Multidimensional scaling and flat split systems

Monika Balvočiūtė

joint work with
David Bryant

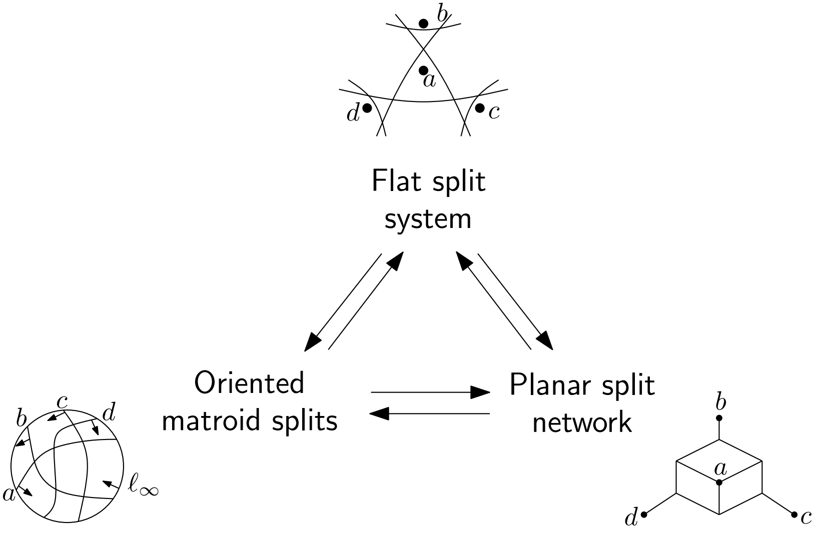
University of Otago

6th Nov 2014

Splits and Split systems

- A **split** $S = A|B$ is a bipartition of a set of taxa $\mathcal{X}$ into two non empty subsets such that $\mathcal{X} = A \cup B$ and $A \cap B = \emptyset$.
- A **split system** $\mathcal{S}$ is set of splits $\{S\}$ over some set of taxa $\mathcal{X}$.

Equivalent representations of flat split systems



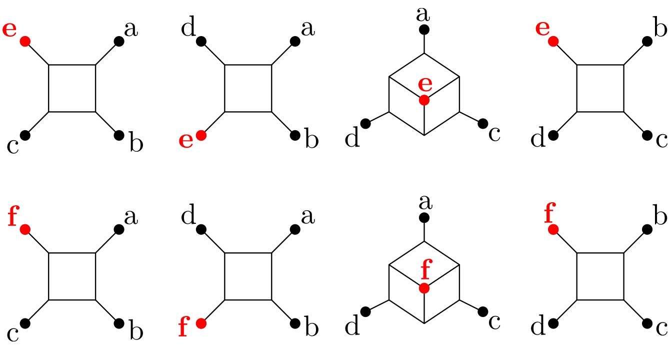
FlatNJ – computing planar split networks

- Compute building blocks
- Identify neighbors
- Agglomerate
- Reverse agglomeration
- Weight and filter

