Correspondence analysis and related methods

Middle East Technical University
Fall semester

Michael Greenacre
Universitat Pompeu Fabra

mmr.upf@gmail.com
www.econ.upf.edu/~michael
www.multivariatestatistics.org

www.globalsong.net
www.youtube.com/StatisticalSongs  ../CARMEnetwork  ../ArcticFrontiers

ADDITIONAL READING

  (Spanish edition available for free download from http://www.fbbva.es)
Website of book **Biplots in Practice**
published by the BBVA Foundation (Madrid) in September 2010 and available online for free download.
Includes R code for all the analyses in the book, as well as data sets, videos, a searchable glossary in English and Spanish, and chapter summaries in Spanish.

**IMPORTANT:**
Homework during the Bayram holiday: Read and study Chapters 1-3 of this book

**Biplot: the fundamental concept.**
**Generalized scatterplot**

[Diagram of scatterplot and biplot]
Some multivariate data

- Let’s start with some simple trivariate data...

**Continuous variables**
- $X_1$ – Purchasing power/capita (euros)
- $X_2$ – GDP/capita (index)
- $X_3$ – Inflation rate (%)

**Count variables**
- $C_1$ – Glance reader
- $C_2$ – Fairly thorough reader
- $C_3$ – Very thorough reader

<table>
<thead>
<tr>
<th>Education</th>
<th>C1</th>
<th>C2</th>
<th>C3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Some primary</td>
<td>E1</td>
<td>5</td>
<td>7</td>
</tr>
<tr>
<td>Primary completed</td>
<td>E2</td>
<td>18</td>
<td>46</td>
</tr>
<tr>
<td>Some secondary</td>
<td>E3</td>
<td>19</td>
<td>29</td>
</tr>
<tr>
<td>Secondary completed</td>
<td>E4</td>
<td>12</td>
<td>40</td>
</tr>
<tr>
<td>Some tertiary</td>
<td>E5</td>
<td>3</td>
<td>7</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Country</th>
<th>Purchasing power/capita (euros)</th>
<th>GDP/capita (index)</th>
<th>Inflation rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Be</td>
<td>19200</td>
<td>115.2</td>
<td>4.5</td>
</tr>
<tr>
<td>De</td>
<td>20400</td>
<td>120.1</td>
<td>3.6</td>
</tr>
<tr>
<td>Ge</td>
<td>19500</td>
<td>115.6</td>
<td>2.8</td>
</tr>
<tr>
<td>Gr</td>
<td>18800</td>
<td>94.3</td>
<td>4.2</td>
</tr>
<tr>
<td>Sp</td>
<td>17600</td>
<td>102.6</td>
<td>4.1</td>
</tr>
<tr>
<td>Fr</td>
<td>19600</td>
<td>108.0</td>
<td>3.2</td>
</tr>
<tr>
<td>Ir</td>
<td>20800</td>
<td>135.4</td>
<td>3.1</td>
</tr>
<tr>
<td>It</td>
<td>18200</td>
<td>101.8</td>
<td>3.5</td>
</tr>
<tr>
<td>Lu</td>
<td>28800</td>
<td>276.4</td>
<td>4.1</td>
</tr>
<tr>
<td>Ne</td>
<td>20400</td>
<td>134.0</td>
<td>2.2</td>
</tr>
<tr>
<td>Po</td>
<td>15000</td>
<td>76.0</td>
<td>2.7</td>
</tr>
<tr>
<td>UK</td>
<td>22600</td>
<td>116.2</td>
<td>3.6</td>
</tr>
</tbody>
</table>

Visualizing trivariate continuous data

**Continuous variables**
- $X_1$ – Purchasing power/capita (euros)
- $X_2$ – GDP/capita (index)
- $X_3$ – Inflation rate (%)
Visualizing trivariate continuous data

Continuous variables

$X_1$ – Purchasing power/capita (euros)
$X_2$ – GDP/capita (index)
$X_3$ – inflation rate (%)

