (which are unobserved) must make **identifying assumptions** (impose identifying restrictions).
- Common to assume common factors are mutually orthogonal (uncorrelated) and are standardized (zero mean, variance 1).
- Several empirical techniques. **Principal components** is the one I teach you.Is oldest and most robust, and very popular.
- Two Factor Structure. For $i = 1, ..., n$

$$r_{t,i} = \delta_{1,i} f_{t,1} + \delta_{2,i} f_{t,2} + r^0_{t,i}$$

That is,

$$\begin{aligned} r_{t,1} &= \delta_{1,1} f_{t,1} + \delta_{2,1} f_{t,2} + r^0_{t,1} \\ r_{t,2} &= \delta_{1,2} f_{t,1} + \delta_{2,2} f_{t,2} + r^0_{t,2} \\ \vdots &= \\ r_{t,n} &= \delta_{1,n} f_{t,1} + \delta_{2,n} f_{t,2} + r^0_{t,n} \end{aligned}$$

- Factors $f_{t,1}, f_{t,2}$ are **common** to each return. $r^0_{t,i}$ is idiosyncratic component of returns.
- δ coefficients are called **factor loadings**.

$$r_{t,i} = \delta_{1,i} f_{t,1} + \delta_{2,i} f_{t,2} + r^0_{t,i}$$

- Procedure chooses PCs sequentially.

- The first PC $f_{1,t}$ and the $\delta_{1,1}, ..., \delta_{1,n}$ are chosen to minimize the sum of squared deviations for every return $r_{t,i}$. Now take each return and control for the first PC.
- The second PC $f_{2,t}$ and the loadings $\delta_{2,1}, \dots, \delta_{2,n}$ are chosen to minimize the sum of squared deviations for every one of these deviations from the first PC.
- If you have n assets, there will be n principal components, but the analysis is only useful if a small set of them explains the data.

- Require factors to be **mutually orthogonal**. Then, sum of squares of the factor loadings tells us the **proportion of variance explained** by the common factors.

$$\delta_{1,i}^2 + \delta_{2,i}^2$$

- Sum of squares of factor 1 loadings tells us the proportion of variance of all returns explained by the first factor

$$\delta_{1,1}^2 + \delta_{1,2}^2 + \ldots + \delta_{1,n}^2$$

- PCs are linear combinations of the data that explain the evolution of the data.
- r is the $T \times n$ matrix containing your data $[r_{t,i}], t = 1, ..., T, i = 1, ..., n$.
- PC describes each individual i by a **linear combination** of a small number of the other variables.

  Start with the **first PC**. It is the $T \times 1$ vector f (is a **time-series**) where

$$\begin{pmatrix} r_{1,1} & \cdots & r_{1,n} \\ r_{2,1} & \cdots & r_{2,n} \\ \vdots & & \vdots \\ r_{T,1} & \cdots & r_{T,n} \end{pmatrix} = \begin{pmatrix} f_1\delta_{1,1} & \cdots & f_1\delta_{1,n} \\ f_2\delta_{1,1} & \cdots & f_2\delta_{1,n} \\ \vdots & & \vdots \\ f_T\delta_{1,1} & \cdots & f_T\delta_{1,n} \end{pmatrix}$$

  In matrix algebra,

$$r = \begin{pmatrix} f_1 \\ f_2 \\ \vdots \\ f_T \end{pmatrix} \begin{pmatrix} \delta_{1,1} & \delta_{1,2} & \ldots & \delta_{1,n} \end{pmatrix} = f\delta'$$

- The **PC is not unique** because we can write

$$r = f\delta' = (fc)(\delta'/c)$$

  for any scalar c. So we **normalize** the observations.

- Sum of squares of every element of $(r - f\delta')$ is

$$Tr(r - f\delta')'(r - f\delta')$$

We want to choose f to **minimize** it. Okay, look. Suppose $n = 2$ and $T = 3$, and we want a single (first) factor. Let

