FINM 331: DATA ANALYSIS FOR FINANCE AND STATISTICS
FALL 2015
FACTOR ANALYSIS

1. POPULATION FACTOR ANALYSIS

- as in population PCA, what we discuss here are the statistical principles behind the actual technique that works on samples/data
- when we say 'population blah blah,' it would be a discussion about random variables $X_1, \dots, X_p$ but when we say 'sample blah blah,' it would about actual data $X \in \mathbb{R}^{n \times p}$
- we begin with the famous *factor model*

$$\begin{aligned} X_1 &= \mu_1 + \ell_{11}F_1 + \ell_{12}F_2 + \cdots + \ell_{1m}F_m + \varepsilon_1 \\ X_2 &= \mu_2 + \ell_{21}F_1 + \ell_{22}F_2 + \cdots + \ell_{2m}F_m + \varepsilon_2 \\ &\vdots \\ X_p &= \mu_p + \ell_{p1}F_1 + \ell_{p2}F_2 + \cdots + \ell_{pm}F_m + \varepsilon_p \end{aligned}$$

  where $X_1, \dots, X_p$; $F_1, \dots, F_m$; $\varepsilon_1, \dots, \varepsilon_p$ are random variables and $\mu_1, \dots, \mu_p \in \mathbb{R}$; $\ell_{11}, \dots, \ell_{pm} \in \mathbb{R}$ are constants
- m is usually much smaller than p, so that the factor model does is an attempt to use a small set of random variables $F_1, \dots, F_m$ to explain a large set of random variables $X_1, \dots, X_p$
- as usual we introduce random vectors

$$\mathbf{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{bmatrix}, \qquad \mathbf{F} = \begin{bmatrix} F_1 \\ F_2 \\ \vdots \\ F_m \end{bmatrix}, \qquad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_p \end{bmatrix}$$

  and constant vectors and matrices

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix} \in \mathbb{R}^p, \qquad L = \begin{bmatrix} \ell_{11} & \ell_{12} & \cdots & \ell_{1p} \\ \ell_{21} & \ell_{22} & & \vdots \\ \vdots & & \ddots & \vdots \\ \ell_{p1} & \cdots & \cdots & \ell_{pp} \end{bmatrix} \in \mathbb{R}^{p \times m}$$

- the factor model may now be written as

$$\mathbf{X} = \boldsymbol{\mu} + L\mathbf{F} + \boldsymbol{\varepsilon}$$

  or some people prefer

$$\mathbf{X} - \boldsymbol{\mu} = L\mathbf{F} + \boldsymbol{\varepsilon}$$

- now the names
  - the random variables $F_1, \dots, F_m$ are called *factors* or *common factors*
  - the random variables $\varepsilon_1, \dots, \varepsilon_p$ are called *errors* or *specific factors*
  - the constant matrix $L \in \mathbb{R}^{p \times m}$ is called the *matrix of factor loadings*
- we will need some mild assumptions if we want to do anything with the factor model

Date: November 7, 2015, version 1.0. Comments, bug reports: `lekheng@galton.uchicago.edu`.

- the common factors $F_1, \dots, F_m$ have zero mean, unit variance, and are uncorrelated with each other

$$E(\mathbf{F}) = \mathbf{0} \in \mathbb{R}^m, \qquad \operatorname{Cov}(\mathbf{F}) = E(\mathbf{F}\mathbf{F}^\mathsf{T}) = I_m \in \mathbb{R}^{m \times m}$$

- the specific factors $\varepsilon_1, \dots, \varepsilon_p$ have zero mean and are uncorrelated with each other

$$E(\boldsymbol{\varepsilon}) = \mathbf{0} \in \mathbb{R}^p, \qquad \operatorname{Cov}(\boldsymbol{\varepsilon}) = E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\mathsf{T}) = \begin{bmatrix} \psi_1 & & & \\ & \psi_2 & & \\ & & \ddots & \\ & & & \psi_p \end{bmatrix} =: \Psi \in \mathbb{R}^{p \times p}$$

