

Measurement in the Social Sciences: Principal Components and Factor Analysis (1)

Dave Armstrong

University of Oxford
Department of Politics and International Relations
Center for Research Methods in the Social Sciences

t: 01865 285956
e: david.armstrong@politics.ox.ac.uk
w: <http://users.ox.ac.uk/~polf0104>

Office: 1101 Manor Road Building (Politics Dept)
Office Hours: TWTh 1400-1500

May 7, 2009

1/28

Outline

- Principal Components: Theory
- Principal Components: Estimation
 - SVD - what is it and what does it do?
- PCA examples
- Intro to Factor Analysis

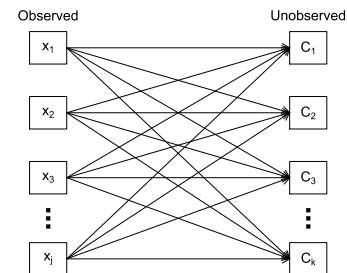
2/28

What you'll learn today

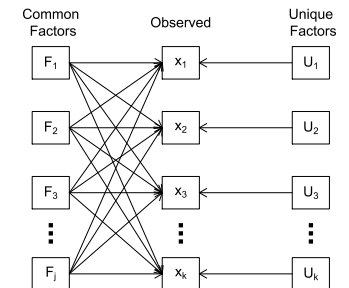
1. The difference between Principal Components and Factor Analysis
2. How Stata estimates the parameters in a PCA
3. How to interpret and use principal components analysis
4. The theoretical model behind Factor Analysis

3/28

Theoretical Models



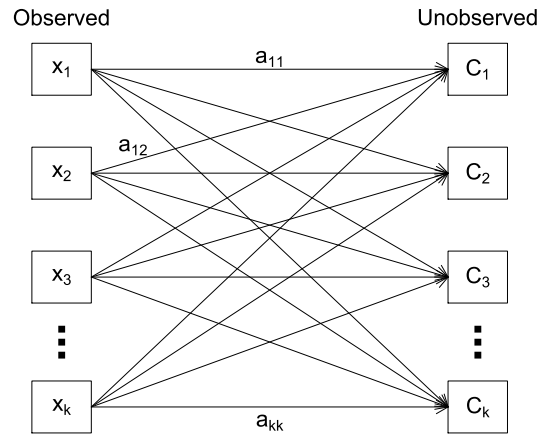
(a) Principal Components



(b) Common Factor Model

4/28

Graphical Representation of Principal Components



5 / 28

Principal Components: Equation Form

$$\begin{aligned}
 C_1 &= a_{11}x_1 + a_{21}x_2 + \dots + a_{k1}x_k \\
 C_2 &= a_{12}x_1 + a_{22}x_2 + \dots + a_{k2}x_k \\
 &\vdots \\
 C_k &= a_{1k}x_1 + a_{2k}x_2 + \dots + a_{kk}x_k
 \end{aligned}$$

6 / 28

Things to note

- There are as many C variables as there are x variables. So Principal Components analysis does not necessarily reduce dimensionality. It is unlikely you can *perfectly* reproduce the correlation matrix with fewer than k components unless one x is a perfect linear combination of the other x 's.
- The causal direction goes from the observed to the unobserved variables, so this is not a "model" that proposes a small set of variables that can account for the covariation in the observed variables.
- Notice, there are no errors here. We are accounting for all of the variance in the observed variables by using k components to summarize k variables.

7 / 28

PCA as Dimension Reductions

- PCA can be used to reduce dimensionality if a relatively small number of the k components account for a disproportionately large proportion of the combined variance in the observed variables.
 - Remember that our variables are all standardized now, so if we have 10 variables, and each one has a variance of 1, we will have a combined variance of 10. If we could produce one component that accounted for a variance of 9, then we could use that one variable without losing *much* of the original information (w.r.t variance).
- This will induce some error because we are using fewer than k components, but note that in the "model" there is no error. So this error and the uniqueness we see in factor analysis are somewhat different.
- Further, PCA is not attempting to estimate some population parameter. PCA produces an orthogonal linear combination of the observed variables.

8 / 28

What we need

- A set of coefficients relating observed variables to a set of latent variables.
- These will be unidentified unless we impose some constraints here.
 - The resulting new variables (components) will be uncorrelated (orthogonal).
 - The first component will capture the most variance in the observed variables, the second component will capture the second most variance, etc ...
- In matrix form, we need: $C = XA$.
- In scalar form: $C_j = a_{1j}x_1 + a_{2j}x_2 + \dots + a_{jj}x_j \forall j = 1, \dots, k$.

