II — Basic Notions of Geometric Data Analysis

Adapted from the course “Statistiques avancées et modélisation”, by Brigitte Le Roux at research master “Politique et sociétés en Europe” (Fondation Nationale des Sciences Politiques, Paris, 2006).

Brigitte.LeRoux@math-info.univ-paris5.fr
rouanet@math-info.univ-paris5.fr

www.math-info.univ-paris5.fr/~lerb/
www.math-info.univ-paris5.fr/~rouanet/
1 Euclidean Cloud

1.1 Target Example

Figure 1. Target Example (10 points)
Figure 1bis: Cloud with origin-point (point O) and initial axes

Initial coordinates of points

<table>
<thead>
<tr>
<th></th>
<th>$x_1$</th>
<th>$x_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>i1</td>
<td>0</td>
<td>-12</td>
</tr>
<tr>
<td>i2</td>
<td>6</td>
<td>-10</td>
</tr>
<tr>
<td>i3</td>
<td>14</td>
<td>-6</td>
</tr>
<tr>
<td>i4</td>
<td>6</td>
<td>-2</td>
</tr>
<tr>
<td>i5</td>
<td>12</td>
<td>0</td>
</tr>
<tr>
<td>i6</td>
<td>-8</td>
<td>2</td>
</tr>
<tr>
<td>i7</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>i8</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>i9</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>i10</td>
<td>12</td>
<td>10</td>
</tr>
</tbody>
</table>
1.2 Deviations between points

The deviation of point M from point P is the vector $\overrightarrow{PM}$.

Deviations add up vectorially: $\overrightarrow{PM} + \overrightarrow{PN} = \overrightarrow{PQ}$

(parallelogram rule)
1.3 Mean point

The **mean point** of a cloud is the point $G$ such that the sum of deviations of the points from $G$ is the null vector (barycentric property).

$$\sum_{i=1}^{10} \vec{GM}_i = \vec{0}$$

**Property:** The coordinates of the mean point are the means of the coordinates.

$$\bar{x}_1 = 6 \text{ and } \bar{x}_2 = 0$$

Figure 2: Mean point and Barycentric Property
1.4 Subclouds

A subset of a cloud defines a subcloud.

\( A \): subcloud of 2 points (dipole)
\[ \{ i1, i2 \} \]

\( B \): subcloud of 1 point
\[ \{ i6 \} \]

\( C \): subcloud of 7 points
\[ \{ i3, i4, i5, i7, i8, i9, i10 \} \]

Figure 3: Subclouds
\( A, B, C \) are the mean points of subclouds \( A, B, C \); their respective weights are 2, 1, 7.

By grouping:
— points “average up”
— weights add up

<table>
<thead>
<tr>
<th>Coordinates</th>
<th>weights</th>
</tr>
</thead>
<tbody>
<tr>
<td>( x_1 )</td>
<td>( x_2 )</td>
</tr>
<tr>
<td>A</td>
<td>3</td>
</tr>
<tr>
<td>B</td>
<td>-8</td>
</tr>
<tr>
<td>C</td>
<td>8.857</td>
</tr>
</tbody>
</table>

\[ \bar{x}_1 = 6 \quad \bar{x}_2 = 0 \quad n = 10 \]
1.5 Between-cloud

Partition of the cloud in 3 classes: the 3 subclouds $A$, $B$ and $C$.

The mean points $(A,2)$, $(B,1)$ et $(C,7)$ of the three subclouds define a derived cloud called the between-cloud associated with the partition.

The between-cloud is a weighted cloud.

Its total weight is $n = 10$; its mean point is $G$.

Barycentric property for the between-cloud:

$$2GA + 1GB + 7GC = \vec{0}$$
1.6 Variance

The **variance** of a Euclidean cloud is the mean of the squares of distances of the points from the mean point.

Squares of distances from the mean point:

\[(GM^i_1)^2 = (0 - 6)^2 + (-12 - 0)^2 = 180;\]
\[(GM^i_2)^2 = 100; \ (GM^i_3)^2 = 100; \ (GM^i_4)^2 = 4;\]
\[(GM^i_5)^2 = 36; \ (GM^i_6)^2 = 200; \ (GM^i_7)^2 = 32;\]
\[(GM^i_8)^2 = 16; \ (GM^i_9)^2 = 116; \ (GM^i_{10})^2 = 136.\]

- Variance of the elementary cloud of 10 points (**total variance**):

\[
\frac{1}{10} \left( GM^i_1 \right)^2 + \frac{1}{10} \left( GM^i_2 \right)^2 + \cdots + \frac{1}{10} \left( GM^i_{10} \right)^2
\]
\[
= \frac{1}{10} \times 180 + \frac{1}{10} \times 100 + \cdots + \frac{1}{10} \times 136
\]
\[
= 92
\]
• Variance of weighted between-cloud (between-variance):
\[
\frac{n_A}{n}(GA)^2 + \frac{n_B}{n}(GB)^2 + \frac{n_C}{n}(GC)^2 = \frac{2}{10} \times 130 + \frac{1}{10} \times 200 + \frac{7}{10} \times 16.29
\]
\[
= 26 + 20 + 11.4 = 57.4
\]