<table>
<thead>
<tr>
<th>Country</th>
<th>$X_1$</th>
<th>$X_2$</th>
<th>$X_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Be</td>
<td>19200</td>
<td>115.2</td>
<td>4.5</td>
</tr>
<tr>
<td>De</td>
<td>20400</td>
<td>120.1</td>
<td>3.6</td>
</tr>
<tr>
<td>Ge</td>
<td>19500</td>
<td>115.6</td>
<td>2.8</td>
</tr>
<tr>
<td>Gr</td>
<td>18800</td>
<td>94.3</td>
<td>4.2</td>
</tr>
<tr>
<td>Sp</td>
<td>17600</td>
<td>102.6</td>
<td>4.1</td>
</tr>
<tr>
<td>Fr</td>
<td>19600</td>
<td>108.0</td>
<td>3.2</td>
</tr>
<tr>
<td>Ir</td>
<td>20800</td>
<td>135.4</td>
<td>3.1</td>
</tr>
<tr>
<td>It</td>
<td>18200</td>
<td>101.8</td>
<td>3.5</td>
</tr>
<tr>
<td>Lu</td>
<td>28800</td>
<td>276.4</td>
<td>4.1</td>
</tr>
<tr>
<td>Ne</td>
<td>20400</td>
<td>134.0</td>
<td>2.2</td>
</tr>
<tr>
<td>Po</td>
<td>15000</td>
<td>76.0</td>
<td>2.7</td>
</tr>
<tr>
<td>UK</td>
<td>22600</td>
<td>116.2</td>
<td>3.6</td>
</tr>
</tbody>
</table>
Continuous variables

X1 – Purchasing power/capita (euros)
X2 – GDP/capita (index)
X3 – inflation rate (%)

Visualizing trivariate continuous data

cor  X1  X2  X3
X1 1.000 0.929 0.243
X2 0.929 1.000 0.207
X3 0.243 0.207 1.000

This is almost a biplot!
Visualizing trivariate count data

Count variables

C1 – Glance reader
C2 – Fairly thorough reader
C3 – Very thorough reader

<table>
<thead>
<tr>
<th>Education</th>
<th>C1</th>
<th>C2</th>
<th>C3</th>
<th>Row Profiles</th>
</tr>
</thead>
<tbody>
<tr>
<td>Primary incomplete</td>
<td>E1</td>
<td>5</td>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td>Primary completed</td>
<td>E2</td>
<td>18</td>
<td>46</td>
<td>20</td>
</tr>
<tr>
<td>Secondary incomplete</td>
<td>E3</td>
<td>19</td>
<td>29</td>
<td>39</td>
</tr>
<tr>
<td>Secondary completed</td>
<td>E4</td>
<td>12</td>
<td>40</td>
<td>49</td>
</tr>
<tr>
<td>Some tertiary</td>
<td>E5</td>
<td>3</td>
<td>7</td>
<td>16</td>
</tr>
</tbody>
</table>

Row profiles

<table>
<thead>
<tr>
<th>Education</th>
<th>C1</th>
<th>C2</th>
<th>C3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Some primary</td>
<td>E1</td>
<td>.36</td>
<td>.50</td>
</tr>
<tr>
<td>Primary completed</td>
<td>E2</td>
<td>.21</td>
<td>.55</td>
</tr>
<tr>
<td>Some secondary</td>
<td>E3</td>
<td>.22</td>
<td>.33</td>
</tr>
<tr>
<td>Secondary completed</td>
<td>E4</td>
<td>.12</td>
<td>.40</td>
</tr>
<tr>
<td>Some tertiary</td>
<td>E5</td>
<td>.12</td>
<td>.27</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>C1</th>
<th>C2</th>
<th>C3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
Visualizing trivariate count data

Count variables

C1 – Glance reader
C2 – Fairly thorough reader
C3 – Very thorough reader

<table>
<thead>
<tr>
<th>Education</th>
<th>C1</th>
<th>C2</th>
<th>C3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Some primary</td>
<td>.36</td>
<td>.50</td>
<td>.14</td>
</tr>
<tr>
<td>Primary completed</td>
<td>.21</td>
<td>.55</td>
<td>.24</td>
</tr>
<tr>
<td>Some secondary</td>
<td>.22</td>
<td>.33</td>
<td>.45</td>
</tr>
<tr>
<td>Secondary completed</td>
<td>.12</td>
<td>.40</td>
<td>.49</td>
</tr>
<tr>
<td>Some tertiary</td>
<td>.12</td>
<td>.27</td>
<td>.62</td>
</tr>
</tbody>
</table>

row profiles

<table>
<thead>
<tr>
<th>C1</th>
<th>C2</th>
<th>C3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

This is almost a correspondence analysis!

And almost a correspondence analysis biplot if vectors drawn to the corners!

