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

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.

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

As usual $\mathbf{v}_1, \ldots, \mathbf{v}_p$ are the eigenvectors and $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$ 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 $z_1, \ldots, z_p$ as

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

- $z_1,\ldots,z_p$, $z_j = \mathbf{v}_j'(\mathbf{x}-\boldsymbol{\mu})$ 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

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

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

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

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

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

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

Define the random vector of m "factors":

- $\underset{m\times 1}{\mathbf{f}} = [f_1, \ldots, f_m]'$,

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

  f_j is standardized PC z_j

and the p by m "loading" matrix

- $\underset{p\times m}{\mathbf{L}} = [\ell_{kj}] \equiv [\sqrt{\lambda_1}\mathbf{v}_1, \sqrt{\lambda_2}\mathbf{v}_2, \ldots, \sqrt{\lambda_m}\mathbf{v}_m]$
  $\equiv [\boldsymbol{\ell}_1, \boldsymbol{\ell}_2, \ldots, \boldsymbol{\ell}_m],\ \ell_{kj} = \sqrt{\lambda_j}v_{kj}$

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

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

In terms of matrices this is

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

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

$$x_k = \mu_k + \textstyle\sum_{1\le j\le m} \ell_{kj} f_j + \varepsilon_k.$$

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 $f_1, \ldots, f_m$, viewed as predictor variables.

Vocabulary

$\mathbf{f} = [f_1, \ldots, f_m]'$ 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 ε_k of $\boldsymbol{\varepsilon}$ as either

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

The matrix of *loadings* is

$$\mathbf{L} = \begin{bmatrix} \ell_{11} & \ell_{12} & \dots & \ell_{1m} \\ \cdot & \cdot & \cdots & \cdot \\ \ell_{k1} & \ell_{k2} & \dots & \ell_{km} \\ \cdot & \cdot & \cdots & \cdot \\ \ell_{p1} & \ell_{p2} & \dots & \ell_{pm} \end{bmatrix} \begin{matrix} \text{Loadings for } x_1 \\ \cdots\cdots \\ \text{Loadings for } x_k \\ \cdots\cdots \\ \text{Loadings for } x_p \end{matrix}$$

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

Elements ℓ_{kj} of **L** characterize the dependence of **x** on the common factors.

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

Facts concerning f and 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)
- $\sum_{1\leq k\leq p} \ell_{kj}^2 = \lambda_j \|\mathbf{v}_j\|^2 = \lambda_j$ (**L** col SS)
  When $1 \leq i \neq j \leq m$,
  $\sum_{1\leq k\leq p} \ell_{ki}\ell_{kj} = \sqrt{\{\lambda_i\lambda_j\}}\mathbf{v}_i'\mathbf{v}_j = 0$, (**L** col SP)

Thus the columns $\boldsymbol{\ell}_j$ of **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 $\boldsymbol{\rho}$,

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

  and

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

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

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

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

This provides one interpretation for thinking about the loadings.

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

$$\boldsymbol{\varepsilon} = \mathbf{x} - \mathbf{Lf} = \sum_{m+1\le j\le p} z_j \mathbf{v}_j.$$

with variance

$$V[\boldsymbol{\varepsilon}] = \sum_{m+1\le j\le p} \lambda_j \mathbf{v}_j \mathbf{v}_j', \; p\times p$$

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

That is, the elements of $\boldsymbol{\varepsilon}$ *cannot* be <u>completely</u> uncorrelated. Common factors *cannot* "explain" *all* the correlation.

This is *not* true of factor analysis.

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

- $V[\boldsymbol{\varepsilon}]$ will be small compared to $\boldsymbol{\Sigma}$
- Covariances of ε_i, $i = 1,\ldots,p$ will be much smaller than the covariances of $\mathbf{x}^{(m)}$, $x_k^{(m)} \equiv \sum_{1\le j\le m} \ell_{kj} f_j$, $i = 1,\ldots,p$

$\Rightarrow$ Common factors $f_1, \ldots, f_m$ "explain" <u>most</u> of $\mathrm{cov}[x_j, x_k]$, $j \ne k$.

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.

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

$$x_{ik} = \mu_k + \ell_k \times IQ_i + \varepsilon_{ik},\ i = 1,\ldots,N,\ k = 1,\ldots,p$$

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

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

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, ε_{i1}, ε_{i2}, ..., ε_{ip} are assumed uncorrelated, that is

$$V[\boldsymbol{\varepsilon}_i] = \boldsymbol{\Psi} = \mathrm{diag}[\psi_1,\psi_2,\ldots,\psi_p],\ \psi_k \geq 0.$$

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 <u>sign</u>, too, is arbitrary in that

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

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

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

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

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

with loadings $\tilde{\ell}_i = -\ell_i \leq 0$.

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

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

where

- $\mathbf{f} = [f_1, f_2, \ldots, f_m]'$ is a m by 1 vector of random *unobservable* common factors with $E[\mathbf{f}] = \mathbf{0}$
- $\boldsymbol{\varepsilon} = [\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_p]'$ is an *unobservable* vector of p ***unique factors*** with
  $E[\boldsymbol{\varepsilon}] = \mathbf{0}$, $V[\varepsilon_j] = \psi_j \geq 0$
  but with $\text{corr}(\varepsilon_j, \varepsilon_k) = 0$, $j \neq k$
  so
  $V[\boldsymbol{\varepsilon}] = \boldsymbol{\Psi} = \text{diag}[\psi_1,\ldots,\psi_p]$, $\text{rank}(\boldsymbol{\Psi}) = p$

Explaining variance is not normally part of factor analysis.

$\mathbf{L} = [\ell_{kj}]$ 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 ($\boldsymbol{\mu}$, $\mathbf{L}$ and $\boldsymbol{\Psi}$) 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 $\boldsymbol{\Sigma}$ 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 ε_k of $\boldsymbol{\varepsilon}$ are *uncorrelated*.

  Therefore *all* correlation among the x_k's must come from having factors in common.

  This *cannot* true in PCA, because $V[\boldsymbol{\varepsilon}] = \sum_{m+1\leq j\leq p}\lambda_j\mathbf{v}_j\mathbf{v}_j'$, a rank p-m 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

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

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

- Elements f_j of $\mathbf{f}$ are *common factors*.
- Elements ε_k of $\boldsymbol{\varepsilon}$ are *unique factors*. and are uncorrelated with $f_1,\ldots,f_m$.
- Elements ℓ_{kj} of $\mathbf{L}$ are *loadings* of variable k on factor j.
- The diagonal elements $\psi_k = V[\varepsilon_k]$ of $\boldsymbol{\Psi}$ are called the *uniquenesses* or *specific variances*.
- $h_k^2 \equiv \sigma_{kk} - \psi_k = V[\sum_{1\leq j\leq m} \ell_{kj} f_j] = V[x_k - \mu_k - \varepsilon_k]$ 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 σ_{kk}, x_k can't be highly correlated x_j $j \neq k$.

In summation notation the factor analytic model is

$$x_k = \mu_k + \sum_{1\leq j\leq m} \ell_{kj} f_j + \varepsilon_k,\ k = 1,\ldots,p.$$

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, ..., 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/σ_{kk} and the smaller ψ_k/σ_{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 $f_1, \ldots, f_m$.

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 (ψ_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 $\rho_{j\ell.k}$ 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 $E[f_i] = 0$ 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 $\boldsymbol{\Gamma} = V[\mathbf{f}] = \mathbf{I}_m$.