Two-step Clustering in older SPSS versions

Compared to SPSS 18 (which is the basis for this book), former versions of SPSS have slightly different menu options and outputs. While the **Main** menu and **Options** menu exhibit the same features, SPSS has a menu called **Plots** where we can request a variable importance plot, which shows each variable's relevance for the construction of each cluster. This can help with interpreting and profiling clusters, as variables may sometimes be important for the creation of one cluster but unimportant for the creation of another. We can control the results display in the **Rank Variables** submenu (since we want to compare the variables per cluster, select **By Variable**) and determine the importance measure. **Chi-square or T-test of significance** produces a Pearson chi-square statistic indicating the importance of a categorical variable and a t-statistic indicating the importance of a continuous variable. The output produced by checking the **Significance** option essentially contains the same information; simply use the first one.

In the **Output** menu, we can choose between different tables of the clustering results, including the cluster distribution (i.e. the number of objects in each cluster), cluster profiles (i.e. the mean values and standard deviations of all clustering variables with respect to each cluster), as well as the results from the auto-clustering process, which is in our case based on the BIC values. Choose all options in the **Statistics** submenu and make sure that you select the option **Create cluster membership variable** in the **Working Data File** submenu. By clicking on **Continue** and finally **OK**, you can run the analysis.

The first output table provides an overview of the auto-clustering procedure (Table A9.1). SPSS computes several models, each with a different number of clusters, and produces values for BIC and the ratio of distance measures for each solution.

SPSS automatically determines the number of clusters on the basis of the ratio of distance measures. More precisely, the program calculates clustering solutions for different numbers of clusters (e.g., one to six) and selects the solution that maximizes the ratio of distance measures. In this case, a two-segment solution is selected. In contrast, BIC (and AIC) are constructed in such a way that the solution

t1.1 **Table A9.1** Auto-clustering results

t1.2	Auto-Clustering				
t1.3	Number of Clusters	Schwarz's Bayesian Criterion (BIC)	BIC Change[a]	Ratio of BIC Changes[b]	Ratio of Distance Measures[c]
t1.4	1	91.829			
t1.5	2	84.291	−7.537	1.000	3.314
t1.6	3	110.790	26.499	−3.516	1.272
t1.7	4	138.458	27.668	−3.671	1.292
t1.8	5	169.615	31.157	−4.134	1.271
t1.9	6	198.929	29.315	−3.889	1.308

t1.10 a. The changes are from the previous number of clusters in the table.
b. The ratios of changes are relative to the change for the two cluster solution.
c. The ratios of distance measures are based on the current number of clusters against the previous number of clusters.

with the lowest value indicates the number of clusters that can be retained from the data (see Chapter 9). This provides us with additional information on a potential solution and there might be cases where AIC or BIC indicate a different number of segments, since SPSS retains segments on the basis of the ratio of distance measures. In such a situation, you should rather compute several models with different numbers of clusters and evaluate each on practical grounds and the solution's interpretability. Do not rely on the automatic model selection when there is a mix of continuous and categorical variables, as this does not always work well. Examine the results very carefully! In our case, both BIC and the ratio of distance measures uniformly indicate a two-segment solution, which differs from our previous analyses (the same holds when using AIC instead of BIC). Table A9.2 indicates the cluster distribution. Upon closer examination of the objects' partitioning, we can see that the first cluster in the two-step procedure is identical to the clusters obtained from the hierarchical and k-means clustering (see Ch. However, the second cluster comprises sports cars as well as limousines, which were previously separate clusters. We can ascertain why this is so by re-running the Chapter 9) previous analysis. We could use different distance measures or hierarchical clustering methods, or carry out k-means clustering for a two- and three-segment solution without pre-specifying the cluster centers. Again, choosing the number of segments should not be exclusively based on the results provided by SPSS. Data only provide a rough guideline. Ultimately, however, we have to take practical issues into account.

t2.1 **Table A9.2** Cluster distribution

t2.2	Cluster Distribution				
t2.3			N	% of Combined	% of Total
t2.4	Cluster	1	7	46.7%	46.7%
t2.5		2	8	53.3%	53.3%
t2.6		Combined	15	100.0%	100.0%
t2.7	Total		15		100.0%