The biplot: The result of a simple matrix decomposition

\[
\begin{pmatrix}
8 & 2 & 2 & -6 \\
5 & 0 & 3 & -4 \\
-2 & -3 & 3 & 1 \\
2 & 3 & -3 & -1 \\
4 & 6 & -6 & -2
\end{pmatrix}
= \begin{pmatrix}
2 & 2 \\
1 & 2 \\
-1 & 1 \\
1 & -1 \\
2 & -2
\end{pmatrix}
\begin{pmatrix}
3 & 2 & -1 & -2 \\
1 & -1 & 2 & -1
\end{pmatrix}
\]

target matrix = left matrix \cdot right matrix

\[S = AB^T\]

(Note: we say that the target matrix is of rank 2)
The biplot: A graphical display of a matrix decomposition

\[
\begin{bmatrix}
8 & 2 & 2 & -6 \\
5 & 0 & 3 & -4 \\
-2 & -3 & 3 & 1 \\
2 & 3 & -3 & -1 \\
4 & 6 & -6 & -2
\end{bmatrix}
= \begin{bmatrix}
2 & 2 \\
1 & 2 \\
-1 & 1 \\
1 & -1 \\
2 & -2
\end{bmatrix}
\begin{bmatrix}
y_1 & y_2 & y_3 & y_4
\end{bmatrix}
\]

target matrix = left matrix \cdot right matrix

\[
\begin{bmatrix}
2 & 2 \\
2 & 1
\end{bmatrix}
\cdot
\begin{bmatrix}
3 & 2 & -1 & -2 \\
1 & -1 & 2 & -1
\end{bmatrix}
= x_1^T y_1 = 2 \cdot 3 + 2 \cdot 1 = 8
\]

(scalar product between two vectors)

Cases usually points. Variables usually vectors

Geometry of scalar products

\[
x^T y = \|x\| \|y\| \cos(\theta) = \|x\| \cos(\theta) \|y\|
\]

= projection of \(x\) onto \(y\) \times length of \(y\)
The biplot:
A graphical display of a matrix decomposition

\[
\begin{pmatrix}
8 & 2 & 2 & -6 \\
5 & 0 & 3 & -4 \\
-2 & -3 & 3 & 1 \\
2 & 3 & -3 & -1 \\
4 & 6 & -6 & -2 \\
\end{pmatrix}
\begin{pmatrix}
2 & 2 \\
1 & -2 \\
\end{pmatrix}
= \begin{pmatrix}
y_1 & y_2 & y_3 & y_4 \\
3 & 2 & -1 & -2 \\
1 & -1 & 2 & -1 \\
\end{pmatrix}
\]

target matrix = left matrix \cdot right matrix

\[
\begin{pmatrix}
2 \\
2 \\
\end{pmatrix}
\cdot \begin{pmatrix}
3 \\
1 \\
\end{pmatrix}
= x_1^\top y_1 = 2\times3 + 2\times1 = 8
\]

(scalar product between two vectors)

Projection of \( x_1 \) onto \( y_1 \) = 2.530 (angle between them is 26.57°)
Length of \( y_1 \) = \( \sqrt{10} \) = 3.162
2.530 \times 3.162 = 8.000  \checkmark
1/3.162 \times 3.162 = 1  \therefore  Calibration unit is 1/3.162 = 0.3162 (1/length of biplot vector)

Calibrated biplot

N.B. It's important to draw a biplot with aspect ratio equal to 1 ("asp=1" option in R)
Regression biplots: Data set “bioenv”

