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 www.math-info.univ-paris5.fr/~lerb/
rouanet@math-info.univ-paris5.fr www.math-info.univ-paris5.fr/~rouanet/

1 Euclidean Cloud

1.1 Target Example

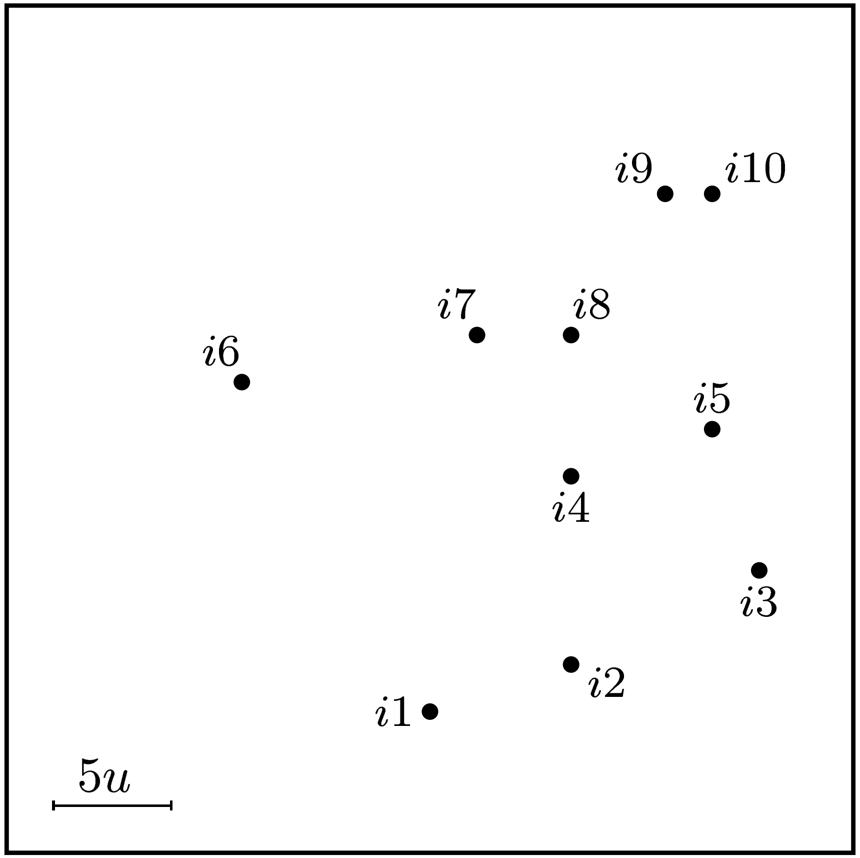


Figure 1. Target Example (10 points)

	$i1$	$i2$	$i3$	$i4$	$i5$	$i6$	$i7$	$i8$	$i9$
$i2$	6.3								
$i3$	15.2	8.9							
$i4$	11.7	8.0	8.9						
$i5$	17.0	11.7	6.3	6.3					
$i6$	16.1	18.4	23.4	14.6	20.1				
$i7$	16.1	14.6	15.6	7.2	10.8	10.2			
$i8$	17.1	14.0	12.8	6.0	7.2	14.1	4.0		
$i9$	24.2	20.4	16.5	12.7	10.2	19.7	10.0	7.2	
$i10$	25.1	20.9	16.1	13.4	10.0	21.5	11.7	8.5	2.0

Table of distances

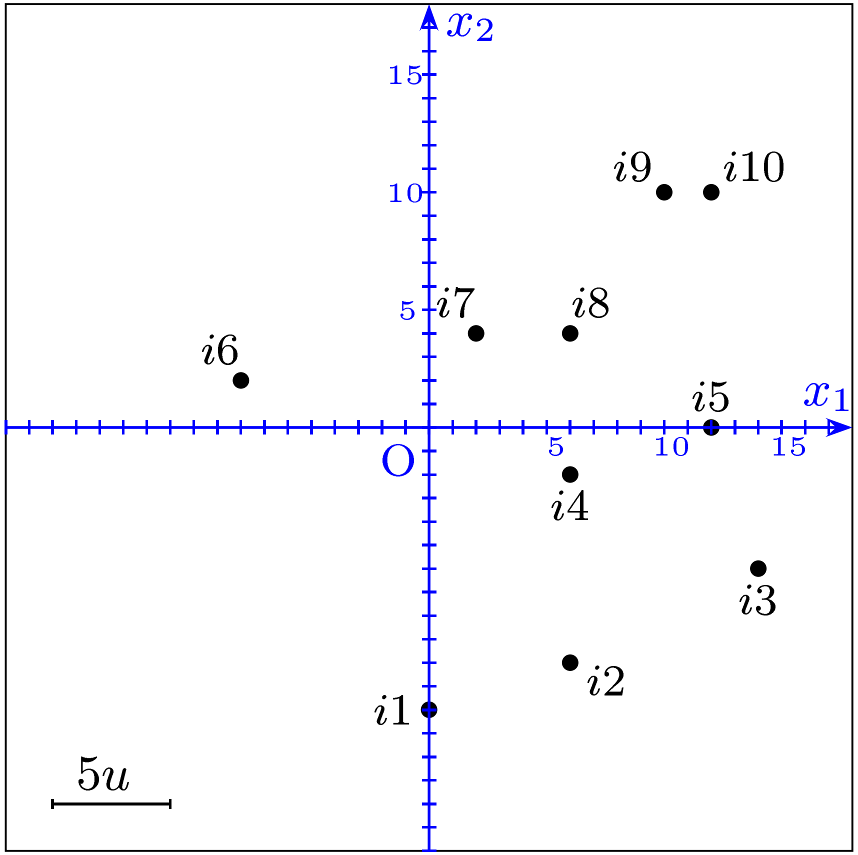


Figure 1bis: Cloud with origin-point (point O) and initial axes

	x_1	x_2
$i1$	0	−12
$i2$	6	−10
$i3$	14	−6
$i4$	6	−2
$i5$	12	0
$i6$	−8	2
$i7$	2	4
$i8$	6	4
$i9$	10	10
$i10$	12	10

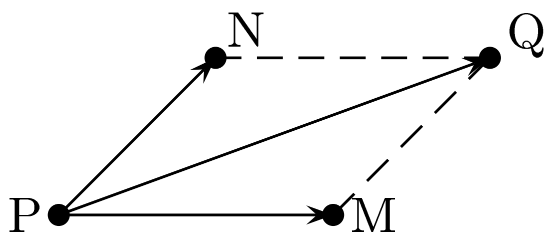
Initial coordinates of points

1.2 Deviations between points

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

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

(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).

$$\overrightarrow{\mathrm{GM}}^{i1} + \overrightarrow{\mathrm{GM}}^{i2} + \cdots \overrightarrow{\mathrm{GM}}^{i10} = \overrightarrow{0}$$