M. Balvočiūtė, A. Spillner and V. Moulton, FlatNJ:..., Syst. Biol. 2014, 63(3): 383–96

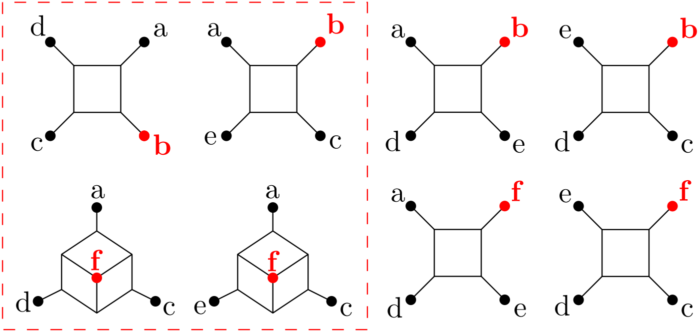
Neighbors

e and f are neighbors

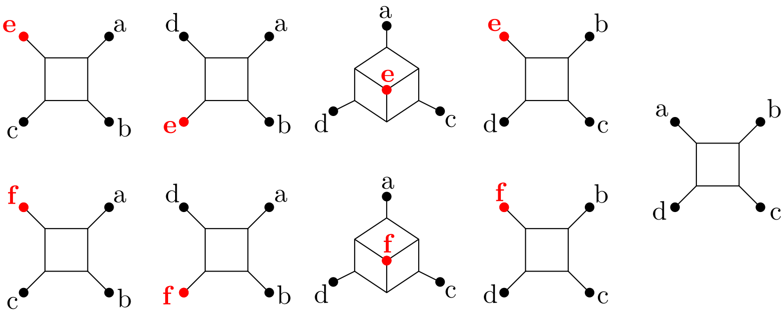


Not Neighbors

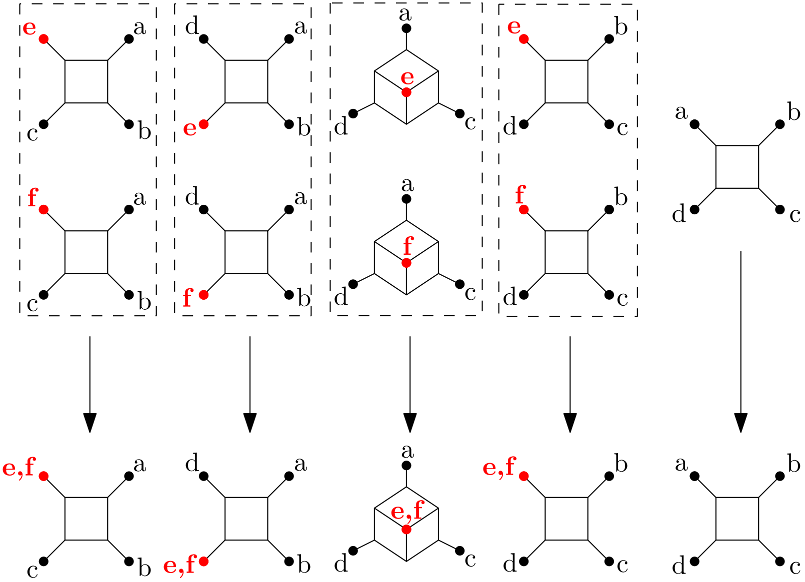
b and f are not neighbors



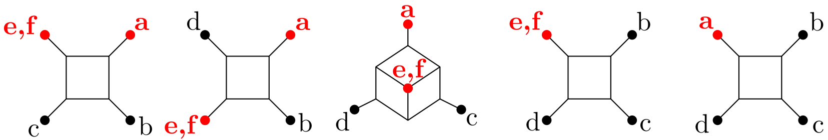
Agglomeration



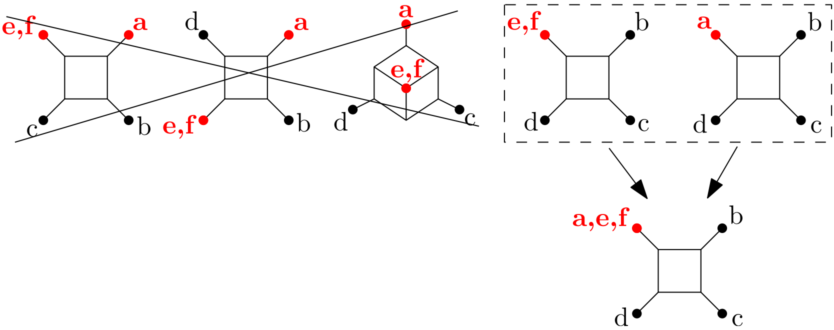
Agglomeration



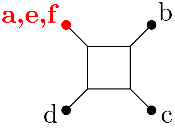
Agglomeration



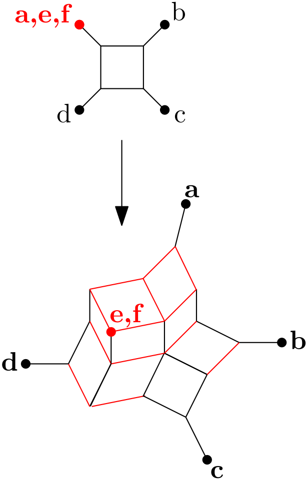
Agglomeration



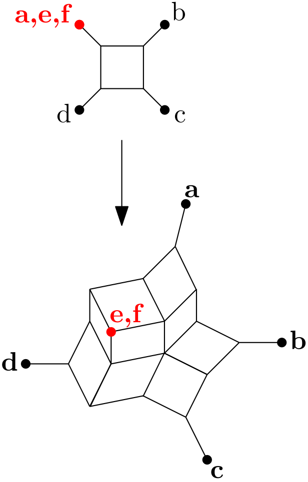
Reversing agglomeration



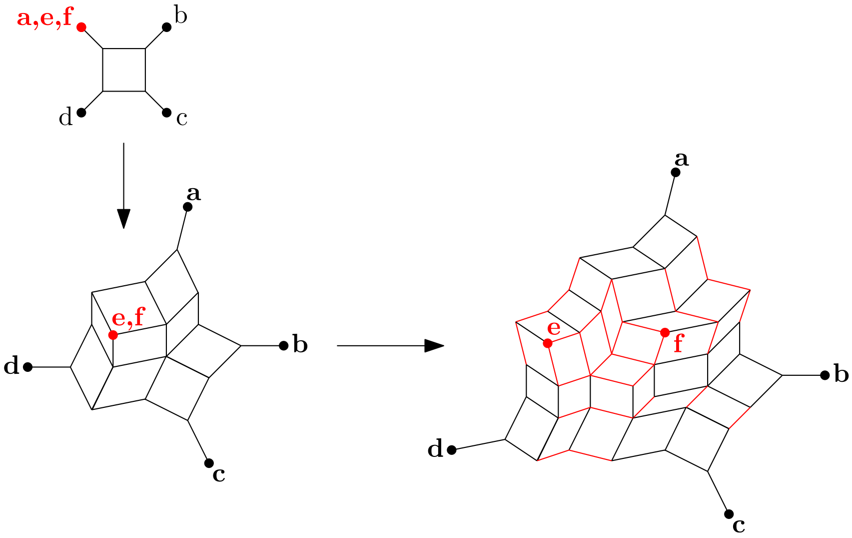
Reversing agglomeration



Reversing agglomeration



Reversing agglomeration

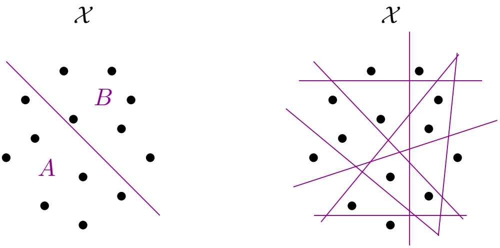


Q: When does it fail?

A: When there are no neighbours.

