Ve406 Lecture 21

Jing Liu

UM-SJTU Joint Institute

July 30, 2018

- Principal component analysis is about replacing a centred and scaled $\mathbf{X}$ by

$$\mathbf{Z}_{n\times\ell} = \mathbf{X}_{n\times k}\mathbf{W}_{k\times\ell} \qquad \text{where } \ell = 1,\ldots,k$$

  and $\mathbf{W}$ has orthonormal eigenvectors of $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ as its columns.

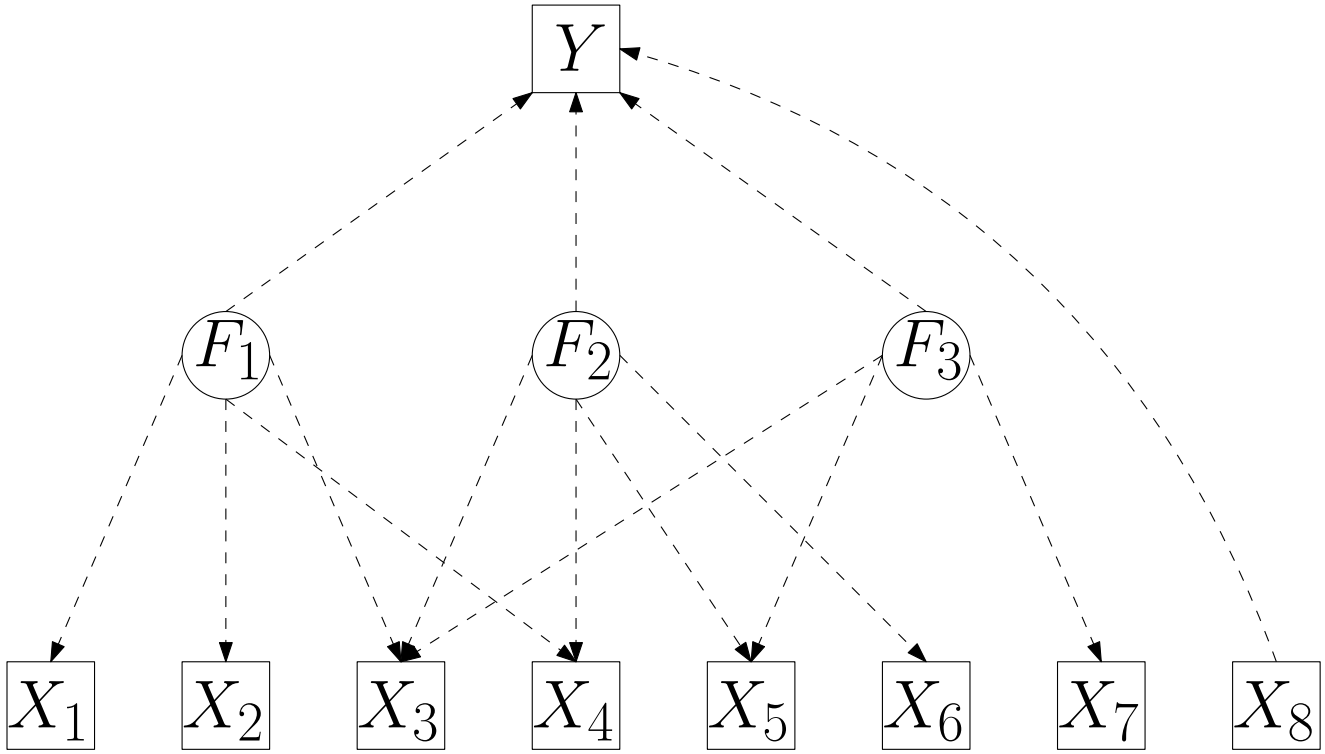
- In reverse, the observed data can be completely reconstructed if $\ell = k$

$$\mathbf{X}_{n\times k} = \mathbf{Z}_{n\times k}\mathbf{W}^{\mathrm{T}}_{k\times k}$$

- If only the first few eigenvectors are used, that is $\ell < k$, then

$$\mathbf{X}_{n\times k} \approx \mathbf{X}^{*}_{n\times k} = \mathbf{Z}_{n\times\ell}\mathbf{W}^{\mathrm{T}}_{k\times\ell}$$

- Note PCA involves absolutely no probabilistic assumptions at all, thus it says nothing about the stochastic processes that generated the data.
- If we assume there is some hidden structure from which the observed $\mathbf{X}$ are generated, then other dimension reduction methods might be reasonable.