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

$$\overline{x_1} = 6 \text{ and } \overline{x_2} = 0$$

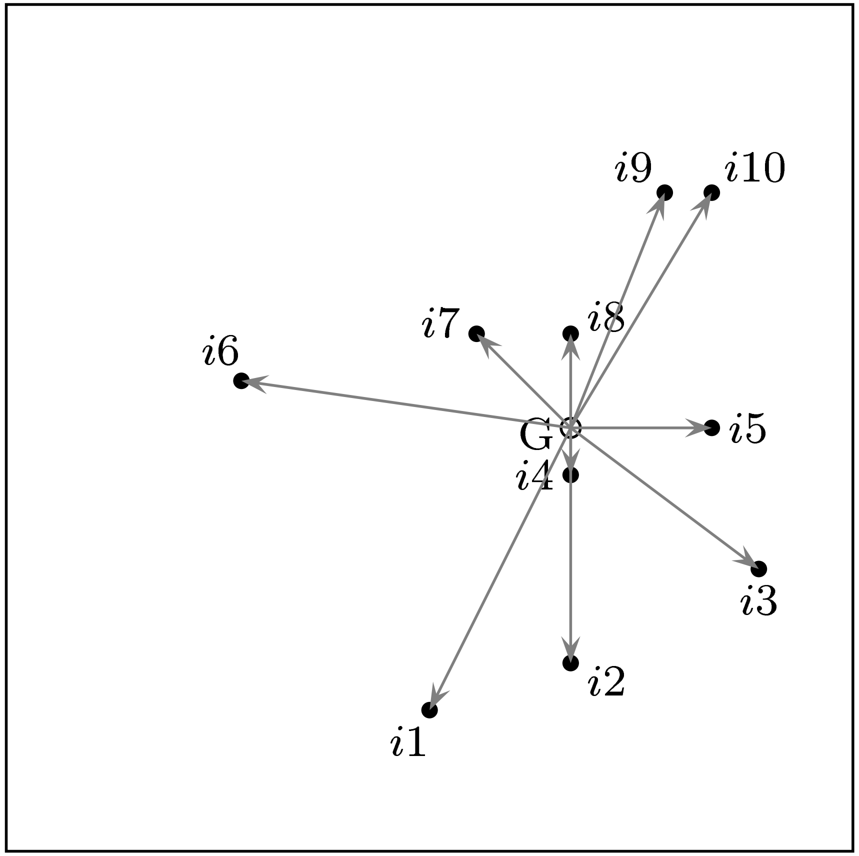


Figure 2: Mean point and Barycentric Property

1.4 Subclouds

A subset of a cloud defines a subcloud.

$\mathcal{A}$: subcloud of 2 points (dipole)

$\{i1, i2\}$

$\mathcal{B}$: subcloud of 1 point

$\{i6\}$

$\mathcal{C}$: subcloud of 7 points

$\{i3, i4, i5, i7, i8, i9, i10\}$

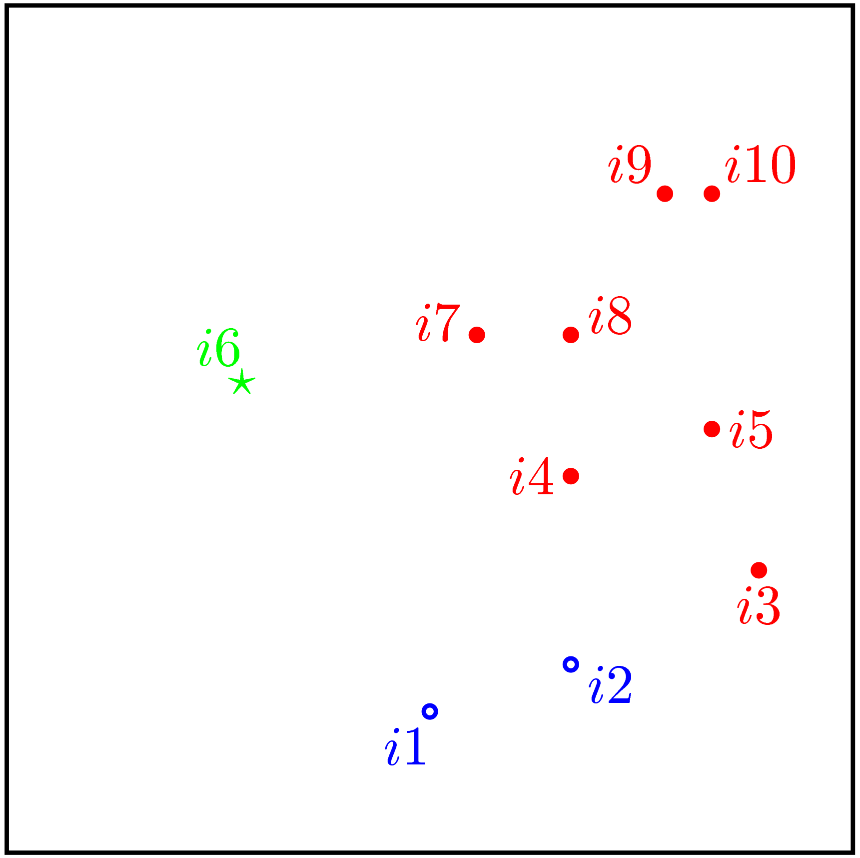


Figure 3: Subclouds

A, B, C are the mean points of subclouds $\mathcal{A}$, $\mathcal{B}$, $\mathcal{C}$;

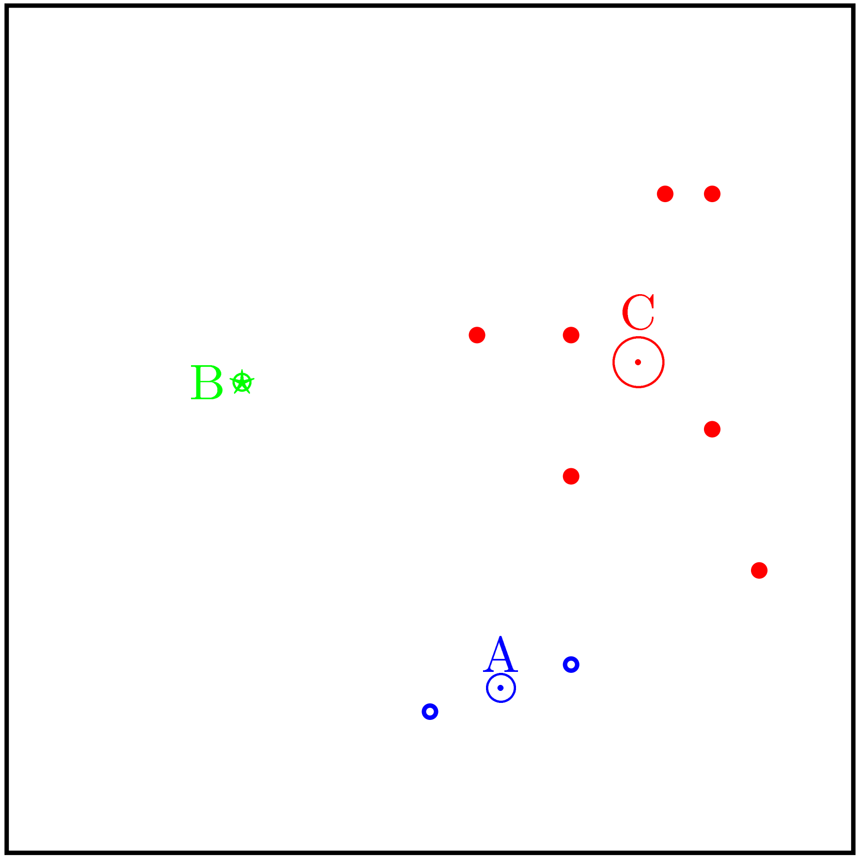
their respective weights are 2, 1, 7.

