

Measurement in the Social Sciences (TT 2009)

Appendix 2.1: PCA Extended Examples - Stata

Dave Armstrong
University of Oxford
Department of Politics and International Relations

e: david.armstrong@politics.ox.ac.uk
w: <http://www.quantoid.net/Oxford.php>

May 7, 2009

1 PCA for the 2004 NES

Open the dataset `nes2004.dta`, the same one we used last week. We will use these same data to investigate the output from a Principal Components Analysis. Before we start, you'll want to make sure that the variables are all standardized. We can do this as follows:

```
. sum spendserv-womrole
```

Variable	Obs	Mean	Std. Dev.	Min	Max
spendserv	1060	3.484906	1.592819	1	7
defspend	1061	4.573044	1.479925	1	7
insurance	1112	3.657374	1.920672	1	7
jobsliv	1103	4.209429	1.865394	1	7
gabblacks	1073	4.540541	1.78679	1	7
envjobs	1019	3.593719	1.575163	1	7
fedgun	1202	2.590682	1.924169	1	5
womrole	1157	1.917891	1.466819	1	7

What we want to see are means of 0 and standard deviations of 1. Since these variables are not standardized, we can create standardized versions as follows:

for var spendserv-womrole: egen ZX=std(X)

Now, we can put all of these into the PCA routine.

pca Z*

```
Principal components/correlation      Number of obs   =    749
                                      Number of comp. =     8
                                      Trace              =     8
Rotation: (unrotated = principal)    Rho              = 1.0000
```

Component	Eigenvalue	Difference	Proportion	Cumulative
Comp1	2.89185	1.88301	0.3615	0.3615
Comp2	1.00884	.0835706	0.1261	0.4876
Comp3	.925266	.0822657	0.1157	0.6032
Comp4	.843001	.144881	0.1054	0.7086
Comp5	.69812	.024348	0.0873	0.7959
Comp6	.673771	.138822	0.0842	0.8801
Comp7	.53495	.110739	0.0669	0.9470
Comp8	.424211	.	0.0530	1.0000

Principal components (eigenvectors)

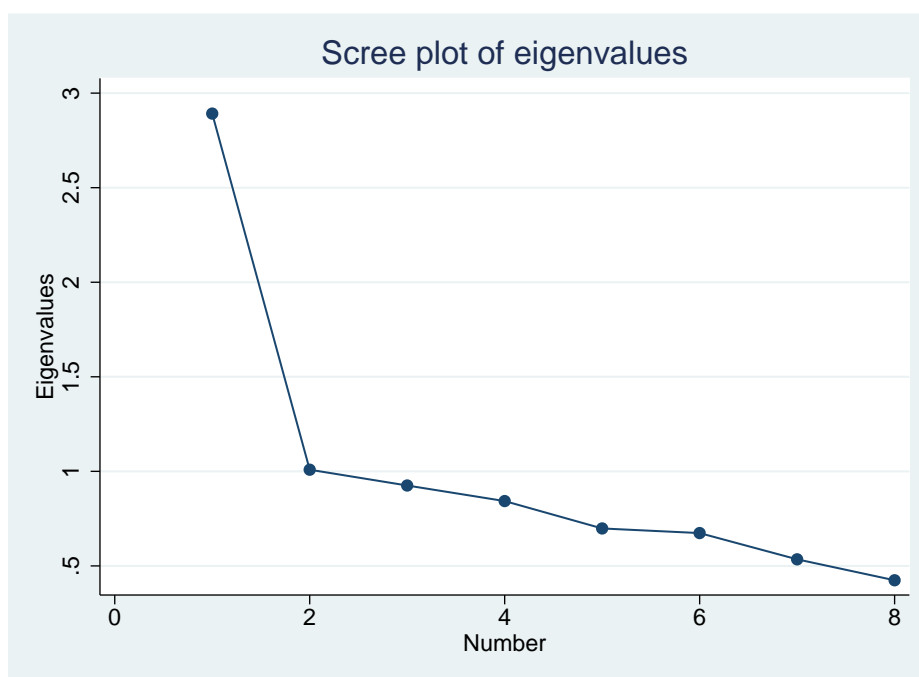
Variable	Comp1	Comp2	Comp3	Comp4	Comp5
Zspendserv	0.4103	-0.1722	-0.1899	-0.2876	-0.2977
Zdefspend	0.3116	0.1388	0.4551	0.5873	0.1309
Zinsurance	0.4298	-0.1663	-0.0695	-0.1279	-0.3725
Zjobsliv	0.4462	-0.3424	-0.0520	-0.0808	0.0967
Zgablacks	0.3852	-0.2779	0.0247	0.0855	0.6468
Zenvjobs	0.2881	0.3705	-0.4122	0.5547	-0.3240
Zfedgun	0.2541	0.3695	0.6875	-0.3269	-0.2186
Zwomrole	0.2343	0.6746	-0.3257	-0.3571	0.4184

