Factor Models and Principal Components

Professor: Hammou El Barmi
Columbia University

Introduction

- Many financial markets are characterized by a high degree of collinearity between returns.
- Variables are highly collinear when there are only a few important sources of information in the data that are common to many variables.
- High-dimensional data can be challenging to analyze.
- They are difficult to visualize, need extensive computer resources, and often require special statistical methodology.
- Fortunately, in many practical applications, high-dimensional data have most of their variation in a lower-dimensional space that can be found using dimension reduction techniques.
- There are many methods designed for dimension reduction, and in this chapter we will study two closely related techniques, factor analysis and principal components analysis, often called PCA.

Introduction

- PCA finds structure in the covariance or correlation matrix and uses this structure to locate low-dimensional subspaces containing most of the variation in the data. Idea: Extract the most important uncorrelated sources of variation in a multivariate system — principal component analysis (PCA)
- Factor analysis explains returns with a smaller number of fundamental variables called factors or risk factors.
- Factor analysis models can be classified by the types of variables used as factors, macroeconomic or fundamental, and by the estimation technique, time series regression, cross-sectional regression, or statistical factor analysis.

Principal Components Analysis (PCA)

- The other main advantage of PCA is that once you have found these patterns in the data, and you compress the data, ie. by reducing the number of dimensions, without much loss of information
- PCA is concerned with explaining the variance covariance of a set of variables
- This explanation comes from a "few" linear combinations of the original variables
- Generally speaking, PCA has two objectives
  1. Data "reduction": moving from many original variables to a few linear combinations of the original variables
  2. Interpretation: which variables play a larger role in the explanation of the total variance

Principal Components Analysis (PCA)

- Principal components are ordered according to their variances.
- The first principal component is the linear combination that encapsulates most of the variability. In other words, the first principal component represents a rotation of the data along the axis representing the largest spread in the multidimensional cluster of data points.
- The second principal component is the linear combination that explains the most of the remaining variability while being uncorrelated (i.e. perpendicular) to the first principal component.
- If there was a third principal component, it would explains most of the remaining variability while being un- correlated to the first two principal components. This pattern continues for all consecutive principal components.

Principal Components Analysis (PCA)

- Let $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_d)^T$ be a random vector with mean $\boldsymbol{\mu}$ and covariance matrix Σ
- One way to measure the total variability in $\mathbf{Y}$ uses $|\Sigma|$, the determinant of Σ or the trace of Σ defined as

$$trace(\Sigma) = \sum_{i=1}^{d} \sigma_{ii}$$

  (We like to use one number as opposed to too many numbers)
- Notice that $trace(\Sigma) = 0$ if and only if $\sigma_{ii} = 0, \forall i$.

Principal Components Analysis (PCA)

- First principal component is $\mathbf{O}_1^T\mathbf{Y}$ where

$$\mathbf{O}_1 = argmax\{Var(\mathbf{O}^T\mathbf{Y}) = \mathbf{O}^T V\mathbf{O}\}$$

  subject $\mathbf{O}^T\mathbf{O} = 1$.

- The Lagrangian corresponding to this situation is

$$L(\mathbf{O}, \lambda) = \mathbf{O}^T V\mathbf{O} + \lambda(1 - \mathbf{O}^T\mathbf{O})$$

- To find **a** we need to solve

$$\begin{aligned} \frac{\partial}{\partial \mathbf{O}} L(\mathbf{O}, \lambda)) &= V\mathbf{O} - \lambda\mathbf{O} = \mathbf{0} \\ \frac{\partial}{\partial \lambda} L(\mathbf{O}, \lambda)) &= \mathbf{O}^T\mathbf{O} - 1 = 0 \end{aligned}$$

- The solution to these equation is $\mathbf{O}_1$ and λ_1
- Since $V\mathbf{O}_1 = \lambda_1\mathbf{O}_1, \mathbf{O}_1$is an eigen vector of V and λ_1 is its corresponding eigen value.

Principal Components Analysis (PCA)

- Second principal component

$$\mathbf{O}_2 = argmax\{Var(\mathbf{O}^T\mathbf{Y}) = \mathbf{O}^T V\mathbf{O}\}$$

subject $\mathbf{O}_2^T\mathbf{O}_2 = 1$ and $\mathbf{O}_1^T\mathbf{O}_2 = 0$.