• Variances of subclouds
\[\mathcal{A}: \quad 10 = \frac{1}{2}(AM^{i1})^2 + \frac{1}{2}(AM^{i2})^2\]
\[\mathcal{B}: \quad 0\]
\[\mathcal{C}: \quad 46.57 = \frac{1}{7}(CM^{i3})^2 + \frac{1}{7}(CM^{i4})^2 + \frac{1}{7}(CM^{i5})^2 + \frac{1}{7}(CM^{i7})^2 + \frac{1}{7}(CM^{i8})^2 + \frac{1}{7}(CM^{i9})^2 + \frac{1}{7}(CM^{i10})^2\]
1.7 Contributions

- Absolute contribution (Cta) = part of variance

Examples:

Absolute contribution of i1: \( \frac{1}{10} (GM^{i1})^2 = \frac{1}{10} \times 180 = 18 \)

Absolute contribution of C: \( \frac{7}{10} (GC)^2 = \frac{7}{10} \times 16.29 = 11.4 \)

- Relative contribution (Ctr) = proportion of variance

\[
= \frac{\text{Absolute contribution}}{\text{Variance}}
\]

Examples:

Relative contribution of i1 to total variance: \( \frac{18}{92} = 0.196 \ (19.6\%) \)

Relative contribution of point C to between-variance: \( \frac{11.4}{57.4} = 19.9\% \)

to total variance: \( \frac{11.4}{92} = 12.4\% \).
1.8 Contributions of a subcloud

The absolute contribution of a subcloud is the sum of the absolute contributions of its points.

— Example: absolute contribution of subcloud $C$.

$$\frac{1}{10}(GM^3)^2 + \cdots + \frac{1}{10}(GM^5)^2 + \frac{1}{10}(GM^7)^2 + \cdots + \frac{1}{10}(GM^{10})^2 = 10 + 0.4 + 3.6 + 3.2 + 1.6 + 11.6 + 13.6 = 44$$

The absolute within-contribution of a subcloud is the product of its weight by its variance.

— Example: absolute within-contribution of subcloud $C$.

$$\frac{7}{10} \times 46.57 = 32.6$$
**Huyghens theorem**: The contribution of a subcloud to the total variance is the sum of the contribution of its mean point and of its within-contribution.

**Example.** Subcloud $C : 11.4 + 32.6 = 44.$

![Figure 5: Huyghens theorem](image)
1.9 Between-within decomposition of variance

<table>
<thead>
<tr>
<th></th>
<th>Absolute contributions</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>mean points within subclouds</td>
<td></td>
</tr>
<tr>
<td>(A)</td>
<td>26.0 2.0 28</td>
<td></td>
</tr>
<tr>
<td>(B)</td>
<td>20.0 0 20</td>
<td></td>
</tr>
<tr>
<td>(C)</td>
<td>11.4 32.6 44</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>57.4 34.6 92</td>
<td></td>
</tr>
<tr>
<td>Variance</td>
<td>between within total</td>
<td></td>
</tr>
</tbody>
</table>

Within-variance = sum of within-contributions \((2.0 + 0 + 32.6)\)

= weighted mean of variances of subclouds \(\left(\frac{2}{10} \times 10 + 0 + \frac{7}{10} \times 46.6\right)\)

= 34.6

Total variance = between-variance + within-variance

\[\eta^2 = \frac{\text{between-variance}}{\text{total variance}}\] (eta-square)
Subcloud of 2 points (dipole)

A and B weighted by $n_A = 2$ and $n_B = 1$ with mean point $G'$.

Weight of dipole: $\hat{n}_{AB} = 1/(\frac{1}{n_A} + \frac{1}{n_B})$

Absolute contribution of the dipole: $p \times d^2$

with $p = \frac{n_{AB}}{n}$ (relative weight) and $d^2 = AB^2$

(square of the deviation).

Example: dipole $\{A, B\}$.

$AB^2 = 290$

$\hat{n}_{AB} = \frac{1}{\frac{1}{2} + \frac{1}{2}} = 2/3, p = \frac{2/3}{10} = 0.06667$

Absolute contribution: $0.06667 \times 290 = 19.33$

Property: The absolute contribution of a dipole is the absolute contribution of the subcloud of its two points.
2 Principal axes of a cloud

2.1 Projection of a cloud

$P'$ projection of point $P$ onto $\mathcal{D}$ along $\mathcal{D}'$. 
Midpoint property: The midpoint is preserved by projection

Mean point property: The mean point of a cloud is preserved by projection.
2.1.1 Orthogonal projection

The orthogonal projection of point \( P \) onto \( \mathcal{D} \) is point \( P' \) such that \( PP' \) is perpendicular to \( \mathcal{D} \).