Affine splits

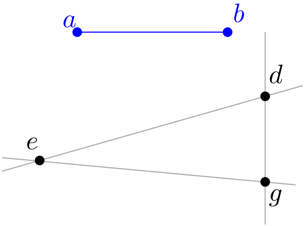
- Split – line ℓ_S in $\mathbb{R}^2 - \mathcal{X}$;
- Split system – arrangement of lines $\mathcal{A}$ in $\mathbb{R}^2 - \mathcal{X}$;



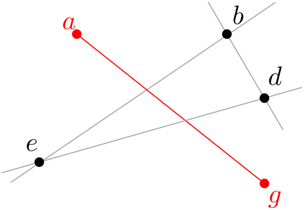
Split

Split system

Neighbours in affine split systems

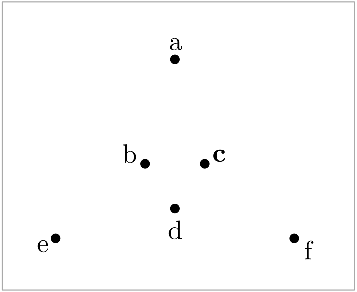


Neighbours



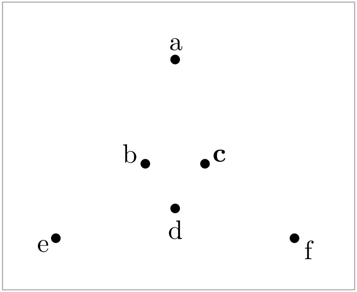
Not neighbours

For example



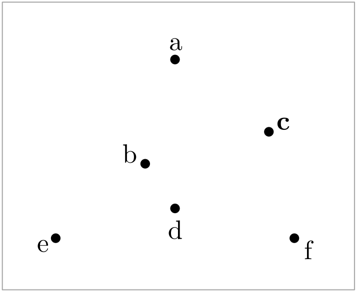
Input

For example



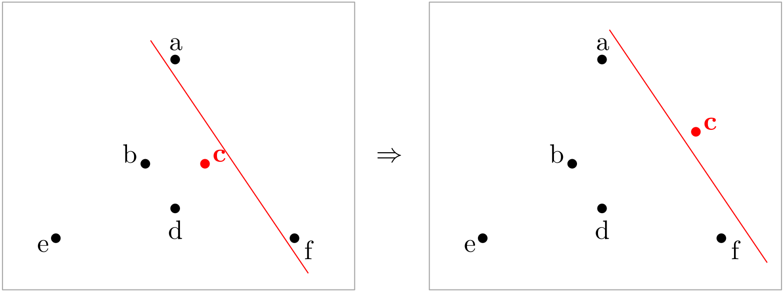
Input

⇒



Output

For example



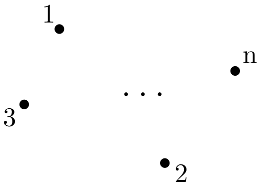
Input Output

Multidimensional scaling (MDS)

- Plot points in low (e.g. two) dimensional space based on their pairwise distances.

	1	2	3	...	n
1	0	d_{12}	d_{13}	$\dots$	d_{1n}
2	d_{12}	0	d_{23}	$\dots$	d_{2n}
3	d_{13}	d_{23}	0	$\dots$	d_{3n}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
n	d_{1n}	d_{2n}	d_{3n}	$\dots$	0

$\Rightarrow$

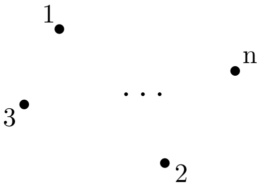


Multidimensional scaling (MDS)

- Plot points in low (e.g. two) dimensional space based on their pairwise distances.

	1	2	3	…	n
1	0	d_{12}	d_{13}	…	d_{1n}
2	d_{12}	0	d_{23}	…	d_{2n}
3	d_{13}	d_{23}	0	…	d_{3n}
⋮	⋮	⋮	⋮	⋱	⋮
n	d_{1n}	d_{2n}	d_{3n}	…	0

$\Rightarrow$



- Minimize the difference between **input** and **output** distances.

Stress

MSD

$\sum_i \sum_{j \neq i} (d_{ij} - \delta_{ij})^2$

↘

Stress

d_{ij} – actual distance; δ_{ij} – plotted distance

MSD

$\sum_i \sum_{j \neq i} (d_{ij} - \delta_{ij})^2$ $\quad\quad$ $\sum_i \sum_{j \neq i} (d_{ij}^2 - \delta_{ij}^2)^2$

Stress

d_{ij} – actual distance; δ_{ij} – plotted distance

MSD

$$\sum_i \sum_{j \neq i} (d_{ij} - \delta_{ij})^2$$

$$\sum_i \sum_{j \neq i} (d_{ij}^2 - \delta_{ij}^2)^2$$

Stress

$$\sqrt{\frac{\sum_i \sum_{j \neq i} (d_{ij} - \delta_{ij})^2}{\sum_i \sum_{j \neq i} d_{ij}^2}}$$

d_{ij} – actual distance; δ_{ij} – plotted distance

MSD

$$\sum_i \sum_{j \neq i} (d_{ij} - \delta_{ij})^2$$

$$\sum_i \sum_{j \neq i} (d_{ij}^2 - \delta_{ij}^2)^2$$

Stress

$$\sqrt{\frac{\sum_i \sum_{j \neq i} (d_{ij} - \delta_{ij})^2}{\sum_i \sum_{j \neq i} d_{ij}^2}}$$

