

## Factor Analysis

Factor analysis is based on a specific model (the factor analytic model) that

- "Explains" the *covariances* or *correlations* between variables in terms of their dependence on one or more underlying *unobservable* or *latent* variables - the factors.
- Attempts to identify and understand the factors that influence the observed variables.

Factor analysis has a lot of similarity to principal components analysis when you view it as a technique to approximate random variables in terms of fewer random variables - that is as a dimension reduction technique.

That's where I begin.

2

Displays for Statistics 5401/8401

Lecture 27

November 9, 2005

Christopher Bingham, Instructor

612-625-1024, kb@umn.edu  
372 Ford Hall

Class Web Page

<http://www.stat.umn.edu/~kb/classes/5401>  
© 2005 by Christopher Bingham

Suppose  $\mathbf{x} = [x_1, x_2, \dots, x_p]'$  is a random vector with mean  $E[\mathbf{x}] = \boldsymbol{\mu}$  and variance matrix  $V[\mathbf{x}] = \boldsymbol{\Sigma}$ .

As usual  $\mathbf{v}_1, \dots, \mathbf{v}_p$  are the eigenvectors and  $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p$  the eigenvalues of  $\boldsymbol{\Sigma} = V[\mathbf{x}]$ .

I start by representing  $\mathbf{x}$  and  $\boldsymbol{\Sigma}$  in terms of the principal components structure.

You can exactly represent  $\mathbf{x}$  in terms of all  $p$  principal components  $z_1, \dots, z_p$  as

$$\mathbf{x} = \boldsymbol{\mu} + \sum_{1 \leq j \leq p} z_j \mathbf{v}_j.$$

- $z_1, \dots, z_p, z_j = \mathbf{v}_j'(\mathbf{x} - \boldsymbol{\mu})$  are uncorrelated population principal components which have  $\mu_{z_j} = 0$  and  $\sigma_{z_j}^2 = \lambda_j$ .

The  $z_j$ 's are random variables. and element  $v_{kj}$  of  $\mathbf{v}_j$  is coefficient of  $z_j$  in  $x_k$ .

For any  $m < p$ , you can split this up as

$$\begin{aligned} \mathbf{x} &= \boldsymbol{\mu} + \sum_{1 \leq j \leq m} z_j \mathbf{v}_j + \sum_{m+1 \leq j \leq p} z_j \mathbf{v}_j \\ &= \boldsymbol{\mu} + \mathbf{x}^{(m)} + \boldsymbol{\epsilon} \end{aligned}$$

- $\mathbf{x}^{(m)} = \sum_{1 \leq j \leq m} z_j \mathbf{v}_j$  is the part of  $\mathbf{x}$  "explained" by the the first  $m$   $z_j$ 's (PC's).
- $\boldsymbol{\epsilon} \equiv \sum_{m+1 \leq j \leq p} z_j \mathbf{v}_j$  is the part of  $\mathbf{x}$  *not* "explained" by the first  $m$   $z_j$ 's.

Because  $z_1, \dots, z_p$  are uncorrelated,

- $\boldsymbol{\epsilon}$  is uncorrelated with  $\mathbf{x}^{(m)}$
- $\boldsymbol{\Sigma}^{(m)} \equiv V[\mathbf{x}^{(m)}] = \sum_{1 \leq j \leq m} \lambda_j \mathbf{v}_j \mathbf{v}_j'$
- $V[\boldsymbol{\epsilon}] = \sum_{m+1 \leq j \leq p} \lambda_j \mathbf{v}_j \mathbf{v}_j'$
- $\boldsymbol{\Sigma} = V[\mathbf{x}] = \sum_{1 \leq j \leq p} \lambda_j \mathbf{v}_j \mathbf{v}_j'$   
 $= \sum_{1 \leq j \leq m} \lambda_j \mathbf{v}_j \mathbf{v}_j' + \sum_{1 \leq j \leq p} \lambda_j \mathbf{v}_j \mathbf{v}_j'$   
 $= \boldsymbol{\Sigma}^{(m)} + \boldsymbol{\Sigma} - \boldsymbol{\Sigma}^{(m)}$   
 $= V[\mathbf{x}^{(m)}] + V[\boldsymbol{\epsilon}].$

$\boldsymbol{\Sigma}^{(m)}$  has rank  $m$ .  $V[\boldsymbol{\epsilon}]$  has rank  $p - m$ .

Here's some new notation to replace the PC notation.

