Probabilistische Variablenselektion in der Clusteranalyse

Gunter Ritter

Faculty of Informatics and Mathematics
University of Passau/Germany

ritter@fim.uni-passau.de

1. Introduction

Literature

Recent approaches to variable selection

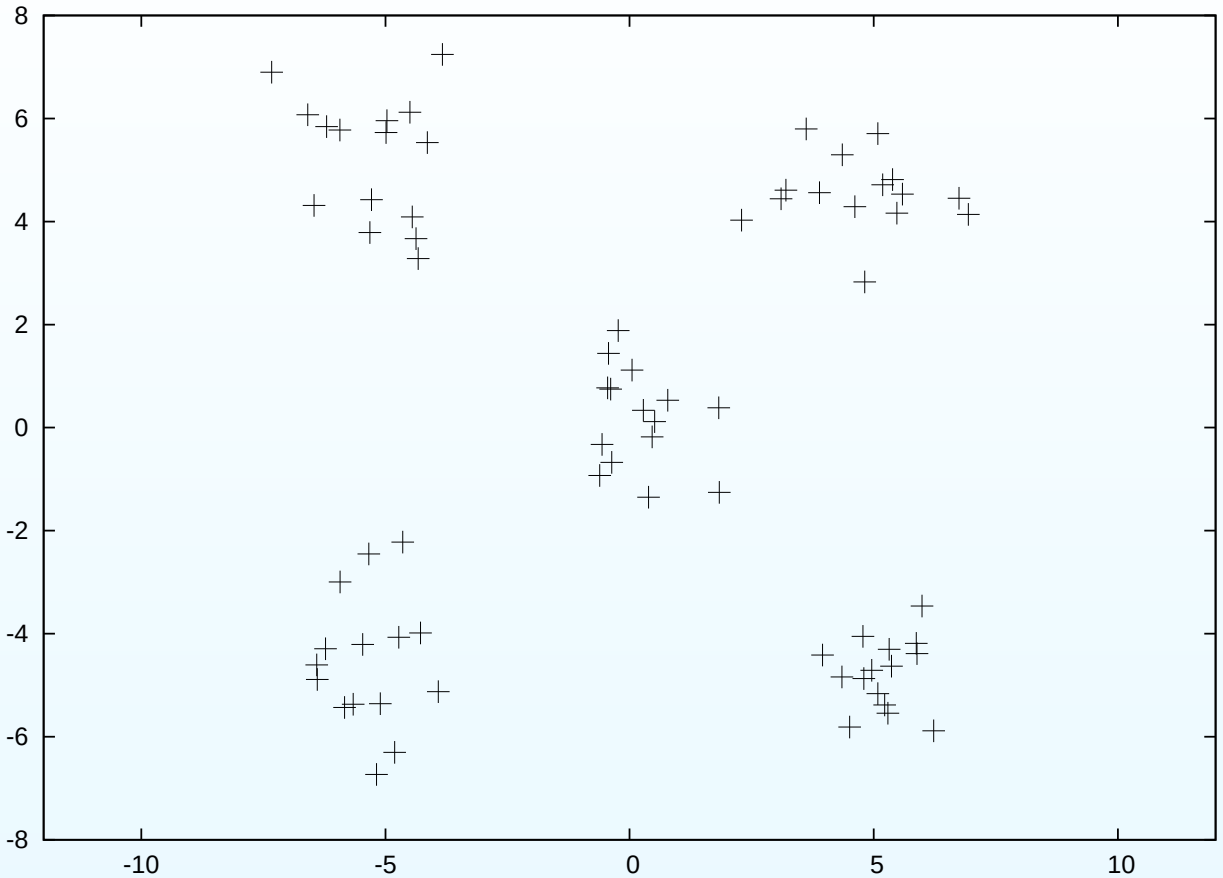
(a) Raftery and Dean (2006)

(b) Tyler et al. (2009)

(c) Hui and Lindsay (2010)