$$\sqrt{\frac{\sum_i \sum_{j \neq i} w_{ij} (d_{ij} - \delta_{ij})^2}{\sum_i \sum_{j \neq i} w_{ij} d_{ij}^2}}$$

d_{ij} – actual distance; δ_{ij} – plotted distance

MSD

Stress

$$\sum_i \sum_{j \neq i} (d_{ij} - \delta_{ij})^2$$

$$\sum_i \sum_{j \neq i} (d_{ij}^2 - \delta_{ij}^2)^2$$

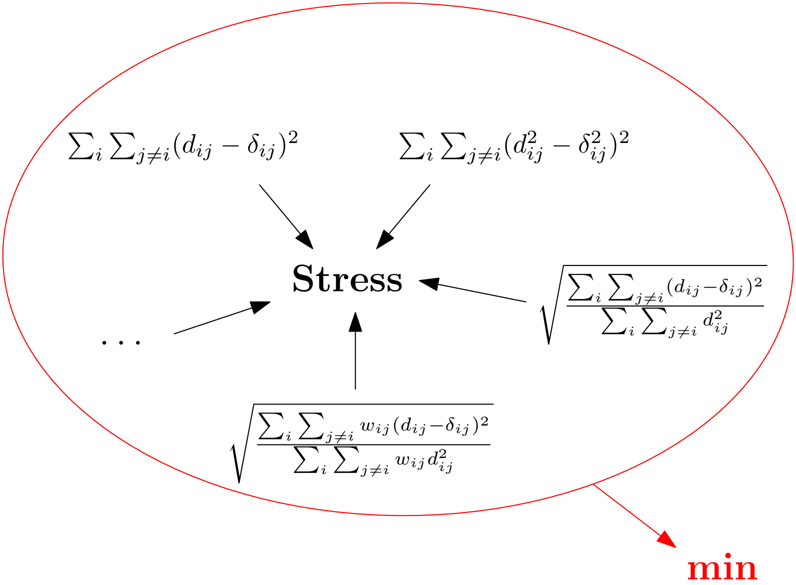
$$\sqrt{\frac{\sum_i \sum_{j \neq i} (d_{ij} - \delta_{ij})^2}{\sum_i \sum_{j \neq i} d_{ij}^2}}$$

$$\sqrt{\frac{\sum_i \sum_{j \neq i} w_{ij} (d_{ij} - \delta_{ij})^2}{\sum_i \sum_{j \neq i} w_{ij} d_{ij}^2}}$$

. . .

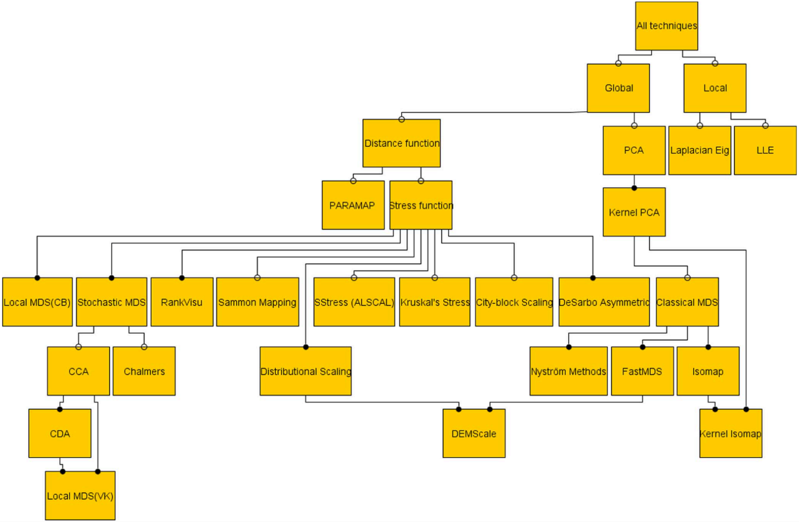
d_{ij} – actual distance; δ_{ij} – plotted distance

MSD



d_{ij} – actual distance; δ_{ij} – plotted distance

MSD

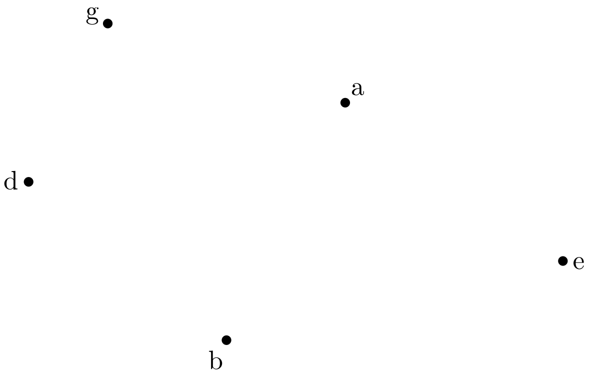


S. L. France & J. D. Carroll, Two-Way Multidimensional Scaling: A Review, IEEE Trans. Syst., Man, Cybern.,Syst 2011, 41(5): 644–61

