Admin

- Assignment 1 due next Tuesday at 3pm in the Psychology course centre.
- Matrix Quiz during the first hour of next lecture.
- Assignment 2 due 13 May at 10am. I will upload and distribute these at the end of this lecture.

Small Group Presentations

This is the second of the small group presentations. These presentations are to take about three to five minutes and *no more than five minutes*.

The discussion of the topics will be general and be illustrated from the analysis of the T&F large sample example.

The topics you should cover are:

1. Overall statistical significance of the relationship. Number of statistically significant discriminant functions and importance of the discriminant functions.

2. Mean differences between the groups on the discriminant variables. Univariate F-ratios and F-to-REMOVE statistics.

3. Importance of variables: Standardised Discriminant function coefficients, Structure coefficients, Relative weights.

4. Centroid Plots; Pairwise F ratios; Classification Table.

You will present in your tutorials.

Tutors will arrange the schedule for the presentations.

T&F Complete Example

- Explores one 3-group categorical variable: WORKSTAT
    - **Group 1**: Women in paid jobs (WORKING)
    - **Group 2**: Happy housewives (HAPHOUSE)
    - **Group 3**: Unhappy housewives (UNHOUSE)
- How do these three groups of women differ in attitudes?
- Predictors are four discriminant variables:
    - **Variable 1**: Measure of control ideology - internal vs external (CONTROL)
    - **Variable 2**: Satisfaction with current marital status (ATTMAR)
    - **Variable 3**: Measure of conservative or liberal attitudes toward the role of women (ATTROLE)
    - **Variable 4**: Frequency of experiencing various favourable and unfavourable attitudes toward housework (ATTHOUSE)

$$Y_1, Y_2, \ldots Y_p \longleftarrow X$$

p continuous variables | categorical, *k* levels

Research Questions

- Is the overall relationship statistically significant and how strong is the relationship?
  - What is the number of significant discriminant functions?
- What variables are individually important in separating (discriminating) between the groups?

CONTROL ATTMAR ATTROLE ATTHOUSE	$\longleftarrow$	WORKSTAT
$p = 4$ continuous variables		categorical $k = 3$ levels

Assumptions of Discriminant Analysis

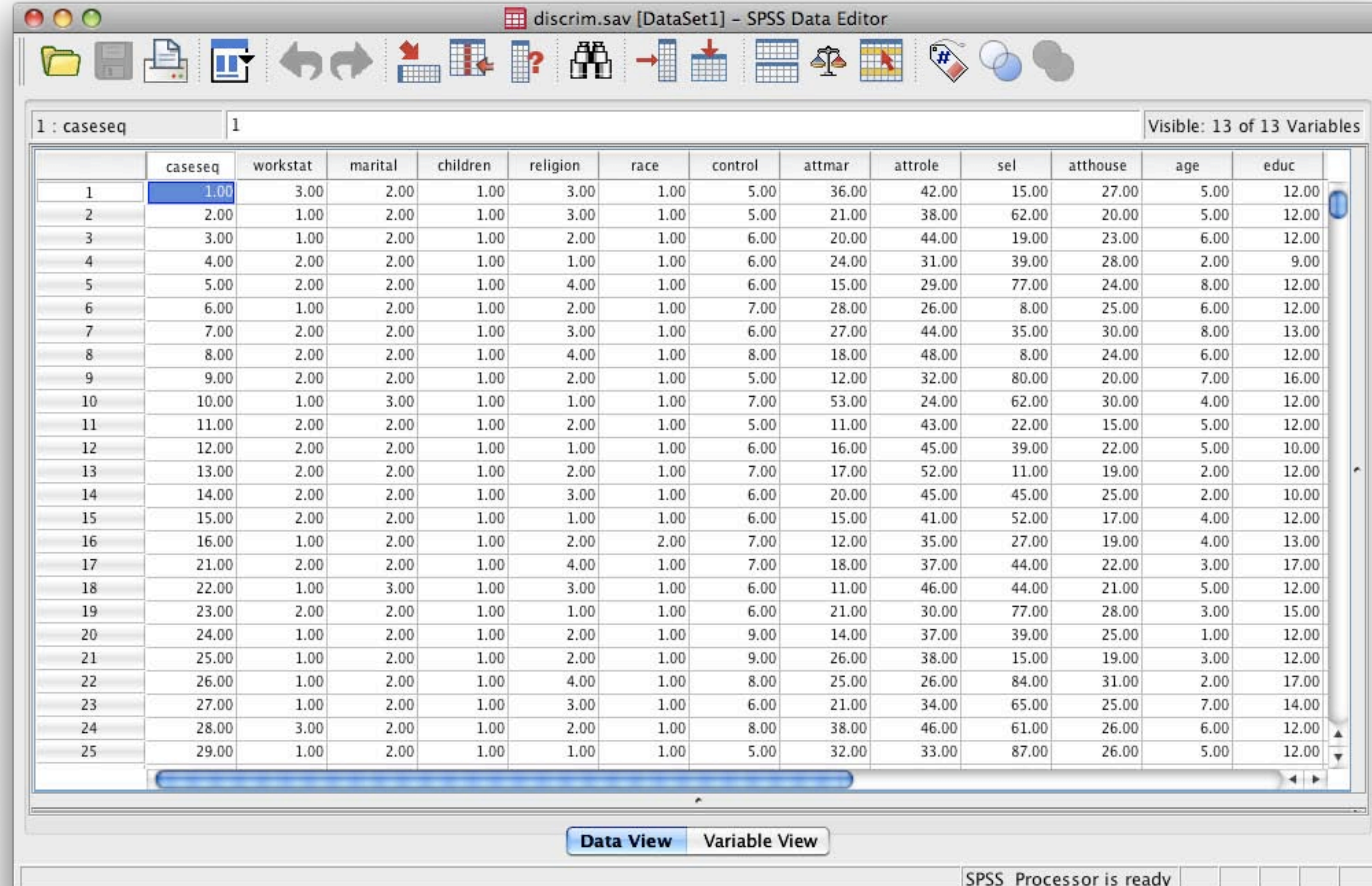
- True Categorical Grouping Variable
  - Discriminant Analysis assumes that the grouping variable is a true categorical variable. The groups must also be mutually exclusive.
- Sample sizes
  - It's acceptable to have unequal sample (group) sizes in Discriminant Analysis. With respect to sample sizes, there are 2 general rules of thumb:
    1. the sample size of the smallest group should exceed the number of predictors.
    2. the sample size of the smallest group should be at least 20 for 4 or more predictors.
- Homoscedasticity
  - Homoscedasticity is the assumption of homogeneity of variances of scores on the response variables within each group formed by the grouping variable. Each group should also have similar co-variances to the other groups for the response variables.

Assumptions of Discriminant Analysis

- Homoscedasticity (con't)
  - A violation of this assumption may indicate the presence of outliers in one or more groups. Discriminant Analysis is very sensitive to outliers. Box's M tests the assumption of homogeneity of variances/co-variances and a significant Box's M indicates that this assumption has been violated. Tabachnick and Fidell state that when sample sizes are large or equal, Discriminant Analysis is robust to the violation of this assumption.
- Outliers
  - Discriminant Analysis is very sensitive to both univariate and multivariate outliers. Data can be screened similar to the screening of data in Regression Diagnostics.
- Multicollinearity, Singularity, and Redundant Variables
  - Due to the need for matrix inversion in Discriminant Analysis, variables that are highly related (multicollinearity), perfectly related (singularity) or completely unrelated (redundant) need to be accounted for. Checking the Tolerance value of the response variables can check for the above.

SPSS commands for discriminant analysis

- We need to convince SPSS to yield ALL the information we need to address the research questions. e.g., F-To-Remove values.
- This means going beyond just the simple menu options in SPSS.
- Data Diagnostics - still important.
  - Strategy as per multiple regression.
  - Diagnostics done by groups.