- A factor model is used to model such hidden structures

$$\mathbf{X}_{n\times k} = \mathbf{F}_{n\times \ell}\mathbf{W}^{\mathrm{T}}_{k\times \ell} + \boldsymbol{\epsilon}_{n\times k} \qquad \text{where } \boldsymbol{\epsilon} \text{ is an error term.}$$

- The observed variables in $\mathbf{X}$ are known as manifest variables and the hidden variables, now denoted collectively as $\mathbf{F}$ are known as latent factors.

- The idea of factor analysis is the variability in a large number of observed variables could be due to a few unobserved factors and unobserved errors.

$$\mathbf{X}_{n\times k} = \mathbf{F}_{n\times \ell}\mathbf{W}^{\mathrm{T}}_{k\times \ell} + \boldsymbol{\epsilon}_{n\times k} \qquad \text{where } \boldsymbol{\epsilon} \text{ is an error term.}$$

- By first identify $\mathbf{W}$ then $\mathbf{F}$, dimension reduction is achieved since $\ell < k$.
- The following specifications are used for a typical factor model:

1. The manifest variables have zero mean and unit variance.
2. The latent factors have zero mean and unit variance.
3. The error terms have zero mean.
4. The latent factors are uncorrelated

   across observations and with other latent factors.
5. The error terms are uncorrelated

   across observations, across manifest variables and across latent factors.

- Under the assumptions 4-5 of the factor model, the manifest variables are correlated with each other solely due to correlation with the latent factors.

$$\operatorname{Cov}\left[X_{i1}, X_{i2}\right] = \sum_{p=1}^{\ell} w_{1p} w_{2p}$$

where w_{1p} and w_{2p} are elements in the first and the second row of $\mathbf{W}$, since weights now are the covariance/correlation of manifest and factor variables

$$\operatorname{Cov}\left[X_{ij}, F_{iq}\right] = w_{jq}$$

- This leads us to consider the covariance matrix $\boldsymbol{\Sigma}$ in order to compute $\mathbf{W}$,

$$\boldsymbol{\Sigma} = \mathbb{E}\left[\mathbf{C}\right] = \frac{1}{n-1}\mathbb{E}\left[\mathbf{X}^{\mathrm{T}}\mathbf{X}\right] = \boldsymbol{\psi} + \mathbf{W}\mathbf{W}^{\mathrm{T}}$$

where $\mathbf{C}$ is the sample covariance matrix and $\boldsymbol{\psi}$ is a diagonal matrix.

- A natural step next is to use the sample covariance matrix,

$$\mathbf{C} = \boldsymbol{\psi} + \mathbf{W}\mathbf{W}^{\mathrm{T}}$$

and try to solve the system of equations obtained from the above equality.

- For example, consider the case $k = 3$ and $\ell = 1$, which corresponds

$$\begin{bmatrix} 1 & c_{12} & c_{13} \\ c_{21} & 1 & c_{23} \\ c_{31} & c_{32} & 1 \end{bmatrix} = \begin{bmatrix} \psi_1 & 0 & 0 \\ 0 & \psi_2 & 0 \\ 0 & 0 & \psi_3 \end{bmatrix} + \begin{bmatrix} w_{11} \\ w_{21} \\ w_{31} \end{bmatrix} \begin{bmatrix} w_{11} & w_{21} & w_{31} \end{bmatrix}$$

which from the weights/covariances/correlations w_{ji} are given by

$$w_{11}^2 = \frac{c_{12}c_{13}}{c_{23}}; \qquad w_{21}^2 = \frac{c_{12}c_{23}}{c_{13}}; \qquad w_{31}^2 = \frac{c_{13}c_{23}}{c_{12}}$$

and

$$\psi_1 = 1 - w_{11}^2; \qquad \psi_2 = 1 - w_{21}^2; \qquad \psi_3 = 1 - w_{31}^2$$