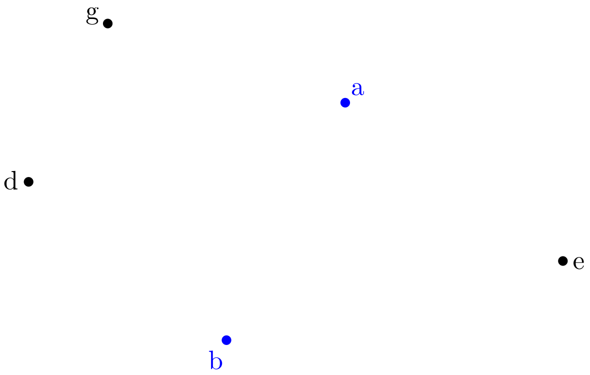
Agglomerative approach to MDS

- Take pairwise distance matrix
- Identify neighbours
- Agglomerate
- Reverse

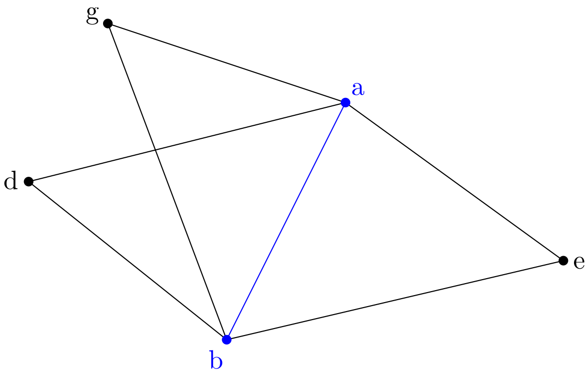
Agglomeration



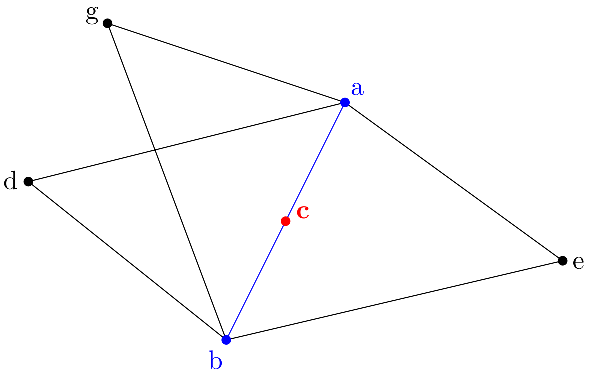
Agglomeration



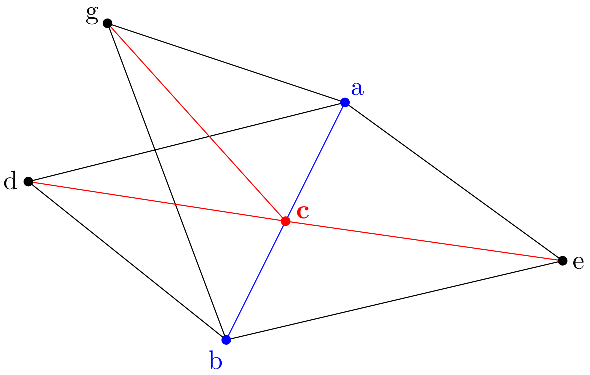
Agglomeration



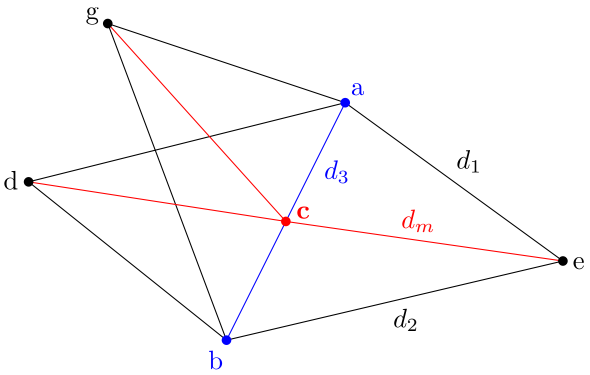
Agglomeration



Agglomeration

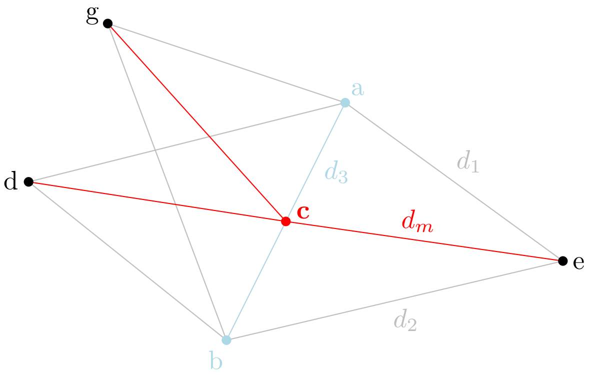


Agglomeration



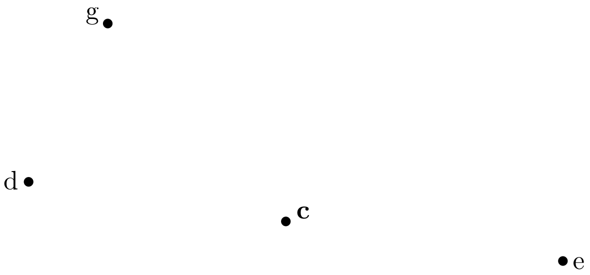
$$d_m = \sqrt{\frac{2d_1^2+2d_2^2-d_3^2}{4}}$$

Agglomeration



$$d_m = \sqrt{\frac{2d_1^2 + 2d_2^2 - d_3^2}{4}}$$

Agglomeration



$$d_m = \sqrt{\frac{2d_1^2+2d_2^2-d_3^2}{4}}$$

Agglomeration

	1	2	...	m	a	b
1	0	d_{12}	...	d_{1m}	d_{a1}	d_{b1}
2	d_{12}	0	...	d_{2m}	d_{a2}	d_{b2}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$
m	d_{1m}	d_{2m}	...	0	d_{am}	d_{bm}
a	d_{a1}	d_{a2}	...	d_{am}	0	d_{ab}
b	d_{b1}	d_{b2}	...	d_{bm}	d_{ab}	0