Define the random vector of  $m$  "factors":

- $\mathbf{f} = [f_1, \dots, f_m]'$ ,  
 $m \times 1$
- $f_j \equiv z_j / \sqrt{\lambda_j} = \mathbf{v}_j'(\mathbf{x} - \boldsymbol{\mu}) / \sqrt{\lambda_j}$   
 $f_j$  is standardized PC  $z_j$

and the  $p$  by  $m$  "loading" matrix

- $\mathbf{L} = [\mathbf{l}_{kj}] \equiv [\sqrt{\lambda_1} \mathbf{v}_1, \sqrt{\lambda_2} \mathbf{v}_2, \dots, \sqrt{\lambda_m} \mathbf{v}_m]$   
 $p \times m$   
 $\equiv [\mathbf{l}_1, \mathbf{l}_2, \dots, \mathbf{l}_m], \mathbf{l}_{kj} = \sqrt{\lambda_j} \mathbf{v}_{kj}$

Then  $f_j \mathbf{l}_j = (z_j / \sqrt{\lambda_j})(\sqrt{\lambda_j} \mathbf{v}_j) = z_j \mathbf{v}_j$  and

$$\begin{aligned} \mathbf{x} &= \boldsymbol{\mu} + \mathbf{x}^{(m)} + \boldsymbol{\varepsilon} \\ &= \boldsymbol{\mu} + \sum_{1 \leq j \leq m} z_j \mathbf{v}_j + \boldsymbol{\varepsilon} \\ &= \boldsymbol{\mu} + \sum_{1 \leq j \leq m} f_j \mathbf{l}_j + \boldsymbol{\varepsilon} \end{aligned}$$

In terms of matrices this is

$$\mathbf{x} = \boldsymbol{\mu} + \mathbf{L} \mathbf{f} + \boldsymbol{\varepsilon}$$

$p \times 1 \quad p \times 1 \quad p \times m \quad m \times 1 \quad p \times 1$

The matrix of *loadings* is

$$\mathbf{L} = \begin{bmatrix} l_{11} & l_{12} & \dots & l_{1m} \\ \dots & \dots & \dots & \dots \\ l_{k1} & l_{k2} & \dots & l_{km} \\ \dots & \dots & \dots & \dots \\ l_{p1} & l_{p2} & \dots & l_{pm} \end{bmatrix}$$

Loadings for  $x_1$   
 .....  
 Loadings for  $x_k$   
 .....  
 Loadings for  $x_p$

Loadings on  $f_1 \quad f_2 \quad \dots \quad f_m$ .

Elements  $l_{kj}$  of  $\mathbf{L}$  characterize the dependence of  $\mathbf{x}$  on the common factors.

- $[l_{k1}, l_{k2}, \dots, l_{km}]$  (row  $k$  of  $\mathbf{L}$ ) goes with variable  $x_k$ , and characterizes how  $x_k$  is affected by each factor
- $\begin{bmatrix} l_{1j} \\ l_{2j} \\ l_{3j} \\ \dots \\ l_{pj} \end{bmatrix}$  (column of  $\mathbf{L}$ ) goes with factor  $f_j$  and characterizes how  $f_j$  affects each  $x_k, 1 \leq k \leq p$

For variable  $x_k$  (element of  $\mathbf{x}$ ) this is

$$x_k = \mu_k + \sum_{1 \leq j \leq m} l_{kj} f_j + \varepsilon_k.$$

Variable  $x_k$  "loads on" factor  $f_j$  with loading  $l_{kj}$ .

This is rather like a multiple regression of  $x_k$  on the factors  $f_1, \dots, f_m$ , viewed as predictor variables.

**Vocabulary**

$\mathbf{f} = [f_1, \dots, f_m]'$  is a vector of common factors  $f_j$ , since potentially all elements  $x_k$  of  $\mathbf{x}$  have them "in common".

You can view the elements  $\varepsilon_k$  of  $\boldsymbol{\varepsilon}$  as either

- non-reproducible "errors" or
- reproducible characteristics, *unique* to the individual and variable, or
- a combination of error and unique characteristic.

**Facts concerning  $\mathbf{f}$  and  $\mathbf{L}$**

Because

$$V[(z_1, z_2, \dots, z_m)'] = \text{diag}[\lambda_1, \lambda_2, \dots, \lambda_m],$$

and  $f_j = z_j / \sqrt{\lambda_j}$ ,

- $V[\mathbf{f}] = \mathbf{I}_m$

That is,  $\sigma_{f_j}^2 = 1, j = 1, \dots, m$ .

