Statistics 5401
21. Factor Models

Gary W. Oehlert
School of Statistics
313B Ford Hall
612-625-1557
gary@stat.umn.edu

Factor analysis is a model-based technique for "explaining" covariance (or correlation) between the components of a random vector.
The idea is that there are one or more unobserved "factors" which affect the observable factors through a linear regression-like model.
Factor analysis seeks to identify these factors and the relationships between the observed variables and the unobservable factors.
It is generally the case that there are fewer factors than variables. Thus factor models bear a strong resemblence to principal components.

1. Overview of factor analysis
2. Contrast with principal components
3. Details and foibles of factor models
4. Extracting factors
5. Factor rotation
6. Factor scores

Factor analysis from 30,000 feet.
Let x be a p-dimensional random variable with mean μ and variance Σ. We may from time to time assume multivariate normality, but the basic ideas are more general.
Factor analysis is concerned with modeling Σ.
Factor analysis assumes that we can write

$$\Sigma = \mathbf{V} + \Psi$$

where $\mathbf{V}$ has rank $m < p$ and Ψ is diagonal with nonnegative diagonal elements.
Having found $\mathbf{V}$, we express it as

$$\mathbf{V} = \mathbf{L}\mathbf{L}' \quad \text{or} \quad \mathbf{V} = \mathbf{L}\Gamma\mathbf{L}'$$

where $\mathbf{L}$ is $p \times m$ and Γ is $m \times m$. This is where things can get ambiguous.
Why this structure? Suppose that the following holds:

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

where $\mathbf{L}$ is $p \times m$ and f is a random m-vector. If $\mathrm{E}(f) = 0$, $\mathrm{Cov}(f) = \mathbf{I}$, $\mathrm{E}(\epsilon) = 0$ and $\mathrm{Cov}(\epsilon) = \Psi$, then

$$\mathrm{Cov}(x) = \mathbf{L}\mathbf{L}' + \Psi$$

$$\mathrm{Cov}(x_i, f_j) = \mathbf{L}_{ij}$$

This is the *Orthogonal Factor Model*.

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

In this setup, f_j is the jth *common factor*, $\mathbf{L}_{ij}$ is the *loading* of the ith variable on the jth factor, and ϵ_i is the ith *specific factor*.
x is observable, but the latent variables f and ϵ are unobservable.
The idea is that we represent the p variables in x as combinations of m variables from f, plus some hopefully small noise in ϵ.
Factor models are sometimes described as doing dimension reduction, because we only use m factors.
On the other hand, we started with p observable variables, and now we $m+p$ unobservable variables (the common and specific factors); is this progress?
Where does this come from? Factor analysis arose originally in attempts to measure intelligence.
Suppose that we have several test responses: verbal skill, musical talent, math skill, analytical competence, and so on. We might believe that there is some underlying intelligence, such that a person who is more intelligent will score higher on the tests than a person who is less intelligent.
Thus test scores will be correlated, with the correlation due to the common intelligence.
Here is one way to model this. Let x be the test responses.

$$x_i = \mu_i + \ell_i(I) + \epsilon_i$$

Thus the ith test score is a linear function of the underlying intelligence I, plus some error. Each test score can have a different mean and slope coefficient, and the ϵ_is cover measurement error and unique individual responses.
Here we have a single factor (intelligence I), with loadings ℓ_i.
Note that there is more than one way to skin a cat:

$$x_i = \mu_i + -\ell_i(S) + \epsilon_i$$

where $S = -I$ is a stupidity factor is just as valid as the expression using intelligence.
We can also rescale I or S to have whatever mean and variance we want; μ_i and ℓ_i can adjust.
Centers, scale, and sign are all arbitrary in this one-factor model; life is even more interesting in multifactor models.
Go back to

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

with Var$(f) = \mathbf{I}$. Let $\mathbf{H}$ be any $m \times m$ orthogonal matrix. Then

$$\begin{aligned} x &= \mu + \mathbf{L}f + \epsilon \\ &= \mu + \mathbf{L}\mathbf{H}\mathbf{H}'f + \epsilon \\ &= \mu + \mathbf{L}^\star f^\star + \epsilon \end{aligned}$$

where $\mathbf{L}^\star = \mathbf{L}\mathbf{H}$ and $f^\star = \mathbf{H}'f$, with Var$(f^\star) = \mathbf{I}$.
Factors and loadings are most definitely not unique.
Is anything well defined? Yes!
x has mean μ and variance Σ with

$$\Sigma = \mathbf{V} + \Psi$$

with rank$(\mathbf{V}) = m$ and Ψ nonnegative and diagonal.
μ, $\mathbf{V}$, and Ψ are all well defined. It's the representation of $\mathbf{V}$ via factors that starts making things ambiguous.
Of course, it is the representation via factors that is generally of explanatory interest, so we can't ignore it.
Factor analysis thus has two steps:

Factor *extraction*, where we estimate $\mathbf{V}$ and Ψ, and possibly a preliminary $\mathbf{L}$.
Factor *rotation*, where we choose a $\mathbf{L}$ with desirable properties. Statistically, all feasible $\mathbf{L}$s are equivalent, so this choice requires additional subject matter (nonstatistical) input and information.
Another aspect of factor analysis is called *confirmatory* factor analysis. In this exercise, we hypothesize that $\mathbf{L}$ takes a special form. For example,

$$\mathbf{L} = \begin{bmatrix} \star & \star & 0 \\ \star & 0 & 0 \\ 0 & 0 & \star \\ 0 & 0 & \star \end{bmatrix}$$

where $\star$ means any nonzero value. We then do a goodness of fit test for the data comparing this form to an arbitrary $\mathbf{L}$ as the alternative.
We have noted that rescaling a random vector can give radically different principal components than the original vector. Furthermore, there is no obvious relationship between the scaled and unscaled components.
For factor analysis, there is a simple relationship.
Let Δ be diagonal with elements $1/\sqrt{\Sigma_{ii}}$.
Then $z = \Delta(x - \mu)$ is the standardized vector.