- Once $\mathbf{W}_{3\times 1}$ is available, we can estimate $\mathbf{F}_{n\times 1}$ using "regression"

$$\mathbf{X}_{n\times 3}\mathbf{W}_{3\times 1} = \hat{\mathbf{F}}_{n\times 1} + \boldsymbol{\epsilon}_{n\times 3}\mathbf{W}_{3\times 1}$$

it can be shown the optimal linear solution which minimises

$$\left\|\boldsymbol{\epsilon}_{n\times 3}\mathbf{W}_{3\times 1}\right\|^2$$

is given by

$$\hat{\mathbf{F}}_{n\times 1} = \mathbf{X}_{n\times 3}\mathbf{W}_{3\times 1}$$

- However, in general it is not that simple, notice for an orthogonal matrix $\mathbf{Q}$,

$$\begin{aligned}\mathbf{X}_{n\times k} &= \mathbf{F}_{n\times \ell}\mathbf{W}^{\mathrm{T}}_{k\times \ell} + \boldsymbol{\epsilon}_{n\times k}\\ \mathbf{X}_{n\times k} &= \left(\mathbf{F}_{n\times \ell}\mathbf{Q}_{\ell\times \ell}\right)\left(\mathbf{W}_{k\times \ell}\mathbf{Q}_{\ell\times \ell}\right)^{\mathrm{T}} + \boldsymbol{\epsilon}_{n\times k}\end{aligned}$$

which means neither $\mathbf{W}$ nor $\mathbf{F}$ is unique when we have more than one factor.

- Having multiple solutions is numerically problematic, but it can be avoided by looking at it slightly differently and employ a numerical stable method

$$\mathbf{C} - \boldsymbol{\psi} = \mathbf{W}\mathbf{W}^{\mathrm{T}}$$

which naturally adds constraints to avoid multiple solutions from rotation.

- For a given ψ, the left-hand side is surely orthogonally diagonalizable

$$\mathbf{C} - \boldsymbol{\psi} = \mathbf{U}\mathbf{D}\mathbf{U}^{\mathrm{T}}$$

since it is real symmetric, and we can reduce to ℓ factors by finding the first ℓ largest eigenvalues and the corresponding eigenvectors to approximate

$$\mathbf{C} - \boldsymbol{\psi} \approx \tilde{\mathbf{A}}_{k\times k} = \mathbf{U}_{k\times \ell}\mathbf{D}_{\ell\times\ell}\mathbf{U}^{\mathrm{T}}_{k\times\ell}$$

from which we obtain

$$\mathbf{W}_{k\times\ell} = \mathbf{U}_{k\times\ell}\mathbf{D}^{1/2}_{\ell\times\ell}$$

- Once we have $\mathbf{W}_{k\times\ell}$, by making a more assumption on data

$$\mathbf{x}_i \sim \text{Normal}\left(\mathbf{0}, \boldsymbol{\psi} + \mathbf{W}\mathbf{W}^{\text{T}}\right)$$

we can update $\boldsymbol{\psi}$ according maximum likelihood estimation, then we iterate by computing $\mathbf{W}$ again and then $\boldsymbol{\psi}$ again and again until it converges.