- The Lagrangian corresponding to this situation is

$$L(\mathbf{O},\lambda) = \mathbf{O}^T V\mathbf{O} + \lambda(1 - \mathbf{O}^T\mathbf{O}) + \gamma\mathbf{O}^T\mathbf{O}_1$$

- To find $\mathbf{O}_2$ we need to solve

$$\begin{aligned} \frac{\partial}{\partial \mathbf{O}} L(\mathbf{O},\lambda)) &= V\mathbf{O} - \lambda\mathbf{O} = \mathbf{0} \\ \frac{\partial}{\partial \lambda} L(\mathbf{O},\lambda)) &= \mathbf{O}^T\mathbf{O} - 1 = 0 \\ \frac{\partial}{\partial \gamma} L(\mathbf{O},\lambda)) &= \mathbf{O}^T\mathbf{O}_1 = 0 \end{aligned}$$

- The solution to these equation is $\mathbf{O}_2$ and λ_2 and γ_2
- Since $V\mathbf{O}_2 = \lambda_2\mathbf{O}_2, \mathbf{O}_2$is an eigen vector of V and λ_2 is its corresponding eigen value.

Principal Components Analysis (PCA)

- We continue until we get d eigen-values and vectors $(\mathbf{O}_1, \mathbf{O}_2, \ldots, \mathbf{O}_d)^T$ and $(\lambda_1, \lambda_2, \ldots, \lambda_d)$.
- Since Σ is symmetric, its eigenvalues (solutions of the polynomial equation $\det(\Sigma - \lambda I) = 0$) are real and can be ordered as $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_d$
- They are all nonnegative since V is nonnegative definite.
- Moreover

$$\sum_{i=1}^{d} Var(Y_i) \equiv \sum_{i=1}^{d} \sigma_{ii} = \sum_{i=1}^{d} \lambda_i \quad \text{and} \quad det(\Sigma) = \prod_{i=1}^{d} \lambda_i$$

- $\sum_{i=1}^{d} Var(Y_i)$ is called the trace of Σ and is used to measure the total variability in $\mathbf{Y}$.
- $\mathbf{O}_j^T Y$ is called the jth principal component

Principal Components Analysis (PCA)

It turns out that

- $\boldsymbol{\Sigma} = \lambda_1 \mathbf{O}_1 \mathbf{O}_1^T + \lambda_2 \mathbf{O}_2 \mathbf{O}_2^T + \ldots + \lambda_d \mathbf{O}_d \mathbf{O}_d^T$
- $Var(\mathbf{O}_j^T \mathbf{Y}) = \lambda_j, \, j = 1, 2, \ldots, d.$
- We hope that only a few principal components account for most of the overall variance. i.e.

  $$\frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{d} \lambda_i}$$

  is near 1 for small k
- Factor loadings are columns giving the elements of the column vectors $\mathbf{O}_i$s for the principal components $\mathbf{O}_i^T \mathbf{Y}$s
- The factor loading for the first principal component are $(O_{11}, O_{12}, \ldots, O_{1d})^T$

Principal Components Analysis (PCA)

- In practice Σ is not known and is estimated by

$$S = \begin{pmatrix} s_{11} & s_{12} & \dots & s_{1d} \\ s_{21} & s_{21} & \dots & s_{22} \\ \vdots & \vdots & \ddots & \vdots \\ s_{d1} & s_{d2} & \dots & s_{dd} \end{pmatrix}.$$

- If the data are not commensurate, we use the correlation matrix

$$S = \begin{pmatrix} 1 & r_{12} & \dots & r_{1d} \\ r_{21} & 1 & \dots & r_{22} \\ \vdots & \vdots & \ddots & \vdots \\ r_{d1} & r_{d2} & \dots & 1 \end{pmatrix}.$$

where

$$r_{ij} = \frac{s_{ij}}{s_{ii} s_{jj}}, \quad \forall (i,j)$$

Principal Components Analysis (PCA)

- Let

$$\mathbf{Y} = \begin{pmatrix} Y_{11} & Y_{12} & \dots & Y_{1d} \\ Y_{21} & Y_{22} & \dots & Y_{2d} \\ \vdots & \vdots & \dots & \vdots \\ Y_{n1} & Y_{n2} & \dots & Y_{nd} \end{pmatrix} = \begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \\ \vdots \\ \mathbf{Y}_d \end{pmatrix}$$