Agglomeration

	1	2	...	m	**a**	**b**
1	0	d_{12}	...	d_{1m}	d_{a1}	d_{b1}
2	d_{12}	0	...	d_{2m}	d_{a2}	d_{b2}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$
m	d_{1m}	d_{2m}	...	0	d_{am}	d_{bm}
a	d_{a1}	d_{a2}	...	d_{am}	0	d_{ab}
b	d_{b1}	d_{b2}	...	d_{bm}	d_{ab}	0

Agglomeration

	1	2	…	m	**a**	**b**
1	0	d_{12}	…	d_{1m}	d_{a1}	d_{b1}
2	d_{12}	0	…	d_{2m}	d_{a2}	d_{b2}
⋮	⋮	⋮	⋱	⋮	⋮	⋮
m	d_{1m}	d_{2m}	…	0	d_{am}	d_{bm}
a	d_{a1}	d_{a2}	…	d_{am}	0	d_{ab}
b	d_{b1}	d_{b2}	…	d_{bm}	d_{ab}	0

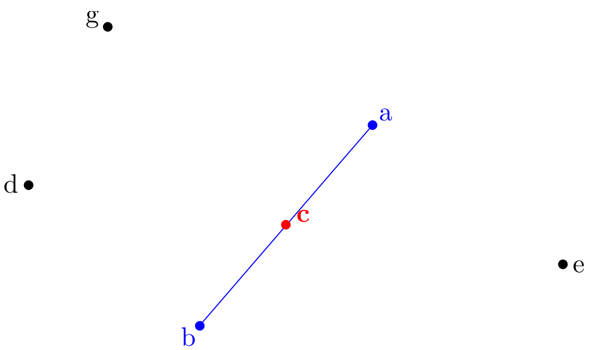
$$\Downarrow$$

	1	2	…	m	**c**
1	0	d_{12}	…	d_{1m}	$d_{c1} = \sqrt{\frac{2d_{a1}^2 + 2d_{b1}^2 - d_{ab}^2}{4}}$
2	d_{12}	0	…	d_{2m}	$d_{c2} = \sqrt{\frac{2d_{a2}^2 + 2d_{b2}^2 - d_{ab}^2}{4}}$
⋮	⋮	⋮	⋱	⋮	⋮
m	d_{1m}	d_{2m}	…	0	$d_{cm} = \sqrt{\frac{2d_{am}^2 + 2d_{bm}^2 - d_{ab}^2}{4}}$
c	d_{c1}	d_{c2}	…	d_{cm}	0

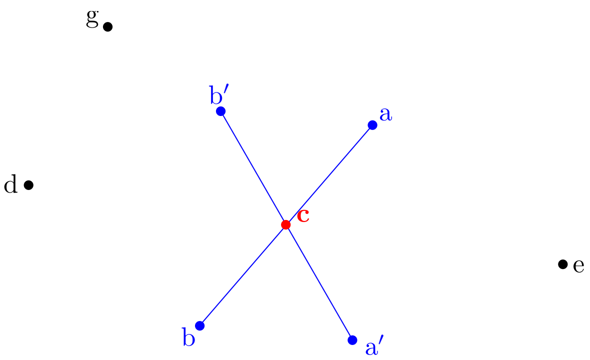
Expansion

Expansion

Expansion

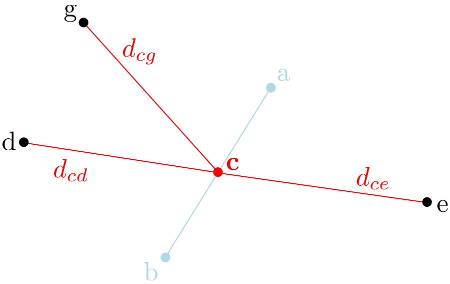


Expansion



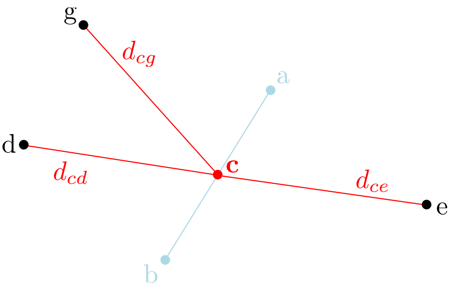
Expansion

We know:



Expansion

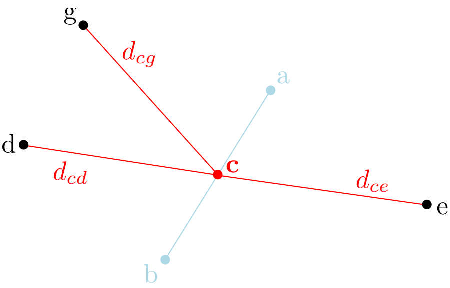
We know:



$c = \{a, b\}$

Expansion

We know:

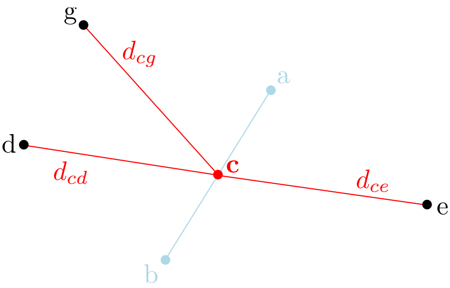


$c = \{a, b\}$

d_{ag}, d_{bg}
d_{ad}, d_{bd}
d_{ae}, d_{be}

Expansion

We know:



