Factor Analysis

- Principal Components Analysis, e.g. of stock price movements, sometimes suggests that several variables may be responding to a small number of underlying forces.

- In the factor model, we assume that such latent variables, or factors, exist.
The Orthogonal Factor Model equation:

\[ X_1 - \mu_1 = l_{1,1}F_1 + l_{1,2}F_2 + \cdots + l_{1,m}F_m + \epsilon_1, \]
\[ X_2 - \mu_2 = l_{2,1}F_1 + l_{2,2}F_2 + \cdots + l_{2,m}F_m + \epsilon_2, \]
\[ \vdots \]
\[ X_p - \mu_p = l_{p,1}F_1 + l_{p,2}F_2 + \cdots + l_{p,m}F_m + \epsilon_p, \]

where:

- \( F_1, F_2, \ldots, F_m \) are the common factors (latent variables);
- \( l_{i,j} \) is the loading of variable \( i, X_i \), on factor \( j, F_j \);
- \( \epsilon_i \) is a specific factor, affecting only \( X_i \).
• In matrix form:

\[ \mathbf{X} - \mu = L \mathbf{F} + \epsilon. \]

• To make this identifiable, we further assume, with no loss of generality:

\[
\begin{align*}
\mathbb{E}(\mathbf{F}) &= \mathbf{0}_{m \times 1} \\
\text{Cov}(\mathbf{F}) &= \mathbf{I}_{m \times m} \\
\mathbb{E}(\epsilon) &= \mathbf{0}_{p \times 1} \\
\text{Cov}(\epsilon, \mathbf{F}) &= \mathbf{0}_{p \times m}
\end{align*}
\]
• and with serious loss of generality:

\[ \text{Cov}(\epsilon) = \Psi = \text{diag}(\psi_1, \psi_2, \ldots, \psi_p). \]

• In terms of the observable variables \(X\), these assumptions mean that

\[ \mathbb{E}(X) = \mu, \]
\[ \text{Cov}(X) = \Sigma = L L' + \Psi. \]

Usually \(X\) is standardized, so \(\Sigma = R\).

• The observable \(X\) and the unobservable \(F\) are related by

\[ \text{Cov}(X, F) = L. \]
• Some terminology: the \((i,i)\) entry of the matrix equation \(\Sigma = LL' + \Psi\) is

\[
\sigma_{i,i} = l_{i,1}^2 + l_{i,2}^2 + \cdots + l_{i,m}^2 + \psi_i
\]

\(\text{Var}(X_i)\) Communality Specific variance

or

\[
\sigma_{i,i} = h_i^2 + \psi_i
\]

where

\[
h_i^2 = l_{i,1}^2 + l_{i,2}^2 + \cdots + l_{i,m}^2
\]

is the \(i^{th}\) communality.

• Note that if \(T\) is \((m \times m)\) orthogonal, then \((LT)(LT)' = LL'\), so loadings \(LT\) generate the same \(\Sigma\) as \(L\): loadings are not unique.
Existence of Factor Representation

• For any $p$, every $(p \times p)$ $\Sigma$ can be factorized as

\[ \Sigma = LL' \]

for $(p \times p)$ $L$, which is a factor representation with $m = p$ and $\Psi = 0$; however, $m = p$ is not much use—we usually want $m \ll p$.

• For $p = 3$, every $(3 \times 3)$ $\Sigma$ can be represented as

\[ \Sigma = LL' + \Psi \]

for $(3 \times 1)$ $L$, which is a factor representation with $m = 1$, but $\Psi$ may have negative elements.
• In general, we can only approximate $\Sigma$ by $LL' + \Psi$.

• Principal components method: the spectral decomposition of $\Sigma$ is

$$\Sigma = EE' = \left(E \Lambda^{1/2}\right) \left(E \Lambda^{1/2}\right)' = LL'$$

with $m = p$.

• If $\lambda_1 + \lambda_2 + \cdots + \lambda_m \gg \lambda_{m+1} + \cdots + \lambda_p$, and $L^{(m)}$ is the first $m$ columns of $L$, then

$$\Sigma \approx L^{(m)}L^{(m)'}$$

gives such an approximation with $\Psi = 0$. 
• The remainder term $\Sigma - L^{(m)}L^{(m)\prime}$ is non-negative definite, so its diagonal entries are non-negative $\Rightarrow$ we can get a closer approximation as

$$\Sigma \approx L^{(m)}L^{(m)\prime} + \Psi^{(m)},$$

where $\Psi^{(m)} = \text{diag} \left( \Sigma - L^{(m)}L^{(m)\prime} \right)$.

• SAS proc factor program and **output**:

```sas
proc factor data = all method = prin;
  var cvx -- xom;
  title 'Method = Principal Components';
proc factor data = all method = prin nfact = 2 plot;
  var cvx -- xom;
  title 'Method = Principal Components, 2 factors';
```