- the common factors $F_1, \dots, F_m$ and the specific factors $\varepsilon_1, \dots, \varepsilon_p$ are uncorrelated with each other too

$$\operatorname{Cov}(\boldsymbol{\varepsilon}, \mathbf{F}) = E(\boldsymbol{\varepsilon}\mathbf{F}^\mathsf{T}) = O \in \mathbb{R}^{p \times m}$$

- we will see that these assumptions have some implications on the covariance structure of $\mathbf{X}$ that would ultimately allow us to estimate the loading matrix from just observations of $\mathbf{X}$
- note that this is pretty amazing, we do not know anything about $\boldsymbol{\mu}$, L, $\mathbf{F}$, and $\boldsymbol{\varepsilon}$ — that's a lot of unknowns — but nonetheless we can still find L based on observations of $\mathbf{X}$ with the mild assumptions above

2. Aside: finance interlude

- factor models are very widely used in financial applications
- the following are some oft-used terminologies in finance to help you relate what we discussed
- $X_1, \dots, X_p$ are *investment/asset returns* (or returns in excess of risk-free rate)
- $\mu_1, \dots, \mu_p$ are called the *alphas* and they have the following interpretations
  - $\mu_i < 0$ means the ith investment return is too little given the amount of assumed risk
  - $\mu_i = 0$ means the ith investment return is adequate given the amount of assumed risk
  - $\mu_i > 0$ means the ith investment return is in excess given the amount of assumed risk
- $\ell_{11}, \dots, \ell_{pm}$ are called the *betas* or more specifically ℓ_{ij} is called the *factor beta* of the ith asset on the jth factor, quantifying risk from exposure to general market movements
  - $\ell_{ij} > 1$: highly correlated with market movements, i.e., moves up or down as market moves up or down
  - $0 \leq \ell_{ij} < 1$: mildy or not correlated with market movements
  - $\ell_{ij} < 0$: negatively correlated with market moves, i.e., moves down or up as market moves up or down
- $F_1, \dots, F_m$ often include *macroeconomic factors*, examples include
  - inflation rate
  - treasury bill rate
  - return on long-term government bonds
  - market indices
  - industrial production
  - consumption
  - oil prices
- there are three common types of factor models in finance:
  - *macroeconomic* factor model: $F_1, \dots, F_m$ are observable economic and financial time series (e.g. Sharpe single factor model)
  - *fundamental* factor model: $F_1, \dots, F_m$ are observable asset characteristics (BARRA single factor model, BARRA industry factor model, Fama-French three-factor model)
  - *statistical* factor model: $F_1, \dots, F_m$ are unobservable, we deduce information about them from the asset returns $X_1, \dots, X_p$ (this is the case we are interested in)

- financial data are usually in the form of time series and so all the random variables are in addition also indexed by time

$$\begin{aligned} X_{1t} &= \mu_1 + \ell_{11}F_{1t} + \ell_{12}F_{2t} + \cdots + \ell_{1m}F_{mt} + \varepsilon_{1t} \\ X_{2t} &= \mu_2 + \ell_{21}F_{1t} + \ell_{22}F_{2t} + \cdots + \ell_{2m}F_{mt} + \varepsilon_{2t} \\ &\vdots \\ X_{pt} &= \mu_p + \ell_{p1}F_{1t} + \ell_{p2}F_{2t} + \cdots + \ell_{pm}F_{mt} + \varepsilon_{pt} \end{aligned}$$