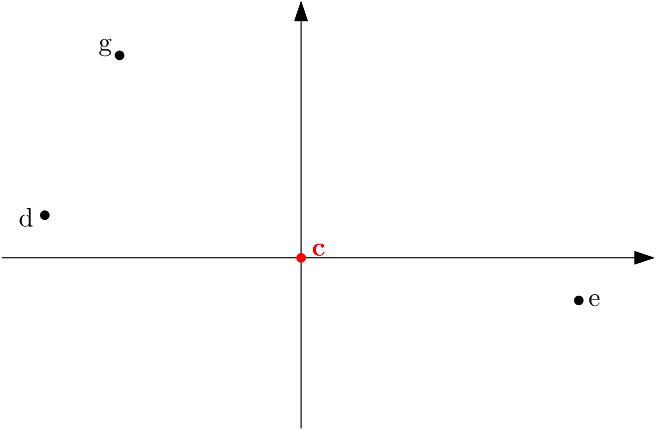
$c = \{a, b\}$

d_{ag}, d_{bg}
d_{ad}, d_{bd}
d_{ae}, d_{be}

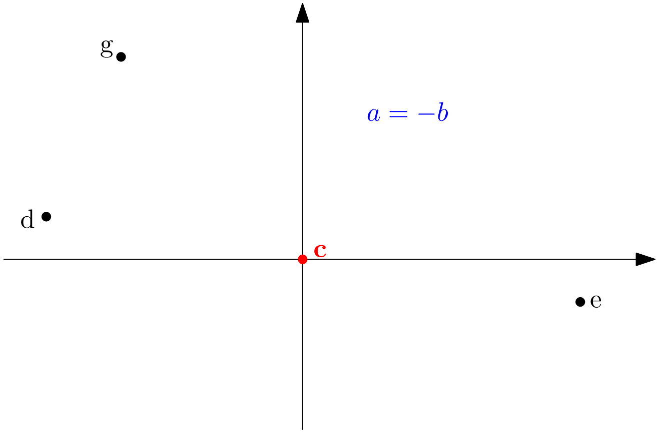
We don't know:
Actual dimension

Expansion

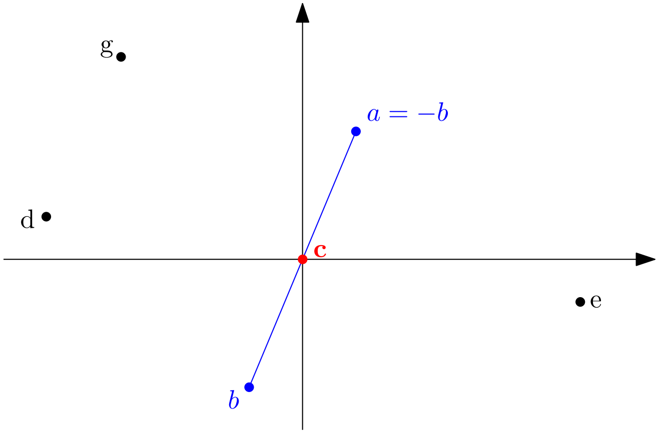
Expansion



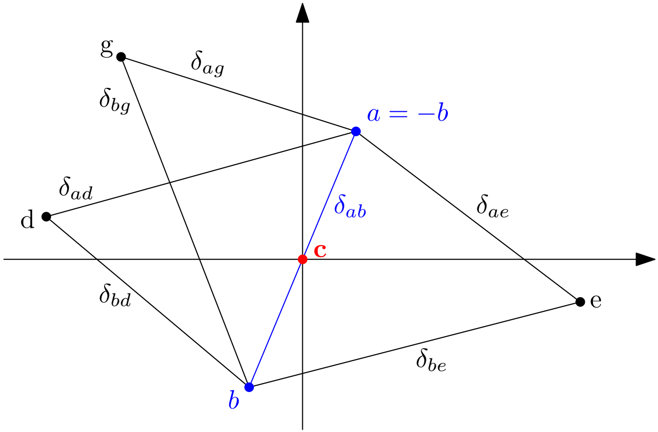
Expansion



Expansion

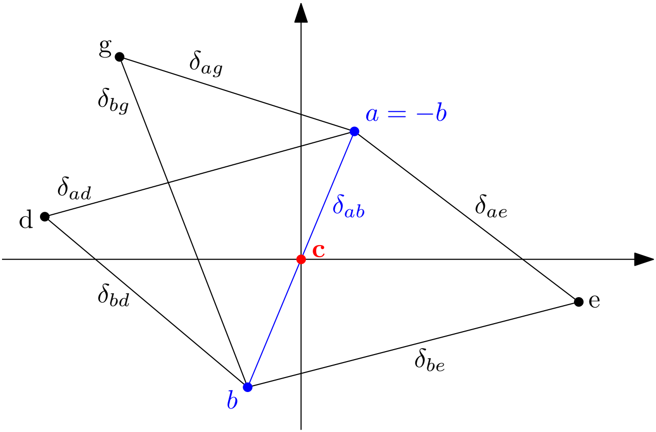


Expansion



g
δ_{ag}
δ_{bg}
$a = -b$
δ_{ad}
δ_{ab}
δ_{ae}
d
c
δ_{bd}
e
δ_{be}
b

Expansion



$$\delta_{ab} \sim d_{ab} \quad \delta_{ag} \sim d_{ag} \quad \delta_{ad} \sim d_{ad} \quad \delta_{ae} \sim d_{ae}$$
$$\delta_{bg} \sim d_{bg} \quad \delta_{bd} \sim d_{bd} \quad \delta_{be} \sim d_{be}$$

