Department of Political Science, University of Aarhus Tel.: 89 42 12 71
Universitetsparken, DK-8000 Århus C Fax.: 86 13 98 39
Søren Risbjerg Thomsen E-mail:srt@ps.au.dk

August 2005

Models for Multi-Level Voting Behaviour

Exercises, Session 6 – Factor analysis of aggregate party shares

In this session we will analyze aggregate geographical data instead of individual level data. First, we will derive an aggregate level model from the individual level model we developed in the previous exercises. This is kind of controversial within the field of electoral ecology (the research field that work with aggregate electoral data) since we need some crude assumptions and approximations to derive explicit solutions. An alternative would be to simulate aggregate consequences of the individual level model, but then we would not have an explicit model for the aggregate level. However, since the explicit model works very well and give plausible results we will use it as a simple approach (the more advanced simulation approach is outside the scope of this course).

To resume the individual level theory we used the following utility model

$$P_{ij} = \frac{e^{U_{ij}}}{e^{U_{i1}} + \cdots + e^{U_{im}}} \tag{6.1}$$

where p_{ij} is the probability that voter no. i chooses party no. j out of all m parties. U_{ij} is the utility of party j for voter i. In Session 5 we found it useful to assume that the individual level utility for a certain party is a function of three aspects: the general sympathy with the party, the identification of the voter with the party, and the relation between the party's and the voter's position in an issue-space. Since the directional model gave the most plausible estimate we will use this model in the following derivations. The equation for this model is

$$U_{ij} = a_j + bB_{ij} + c_1 Y_{1j} X_{1i} + \cdots + c_K Y_{Kj} X_{Ki} \tag{6.2}$$

where a_j is the general sympathy, B_{ij} is the identification of voter i with party j, Y_{1j} ... Y_{Kj} is the position of the party in the K-dimensional issue-space, and X_{1i} ... X_{Ki} is the position of the voter in the same issue-space. b is a coefficient for the importance of party identification and c_1 ... c_K is the importance of each of the K dimensions in the issue space.

In the situation where we do not know the position of the parties in the issue space this position can be estimated combined with the importance of each dimension with the following substitutions

$$c_{1j} = c_1 Y_{1j} \; ; \; \ldots \; ; c_{Kj} = c_K Y_{Kj} \, , \tag{6.3}$$

and equation (6.2) becomes

$$U_{ij} = a_j + bB_{ij} + c_{1j} X_{1i} + \cdots + c_{Kj} X_{Ki} \; . \tag{6.4}$$

We now need some crude approximations to derive the aggregate level model for the support of each party (including the ”party” of abstention) in a number of geographical districts with in a certain geographical region nr. *r*. We use the common observation that a party usually have special strongholds within certain geographical regions of the country. Apart from that we assume that the party identification is also a party-specific function of the position of the voter in the issue space. Thus we assume the following model for party-identification:

$$B_{ij}^{r} = b_{oj}^{r} + b_{1j}X_{1i} + \cdots + b_{Kj}X_{Ki} \ , \tag{6.5}$$

where b_{oj}^{r} is a parameter for the regional strength of the party and b_{1j}, ..., b_{Kj} is the position of the party identification in the party space. Inserting equation(6.5) into equation(6.4) we get

$$U_{ij} = (a_j + b_{oj}^{r}) + (b_{1j} + c_{1j})X_{1i} + \cdots + (b_{Kj} + c_{Kj})X_{Ki} \, . \tag{6.6}$$

Since we cannot separate the two constants in each parenthesis in equation(6.6), we use the following substitutions

$$a_{oj}^{r} = a_j + b_{oj}^{r}; \quad a_{1j} = b_{1j} + c_{1j}; \quad \ldots \quad a_{Kj} = b_{Kj} + c_{Kj} \tag{6.7}$$

and equation(6.6) can now be written

$$U_{ij} = a_{oj}^{r} + a_{1j}X_{1i} + \cdots + a_{Kj}X_{Ki} \, . \tag{6.8}$$

To begin with we will drop the region-specific index *r* from the party constant a_{oj}^{r} and just assume that it has about the same value a_{oj} in the whole country. Later, in session 7, we will investigate if it might differ between regions. For now we will just write equation(6.8) as the familiar multinomial logit model

$$U_{ij} = a_{oj} + a_{1j}X_{1i} + \cdots + a_{Kj}X_{Ki} \, . \tag{6.9}$$

We will consider the case where we have no access to individual level data and only have aggregate date about the voting behaviour in several districts. In this situation we just assume that the individual utility with respect to party *j* is the following function of a latent vector $\boldsymbol{\theta}$, (i.e., a set of unknown latent variables θ_1, θ_2, ... , θ_K, representing the different issue-dimensions)

$$U_j(\boldsymbol{\theta}) = \alpha_{oj} + \alpha_{1j}\theta_1 + \cdots + \alpha_{Kj}\theta_K \, , \tag{6.10}$$

and the individual level probability is

$$P_j(\boldsymbol{\theta}) = \frac{e^{U_j(\boldsymbol{\theta})}}{e^{U_1(\boldsymbol{\theta})} + \cdots + e^{U_m(\boldsymbol{\theta})}} \, . \tag{6.11}$$