  or

$$\mathbf{X}_t = \boldsymbol{\mu} + L\mathbf{F}_t + \boldsymbol{\varepsilon}_t$$

- see the file `zivot.pdf` for further information if you're interested

3. COVARIANCE STRUCTURE

- note that $\mathrm{Cov}(\mathbf{X}) \in \mathbb{R}^{p\times p}$ requires $p(p+1)/2$ parameters (it is a symmetric matrix) to describe whereas $\mathrm{Cov}(\mathbf{F}) \in \mathbb{R}^{m\times m}$ requires $m(m+1)/2$ parameters (again symmetric matrix) to describe
- since m is small relative to p, $m(m+1)/2$ is much smaller than $p(p+1)/2$ and if we could explain the covariation in $\mathbf{X}$ with that in $\mathbf{F}$, we would use much fewer parameters
- this would be our goal but first we describe the relation between $\mathrm{Cov}(\mathbf{X})$ and $\mathrm{Cov}(\mathbf{F})$

$$\begin{aligned} \Sigma = \mathrm{Cov}(\mathbf{X}) &= E(\mathbf{X}-\boldsymbol{\mu})(\mathbf{X}-\boldsymbol{\mu})^\mathsf{T} = E(L\mathbf{F}+\boldsymbol{\varepsilon})(L\mathbf{F}+\boldsymbol{\varepsilon})^\mathsf{T} \\ &= E(L\mathbf{F}\mathbf{F}^\mathsf{T}L^\mathsf{T} + \boldsymbol{\varepsilon}\mathbf{F}^\mathsf{T}L^\mathsf{T} + L\mathbf{F}\boldsymbol{\varepsilon}^\mathsf{T} + \boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\mathsf{T}) \\ &= LE(\mathbf{F}\mathbf{F}^\mathsf{T})L^\mathsf{T} + E(\boldsymbol{\varepsilon}\mathbf{F}^\mathsf{T})L^\mathsf{T} + LE(\mathbf{F}\boldsymbol{\varepsilon}^\mathsf{T}) + E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\mathsf{T}) \\ &= LI_mL^\mathsf{T} + O_{p\times m}L^\mathsf{T} + LO_{m\times p} + \Psi \\ &= LL^\mathsf{T} + \Psi \end{aligned}$$

- this says

$$\begin{aligned} \mathrm{Var}(X_i) = \sigma_{ii} &= \ell_{i1}^2 + \cdots + \ell_{im}^2 + \psi_i, \qquad i = 1,\ldots,p \\ \mathrm{Cov}(X_i, X_j) = \sigma_{ij} &= \ell_{i1}\ell_{j1} + \cdots + \ell_{im}\ell_{jm}, \qquad i = 1,\ldots,p,\ j = 1,\ldots,m,\ i \neq j, \end{aligned}$$

- we have the following names

$$\begin{aligned} \text{communality} &= \ell_{i1}^2 + \cdots + \ell_{im}^2 \\ \text{uniqueness} &= \psi_i \end{aligned}$$

- the expression of $\mathrm{Var}(X_i)$ into a sum of *communality* and *uniqueness* is often called the decomposition of variance
  - communality, also known as common variance, captures the part of the variance shared among all $X_1, \ldots, X_p$
  - uniqueness, also known as specific variance, captures the part of the variance unique or specific to X_i
- a simple calculation shows that the loading matrix L measures precisely the covariance of $\mathbf{X}$ and $\mathbf{F}$

$$\begin{aligned} \mathrm{Cov}(\mathbf{X},\mathbf{F}) &= E(\mathbf{X}-\boldsymbol{\mu})\mathbf{F}^\mathsf{T} = E(L\mathbf{F}+\boldsymbol{\varepsilon})\mathbf{F}^\mathsf{T} = E(L\mathbf{F}\mathbf{F}^\mathsf{T} + \boldsymbol{\varepsilon}\mathbf{F}^\mathsf{T}) \\ &= LE(\mathbf{F}\mathbf{F}^\mathsf{T}) + E(\boldsymbol{\varepsilon}\mathbf{F}^\mathsf{T}) = LI_m + O_{p\times m}L^\mathsf{T} \\ &= L \end{aligned}$$

