Factor Analysis

UBCO MDS — DATA 573

Factor Analysis

- Assumes a small number of unobserved (latent) variables are driving the observed (manifest) variables.
- Example (that may have no basis in reality):
  - Observed/Manifest — University students' exam scores from Math, Biology, English, Chemistry, History
  - Unobserved/Latent — Perhaps driven by "Science" and "Arts" aptitudes/interests/etc

Latent Variable Models

- Factor analysis is a special case of latent variable models
  - Factor analysis — continuous latent, continuous manifest
  - Latent trait analysis — continuous latent, categorical manifest
  - Latent profile analysis — categorical latent, continuous manifest
  - Latent class analysis — categorical latent, categorical manifest

Factor Analysis Model

- While some of the goals and language are similar to PCA, FA is a 'true' statistical model

$$\mathbf{x}_i = \boldsymbol{\mu} + \boldsymbol{\Gamma}\mathbf{u}_i + \boldsymbol{\epsilon}_i$$

  - where $\mathbf{x}_i$ is the i^{th} observed data vector
  - $\boldsymbol{\mu}$ is the mean vector
  - $\boldsymbol{\Gamma}$ are the factor loadings/coefficients
  - $\mathbf{u}_i$ is the i^{th} latent vector, assumed $\text{MVN}(0, I_q)$, where q is the number of latent variables
  - ϵ_i is the i^{th} error, assumed $\text{MVN}(0, \boldsymbol{\Psi})$, where $\boldsymbol{\Psi}$ is a diagonal matrix of error variance

Factor Analysis Model

- Under the assumptions outlined on the previous slide, one can show that the marginal distribution of $\mathbf{X}$ is

$$\mathbf{X} \sim MVN(\boldsymbol{\mu}, \boldsymbol{\Gamma}\boldsymbol{\Gamma}' + \boldsymbol{\Psi})$$

- So then the key to factor analysis is the following decomposition:

$$\boldsymbol{\Sigma} \approx \boldsymbol{\Gamma}\boldsymbol{\Gamma}' + \boldsymbol{\Psi}$$

- Note that the covariance $\boldsymbol{\Sigma}$ has $\frac{p(p+1)}{2}$ free parameters
- While $\boldsymbol{\Gamma}\boldsymbol{\Gamma}' + \boldsymbol{\Psi}$ has $pq - \frac{q(q-1)}{2} + p$ free parameters

Factor Analysis Model

- Thus, factor analysis requires

$$\frac{p(p+1)}{2} \geq pq - \frac{q(q-1)}{2} + p$$

- Which implies

$$q \leq \frac{2p+1-\sqrt{8p-1}}{2}$$

as the max number of latent variables

- Quick guide:

Max q	For p
$q=1$	$p=3,4$
$q=2$	$p=5$
$q=3$	$p=6,7$
$\vdots$	$\vdots$

Non-uniqueness

- See board.
- This has big implications
- If loadings are 'uninterpretable', then they can be rotated.
- Negative view: this can be abused in applications where people seek to justify the latent variables that they expect
- There are also semi-objective rotations
  - Varimax — seek $\boldsymbol{\Gamma}$ such that each variable loads heavily on only one factor.
  - Quartimax — seek $\boldsymbol{\Gamma}$ such that q (number of latent variables) is minimized.
  - Promax — non-orthogonal rotation (called oblique rotation), popular for increasing the amount of variance explained by small q...allows correlation among factors.

Comments

- Factor analysis and PCA are often confused — similar terminology (loadings, scores, etc), similar interpretation

- To add to the confusion, PCA can often be used to aid the estimation process of FA (and therefore often appears in FA software).

- BUT

    - PCA is simply a rotation of X, with closed form solutions, with no loss of information (until components are removed), and no distributional assumptions
    - FA is a statistical model, with no closed form solutions, with an expected loss of information (hopefully minimal), and assumed distributions (MVN $\mathbf{u}$ and ϵ)