The aggregate share choosing party no. *j* in district no. *g* depends on the populations density $\varphi_g(\boldsymbol{\theta})$ of $\boldsymbol{\theta}$ within district *g*. We assume that it is *K*-dimensional normal distributed with a vector $\boldsymbol{\mu}_g$ of mean values and a variance-covariance matrix $\boldsymbol{\sigma}$:

$$\boldsymbol{\theta} \sim N_K(\boldsymbol{\mu}_g, \boldsymbol{\sigma}). \tag{6.12}$$

This means that we assume that each of the latent variables is normal distributed within each district. Thus, the proportion p_{gj} voting for party no. *j* in district no. *g* is derived by integrating (11) over this distribution:

$$p_{gj} = \int_{\boldsymbol{\theta}} P_j(\boldsymbol{\theta}) \varphi_g(\boldsymbol{\theta}) d\boldsymbol{\theta} \tag{6.13}$$

By raising assumptions about $\boldsymbol{\sigma}$**,** one can estimate the unknown individual level parameters from the aggregate data. Unfortunately, with more than two parties there is no simple solution to (13) even with very simplified assumptions about $\boldsymbol{\sigma}$. So, in the multiparty case one must either use simulation methods, or one must somehow reduce the complexity of the multiparti model.

Crude binary choice

It turns out that if we consider the crude binary choice between a party *j* and all the rest of the parties (including abstention) a simple aproximate relation holds between logit transformed party shares and the mean values μ_1 , μ_2 , ... , μ_K , of the latent variables (Thomsen 2000, pp. 8-10). The logit share of the proportion p_{gj} of voters choosing party *j* within district *g* is

$$x_{gj} = \ln\left(\frac{p_{gj}}{1-p_{gj}}\right) \tag{6.14}$$

and the approximate relation between the logit share and the latent mean values is[1]

$$x_{gj} = a_{oj} + a_{1j}\mu_{g1} + a_{2j}\mu_{g2} + \cdots + a_{Kj}\mu_{gK} . \tag{6.15}$$

Thus, logit shares of the parties can be assumed a linear function of a set of latent variables representing various (unknown) issue-dimensions. A method for estimating the coefficients in this model is Factor Analysis. Factor Analysis is usually done with standardized *x*-variables so that a_{oj} is equal to 0 and the other *a*-coefficients are between -1 and +1. In our case the logit shares are not standardized so we are interested in finding an unstandardized or "raw" solution[2].

Preparing data (Stata commands in DKdis01.do)

[1] The a-coefficients are proportional to the α-coefficients in equation (6.10), see Thomsen (2000, pp. 10-11).

[2] In the raw solution each latent μ-variable is still standardised, but the *a*-coefficients are not limited between -1 and +1. The Stata program **pca** can only find the raw principal component solution but a special Stata program **obliraw**, developed by the author can also find an oblique solution.

DKdistricts.dta is a Stata file with a lot of information about the 103 Danish nomination districts, including variables about socio-economic structure and election results 1979-2001. The variables are described in the file DKdistricts.doc. The socio-economic variables cover citizenship, country origin, education, income, and occupation. The variables about election results cover European Parliament (EP) elections, EU referendums (ER), and Danish Parliament (DP) elections.

Most variables report the number of voters belonging to a certain socio-economic category or voting for at certain party or political option (or spoiling the vote, or abstaining). The only exception, where a variable is not indicating number of voters, is income reported as Gross income per household. [3]

To get familiar with the data we will first do a few simple tasks such as inspecting the voting data from the Danish Parliament election of 1998, and compute percentages from absolute numbers. In Stata we open the file DKdistricts.dta and issue the command

```
sum dpoe98-votdp98
```

to inspect summary statistics for the aggregate voting behavior at the DP election of 1998. The results are listed in Table 6.1. [Change to same party abbreviations as in survey analysis].