- this says that

$$\mathrm{Cov}(X_i, F_j) = \ell_{ij}, \qquad i = 1,\ldots,p,\ j = 1,\ldots,m$$

- the most important equation for us is

$$\boxed{\Sigma = LL^{\mathsf{T}} + \Psi} \tag{3.1}$$

  which we will use to estimate $L \in \mathbb{R}^{p\times m}$ and $\Psi = \operatorname{diag}(\psi_1, \dots, \psi_p) \in \mathbb{R}^{p\times p}$
- now the problem is to find L and Ψ given Σ, this problem is unfortunately ill-posed[1] in both sense of the word:
  - existence: Σ may not have the requisite form if m is chosen too small
    - by (**??**), $\operatorname{rank}(\Sigma - \Psi) = \operatorname{rank}(LL^{\mathsf{T}}) = m$ so if $\operatorname{rank}(\Sigma - \Psi) < m$, then we don't have a solution
    - by (**??**), $\Sigma - \Psi = LL^{\mathsf{T}}$ is positive semidefinite so if $\Sigma - \Psi$ is not positive semidefinite, then we don't have a solution
  - uniqueness: note that for any orthogonal matrix $Q \in \mathbb{R}^{m\times m}$, since $QQ^{\mathsf{T}} = I$, we always have that

$$\Sigma = LL^{\mathsf{T}} + \Psi = LQQ^{\mathsf{T}}L^{\mathsf{T}} + \Psi = (LQ)(LQ)^{\mathsf{T}} + \Psi$$

  - in other words, if L is a solution to (**??**), then LQ is also a solution for any orthogonal matrix Q and we have infinitely many possible choices
  - this nonuniqueness is called *rotation ambiguity* or *factor indeterminacy*
- there are various ways to deal with these issues but none are completely satisfactory

4. sample factor analysis

- instead of random variables $X_1, \dots, X_p$ we have sample data $\mathbf{x}_1, \dots, \mathbf{x}_n \in \mathbb{R}^p$
- again, as in the case of population vs sample PCA, the relation between population and sample factor analysis is that we regard

$$x_{ij} = i\text{th observation of } X_j$$

  for $i = 1, \dots, n$ and $j = 1, \dots, p$
- and, as in the case of PCA, we shall use the sample covariance matrix

$$S = \sum_{i=1}^{n} (\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})^{\mathsf{T}} \in \mathbb{R}^{p\times p}$$

  in place of the population covariance matrix $\Sigma = \operatorname{Cov}(\mathbf{X}) \in \mathbb{R}^{p\times p}$
- so the basic problem is: given a covariance matrix $S \in \mathbb{R}^{p\times p}$, find $L \in \mathbb{R}^{p\times m}$ and $\Psi = \operatorname{diag}(\psi_1, \dots, \psi_p) \in \mathbb{R}^{p\times p}$ such that

$$\boxed{S = LL^{\mathsf{T}} + \Psi}$$

- again this is an impossible problem in general: there are too many unknowns
- so we approximate, i.e., replace $=$ by $\approx$, and use some heuristics to make it solvable
- the main and most common method is called the *principal components solution* although *principal components approximations* would have been more appropriate
- as the name implies, it relies on sample PCA and the method goes as follows
  (1) fix $m < p$
  (2) compute the EVD of S

$$S = Q\Lambda Q^{\mathsf{T}} \tag{4.1}$$

  where $Q = [\mathbf{q}_1, \dots, \mathbf{q}_p] \in \mathbb{R}^{p\times p}$ is orthogonal and $\Lambda = \operatorname{diag}(\lambda_1, \dots, \lambda_p) \in \mathbb{R}^{p\times p}$ with $\lambda_1 \geq \cdots \geq \lambda_p$

[1] A problem is *well-posed* if a solution exists and is unique, if either of these conditions fail, then it's called *ill-posed*.

