Dendrogram

***K. Gibert*[(1)]**

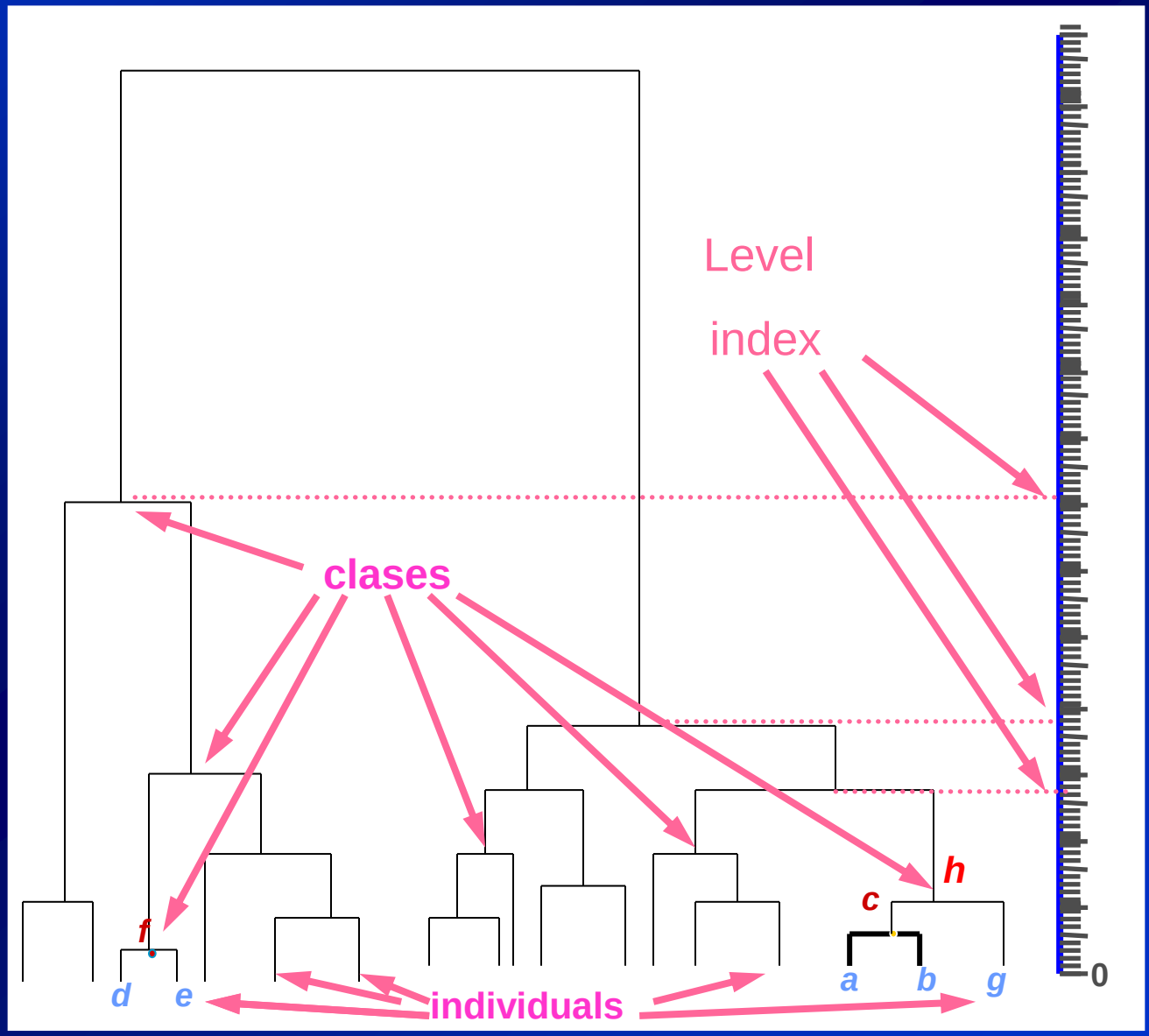
[(1)]Department of Statistics and Operation Research

Knowledge Engineering and Machine Learning group
Universitat Politècnica de Catalunya, Barcelona

Ascendant hierarchical clustering

Dendrogram structure:

- Leaves
- Internal nodes
- Node height



How to cut a dendrogram

- Always cut HORIZONTALLY
- Look for levels with long branches *(biggest gaps)*
- Maximize the ratio (Calinski-Harabatz)