Table 6.1 Inspection of the aggregate voting behavior at DP 1998

Variable	Obs	Mean	Std. Dev.	Min	Max
dpoe98	103	892.5534	574.8467	137	2409
dpf98	103	2499.087	1231.011	657	6737
dpu98	103	104.5437	38.36966	37	202
dpa98	103	11879.81	5185.938	3310	30351
dpb98	103	1274.311	583.88	285	3117
dpd98	103	1425.262	769.8384	329	4167
dpq98	103	831.6117	638.8942	171	3841
dpc98	103	2951.117	2018.123	358	13366
dpv98	103	7940.718	3948.116	1549	20013
dpo98	103	2450.767	1332.618	590	8762
dpz98	103	800.3592	941.8161	102	4622
xdp98	70	26.18571	43.99351	2	262
spldp98	103	251.7379	89.63094	98	619
absdp98	103	5448.282	1903.381	2032	12244
votdp98	103	38767.95	15270.65	15161	89606

Notice that the 103 observations are the 103 Danish nomination districts, and for example that the first party oe (United List) on average across districts got 892.5 votes.[4] Also notice that no category had zero votes in any district (except independent candidates, xdp98, but they only ran in 70 districts, and outside these districts the number of votes is set to missing (.)).

[3] At DP elections only people with Danish citizenship can vote. At EP elections foreign EU citizens can choose to vote in Denmark.

[4] In the variable names dp indicates Danish Parliament election, and 98 indicate 1998. The variable label for votes for the Unite List is "DP98: Red-Green Coalition", where Red-Green Coalition is the more official name of the United List. In Denmark, each party has a special "list letter" and the list letter for the United List is the Danish letter Ø, here written as oe. In the previous exercises we instead used two letters to identify the parties. The list letters of each party in a rough left-right order (with the previous two letter identification in parentheses) is oe (ul), f (sp), u (this party, Democratic Renewal, was omitted in the previous exercises because they got very few votes and did not run again), a (sd), b (sl), d (cd), q (cp), c (co), v (li), o (dp), z (pp). x indicates (the very few) votes for independent candidates. Finally, spl indicates spoilt votes (blank and other invalid votes), abs indicates abstainers (voters eligible to vote, but did not vote), and vot indicates all voters eligible to vote.

In the following we will always compute party shares in percent of all voters, so that independent candidates (x), spoilt votes (spl) and abstainers (abs) are always included in the "party distribution". This is because we want to consider all kinds of voting behavior including non-voting, and also because non-party choice is involved in voter transitions between elections.

The share in percent of all voters voting for the United List could be computed by the command

```
gen pdpoe98 = dpoe98/votdp98*100
```

but it is more convenient to compute all party shares in a single loop across all parties. Table 6.2 shows loop and the previous commands.

Table 6.2 Stata commands for computing percentages

```
* dkdis01.do - Exercises, Set 6

use "C:\Data\DKdistricts.dta", clear

* Inspect DP98
sum dpoe98-votdp98

* Compute percent of all voters
* for a single party
gen pdpoe98 = dpoe98/votdp98*100
drop pdpoe98

* for all parties
foreach v of varlist dpoe98 dpf98 dpu98 dpa98 dpb98 dpd98 dpq98 dpc98 /*
  */ dpv98 dpo98 dpz98 xdp98 spldp98 absdp98 {
gen p`v' = `v'/votdp98*100
  }
sum pdpoe98-pabsdp98
```

Since gen is only used to create new variables pdpoe98 is dropped before the percentage loop. Table 6.2 present the summary statistics for the percentages,

Table 6.2 Summary statistics for percentages

```
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
     pdpoe98 |       103     2.60524    2.249272   .6074514   11.72495
      pdpf98 |       103    6.613032    2.560778   2.443833   14.80556
      pdpu98 |       103    .2821746    .0871124   .1648805   .6435055
      pdpa98 |       103    30.33298    5.052491   12.77203   40.50758
      pdpb98 |       103    3.455751     1.51376   1.710992   8.497404
-------------+--------------------------------------------------------
      pdpd98 |       103    3.567684    .8279749   1.874644   6.587191
      pdpq98 |       103    2.143483    1.386776   .8374963   10.36987
      pdpc98 |       103    7.367969    3.150126   2.039886   20.45069
      pdpv98 |       103    20.21613    6.417162   8.242727   34.35627
      pdpo98 |       103    6.220474    1.654031   2.875793   10.20691
-------------+--------------------------------------------------------
      pdpz98 |       103    2.069491    2.314561   .6727788   12.04352
      pxdp98 |        70    .0545952    .0700412   .0050642   .2923911
    pspldp98 |       103    .6716993    .1287033    .374677   .9910563
    pabsdp98 |       103    14.41678    2.336659   9.505466   22.40156
```

Notice that the mean percentage voting for a party is not necessarily the same as the percent in the whole country voting for the party unless all districts have equal number of voters. To

compute country-wide percentages one must weight the percentages with the number of voters ("the unit size") in each district by the command

```
sum pdpoe98-pabsdp98 [aw=votdp98]
```

With "analytical weights" indicated by aw the weights are proportional to the number of voters (but they only add up to the number of units in tests of significance). The weighted summary statistics are shown in Table 6.3.