Variable	Comp6	Comp7	Comp8	Unexplained
Zspendserv	-0.0551	0.7685	0.0320	0
Zdefspend	-0.5184	0.2118	-0.0285	0
Zinsurance	-0.3263	-0.5203	0.5005	0
Zjobsliv	-0.0127	-0.2897	-0.7622	0
Zgablacks	0.4313	0.0350	0.4011	0
Zenvjobs	0.4369	-0.0569	-0.0553	0
Zfedgun	0.4078	-0.0590	-0.0415	0
Zwomrole	-0.2800	-0.0447	-0.0311	0

Now, we have to evaluate this solution. First, when we look at the eigenvalues, we can see that the first component accounts for 36% of the combined variance $\frac{2.89}{8}$. The second component accounts for 12.6% of the combined variance. So, if we wanted to keep a lower-dimensional solution, we could keep 1 component and we would retain 36% of the variance in the original variables. If this doesn't seem like enough, then you can take more than 1 component until you think you have enough of the variance of the original variables. Just remember, that substantive interpretation gets harder the more of these components you include.

1.1 How to pick the right number of components?

There are a couple of different possibilities here. We will use this same strategy for factor analysis (and it probably makes more sense in that context), but we'll talk about it here, too. We will need to construct a plot called a "scree plot". This graphs the eigenvalue on the y-axis and the component number on the x-axis. You look for the "elbow" in the chart and the component before the elbow is the "optimal" solution. Stata will make this plot for you automatically with the command `greigen`.



The advice of the scree plot would be to pick 1 component because the elbow occurs at component 2. This would suggest that one component accounts for a disproportionately large amount of the combined variance.

Another widely used method for deciding how many components to keep is to just keep every component with an eigenvalue over 1. The reasoning here is that each individual variable accounts for a variance of 1. If a component accounts for a variance of

more than 1, then it accounts for more variance than any one of the original observed variables could. You can see above, however, that there is a reasonable one component solution, but the second factor's eigenvalue is slightly above 1.

Personally, I would go with the scree plot method, but it is unlikely that you would get much grief for using the eigenvalues over 1 strategy.

1.2 Interpretation

The one constant in scaling solutions of any flavor is that they do not interpret themselves. Principal components are no exception. You have to impose some meaning on the components so you can substantively interpret the effects they have in a predictive model. We can use the information in the “Principal Components (eigenvectors)” table for this task. These are the coefficients relating the observed variables to the coefficients and can be interpreted like linear model coefficients. We can call these coefficients “component loadings”. If one variable has a high component loading, that means as that variable gets bigger, the Component gets bigger. For example, the jobs - standard of living variable has the biggest component loading. Therefore, as the people think government should spend more money to provide more services, component 1 gets bigger. We can think of this as a left-right ideology variable because as people get more left, component 1 gets bigger.

Now someone could come along and propose a different meaning for this latent variable. Nothing makes my interpretation more “right” than anyone else's. The argument has to be made on claims of validity - that we would expect certain things from a left-right ideological placement variable (w.r.t. relationships with other variables) and to the extent that this variable exhibits those properties, it makes sense as a left-right ideological placement variable.

1.3 What is Stata doing?

Some may be interested to know what Stata is actually *doing* behind the scenes. First, Stata is getting the correlation matrix of the data, so it is not essential that the variables are standardized as this will happen in the computation of the correlation matrix. Then, Stata performs a singular value decomposition on the correlation matrix.¹ It would also be possible to perform a singular value decomposition on the $n \times k$ data matrix. The results are not identical, but are substantively similar. It is unlikely that one would provide a different understanding of what is going on in the data, even though the numbers may not be exactly the same.