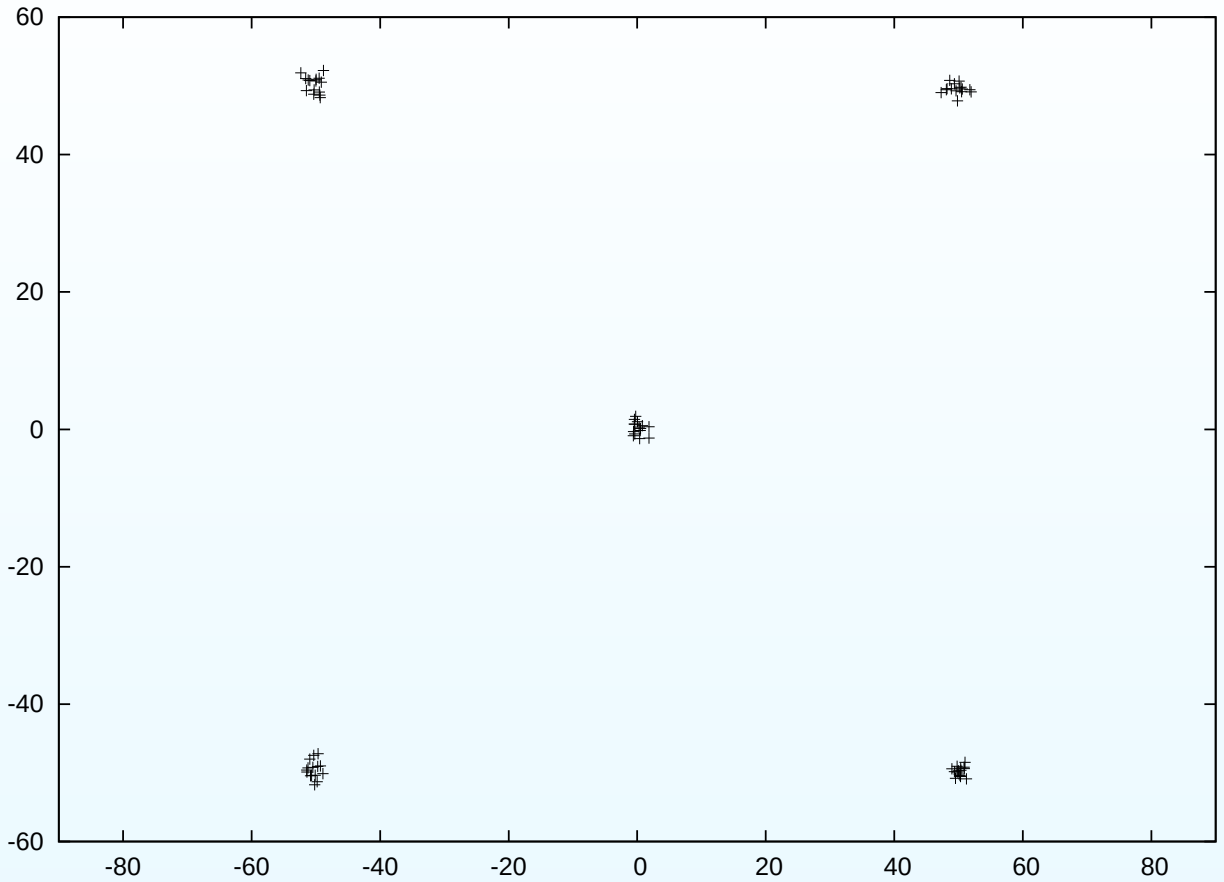
Experiment

Milligan 1980, Fowlkes, Gnanadesikan, and Kettenring 1988:



Det criterion: three additional noisy variables $\rightsquigarrow$ 28 errors.

Experiment



Det criterion $\rightsquigarrow$ 14 errors

2. Irrelevance in clustering

Irrelevance in clustering

Model for irrelevance (redundancy and noise)
John et al. 1994, Koller and Sahami 1996:

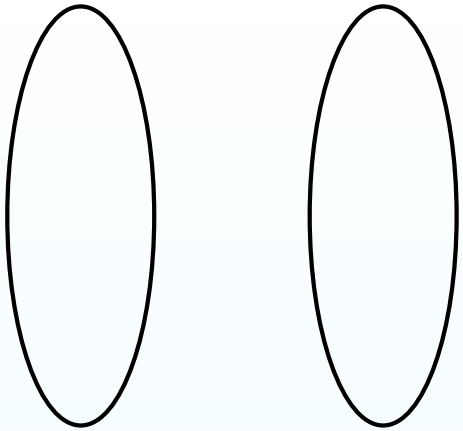
Let $F, E \subseteq 1..D$ be disjoint subsets of variables,
$L(i) = \ell_i$ label of object $i \in 1..n$.