Because the  $z_j$ 's are uncorrelated

- $\text{Cov}[\mathbf{f}, \boldsymbol{\varepsilon}] = \mathbf{0}$  ("errors" and factors are uncorrelated)
- $\sum_{1 \leq k \leq p} l_{kj}^2 = \lambda_j \|\mathbf{v}_j\|^2 = \lambda_j$  (L col SS)  
 When  $1 \leq i \neq j \leq m$ ,  
 $\sum_{1 \leq k \leq p} l_{ki} l_{kj} = \sqrt{\{\lambda_i \lambda_j\}} \mathbf{v}_i' \mathbf{v}_j = 0$ , (L col SP)

Thus the columns  $\mathbf{l}_j$  of  $\mathbf{L}$  are *orthogonal* and  $\mathbf{L}'\mathbf{L} = \boldsymbol{\Lambda} = \text{diag}(\lambda_1, \dots, \lambda_m)$

- When the  $z_i$ 's are *correlation* PC's, defined using the eigenvectors of  $\rho$ ,

$$\tilde{x}_k \equiv (x_k - \mu_k) / \sqrt{\sigma_{kk}} = \sum_{1 \leq j \leq m} \ell_{kj} f_j + \epsilon_k$$

and

$$\text{Cov}(\tilde{x}_k, f_j) = \ell_{kj} V[f_j] = \ell_{kj}$$

$\tilde{x}_k$  is the standardized form of  $x_k$ .

Because  $V[f_j] = 1$ , this implies that

$$\ell_{kj} = \text{Corr}[x_k, f_j] \text{ and } |\ell_{kj}| \leq 1.$$

This provides one interpretation for thinking about the loadings.

### The factor analytic model

Factor analysis originated as an intellectual effort to measure or quantify "intelligence".

This is related to an empirical phenomenon:

When a sample of people take  $p$  cognitive ability tests (math skills, reasoning, reading comprehension, etc.), the following usually happens:

- People who have a high score on one test *tend* to have high scores on all, that is, scores on different tests are highly positively correlated.
- Those who score highly are the people who are generally regarded as "smarter" (more intelligent).

#### Caution:

There may be some circularity in the latter statement.

Because we got to the "model"  $\mathbf{x} = \boldsymbol{\mu} + \mathbf{L}\mathbf{f} + \boldsymbol{\epsilon}$  by way of PC's, the "error" is

$$\boldsymbol{\epsilon} = \mathbf{x} - \mathbf{L}\mathbf{f} = \sum_{m+1 \leq j \leq p} z_j \mathbf{v}_j.$$

with variance

$$V[\boldsymbol{\epsilon}] = \sum_{m+1 \leq j \leq p} \lambda_j \mathbf{v}_j \mathbf{v}_j', \quad p \times p$$

which has rank  $p-m < p$  and *cannot* be a diagonal matrix.

That is, the elements of  $\boldsymbol{\epsilon}$  *cannot* be completely uncorrelated. Common factors *cannot* "explain" *all* the correlation.

This is *not* true of factor analysis.

*However*, when  $\sum_{m+1 \leq j \leq p} \lambda_j / \sum_{1 \leq j \leq p} \lambda_j$  is small:

- $V[\boldsymbol{\epsilon}]$  will be small compared to  $\boldsymbol{\Sigma}$
  - Covariances of  $\epsilon_i, i = 1, \dots, p$  will be much smaller than the covariances of  $\mathbf{x}^{(m)}, x_k^{(m)} \equiv \sum_{1 \leq j \leq m} \ell_{kj} f_j, i = 1, \dots, p$
- $\Rightarrow$  Common factors  $f_1, \dots, f_m$  "explain" most of  $\text{cov}[x_j, x_k], j \neq k$ .

It was natural to suppose that the correlation came because the test scores were largely dependent on a *real*, but not directly observed, "lurking" variable - individuals' *intelligence* level.

The supposed intelligence level was named *Intelligence Quotient* or **IQ**.

If  $x_{ij}$  is the score of person  $i$  on test  $j$ , then the implicit model was that

$$x_{ik} = \mu_k + \ell_k \times IQ_i + \epsilon_{ik}, \quad i = 1, \dots, N, \quad k = 1, \dots, p$$

- $IQ_i$  is  $i^{\text{th}}$  person's IQ
- $\mu_k$  is the mean and  $\ell_k$  is a loading on  $IQ_i$  for test score  $x_k$ .  $\ell_k$  determines how much effect intelligence has on  $x_k$ .
- $\epsilon_{ik} = \epsilon_{ik}^{(1)} + \epsilon_{ik}^{(2)}$  is a random quantity reflecting both "measurement error"  $\epsilon_{ik}^{(2)}$  and the *unique* reproducible response  $\epsilon_{ik}^{(1)}$  of person  $i$  to test  $k$ .