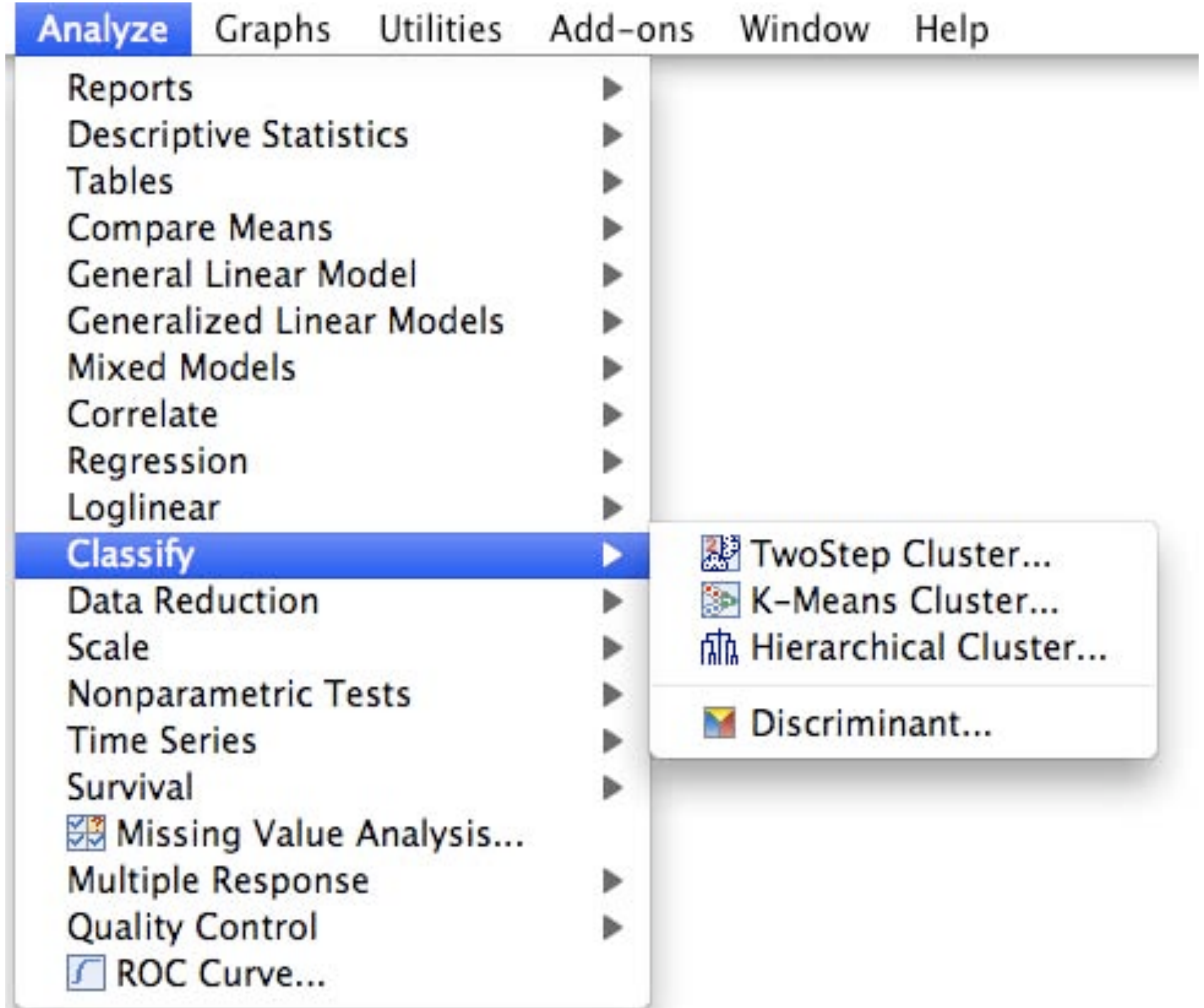
	caseseq	workstat	marital	children	religion	race	control	attmar	attrole	sel	atthouse	age	educ
1	1.00	3.00	2.00	1.00	3.00	1.00	5.00	36.00	42.00	15.00	27.00	5.00	12.00
2	2.00	1.00	2.00	1.00	3.00	1.00	5.00	21.00	38.00	62.00	20.00	5.00	12.00
3	3.00	1.00	2.00	1.00	2.00	1.00	6.00	20.00	44.00	19.00	23.00	6.00	12.00
4	4.00	2.00	2.00	1.00	1.00	1.00	6.00	24.00	31.00	39.00	28.00	2.00	9.00
5	5.00	2.00	2.00	1.00	4.00	1.00	6.00	15.00	29.00	77.00	24.00	8.00	12.00
6	6.00	1.00	2.00	1.00	2.00	1.00	7.00	28.00	26.00	8.00	25.00	6.00	12.00
7	7.00	2.00	2.00	1.00	3.00	1.00	6.00	27.00	44.00	35.00	30.00	8.00	13.00
8	8.00	2.00	2.00	1.00	4.00	1.00	8.00	18.00	48.00	8.00	24.00	6.00	12.00
9	9.00	2.00	2.00	1.00	2.00	1.00	5.00	12.00	32.00	80.00	20.00	7.00	16.00
10	10.00	1.00	3.00	1.00	1.00	1.00	7.00	53.00	24.00	62.00	30.00	4.00	12.00
11	11.00	2.00	2.00	1.00	2.00	1.00	5.00	11.00	43.00	22.00	15.00	5.00	12.00
12	12.00	2.00	2.00	1.00	1.00	1.00	6.00	16.00	45.00	39.00	22.00	5.00	10.00
13	13.00	2.00	2.00	1.00	2.00	1.00	7.00	17.00	52.00	11.00	19.00	2.00	12.00
14	14.00	2.00	2.00	1.00	3.00	1.00	6.00	20.00	45.00	45.00	25.00	2.00	10.00
15	15.00	2.00	2.00	1.00	1.00	1.00	6.00	15.00	41.00	52.00	17.00	4.00	12.00
16	16.00	1.00	2.00	1.00	2.00	2.00	7.00	12.00	35.00	27.00	19.00	4.00	13.00
17	21.00	2.00	2.00	1.00	4.00	1.00	7.00	18.00	37.00	44.00	22.00	3.00	17.00
18	22.00	1.00	3.00	1.00	3.00	1.00	6.00	11.00	46.00	44.00	21.00	5.00	12.00
19	23.00	2.00	2.00	1.00	1.00	1.00	6.00	21.00	30.00	77.00	28.00	3.00	15.00
20	24.00	1.00	2.00	1.00	2.00	1.00	9.00	14.00	37.00	39.00	25.00	1.00	12.00
21	25.00	1.00	2.00	1.00	2.00	1.00	9.00	26.00	38.00	15.00	19.00	3.00	12.00
22	26.00	1.00	2.00	1.00	4.00	1.00	8.00	25.00	26.00	84.00	31.00	2.00	17.00
23	27.00	1.00	2.00	1.00	3.00	1.00	6.00	21.00	34.00	65.00	25.00	7.00	14.00
24	28.00	3.00	2.00	1.00	2.00	1.00	8.00	38.00	46.00	61.00	26.00	6.00	12.00
25	29.00	1.00	2.00	1.00	1.00	1.00	5.00	32.00	33.00	87.00	26.00	5.00	12.00

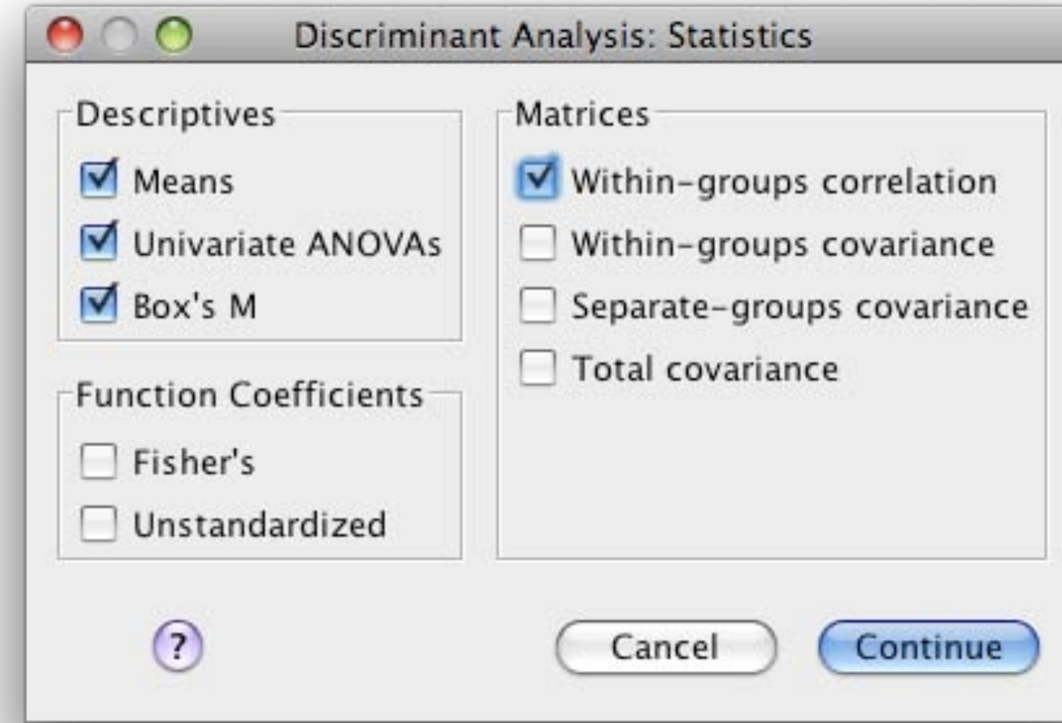
To be consistent with Tabachnick and Fidell’s reported analyses, we will analyse the data with the changes they recommend. Of course, the analysis should be run with all the data and a check whether the substantive interpretation changes, (i.e., regression diagnostics strategy and policy).

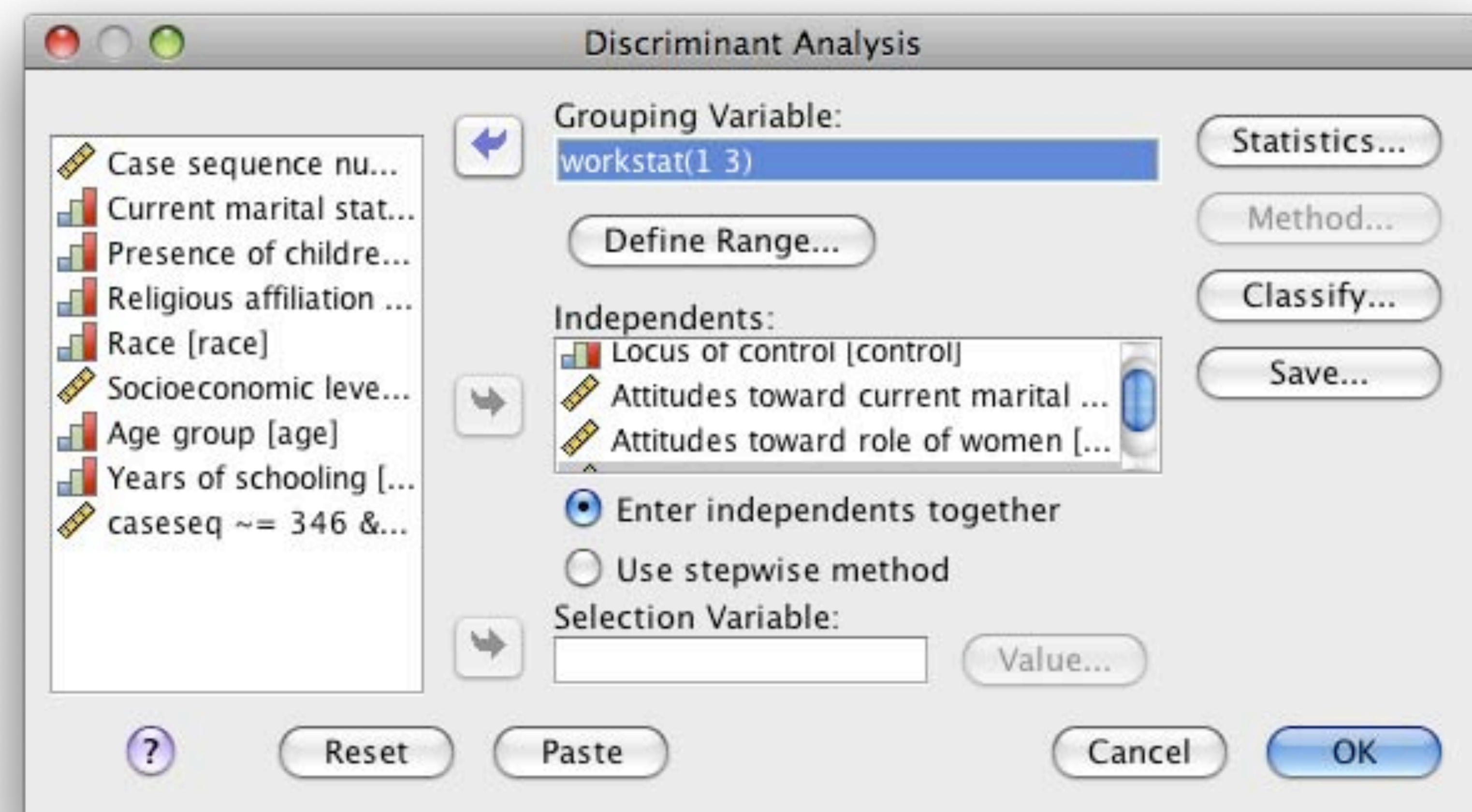
As recommended by Tabachnick and Fidell (2007) diagnostic checks were performed by groups. These indicated two multivariate outliers, cases 346 and 407.