(a) The subset E is irrelevant w.r.t. F if L is conditionally independent of X_E given X_F, that is, P-a.s for all j,

$$P[L = j \,|\, X_F, X_E] = P[L = j \,|\, X_F].$$

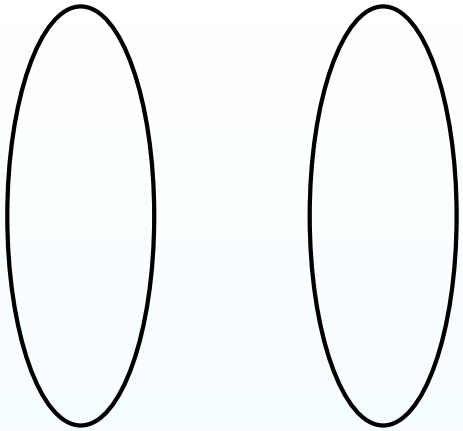
(b) The subset E is irrelevant if it is irrelevant w.r.t. its complement.

Examples

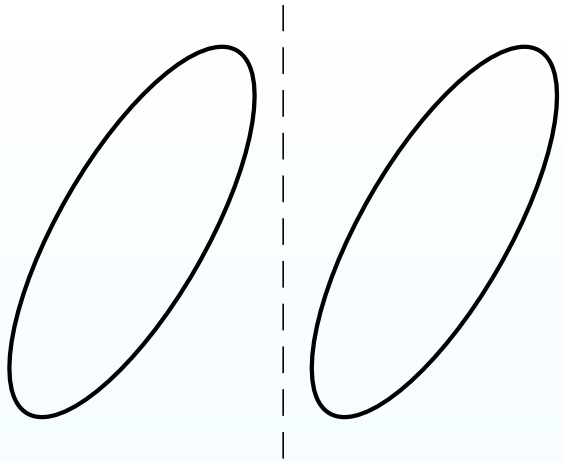


y uninformative, irrelevant | x

Examples

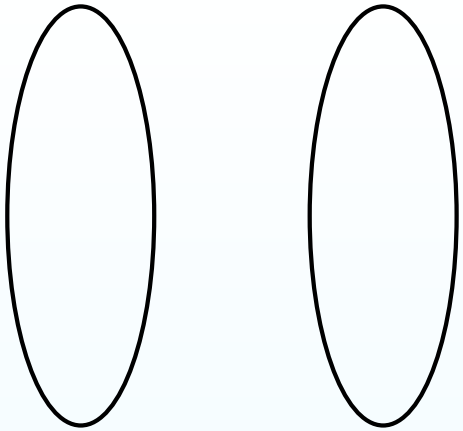


y uninformative, irrelevant | x

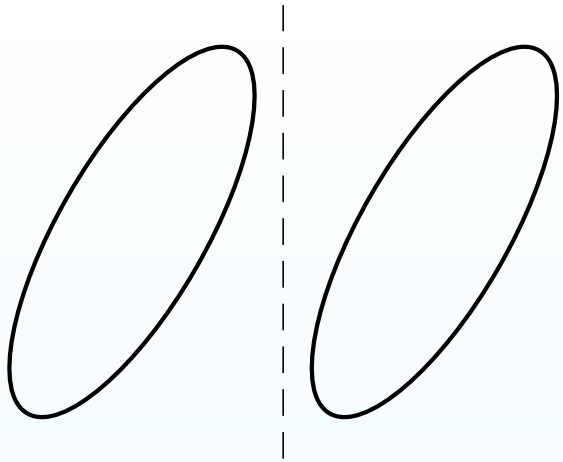


y uninformative, relevant | x

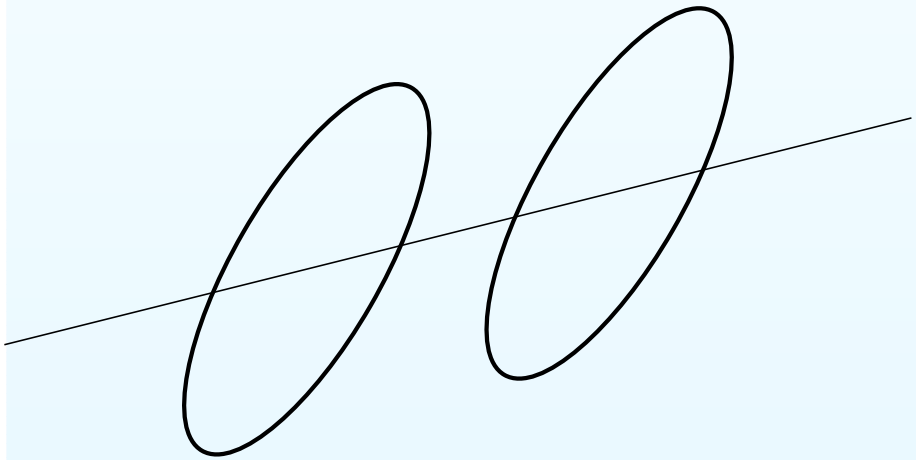
Examples



y uninformative, irrelevant | x



y uninformative, relevant | x



y informative, irrelevant | x

$$V = \begin{pmatrix} v_x & v_{y,x} \\ v_{y,x} & v_y \end{pmatrix}$$

$$\frac{m_y^{(2)} - m_y^{(1)}}{m_x^{(2)} - m_x^{(1)}} = \frac{v_{y,x}}{v_x}$$

Relevance decomposition

The subset $F \subseteq 1..D$ of variables is called structural if no subset $\emptyset \neq C \subseteq F$ is irrelevant w.r.t. $F \setminus C$.

Theorem. (Gallegos & R. 2015)

"Let the real random variables X_i, $i \in 1..D$, have a strictly positive and continuous joint Lebesgue density $f_{(X_1,...,X_D)}$. There exists exactly one structural subset $F \subseteq 1..D$ with irrelevant complement."

Normal case

Assume $\emptyset \neq F \subset 1..D,\ E = \complement F,\ x = (x_F, x_E) \ \Rightarrow\ f(x) = f(x_E \mid x_F) \cdot f(x_F)$