- To illustrate factor analysis, consider the following dataset

```
> ct.df = read.table("~/Desktop/cities.csv",
+                    sep = ",", header = TRUE)
>
> dim(ct.df)
```

```
[1] 21 12
```

```
> names(ct.df)
```

```
 [1] "City"      "Area"      "Pop.1980"  "Pop.1990"
 [5] "Pop.2000"  "Growth"    "Food"      "PersRoom"
 [9] "Water"     "Elec"      "Phones"    "Vehicles"
```

```
> str(ct.df)
```

```
'data.frame':   21 obs. of  12 variables:
 $ City     : Factor w/ 21 levels "Beijing","Bombay",..: 3 7 16 17 1 19 20 4 2 5 ...
 $ Area     : num  700 100 650 800 1680 ...
 $ Pop.1980: int  9918 3290 8789 12101 9029 11739 7268 6852 8067 9030 ...
 $ Pop.1990: int  11448 6578 10948 18119 10867 13447 9249 8633 12223 10741 ...
 $ Pop.2000: int  12822 11511 12162 22552 14366 17407 12508 10761 18142 12675 ...
 $ Growth   : num  1.4 7.2 2.2 4.1 1.9 ...
 $ Food     : int  40 63 26 50 52 55 52 47 57 60 ...
 $ PersRoom: num  1.3 2.4 0.8 0.8 1.2 ...
 $ Water    : int  80 60 86 100 88 95 80 91 92 51 ...
 $ Elec     : int  91 85 98 100 90 95 84 98 78 63 ...
 $ Phones   : int  14 2 8 16 2 4 4 4 5 2 ...
 $ Vehicles: int  1000 600 4000 4000 308 148 200 939 588 500 ...
```

```
> # remove city names
> X = data.matrix(ct.df[, 2:12])

> # use city names as row names
> rownames(X) = ct.df[, 1]

> ct.FA = factanal(X, factors = 2, rotation="none",
+                  scores="regression")
```

```
> ct.FA
```

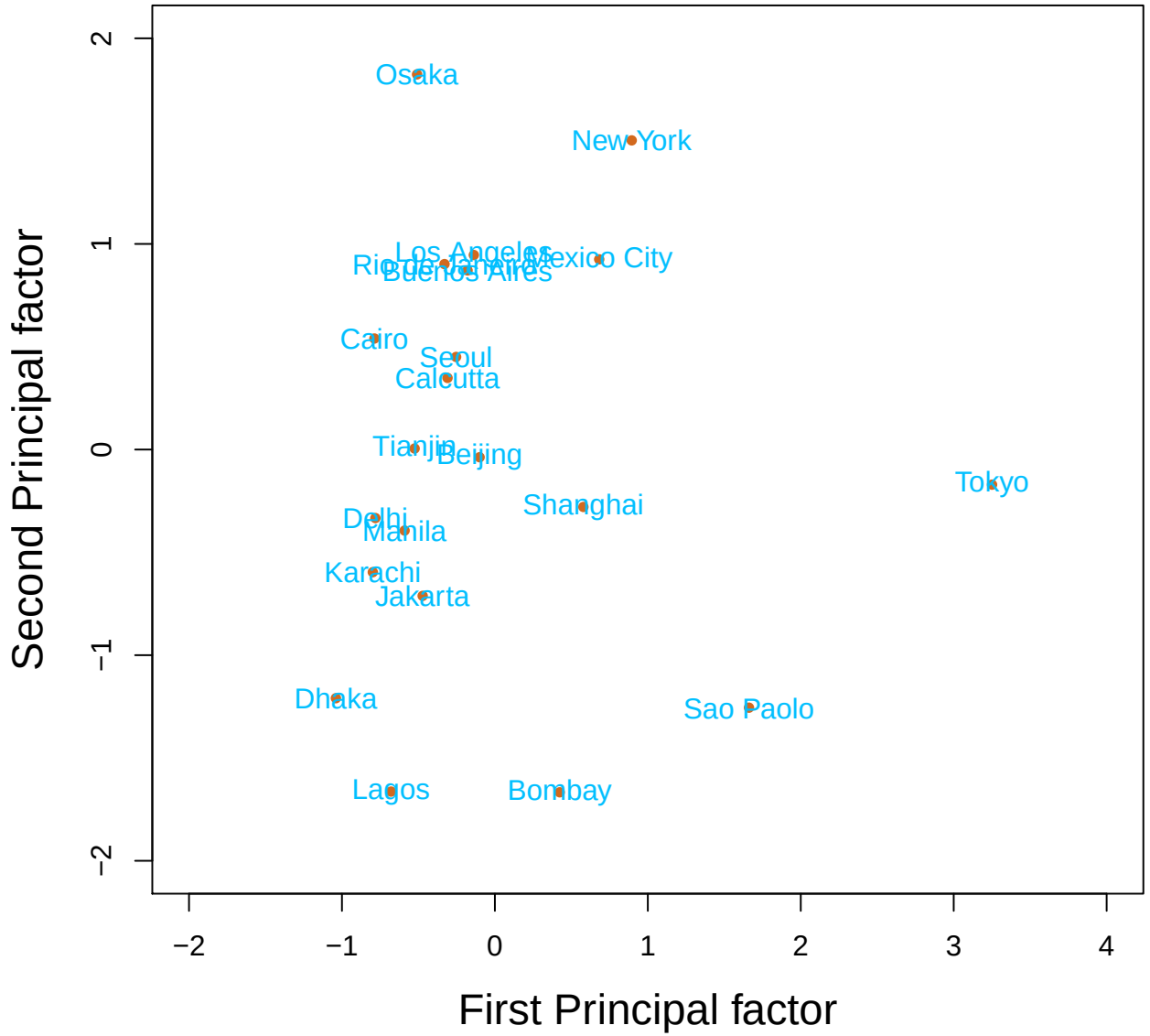
```
Call:
factanal(x = X, factors = 2, scores = "regression", rotation = "none")

Uniquenesses:
    Area Pop.1980 Pop.1990 Pop.2000   Growth     Food PersRoom    Water
   0.933    0.022    0.005    0.005    0.179    0.437    0.387    0.541
    Elec   Phones Vehicles
   0.605    0.426    0.706

Loadings:
          Factor1 Factor2
Area       0.129   0.225
Pop.1980   0.914   0.377
Pop.1990   0.988   0.136
Pop.2000   0.967  -0.243
Growth    -0.386  -0.820
Food      -0.316  -0.681
PersRoom  -0.368  -0.691
Water      0.558   0.384
Elec       0.366   0.510
Phones     0.518   0.553
Vehicles   0.356   0.409
```