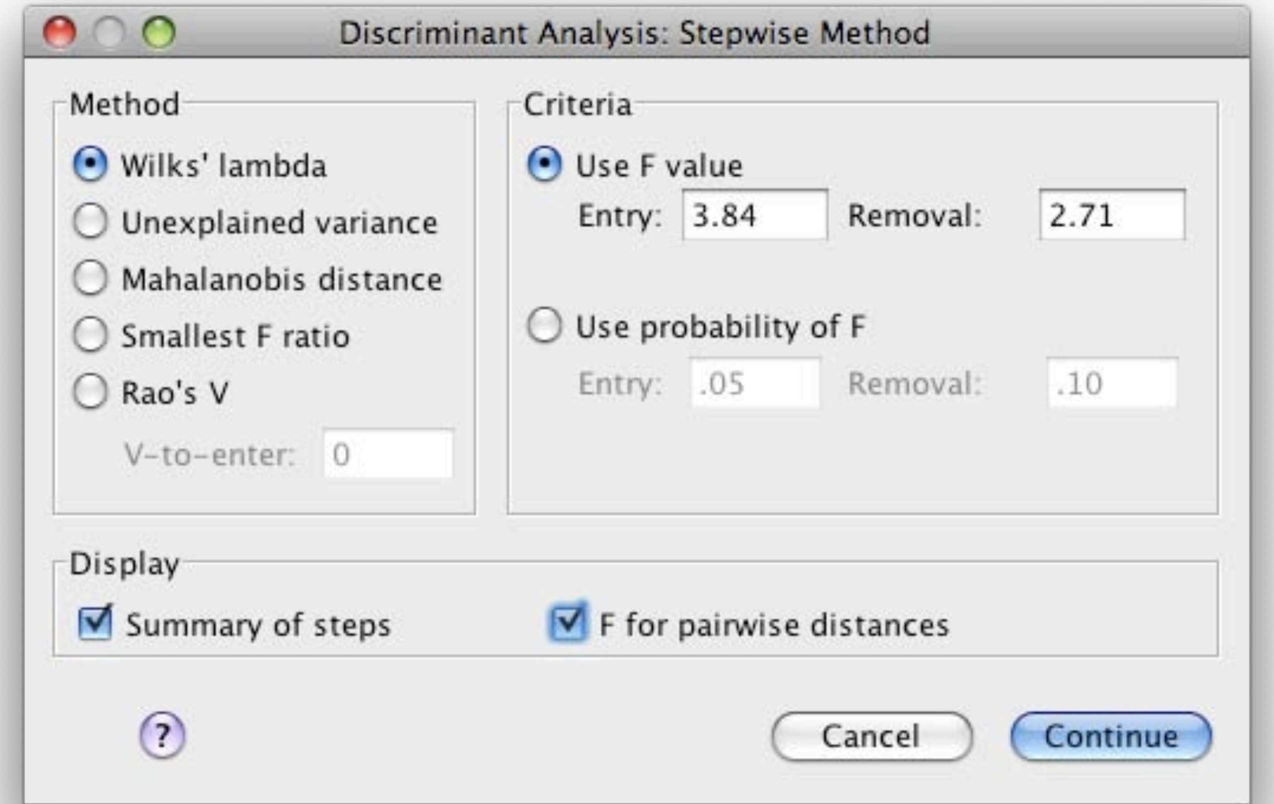
The select if command is used to select all cases not equal to the case sequence numbers using the variable ‘caseseq’.

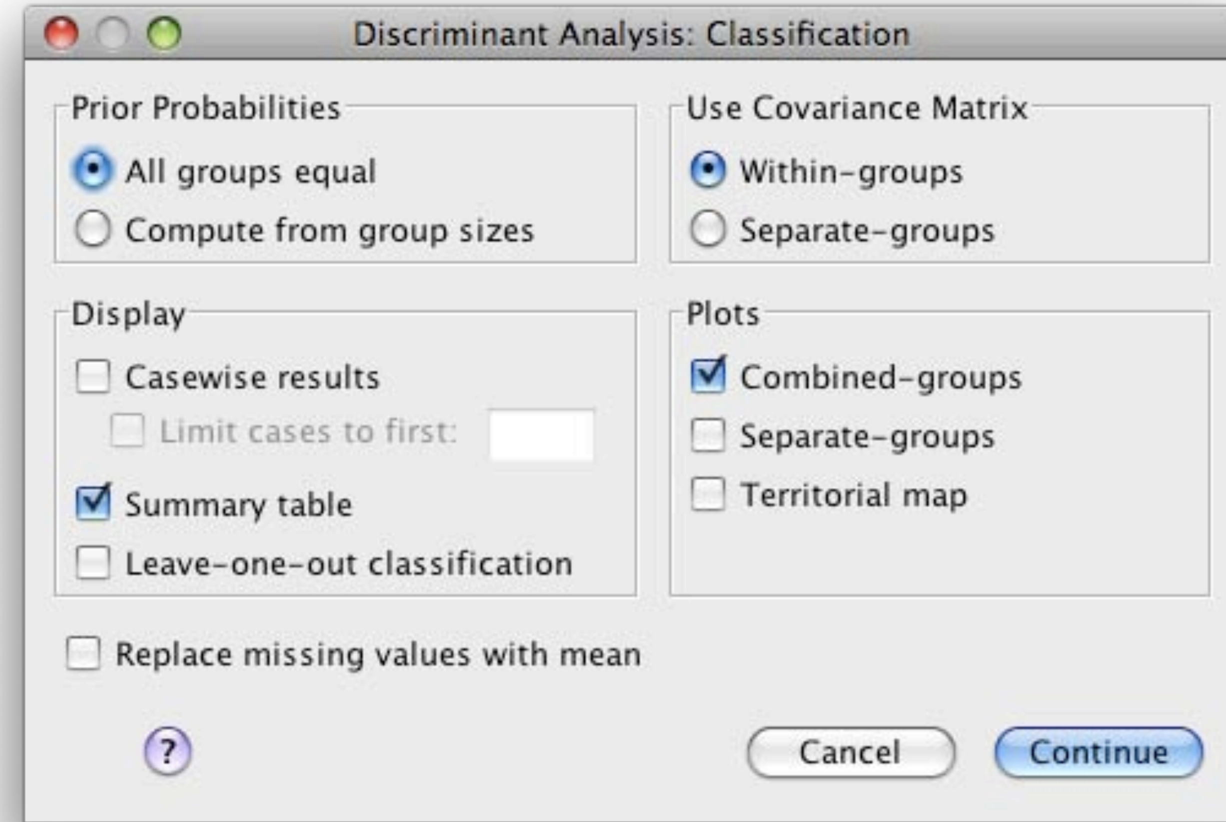
~~261~~	346.00	1.00	1.00	0.00	4.00	1.00	5.00	20.00	41.00	62.00	2.00	1.00	15.00
~~209~~	407.00	1.00	1.00	0.00	3.00	1.00	6.00	20.00	42.00	44.00	2.00	1.00	14.00