$X^{(j)} = (X_F^{(j)}, X_E^{(j)}) \sim N_{m_j, V_j}$ normal $\Rightarrow$

$$f(x_E \mid x_F) \sim X_E^{(j)} \mid x_F = m_{j,E|F} + G_{j,E|F} x_F + U_{E|F}^{(j)}, \qquad U_{E|F}^{(j)} \sim N_{0, V_{j,E|F}}$$

Normal case

Assume $\emptyset \neq F \subset 1..D,\ E = \complement F,\ x = (x_F, x_E) \ \Rightarrow\ f(x) = f(x_E \mid x_F) \cdot f(x_F)$

$X^{(j)} = (X_F^{(j)}, X_E^{(j)}) \sim N_{m_j, V_j}$ normal $\Rightarrow$

$$f(x_E \mid x_F) \sim X_E^{(j)} \mid x_F = m_{j,E|F} + G_{j,E|F} x_F + U_{E|F}^{(j)}, \qquad U_{E|F}^{(j)} \sim N_{0, V_{j,E|F}}$$

Theorem.

(a) If X is a normal mixture, covariance matrix $\mathrm{V}X_F$ invertible, then the following statements are equivalent.

(i) The subset E is irrelevant w.r.t. F;

(ii) the parameters $G_{j,E|F}$, $m_{j,E|F}$, and $V_{j,E|F}$ do not depend on j.

(b) In this case, these common parameters have the representations

(iii) $G_{E|F} = \mathrm{Cov}(X_E, X_F)(\mathrm{V}X_F)^{-1}$;

(iv) $m_{E|F} = m_E - G_{E|F} m_F$;

(v) $V_{E|F} = V_E - G_{E|F}\mathrm{Cov}(X_F, X_E)$.

3. Variable selection algorithm

Digression: Clustering, normal case

(a) Determinant criterion (Symons 1981)

$$\frac{1}{2}\sum_{j=1}^{g} n_j(\boldsymbol{\ell}) \log\det S_j(\boldsymbol{\ell}) + n\mathrm{H}\Big(\frac{n_1(\boldsymbol{\ell})}{n}, \ldots, \frac{n_g(\boldsymbol{\ell})}{n}\Big)$$