Table 6.3 Summary statistics for percentages weighted by voters in each district

```
. sum pdpoe98-pabsdp98 [aw=votdp98]

    Variable |     Obs      Weight        Mean   Std. Dev.       Min        Max
-------------+--------------------------------------------------------------------
     pdpoe98 |     103     3993099    2.302297    1.78221   .6074514   11.72495
      pdpf98 |     103     3993099    6.446271      2.204   2.443833   14.80556
      pdpu98 |     103     3993099    .2696652   .0772657   .1648805   .6435055
      pdpa98 |     103     3993099    30.64337    4.87563   12.77203   40.50758
      pdpb98 |     103     3993099    3.287021    1.27375   1.710992   8.497404
-------------+--------------------------------------------------------------------
      pdpd98 |     103     3993099    3.676393     .82758   1.874644   6.587191
      pdpq98 |     103     3993099    2.145101   1.396594   .8374963   10.36987
      pdpc98 |     103     3993099    7.612258     3.1214   2.039886   20.45069
      pdpv98 |     103     3993099    20.48269   5.908321   8.242727   34.35627
      pdpo98 |     103     3993099    6.321631   1.702716   2.875793   10.20691
-------------+--------------------------------------------------------------------
      pdpz98 |     103     3993099    2.064487   2.303205   .6727788   12.04352
      pxdp98 |      70     2964548    .0618307   .0794869   .0050642   .2923911
    pspldp98 |     103     3993099    .6493453    .122411    .374677   .9910563
    pabsdp98 |     103     3993099    14.05357   2.110486   9.505466   22.40156
```

The mean percent for at party is now showing the percent in the whole country voting for the party (except that the mean percent of pxdp98 shows the percentage voting for independent candidates within the 70 districts with such candidates).

Computing logits

We are now ready to compute logit transformed percentages, also called logit shares, using equation(6.14). The logit share for United List is computed by the command

```
gen ldpoe98 = ln(pdpoe98/(100-pdpoe98))
```

A problem could appear if the percentage is equal to either 0 or 100, where the logit is not defined since one cannot take the logarithm to zero or divide by zero. We know from table 6.3 that this is not the case here but it could easily happen for smaller units such as precincts. In this situation the standard procedure is as follows. In case the percentage is 0 it is instead assumed that just a "half voter" voted for the party, and in case the percentage is 100 it is instead assumed that all voters expect the half voter voted for the party. The set of stata commands to deal with the situation is shown in Table 6.4

Table 6.4 Commands for computing logit in case the percent can be either 0 or 100

```
gen temp = pdpoe98
replace temp = 0.5/votdp98*100 if dpoe98 == 0
replace temp = (votdp98-0.5)/votdp98*100 if dpoe98 == votdp98
gen ldpoe98 = ln(temp/(100-temp))
drop temp
```

Fortunately, we know that this is not the case with the 1998 district results so we can just compute all the logits in a single loop with the commands in Table 6.5, also showing the command for weighted summary statistics of logits.

Table 6.5 Commands for computing logits in a single loop

```
* for all parties
foreach v of varlist dpoe98 dpf98 dpu98 dpa98 dpb98 dpd98 dpq98 dpc98 /*
  */ dpv98 dpo98 dpz98 xdp98 spldp98 absdp98 {
gen l`v' = ln(p`v'/(100-p`v'))
  }
sum ldpoe98-labsdp98 [aw=votdp98]
```

Table 6.6 shows that all logits are negative since no percentage is higher than 50. An important characteristics of logits is that the standard deviation is more similar between parties than is the case with percentages (compare with Table 6.3).

Table 6.6 Summary statistics for logits weighted by voters in each district

Variable	Obs	Weight	Mean	Std. Dev.	Min	Max
ldpoe98	103	3993099	-3.951925	.6270306	-5.09756	-2.018739
ldpf98	103	3993099	-2.733283	.3723625	-3.686861	-1.749933
ldpu98	103	3993099	-5.945266	.243788	-6.406054	-5.039539
ldpa98	103	3993099	-.8286939	.2443441	-1.921267	-.38436
ldpb98	103	3993099	-3.442811	.3513022	-4.050838	-2.376607
ldpd98	103	3993099	-3.288471	.2189524	-3.957827	-2.651901
ldpq98	103	3993099	-3.952597	.4884855	-4.774098	-2.156787
ldpc98	103	3993099	-2.560236	.3781596	-3.871666	-1.358361
ldpv98	103	3993099	-1.398586	.3841568	-2.409816	-.6474577
ldpo98	103	3993099	-2.73519	.3095327	-3.519662	-2.174443
ldpz98	103	3993099	-4.156305	.6695599	-4.994758	-1.988316
lxdp98	70	2964548	-7.878849	.8806536	-9.890681	-5.831905
lspldp98	103	3993099	-5.048105	.190885	-5.583107	-4.604194
labsdp98	103	3993099	-1.820934	.1676212	-2.253422	-1.242417

Percentages and logits are also computed for the DP election of 2001 with the commands in Table 6.7.