<table>
<thead>
<tr>
<th>SITE NO</th>
<th>SPECIES COUNTS</th>
<th>ENVIRONMENTAL VARS.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>a</td>
<td>b</td>
</tr>
<tr>
<td>s1</td>
<td>2</td>
<td>9</td>
</tr>
<tr>
<td>s2</td>
<td>26</td>
<td>4</td>
</tr>
<tr>
<td>s3</td>
<td>0</td>
<td>10</td>
</tr>
<tr>
<td>s4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>s5</td>
<td>13</td>
<td>5</td>
</tr>
<tr>
<td>s6</td>
<td>31</td>
<td>21</td>
</tr>
<tr>
<td>s7</td>
<td>9</td>
<td>6</td>
</tr>
<tr>
<td>s8</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>s9</td>
<td>17</td>
<td>7</td>
</tr>
<tr>
<td>s10</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>s11</td>
<td>0</td>
<td>8</td>
</tr>
<tr>
<td>s12</td>
<td>14</td>
<td>11</td>
</tr>
<tr>
<td>s13</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>s14</td>
<td>13</td>
<td>0</td>
</tr>
<tr>
<td>s15</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>s16</td>
<td>42</td>
<td>20</td>
</tr>
<tr>
<td>s17</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>s18</td>
<td>21</td>
<td>15</td>
</tr>
<tr>
<td>s19</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>s20</td>
<td>0</td>
<td>10</td>
</tr>
<tr>
<td>s21</td>
<td>8</td>
<td>0</td>
</tr>
<tr>
<td>s22</td>
<td>35</td>
<td>10</td>
</tr>
<tr>
<td>s23</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>s24</td>
<td>18</td>
<td>12</td>
</tr>
<tr>
<td>s25</td>
<td>32</td>
<td>26</td>
</tr>
<tr>
<td>s26</td>
<td>32</td>
<td>21</td>
</tr>
<tr>
<td>s27</td>
<td>24</td>
<td>17</td>
</tr>
<tr>
<td>s28</td>
<td>16</td>
<td>3</td>
</tr>
<tr>
<td>s29</td>
<td>11</td>
<td>0</td>
</tr>
<tr>
<td>s30</td>
<td>24</td>
<td>37</td>
</tr>
</tbody>
</table>

Simple linear regression
species $d$ versus pollution ($y$)

$$\hat{d} = 19.10 - 1.815y \quad (R^2 = 0.34)$$

![Regression biplot for species $d$ versus pollution ($y$)](image_url)
Multiple linear regression

\[ \hat{d} = 6.135 - 1.388y + 0.148x \]

\[ R^2 = 0.442 \]

Regression model is a (hyper)plane

<table>
<thead>
<tr>
<th>d</th>
<th>y</th>
<th>x</th>
</tr>
</thead>
<tbody>
<tr>
<td>14</td>
<td>4.8</td>
<td>72</td>
</tr>
<tr>
<td>11</td>
<td>2.8</td>
<td>75</td>
</tr>
<tr>
<td>8</td>
<td>5.4</td>
<td>59</td>
</tr>
<tr>
<td>3</td>
<td>8.2</td>
<td>64</td>
</tr>
<tr>
<td>10</td>
<td>3.9</td>
<td>61</td>
</tr>
<tr>
<td>16</td>
<td>2.6</td>
<td>94</td>
</tr>
<tr>
<td>11</td>
<td>4.6</td>
<td>53</td>
</tr>
<tr>
<td>0</td>
<td>5.1</td>
<td>61</td>
</tr>
<tr>
<td>14</td>
<td>3.9</td>
<td>68</td>
</tr>
<tr>
<td>9</td>
<td>10.0</td>
<td>69</td>
</tr>
<tr>
<td>6</td>
<td>6.5</td>
<td>57</td>
</tr>
<tr>
<td>15</td>
<td>3.8</td>
<td>84</td>
</tr>
<tr>
<td>0</td>
<td>9.4</td>
<td>53</td>
</tr>
<tr>
<td>9</td>
<td>4.7</td>
<td>83</td>
</tr>
<tr>
<td>12</td>
<td>6.7</td>
<td>100</td>
</tr>
<tr>
<td>3</td>
<td>2.8</td>
<td>84</td>
</tr>
<tr>
<td>0</td>
<td>6.4</td>
<td>96</td>
</tr>
<tr>
<td>20</td>
<td>4.4</td>
<td>74</td>
</tr>
<tr>
<td>16</td>
<td>3.1</td>
<td>79</td>
</tr>
<tr>
<td>9</td>
<td>5.6</td>
<td>73</td>
</tr>
<tr>
<td>4</td>
<td>4.3</td>
<td>59</td>
</tr>
<tr>
<td>9</td>
<td>1.9</td>
<td>54</td>
</tr>
<tr>
<td>17</td>
<td>2.4</td>
<td>95</td>
</tr>
<tr>
<td>7</td>
<td>4.3</td>
<td>64</td>
</tr>
<tr>
<td>23</td>
<td>2.0</td>
<td>97</td>
</tr>
<tr>
<td>10</td>
<td>2.5</td>
<td>78</td>
</tr>
<tr>
<td>25</td>
<td>2.1</td>
<td>85</td>
</tr>
<tr>
<td>20</td>
<td>3.4</td>
<td>92</td>
</tr>
<tr>
<td>8</td>
<td>6.0</td>
<td>51</td>
</tr>
<tr>
<td>18</td>
<td>1.9</td>
<td>99</td>
</tr>
</tbody>
</table>