$$\text{entropy } \mathrm{H}(p_1, \ldots, p_g) = -\sum_j p_j \log p_j$$

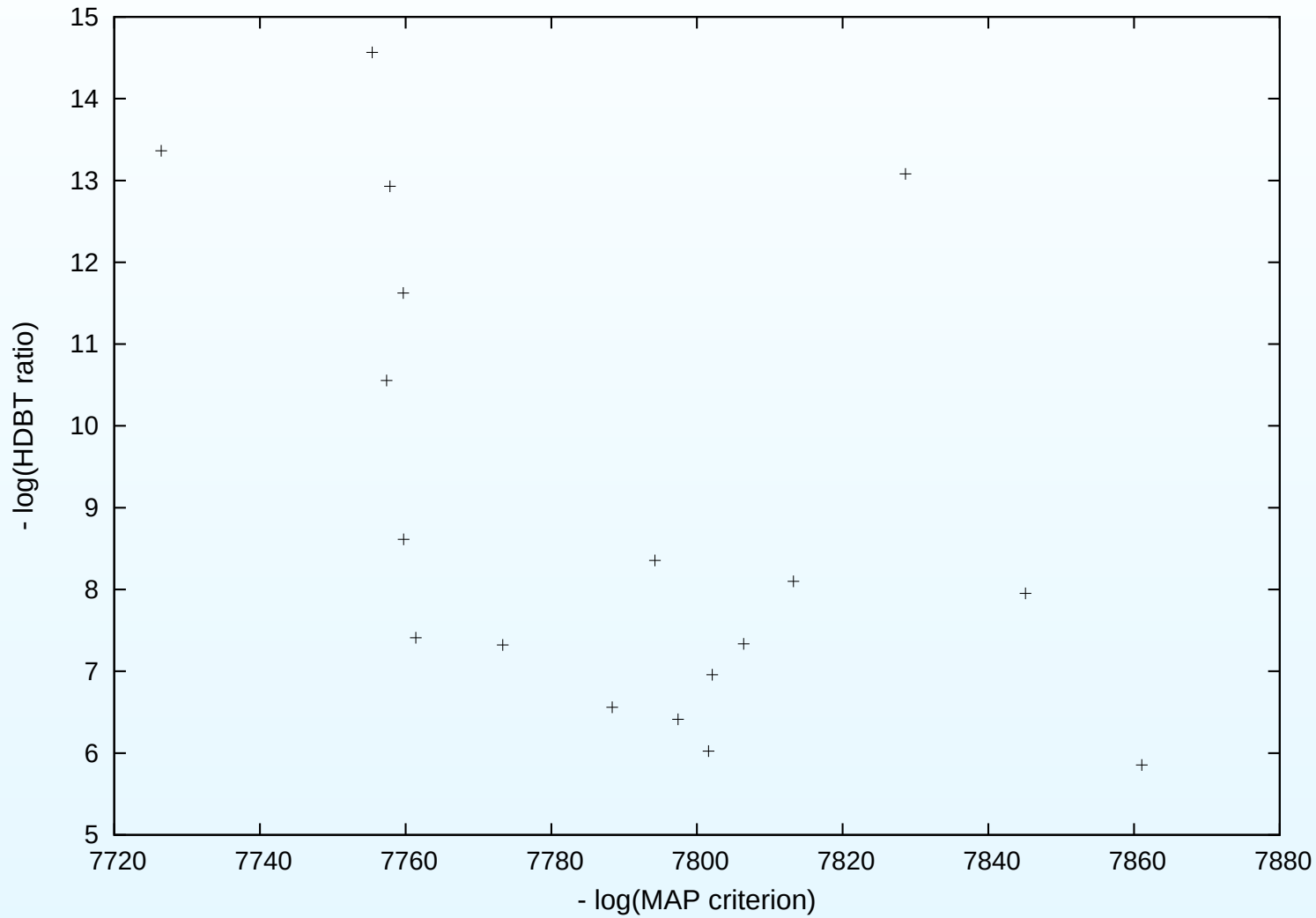
(b) "k-Parameters algorithm" (Schroeder (1976)

Alternate (normal) parameter estimation and Bayesian discriminant analysis (assignment) until stationarity.

(c) Special case: Ward's sum-of-squares criterion (1963), "k-means".

Which "local" minimum? SBF plot

Multiple solutions $\rightsquigarrow$ Random starts



Clustering and selection, normal case

(a) Model

$$\sum_{j=1}^{g} \sum_{i:\ell_i=j} \log N_{m_j,V_j}(x_{i,F}) - n\mathrm{H}\big(\tfrac{n_1(\boldsymbol{\ell})}{n}, \ldots, \tfrac{n_g(\boldsymbol{\ell})}{n}\big)$$
$$+ \sum_{i} \log N_{m_{E|F},V_{E|F}}(x_{i,E} - G_{E|F}x_{i,F}).$$

Parameter $m_j,\ V_j,\ m_{E|F},\ V_{E|F},\ G_{E|F},\ F$

Clustering and selection, normal case

(a) Model

$$\sum_{j=1}^{g} \sum_{i:\ell_i=j} \log N_{m_j,V_j}(x_{i,F}) - n\mathrm{H}\big(\tfrac{n_1(\boldsymbol{\ell})}{n}, \ldots, \tfrac{n_g(\boldsymbol{\ell})}{n}\big)$$
$$+ \sum_{i} \log N_{m_{E|F},V_{E|F}}(x_{i,E} - G_{E|F} x_{i,F}).$$