$$\tilde{r} = (r - f\delta)$$

$$\begin{aligned} \tilde{r}'\tilde{r} &= \begin{pmatrix} \tilde{r}_{11} & \tilde{r}_{21} & \tilde{r}_{31} \\ \tilde{r}_{12} & \tilde{r}_{22} & \tilde{r}_{32} \end{pmatrix} \begin{pmatrix} \tilde{r}_{11} & \tilde{r}_{12} \\ \tilde{r}_{21} & \tilde{r}_{22} \\ \tilde{r}_{31} & \tilde{r}_{32} \end{pmatrix} \\ &= \begin{pmatrix} \tilde{r}_{11}^2 + \tilde{r}_{21}^2 + \tilde{r}_{31}^2 & \tilde{r}_{11}\tilde{r}_{12} + \tilde{r}_{21}\tilde{r}_{22} + \tilde{r}_{31}\tilde{r}_{32} \\ \tilde{r}_{11}\tilde{r}_{12} + \tilde{r}_{21}\tilde{r}_{22} + \tilde{r}_{31}\tilde{r}_{32} & \tilde{r}_{12}^2 + \tilde{r}_{22}^2 + \tilde{r}_{32}^2 \end{pmatrix} \end{aligned}$$

Trace of matrix X, $Tr(X)$, is **sum of diagonal elements,**

$$\begin{aligned} &(r_{11} - f_1\delta_1)^2 + (r_{21} - f_2\delta_1)^2 + (r_{31} - f_3\delta_1)^2 + \\ &(r_{12} - f_2\delta_2)^2 + (r_{22} - f_2\delta_2) + (r_{32} - f_3\delta_2)^2 \end{aligned}$$

Principal components wants to choose f_t and δ_i to minimize this thing.

Is like minimizing the residual variance.

The solution is **first principal component.** Label it f_1. It is $T \times 1$ vector.

- To get the **second PC**, control for f_1. Given f_1, choose f_2 to minimize

$$Tr(r - f_1\delta_1' - f_2\delta_2')'(r - f_1\delta_1' - f_2\delta_2')$$

Solution, is the eigen vector associated with the largest eigen value of $(r - f_1\delta_1')(r - f_1\delta_1')'$ is the second PC, which we call f_2.

- There are n principal components. We can keep going on this way all the way to n, but the point of this is to have a low dimension collection of variables to describe the behavior of returns, so in finance, it doesn't make any sense to compute beyond 3 or item.

- PC in Eviews. Note PC only works if $T > n$. We interpret the PCs as the factors. Eviews calls the PCs 'scores'.

- Dow30.wf1. Converted to returns on sheet 2 (r02 through r26)

  - Open the returns as a group

- Click Proc, then click Make Principal Components
- Give score series names: pc1 pc2.
- Give loadings matrix name: loadings
- That is all. Now inspect the loadings and the PCs

16.3 Application to Term Structure of Interest Rates with McCulloch Data

- People like PC for studying the term structure of interest rates. We find 3 PCs. They have an interpretation of level, slope, and curvature.
- McCulloch's Yield Curve: Hu McCulloch was my colleague at Ohio State. He did something like non-linear least squares (cubic splines) to approximate a continuous yield curve.
- Open Mcculoch_TS.wf1. First page
  - Look at graph of i_01. This is time series of the 0-time to maturity bond
    This is close to the rate the Fed controls
- Look at graph of i_482. This is time series of the 482-month to maturity bond.
  - This is the rate that we think is important for investment. Think about mortgage rates.
- Look at graph of both i_01 and i_482.
  - This is the term premium.
  - Distance between curves is the yield curve slope
- Open second workfile page (Transposed)
  - Plot date_90 (February 2004). This is the yield curve on Feb. 2004.
    * Is upward sloping. Normal state is economic growth
  - Plot date_44 (August 2000).
    * Is downward sloping.
    * Recession dates (FRED): https://fred.stlouisfed.org/series/JHDUSRGDPBR
- **PC in EViews**. Go back to the first sheet. **Create group. Write a little program that says**

  ```
  ' Group.prg

  group all_TS
  ```

```
for !j = 1 to 9
all_TS.add i_0{!j}
next
for !j = 10 to 482
all_TS.add i_{!j}
next
```

- **Open the group all_TS**
- Click on **Proc**
- Make 3 principal components
- Choose option to **normalize scores**
- Score series, give names f_1, f_2, f_3 (stand for factors).
- Loadings matrix. Call it **Lambda**

$$i_{i,t} = \lambda_{i,1} f_{1,t} + \lambda_{i,2} f_{2,t} + \lambda_{i,3} f_{3,t} + i^o_{i,t}$$

- Look at elements of lambda.
- Take i_12.
- In Eviews, series fit_12=lambda(12,1)*f_1+lambda(12,2)*f_2+lambda(12,3)*f_3
- Plot fit_12 with i_12
- Regress i_12 on fit_12
- Take row 12 of Lambda. Write to Excel, transpose to get column. Square the elements. 98 percent of i_12 variation is explained by the first 3 principal components.