9/28

Singular Value Decomposition

- We employ a mathematical tool called the Singular Value Decomposition.
 - This uncovers the “basic structure” of a matrix.
 - It has three components: $\underbrace{U}_{n \times k}$, $\underbrace{D}_{k \times k}$ and $\underbrace{V'}_{k \times k}$ where:
 - n is the number of rows in the original data, and
 - k is the number of columns in the original data
- U gives information about the rows of the original matrix, V gives information about the columns of the original matrix, and D gives the variance accounted for by the rows of U and V .
- The d 's are what we call the “eigenvalues” when the original matrix is square and symmetric.
- Here, $X = UDV'$

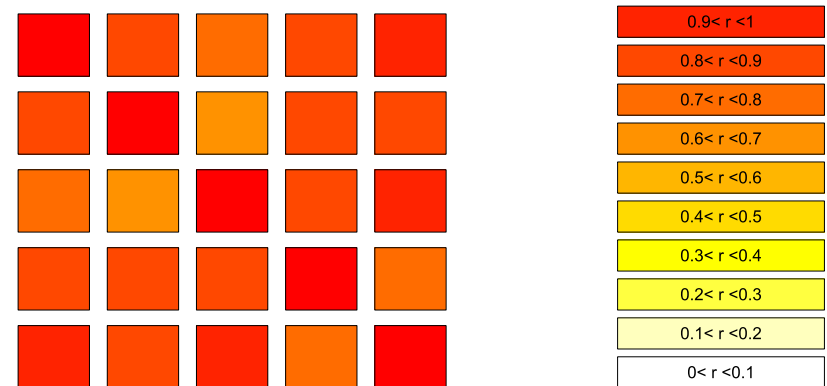
10/28

A Small Digression about SVD

- Sometimes, when you're unsure about what a particular “thing” does, it might make sense to see what it does on made-up data, data where you know the properties. So, if SVD *actually* uncovers the basic structure of a matrix, let's see what kind of results we get in different situations.
- I've got two different situations below - one where there is one underlying trait and one where there are two, mostly uncorrelated underlying traits.

11/28

One Underlying Trait: Graphical Correlation Matrix



12/28

One Underlying Trait: SVD

Eigenvalues

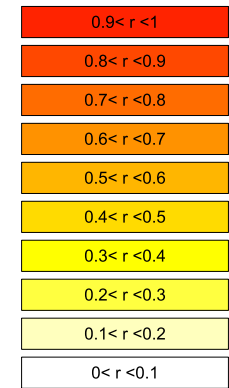
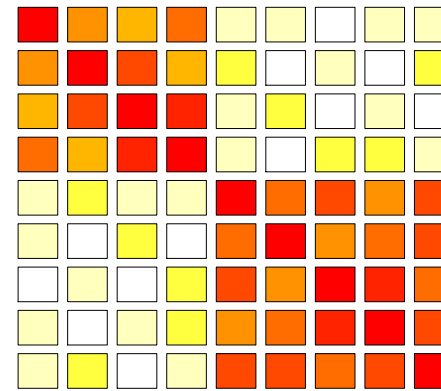
4.12 0.43 0.31 0.17 0.04

Eigenvectors

-0.45 0.21 -0.33 -0.72 -0.33
 -0.43 0.61 -0.03 0.59 -0.26
 -0.43 -0.70 0.18 0.20 -0.48
 -0.44 0.12 0.75 -0.22 0.40
 -0.46 -0.24 -0.52 0.17 0.64

13/28

Two Underlying Traits: Graphical Correlation Matrix



14/28

Two Underlying Traits: SVD

Eigenvalues

4.11 2.84 0.61 0.55 0.47 0.34 0.20 0.12 0.01

Eigenvectors

-0.16 0.43 0.07 -0.80 -0.15 -0.01 0.22 0.26 -0.02
 -0.18 0.45 -0.55 0.17 -0.37 0.40 0.02 -0.37 0.01
 -0.19 0.50 -0.06 0.46 0.33 -0.07 0.09 0.60 0.13
 -0.20 0.48 0.52 0.08 0.13 -0.27 -0.29 -0.50 -0.17
 -0.41 -0.15 -0.25 0.01 -0.31 -0.70 -0.06 -0.04 0.40
 -0.40 -0.16 -0.30 -0.13 0.63 -0.06 0.42 -0.29 -0.23
 -0.42 -0.18 0.26 0.25 -0.45 0.03 0.29 0.19 -0.58
 -0.43 -0.16 0.41 0.02 0.03 0.47 0.15 -0.09 0.61
 -0.43 -0.17 -0.17 -0.17 0.13 0.23 -0.76 0.22 -0.18