Table 6.7 Commands for percentages and logits for DP 2001

```
* Inspect DP01
sum dpoe01-votdp01

* Compute percentages for all parties
foreach v of varlist dpoe01 dpf01 dpa01 dpb01 dpd01 dpq01 dpc01 /*
  */ dpv01 dpo01 dpz01 xdp01 spldp01 absdp01 {
gen p`v' = `v'/votdp01*100
  }
sum pdpoe01-pabsdp01 [aw=votdp01]

* Compute logits for all parties
foreach v of varlist dpoe01 dpf01 dpa01 dpb01 dpd01 dpq01 dpc01 /*
  */ dpv01 dpo01 dpz01 xdp01 spldp01 absdp01 {
gen l`v' = ln(p`v'/(100-p`v'))
  }
sum ldpoe01-labsdp01 [aw=votdp01]
```

Factor analysis

As mentioned above we expect the logit shares to be linear related to a set of latent variables representing various (unknown) issue-dimensions, i.e.

$$x_{gj} = a_{oj} + a_{1j}\mu_{g1} + a_{2j}\mu_{g2} + \cdots + a_{Kj}\mu_{gK}\,. \qquad \text{(6.15, repeated)}$$

This can be investigated with unstandardized (raw) factor analysis. The factor analysis will be more interesting if we include more than one election in the analysis. This is because we assume that the district positions on the different latent variables are relatively stable and thus can be considered constant from one election to the next, while the coefficients characterizing the parties might change over time because of change in party policy. At the next election (2001) we expect the following relation between the logit share y_{gj} and the latent variables:

$$y_{gj} = b_{oj} + b_{1j}\mu_{g1} + b_{2j}\mu_{g2} + \cdots + b_{Kj}\mu_{gK}\,. \qquad \text{(6.16)}$$

Notice when comparing equation(6.15) and (6.16) that the latent variables μ_{g1}, μ_{g2}, ... , μ_{gK}, only varies between districts but are assumed constant in time, while the party coefficient can be different between elections. A special Stata program, **obliraw**, estimates the coefficients in both equations under different assumptions about the factor structure. An example of the command for doing the analysis with logit shares from both elections is

```
obliraw ldpoe98-ldpz98 labsdp98 ldpoe01-ldpz01 labsdp01 [aw=votdp01], factors(2)
```

In this example we exclude votes for independent candidates and spoilt votes from both elections, because these categories are of little importance. Further, we weight with the number of voters in 2001. Finally, we only ask for a solution with two factors. Stata produces several tables in the Result window. The first table appears in Table 6.8, showing univariate summary statistics.

Table 6.8 Univariate summary statistics produced by **obliraw**

Variable	Mean	Std. Dev.	Min	Max
ldpoe98	-3.949131	.6273948	-5.09756	-2.018739
ldpf98	-2.731492	.3720456	-3.686861	-1.749933
ldpu98	-5.945389	.2432846	-6.406054	-5.039539
ldpa98	-.8287704	.2433002	-1.921267	-.38436
ldpb98	-3.441834	.3509648	-4.050838	-2.376607
ldpd98	-3.287699	.2185839	-3.957827	-2.651901
ldpq98	-3.954111	.487799	-4.774098	-2.156787
ldpc98	-2.55945	.3774619	-3.871666	-1.358361
ldpv98	-1.399192	.3842703	-2.409816	-.6474577
ldpo98	-2.73461	.3092678	-3.519662	-2.174443
ldpz98	-4.158262	.6677166	-4.994758	-1.988316
labsdp98	-1.821459	.168088	-2.253422	-1.242417
ldpoe01	-4.047763	.6187751	-5.315038	-2.253395
ldpf01	-2.911328	.3924525	-3.840185	-1.893499
ldpa01	-1.106365	.2251146	-2.072562	-.584183
ldpb01	-3.151438	.4311961	-4.177647	-1.847891
ldpd01	-4.211878	.3027488	-4.959069	-3.484507
ldpq01	-4.005219	.4250153	-4.836054	-2.27696
ldpc01	-2.521704	.3550426	-3.509604	-1.525869
ldpv01	-1.026787	.361188	-1.985159	-.3592598
ldpo01	-2.179563	.2351524	-3.019143	-1.734196
ldpz01	-5.440397	.4322493	-6.262027	-2.932924
labsdp01	-1.9292	.2026838	-2.433905	-1.237175

Since the latent variables are assumed to be standardized the mean values estimates the constants a_{oj} and b_{oj} for each party at the two elections. The interpretation of these constants is similar to the party intercepts in the individual level issue models, i.e. they measure the relative general sympathy for the parties. For example, the general sympathy of the United List declined slightly from -3.95 in 1998 to -4.05 in 2001. The next table with eigenvalues (Table 6.9) indicates how many latent variables (components) are necessary to describe the data.