Usually no attempt is made to try to separate  $\epsilon_{ik}^{(1)}$  and  $\epsilon_{ik}^{(2)}$ .

So far, this model could be describing PCA with only  $m = 1$  PC identified as IQ is retained.

The difference comes from the supposition that *all* the correlation among the scores can be explained by their dependence on the (unobserved)  $f_1 = IQ$ .

In factor analysis, unlike PCA,  $\epsilon_{i1}, \epsilon_{i2}, \dots, \epsilon_{ip}$  are assumed uncorrelated, that is

$$V[\epsilon_i] = \Psi = \text{diag}[\psi_1, \psi_2, \dots, \psi_p], \psi_k \geq 0.$$

**Remark** following up on signs: You might view correlations of test scores as arising from a *negative* dependence of scores on

$$SQ = \text{"stupidity quotient"} \equiv -IQ$$

with loadings  $\tilde{\lambda}_i = -\lambda_i \leq 0$ .

There is *no way* to distinguish an explanation in terms of intelligence from one in terms of stupidity.

The fact you could use the same model to explain test scores both in terms of "intelligence" and of "stupidity" reflects what some see as an arbitrary quality in factor analysis.

An important part of a factor analysis is often "identifying" factors, that is, *giving them names*. Even this one factor example shows this can be arbitrary. And the choice of names can have a large effect on how people interpret research results.

The mean and SD of IQ are arbitrary and could be given any convenient values such as 100 and 15 or 0 and 1.

This is true in general of factors.

Also, note that aside from scaling, the sign, too, is arbitrary in that

$$\lambda_i \times IQ_j = (-\lambda_i) \times (-IQ_j)$$

If you are looking for a measurement of intelligence which has positive correlation with test scores, you would presume  $\lambda_i \geq 0$ .

The choice of sign is the simplest example of "rotation" of factors and factor loadings.

The **factor analytic model** is designed to "explain" all the correlation among the observable variables  $x_1, \dots, x_p$  by their dependence on  $m < p$  common factors  $f_1, \dots, f_m$ . It has the same form as for PCA:

$$\underset{p \times 1}{\mathbf{x}} = \underset{p \times 1}{\boldsymbol{\mu}} + \underset{p \times m}{\mathbf{L}} \underset{m \times 1}{\mathbf{f}} + \underset{p \times 1}{\boldsymbol{\epsilon}},$$

Where

- $\mathbf{f} = [f_1, f_2, \dots, f_m]'$  is a  $m$  by 1 vector of random *unobservable* common factors with  $E[\mathbf{f}] = \mathbf{0}$
- $\boldsymbol{\epsilon} = [\epsilon_1, \epsilon_2, \dots, \epsilon_p]'$  is an *unobservable* vector of  $p$  **unique factors** with  $E[\boldsymbol{\epsilon}] = \mathbf{0}, V[\epsilon_j] = \psi_j \geq 0$  but with  $\text{corr}(\epsilon_j, \epsilon_k) = 0, j \neq k$

so

$$V[\boldsymbol{\epsilon}] = \Psi = \text{diag}[\psi_1, \dots, \psi_p], \text{rank}(\Psi) = p$$

*Explaining variance is not normally part of factor analysis.*

$L = [l_{kj}]$  is the  $p$  by  $m$  *loading matrix* or matrix of *factor loadings*.

This model differs from the principal component representation in two ways:

- Factors  $f_j$  are *unobservable*, even when you know all parameters ( $\mu$ ,  $L$  and  $\Psi$ ) exactly.

In the PC representation,

$$f_j = z_j / \sqrt{\lambda_j} = \mathbf{v}_j'(\mathbf{x} - \boldsymbol{\mu}) / \sqrt{\lambda_j}.$$

so when you know  $\Sigma$  and  $\boldsymbol{\mu}$ , you can find  $\mathbf{v}_j$  and compute  $z_j$  from  $\mathbf{x}$  and thereby "observe"  $f_j = z_j / \sqrt{\lambda_j}$ .

- The elements  $\epsilon_k$  of  $\boldsymbol{\epsilon}$  are *uncorrelated*. Therefore *all* correlation among the  $x_k$ 's must come from having factors in common.