By grouping:

— points "average up"

— weights add up

	Coordinates x_1	x_2	weights
A	3	-11	$n_A = 2$
B	-8	2	$n_B = 1$
C	8.857	2.857	$n_C = 7$
	$\overline{x_1} = 6$	$\overline{x_2} = 0$	$n = 10$



1.5 Between-cloud

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

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:

$$2\overrightarrow{\mathrm{GA}} + 1\overrightarrow{\mathrm{GB}} + 7\overrightarrow{\mathrm{GC}} = \overrightarrow{0}$$

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:

$(\mathrm{GM}^{i1})^2 = (0-6)^2 + (-12-0)^2 = 180;$

$(\mathrm{GM}^{i2})^2 = 100;\ (\mathrm{GM}^{i3})^2 = 100;\ (\mathrm{GM}^{i4})^2 = 4;$

$(\mathrm{GM}^{i5})^2 = 36;\ (\mathrm{GM}^{i6})^2 = 200;\ (\mathrm{GM}^{i7})^2 = 32;$

$(\mathrm{GM}^{i8})^2 = 16;\ (\mathrm{GM}^{i9})^2 = 116;\ (\mathrm{GM}^{i10})^2 = 136.$

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

$$\begin{aligned} &\tfrac{1}{10}(\mathrm{GM}^{i1})^2 + \tfrac{1}{10}(\mathrm{GM}^{i2})^2 + \cdots + \tfrac{1}{10}(\mathrm{GM}^{i10})^2 \\ &\qquad = \tfrac{1}{10} \times 180 + \tfrac{1}{10} \times 100 + \cdots \tfrac{1}{10} \times 136 \\ &\qquad = 92 \end{aligned}$$

• Variance of weighted between-cloud (*between-variance*):

$$\frac{n_A}{n}(\mathrm{GA})^2 + \frac{n_B}{n}(\mathrm{GB})^2 + \frac{n_C}{n}(\mathrm{GC})^2 = \frac{2}{10} \times 130 + \frac{1}{10} \times 200 + \frac{7}{10} \times 16.29$$

$$= 26 + 20 + 11.4 = 57.4$$

• Variances of subclouds

$\mathcal{A}$: $10 = \frac{1}{2}(\mathrm{AM}^{i1})^2 + \frac{1}{2}(\mathrm{AM}^{i2})^2$

$\mathcal{B}$: 0

$\mathcal{C}$: $46.57 = \frac{1}{7}(\mathrm{CM}^{i3})^2 + \frac{1}{7}(\mathrm{CM}^{i4})^2 + \frac{1}{7}(\mathrm{CM}^{i5})^2 + \frac{1}{7}(\mathrm{CM}^{i7})^2 + \frac{1}{7}(\mathrm{CM}^{i8})^2 + \frac{1}{7}(\mathrm{CM}^{i9})^2 + \frac{1}{7}(\mathrm{CM}^{i10})^2$

1.7 Contributions

• Absolute contribution (Cta) = part of variance

Examples:

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

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

• 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 $\mathcal{C}$.

$$\tfrac{1}{10}(\mathrm{GM}^{i3})^2 + \cdots + \tfrac{1}{10}(\mathrm{GM}^{i5})^2 + \tfrac{1}{10}(\mathrm{GM}^{i7})^2 + \cdots + \tfrac{1}{10}(\mathrm{GM}^{i10})^2 =$$
$$10 + 0.4 + 3.6 + 3.2 + 1.6 + 11.6 + 13.6 = 44$$

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

— *Example*: absolute *within-contribution* of subcloud $\mathcal{C}$.

$$\tfrac{7}{10} \times 46.57 = 32.6$$

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 $\mathcal{C}$: $11.4 + 32.6 = 44$.

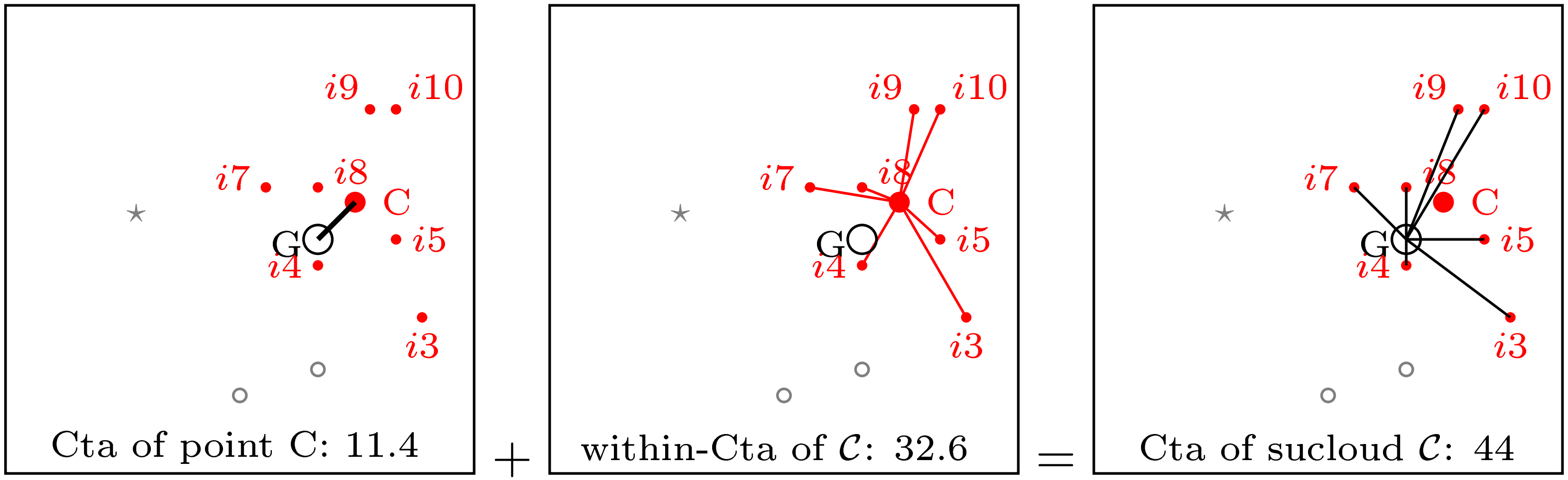


Figure 5 : Huyghens theorem

1.9 Between-within decomposition of variance

	Absolute contributions		
	mean points	within	subclouds
$\mathcal{A}$	26.0	2.0	28
$\mathcal{B}$	20.0	0	20
$\mathcal{C}$	11.4	32.6	44
Total	57.4	34.6	92
Variance	between	within	total