(3) set

$$Q_m = [\mathbf{q}_1, \dots, \mathbf{q}_m] \in \mathbb{R}^{p \times m}, \qquad \Lambda_m = \operatorname{diag}(\lambda_1, \dots, \lambda_m) \in \mathbb{R}^{m \times m}$$

i.e., truncating Q and Λ to just the first m columns of Q and the first m diagonal entries of Λ

(4) set

$$L = Q_m \Lambda_m^{1/2} = [\sqrt{\lambda_1}\mathbf{q}_1, \dots, \sqrt{\lambda_m}\mathbf{q}_m] \in \mathbb{R}^{p \times m}$$

and set

$$\Psi = \operatorname{diag}(S - LL^{\mathsf{T}}) \in \mathbb{R}^{p \times p}$$

- note that when we write $\operatorname{diag}(A)$ for some matrix A we mean the diagonal matrix whose diagonal is that of A

$$\operatorname{diag}\left(\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1p} \\ a_{21} & a_{22} & & \vdots \\ \vdots & & \ddots & \vdots \\ a_{p1} & \cdots & \cdots & a_{pp} \end{bmatrix}\right) = \begin{bmatrix} a_{11} & & & \\ & a_{22} & & \\ & & \ddots & \\ & & & a_{pp} \end{bmatrix}$$

and when we write $\operatorname{diag}(a_1, \dots, a_p)$ we mean the diagonal matrix whose diagonal elements are $a_1, \dots, a_p$

$$\operatorname{diag}(a_1, \dots, a_p) = \begin{bmatrix} a_1 & & & \\ & a_2 & & \\ & & \ddots & \\ & & & a_p \end{bmatrix}$$

- in other words, the principal components solution is given by:
  - L is the matrix whose columns are the first m eigenvectors of S scaled by the positive square roots of the corresponding eigenvalues
  - Ψ is the diagonal matrix $\operatorname{diag}(\psi_1, \dots, \psi_p)$ whose diagonal entries are

$$\psi_i = s_{ii} - \sum_{j=1}^{m} \ell_{ij}^2, \quad i = 1, \dots, p$$

- note that this is only an approximation

$$S \approx LL^{\mathsf{T}} + \Psi$$

since the *residual*

$$E = S - LL^{\mathsf{T}} - \Psi$$

has zero on the diagonals but in general has nonzero off-diagonal entries

- why is this a reasonable solution?
- the answer is Eckhart–Young theorem: note that the EVD and the SVD of a symmetric positive semidefinite matrix are one and the same thing, so (**??**) is also the SVD of the covariance matrix $S \in \mathbb{R}^{p \times p}$ and we may apply Eckhart-Young theorem to see that

$$LL^{\mathsf{T}} = Q_m \Lambda_m^{1/2} (Q_m \Lambda_m^{1/2})^{\mathsf{T}} = Q_m \Lambda_m Q_m^{\mathsf{T}}$$

is a best rank-m approximation to Σ

- this gives us the following theorem (cf. Eckhart–Young theorem from handout)

Theorem 1. *Let $E = S - LL^{\mathsf{T}} - \Psi \in \mathbb{R}^{p \times p}$ be the residual, i.e., the error in approximation $S \approx LL^{\mathsf{T}} + \Psi$, where L and Ψ are given by the principal components solution. Then*

$$\|E\|_2 \leq \lambda_{m+1} \quad \textit{and} \quad \|E\|_F \leq \sqrt{\lambda_{m+1}^2 + \cdots + \lambda_p^2}$$

- we may define similar quantities as in sample PCA, e.g. the proportion of total sample variance due to the jth factor is

$$\frac{\lambda_j}{s_{11} + \cdots + s_{pp}}$$

the proportion of total sample variance due to the ith and jth factors is

$$\frac{\lambda_i + \lambda_j}{s_{11} + \cdots + s_{pp}}$$

and so on
- these are usually expressed as percentages
- we may also do factor analysis with the correlation matrix in place of the covariance matrix as in the case of sample PCA
- again this is equivalent to first standardizing our data $\mathbf{x}_1, \dots, \mathbf{x}_n$,

$$\mathbf{z}_i = D^{-1/2}(\mathbf{x}_i - \overline{\mathbf{x}}) = \begin{bmatrix} \dfrac{x_{i1} - \overline{x}_1}{\sqrt{s_{11}}} \\ \dfrac{x_{i2} - \overline{x}_2}{\sqrt{s_{22}}} \\ \vdots \\ \dfrac{x_{ip} - \overline{x}_p}{\sqrt{s_{pp}}} \end{bmatrix} \in \mathbb{R}^p, \qquad i = 1, \dots, n$$