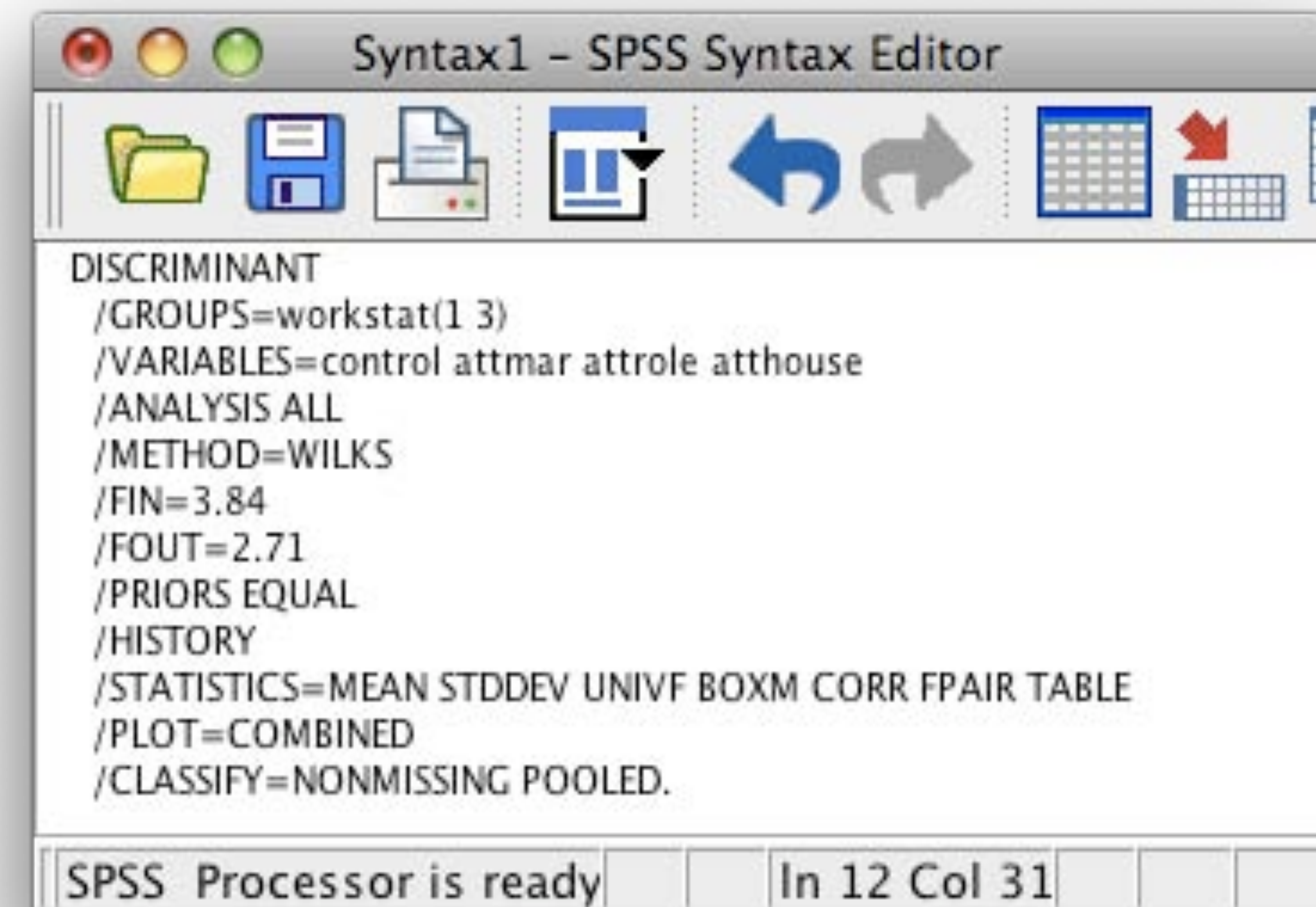

Analyze Graphs Utilities Add-ons Window Help
Reports
Descriptive Statistics
Tables
Compare Means
General Linear Model
Generalized Linear Models
Mixed Models
Correlate
Regression
Loglinear
Classify
Data Reduction
Scale
Nonparametric Tests
Time Series
Survival
Missing Value Analysis...
Multiple Response
Quality Control
ROC Curve...
TwoStep Cluster...
K-Means Cluster...
Hierarchical Cluster...
Discriminant...


Discriminant Analysis: Statistics
Descriptives
Means
Univariate ANOVAs
Box's M
Function Coefficients
Fisher's
Unstandardized
Matrices
Within-groups correlation
Within-groups covariance
Separate-groups covariance
Total covariance
?
Cancel
Continue


Discriminant Analysis
Case sequence nu...
Current marital stat...
Presence of childre...
Religious affiliation ...
Race [race]
Socioeconomic leve...
Age group [age]
Years of schooling [...
caseseq ~= 346 &...
Grouping Variable:
workstat(1 3)
Define Range...
Independents:
Locus of control [control]
Attitudes toward current marital ...
Attitudes toward role of women [...
Enter independents together
Use stepwise method
Selection Variable:
Value...
Statistics...
Method...
Classify...
Save...
?
Reset
Paste
Cancel
OK


Discriminant Analysis: Stepwise Method
Method
Wilks' lambda
Unexplained variance
Mahalanobis distance
Smallest F ratio
Rao's V
V-to-enter: 0
Criteria
Use F value
Entry: 3.84 Removal: 2.71
Use probability of F
Entry: .05 Removal: .10
Display
Summary of steps
F for pairwise distances
?
Cancel
Continue


Discriminant Analysis: Classification
Prior Probabilities
All groups equal
Compute from group sizes
Use Covariance Matrix
Within-groups
Separate-groups
Display
Casewise results
Limit cases to first:
Summary table
Leave-one-out classification
Plots
Combined-groups
Separate-groups
Territorial map
Replace missing values with mean
?
Cancel
Continue


Syntax1 – SPSS Syntax Editor
DISCRIMINANT
/GROUPS=workstat(1 3)
/VARIABLES=control attmar attrole atthouse
/ANALYSIS ALL
/METHOD=WILKS
/FIN=3.84
/FOUT=2.71
/PRIORS EQUAL
/HISTORY
/STATISTICS=MEAN STDDEV UNIVF BOXM CORR FPAIR TABLE
/PLOT=COMBINED
/CLASSIFY=NONMISSING POOLED.
SPSS Processor is ready
In 12 Col 31

```
DISCRIMINANT
  /GROUPS=workstat(1 3)
  /VARIABLES=control attmar attrole atthouse
  /ANALYSIS ALL (2)
  /METHOD=WILKS
  /FIN=3.84
  /FOUT=2.71
  /PRIORS EQUAL
  /HISTORY
  /STATISTICS=MEAN STDDEV UNIVF BOXM CORR FPAIR TABLE
  /PLOT=COMBINED
  /CLASSIFY=NONMISSING POOLED.
```

This seems rather mystical and cryptic – it is – This tells SPSS to force entry of every discriminant variable. This will give us give us F-TO-REMOVE values.

`/GROUPS` specifies the grouping variable and the range of values to be used in the analysis.

`/VARIABLES` lists all the variables to be used as discriminating (predictor, independent) variables.

`/ANALYSIS` and `/METHOD` : The default method of analysis performed by the DISCRIMINANT procedure is the direct method. However the direct method doesn't calculate the F-TO-REMOVE values which are needed for the interpretation. They are available by specifying a stepwise method, Wilks, when all the variables are forced to enter the analysis. The analysis subcommand specifies the variables to be used in the analysis and the (2) specifies the inclusion number for the variables. This particular value is even numbered and forces the variables entered together. The result of these two subcommands is to achieve the same results as for the direct method but allows the calculation of the F-TO-REMOVE values.

`/PLOT` produces a scatterplot of the discriminant scores (the linear composite) which also shows the group centroids. `COMBINED` provides a plot with all the cases.

In the `/STATISTICS` subcommand:

- `MEAN` and `STDDEV` give the means and standard deviations for each group and discriminating variable.
- `CORR` gives the pooled within groups correlation matrix.
- `UNIVF` produces the F tests for the differences between the groups on each variable.
- `BOXM` tests the equality of the group covariance matrices.
- `TABLE` produces a classification table.
- `FPAIR` produces a matrix of pairwise F ratios for the groups based on Mahalanobis distance between groups.

Interpretation of discriminant analysis

- Overall relationship
  - overall strength & statistical significance
  - number of significant functions
  - importance of each function
- Importance of each variable
  - overall importance
  - importance on each function
- Group separation

Test for Homogeneity

Box’s M

Test Results

Box's M		51.563
F	Approx.	2.537
	df1	20
	df2	245858
	Sig.	.000

Tests null hypothesis of equal population covariance matrices.

A significant Box’s M indicates a violation of the assumption of homogeneity of variances/co-variances. T&F state that when group sample sizes are equal or large, discriminant analysis is robust to violations of this assumption. They give further advice when sample sizes are small and/or unequal. Essentially the levels for the overall significance test of Wilk’s are not correct and care is needed with the interpretation of the overall significance test (i.e. be somewhat conservative).

However, although inferential (descriptive) Discriminant Analysis is usually robust to violation of this assumption, when the purpose of the Discriminant Analysis is classification (predictive discriminant analysis), it is not.

Overall statistical significance

Wilk's Lambda

$$\Lambda = \frac{|\mathbf{W}|}{|\mathbf{T}|} = \frac{|\mathbf{W}|}{|\mathbf{W}+\mathbf{B}|} = \prod_{j=1}^{r} \frac{1}{1+\lambda_j} = \prod_{j=1}^{r} \left(1 - R_{Cj}^2\right)$$

1. In terms of within and between group variance.
   - This is similar to the reciprocal of an F value: $\left(\frac{1}{F}\right)$
   - The bigger the effects of differences between groups the smaller the value of Λ.

Recall from last lecture...

Overall statistical significance

Wilk’s Lambda

Wilk’s Lambda is used to test the overall statistical significance of the discriminant model. Wilk’s Lambda varies between 0 and 1, with 0 meaning that the groups differ and 1 meaning that the groups are the same. However, Bartlett’s V, a transformation of Wilk’s Lamba that approximates a Chi-square distribution, is what is actually tested.

Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1 through 2	.897	49.002	8	.000
2	.966	15.614	3	.001

In the first step (1 through 2 in our example in the table; 1 through k-1 in general), both (all) functions are being tested. This is the overall test. If this is not significant then our discriminant variables are not able to distinguish between our groups.

Number of significant discriminant functions

Wilk's Lambda again

Wilk's Lambda is used to test the overall statistical significance of the discriminant model. Wilk's Lambda varies between 0 and 1, with 0 meaning that the groups differ and 1 meaning that the groups are the same. However, Bartlett's V, a transformation of Wilk's Lamba that approximates a Chi-square distribution, is what is actually tested.

Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1 through 2	.897	49.002	8	.000
2	.966	15.614	3	.001

There are 2 possible discriminant functions

With both functions in, there is a statistically significant effect.

For the second function there are still significant differences between groups.
So two functions needed to describe the between group differences.

Importance of the discriminant functions

Canonical correlations squared

The square of canonical correlation coefficient reported for each discriminant function estimates the amount of between group variability accounted for by each discriminant function.

Eigenvalues

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	.077[a]	68.6	68.6	.267
2	.035[a]	31.4	100.0	.184

a. First 2 canonical discriminant functions were used in the analysis.

$R_1^2 = .267^2 = .071 = 7.1\%$ of the between group variability that is explained by the first discriminant function.

$R_2^2 = .184^2 = .034 = 3.4\%$ of the between group variability that is explained by the second discriminant function.

Importance of the discriminant functions

Canonical correlations squared

The square of canonical correlation coefficient reported for each discriminant function estimates the amount of between group variability accounted for by each discriminant function.

Eigenvalues

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	.077[a]	68.6	68.6	.267
2	.035[a]	31.4	100.0	.184

a. First 2 canonical discriminant functions were used in the analysis.

Note: This is different to the '% of Variance' reported in the table. '% of Variance' looks at the contribution of that discriminant function relative to all other functions. From the table we can see that the 7.1% of between group variability explained by the first discriminant function makes up 68.6% (% of Variance column) of the amount of between group variance that the two modelled functions are together able to explain.

Canonical Correlations

Interpretation

Be sure not to confuse R^2_{Cj} with the '% variance' reported in SPSS.

R^2_{Cj}	% variance
$\sqrt{\frac{\lambda_j}{(1+\lambda_j)}}$	$\frac{\lambda_j}{(\sum \lambda_j)}$
How much of the between groups variability is accounted for by that function.	How well one discriminant function discriminates between groups in comparison to the all other discriminant functions in the analysis

Importance of the discriminant functions

Overall multivariate effect size – Pillai's measure η^2

$R_1^2 = .267^2 = .071 = 7.1\%$ $\qquad$ $R_2^2 = .184^2 = .034 = 3.4\%$

Eigenvalues

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	.077[a]	68.6	68.6	.267
2	.035[a]	31.4	100.0	.184

a. First 2 canonical discriminant functions were used in the analysis.