Table 6.9 Eigenvalues with principal components solution

(principal components; 2 components retained)

Component	Eigenvalue	Difference	Proportion	Cumulative
1	1.86859	1.37526	0.5332	0.5332
2	0.49333	0.11908	0.1408	0.6739
3	0.37425	0.10753	0.1068	0.7807
4	0.26672	0.14536	0.0761	0.8568
5	0.12137	0.02742	0.0346	0.8914
6	0.09395	0.03150	0.0268	0.9182
7	0.06245	0.01180	0.0178	0.9361
8	0.05065	0.00723	0.0145	0.9505
9	0.04342	0.00668	0.0124	0.9629
10	0.03674	0.00725	0.0105	0.9734
11	0.02950	0.00815	0.0084	0.9818
12	0.02134	0.00751	0.0061	0.9879
13	0.01383	0.00616	0.0039	0.9918
14	0.00767	0.00187	0.0022	0.9940
15	0.00580	0.00168	0.0017	0.9957
16	0.00413	0.00072	0.0012	0.9969
17	0.00341	0.00086	0.0010	0.9978
18	0.00255	0.00082	0.0007	0.9986
19	0.00174	0.00044	0.0005	0.9991
20	0.00130	0.00034	0.0004	0.9994
21	0.00096	0.00013	0.0003	0.9997
22	0.00083	0.00059	0.0002	0.9999
23	0.00024	.	0.0001	1.0000

The eigenvalues for the different components have no simple interpretation, but the ratio of an eigenvalue in relation to the sum of all eigenvalues indicates the proportion of variation in the data explained by the component. Thus, the first component explains 53.3 percent of all varia-

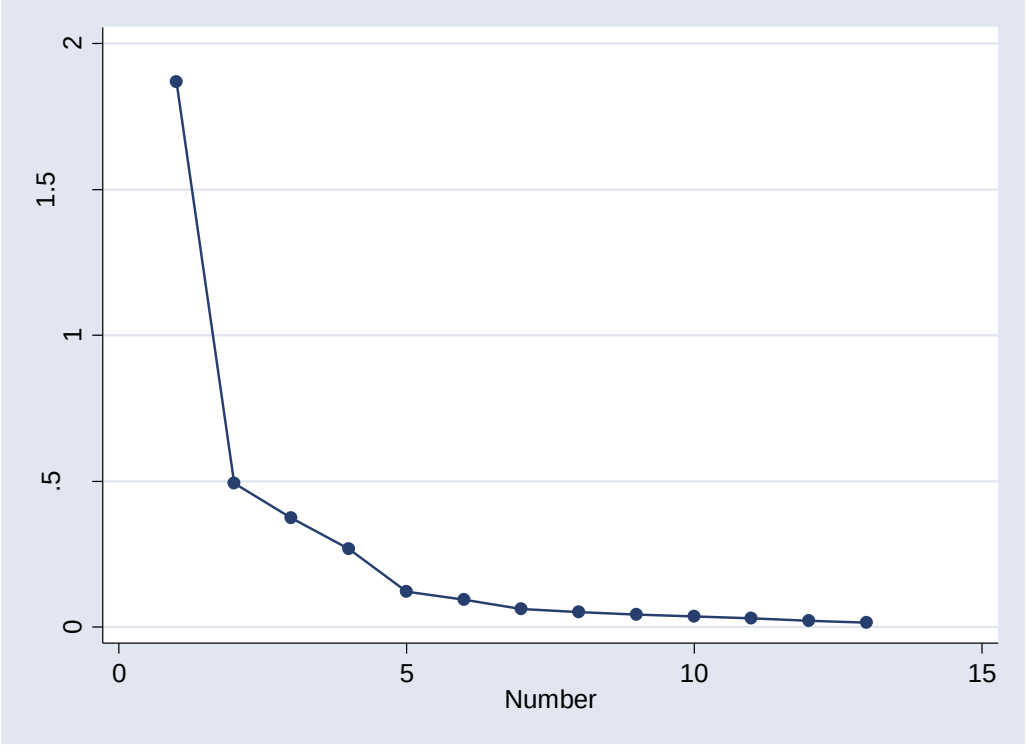
tion; the second component explains 14.1 percent etc.[5] The principal component solution means that the first component explains as much as possible that could be explained with only a single component. The second component explains as much as possible of the variation that is not already explained with the first component, etc. Thus, every next component always explains less than the previous one. The last cumulative column in Table 6.9 adds the contribution of each component and for example it shows that one can explain 89 percent of all variation with 5 components.

