Completely elaborated example of data order instability in a Between-group Average hierarchical cluster analysis

Willem A. van der Kloot
Department of Psychology
Leiden University
The Netherlands

November 15, 2005
Given the matrix $\mathbf{D}$ of distances between six objects, A through $F$, a hierarchical agglomerative cluster analysis using the Between-group average\(^1\) method (Baverage; also known as UPGMA) is performed.

<table>
<thead>
<tr>
<th></th>
<th>$A$</th>
<th>$B$</th>
<th>$C$</th>
<th>$D$</th>
<th>$E$</th>
<th>$F$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A$</td>
<td>0</td>
<td>25</td>
<td>17</td>
<td>115</td>
<td>102</td>
<td>64</td>
</tr>
<tr>
<td>$B$</td>
<td>25</td>
<td>0</td>
<td>19</td>
<td>90</td>
<td>76</td>
<td>38</td>
</tr>
<tr>
<td>$C$</td>
<td>17</td>
<td>19</td>
<td>0</td>
<td>101</td>
<td>89</td>
<td>48</td>
</tr>
<tr>
<td>$D$</td>
<td>115</td>
<td>90</td>
<td>101</td>
<td>0</td>
<td>22</td>
<td>51</td>
</tr>
<tr>
<td>$E$</td>
<td>102</td>
<td>76</td>
<td>89</td>
<td>22</td>
<td>0</td>
<td>49</td>
</tr>
<tr>
<td>$F$</td>
<td>64</td>
<td>38</td>
<td>48</td>
<td>51</td>
<td>49</td>
<td>0</td>
</tr>
</tbody>
</table>


Step 1. Find the smallest distance in $\mathbf{D}$. This is $d_{AC} = 17$, the distance between Objects $A$ and $C$.

Merge Objects $A$ and $C$ into Cluster $\{A, C\}$. Compute the distances between Cluster $\{A, C\}$ and the remaining objects by averaging the distances between each object outside and the two objects in the cluster: $d_{B\{A,C\}} = (d_{AB} + d_{BC})/2 = (25 + 19)/2 = 22$; $d_{D\{A,C\}} = (d_{AD} + d_{CD})/2 = (115 + 101)/2 = 108$; $d_{E\{A,C\}} = (d_{AE} + d_{CE})/2 = (102 + 89)/2 = 95.5$; $d_{F\{A,C\}} = (d_{AF} + d_{CF})/2 = (64 + 48)/2 = 56$. Update distance matrix $\mathbf{D}$ into $\mathbf{D}^{(1)}$.

\[ \mathbf{D}^{(1)}, \text{updated matrix of distances between one cluster and five remaining objects after Step 1 in the agglomeration process.} \]

<table>
<thead>
<tr>
<th></th>
<th>$A, C$</th>
<th>$B$</th>
<th>$D$</th>
<th>$E$</th>
<th>$F$</th>
</tr>
</thead>
<tbody>
<tr>
<td>${A, C}$</td>
<td>0</td>
<td>22</td>
<td>108</td>
<td>95.5</td>
<td>56</td>
</tr>
<tr>
<td>$B$</td>
<td>22</td>
<td>0</td>
<td>90</td>
<td>76</td>
<td>38</td>
</tr>
<tr>
<td>$D$</td>
<td>108</td>
<td>90</td>
<td>0</td>
<td>22</td>
<td>51</td>
</tr>
<tr>
<td>$E$</td>
<td>95.5</td>
<td>76</td>
<td>22</td>
<td>0</td>
<td>49</td>
</tr>
<tr>
<td>$F$</td>
<td>56</td>
<td>38</td>
<td>51</td>
<td>49</td>
<td>0</td>
</tr>
</tbody>
</table>

Step 2. Find the smallest distance in $\mathbf{D}^{(1)}$. There are two such values: $d_{B\{A,C\}} = 22$ and $d_{DE} = 22$.

Therefore, there are two options to form a new cluster:

**Option 1: Merge $D$ and $E$.**

Compute distances as above: $d_{D,E\{A,C\}} = (d_{AD} + d_{CD} + d_{AE} + d_{CE})/4 = (115 + 101 + 102 + 89)/4 = 101.75$, $d_{D,E,B} = (d_{BD} + d_{BE})/2 = (90 + 76)/2 = 83$, and $d_{D,E,F} = (d_{DF} + d_{EF})/2 = (51 + 49)/2 = 50$. Update $\mathbf{D}^{(1)}$ into $\mathbf{D}^{(2, \text{Option 1})}$.