Multiple linear regression, variables standardized

\[ \hat{d}^* = -0.446y^* + 0.347x^* \]

\[ R^2 = 0.442 \]

Explanatory variables \( x \) and \( y \) and response variable \( d \) standardized
Another geometry of regression & prediction

\[ \hat{d}^* = -0.446y^* + 0.347x^* \]

Regression biplot

Variance explained \((R^2)\): 44.2%

Variance explained:
\[ a: 52.9\% \]
\[ b: 39.1\% \]
\[ c: 21.8\% \]
\[ d: 44.2\% \]
\[ e: 23.5\% \]

Overall:
41.5%

Significance:
\[ a: ** \]
\[ b: ** \]
\[ c: * \]
\[ d: ** * \]
\[ e: * * \]

\(*=p<0.05 \quad **=p<0.01\)
Regression is a matrix decomposition

For $d$, regression model was:

$$\hat{d}^* = -0.446y^* + 0.347x^*$$

For all five variables, the five regression models can be written as:

$$\begin{bmatrix} \hat{a}^* & \hat{b}^* & \hat{c}^* & \hat{d}^* & \hat{e}^* \end{bmatrix} = \begin{bmatrix} y^* & x^* \end{bmatrix} \begin{bmatrix} -0.727 & -0.449 & 0.491 & -0.446 & -0.475 \\ 0.000 & 0.229 & 0.074 & 0.347 & -0.400 \end{bmatrix}$$

$$\hat{S} = UB^T$$

Target matrix: the predicted values from the regression models

Left matrix: the explanatory variables (fixed!)

Right matrix: the regression coefficients

$$\begin{bmatrix} a^* & b^* & c^* & d^* & e^* \end{bmatrix} \rightarrow S = UB^T + E$$

$$B^T = (U^TU)^{-1}U^TS$$

Classic notation:

$$y = X\beta + e$$

$$\hat{\beta} = (X^TX)^{-1}X^Ty$$

$$\hat{y} = X\hat{\beta}$$

What happens for three predictors?

- Each regression model can be represented as a point vector in three-dimensional space.
- Reconstruct the data from projections of cases onto variable directions, but only as well as measured by $R^2$; in this example the increase in explained variance from two-dimensional to three-dimensional (adding temperature as an explanatory variable) is from 41.5% to 42.7%, hence temperature is explaining very little extra variance.

There will be a particular orientation of the vectors that gives maximum variance explained in the two-dimensional projection...

Dimension reduction coming...
Generalized linear model biplots

e.g., logistic regression

\[ \logit(p_d) = \log \left( \frac{p_d}{1 - p_d} \right) = 2.712 - 1.177y^* - 0.137x^* \]

<table>
<thead>
<tr>
<th>SITE</th>
<th>SPECIES COUNTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>NO.</td>
<td>a</td>
</tr>
<tr>
<td>s1</td>
<td>0</td>
</tr>
<tr>
<td>s2</td>
<td>26</td>
</tr>
<tr>
<td>s3</td>
<td>0</td>
</tr>
<tr>
<td>s4</td>
<td>0</td>
</tr>
<tr>
<td>s5</td>
<td>13</td>
</tr>
<tr>
<td>s6</td>
<td>31</td>
</tr>
<tr>
<td>s7</td>
<td>9</td>
</tr>
<tr>
<td>s8</td>
<td>2</td>
</tr>
<tr>
<td>s9</td>
<td>17</td>
</tr>
<tr>
<td>s10</td>
<td>0</td>
</tr>
<tr>
<td>s11</td>
<td>0</td>
</tr>
<tr>
<td>s12</td>
<td>14</td>
</tr>
<tr>
<td>s13</td>
<td>0</td>
</tr>
</tbody>
</table>

Transform to presence/absence (0/1) data

Logistic regression biplot

\[ \logit(p_d) = \log \left( \frac{p_d}{1 - p_d} \right) = 2.712 - 1.177y^* - 0.137x^* \]

Error deviances:

- a: 0.464
- b: 0.756
- c: 0.911
- d: 0.798
- e: 0.832

Similarly for Poisson regression and any generalized linear model...