The big question is how many components one should pick for the final model. It is tempting to include a fairly large number of components to explain a large share of the variation in the data. However, the problem is that it can be difficult to interpret a large number of components. One approach for deciding the number of components is to consider the additional contribution from every new factor by inspection of the so called "Scree plot" of declining importance of the components. This plot is produced in Stata[6] by the command

```
greigen
```

and the Scree plot is presented in Figure 6.1.

Figure 6.1 Scree plot from the raw factor analysis



The scree plot shows that the first component is much more important than the rest of the components. The pattern is similar to the heap of stones (the scree) beneath a steep mountain side and one criterion is to ignore all components from the scree because the additional expla-

[5] Notice that the different components with raw factor analysis describe the unstandardized (raw) variation in the data, while the components in standardized factor analysis describe the standardized variation. In practice, raw factor analysis estimate the components from a matrix of covariances between the variables, while standardized factor analysis uses the correlation matrix.
[6] Notice that graphics in Stata can be very time consuming, so be patient.

nation of each next component is quite marginal. With this criterion one should only pick one component. However, in my experience it is often useful to include the first component on the scree and thus in this case I pick two components for the final solution.

The next interesting table of results from the obliraw run is not the eigenvectors but the table of coefficients to the raw principal components under the heading "Raw components", shown in Table 6.10. These are the *a* and *b* coefficients with respect to the latent variables from equation(6.15) and (16). The first task is to find an interpretation of the two factors from the values of the coefficients. This is difficult without knowledge of the political geography of Denmark, but the first component is clearly connected to urban-rural contrast with relatively high positive coefficients for the (urban) extreme left parties oe and f (United List and Socialist People's Party) at both elections and relatively high negative coefficients for q and v (the Christian People's Party and the agrarian Liberal Party). Sometimes, it can be difficult to interpret other components than the first component because all components are forced to be uncorrelated (orthogonal) and the rest of the components are only explaining the residual variation after the first component has explained as much as possible. However, in this case it is quite clear that the second component is related to the working class – middle class contrast with relatively high positive coefficients for a and abs (the Social Democrats and abstainers) and relatively high negative coefficients for c (Conservatives).

Table 6.10 Coefficients to the raw principal components

Raw Components

	F1	F2
ldpoe98	0.6011	0.0903
ldpf98	0.3446	0.0087
ldpu98	-0.0087	0.0480
ldpa98	0.0065	0.1592
ldpb98	0.2422	-0.0956
ldpd98	0.0500	-0.1458
ldpq98	-0.3848	-0.0795
ldpc98	0.1223	-0.2510
ldpv98	-0.3049	-0.1809
ldpo98	0.0259	-0.0985
ldpz98	-0.3789	0.4037
labsdp98	0.0454	0.1023
ldpoe01	0.5804	0.1500
ldpf01	0.3652	0.0247
ldpa01	0.0014	0.1423
ldpb01	0.3467	-0.0949
ldpd01	0.2175	-0.0788
ldpq01	-0.2910	-0.0510
ldpc01	0.0670	-0.1456
ldpv01	-0.2818	-0.1681
ldpo01	-0.0774	0.0336
ldpz01	-0.3290	0.1114
labsdp01	0.0665	0.1330

To ease the interpretation of the two components one can also look at the rescaled coefficients to the raw components in Table 6.11. These coefficients are similar to the coefficients in standardized factor analysis, since they are rescaled as if the logit shares of the parties were standardised.[7]

[7] These are not the same coefficients as those obtained with standardized factor analysis, because they are derived from a solution based on the covariance matrix.

Table 6.11 Standardized coefficients to the principal components

Rescaled Components

	F1	F2
ldpoe98	0.9582	0.1439
ldpf98	0.9262	0.0234
ldpu98	-0.0357	0.1973
ldpa98	0.0266	0.6544
ldpb98	0.6901	-0.2725
ldpd98	0.2289	-0.6668
ldpq98	-0.7889	-0.1631
ldpc98	0.3239	-0.6650
ldpv98	-0.7934	-0.4707
ldpo98	0.0838	-0.3185
ldpz98	-0.5674	0.6046
labsdp98	0.2703	0.6089
ldpoe01	0.9380	0.2425
ldpf01	0.9306	0.0628
ldpa01	0.0063	0.6322
ldpb01	0.8039	-0.2200
ldpd01	0.7186	-0.2602
ldpq01	-0.6846	-0.1199
ldpc01	0.1888	-0.4101
ldpv01	-0.7802	-0.4655
ldpo01	-0.3290	0.1428
ldpz01	-0.7611	0.2576
labsdp01	0.3281	0.6562