Table A9.3 shows the cluster centroids of some of the clustering variables. As one might expect, the merging of clusters two and three (see previous analyses in

t3.1 **Table A9.3** Centroids

t3.2			Centroids											
t3.3			Turning moment in Nm		Width in mm		Net weight in kg		Trunk volume in litres		Maximum speed in km/h		Acceleration 0–100 km/h in seconds	
t3.4			Mean	Std. Deviation	Mean	Std. Deviation	Mean	Std. Deviation	Mean	Std. Deviation	Mean	Std. Deviation	Mean	Std. Deviation
t3.5	Cluster	1	116.86	19.334	1699.14	52.806	1115.57	100.528	249.43	69.670	170.00	9.950	12.9571	1.50317
t3.6		2	297.88	43.871	1812.25	31.422	1538.88	107.039	488.00	116.899	233.38	26.224	8.2375	2.05908
t3.7		Combined	213.40	99.300	1759.47	71.416	1341.33	240.505	376.67	155.212	203.80	38.175	10.4400	3.00471

the book) comes at the expense of higher standard deviations in most cases. For example, in the previous analysis, *speed* exhibited a standard deviation of 17.678 (second segment) and 21.163 (third segment) units, respectively (Table 9.14), but 26.224 units in the present analysis, indicating a higher degree of fuzziness and uncertainty. However, some of this uncertainty can be resolved by looking at the variables' importance for each of the clusters (Figures A9.1 and A9.2). For example, Figure A9.1 lists the variables in descending order of importance for the creation of the first cluster, based on absolute t-values. Higher (absolute) t-values denote the variable's greater importance for the clustering solution. Comparing the results, we can see that *moment* is the most important variable for the first cluster, followed by *speed*, *weight*, and *trunk*. Whereas *moment* is also of importance for the second cluster, *weight* and *width* play a greater role in this segment.

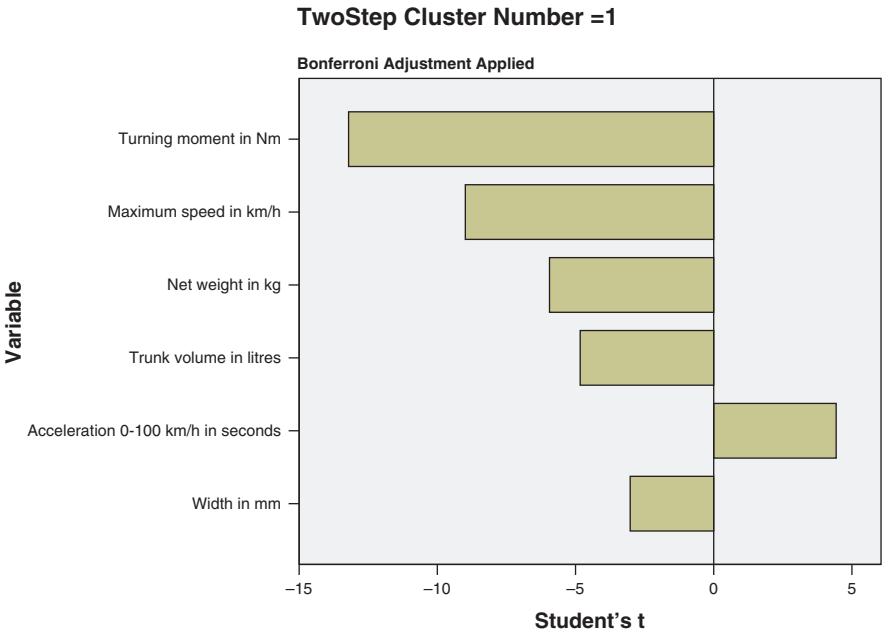

Figure A9.1 Variables' importance for the first cluster.

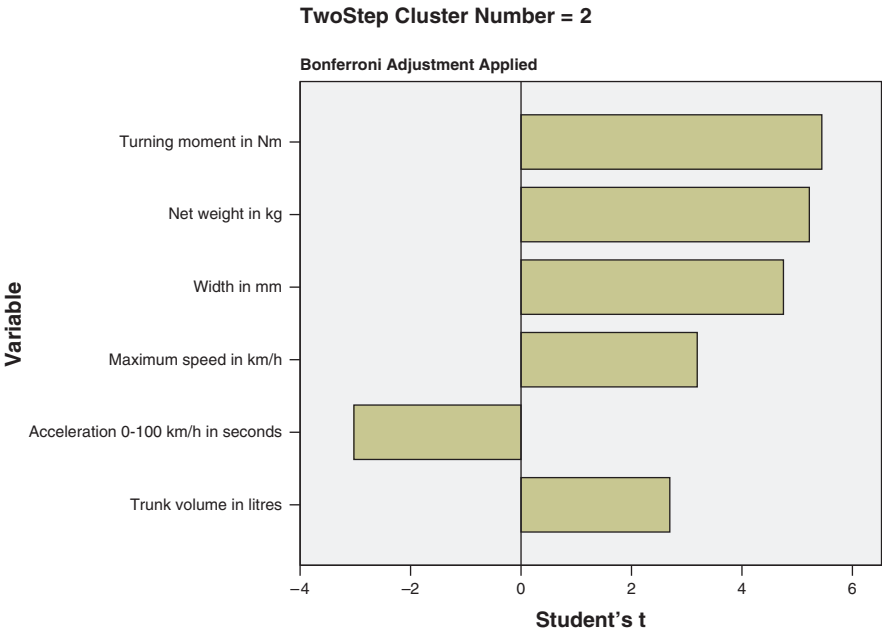

Figure A9.2 Variables' importance for the second cluster.