15/28

PCA and SVD

We have the following pieces of information:

$$\mathbf{X} = \mathbf{UDV}'$$

$$\mathbf{C} = \mathbf{XA}$$

What we need is $\underbrace{\mathbf{C}}_{n \times k}$ (something we don't know) = $\underbrace{\mathbf{X}}_{n \times k} \times \underbrace{\mathbf{A}}_{k \times k}$ (something else we don't know). Let's do the following:

$$\mathbf{X} = \mathbf{UDV}'$$

$$\mathbf{XV} = \mathbf{UDV}'\mathbf{V}$$

$$\underbrace{\mathbf{X}}_{\mathbf{A}} \underbrace{\mathbf{V}}_{\mathbf{C}} = \underbrace{\mathbf{UD}}_{\mathbf{C}}$$

We can use the Singular Value Decomposition to solve for the coefficients relating the observed variables to the latent component and to generate an estimate for the latent component.

16/28

How do we *do* this?

- `pca` in Stata will do this for you. If we look at the file `srn.dta` and issue the command `pca X*`, we'll get the following:

```
Principal components/correlation      Number of obs   =   1000
                                     Number of comp. =    4
                                     Trace            =    4
                                     Rho              =   1.0000
```

Rotation: (unrotated = principal)

Component	Eigenvalue	Difference	Proportion	Cumulative
Comp1	3.41316	3.20048	0.8533	0.8533
Comp2	.212683	.0117253	0.0532	0.9065
Comp3	.200958	.0277595	0.0502	0.9567
Comp4	.173199	.	0.0433	1.0000

Principal components (eigenvectors)

Variable	Comp1	Comp2	Comp3	Comp4	Unexplained
X1	0.4981	-0.4465	0.7386	-0.0837	0
X2	0.4964	0.8316	0.1866	0.1648	0
X3	0.5021	-0.3260	-0.4614	0.6548	0
X4	0.5034	-0.0532	-0.4547	-0.7328	0

17/28

Interpreting Output

- The first pane gives some basic information.
- The second pane provides eigenvalues as well as the proportion and cumulative proportion of variance accounted for by each component.
 - What we see here is a relatively large eigenvalue on the first component and relatively small ones on the remaining components. We can interpret this to mean these data would be relatively well represented by a single dimension (instead of 4). The first component accounts for 85% of the variance in the four variables $\frac{3.413}{4}$.
- The third pane gives the coefficients by which we must multiply the data to obtain the components.
- It is important to note here, that we are defining the components as a variance-maximizing linear transformation of the original data. Therefore, we obtain actual values of the components and not estimates of these. With factor analysis, we will get estimates of the factors, but not the actual factors themselves.

18/28

Interpreting Principal Components

- After you have performed the PCA, you can then create a new variable or new variables that summarize(s) the information in the observed variables. This new variable has no inherent metric, it will be scaled to have mean = 0 and variance=1.
- We can interpret this new variable as a function of all the observed variables. It is, in fact, a weighted average of the observed variables and variables with higher coefficients have more weight in determining the score on that component.
- Further, we can also look at the variance of the components.

$$\begin{aligned}
 C_j &= a_{1j}x_1 + a_{2j}x_2 + \dots + a_{jj}x_j \\
 \text{var}(C_j) &= \text{var}(a_{1j}x_1 + a_{2j}x_2 + \dots + a_{jj}x_j) \\
 &= a_{1j}^2 \text{var}(x_1) + a_{2j}^2 \text{var}(x_2) + \dots + a_{jj}^2 \text{var}(x_j)
 \end{aligned}$$

So the variance in the component C_j is a weighted sum of the variances of each of the individual observed variables. The variance in the component is more a function of the variance in observed variables with big coefficients than those with small coefficients.