Contracting property: The orthogonal projection contracts distances \((P'Q' \leq PQ)\).

Consequence: variance of projected cloud \( \leq \) variance of initial cloud
2.2 Projected clouds on several directions

Orthogonal additive decomposition: the sum of the variances of projected clouds onto perpendicular directions is the variance of initial cloud: $40 + 52 = 92$. 

\[
\begin{align*}
\text{Projection onto horizontal line:} & \quad \text{Projection onto vertical line:} \\
\text{variance}=40 & \quad \text{variance}=52
\end{align*}
\]
Projection onto skew line (60 degrees): variance = 55.9
2.3 Principal Axes

The oriented line for which the variance of the projected cloud is maximum is the first principal axis $A_1$.

Projected cloud on $A_1$: first principal cloud, best fit of the cloud by a unidimensional cloud, in the sense of orthogonal least squares.

Here: $\alpha = 63^\circ$ and variance (eigenvalue) $\lambda_1 = 56$. 
Principal lines and principal clouds (empty circles)
Representation of the cloud in the principal plane.
2.4 Results of analysis

$\lambda_1 = 56$ (variance of axis, eigenvalue); Variance rate: $\lambda_1$ divided by total variance: $\frac{56}{92} = 61\%$

<table>
<thead>
<tr>
<th>Results for Axis 1 ($\lambda_1 = 56$)</th>
<th>Results for Axis 2 ($\lambda_2 = 36$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Co-</td>
<td>Ctr (%)</td>
</tr>
<tr>
<td>ordinates</td>
<td></td>
</tr>
<tr>
<td>$i1$</td>
<td>-13.41</td>
</tr>
<tr>
<td>$i2$</td>
<td>-8.94</td>
</tr>
<tr>
<td>$i3$</td>
<td>-1.79</td>
</tr>
<tr>
<td>$i4$</td>
<td>-1.79</td>
</tr>
<tr>
<td>$i5$</td>
<td>+2.68</td>
</tr>
<tr>
<td>$i6$</td>
<td>-4.47</td>
</tr>
<tr>
<td>$i7$</td>
<td>+1.79</td>
</tr>
<tr>
<td>$i8$</td>
<td>+3.58</td>
</tr>
<tr>
<td>$i9$</td>
<td>+10.73</td>
</tr>
<tr>
<td>$i10$</td>
<td>+11.63</td>
</tr>
</tbody>
</table>
• **Absolute contribution of point to axis**: weight×square of coordinate

Example. For \( i1 \): weight= 1/10, coordinate= −13.41,  
\( \text{Cta} = \frac{1}{10} \times (-13.41)^2 = 17.96. \)

• **Relative contribution to axis**: absolute contribution divided by variance of axis.

Example. For \( i1 \), \( \text{Ctr} = 17.96/56 = 32.1\% \).

• **Quality of representation of point on axis**: \( \cos^2 \theta = \frac{GP^2}{GM^2} \)

Example: for \( i2 \), \( \cos^2 \theta = \frac{(-8.94)^2}{100} = 0.80 \)

**Reconstitution of distances:**

\[
d^2(i1,i2) = (-13.4 + 8.9)^2 + (0 - 4.47)^2 = 4.23 = (6.3)^2
\]
SUMMARY of FORMULAS

- Elementary Euclidean distance between points M and M’ with coordinates \((x_1, x_2) (x'_1, x'_2)\) :
  \[\text{MM} = \sqrt{(x_1 - x'_1)^2 + (x_2 - x'_2)^2}\]

- Variance of abscissas: \(v_1 = \sum n_i(x^i_1 - \overline{x_1})^2/n : v_1 = 40\)
  Variance of ordinates: \(v_2 = \sum n_i(x^i_2 - \overline{x_2})^2/n : v_2 = 52\)
  Covariance: \(c = \sum n_i(x^i_1 - \overline{x_1})(x^i_2 - \overline{x_2})/n : c = 8\)

- Equation for eigenvalues: \(\lambda^2 - (v_1 + v_2)\lambda + v_1v_2 - c^2 = 0\)
  Variance of the first principal axis: \(\lambda_1 = \frac{v_1 + v_2}{2} + \frac{1}{2} \sqrt{(v_1 - v_2)^2 + 4c^2}\)
  Example: \(\lambda_1 = \frac{40 + 52}{2} + \frac{1}{2} \sqrt{(40 - 52)^2 + 4 \times 8^2} = 56\)

- Angle \(\alpha\) for a principal axis: \(\tan \alpha = (\lambda - v)/c\).
  Example: \(\tan \alpha = \frac{56 - 40}{8} = 2\). Equation of axis 1: \(x_2 - 0 = 2(x_1 - 6)\).

- Principal coordinate of point \(M^i\): \(y^i = (x^i_1 - \overline{x_1}) \cos \alpha + (x^i_2 - \overline{x_2}) \sin \alpha\)
References