Parameter $m_j,\ V_j,\ m_{E|F},\ V_{E|F},\ G_{E|F},\ F$

(b) Determinant criterion with selection

$$\tfrac{1}{2} \sum_{j=1}^{g} n_j(\boldsymbol{\ell}) \log\det S_{j,F}(\boldsymbol{\ell}) + n\mathrm{H}\big(\tfrac{n_1(\boldsymbol{\ell})}{n}, \ldots, \tfrac{n_g(\boldsymbol{\ell})}{n}\big) + \tfrac{n}{2} \log\det S_{E|F}.$$

$S_{j,F}(\boldsymbol{\ell})$ scatter matrix of cluster j, $S_{E|F}$ residual scatter matrix

$$\det S = \det S_F \cdot \det S_{E|F} \quad \Rightarrow \quad \det S_{E|F} \sim -\det S_F$$

$$\tfrac{1}{2} \sum_{j=1}^{g} n_j(\boldsymbol{\ell}) \log\det S_{j,F}(\boldsymbol{\ell}) + n\mathrm{H}\big(\tfrac{n_1(\boldsymbol{\ell})}{n}, \ldots, \tfrac{n_g(\boldsymbol{\ell})}{n}\big) - \tfrac{n}{2} \log\det S_F.$$

$$\big(\boldsymbol{\ell}^{(0)}, F^{(0)}\big) \longrightarrow \big(\boldsymbol{\ell}^{(0)}, F^{(1)}\big) \longrightarrow \big(\boldsymbol{\ell}^{(1)}, F^{(1)}\big) \longrightarrow \ldots.$$

Clustering and selection procedure (wrapper)

$\ell : 1..n \to 1..g$

// Input: Subset $F \subseteq 1..D$, $|F| = d$, admissible ℓ, and value of the criterion.
// Output: New quantities F_{new} and ℓ_{new}, with improved criterion *or* "stop."

1. *(Estimation)* Compute the sample mean vectors $\overline{x}_j(\ell)$ and scatter matrices $S_j(\ell)$, $1 \le j \le g$, and the total scatter matrix S.

2. *(Selection)* Minimize

$$h(F') = \sum_{j=1}^{g} n_j(\ell) \log\det S_{j,F'}(\ell) - n \log\det S_{F'} \quad (\le h(F))$$

w.r.t. F', $|F'| = d$. Denote the minimizer by F_{new}.

Easily attained by sorting if variables independent!

Selection procedure (wrapper)

3. Use the quantities from step 1 to compute the MLE's of the regression parameters (G, m, V) w.r.t. ℓ and the new subsets F_{new} and $E_{\text{new}} = \complement F_{\text{new}}$.

   Let

$$\begin{aligned} u_{i,j} = & \log n_j - \tfrac{1}{2} \log\det S_{j,F_{\text{new}}}(\ell) \\ & - \tfrac{1}{2}(x_{i,F_{\text{new}}} - \overline{x}_{j,F_{\text{new}}}(\ell))^\top S_{j,F_{\text{new}}}(\ell)^{-1}(x_{i,F_{\text{new}}} - \overline{x}_{j,F_{\text{new}}}(\ell)) \\ & - \tfrac{1}{2}(x_{i,E_{\text{new}}} - m - Gx_{i,F})^\top V^{-1}(x_{i,E_{\text{new}}} - m - Gx_{i,F}). \end{aligned}$$

4. *(Assignment and trimming)* Compute an admissible assignment ℓ_{new} using a reduction step based on the statistics $u_{i,j}$.

5. *(Decision)*
   If F_{new} and ℓ_{new} improve the criterion then return F_{new}, ℓ_{new},
   else "stop".

Literatur

Guodong Hui and Bruce G. Lindsay. Projection pursuit via white noise matrices. *Sankhyā, Series B*, 72:123–153, 2010.

Adrian E. Raftery and Nema Dean. Variable selection for model-based clustering. *J. Amer. Stat. Assoc.*, 101:168–178, 2006.

Gunter Ritter. *Robust Cluster Analysis and Variable Selection*, volume 137 of *Monographs in Statistics and Applied Probability*. Chapman & Hall/CRC, Boca Raton, London, New York, 2015.

David E. Tyler, Frank Critchley, Lutz Dümbgen, and Hannu Oja. Invariant co-ordinate selection. *J. Royal Statist. Soc., Series B*, 71:549–592, 2009. With discussion and rejoinder.