This *cannot* true in PCA, because

$$V[\boldsymbol{\epsilon}] = \sum_{m+1 \leq j \leq p} \lambda_j \mathbf{v}_j \mathbf{v}_j', \text{ a rank } p-m \text{ matrix.}$$

**Comment:**

Factor analysis is usually described as a dimension *reduction* technique since you "boil down"  $p$  variables to  $m$  common factors.

However, perversely, you can view it as a *dimension augmentation* method, since you still have  $p$  unique factors in addition to the  $m$  common factors. Thus, in a certain sense, you go from  $p$  variables to  $m + p$  factors.

This is at least part of the reason for the non-uniqueness of the factor analytic model.

**Summary of terminology and notation**

The *factor analysis model* with  $m$  factors

$$\mathbf{x} = \boldsymbol{\mu} + \mathbf{L} \mathbf{f} + \boldsymbol{\epsilon}$$

$\begin{matrix} p \times 1 & p \times 1 & p \times m & m \times 1 & p \times 1 \end{matrix}$

$$V[\boldsymbol{\epsilon}] = \boldsymbol{\Psi} = \text{diag}[\psi_1, \psi_2, \dots, \psi_p]$$

- Elements  $f_j$  of  $\mathbf{f}$  are *common factors*.
- Elements  $\epsilon_k$  of  $\boldsymbol{\epsilon}$  are *unique factors* and are uncorrelated with  $f_1, \dots, f_m$ .
- Elements  $l_{kj}$  of  $L$  are *loadings* of variable  $k$  on factor  $j$ .
- The diagonal elements  $\psi_k = V[\epsilon_k]$  of  $\boldsymbol{\Psi}$  are called the *uniquenesses* or *specific variances*.
- $h_k^2 \equiv \sigma_{kk} - \psi_k = V[\sum_{1 \leq j \leq m} l_{kj} f_j] = V[x_k - \mu_k - \epsilon_k]$  are the *communalities*. You can show that  $|\rho_{kj}| \leq (h_k / \sqrt{\sigma_{kk}})(h_j / \sqrt{\sigma_{jj}})$ , so when  $h_k^2$  is small relative to  $\sigma_{kk}$ ,  $x_k$  can't be highly correlated  $x_j$   $j \neq k$ .

In summation notation the factor analytic model is

$$x_k = \mu_k + \sum_{1 \leq j \leq m} l_{kj} f_j + \epsilon_k, \quad k = 1, \dots, p.$$

This has the appearance of  $p$  *multiple regressions* of each element of  $\mathbf{x}$  as a dependent variable on the  $m$  factors playing the role of independent variables.

This is deceptive. The "independent variables" *are not* and *cannot* be directly observed.

- $h_k^2$  is analogous to the *regression SS* in a regression of  $x_k$  on  $f_1, \dots, f_m$ .
- $h_k^2 / \sigma_{kk} \leq 1$  is analogous to multiple  $R^2$ .
- $\psi_k / \sigma_{kk} \leq 1$  is analogous to  $1 - R^2$

The larger  $h_k^2 / \sigma_{kk}$  and the smaller  $\psi_k / \sigma_{kk}$ , the more completely you can explain the behavior of  $x_k$  in terms of the common factors.

### Heywood case

When  $h_k^2 = \sigma_{kk}$ ,  $\psi_k = 0$ . This means that  $x_k$  is completely predictable from the common factors  $f_1, \dots, f_m$ .

This situation is referred to as the Heywood case.

The Heywood case can cause problems for estimation algorithms since it is a situation where a parameter ( $\psi_k$ ) is at the edge of its permissible region ( $\psi_k \geq 0$ ).

One way out is to take  $x_k$  itself as a factor and then analyze partial correlations  $\rho_{j \cdot k}$  assuming  $m - 1$  additional factors.

**Q.** What can you say about the expectation vector  $\boldsymbol{\mu}_f = E[\mathbf{f}]$  and the  $m$  by  $m$  matrix  $\boldsymbol{\Gamma} \equiv V[\mathbf{f}]$ ?

**A.** *Nothing*, except by convention or subject matter theory.

- Without losing any generality, you can assume that  $E[f_j] = 0$  and  $V[f_j] = 1$ . Once you have identified factors, you can rescale and re-center them if you want.
- Often, factors are assumed to be *uncorrelated* so that  $\boldsymbol{\Gamma} = V[\mathbf{f}] = \mathbf{I}_m$ .