¹An Eigen decomposition is a special case of a singular value decomposition performed on a square, symmetric matrix (like a correlation matrix).

2 An Ambiguous PCA Solution

Remember back to last week where we looked at data that were not appropriate for the summated rating model (from `bad.dta`). We know that they were not right for the SRM, but let's see what happens when we subject these data to the Principal Components model.

```
. pca Z*
```

```
Principal components/correlation      Number of obs   =      1000
                                      Number of comp. =         6
                                      Trace             =         6
Rotation: (unrotated = principal)    Rho              =      1.0000
```

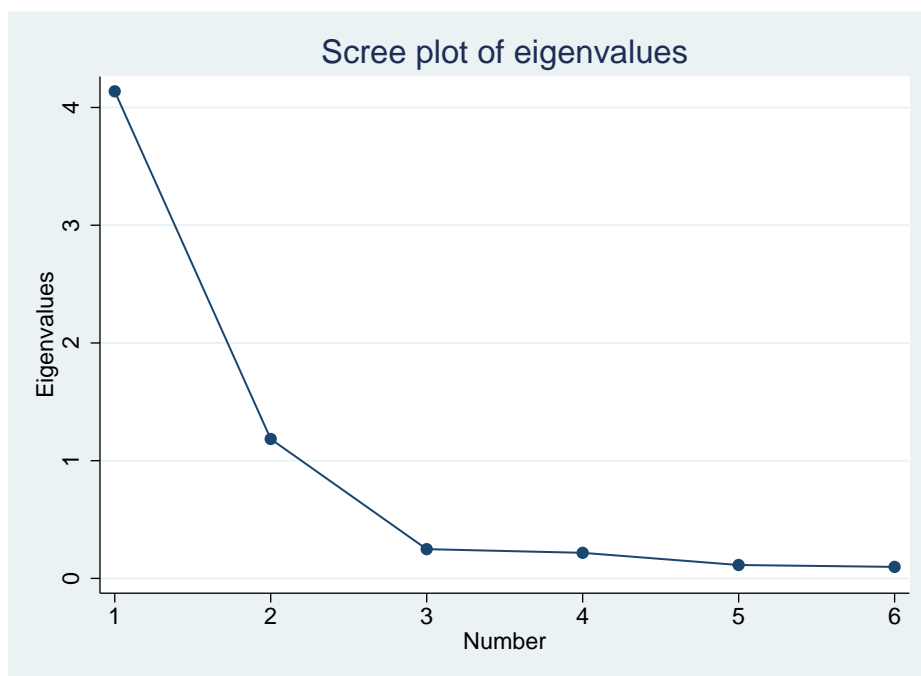
Component	Eigenvalue	Difference	Proportion	Cumulative
Comp1	4.13751	2.95309	0.6896	0.6896
Comp2	1.18442	.935786	0.1974	0.8870
Comp3	.248634	.0310505	0.0414	0.9284
Comp4	.217583	.103685	0.0363	0.9647
Comp5	.113898	.0159392	0.0190	0.9837
Comp6	.0979591	.	0.0163	1.0000

```
Principal components (eigenvectors)
```

Variable	Comp1	Comp2	Comp3	Comp4
ZV1	-0.3921	0.4294	-0.2243	0.7819
ZV2	-0.3914	0.4096	0.8004	-0.1918
ZV3	-0.3875	0.4370	-0.5530	-0.5927
ZV4	0.4250	0.3859	0.0287	0.0006
ZV5	0.4247	0.3943	0.0334	0.0177
ZV6	0.4265	0.3905	-0.0361	-0.0138

Variable	Comp5	Comp6	Unexplained
ZV1	0.0097	0.0065	0
ZV2	-0.0114	0.0393	0
ZV3	0.0082	-0.0408	0
ZV4	0.8064	-0.1392	0
ZV5	-0.5217	-0.6250	0
ZV6	-0.2782	0.7660	0

You can see that the first component explains a large amount of the combined variance (around 70%). We might be tempted to look only at the first component, but let's look at the scree plot first.



The scree plot would probably suggest 2 components, but there are two elbows in this plot. One at 2 and a bigger one at 3. It would seem like keeping 2 components would probably be better, but you may be able to make an argument for 1.