Inertia between classes *distinguishability*

versus inertia within classes *homogeneity*

$$CH_k = \frac{B_k / (k-1)}{W_k / (n-k)}$$

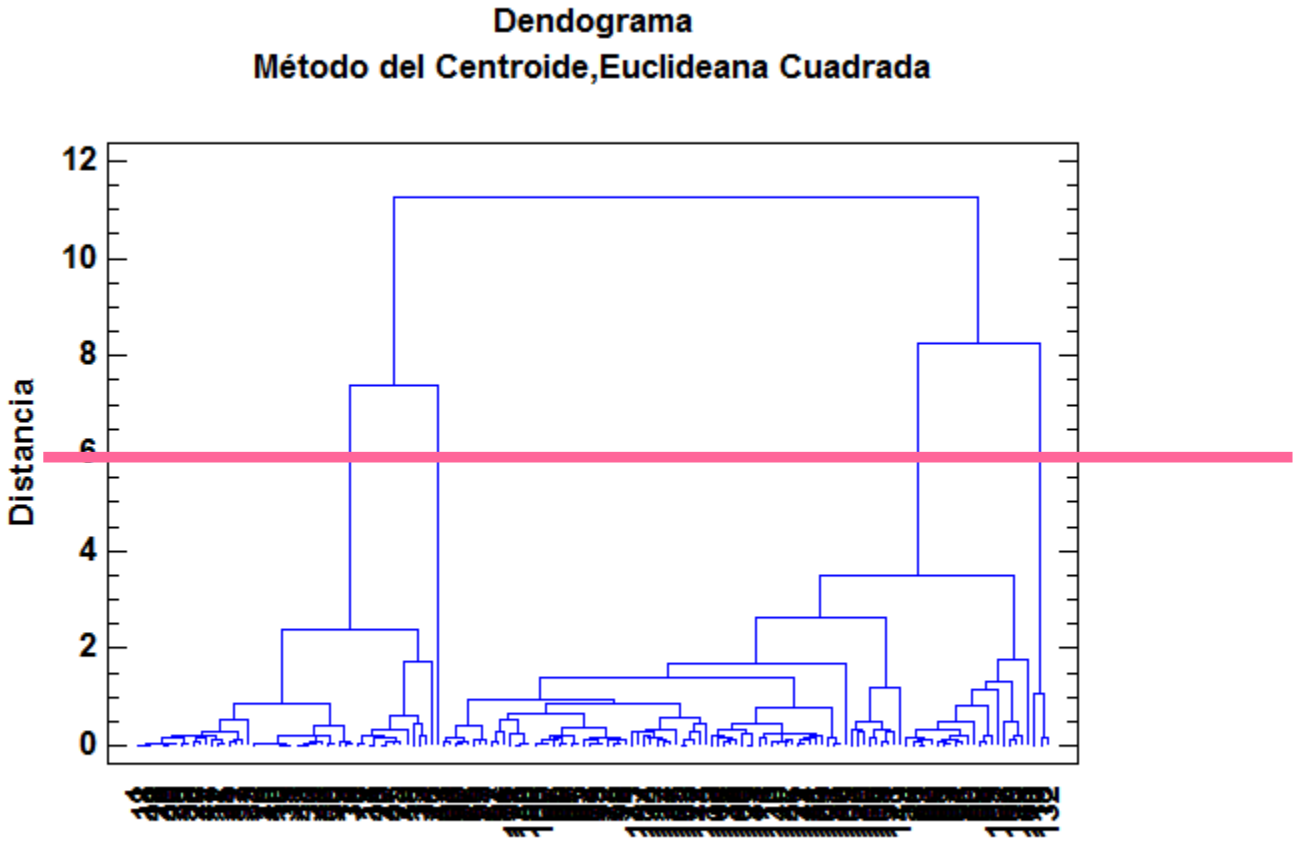
(often the optimal is a 2 classes-cut, TRIVIAL!!!!)

