Factor Analysis comes in two forms: Exploratory (EFA) and Confirmatory (CFA)

Factor Analysis comes in two forms: Exploratory (EFA) and Confirmatory (CFA)

- EFA: see what factors might be in the data

Factor Analysis comes in two forms: Exploratory (EFA) and Confirmatory (CFA)

- EFA: see what factors might be in the data
- CFA: write down a model and use the data to test it

Factor Analysis comes in two forms: Exploratory (EFA) and Confirmatory (CFA)

- EFA: see what factors might be in the data

- CFA: write down a model and use the data to test it

In structural econometrics, we pretty much only do CFA

FA is used extensively in psychometrics

FA is used extensively in psychometrics

It is a natural tool for analyzing cognitive or behavioral tests

FA is used extensively in psychometrics

It is a natural tool for analyzing cognitive or behavioral tests

- Each test measures some set of skills, but does so noisily

FA is used extensively in psychometrics

It is a natural tool for analyzing cognitive or behavioral tests

- Each test measures some set of skills, but does so noisily
- Tests tend to measure the same set of skills, so they are correlated

Suppose our J columns of M correspond to measurements (e.g. test scores)

Suppose our J columns of M correspond to measurements (e.g. test scores)

FA tries to find some underlying unobservables that commonly affect M

Suppose our J columns of M correspond to measurements (e.g. test scores)

FA tries to find some underlying unobservables that commonly affect M

We assume that we cannot observe $\boldsymbol{\theta}$

Suppose our J columns of M correspond to measurements (e.g. test scores)

FA tries to find some underlying unobservables that commonly affect M

We assume that we cannot observe $\boldsymbol{\theta}$

If we assume that M is standardized (mean-zero, unit-variance), then

$$M = \underbrace{\boldsymbol{\theta}_k \Lambda_k + \varepsilon}_{\boldsymbol{u}}$$

Suppose our J columns of M correspond to measurements (e.g. test scores)

FA tries to find some underlying unobservables that commonly affect M

We assume that we cannot observe $\boldsymbol{\theta}$

If we assume that M is standardized (mean-zero, unit-variance), then

$$M = \underbrace{\boldsymbol{\theta}_k \Lambda_k + \boldsymbol{\varepsilon}}_{\boldsymbol{u}}$$

$\boldsymbol{u}$ is a composite error term (since both $\boldsymbol{\theta}$ and ε are unobservable)

Suppose our J columns of M correspond to measurements (e.g. test scores)

FA tries to find some underlying unobservables that commonly affect M

We assume that we cannot observe $\boldsymbol{\theta}$

If we assume that M is standardized (mean-zero, unit-variance), then

$$M = \underbrace{\boldsymbol{\theta}_k \Lambda_k + \boldsymbol{\varepsilon}}_{\boldsymbol{u}}$$

$\boldsymbol{u}$ is a composite error term (since both $\boldsymbol{\theta}$ and ε are unobservable)

In FA, we call $\boldsymbol{\theta}$ factors, and we call Λ factor loadings and $\boldsymbol{\varepsilon}$ uniquenesses

Clearly, PCA and FA are related, but there are important differences

Clearly, PCA and FA are related, but there are important differences

- When we drop components, both have error: $M = \boldsymbol{\theta}_k \Lambda_k + \boldsymbol{\varepsilon}$

Clearly, PCA and FA are related, but there are important differences

- When we drop components, both have error: $M = \boldsymbol{\theta}_k \Lambda_k + \boldsymbol{\varepsilon}$

- But the $\boldsymbol{\theta}_k$, Λ_k, and $\boldsymbol{\varepsilon}$ are all different between PCA and FA

Clearly, PCA and FA are related, but there are important differences

- When we drop components, both have error: $M = \boldsymbol{\theta}_k \Lambda_k + \boldsymbol{\varepsilon}$
- But the $\boldsymbol{\theta}_k$, Λ_k, and ε are all different between PCA and FA
- PCA error: approximation from dimensionality reduction ($\rightarrow 0$ as $K \rightarrow J$)