- recall that transforming

$$X = \begin{bmatrix} \mathbf{x}_1^\mathsf{T} \\ \mathbf{x}_2^\mathsf{T} \\ \vdots \\ \mathbf{x}_n^\mathsf{T} \end{bmatrix} \longrightarrow Z = \begin{bmatrix} \mathbf{z}_1^\mathsf{T} \\ \mathbf{z}_2^\mathsf{T} \\ \vdots \\ \mathbf{z}_n^\mathsf{T} \end{bmatrix}$$

has the effect of transforming

$$S = \frac{1}{n-1} \sum_{i=1}^{n} (\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})^\mathsf{T} \longrightarrow R = \frac{1}{n-1} \sum_{i=1}^{n} \mathbf{z}_i \mathbf{z}_i^\mathsf{T}$$

i.e, the sample covariance matrix of Z is the sample correlation matrix of X
- we use R in place of S in (**??**)
- there is another popular method called the *maximum likelihood method for factor analysis* that we would briefly describe but not discuss in detail since it relies on optimization techniques that are beyond the scope of this course (take Stat 310 if you want to learn this)
- this does the following
  (1) compute t_{ii}, the ith diagonal entry of S^{-1} by solving

$$S\mathbf{t} = \mathbf{e}_i, \quad i = 1, \dots, p$$

  and taking the ith entry of the solution vector (why does this work?)
  (2) set

$$\widehat{\psi}_i = \left(1 - \frac{m}{2p}\right) \frac{1}{t_{ii}}$$

  and fset $\widehat{\Psi} = \operatorname{diag}(\widehat{\psi}_1, \dots, \widehat{\psi}_p)$
  (3) set

$$\widehat{S} = \widehat{\Psi}^{-1/2} S \widehat{\Psi}^{-1/2}$$

  (4) compute the EVD of $\widehat{S}$

$$\widehat{S} = Q \Lambda Q^\mathsf{T}$$

(5) set

$$\widehat{L} = \widehat{\Psi}^{1/2} Q_m (\Lambda_m - I)^{1/2}$$

where

$$Q_m = [\mathbf{q}_1, \ldots, \mathbf{q}_m] \in \mathbb{R}^{p \times m}, \qquad \Lambda_m = \operatorname{diag}(\lambda_1, \ldots, \lambda_m) \in \mathbb{R}^{m \times m}$$

(6) solve the optimization problem

$$\underset{\Psi = \operatorname{diag}(\psi_1, \ldots, \psi_p)}{\text{minimize}} \log\det(\widehat{L}\widehat{L}^{\mathsf{T}} + \Psi) + \operatorname{tr}(\widehat{L}\widehat{L}^{\mathsf{T}} + \Psi)^{-1} S \tag{4.2}$$

to get $\psi_1, \ldots, \psi_p$

(7) repeat steps 2–6 with $\psi_1, \ldots, \psi_p$ in place of $\widehat{\psi}_1, \ldots, \widehat{\psi}_p$ until convergence

- this method assumes that $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are sampled from a normal distribution and the optimization problem (**??**) comes from applying the maximum likelihood method to estimate the covariance matrix Σ subjected to assumptions
  (i) $\Sigma = LL^{\mathsf{T}} + \Psi$ with Ψ a diagonal matrix
  (ii) $L^{\mathsf{T}}\Psi^{-1}L$ is a diagonal matrix
- the two assumptions together imply that

$$(\Psi^{-1/2} S \Psi^{-1/2})(\Psi^{-1/2} L) = (\Psi^{-1/2} L)(I + L^{\mathsf{T}}\Psi^{-1} L)$$

which motivates steps 3 and 5
- we will not use this method in our class