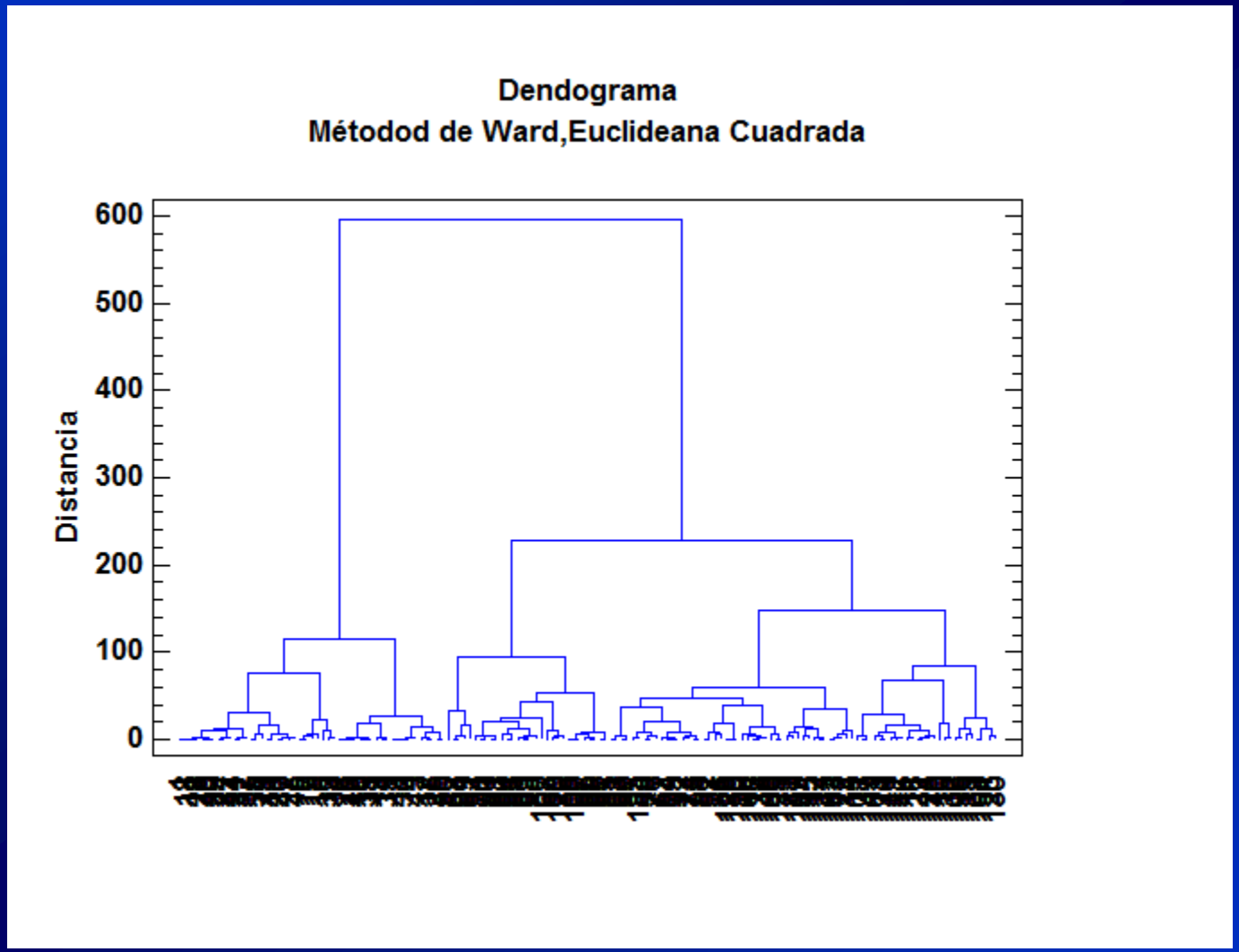
(There are tests, but do they work with large samples?)