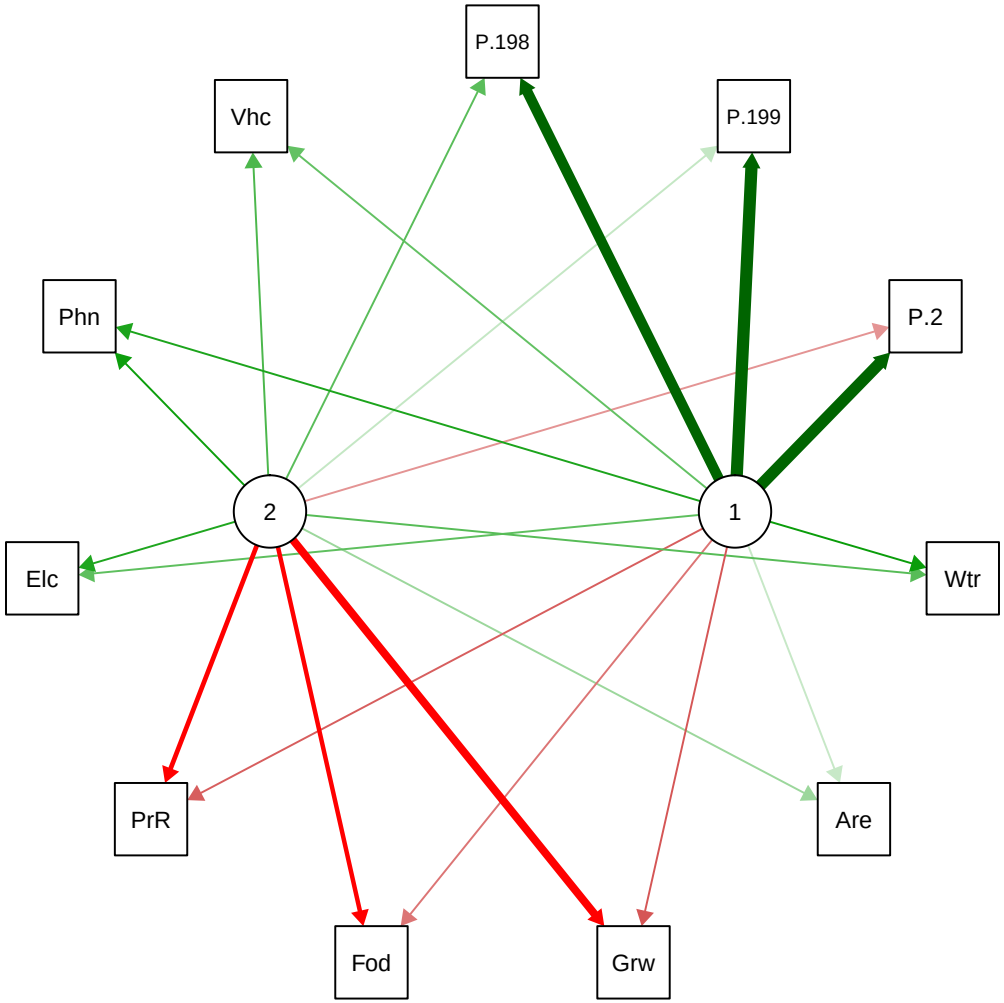
- The output `Uniquenesses` tells us what fraction of the variance in each observable comes from its own error/noise, that is, the diagonals of ψ.

```
> # The two factors are stored under scores
> ct.FA$scores
```