---

**Step 3.2.** Find the smallest distance in $D_{(2, \text{Option 1})}$. This is $d_{B\{A,C\}} = 22$. Merge Object $B$ with Cluster $\{A,C\}$. Compute $d_{(D,E)\{A,B,C\}} = (d_{AD} + d_{BD} + d_{CD} + d_{AE} + d_{BE} + d_{CE})/6 = (115 + 90 + 101 + 102 + 76 + 89)/6 = 573/6 = 95.5$ and $d_{(A,B,C)F} = (d_{AF} + d_{BF} + d_{CF})/3 = (64 + 38 + 48)/3 = 50$.

Update $D_{(2, \text{Option 1})}$ into $D_{(3, \text{Option 1})}$.

**Step 4.2.** Find the smallest distance in $D_{(3, \text{Option 1})}$. There are two smallest distances: $d_{F\{A,B,C\}} = d_{F\{D,E\}} = 50$ and therefore again two options:

**Option 1.A: Merge Object $F$ with Cluster $\{A,B,C\}$.**

Now we have two clusters left: $\{A,B,C,F\}$ and $\{D,E\}$. The distance between those two clusters is


$$= (3 \times 2 \times 95.5 + 1 \times 2 \times 50)/(3 \times 2 + 1 \times 2) = 673/8 = 84.125$$

(this is an equivalent and simpler formula that only needs the latest updated matrix $D$ and the number of elements in each of the latest clusters).

The next and last fusion necessarily merges all objects into one cluster, with fusion coefficient 84.125.

**Option 1.B: Merge Object $F$ with Cluster $\{D,E\}$.**

Again, we have two clusters left: $\{A,B,C\}$ and $\{D,E,F\}$.


$$= (3 \times 2 \times 95.5 + 3 \times 1 \times 50)/(3 \times 2 + 3 \times 1) = 723/9 = 80.333$$

The next and last fusion necessarily merges all objects into one cluster, with fusion coefficient 80.333.
Option 2: Merge $B$ with $A$ and $C$.

Compute the distances between $\{A,B,C\}$ and the remaining objects and clusters:

\[
d_{D(A,B,C)} = (d_{AD} + d_{BD} + d_{CD})/3 = (115 + 90 + 101)/3 = 102; \quad d_{E(A,B,C)} = (d_{AE} + d_{BE} + d_{CE})/3 = (102 + 76 + 89)/3 = 89; \quad d_{F(A,B,C)} = (d_{AF} + d_{BF} + d_{CF})/3 = (64 + 38 + 48)/3 = 50.
\]

Update $D_{(1)}$ into $D_{(2,\text{Option 2})}$.

\[
D_{(2,\text{Option 2})}, \text{ updated matrix of distances between one cluster and three objects after Step 2 in the agglomeration process after choosing Option 1 of breaking two ties}
\]

<table>
<thead>
<tr>
<th></th>
<th>${A,B,C}$</th>
<th>$D$</th>
<th>$E$</th>
<th>$F$</th>
</tr>
</thead>
<tbody>
<tr>
<td>${A,B,C}$</td>
<td>0</td>
<td>102</td>
<td>89</td>
<td>50</td>
</tr>
<tr>
<td>$D$</td>
<td>102</td>
<td>0</td>
<td>22</td>
<td>51</td>
</tr>
<tr>
<td>$E$</td>
<td>89</td>
<td>22</td>
<td>0</td>
<td>49</td>
</tr>
<tr>
<td>$F$</td>
<td>50</td>
<td>51</td>
<td>49</td>
<td>0</td>
</tr>
</tbody>
</table>

Step 3.1. Find the smallest distance in $D_{(2,\text{Option 2})}$. This is $d_{DE} = 22$. Merge Objects $D$ and $E$.

Compute the distances between $\{D,E\}$ and the remaining objects and clusters:

\[
d_{D\{D,E\}}(A,B,C) = (d_{AD} + d_{BD} + d_{CD} + d_{AE} + d_{BE} + d_{CE})/6 = (115 + 90 + 101 + 102 + 76 + 89)/6 = 573/6 = 95.5 \quad \text{and} \quad d_{F\{D,E\}}(A,B,C) = (d_{DF} + d_{EF})/2 = (51 + 49)/2 = 50.
\]

Update $D_{(2,\text{Option 2})}$ into $D_{(3,\text{Option 2})}$.

\[
D_{(3,\text{Option 2})}, \text{ updated matrix of distances between two clusters and one object after Step 2 in the agglomeration process after choosing Option 1 of breaking two ties}
\]

<table>
<thead>
<tr>
<th></th>
<th>${A,B,C}$</th>
<th>${D,E}$</th>
<th>$F$</th>
</tr>
</thead>
<tbody>
<tr>
<td>${A,B,C}$</td>
<td>0</td>
<td>95.5</td>
<td>50</td>
</tr>
<tr>
<td>${D,E}$</td>
<td>95.5</td>
<td>0</td>
<td>50</td>
</tr>
<tr>
<td>$F$</td>
<td>50</td>
<td>50</td>
<td>0</td>
</tr>
</tbody>
</table>