Expansion [minimizing stress function]

We have m points and want to separate a and b:

$$\sum_{i=1}^{m} \left[(\delta_{ai} - d_{ai})^2 + (\delta_{bi} - d_{bi})^2 \right] + (\delta_{ab} - d_{ab})^2 \to min$$

Expansion [minimizing stress function]

We have m points and want to separate a and b:

$$\sum_{i=1}^{m} \left[(\delta_{ai} - d_{ai})^2 + (\delta_{bi} - d_{bi})^2\right] + (\delta_{ab} - d_{ab})^2 \rightarrow min$$

Substitute distances (δ's) with coordinates
(remember that $a = -b$):

$$\sum_{i=1}^{m}[(\sqrt{(x_i - x_a)^2 + (y_i - y_a)^2} - d_{ai})^2 +$$
$$(\sqrt{(x_i + x_a)^2 + (y_i + y_a)^2} - d_{bi})^2] +$$
$$(2\sqrt{(x_a)^2 + (y_a)^2} - d_{ab})^2 \rightarrow min$$

Expansion [minimizing stress function]

We have m points and want to separate a and b:

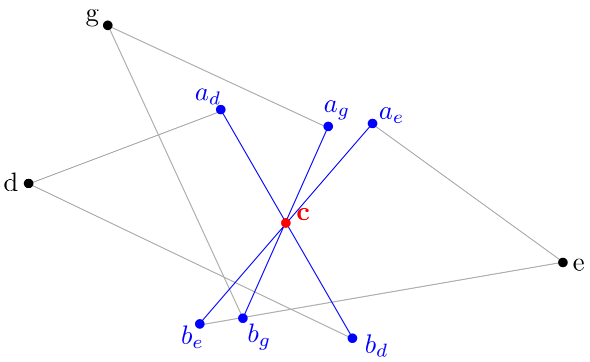
$$\sum_{i=1}^{m} \left[(\delta_{ai} - d_{ai})^2 + (\delta_{bi} - d_{bi})^2\right] + (\delta_{ab} - d_{ab})^2 \rightarrow min$$

Substitute distances (δ's) with coordinates
(remember that $a = -b$):

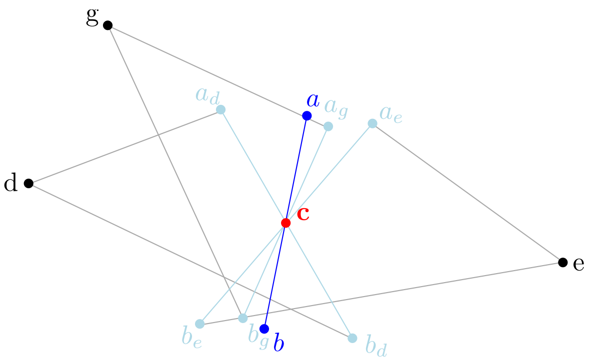
$$\sum_{i=1}^{m}[(\sqrt{(x_i - x_a)^2 + (y_i - y_a)^2} - d_{ai})^2 + \\ (\sqrt{(x_i + x_a)^2 + (y_i + y_a)^2} - d_{bi})^2] + \\ (2\sqrt{(x_a)^2 + (y_a)^2} - d_{ab})^2 \rightarrow min$$

And that is hard.

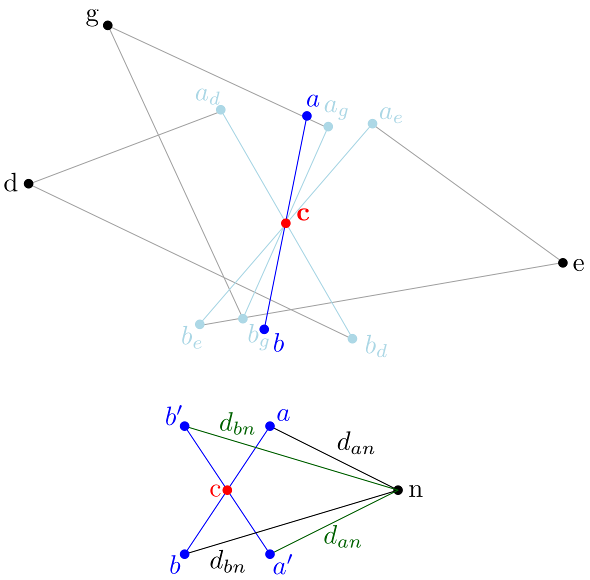
Expansion (Solution no.1)



Expansion (Solution no.1)



Expansion (Solution no.1)



Expansion (Solution no.2)

Solve numerically.

How to evaluate what we get?

- Compute overall stress.

How to evaluate what we get?

- Compute overall stress.
- Compare neighbourhoods (n nearest neighbours).

How to select neighbours?

- Minimum/maximum **distance**
- Minimum/maximum **variance**

Thanks to

UNIVERSITY of OTAGO
Te Whare Wānanga o Otāgo
NEW ZEALAND
SAPERE AUDE

ALLAN WILSON CENTRE

Thank **you** for attention!