$$\begin{aligned}\text{Within-variance} &= \text{sum of within-contributions } (2.0+0+32.6)\\ &= \text{weighted mean of variances of subclouds } (\tfrac{2}{10}\times 10 + 0 + \tfrac{7}{10}\times 46.6)\\ &= 34.6\end{aligned}$$

$$\text{Total variance} = \text{between-variance} + \text{within-variance}$$

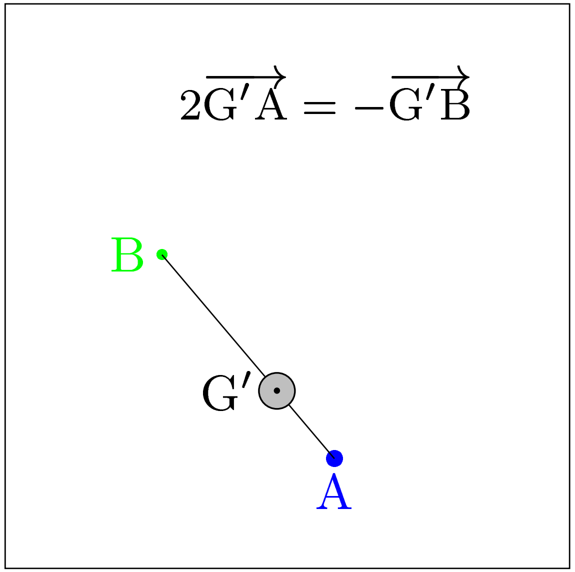
$$\eta^2 = \frac{\text{between-variance}}{\text{total variance}} \quad \text{(eta-square)}$$

Subcloud of 2 points (dipole)

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

Weight of dipole : $\widetilde{n_{AB}} = 1/(\frac{1}{n_A} + \frac{1}{n_B})$
Absolute contribution of the dipole: $p \times d^2$
with $p = \frac{\widetilde{n_{AB}}}{n}$ (relative weight) and $d^2 = \mathrm{AB}^2$
(square of the deviation).

Example: dipole $\{\mathrm{A}, \mathrm{B}\}$.
$\mathrm{AB}^2 = 290$
$\widetilde{n_{AB}} = \frac{1}{\frac{1}{1}+\frac{1}{2}} = 2/3$, $p = \frac{2/3}{10} = 0.06667$
Absolute contribution: $0.06667 \times 290 = 19.33$

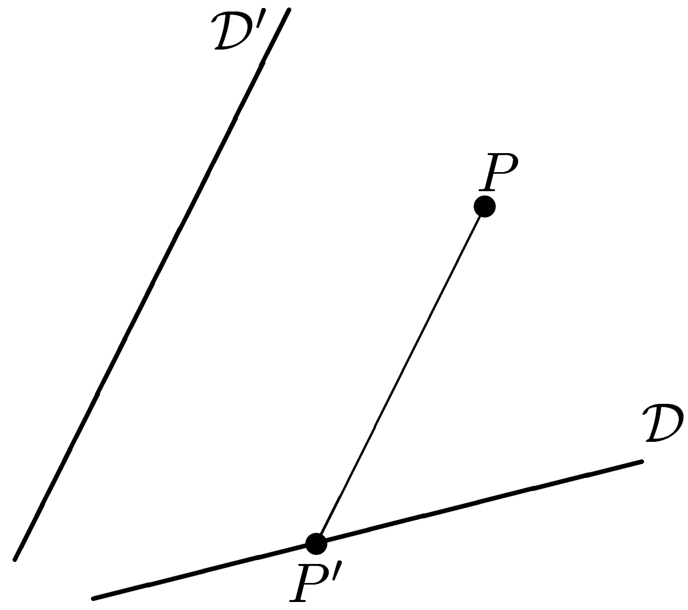


Property: The absolute contribution of a dipole is the absolute contribution of the subcloud of its two points.

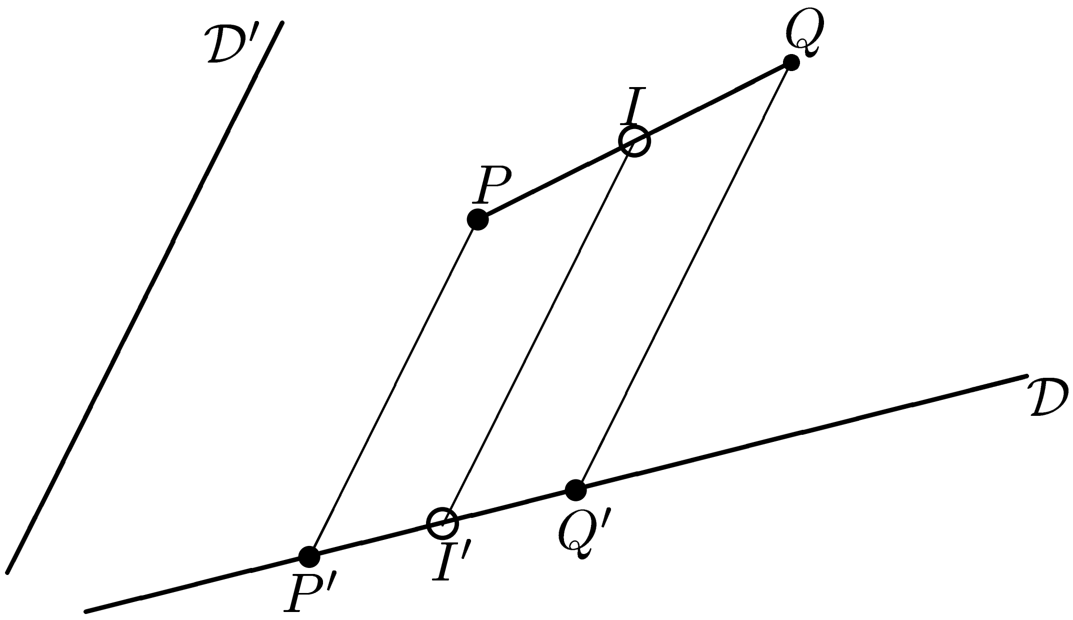
2 Principal axes of a cloud

2.1 Projection of a cloud

P' projection of point P onto $\mathcal{D}$ along $\mathcal{D}'$.