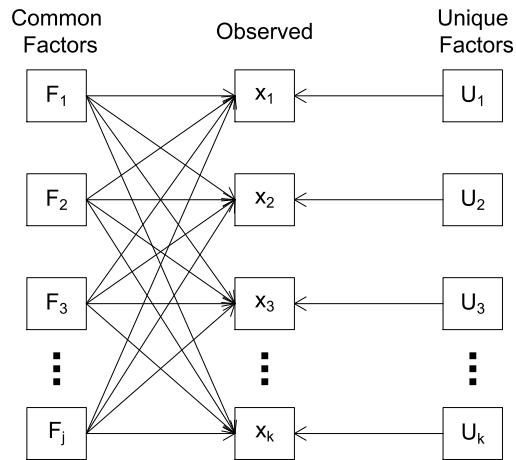
19/28

PCA: Conclusions and Caveats

- Principal Components is not a “latent variable model” the same way as Factor Analysis and the SRM are. At base, it takes k variables and produces k new uncorrelated variables that are linear combinations of the original variables.
- To the extent that m (where $m \ll k$) of these components account for a disproportionately large amount of the variance in the original variables, then these m components can be used as a proxy for the original variables, but you will be losing information.
- You should *never* rotate a principal components solution. The PCA solution is defined as one where the components are orthogonal and the components are extracted in order of variance explained. Rotating the solution could result in changing these properties.
- This model is appropriate for variables where correlations make sense. If correlations are meaningless, then this model doesn't make sense either.

20/28

Graphical Representation of Common Factor Model



21 / 28

Common Factor Model in Equation Form

$$\begin{aligned} x_{i1} &= a_{11}F_{i1} + a_{12}F_{i2} + \dots + a_{1m}F_{im} \\ x_{i2} &= a_{21}F_{i1} + a_{22}F_{i2} + \dots + a_{2m}F_{im} \\ &\vdots \\ x_{ik} &= a_{k1}F_{i1} + a_{k2}F_{i2} + \dots + a_{km}F_{im} \end{aligned}$$

For each observation $i = 1, 2, \dots, n$ on each observed variable in $1, 2, \dots, k$ and each factor $1, 2, \dots, m$, where always $m \leq k$ and usually $m \ll k$.

22 / 28

Factor Analysis Glossary (1)

Factor Loading Coefficient relating the unobserved variable to the observed variable a_{km} .

Factor Pattern Matrix Matrix of factor loadings, usually referred to as **A**

$$\begin{bmatrix} a_{1,1} & a_{1,2} & \dots & a_{1,m} \\ a_{2,1} & a_{2,2} & \dots & a_{2,m} \\ \vdots & \vdots & \vdots & \vdots \\ a_{k,1} & a_{k,2} & \dots & a_{k,m} \end{bmatrix}$$

Communality Amount of an observed variable's variance shared with the other variables; usually denoted as $h_k^2 = \sum_{j=1}^m a_{kj}$.

Uniqueness Amount of an observed variable's variance *not* shared with the other variables; usually denoted U_k .

23 / 28

Factor Analysis Glossary (2)

Eigenvalue In an un-rotated factor solution, the amount of variance explained by each factor.

Rotation Factor solutions are only identified up to a rotation, meaning there are infinitely many solutions that are equally "good" in terms of variance explained (ability to reproduce the correlation matrix). Rotating means moving the factors around in space often so they explain the same amount of variance, but so they also have other desirable properties.

Factor Structure Matrix Asymmetric matrix of correlations between the observed variables and the factors. This is the same as the Factor Pattern Matrix for orthogonal factors, but these are not the same when we allow the factors to be correlated.

24 / 28

Factor Pattern Matrix

					$\sum_{j=1}^k a_{jm}^2$
	$a_{1,1}$	$a_{1,2}$	\cdots	$a_{1,m}$	h_1^2
	$a_{2,1}$	$a_{2,2}$	\cdots	$a_{2,m}$	h_2^2
	\vdots	\vdots	\vdots	\vdots	\vdots
	$a_{k,1}$	$a_{k,2}$	\cdots	$a_{k,m}$	h_k^2
$\sum_{l=1}^m a_{kl}^2$	λ_1^2	λ_2^2	\cdots	λ_m^2	

25 / 28

Things to note

- This model proposes that there is potentially less variance than the total combined variance to explain. In math, we would say: $\sum_{l=1}^m h_l^2 < k$. Remember, PCA is accounting for *all* of the variance in the observed variables.
 - This will be a source of uncertainty for us as we will have to get an estimate of the communality for each variable.
- This is a model of *linear* structure. Factor Analysis models the underlying linear association among the variables with a smaller set of factors.

26 / 28

What you learned today

1. The difference between Principal Components and Factor Analysis
2. How Stata estimates the parameters in a PCA
3. How to interpret and use principal components analysis
4. The theoretical model behind Factor Analysis

27 / 28

Next Week ...

Factor Analysis: Operational Model

28 / 28