Clearly, PCA and FA are related, but there are important differences

- When we drop components, both have error: $M = \boldsymbol{\theta}_k \Lambda_k + \boldsymbol{\varepsilon}$
- But the $\boldsymbol{\theta}_k$, Λ_k, and ε are all different between PCA and FA
- PCA error: approximation from dimensionality reduction ($\rightarrow 0$ as $K \rightarrow J$)
- FA error: measurement error (persists even with all factors retained)

Clearly, PCA and FA are related, but there are important differences

- When we drop components, both have error: $M = \boldsymbol{\theta}_k \Lambda_k + \boldsymbol{\varepsilon}$
- But the $\boldsymbol{\theta}_k$, Λ_k, and $\boldsymbol{\varepsilon}$ are all different between PCA and FA
- PCA error: approximation from dimensionality reduction ($\to 0$ as $K \to J$)
- FA error: measurement error (persists even with all factors retained)

FA separates common variance from unique variance; PCA explains total variance

Clearly, PCA and FA are related, but there are important differences

- When we drop components, both have error: $M = \boldsymbol{\theta}_k \Lambda_k + \varepsilon$
- But the $\boldsymbol{\theta}_k$, Λ_k, and ε are all different between PCA and FA
- PCA error: approximation from dimensionality reduction ($\rightarrow 0$ as $K \rightarrow J$)
- FA error: measurement error (persists even with all factors retained)

FA separates common variance from unique variance; PCA explains total variance

For many more excellent details, see Shalizi (2019)

In econometrics, we often want to allow covariates X to affect our measurements

In econometrics, we often want to allow covariates X to affect our measurements

We can extend FA to incorporate observable characteristics:

$$M = \boxed{X\beta} + \underbrace{\boldsymbol{\theta}_k \Lambda_k + \varepsilon}_{\boldsymbol{u}}$$

In econometrics, we often want to allow covariates X to affect our measurements

We can extend FA to incorporate observable characteristics:

$$\begin{aligned} M &= X\boldsymbol{\beta} + \underbrace{\boldsymbol{\theta}_k \Lambda_k + \varepsilon}_{\boldsymbol{u}} \\ &= \underbrace{X}_{N \times L} \times \underbrace{\boldsymbol{\beta}}_{L \times J} + \underbrace{\boldsymbol{\theta}_k}_{N \times K} \times \underbrace{\Lambda_k}_{K \times J} + \underbrace{\varepsilon}_{N \times J} \end{aligned}$$

In econometrics, we often want to allow covariates X to affect our measurements

We can extend FA to incorporate observable characteristics:

$$
\begin{aligned}
M &= X\boldsymbol{\beta} + \underbrace{\boldsymbol{\theta}_k \Lambda_k + \boldsymbol{\varepsilon}}_{\boldsymbol{u}} \\
&= \underbrace{X}_{N\times L} \times \underbrace{\boldsymbol{\beta}}_{L\times J} + \underbrace{\boldsymbol{\theta}_k}_{N\times K} \times \underbrace{\Lambda_k}_{K\times J} + \underbrace{\boldsymbol{\varepsilon}}_{N\times J}
\end{aligned}
$$

- The factors $\boldsymbol{\theta}_k$ now capture common variation *after* conditioning on X

In econometrics, we often want to allow covariates X to affect our measurements

We can extend FA to incorporate observable characteristics:

$$
\begin{aligned}
M &= X\boldsymbol{\beta} + \underbrace{\boldsymbol{\theta}_k \Lambda_k + \boldsymbol{\varepsilon}}_{\boldsymbol{u}} \\
&= \underbrace{X}_{N\times L} \times \underbrace{\boldsymbol{\beta}}_{L\times J} + \underbrace{\boldsymbol{\theta}_k}_{N\times K} \times \underbrace{\Lambda_k}_{K\times J} + \underbrace{\boldsymbol{\varepsilon}}_{N\times J}
\end{aligned}
$$

- The factors $\boldsymbol{\theta}_k$ now capture common variation *after* conditioning on X
- Key challenge: We need additional assumptions for identification