

Measurement in the Social Sciences (TT 2008)

Appendix 2.2: PCA Extended Examples - SPSS

Dave Armstrong
University of Oxford
Department of Politics and International Relations

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

May 1, 2008

1 PCA for the 2004 NES

Open the dataset `nes2004.sav`, 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 standardize your variables and drop out any of the missing values as follows:

```
DESCRIPTIVES
```

```
VARIABLES=spendserv defspend insurance jobsliv gablacks envjobs fedgun  
womrole /SAVE  
/STATISTICS=MEAN STDDEV MIN MAX .
```

```
DATASET COPY nes2004sub.sav.
```

```
DATASET ACTIVATE nes2004sub.sav.
```

```
FILTER OFF.
```

```
USE ALL.
```

```
SELECT IF(~MISSING(Zspendserv) & ~MISSING(Zdefspend) &  
~MISSING(Zinsurance) &  
~MISSING(Zjobsliv) & ~MISSING(Zgablacks) & ~MISSING(Zenvjobs) &  
~MISSING(Zfedgun) & ~MISSING(Zwomrole)).
```

```
DATASET ACTIVATE DataSet1.
```

```
EXECUTE .
```

This will create a new dataset called `nes2004sub.sav` which is a subset of the original data for which all of the observed variables have complete data. We can then use this

dataset to do the PCA.

It is important to note here that SPSS does not have a “canned” PCA routine. It has a zero-uniqueness factor analysis model (which we’ll talk more about next week), but this is slightly different from the PCA model and if you look at PCA in Stata versus the zero-uniqueness factor model in SPSS, you will see different results. I will provide the SPSS syntax needed to compute a principal components analysis, though, it will just be a bit more difficult than in Stata.

Open the data file `nes2004sub.sav` and then the syntax files `pca.sps`. This file gives the syntax for performing a principal components analysis. The things you’ll need to change in the syntax (when you do your own work) are:

- Line 2:

```
GET X /VARIABLES = Zspendserv Zdefspend Zinsurance Zjobsliv  
Zgablacks Zenvjobs Zfedgun Zwomrole.
```

Here, you’ll need to input all the variables you want to include in the PCA.

- Line 4:

```
COMPUTE VNames={"SpendServ","DefSpend","Insurance","JobsLiving",  
"GABlacks","EnvJobs","FedGun","WomRole"}.
```

Here, you can give the variable names, something more descriptive and I think you can include spaces here.

- Line 5:

```
COMPUTE PCRNAMES={"PCR1","PCR2","PCR3","PCR4","PCR5","PCR6",  
"PCR7","PCR8"}.
```

Here, you need to name the components. You need one component name for every observed variable in the model.

Now, we can run the `pca` syntax file on the subset dataset and we get the following results.

Run MATRIX procedure:

Correlation Matrix

	SpendSer	DefSpend	Insuranc	JobsLivi	GABlacks	EnvJobs	FedGun	WomRole
SpendSer	1.000	.202	.464	.468	.350	.242	.201	.209
DefSpend	.202	1.000	.286	.282	.268	.225	.240	.123

Insuranc	.464	.286	1.000	.520	.327	.255	.220	.196
JobsLivi	.468	.282	.520	1.000	.491	.227	.194	.157
GABlacks	.350	.268	.327	.491	1.000	.218	.183	.140
EnvJobs	.242	.225	.255	.227	.218	1.000	.107	.230
FedGun	.201	.240	.220	.194	.183	.107	1.000	.176
WomRole	.209	.123	.196	.157	.140	.230	.176	1.000

Principal Components Results Using the Correlation

Variance of Components (Using COR)

PCR1	2.892
PCR2	1.009
PCR3	.925
PCR4	.843
PCR5	.698
PCR6	.674
PCR7	.535
PCR8	.424

Proportion of Variance Components (Using COR)

PCR1	.361
PCR2	.126
PCR3	.116
PCR4	.105
PCR5	.087
PCR6	.084
PCR7	.067
PCR8	.053

Loadings (Using COR)