A measure of overall multivariate effect size is given by the average of the R_j^2. This is Pillai's measure and is called η^2. In general it should be calculated from all discriminant functions. In this example:

$$\eta^2 = \frac{.071 + .034}{2} = .0525 = 5.3\%$$

That is, on average, the discriminant functions each explain 5.3% of the between group variability. This effect is not overly strong but this will depend on the field of research.

Pooled Within-groups Correlation Matrix

Pooled Within-Groups Matrices

		Locus of control	Attitudes toward current marital status	Attitudes toward role of women	Attitudes toward housework
Correlation	Locus of control	1.000	.172	.009	.155
	Attitudes toward current marital status	.172	1.000	-.070	.282
	Attitudes toward role of women	.009	-.070	1.000	-.291
	Attitudes toward housework	.155	.282	-.291	1.000

The pooled within-group correlation matrix provides estimates of the correlations between variables with the effects of the grouping variable removed. In effect, this is as if the variables were correlated separately for each of the groups and these correlations were averaged.

This shows the correlation between the variables and shows the need to take any shared variance into account.

Relative importance of variables

Like multiple regression this is not an easy question to answer because there are many different statistics suggested.

In this course we will consider five of them:

- *Overall* importance of each variable
  - Each variable is considered separately

1. Univariate F-ratio
2. F-TO-REMOVE statistics and pr^2

- Importance of each variable for *each function*
  - Variables are considered in combination

3. Structure Coefficients
4. Standardised discriminant function coefficients
5. Relative Weights

Relative importance of variables

Univariate F-ratio

Tests of Equality of Group Means

	Wilks' Lambda	F	df1	df2	Sig.
Locus of control	.987	2.957	2	453	.053
Attitudes toward current marital status	.959	9.805	2	453	.000
Attitudes toward role of women	.953	11.261	2	453	.000
Attitudes toward housework	.962	8.911	2	453	.000

Three variables show statistically significant differences univariately ($p<.001$).

The way in which the groups differ on specific variables is found by looking at the means for each group. The univariate F ratios test for the difference between these means. These are simply a series of ANOVA's for each discriminant variable. These statistics don't take into account the interrelationships between the variables or the effect on the familywise error rate with multiple tests. The degrees of freedom are [k-1, N-k].

Relative importance of variables

F-TO-REMOVE statistics for each variable

- Provide similar information to squared semi-partial correlations.
  - measure how much the variable adds to the discrimination between groups after the other variables are in the equation.
- Obtained from SPSS sneakily by specifying a stepwise analysis but forcing all the variables into the analysis.
- Values are taken from the FINAL step in stepwise analysis.

Relative importance of variables

F-TO-REMOVE statistics for each variable

Variables in the Analysis

Step		Tolerance	F to Remove	Wilks' Lambda
1	Locus of control	1.000	2.957	
2	Locus of control	.971	1.652	.959
	Attitudes toward current marital status	.971	8.446	.987
3	Locus of control	.970	1.620	.917
	Attitudes toward current marital status	.965	7.518	.940
	Attitudes toward role of women	.995	10.301	.952
4	Locus of control	.955	1.076	.901
	Attitudes toward current marital status	.904	4.903	.917
	Attitudes toward role of women	.912	9.313	.934
	Attitudes toward housework	.833	3.218	.910

Final Step

Relative importance of variables

F-TO-REMOVE statistics for each variable

4	Locus of control	.955	1.076	.901
	Attitudes toward current marital status	.904	$\boxed{4.903}$	.917
	Attitudes toward role of women	.912	$\boxed{9.313}$	.934
	Attitudes toward housework	.833	$\boxed{3.218}$	.910

From Tables: the critical value of F for = .05 for testing F-TO-REMOVE is $F(2,450) = 2.99$. The degrees of freedom are $[k-1, N - k - p + 1]$. No Bonferonni adjustment.

Three variables are statistically significant using this critical value and contribute uniquely to the separation of the groups in addition to the other variables.

Relative importance of variables

partial η^2 ($pr^2\%$)

4	Locus of control	.955	1.076	.901	0.48
	Attitudes toward current marital status	.904	4.903	.917	2.13
	Attitudes toward role of women	.912	9.313	.934	3.97
	Attitudes toward housework	.833	3.218	.910	1.41

We can use the F-TO-REMOVE values to calculate an estimate of the effect size for the difference between groups for a variable controlling for the other variables. It's equivalent to pr^2, the squared partial-correlation coefficient. For the ith variable controlling for the other variables:

$$pr_i^2 = \frac{SS_{B_i}}{SS_{T_i}} \quad \text{for the ith variable.}$$

This is the proportion of total variance for a variable that is accounted for by the grouping variable controlling for the other variables. The formula for calculating this from the F-TO-REMOVE values is, where F_{tri} = F-TO-REMOVE for the ith variable,

$$pr_i^2 = \frac{\frac{(k-1)F_{tr_i}}{(N-k-p+1)}}{\left(\frac{(k-1)F_{tr_i}}{(N-k-p+1)} + 1\right)}$$

Relative importance of variables

Structure Coefficients (*s*)

These are the "pooled within group correlations between the discriminant functions and the discriminating variables". That is, they are the correlations between the four discriminant variables and each of the two discriminant functions, (Composite 1 and Composite 2). The correlations are calculated within each group and then pooled.

Structure Matrix

	Function	
	1	2
Attitudes toward current marital status	.718*	.323
Attitudes toward housework	.679*	.333
Attitudes toward role of women	-.639	.722*
Locus of control	.282	.445*

Pooled within-groups correlations between discriminating variables and standardized canonical discriminant functions
Variables ordered by absolute size of correlation within function.
*. Largest absolute correlation between each variable and any discriminant function

Forget the * in SPSS

An advantage of structure coefficients is that they have a range from –1 to 1.The 'meaning' of the variables can be used to place a meaning or an interpretation on the discriminant function. The definition of a high value for these correlations is problematic. T&F employ a variety of criteria, e.g. structure coefficients greater than .50, or .30. There is no agreed value for the cutoff and there are no parametric tests of significance.

Relative importance of variables

Standardised Discriminant Function Coefficients (*d*)

Standardized Canonical Discriminant Function Coefficients

	Function	
	1	2
Locus of control	.135	.329
Attitudes toward current marital status	.560	.191
Attitudes toward role of women	-.498	.873
Attitudes toward housework	.355	.483

These are similar to beta weights in multiple regression.

These represent the unique contribution of each variable to the discriminant functions, taking into account any shared variance between variables.

T&F state that using the magnitude of these coefficients can be misleading. This is because their theoretical range is from minus to plus infinity.

Relative importance of variables

Relative Weights (d x s)

Structure Matrix

	Function	
	1	2
Attitudes toward current marital status	.718*	.323
Attitudes toward housework	.679*	.333
Attitudes toward role of women	-.639	.722*
Locus of control	.282	.445*

Pooled within-groups correlations between discriminating variables and standardized canonical discriminant functions
Variables ordered by absolute size of correlation within function.

*. Largest absolute correlation between each variable and any discriminant function

Standardized Canonical Discriminant Function Coefficients

	Function	
	1	2
Locus of control	.135	.329
Attitudes toward current marital status	.560	.191
Attitudes toward role of women	-.498	.873
Attitudes toward housework	.355	.483

	Function	
	1	2
Attitudes toward current marital status	40.24%	6.18%
Attitudes toward housework	24.15%	16.09%
Attitudes toward role of women	31.81%	63.08%
Locus of control	3.80%	14.65%
Total	100%	100%

They indicate for each function the proportion of between group variability accounted for by a variable. Like RW in multiple regression they could also be expressed as percentages.

Relative importance of variables

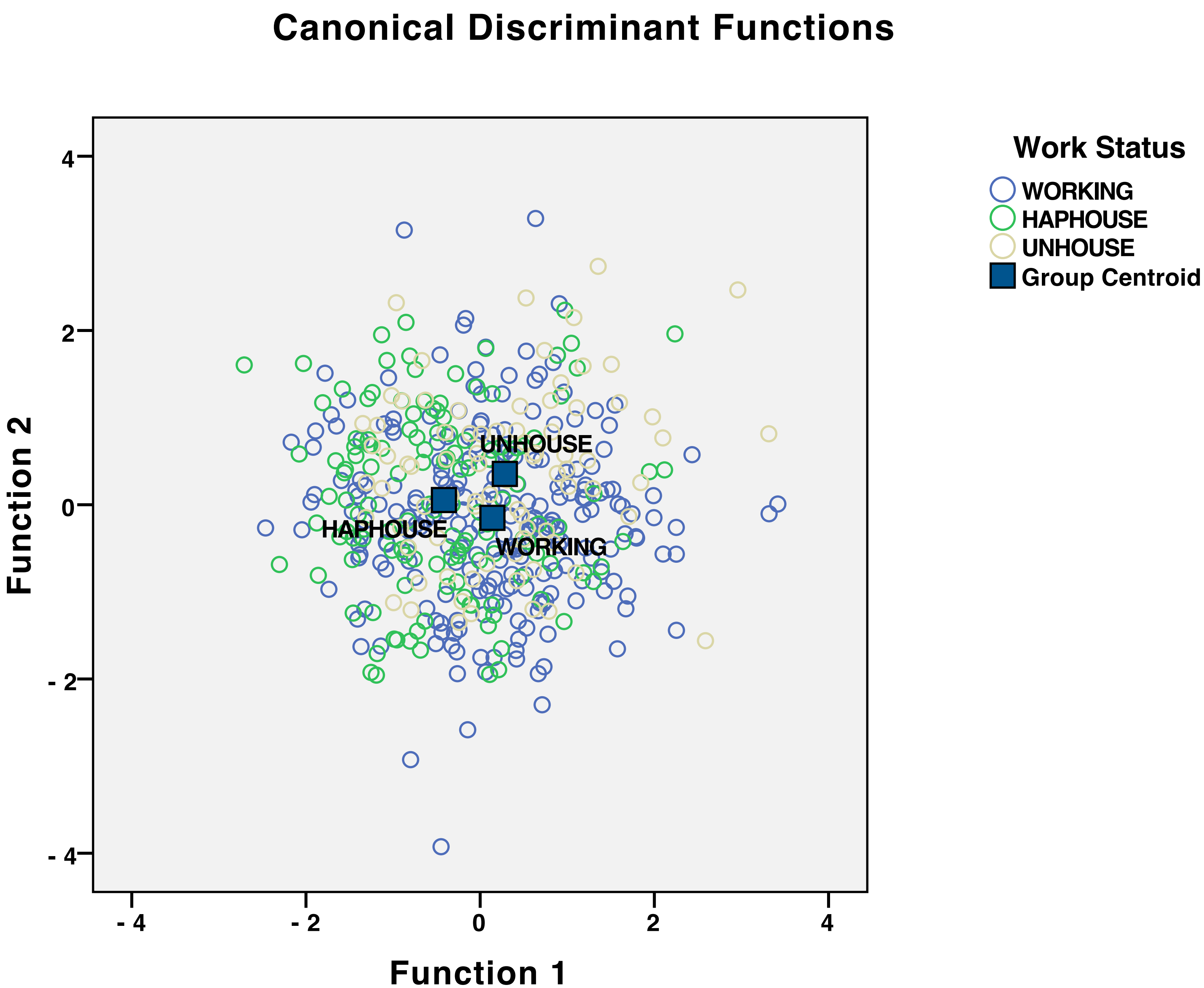
1. Univariate F-ratio
2. F-TO-REMOVE statistics and pr^2
3. Structure Coefficients
4. Standardised discriminant function coefficients
5. Relative Weights