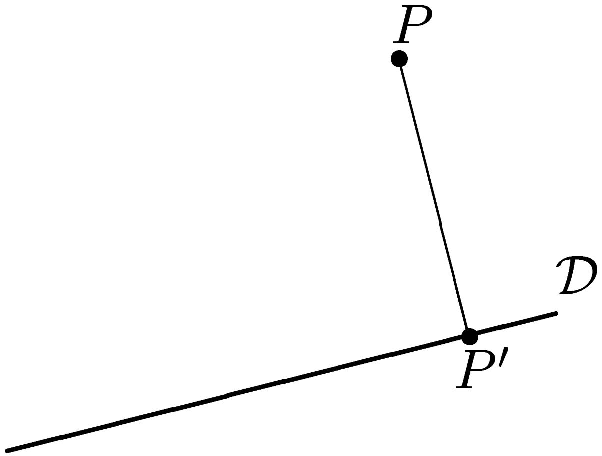
Midpoint property: The midpoint is preserved by projection

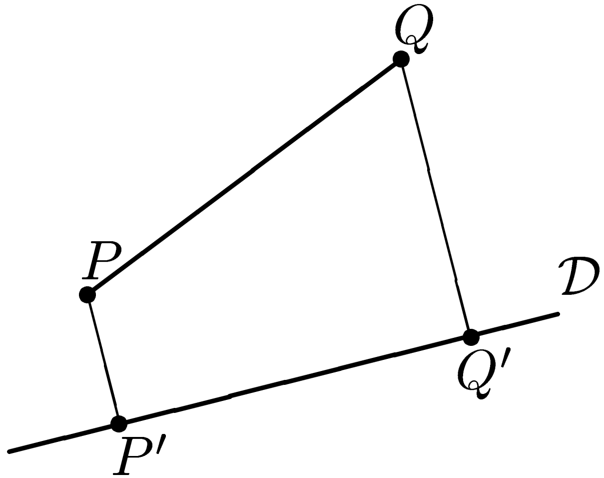


Mean point property : *The mean point of a cloud is preserved by projection.*

2.1.1 Orthogonal projection

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

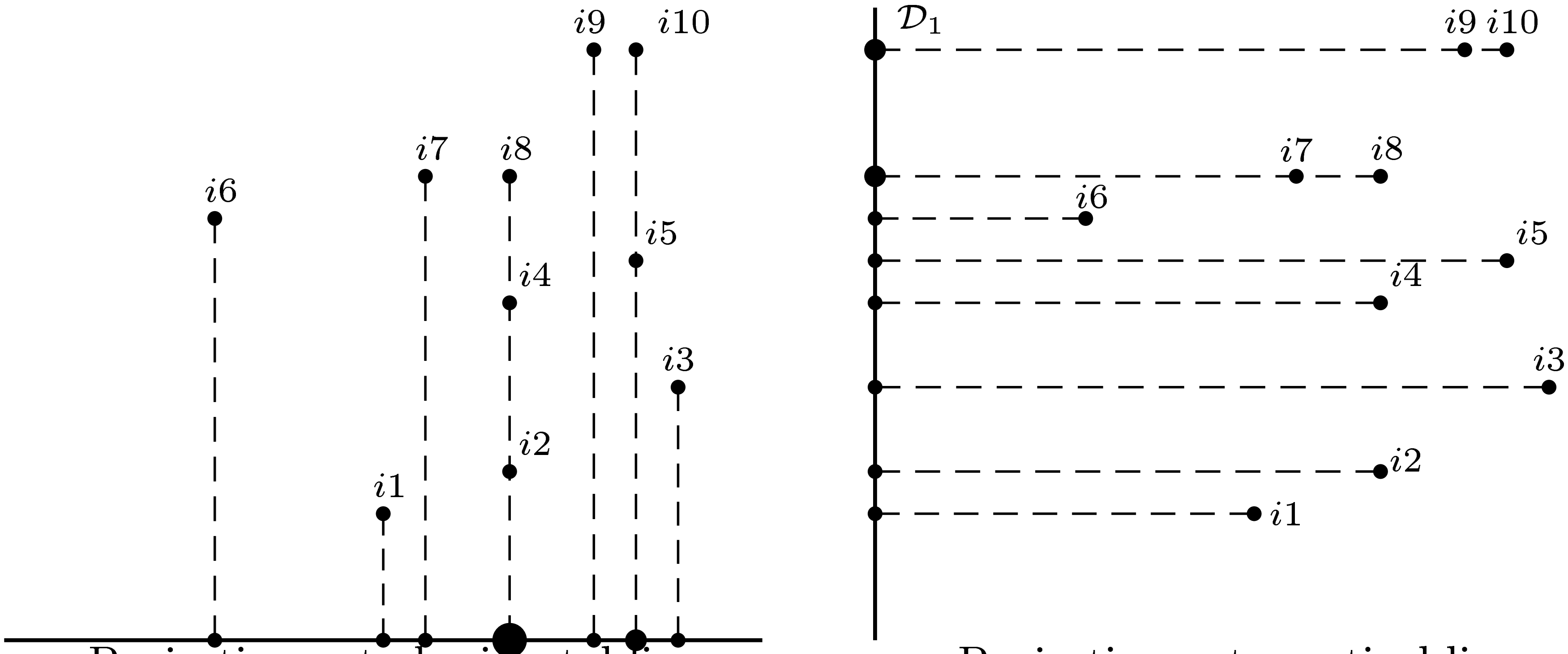




Contracting property: The orthogonal projection contracts distances ($P'Q' \leq PQ$).