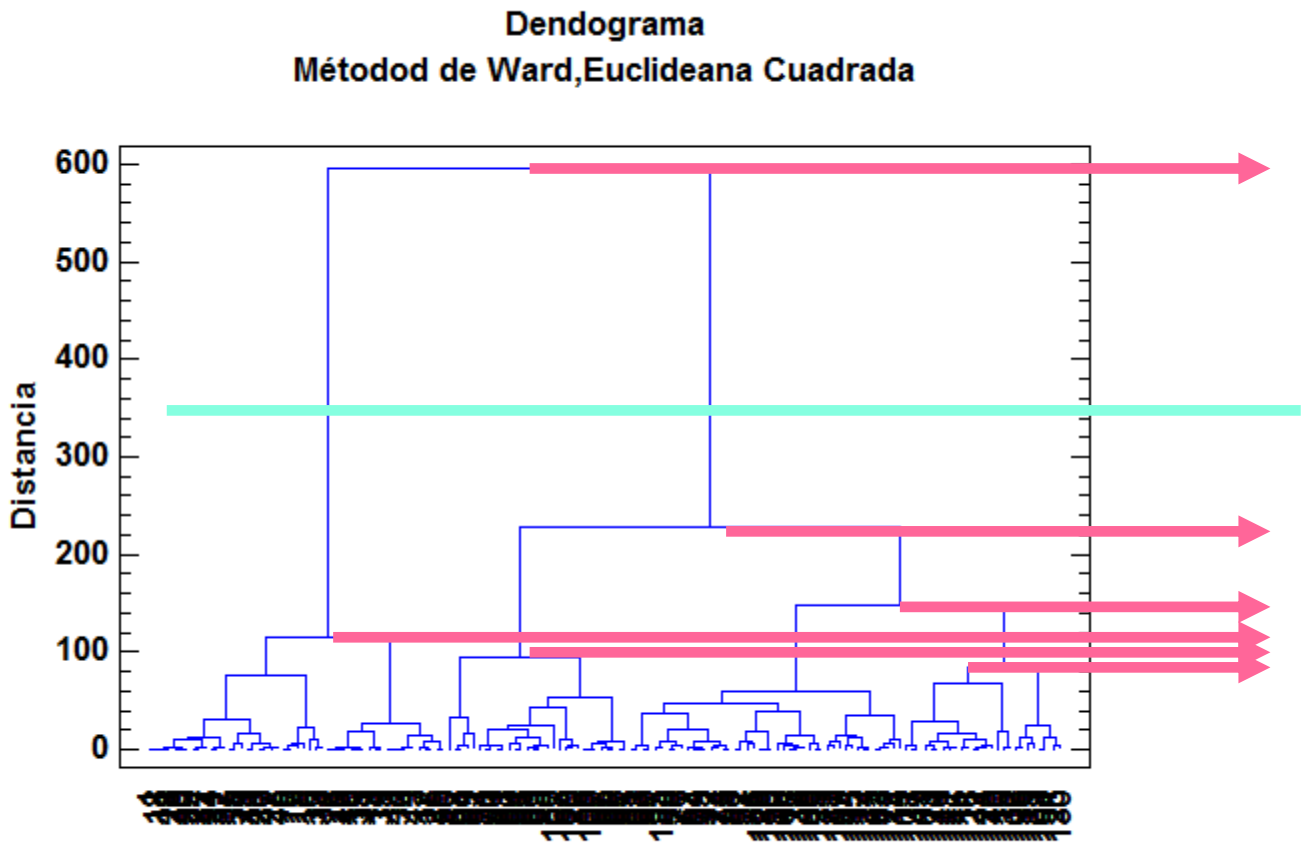
TRADE-OFF *Technical precision vs Interpretability*

Contribute with relevant knowledge for the domain

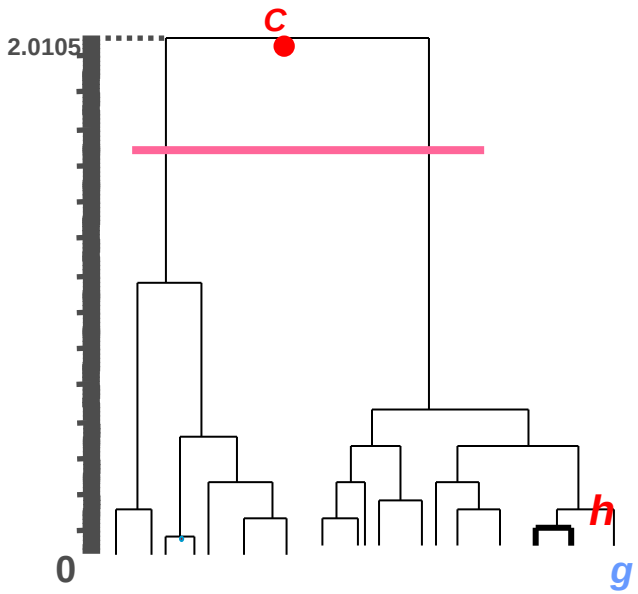
Calinski, T. and Harabasz, J. (1974) A dendrite method for cluster analysis. Communications in Statistics - Simulation and Computation, 3(1):1—27, Taylor and Francis


Dendograma
Método del Centroide,Euclideana Cuadrada
Distancia
12
10
8
6
4
2
0


Dendograma
Métodod de Ward,Euclideana Cuadrada
Distancia
600
500
400
300
200
100
0


Dendograma
Métodod de Ward,Euclideana Cuadrada
Distancia
600
500
400
300
200
100
0

Ascendant hierarchical clustering (cutting criteria)



Level indexes histogram

←idn(c) ©K. Gibert

How to cut a dendrogram *[Husson 2011]*

- Always cut HORIZONTALLY
- Look for levels with long branches *(biggest gaps)*
- Maximize the ratio (Husson 2011)

$$\min_{q_{min} \leq q \leq q_{max}} = \frac{\Delta(q)}{\Delta(q+1)}$$

$\Delta(q)$ Increase of inertia between classes of moving from q – 1 to q clusters

function HCPC (Hierarchical Clustering on Principal Components) in R

TRADE-OFF *Technical precision vs Interpretability*

Contribute with relevant knowledge for the domain

Husson et al. Exploratory multivariate analysis by example using R. Boca Raton: CRCPress, 2011