The process of deciding what variables are important takes into account the pattern of results across the above five statistics. This is because no single statistic tells the ‘full’ story; they each view the group differences from different angles.

Group separation

Centroid Plots in reduced discriminant space

How are the groups separated? This is answered by plotting the group centroids (looking at the combined-groups plot or plotting them yourselves from the table) and by labelling the discriminant functions with the names of the important variables. This shows the use of discriminant analysis as a data reduction method.



Note the considerable overlap of the groups!

Group separation

Centroid Plots in reduced discriminant space

The group centroids are the means for each group on each discriminant function.

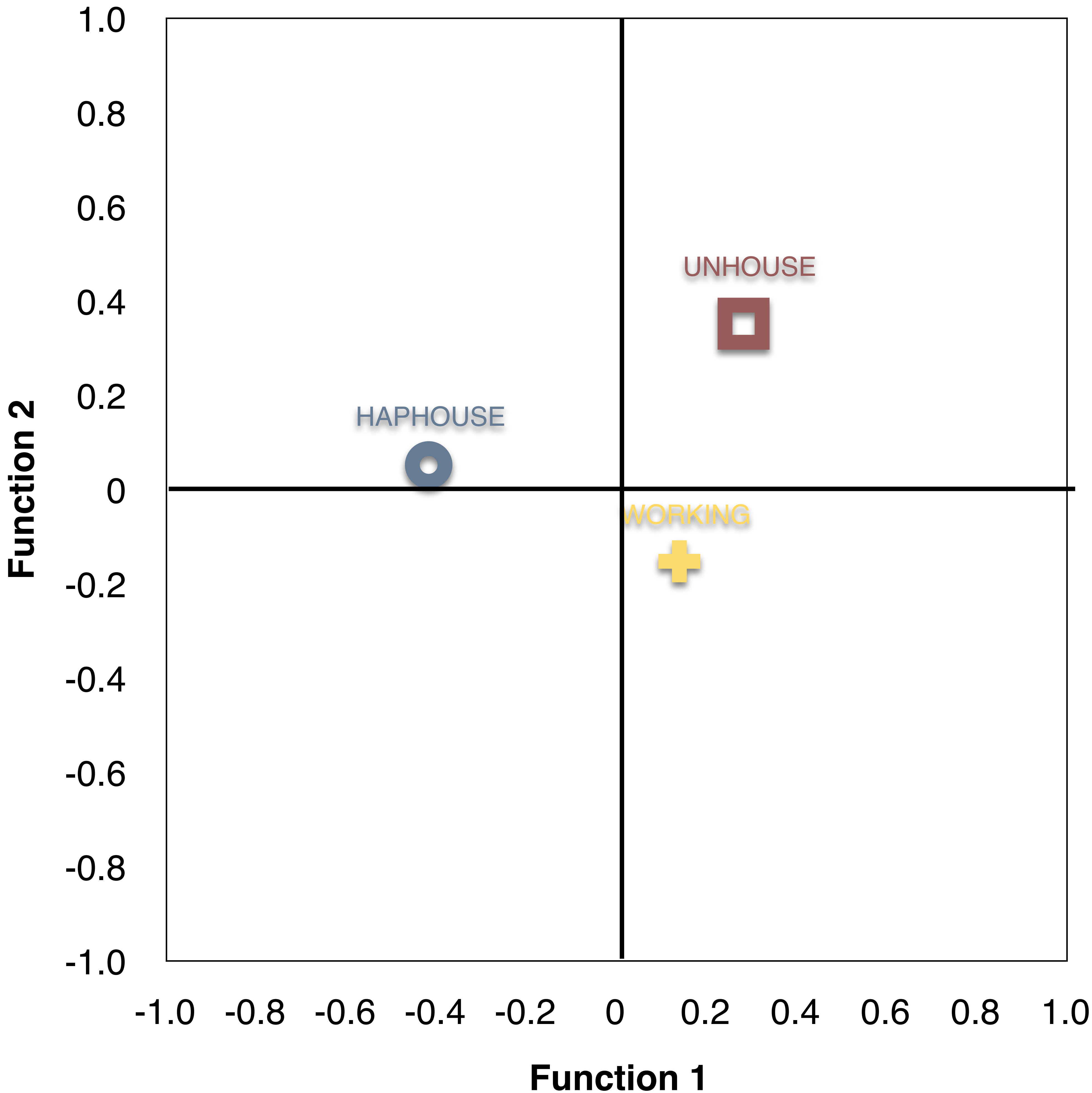
Since the group centroids are a linear combination of the means for each variable, there may be some discrepancies in an interpretation based on the group centroids and the means for each variable. Which is used depends on the focus of the interpretation; whether each variable separately or the combination of the variables is of interest.

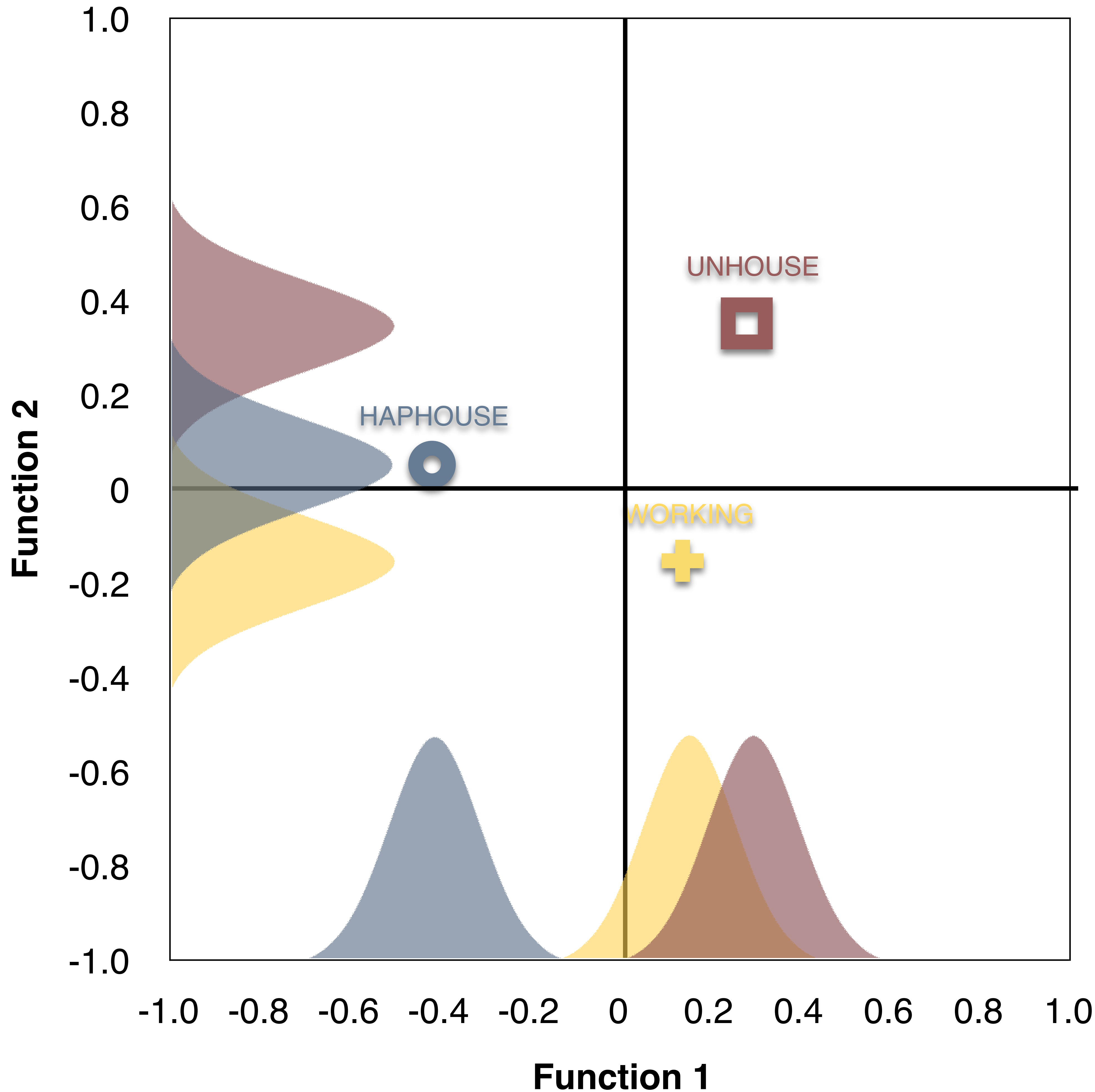
Functions at Group Centroids

	Function	
Work Status	1	2
WORKING	.141	-.151
HAPHOUSE	-.416	5.393E-02
UNHOUSE	.283	.354