Step 4.1. Find the smallest distance in $D_{(3,\text{Option 2})}$. There are two smallest distances: $d_{F\{A,B,C\}} = d_{F\{D,E\}} = 50$. Therefore, there are again two options:

**Option 2.A: Merge Object $F$ with Cluster $A,B,C$.**

We have the same clusters left as in Option 1.A: $\{A,B,C,F\}$ and $\{D,E\}$.

\[
d_{\{A,B,C,F\}\{D,E\}} \text{ = 84.125}. \text{The next and last fusion necessarily merges all objects into one cluster, with fusion coefficient 84.125.}
\]

**Option 2.B: Merge Object $F$ with Cluster $D,E$.**

We have the same clusters left as in Option 1.B: $\{A,B,C\}$ and $\{D,E,F\}$.

\[
d_{\{A,B,C\}\{D,E,F\}} \text{ = 80.333}. \text{The next and last fusion merges all objects into one cluster, with fusion coefficient 80.333.}
\]
The analysis of \( D \) thus may yield four different solutions. Which option is chosen, depends on the order in which the rows and columns of \( D \) are read into the computer program. Figure 1 shows the four trees that were obtained by SPSS after permuting the rows and columns of \( D \). The corresponding agglomeration schedules are listed in Table 1. The SPSS data matrix and the SPSS syntax are listed in Appendix A.

\[
\begin{array}{c}
\text{Solution 1} \\
\text{input order: ABCDEF} \\
\begin{array}{c}
A \\
C \\
B \\
F \\
D \\
E
\end{array}
\end{array} \\
\begin{array}{c}
\text{Solution 2} \\
\text{input order: FDEACB} \\
\begin{array}{c}
A \\
C \\
B \\
F \\
D \\
E
\end{array}
\end{array}
\]

\[
\begin{array}{c}
\text{Solution 3} \\
\text{input order: FABCDE} \\
\begin{array}{c}
A \\
C \\
B \\
F \\
D \\
E
\end{array}
\end{array} \\
\begin{array}{c}
\text{Solution 4} \\
\text{input order: FACDEB} \\
\begin{array}{c}
A \\
C \\
B \\
F \\
D \\
E
\end{array}
\end{array}
\]

Figure 1. Four dendrograms of the Objects A through F obtained by SPSS' Baverage analysis of the distance matrix \( D \) in four different input orders. The connections between the horizontal lines indicate the stages at which the clusters were formed; usually, they indicate the values of the fusion coefficients.
Table 1.
Agglomeration Schedules Corresponding to the Dendrograms of Figure 1.

1. Input Order **ABCDEF**

<table>
<thead>
<tr>
<th>Stage</th>
<th>Cluster Combined: Cluster 1</th>
<th>Cluster Combined: Cluster 2</th>
<th>Coefficients</th>
<th>Stage Cluster First Appears: Cluster 1</th>
<th>Stage Cluster First Appears: Cluster 2</th>
<th>Next Stage</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1 (A)</td>
<td>3 (C)</td>
<td>17.000</td>
<td>0</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>4 (D)</td>
<td>5 (E)</td>
<td>22.000</td>
<td>0</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>3</td>
<td>1 (A, C)</td>
<td>2 (B)</td>
<td>22.000</td>
<td>1</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>1 (A, B, C)</td>
<td>6 (F)</td>
<td>50.000</td>
<td>3</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>1 (A, B, C, F)</td>
<td>4 (D, E)</td>
<td>84.125</td>
<td>4</td>
<td>2</td>
<td>0</td>
</tr>
</tbody>
</table>

2. Input Order **FDEACB**

<table>
<thead>
<tr>
<th>Stage</th>
<th>Cluster Combined: Cluster 1</th>
<th>Cluster Combined: Cluster 2</th>
<th>Coefficients</th>
<th>Stage Cluster First Appears: Cluster 1</th>
<th>Stage Cluster First Appears: Cluster 2</th>
<th>Next Stage</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1 (A)</td>
<td>3 (C)</td>
<td>17.000</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>1 (A, C)</td>
<td>2 (B)</td>
<td>22.000</td>
<td>1</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>4 (D)</td>
<td>5 (E)</td>
<td>22.000</td>
<td>0</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>4</td>
<td>1 (A, B, C)</td>
<td>6 (F)</td>
<td>50.000</td>
<td>2</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>1 (A, B, C, F)</td>
<td>4 (D, E)</td>
<td>84.125</td>
<td>4</td>
<td>3</td>
<td>0</td>
</tr>
</tbody>
</table>