$$\begin{aligned} Cov(z) &= \Delta\Sigma\Delta' \\ &= \Delta(\mathbf{V} + \Psi)\Delta' \\ &= \Delta(\mathbf{L}\mathbf{L}' + \Psi)\Delta' \\ &= \Delta\mathbf{L}\mathbf{L}'\Delta' + \Delta\Psi\Delta' \\ &= \mathbf{L}^\star\mathbf{L}^{\star\prime} + \Psi^\star \end{aligned}$$

where $\mathbf{L}^\star = \Delta\mathbf{L}$ and $\psi_i^\star = \psi_i/\Sigma_{ii}$.
Thus there is an easy translation back and forth between factor models on scaled and unscaled data.
A little more terminology and notation.

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

and

$$\Sigma = \mathbf{L}\mathbf{L}' + \Psi$$

Thus

$$\begin{array}{rcccc} \Sigma_{ii} & = & \mathbf{L}_{i1}^2 + \mathbf{L}_{i2}^2 + \ldots + \mathbf{L}_{im}^2 & + & \psi_i \\ & = & h_i^2 & + & \psi_i \\ \text{Var}(x_i) & = & \text{communality} & + & \text{uniqueness} \end{array}$$

The uniqueness is also called the specific variance.
Relation with principal components.
Let's now compare and contrast principal components with factor models.
As before, x has mean μ and variance Σ.
The eigenvalues of Σ are $\eta_1 \geq \eta_2 \geq \ldots \geq \eta_p \geq 0$, with corresponding eigenvectors $v_1, v_2, \ldots, v_p$.
With this, we have

$$w = \mathbf{V}'(x - \mu)$$

and

$$x = \mu + \mathbf{V}w$$

where w is x expressed in principal components.
Note that Var$(w_j) = \eta_j$.
Now rescale. Let $\check{\mathbf{U}}_j = \check{\mathbf{V}}_j \eta_j$, and let $z_j = w_j / \eta_j$. Then

$$x = \mu + \mathbf{U}z$$

with E$(z) = 0$ and Var$(z) = \mathbf{I}$.
This expression of principal components is an orthogonal factor model with p components and all uniquenesses equal to 0.
However, what happens if we use fewer than p components?

$$\begin{aligned} x &= \mu + \sum_{j=1}^{m} \check{\mathbf{U}}_j z_j + \sum_{j=m+1}^{p} \check{\mathbf{U}}_j z_j \\ &= \mu + \mathbf{L}f + \omega \end{aligned}$$

The first part looks like a factor model, but what about ω?

$$Var(\omega) = \sum_{j=m+1}^{p} \check{\mathbf{U}}_j \check{\mathbf{U}}_j'$$

This will only be diagonal (like Ψ) if the rows of the last $p - m$ columns of $\mathbf{U}$ are orthogonal, and there is no particular reason that they should be.
In the factor model, the $\mathbf{L}f$ part explains the covariance in the x exactly, but some addtional variance may be needed along the diagonal to match Σ exactly.
In the principal components model, the $\mathbf{L}f$ part explains the covarariance in x approximately, and the additional bits ϵ you need to add on will usually be correlated.
On the other hand, the principal component are decomposition is uniquely defined.
Principal components shows us that any Σ can be decomposed in a factor model with $m = p$ factors.
However, not all variances Σ follow a factor model with $m < p$ factors.
For example, try to fit

$$\Sigma = \begin{bmatrix} 1 & .9 & .7 \\ .9 & 1 & .4 \\ .7 & .4 & 1 \end{bmatrix}$$

using $m = 1$.

$$\mathbf{L} = \begin{bmatrix} \ell_1 \\ \ell_2 \\ \ell_3 \end{bmatrix}$$

$$\mathbf{L}\mathbf{L}' = \begin{bmatrix} \ell_1^2 & \ell_1\ell_2 & \ell_1\ell_3 \\ \ell_1\ell_2 & \ell_2^2 & \ell_2\ell_3 \\ \ell_1\ell_3 & \ell_2\ell_3 & \ell_3^2 \end{bmatrix}$$

$$\Psi = \begin{bmatrix} \psi_1 & 0 & 0 \\ 0 & \psi_2 & 0 \\ 0 & 0 & \psi_3 \end{bmatrix}$$

$$\begin{aligned} 1 &= \ell_1^2 + \psi_1 \\ 1 &= \ell_2^2 + \psi_2 \\ 1 &= \ell_3^2 + \psi_3 \\ .9 &= \ell_1 \ell_2 \\ .7 &= \ell_1 \ell_3 \\ .4 &= \ell_2 \ell_3 \end{aligned}$$

$$\frac{.4}{.7} = \frac{\ell_2 \ell_3}{\ell_1 \ell_3} = \frac{\ell_2}{\ell_1}$$

so $\ell_2 = 4\ell_1/7$.

$$.9 = \ell_1 \ell_2 = \frac{4\ell_1^2}{7}$$

or $\ell_1^2 = .9 \times 7/4 = 1.575$.

$$1 = \ell_1^2 + \psi_1 = 1.575 + \psi_1$$

so $\psi_1 = -.575$, but ψ_1 is a variance. Oops.