FA on Decathlon Data

```
> od
         100m     LJ      SP     HJ   400m  100mH     DS     PV     JV   1500m
100m   1.0000 0.6386  0.4752 0.3227 0.5520 0.3262 0.3509 0.4008 0.1821 -0.0352
LJ     0.6386 1.0000  0.4953 0.5668 0.4706 0.3520 0.3998 0.5167 0.3102  0.1012
SP     0.4752 0.4953  1.0000 0.4357 0.2539 0.2812 0.7926 0.4728 0.4682 -0.0120
HJ     0.3227 0.5668  0.4357 1.0000 0.3449 0.3503 0.3657 0.6040 0.2344  0.2380
400m   0.5520 0.4706  0.2539 0.3449 1.0000 0.1546 0.2100 0.4213 0.2116  0.4125
100mH  0.3262 0.3520  0.2812 0.3503 0.1546 1.0000 0.2553 0.4163 0.1712  0.0002
DS     0.3509 0.3998  0.7926 0.3657 0.2100 0.2553 1.0000 0.4036 0.4179  0.0109
PV     0.4008 0.5167  0.4728 0.6040 0.4213 0.4163 0.4036 1.0000 0.3151  0.2395
JV     0.1821 0.3102  0.4682 0.2344 0.2116 0.1712 0.4179 0.3151 1.0000  0.0983
1500m -0.0352 0.1012 -0.0120 0.2380 0.4125 0.0002 0.0109 0.2395 0.0983  1.0000
```

FA on Decathlon Data

- Covariance matrix from 44 years of olympic decathlons (so I guess 11 competitions).
- The factanal command can be run on a standard matrix of predictors, or just on a covariance matrix.
- Let's run it and look at the output

FA on Decathlon Data

```
> decfa <- factanal(covmat=od, factors=2, n.obs=280, rotation="none")
> decfa

Uniquenesses:
 100m    LJ    SP    HJ  400m 100mH    DS    PV    JV 1500m 
0.531 0.374 0.063 0.540 0.559 0.804 0.330 0.490 0.758 0.885 

Loadings:
      Factor1 Factor2
100m   0.552   0.405 
LJ     0.602   0.513 
SP     0.956  -0.151 
HJ     0.525   0.430 
400m   0.353   0.563 
100mH  0.348   0.275 
DS     0.806  -0.146 
PV     0.566   0.436 
JV     0.492         
1500m          0.337 
```

FA on Decathlon Data

```
               Factor1 Factor2
SS loadings      3.313   1.354
Proportion Var   0.331   0.135
Cumulative Var   0.331   0.467

Test of the hypothesis that 2 factors are sufficient.
The chi square statistic is 192.36 on 26 degrees of freedom.
The p-value is 2.53e-27
```

FA

- As FA is a statistical model, there is much more output associated here than with PCA.
- First up: uniqueness.
- This provides the values along the diagonal of $\mathbf{\Psi}$. AKA the corrective terms for the variances.
- Hence, larger values would indicate that the variable in question is not being captured with the current factors.

FA

- Loadings are the coefficients of the linear combinations (similar to PC loadings), we'll talk about these more shortly.

- We are also provided information on the proportion of variance explained by each factor

- Finally, we have a hypothesis test!

- H_0 is that the number of factors are sufficient to describe the data, the alternative being that more are needed. So we are hoping to fail to reject! In this case, we are not...how about 3 factors?

FA on Decathlon Data

```
> decfa <- factanal(covmat=od, factors=3, n.obs=280, rotation="none")
> decfa

Uniquenesses:
 100m    LJ    SP    HJ  400m 100mH    DS    PV    JV 1500m
0.005 0.399 0.083 0.414 0.549 0.796 0.311 0.401 0.736 0.766

Loadings:
      Factor1 Factor2 Factor3
100m   0.997
LJ     0.652   0.263   0.327
SP     0.505   0.803  -0.130
HJ     0.342   0.413   0.546
400m   0.556           0.376
100mH  0.335   0.192   0.234
DS     0.378   0.732
PV     0.419   0.408   0.507
JV     0.200   0.467
1500m                  0.477
```

FA on Decathlon Data

```
               Factor1 Factor2 Factor3
SS loadings      2.572   1.849   1.118
Proportion Var   0.257   0.185   0.112
Cumulative Var   0.257   0.442   0.554

Test of the hypothesis that 3 factors are sufficient.
The chi square statistic is 85.03 on 18 degrees of freedom.
The p-value is 1.11e-10
```

FA on Decathlon Data

```
> decfa <- factanal(covmat=od, factors=4, n.obs=280, rotation="none")
> decfa

Uniquenesses:
 100m    LJ    SP    HJ  400m 100mH    DS    PV    JV 1500m
0.010 0.388 0.089 0.327 0.196 0.734 0.304 0.420 0.725 0.600

Loadings:
      Factor1 Factor2 Factor3 Factor4
100m   0.993
LJ     0.665   0.252   0.239   0.220
SP     0.530   0.777  -0.141
HJ     0.363   0.428   0.421   0.425
400m   0.571           0.620  -0.304
100mH  0.343   0.190           0.323
DS     0.401   0.718  -0.102
PV     0.439   0.407   0.390   0.263
JV     0.218   0.461
1500m                  0.609  -0.145
```