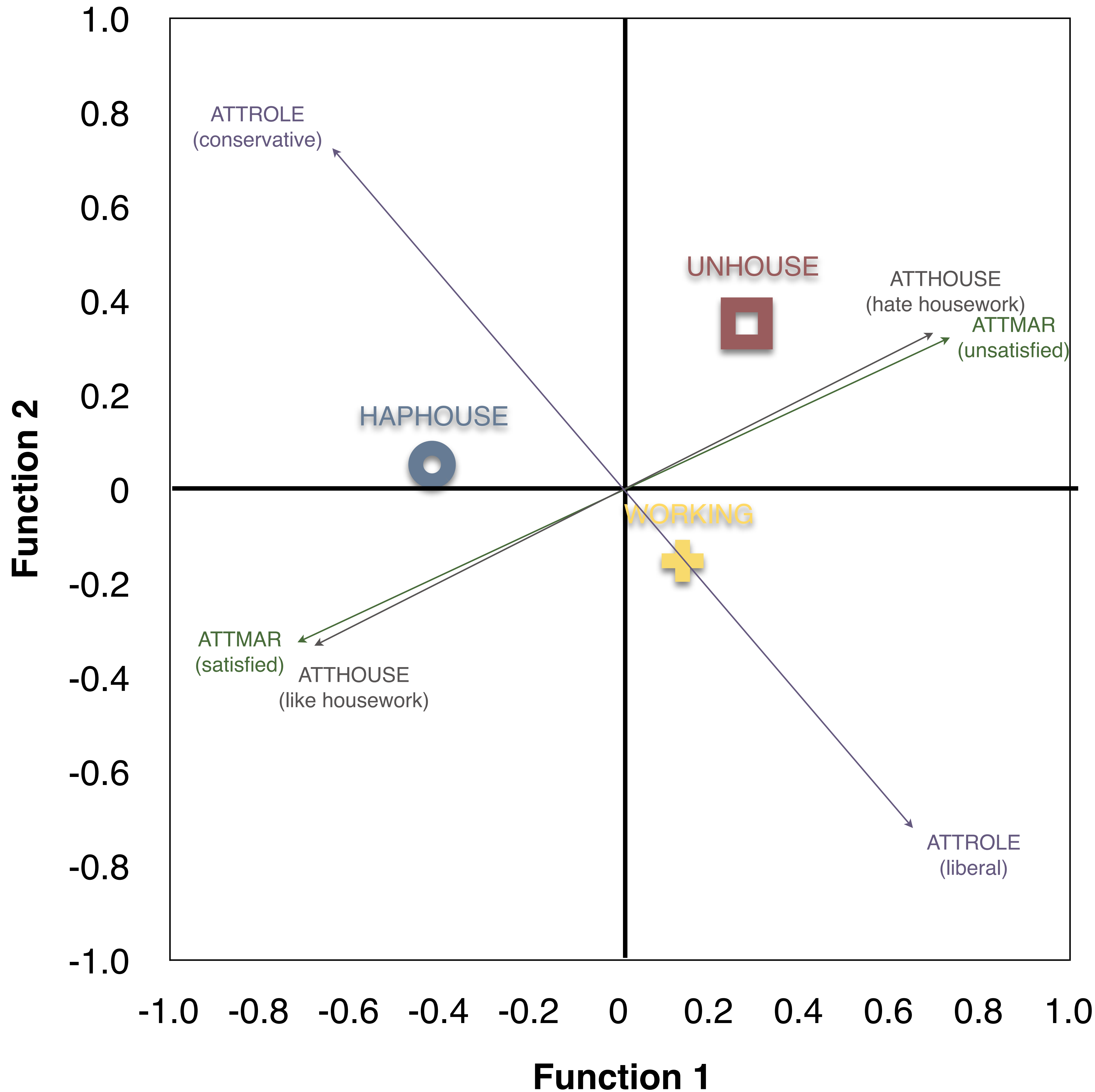
Unstandardized canonical discriminant functions evaluated at group means

Another approach, is to superimpose a plot of the variables in the discriminant function space.


1.0
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1.0
Function 2
UNHOUSE
HAPHOUSE
WORKING
-1.0
-0.8
-0.6
-0.4
-0.2
0
0.2
0.4
0.6
0.8
1.0
Function 1

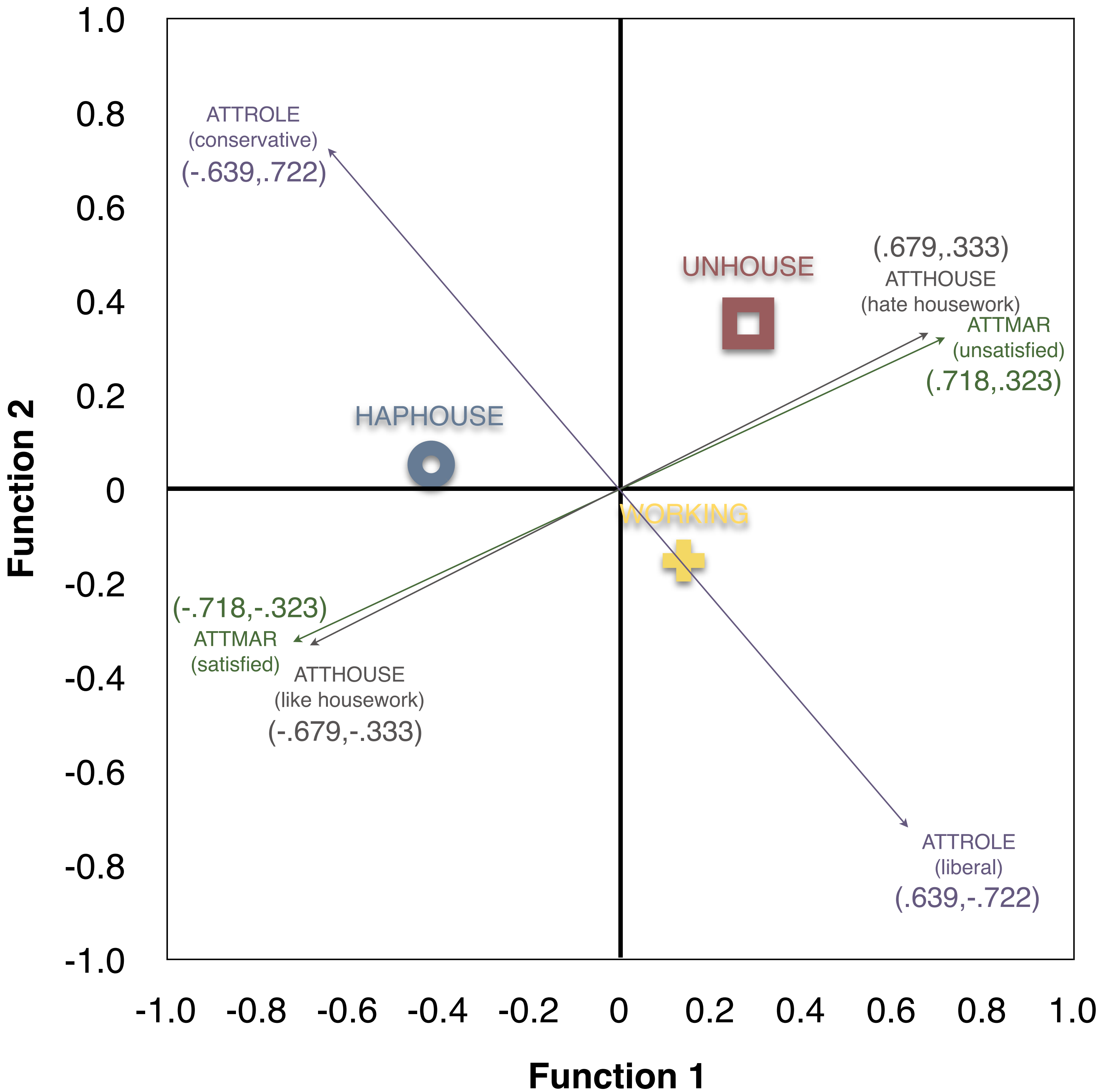


1.0
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1.0
Function 2
UNHOUSE
HAPHOUSE
WORKING
-1.0 -0.8 -0.6 -0.4 -0.2 0 0.2 0.4 0.6 0.8 1.0
Function 1


1.0
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1.0
Function 2
-1.0
-0.8
-0.6
-0.4
-0.2
0
0.2
0.4
0.6
0.8
1.0
Function 1
ATTROLE
(conservative)
UNHOUSE
ATTHOUSE
(hate housework)
ATTMAR
(unsatisfied)
HAPHOUSE
WORKING
ATTMAR
(satisfied)
ATTHOUSE
(like housework)
ATTROLE
(liberal)

Group separation

Group Centroid Plot with variables as bipolar vectors



Structure Matrix

	Function	
	1	2
Attitudes toward current marital status	.718*	.323
Attitudes toward housework	.679*	.333
Attitudes toward role of women	-.639	.722*
Locus of control	.282	.445*

Pooled within-groups correlations between discriminating variables and standardized canonical discriminant functions
Variables ordered by absolute size of correlation within function.

*. Largest absolute correlation between each variable and any discriminant function

To map each SIGNIFICANT variable onto the functions, use the structure coefficient as coordinates for each variable and then reflecting the line through the origin to make it a bipolar vector.

Group separation

Matrix of pairwise F values and Group means

Pairwise Group Comparisons[a,b,c,d]

Step	Work Status		WORKING	HAPHOUSE	UNHOUSE
1	WORKING	F		.376	4.231
		Sig.		.540	.040
	HAPHOUSE	F	.376		5.539
		Sig.	.540		.019
	UNHOUSE	F	4.231	5.539	
		Sig.	.040	.019	
2	WORKING	F		4.826	3.614
		Sig.		.008	.028
	HAPHOUSE	F	4.826		10.443
		Sig.	.008		.000
	UNHOUSE	F	3.614	10.443	
		Sig.	.028	.000	
3	WORKING	F		9.882	4.064
		Sig.		.000	.007
	HAPHOUSE	F	9.882		7.581
		Sig.	.000		.000
	UNHOUSE	F	4.064	7.581	
		Sig.	.007	.000	
4	WORKING	F		7.572	4.124
		Sig.		.000	.003
	HAPHOUSE	F	7.572		7.297
		Sig.	.000		.000
	UNHOUSE	F	4.124	7.297	
		Sig.	.003	.000	

a. 1, 453 degrees of freedom for step 1.

b. 2, 452 degrees of freedom for step 2.

c. 3, 451 degrees of freedom for step 3.

d. 4, 450 degrees of freedom for step 4.