Consequence: variance of projected cloud $\leq$ variance of initial cloud

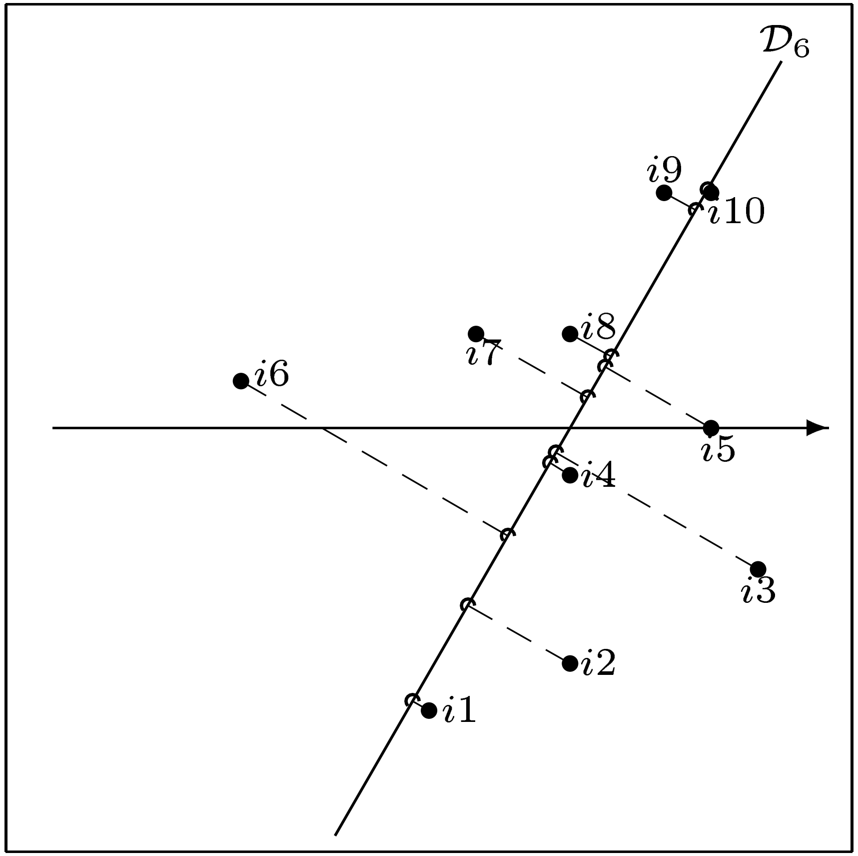
2.2 Projected clouds on several directions



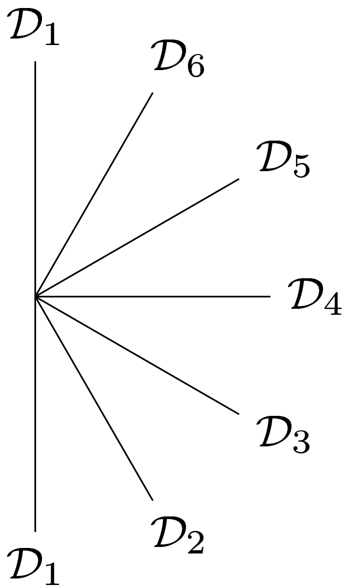
Projection onto horizontal line: variance=40

Projection onto vertical line: variance = 52

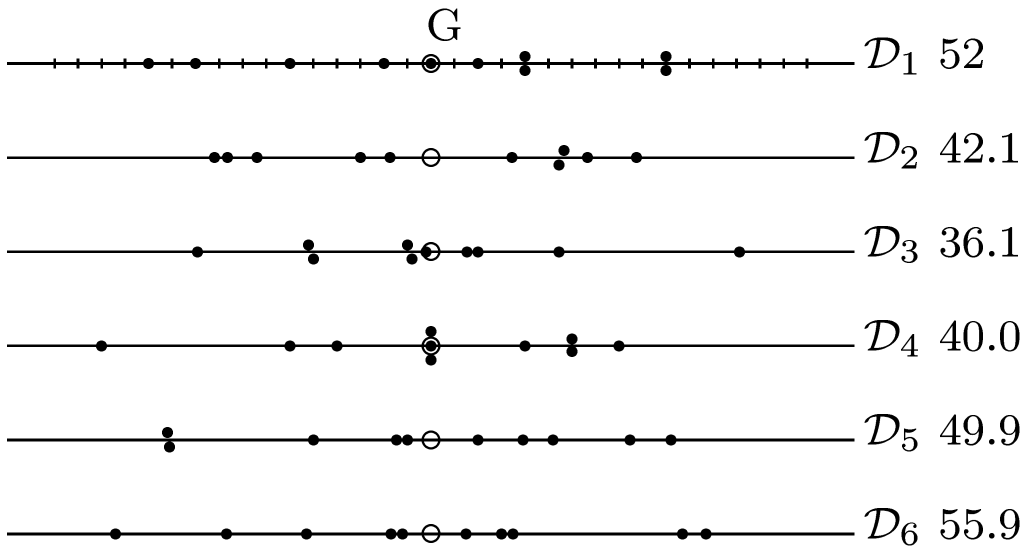
Orthogonal additive decomposition : the sum of the variances of projected clouds onto perpendicular directions is the variance of initial cloud: $40 + 52 = 92$.



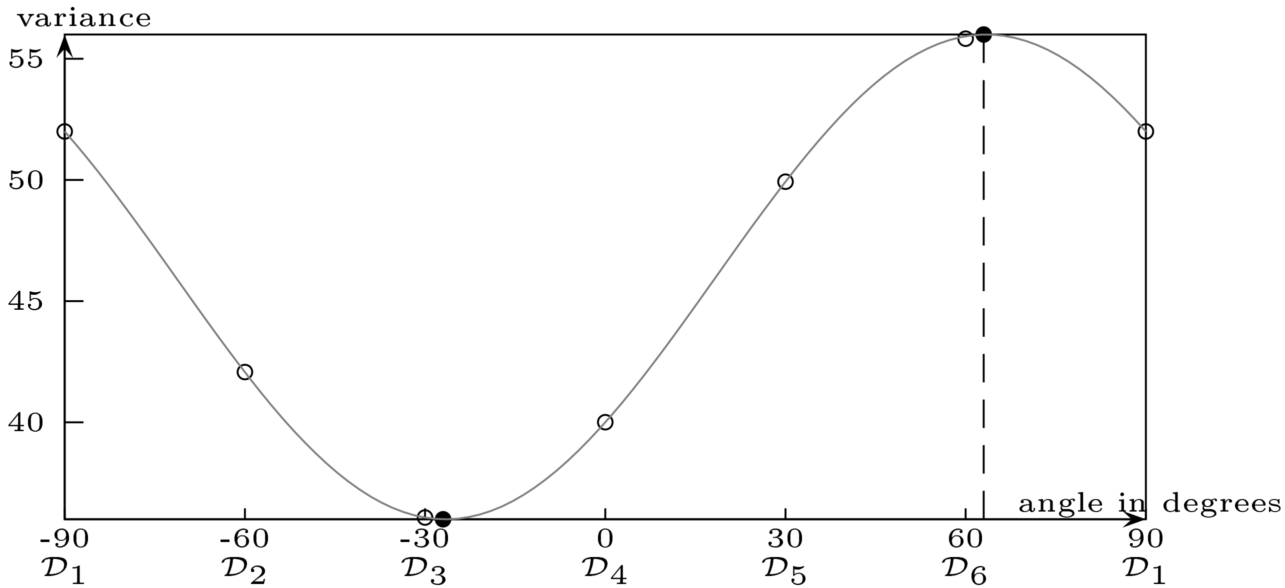
Projection onto skew line (60 degrees): variance = 55.9



$\mathcal{D}_1$
$\mathcal{D}_6$
$\mathcal{D}_5$
$\mathcal{D}_4$
$\mathcal{D}_3$
$\mathcal{D}_2$
$\mathcal{D}_1$