FA on Decathlon Data

```
               Factor1 Factor2 Factor3 Factor4
SS loadings      2.686   1.794   1.187   0.539
Proportion Var   0.269   0.179   0.119   0.054
Cumulative Var   0.269   0.448   0.567   0.621

Test of the hypothesis that 4 factors are sufficient.
The chi square statistic is 15.74 on 11 degrees of freedom.
The p-value is 0.151
```

FA on Decathlon Data

```
> decfa <- factanal(covmat=od, factors=4, n.obs=280, rotation="varimax")
> decfa

Uniquenesses:
 100m    LJ    SP    HJ  400m 100mH    DS    PV    JV 1500m
0.010 0.388 0.089 0.327 0.196 0.734 0.304 0.420 0.725 0.600

Loadings:
      Factor1 Factor2 Factor3 Factor4
100m   0.205   0.296   0.928
LJ     0.280   0.554   0.451   0.155
SP     0.883   0.278   0.228
HJ     0.254   0.739           0.242
400m   0.142   0.151   0.519   0.701
100mH  0.136   0.465   0.173
DS     0.794   0.220   0.133
PV     0.314   0.612   0.168   0.279
JV     0.477   0.160           0.139
1500m          0.111           0.619
```

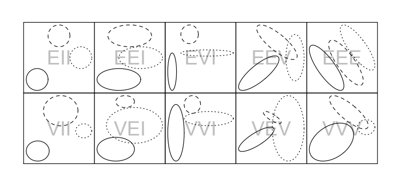
FA on Decathlon Data

```
               Factor1 Factor2 Factor3 Factor4
SS loadings      1.958   1.719   1.471   1.057
Proportion Var   0.196   0.172   0.147   0.106
Cumulative Var   0.196   0.368   0.515   0.621

Test of the hypothesis that 4 factors are sufficient.
The chi square statistic is 15.74 on 11 degrees of freedom.
The p-value is 0.151
```

Recall: Mixture Model Families

- To combat overparameterization, we often develop ‘families’ of mixture distributions via various decompositions of the covariance structure.
  - Facilitates parsimony and clustering flexibility
  - Occasionally gives interesting geometric properties (MCLUST)



- $\mathbf{\Sigma}_g = \lambda_g \mathbf{D}_g \mathbf{A}_g \mathbf{D}_g'$ (GPCM, MCLUST, tEIGEN)
- $\mathbf{\Sigma}_g = \mathbf{\Gamma}_g \mathbf{\Gamma}_g' + \mathbf{\Psi}_g$ (MCFA, PGMM, MMtFA)

An Argument for Factor Analyzers

- We've noted the decomposition $\mathbf{\Sigma}_g = \lambda_g \mathbf{D}_g \mathbf{A}_g \mathbf{D}_g'$.
- Biggest parameter savings come from models that assume local independence among the variables (ie, $\mathbf{\Sigma}_g = \lambda \mathbf{A}$) — a strong, and usually invalid, assumption on real data.
- Fortunately, there is another option...

Factor Analyzers

- Factor analysis model: $\mathbf{X}_i = \mu + \mathbf{\Gamma}\mathbf{U}_i + \epsilon_i$
  - $\mathbf{\Gamma}$ is a $p \times q$ matrix of factor loadings (where $q << p$)
  - $\epsilon \sim N(0, \mathbf{\Psi})$ are the error terms
    ($\mathbf{\Psi}$ a diagonal $p \times p$ matrix called error variance)
  - Then the marginal of $\mathbf{X}_i$ is $N(\boldsymbol{\mu}, \mathbf{\Gamma}\mathbf{\Gamma}' + \mathbf{\Psi})$

- This leads to mixtures of factor analyzers. In the Gaussian case:

$$f(\mathbf{x}) = \sum_{g=1}^{G} \pi_g \phi(\mathbf{x} \mid \boldsymbol{\mu}_g, \mathbf{\Gamma}_g\mathbf{\Gamma}_g' + \mathbf{\Psi}_g),$$

Why Factor Analyzers?

- For technical reasons, the factor analysis covariance structure is attractive.

- Normally, we have $Gp(p+1)/2$ parameters to estimate within the covariance structure.