- The new data is

$$\mathbf{Z} = \begin{pmatrix} Z_{11} & Z_{12} & \dots & Z_{1d} \\ Z_{21} & Z_{22} & \dots & Z_{2d} \\ \vdots & \vdots & \dots & \vdots \\ Z_{n1} & Z_{n2} & \dots & Z_{nd} \end{pmatrix} = \begin{pmatrix} \mathbf{Z}_1 \\ \mathbf{Z}_2 \\ \vdots \\ \mathbf{Z}_d \end{pmatrix}$$

where $Z_{ij} = \mathbf{Y}_i^T \mathbf{O}_j$. That is

$$\mathbf{Z} = \begin{pmatrix} \mathbf{Y}_1^T \mathbf{O}_1 & \mathbf{Y}_1^T \mathbf{O}_2 & \dots & \mathbf{Y}_1^T \mathbf{O}_d \\ \mathbf{Y}_2^T \mathbf{O}_1 & \mathbf{Y}_2^T \mathbf{O}_2 & \dots & \mathbf{Y}_2^T \mathbf{O}_d \\ \vdots & \vdots & \dots & \vdots \\ \mathbf{Y}_n^T \mathbf{O}_1 & \mathbf{Y}_n^T \mathbf{O}_2 & \dots & \mathbf{Y}_n^T \mathbf{O}_d \end{pmatrix}$$

Principal Components Analysis (PCA)

- If the components of $\mathbf{Y}_i$ are comparable (e.g. are all daily returns on equities or are yield on bonds) then working with the original variables should be ok.
- If the components are not comparable (e.g. one in an unemployment rate and another is GDP in dollars), then some variables may many orders of magnitude larger than the others. In such a case, the large variables could completely dominate the PCA so that the first principal component is the direction if the variable with the largest standard deviation. To eliminate this, we should standardize the variables.

Principal Components Analysis (PCA)

- As a simple illustration of the difference between using standardized and unstandardized variables, suppose there are two variables ($d = 2$) with a correlation of 0.9. Then the correlation matrix is

$$\begin{pmatrix} 1\ 0 & .9 \\ 0.9 & 1 \end{pmatrix}$$

- eigenvectors $(0.71, 0.71)^T$ and $(?0.71, 0.71^T\ )$ and eigenvalues 1.9 and 0.1. Most of the variation is in the direction (1, 1), which is consistent with the high correlation between the two variables.

Principal Components Analysis (PCA)

- However, suppose that the first variable has variance 1,000,000 and the second has variance 1. The covariance matrix is

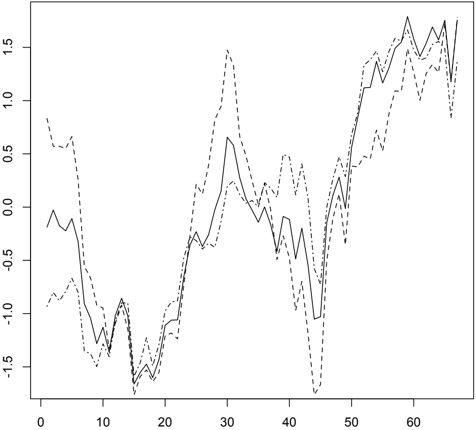
$$\begin{pmatrix} 1,000,000 & 900 \\ 900 & 1 \end{pmatrix}$$

- The eigenvectors, after rounding, equal to (1.0000,0.0009) and (?0.0009, 1) and eigenvalues 1,000,000 and 0.19.
- The first variable dominates the principal components analysis based on the covariance matrix.
- This principal components analysis does correctly show that almost all of the variation is in the first variable, but this is true only with the original units.
- Suppose that variable 1 had been in dollars and is now converted to millions of dollars.
- Then its variance is equal to 10^{-6}, so that the principal components analysis using the covariance matrix will now show most of the variation to be due to variable 2.
- In contrast, principal components analysis based on the correlation matrix does not change as the variables units change.

Principal Components Analysis (PCA)

Here we look the the indices SP500, DOW JONES and NASDAQ (daily historical prices from 8/7/13-11/8/2013)

Figure: Indices SP500 Dow Jones and Nasdaq



Principal Components Analysis (PCA)

Standard deviations:

1.6506196 0.5151970 0.1001344

Rotation:

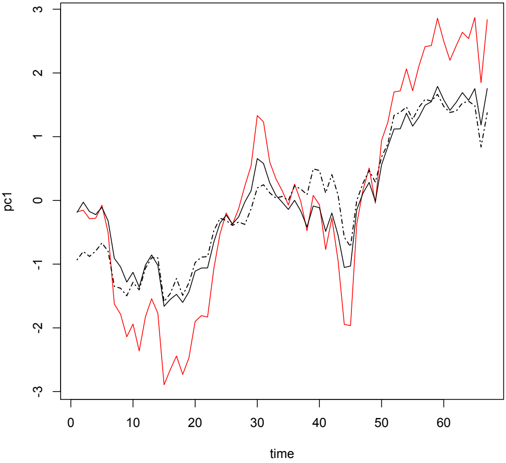
	PC1	PC2	PC3
sp500	0.6033967	-0.08039077	-0.7933787
dowjones	0.5579345	0.75339939	0.3479921
nasdaq	0.5697556	-0.65263058	0.4994515

