Hierarchical Clustering on Principal Components (HCPC)

LE RAY Guillaume
MOLTO Quentin
Students of AGROCAMPUS OUEST majored in applied statistics
Context

- **R**: A free, opensource software for statistics (1875 packages).
- **FactoMineR**: a R package, developed in Agrocampus-Ouest, dedicated to factorial analysis.
- The aim is to create a complementary tool to this package, **dedicated to clustering**, especially **after a factorial analysis**.
- Wide range of choices and uses, results, and graphical representations.
Clustering and factorial analysis

• Factorial analysis and hierarchical clustering are very complementary tools to explore data.

• Removing the last factors of a factorial analysis remove noise and makes the clustering robuster.

Analyses factorielles simples et multiples 4ème édition, Escofier, Pagès 2008
Program structure

Factorial analysis
PCA, MCA, MFA…

Hierarchical Clustering
Ward, Euclidean

Cutting the tree
partition

Consolidation
K-means

Description of clusters and factor maps

Consolidation
Description of clusters and factor maps
Statistic methods (1)

• **Hierarchical clustering:**
  – Function *agnes*
  – Euclidean distance
  – Ward criterion = \(d^2(i,j)x(mi.mj)/(mi+mj)\)

• **Suggested level to cut the tree:**
  – Intra-cluster inertia
  – Partition comparison: \(Q=(I_{n+1} - I_n)/I_{n+1}\)
  – Max = nb of individuals/2
Statistic methods (2)

Consolidation

- Non optimal partition
- K means with the cluster centers
Statistic methods (2)
Consolidation

- Non optimal partition
- K means with the cluster centers

Factorial analysis
Hierarchical clustering
Cutting the tree
Consolidation
Description of clusters and factor maps
Statistic methods (3)

Clusters description

• Description by individuals:
  – Use real individuals to characterise clusters.

• Description by variables:
  – Give list of typical variable of clusters.