	Factor1	Factor2
Buenos Aires	-0.17889040	0.871144421
Dhaka	-1.03956967	-1.209158145
Rio de Janeiro	-0.32938587	0.902073847
Sao Paolo	1.66151441	-1.255368200
Beijing	-0.09985343	-0.037292081
Shanghai	0.57846585	-0.279218251
Tianjin	-0.52473954	0.005131718
Cairo	-0.78794455	0.540001986
Bombay	0.42386495	-1.666913485
Calcutta	-0.30956389	0.348401654
Delhi	-0.78137675	-0.333084139
Jakarta	-0.47221026	-0.711526317
Osaka	-0.50879573	1.824775539
Tokyo	3.25066807	-0.170749103
Seoul	-0.25369592	0.450506136
Mexico City	0.68217916	0.925132375
Lagos	-0.67890390	-1.663405574
Karachi	-0.79825159	-0.596300773
Manila	-0.59030193	-0.393923970
Los Angeles	-0.13777899	0.946137238
New York	0.89456997	1.503635125

Large cities




P.198
Vhc
P.199
Phn
P.2
2
1
Elc
Wtr
PrR
Are
Fod
Grw

- Notice the first factor from the following two models are not the same,

```
> ct.FA = factanal(X, factors = 2, rotation="none",
+                  scores="regression")
> ct.1.FA = factanal(X,factors = 1,rotation="none",
+                  scores="regression")
> cbind(ct.FA$loadings, ct.1.FA$loadings)
```

	Factor1	Factor2	Factor1
Area	0.1288805	0.2253654	0.1840646
Pop.1980	0.9142481	0.3771544	0.9959482
Pop.1990	0.9884345	0.1356698	0.9589555
Pop.2000	0.9673852	-0.2433919	0.7995058
Growth	-0.3856607	-0.8198362	-0.6800145
Food	-0.3156484	-0.6809724	-0.5238999
PersRoom	-0.3677294	-0.6914914	-0.5697535
Water	0.5581297	0.3836780	0.6317289
Elec	0.3662733	0.5103482	0.4915450
Phones	0.5177603	0.5533497	0.6758071
Vehicles	0.3558366	0.4094097	0.4306993

- Similar to what we have defined for PCA, we can define

$$R^2 = \frac{\sum_{j}^{k}\sum_{q}^{\ell} w_{jq}^2}{k} = \frac{\sum_{j}^{k} \lambda_j}{k}$$

- However, R^2 should not be used to assess the fit of a factor model
- A likelihood ratio test is provided by the output

```
> ct.FA
```

```
Call:
factanal(x = X, factors = 2, scores = "regression", rotation = "none")

Test of the hypothesis that 2 factors are sufficient.
The chi square statistic is 83.44 on 34 degrees of freedom.
The p-value is 4.86e-06
```

which is similar to the deviance χ^2 test we had for glm, and can be used to test the fit of a factor model.