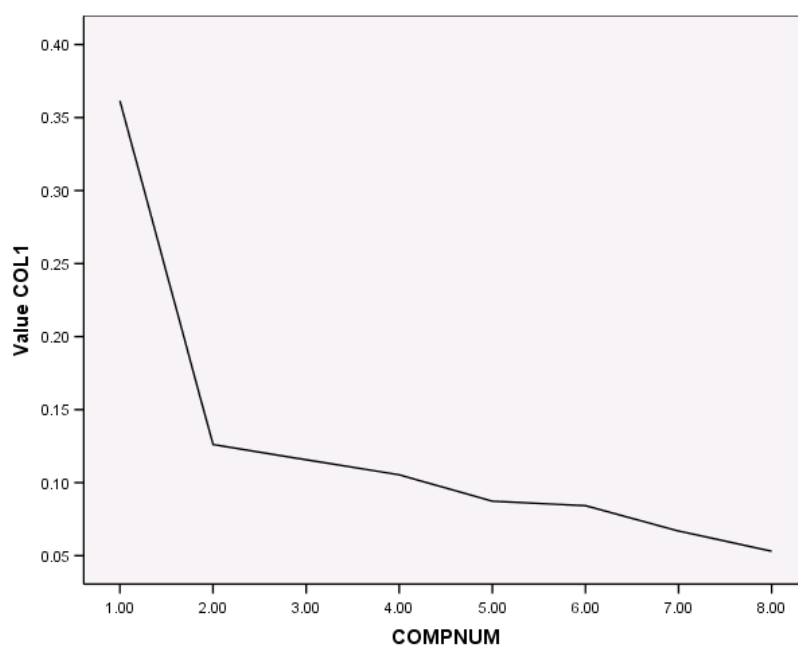
	PCR1	PCR2	PCR3	PCR4	PCR5	PCR6	PCR7	PCR8
SpendSer	-.410	-.172	-.190	.288	.298	.055	-.768	.032
DefSpend	-.312	.139	.455	-.587	-.131	.518	-.212	-.029
Insuranc	-.430	-.166	-.070	.128	.373	.326	.520	.501
JobsLivi	-.446	-.342	-.052	.081	-.097	.013	.290	-.762
GABlacks	-.385	-.278	.025	-.086	-.647	-.431	-.035	.401
EnvJobs	-.288	.370	-.412	-.555	.324	-.437	.057	-.055
FedGun	-.254	.370	.688	.327	.219	-.408	.059	-.042
WomRole	-.234	.675	-.326	.357	-.418	.280	.045	-.031

----- END MATRIX -----

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-dimensioned 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 the government should spend more money and 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.

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.

Run MATRIX procedure:

```
Correlation Matrix
      V1      V2      V3      V4      V5      V6
V1    1.000    .766    .781   -.494   -.488   -.494
V2    .766    1.000    .754   -.497   -.492   -.505
V3    .781    .754    1.000   -.484   -.482   -.478
V4   -.494   -.497   -.484    1.000    .888    .892
```

V5	-.488	-.492	-.482	.888	1.000	.901
V6	-.494	-.505	-.478	.892	.901	1.000

Principal Components Results Using the Correlation

Variance of Components (Using COR)

PCR1	4.138
PCR2	1.184
PCR3	.249
PCR4	.218
PCR5	.114
PCR6	.098

Proportion of Variance Components (Using COR)

C

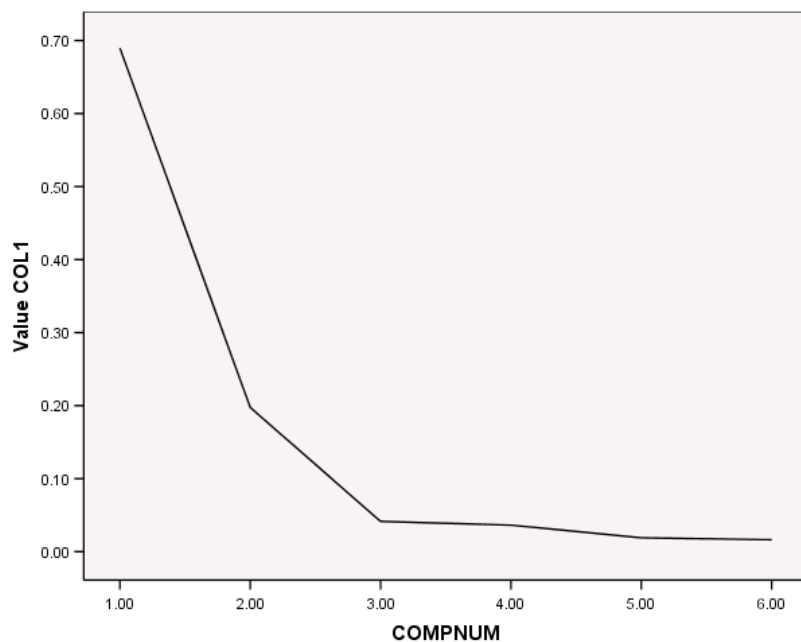
PCR1	.690
PCR2	.197
PCR3	.041
PCR4	.036
PCR5	.019
PCR6	.016

Loadings (Using COR)

	PCR1	PCR2	PCR3	PCR4	PCR5	PCR6
V1	-.392	.429	.224	-.782	-.010	-.007
V2	-.391	.410	-.800	.192	.011	-.039
V3	-.387	.437	.553	.593	-.008	.041
V4	.425	.386	-.029	-.001	-.806	.139
V5	.425	.394	-.033	-.018	.522	.625
V6	.427	.390	.036	.014	.278	-.766

----- END MATRIX -----

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 factors, 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.