• Description by axes:
  – Like in factorial analysis.
# Dataset presentation

<table>
<thead>
<tr>
<th>row.names</th>
<th>January</th>
<th>February</th>
<th>March</th>
<th>April</th>
<th>May</th>
<th>June</th>
<th>July</th>
<th>August</th>
<th>September</th>
<th>October</th>
<th>November</th>
<th>December</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amsterdam</td>
<td>2.9</td>
<td>2.5</td>
<td>5.7</td>
<td>8.2</td>
<td>12.5</td>
<td>14.8</td>
<td>17.1</td>
<td>17.1</td>
<td>14.5</td>
<td>11.4</td>
<td>7</td>
<td>4.4</td>
</tr>
<tr>
<td>Athens</td>
<td>9.1</td>
<td>9.7</td>
<td>11.7</td>
<td>15.4</td>
<td>20.1</td>
<td>24.5</td>
<td>27.4</td>
<td>27.2</td>
<td>23.8</td>
<td>19.2</td>
<td>14.6</td>
<td>11</td>
</tr>
<tr>
<td>Berlin</td>
<td>-0.2</td>
<td>0.1</td>
<td>4.4</td>
<td>8.2</td>
<td>13.8</td>
<td>16</td>
<td>18.3</td>
<td>18</td>
<td>14.4</td>
<td>10</td>
<td>4.2</td>
<td>1.2</td>
</tr>
<tr>
<td>Brussels</td>
<td>3.3</td>
<td>3.3</td>
<td>6.7</td>
<td>8.9</td>
<td>12.8</td>
<td>15.6</td>
<td>17.8</td>
<td>17.8</td>
<td>15</td>
<td>11.1</td>
<td>6.7</td>
<td>4.4</td>
</tr>
<tr>
<td>Budapest</td>
<td>-1.1</td>
<td>0.8</td>
<td>5.5</td>
<td>11.6</td>
<td>17</td>
<td>20.2</td>
<td>22</td>
<td>21.3</td>
<td>16.9</td>
<td>11.3</td>
<td>5.1</td>
<td>0.7</td>
</tr>
<tr>
<td>Copenhagen</td>
<td>-0.4</td>
<td>-0.4</td>
<td>1.3</td>
<td>5.8</td>
<td>11.1</td>
<td>15.4</td>
<td>17.1</td>
<td>16.6</td>
<td>13.3</td>
<td>8.8</td>
<td>4.1</td>
<td>1.3</td>
</tr>
<tr>
<td>Dublin</td>
<td>4.8</td>
<td>5</td>
<td>5.9</td>
<td>7.8</td>
<td>10.4</td>
<td>13.3</td>
<td>15</td>
<td>14.6</td>
<td>12.7</td>
<td>9.7</td>
<td>6.7</td>
<td>5.4</td>
</tr>
<tr>
<td>Helsinki</td>
<td>-5.8</td>
<td>-6.2</td>
<td>-2.7</td>
<td>3.1</td>
<td>10.2</td>
<td>14</td>
<td>17.2</td>
<td>14.9</td>
<td>9.7</td>
<td>5.2</td>
<td>0.1</td>
<td>-2.3</td>
</tr>
<tr>
<td>Kiev</td>
<td>-5.9</td>
<td>-5</td>
<td>-0.3</td>
<td>7.4</td>
<td>14.3</td>
<td>17.8</td>
<td>19.4</td>
<td>18.5</td>
<td>13.7</td>
<td>7.5</td>
<td>1.2</td>
<td>-3.6</td>
</tr>
<tr>
<td>Krakow</td>
<td>-3.7</td>
<td>-2</td>
<td>1.9</td>
<td>7.9</td>
<td>13.2</td>
<td>16.9</td>
<td>18.4</td>
<td>17.6</td>
<td>13.7</td>
<td>8.6</td>
<td>2.6</td>
<td>-1.7</td>
</tr>
<tr>
<td>Lisbon</td>
<td>10.5</td>
<td>11.3</td>
<td>12.8</td>
<td>14.5</td>
<td>16.7</td>
<td>19.4</td>
<td>21.5</td>
<td>21.9</td>
<td>20.4</td>
<td>17.4</td>
<td>13.7</td>
<td>11.1</td>
</tr>
<tr>
<td>London</td>
<td>3.4</td>
<td>4.2</td>
<td>5.5</td>
<td>8.3</td>
<td>11.9</td>
<td>15.1</td>
<td>16.9</td>
<td>16.5</td>
<td>14</td>
<td>10.2</td>
<td>6.3</td>
<td>4.4</td>
</tr>
<tr>
<td>Madrid</td>
<td>5</td>
<td>6.6</td>
<td>9.4</td>
<td>12.2</td>
<td>16</td>
<td>20.8</td>
<td>24.7</td>
<td>24.3</td>
<td>19.8</td>
<td>13.9</td>
<td>8.7</td>
<td>5.4</td>
</tr>
<tr>
<td>Minsk</td>
<td>-6.9</td>
<td>-6.2</td>
<td>-1.9</td>
<td>5.4</td>
<td>12.4</td>
<td>15.9</td>
<td>17.4</td>
<td>16.3</td>
<td>11.6</td>
<td>5.8</td>
<td>0.1</td>
<td>-4.2</td>
</tr>
<tr>
<td>Moscow</td>
<td>-9.3</td>
<td>-7.6</td>
<td>-2</td>
<td>6</td>
<td>13</td>
<td>16.6</td>
<td>18.3</td>
<td>16.7</td>
<td>11.2</td>
<td>5.1</td>
<td>-1.1</td>
<td>-6</td>
</tr>
<tr>
<td>Oslo</td>
<td>-4.3</td>
<td>-3.8</td>
<td>-0.6</td>
<td>4.4</td>
<td>10.3</td>
<td>14.9</td>
<td>16.9</td>
<td>15.4</td>
<td>11.1</td>
<td>5.7</td>
<td>0.5</td>
<td>-2.9</td>
</tr>
<tr>
<td>Paris</td>
<td>3.7</td>
<td>3.7</td>
<td>7.3</td>
<td>9.7</td>
<td>13.7</td>
<td>16.5</td>
<td>19</td>
<td>18.7</td>
<td>16.1</td>
<td>12.5</td>
<td>7.3</td>
<td>5.2</td>
</tr>
<tr>
<td>Prague</td>
<td>-1.3</td>
<td>0.2</td>
<td>3.6</td>
<td>8.8</td>
<td>14.3</td>
<td>17.6</td>
<td>19.3</td>
<td>18.7</td>
<td>14.9</td>
<td>9.4</td>
<td>3.8</td>
<td>0.3</td>
</tr>
<tr>
<td>Reykjavik</td>
<td>-0.3</td>
<td>0.1</td>
<td>0.8</td>
<td>2.9</td>
<td>6.5</td>
<td>9.3</td>
<td>11.1</td>
<td>10.6</td>
<td>7.9</td>
<td>4.5</td>
<td>1.7</td>
<td>0.2</td>
</tr>
<tr>
<td>Rome</td>
<td>7.1</td>
<td>8.2</td>
<td>10.5</td>
<td>13.7</td>
<td>17.8</td>
<td>21.7</td>
<td>24.4</td>
<td>24.1</td>
<td>20.9</td>
<td>16.5</td>
<td>11.7</td>
<td>8.3</td>
</tr>
<tr>
<td>Sarajevo</td>
<td>-1.4</td>
<td>0.8</td>
<td>4.9</td>
<td>9.3</td>
<td>13.8</td>
<td>17</td>
<td>18.9</td>
<td>18.7</td>
<td>15.2</td>
<td>10.5</td>
<td>5.1</td>
<td>0.8</td>
</tr>
<tr>
<td>Sofia</td>
<td>-1.7</td>
<td>0.2</td>
<td>4.3</td>
<td>9.7</td>
<td>14.3</td>
<td>17.7</td>
<td>20</td>
<td>19.5</td>
<td>15.8</td>
<td>10.7</td>
<td>5</td>
<td>0.6</td>
</tr>
<tr>
<td>Stockholm</td>
<td>-3.5</td>
<td>-3.5</td>
<td>-1.3</td>
<td>3.5</td>
<td>9.2</td>
<td>14.6</td>
<td>17.2</td>
<td>16</td>
<td>11.7</td>
<td>6.5</td>
<td>1.7</td>
<td>-1.6</td>
</tr>
</tbody>
</table>
Factorial Analysis