The attractive feature of these coefficients is that they in principle can vary from -1 to +1 and thus are more convenient for interpreting the components. One just has to look for coefficients close to either -1 og +1 to identify the components. On the first component the standardized coefficents (also called loadings) are very high and positive for the two left wing parties and also quite high negative coefficients for the two rural parties q and v. On the second component one still find quite high positive loading for a and abs and quite high negative loadings for c, although the extreme loadings on the second component are not as impressive as the extreme loadings on the first component.

Although the interpretation of the components is easier with the standardized coefficients in Table 6.11 the main problem with these coefficients is that they are not as easy to understand from the policy of the parties. Thus, the more extreme left wing party oe (United List) has about the same standardized coefficient has the less extreme left wing party f (Socialist People's Party), precisely because of the standardization. If one instead looks at the unstandardized coefficients in Table 6.10 it appears that oe has nearly twice the value than f, because of higher geographical variance in the logit share of oe compared to f. Thus, the unstandardized coefficients in Table 6.10 can better reflect the policy of the parties, although the standardized coefficients are more useful for interpreting the components.

A final problem is that the principal component solution as mentioned is more difficult to interpret for other components than the first one. For this reason it is customary to rotate the original solution to ease the interpretation of the selected number of components. obliraw seeks a solution where each party, if possible, have high absolute loading on only one component and relatively low absolute loading on the other components. Further with the "oblimin" solution it allows the components to be correlated. To get the clearest picture the rotation is done on the standardized loadings in Table 6.11 obtaining the standardized loadings of the rotated solution in Table 6.12. Finally, this solution is transformed to the raw rotated solution in Table 6.13.

Table 6.12 Standardized coefficients after rotation

```
Factor loadings after rotation

              c1      c2
 ldpoe98   0.967  -0.020
  ldpf98   0.907  -0.134
  ldpu98   0.012   0.201
  ldpa98   0.182   0.642
  ldpb98   0.607  -0.386
  ldpd98   0.064  -0.698
  ldpq98  -0.807  -0.027
  ldpc98   0.157  -0.712
  ldpv98  -0.885  -0.331
  ldpo98   0.006  -0.329
  ldpz98  -0.408   0.694
labsdp98   0.408   0.556
 ldpoe01   0.971   0.081
  ldpf01   0.921  -0.096
  ldpa01   0.157   0.624
  ldpb01   0.730  -0.354
  ldpd01   0.638  -0.379
  ldpq01  -0.695  -0.002
  ldpc01   0.086  -0.437
  ldpv01  -0.871  -0.328
  ldpo01  -0.286   0.197
  ldpz01  -0.680   0.384
labsdp01   0.476   0.593
```

Table 6.13 Raw coefficients after rotation

```
Raw rotated Components

               F1       F2
 ldpoe98   0.6069  -0.0127
  ldpf98   0.3376  -0.0498
  ldpu98   0.0030   0.0489
  ldpa98   0.0442   0.1562
  ldpb98   0.2130  -0.1355
  ldpd98   0.0140  -0.1525
  ldpq98  -0.3937  -0.0134
  ldpc98   0.0592  -0.2687
  ldpv98  -0.3400  -0.1270
  ldpo98   0.0018  -0.1017
  ldpz98  -0.2727   0.4631
labsdp98   0.0686   0.0934
 ldpoe01   0.6009   0.0499
  ldpf01   0.3615  -0.0375
  ldpa01   0.0353   0.1404
  ldpb01   0.3149  -0.1525
  ldpd01   0.1931  -0.1147
  ldpq01  -0.2955  -0.0010
  ldpc01   0.0306  -0.1552
  ldpv01  -0.3144  -0.1184
  ldpo01  -0.0673   0.0463
  ldpz01  -0.2938   0.1658
labsdp01   0.0964   0.1201
```

In this example the difference between the original principal solution component solution in Table 6.10 and the final rotated solution in Table 6.13 is not very great, although the standardized coefficients are a little higher after (Table 6.11) than before rotation (Table 6.12).

One can further ease the interpretation of the components by including socio-economic variables in the factor analysis. We leave that for a problem.

Problems

Problem 1
Make the same factor analysis as in the exercise, but choose three components instead. Make interpretation of all three factors.

Problem 2
Make the same factor analysis as in the exercises, but include several socio-economic variables to ease the interpretation of the components.

Problem 3
Make a common factor analysis of party choice in all DP election since 1990.