G
$\mathcal{D}_1$ 52
$\mathcal{D}_2$ 42.1
$\mathcal{D}_3$ 36.1
$\mathcal{D}_4$ 40.0
$\mathcal{D}_5$ 49.9
$\mathcal{D}_6$ 55.9



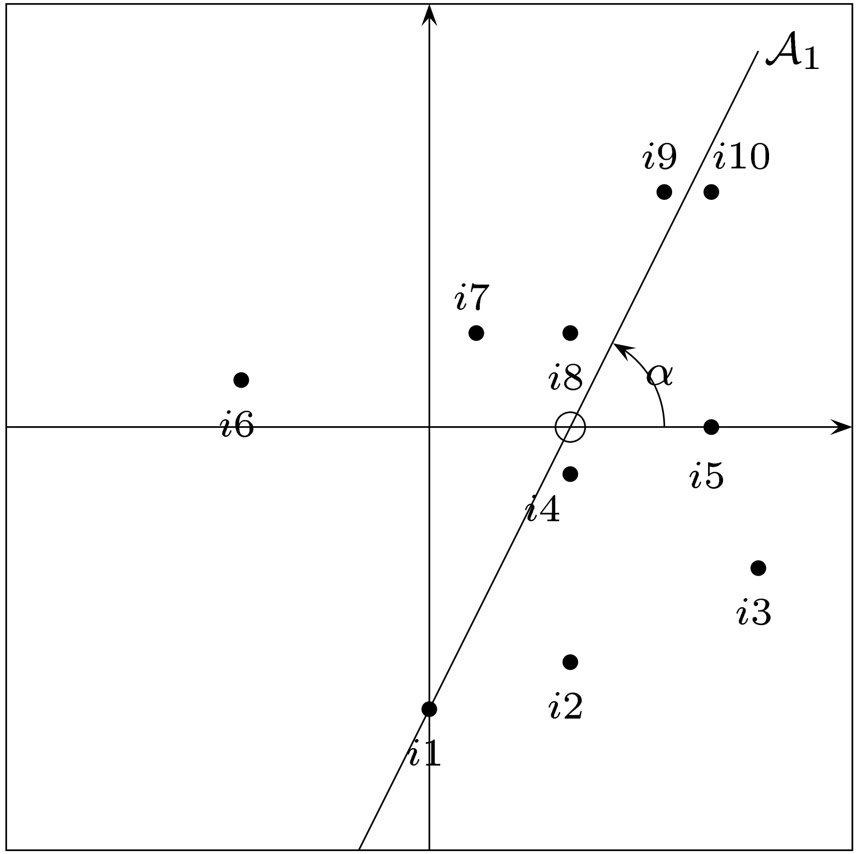
variance
55
50
45
40
angle in degrees
-90 $\mathcal{D}_1$
-60 $\mathcal{D}_2$
-30 $\mathcal{D}_3$
0 $\mathcal{D}_4$
30 $\mathcal{D}_5$
60 $\mathcal{D}_6$
90 $\mathcal{D}_1$

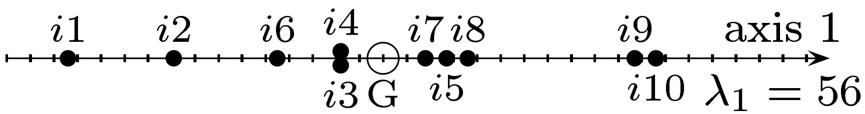
2.3 Principal Axes

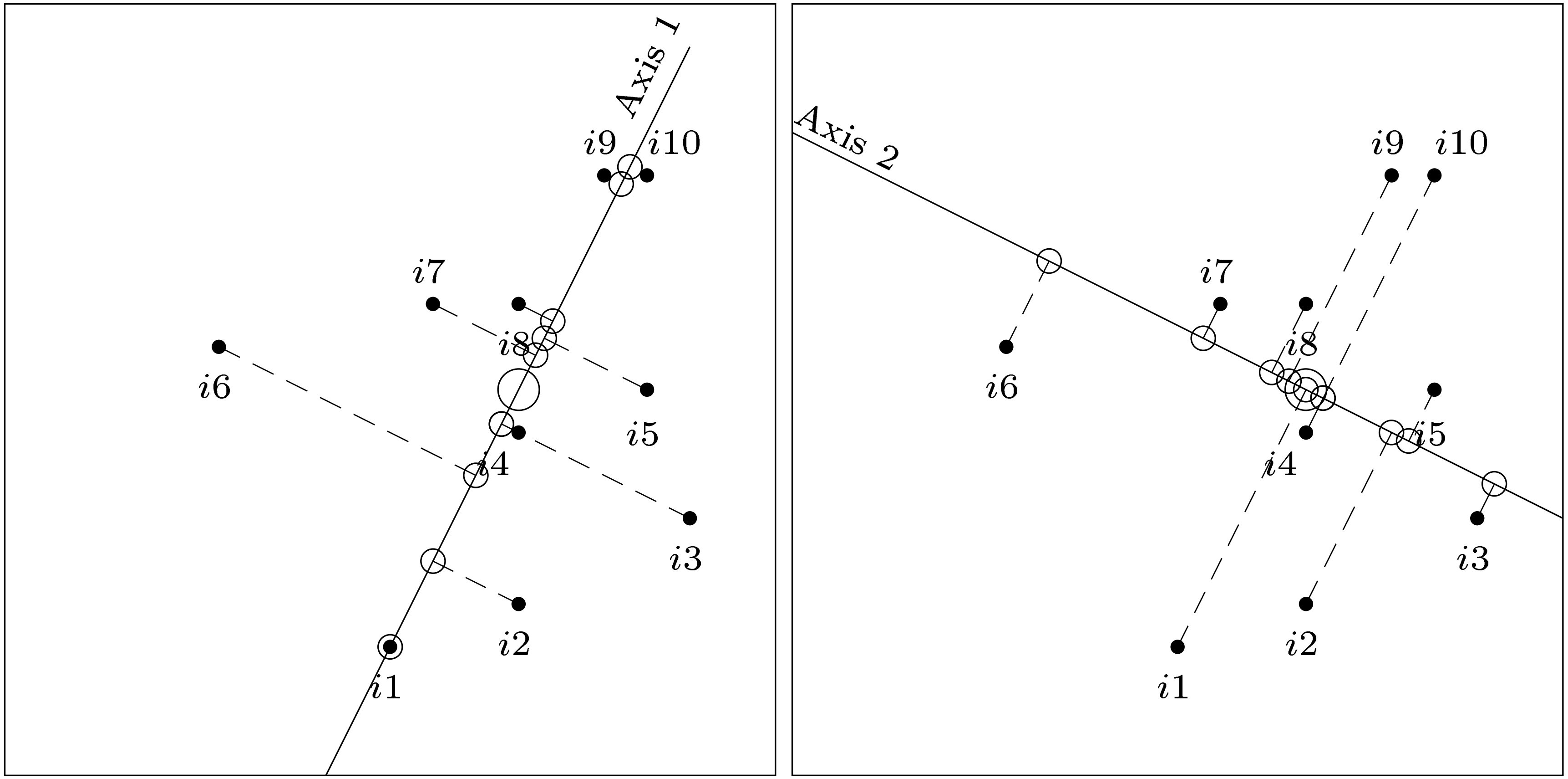
The oriented line for which the variance of the projected cloud is maximum is the *first principal axis* $\mathcal{A}_1$.