Final Step

Group separation

Matrix of pairwise F values and Group means

Pairwise Group Comparisons[a,b,c,d]

Step	Work Status		WORKING	HAPHOUSE	UNHOUSE
4	WORKING	F		7.572	4.124
		Sig.		.000	.003
	HAPHOUSE	F	7.572		7.297
		Sig.	.000		.000
	UNHOUSE	F	4.124	7.297	
		Sig.	.003	.000	

d. 4, 450 degrees of freedom for step 4.

The matrix of pairwise F values between the groups tests which groups are different from one another over all the variables. This can be useful when describing the differences between the groups in the group-centroid plot.

Group separation

Discriminant variable mean differences at the group level

Another aid to interpretation is the difference between the means for each of the 'important' variables. This breaks down the group centroids into group means for each discriminant variable. The focus of interpretation should be on means for variables earlier determined to be an important part of a discriminant function.

Group Statistics

Work Status		Mean	Std. Deviation	Valid N (listwise) Unweighted	Weighted
WORKING	Locus of control	6.7155	1.23780	239	239.000
	Attitudes toward current marital status	23.3975	8.53004	239	239.000
	Attitudes toward role of women	33.8619	6.95618	239	239.000
	Attitudes toward housework	23.8117	4.45544	239	239.000
HAPHOUSE	Locus of control	6.6324	1.30984	136	136.000
	Attitudes toward current marital status	20.6029	6.62350	136	136.000
	Attitudes toward role of women	37.1912	6.45843	136	136.000
	Attitudes toward housework	22.5074	3.88348	136	136.000
UNHOUSE	Locus of control	7.0494	1.25401	81	81.000
	Attitudes toward current marital status	25.6173	10.29753	81	81.000
	Attitudes toward role of women	35.6667	5.75977	81	81.000
	Attitudes toward housework	24.9259	3.95846	81	81.000
Total	Locus of control	6.7500	1.26795	456	456.000
	Attitudes toward current marital status	22.9583	8.52871	456	456.000
	Attitudes toward role of women	35.1754	6.75895	456	456.000
	Attitudes toward housework	23.6206	4.27859	456	456.000

Editing this table, rearrange the columns and rows and delete other information to produce...

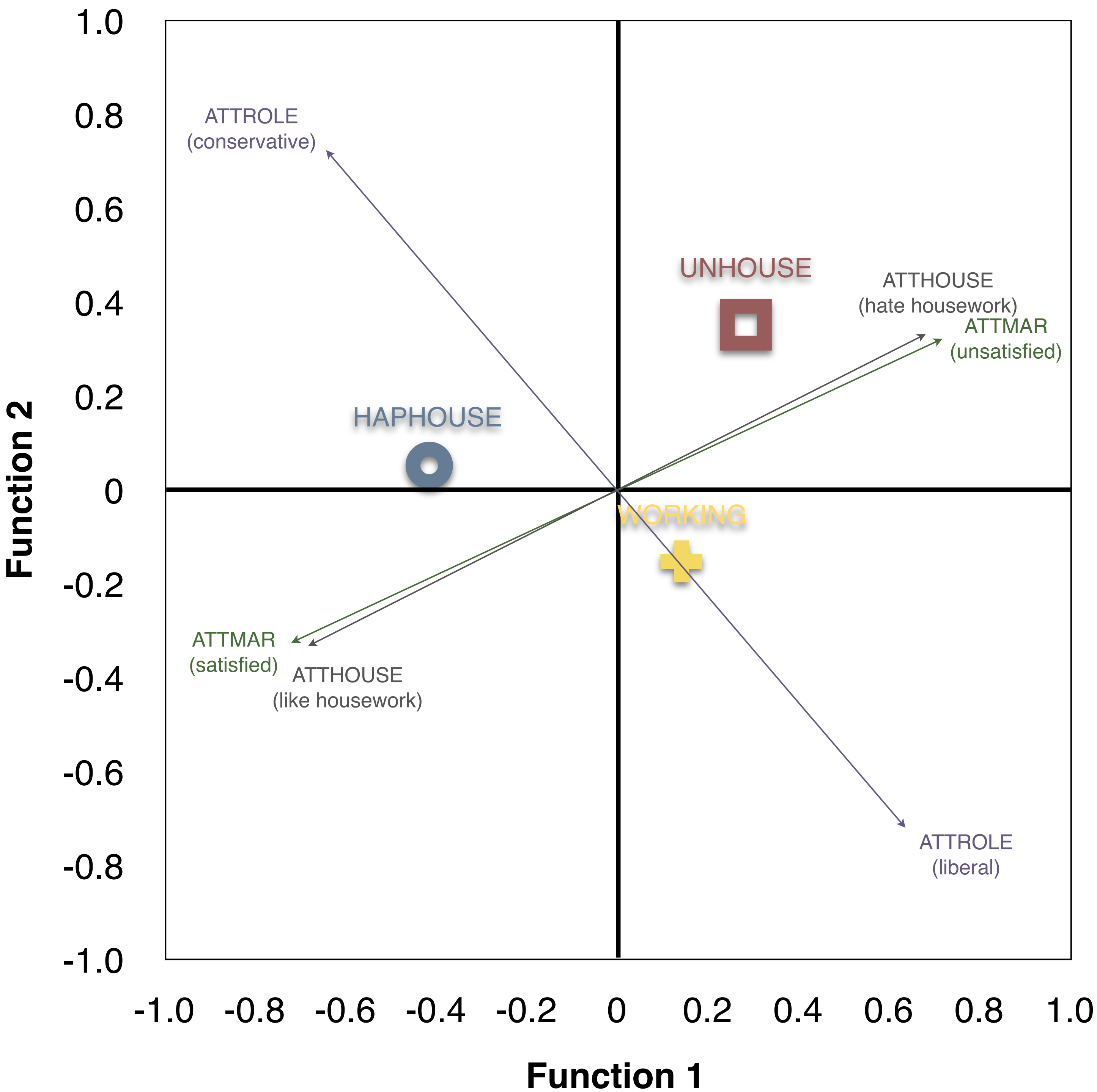
Group separation

Discriminant variable mean differences at the group level

Another aid to interpretation is the difference between the means for each of the ‘important’ variables. This breaks down the group centroids into group means for each discriminant variable. The focus of interpretation should be on means for variables earlier determined to be an important part of a discriminant function.

WORKSTAT Work Status	CONTROL Locus of control	ATTMAR Attitudes toward current marital status	ATTROLE Attitudes toward role of women	ATTHOUSE Attitudes toward housework
WORKING	6.7155	23.3975	33.8619	23.8117
HAPHOUSE	6.6324	20.6029	37.1912	22.5074
UNHOUSE	7.0494	25.6173	35.6667	24.9259
Total	6.7500	22.9583	35.1754	23.6206

...something like this.



You can clearly see by looking at one's attitude toward housework (along the ATTHOUSE vector), that **unhappy** housewives are at one end, and **happy** housewives are at the other end. Those who are **working** fall in the middle. This is reflected in the difference in group means in the ATTHOUSE column. Use these means to help work out the direction.

WORKSTAT Work Status	CONTROL Locus of control	ATTMAR Attitudes toward current marital status	ATTROLE Attitudes toward role of women	ATTHOUSE Attitudes toward housework
WORKING	6.7155	23.3975	33.8619	23.8117
HAPHOUSE	6.6324	20.6029	37.1912	**22.5074**
UNHOUSE	7.0494	25.6173	35.6667	**24.9259**
Total	6.7500	22.9583	35.1754	23.6206

Classification

Prediction of group membership

How well do the discriminant functions predict group membership? The classification table provides this information. Not only is the overall percent of correctly classified important, but also by looking at the miss-classifications, groups that overlap can be identified.

Classification Results[a]

			Predicted Group Membership			
		Work Status	WORKING	HAPHOUSE	UNHOUSE	Total
Original	Count	WORKING	98	70	71	239
		HAPHOUSE	37	74	25	136
		UNHOUSE	22	22	37	81
	%	WORKING	41.0	29.3	29.7	100.0
		HAPHOUSE	27.2	54.4	18.4	100.0
		UNHOUSE	27.2	27.2	45.7	100.0

a. 45.8% of original grouped cases correctly classified.

The accuracy of the classification is influenced by the decisions about the ‘prior probability’ of group membership. Sometimes it might be plausible that each case has an equally likely chance of being in each group. Other times, the group size gives an estimate of the population proportions. Other times, the user may have theoretical reasons for specifying other prior probabilities of group membership.

Comparing Multiple Regression and Discriminant Analysis

	Multiple Regression	**Discriminant Analysis**
Overall significance of the relationship	F test $H_0 : R = 0$ or $H_0 : (1 - R^2) = 1$	χ^2 test $H_0 : V = 0$ or $H_0 : \prod(1 - R_i) = 1$
Importance of Relationship	Squared Multiple Correlation = R^2	Squared Canonical Correlation = R_{ci}^2
Number of dimensions	Only one dimension	Tested using a stepwise analysis

Comparing Multiple Regression and Discriminant Analysis

	Multiple Regression	**Discriminant Analysis**
What variables are important in the relationship?	Simple r_{yi}	Univariate F test for each variable
	sr^2	F-TO-REMOVE for each variable
	beta weights	matrix of standardised discriminant function coefficients (d_i)
	not used	matrix of structure coefficients (s_i)
	Relative Weights $\left(\frac{\beta_{r_{yi}}}{R^2}\right)$	matrix of relative weights $(d_i s_i)$

Comparing Multiple Regression and Discriminant Analysis

- Description of how the predictors explain differences in the criterion:
  - Multiple Regression
    - description of prediction equation (not often used in psychology)
  - Discriminant Analysis
    - description of group separation on the basis of group centroid plot, classification table, pairwise F-tests, mean differences on important predictors.

Which parts of the results of a discriminant analysis are used for interpretation depends on the kind of research question addressed and whether the focus is on the multivariate nature of the variables or on variables considered individually.

Questions

1. Describe how the number of significant discriminant functions is determined.
2. How do outliers affect Discriminant Analysis?
3. Explain the distinctions between using Univariate F's, F-TO-REMOVE statistics, structure coefficients, standardised discrimination function coefficients and relative weights for interpretation of discriminant analysis.
4. What issues need to be addressed if the purpose of Discriminant Analysis is classification?