- Under this structure, we instead have $G(pq - q(q-1)/2 + Gp$ where q is the number of latent factors...and $p >> q$

- This means our covariance parameters grow *linear* with respect to the dimension of the data p rather than *quadratic*.

Why Factor Analyzers? (continued)

- Put simply, the majority of mixture models (including non-constrained $\boldsymbol{\Sigma}_g$ and most of the MCLUST models) cannot be run on ‘small n large p’ data (like in bioinformatics)
- Using mixtures of factor analyzers lessens the blow of the ‘curse of dimensionality’.
- In fact, we will also see that factor analyzers can give better clustering performance even on relatively low dimensional data sets.

The PGMM/MODt Families

Modified Covariance Structure	Covariance Parameters
$\boldsymbol{\Sigma}_g = \boldsymbol{\Gamma}\boldsymbol{\Gamma}' + \varpi\mathbf{I}_p$	$[pq - q(q-1)/2] + 1$
$\boldsymbol{\Sigma}_g = \boldsymbol{\Gamma}\boldsymbol{\Gamma}' + \varpi_g\mathbf{I}_p$	$[pq - q(q-1)/2] + G$
$\boldsymbol{\Sigma}_g = \boldsymbol{\Gamma}_g\boldsymbol{\Gamma}_g' + \varpi\mathbf{I}_p$	$G[pq - q(q-1)/2] + 1$
$\boldsymbol{\Sigma}_g = \boldsymbol{\Gamma}_g\boldsymbol{\Gamma}_g' + \varpi_g\mathbf{I}_p$	$G[pq - q(q-1)/2] + G$
$\boldsymbol{\Sigma}_g = \boldsymbol{\Gamma}\boldsymbol{\Gamma}' + \varpi\boldsymbol{\Delta}$	$[pq - q(q-1)/2] + p$
$\boldsymbol{\Sigma}_g = \boldsymbol{\Gamma}\boldsymbol{\Gamma}' + \varpi_g\boldsymbol{\Delta}$	$[pq - q(q-1)/2] + [G + (p-1)]$
$\boldsymbol{\Sigma}_g = \boldsymbol{\Gamma}_g\boldsymbol{\Gamma}_g' + \varpi\boldsymbol{\Delta}$	$G[pq - q(q-1)/2] + p$
$\boldsymbol{\Sigma}_g = \boldsymbol{\Gamma}_g\boldsymbol{\Gamma}_g' + \varpi_g\boldsymbol{\Delta}$	$G[pq - q(q-1)/2] + [G + (p-1)]$
$\boldsymbol{\Sigma}_g = \boldsymbol{\Gamma}\boldsymbol{\Gamma}' + \varpi\boldsymbol{\Delta}_g$	$[pq - q(q-1)/2] + [1 + G(p-1)]$
$\boldsymbol{\Sigma}_g = \boldsymbol{\Gamma}\boldsymbol{\Gamma}' + \varpi_g\boldsymbol{\Delta}_g$	$[pq - q(q-1)/2] + Gp$
$\boldsymbol{\Sigma}_g = \boldsymbol{\Gamma}_g\boldsymbol{\Gamma}_g' + \varpi\boldsymbol{\Delta}_g$	$G[pq - q(q-1)/2] + [1 + G(p-1)]$
$\boldsymbol{\Sigma}_g = \boldsymbol{\Gamma}_g\boldsymbol{\Gamma}_g' + \varpi_g\boldsymbol{\Delta}_g$	$G[pq - q(q-1)/2] + Gp$

Clustering Wine

	MODt			EPGMM			
	1	2	3	1	2	3	4
Barolo	59			57		2	
Grignolino	2	69			46	25	
Barbera			48				48

Method	ARI	G	q	Cov	DF
MODt	0.96	3	4	$\boldsymbol{\Gamma}\boldsymbol{\Gamma}' + \varpi_g \boldsymbol{\Delta}_g$	16.6
MCLUST	0.90	3	–	$\lambda_g \mathbf{A}_g$	–
tEIGEN	0.83	3	–	$\lambda \mathbf{A}_g$	13.6
EPGMM	0.80	4	4	$\boldsymbol{\Gamma}\boldsymbol{\Gamma}' + \varpi_g \boldsymbol{\Delta}_g$	–

Comments

- As is the case most mixture modelling, fitting mixtures of factor analyzers can be time consuming.

- Some more flexible options (beyond Gaussian and t) are available, such as mixtures of generalized hyperbolic factor analyzers — package MGHFA

UBC THE UNIVERSITY OF BRITISH COLUMBIA