Projected cloud on $\mathcal{A}_1$: *first principal cloud*, best fit of the cloud by a unidimensional cloud, in the sense of orthogonal least squares.

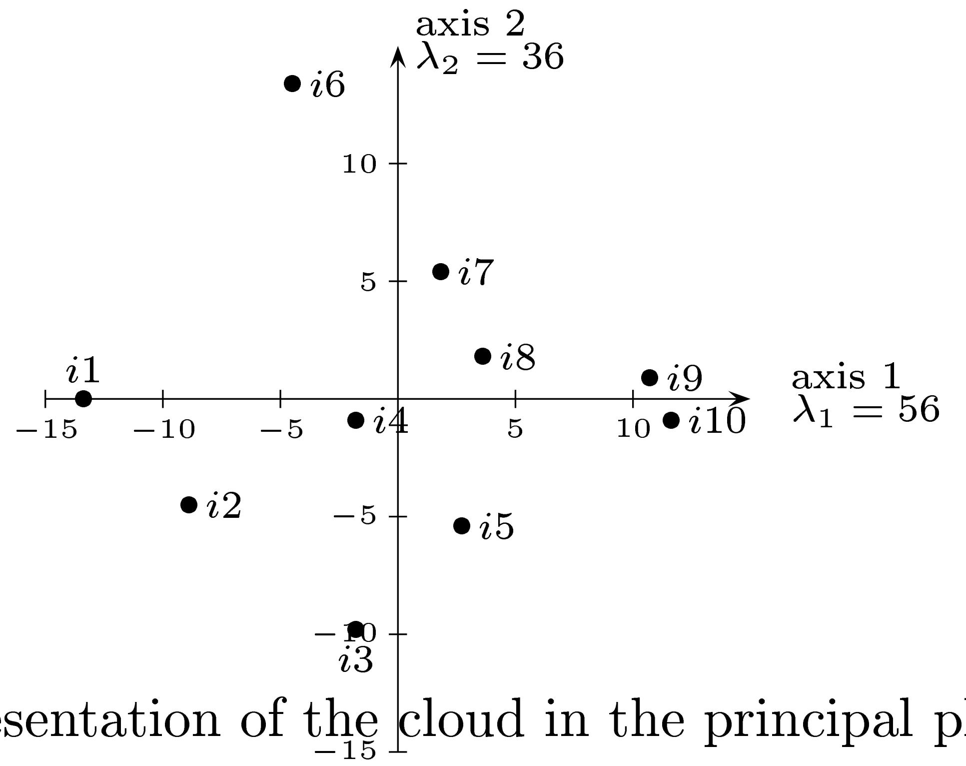
Here: $\alpha = 63°$ and variance (eigenvalue) $\lambda_1 = 56$.







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: λ_1 divided by total variance: $\dfrac{56}{92} = 61\%$

Results for Axis 1 ($\lambda_1 = 56$)

	Coordinates	Ctr (%)	Squared cosines
$i1$	-13.41	32.1	1.00
$i2$	-8.94	14.3	0.80
$i3$	-1.79	0.6	0.03
$i4$	-1.79	1.3	0.80
$i5$	$+2.68$	3.6	0.20
$i6$	-4.47	3.6	0.10
$i7$	$+1.79$	0.6	0.10
$i8$	$+3.58$	2.3	0.80
$i9$	$+10.73$	20.6	0.99
$i10$	$+11.63$	24.1	0.99

Results for Axis 2 ($\lambda_2 = 36$)

	Coordinates	Ctr (%)	Squared cosines
$i1$	0.00	0	0.00
$i2$	$+4.47$	5.6	0.20
$i3$	$+9.84$	26.9	0.97
$i4$	$+0.89$	0.2	0.20
$i5$	$+5.37$	8	0.80
$i6$	-13.42	50.0	0.90
$i7$	-5.37	8	0.90
$i8$	-1.79	0.9	0.20
$i9$	-0.89	0.2	0.01
$i10$	$+0.89$	0.2	0.01

• *Absolute contribution of point to axis*: weight×square of coordinate

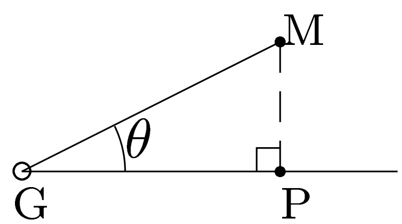
Example. For $i1$: weight$= 1/10$, coordinate$= -13.41$,
Cta $= \frac{1}{10} \times (-13.41)^2 = 17.96$.

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

Example. For $i1$, Ctr $= 17.96/56 = 32.1\%$.

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

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) :

$$\mathrm{MM}' = \sqrt{(x_1 - x'_1)^2 + (x_2 - x'_2)^2}$$

- Variance of abscissas: $v_1 = \sum n_i (x_1^i - \overline{x_1})^2/n$: $v_1 = 40$

  Variance of ordinates: $v_2 = \sum n_i (x_2^i - \overline{x_2})^2/n$: $v_2 = 52$

  Covariance: $c = \sum n_i (x_1^i - \overline{x_1})(x_2^i - \overline{x_2})/n$: $c = 8$

- Equation for eigenvalues: $\lambda^2 - (v_1 + v_2)\lambda + v_1 v_2 - c^2 = 0$

  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 α 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: $y^i = (x_1^i - \overline{x_1})\cos\alpha + (x_2^i - \overline{x_2})\sin\alpha$

References

— BENZÉCRI J-P. (1992) *Correspondence Analysis Handbook*, (Part I), New York: Dekker .

— LE ROUX B. & ROUANET H. (2004), *Geometric Data Analysis: from Correspondence Analysis to Structured Data Analysis* (chapter 3), Dordrecht: Kluwer.

— RICHARD J-F. , LE ROUX B. & ROUANET H. (1996). "Introduction à l'analyse des données et à l'analyse des comparaisons". *Cours de Psychologie 6* (chapter 3 §4), Paris: CNED-Dunod.

— ROUANET H. & LE ROUX B. (1995). *Exercices & Solutions, Statistique en Sciences Humaines* (chapter IV), Paris: Dunod.

— ROUANET H. & LE ROUX B. (1993). *Analyse des données multidimensionnelles* (chapters V & VI), Paris: Dunod.