Dimension 1 (82.9%)

Dimension 2 (15.4%)

Factorial analysis

Hierarchical clustering

Cutting the tree

Consolidation

Description of clusters and factor maps
Hierarchical Clustering

Click to cut the tree

- Suggested level of cutting.

Option:
- Sort the individuals as on the first component.

Factorial analysis

Hierarchical clustering

Cutting the tree

Consolidation

Description of clusters and factor maps

Le Ray, Molto - Agrocampus-Ouest Students - Feb. 09 - "HCPC"
Hierarchical Clustering

Colored rectangles are drawn around the clusters. We keep the same color for each cluster in the next graphs (function rect).

Options:
- cut automatically the tree at the suggested level,
- Cut at level with a choosen number of clusters.

Factorial analysis

Hierarchical clustering

Cutting the tree

Consolidation

Description of clusters and factor maps
Factor map and clusters

Options:
• Draw other axes,
• Remove the names, the centers.

Le Ray, Molto - Agrocampus-Ouest Students - Feb. 09 - "HCPC"
Factor map, clusters, and tree

Options:
- Draw only a part of the tree,
- Draw other axes,
- Remove the names
- Change the height

Description of clusters and factor maps
Cluster description (1)
By individuals

Option: the number of individuals for each cluster (here 2)
Cluster description (1)

By individuals

Option: the number of individuals for each cluster (here 2)

Cluster 1:
- Oslo
- Stockholm

Cluster 2:
- Berlin
- Sarajevo

Cluster 3:
- Rome
- Lisbon

Factorial analysis:
Hierarchical clustering
Cutting the tree
Consolidation
Description of clusters and factor maps
Cluster description (2)
By individuals

Option: the number of individuals for each cluster (here 2)

factorial analysis
Hierarchical clustering
Cutting the tree
Consolidation
Description of clusters and factor maps
Cluster description (3)

By variables

This is the result of a catdes, it describes the different clusters by the variables (the mean in the category, the v.test...)

Option:
the p.value (here 0.05).

factorial analysis

Hierarchical clustering

Cutting the tree

Consolidation

Description of clusters and factor maps
Cluster description (3)
By axes

This is the result of a catdes, it describes the different clusters by the axes (the mean in the category, the v.test...) 

**Option:**
the p.value (here 0.05).

Le Ray, Molto - Agrocampus-Ouest Students - Feb. 09 - "HCPC"
Conclusion

This function was presented with a PCA, but it also accepts:

- MCA and MFA results,
- directly a quantitative dataset (non-scaled PCA),
- a continuous variables to divide into modalities.
Function plot.catdes

It is a graphical representation of the desc.var results

Option:
- show only the quantitative, qualitative variables or all

Le Ray, Molto - Agrocampus-Ouest Students - Feb. 09 - "HCPC"