The principal components are

$$\begin{aligned} PC1 &= 0.6033967\text{sp500} + 0.5579345\text{dowjones} + 0.5697556\text{nasdaq} \\ PC2 &= -0.08039077\text{sp500} + 0.75339939\text{dowjones} - 0.65263058\text{nasdaq} \\ PC3 &= -0.7933787\text{sp500} + 0.3479921\text{dowjones} + 0.4994515\text{nasdaq} \end{aligned}$$

Principal Components Analysis (PCA)

Figure: Indices SP500 Dow Jones and Nasdaq



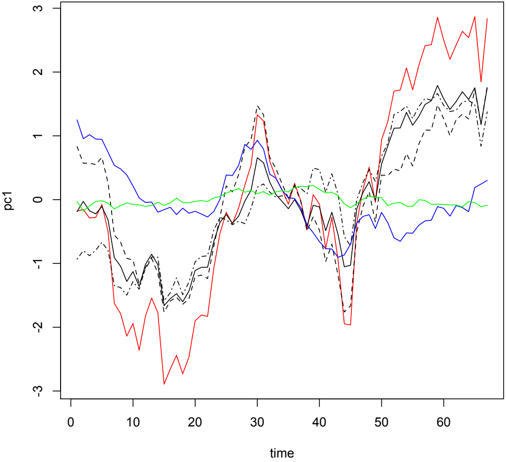
Principal Components Analysis (PCA)

Importance of components:

	PC1	PC2	PC3
Standard deviation	1.6506	0.51520	0.10013
Proportion of Variance	0.9082	0.08848	0.00334
Cumulative Proportion	0.9082	0.99666	1.00000

Principal Components Analysis (PCA)

Figure: Indices SP500 Dow Jones and Nasdaq. PC1 (red), PC2 (blue) and PC3 (green)



- A factor model for excess equity returns is

  $$Rj, t = \beta_{0,j} + \beta_{1,j} F_{1,t} + \ldots + \beta_{p,j} F_{p,t} + \epsilon_{j,t}$$

  where $R_{j,t}$ is either the return or the excess return on the *jth* asset at time $t, F_{1,t}, \ldots, F_{p,t}$ are variables, called factors or risk factors, that represent the state of the financial markets and world economy at time t
- $\epsilon_{1,t}, \ldots, \epsilon_{n,t}$ are uncorrelated, mean-zero random variables called the unique risks of the individual stocks.
- The assumption that unique risks are uncorrelated means that all cross-correlation between the returns is due to the factors
- Notice that the factors do not depend on j since they are common to all returns
- The parameter β_{ij} is called a factor loading and specifies the sensitivity of the return of the *jth* asset to the *ith* factor
- Depending on the type of factor model, either the loadings, the factors, or both the factors and the loadings are unknown and must be estimated.

Factor Models)

- The CAPM is a factor model where $p = 1$ and $F_{1,t}$ is the excess return on the market portfolio.
- In the CAPM, the market risk factor is the only source of risk besides the unique risk of each asset.
- Because the market risk factor is the only risk that any two assets share, it is the sole source of correlation between asset returns.
- Factor models generalize the CAPM by allowing more factors than simply the market risk and the unique risk of each asset.

Factor Models)

- 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);

Factor Models

This example uses the berndtInvest data set in Rs fEcofin package. This data set has monthly returns on 15 stocks over 10 years, 1978 to 1987. The 15 stocks were classified into three industries,Tech, Oil, and Other, as follows:

	tech	oil	other
CITCRP	0	0	1
CONED	0	0	1
CONTIL	0	1	0
DATGEN	1	0	0
DEC	1	0	0
DELTA	0	1	0
GENMIL	0	0	1
GERBER	0	0	1
IBM	1	0	0
MOBIL	0	1	0
PANAM	0	1	0
PSNH	0	0	1
TANDY	1	0	0
TEXACO	0	1	0
WEYER	0	0	1

Factor Models

- One possible model is a model that uses tech and oil as loadings and fit the model

  $$R_j = \beta_0 + \beta_1 \text{tech}_j + \beta_2 \text{oil}_j + \epsilon_j$$

  where R_j is the return on the *jth* stock, tech_j equals 1 if the *jth* stock is a technology stock and equals 0 otherwise, and oil_j is defined similarly.
- the value of β_0, can be viewed as an overall market factor, since it affects all 15 returns. Factors 2 and 3 are the technology and oil factors. For example, if the value of β_2 is positive in any given month, then Tech stocks have better-than-market returns that month.