3. Input Order **FABCDE**

<table>
<thead>
<tr>
<th>Stage</th>
<th>Cluster Combined: Cluster 1</th>
<th>Cluster Combined: Cluster 2</th>
<th>Coefficients</th>
<th>Stage Cluster First Appears: Cluster 1</th>
<th>Stage Cluster First Appears: Cluster 2</th>
<th>Next Stage</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1 (A)</td>
<td>3 (C)</td>
<td>17.000</td>
<td>0</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>4 (D)</td>
<td>5 (E)</td>
<td>22.000</td>
<td>1</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>1 (A, C)</td>
<td>2 (B)</td>
<td>22.000</td>
<td>1</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>4</td>
<td>4 (D, E)</td>
<td>6 (F)</td>
<td>50.000</td>
<td>2</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>1 (A, B, C)</td>
<td>4 (D, E, F)</td>
<td>80.330</td>
<td>3</td>
<td>4</td>
<td>0</td>
</tr>
</tbody>
</table>

4. Input order **FACDEB**

<table>
<thead>
<tr>
<th>Stage</th>
<th>Cluster Combined: Cluster 1</th>
<th>Cluster Combined: Cluster 2</th>
<th>Coefficients</th>
<th>Stage Cluster First Appears: Cluster 1</th>
<th>Stage Cluster First Appears: Cluster 2</th>
<th>Next Stage</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1 (A)</td>
<td>3 (C)</td>
<td>17.000</td>
<td>0</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>1 (A, C)</td>
<td>2 (B)</td>
<td>22.000</td>
<td>1</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>4 (D)</td>
<td>5 (E)</td>
<td>22.000</td>
<td>0</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>4</td>
<td>4 (D, E)</td>
<td>6 (F)</td>
<td>50.000</td>
<td>3</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>1 (A, B, C)</td>
<td>4 (D, E, F)</td>
<td>80.330</td>
<td>2</td>
<td>4</td>
<td>0</td>
</tr>
</tbody>
</table>
Appendix A

SPSS data matrix

```
<table>
<thead>
<tr>
<th>rowtype</th>
<th>label</th>
<th>varname</th>
<th>var1</th>
<th>var2</th>
<th>var3</th>
<th>var4</th>
<th>var5</th>
<th>var6</th>
</tr>
</thead>
<tbody>
<tr>
<td>PROX</td>
<td>A</td>
<td>var1</td>
<td>.0000</td>
<td>.0000</td>
<td>.0000</td>
<td>17.000</td>
<td>115.000</td>
<td>102.000</td>
</tr>
<tr>
<td>PROX</td>
<td>B</td>
<td>var2</td>
<td>25.000</td>
<td>.0000</td>
<td>.0000</td>
<td>19.000</td>
<td>90.000</td>
<td>76.000</td>
</tr>
<tr>
<td>PROX</td>
<td>C</td>
<td>var3</td>
<td>17.000</td>
<td>19.000</td>
<td>.0000</td>
<td>.0000</td>
<td>101.000</td>
<td>69.000</td>
</tr>
<tr>
<td>PROX</td>
<td>D</td>
<td>var4</td>
<td>115.000</td>
<td>90.000</td>
<td>101.000</td>
<td>.0000</td>
<td>.0000</td>
<td>22.000</td>
</tr>
<tr>
<td>PROX</td>
<td>E</td>
<td>var5</td>
<td>102.000</td>
<td>76.000</td>
<td>89.000</td>
<td>22.000</td>
<td>.0000</td>
<td>.0000</td>
</tr>
<tr>
<td>PROX</td>
<td>F</td>
<td>var6</td>
<td>64.0000</td>
<td>38.000</td>
<td>48.000</td>
<td>51.000</td>
<td>46.000</td>
<td>.0000</td>
</tr>
</tbody>
</table>
```

SPSS syntax

```
* input order ABCDEF.
CLUSTER  var1 var2 var3 var4 var5 var6
   /METHOD BAVERAGE
   /ID=label
   /PRINT SCHEDULE
   /PLOT DENDROGRAM
   /matrix=in(*) .

* input order FDEACB.
CLUSTER  var6 var4 var5 var1 var3 var2
   /METHOD BAVERAGE
   /ID=label
   /PRINT SCHEDULE
   /PLOT DENDROGRAM
   /matrix=in(*) .

* input order FABCDE .
CLUSTER  var6 var1 var2 var3 var4 var5
   /METHOD BAVERAGE
   /ID=label
   /PRINT SCHEDULE
   /PLOT DENDROGRAM
   /matrix=in(*) .

* input order FACDEB .
CLUSTER  var6 var1 var3 var4 var5 var2
   /METHOD BAVERAGE
   /ID=label
   /PRINT SCHEDULE
   /PLOT DENDROGRAM
   /matrix=in(*) .
```