A history of the k-means algorithm

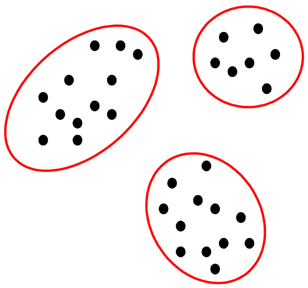
Hans-Hermann Bock, RWTH Aachen, Allemagne

RWTH RHEINISCH-WESTFÄLISCHE TECHNISCHE HOCHSCHULE AACHEN

CNAM, Paris, September 4, 2007

Published version: H.-H. Bock: Clustering methods: a history of k-means algorithms.
In: P. Brito et al. (eds.): Selected contributions in data analysis and classification. Springer Verlag, Heidelberg, 2007, 161-172

1. Clustering with SSQ and the k-means algorithm



Given: $\mathcal{O} = \{1, ..., n\}$ set of n objects

$x_1, ..., x_n \in \mathbb{R}^p$ n data vectors

Problem: Determine a partition $\mathcal{C} = (C_1, ..., C_k)$ of $\mathcal{O}$

with k classes $C_i \subset \mathcal{O}$, $i = 1, ..., k$

characterized by class prototypes: $\mathcal{Z} = (z_1, ..., z_k)$

Clustering criterion: SSQ, variance criterion, trace criterion, inertie,...

$$\boxed{g(\mathcal{C}) \; := \sum_{i=1}^{k} \sum_{\ell \in C_i} ||x_\ell - \overline{x}_{C_i}||^2 \quad \longrightarrow \quad \min_{\mathcal{C}}}$$

with class centroids (class means) $z_1^* = \overline{x}_{C_1}, ..., \; z_k^* = \overline{x}_{C_k}$.

Two-parameter form:

$$\boxed{g(\mathcal{C}, \mathcal{Z}) \; := \sum_{i=1}^{k} \sum_{\ell \in C_i} ||x_\ell - z_i||^2 \quad \longrightarrow \quad \min_{\mathcal{C}, \mathcal{Z}}}$$

Remark: $g(\mathcal{C}) \equiv g(\mathcal{C}, \mathcal{Z}^*)$

The well-known k-means algorithm

- produces a sequence of partitions/prototype systems: $\mathcal{C}^{(0)}, \mathcal{Z}^{(0)}, \mathcal{C}^{(1)}, \mathcal{Z}^{(1)}, \ldots$

$t = 0$:

Start from an arbitrary initial partition $\mathcal{C}^{(0)} = (C_1^{(0)}, \ldots, C_k^{(0)})$ of $\mathcal{O}$

$t \longrightarrow t+1$:

(I) Calculate system $\mathcal{Z}^{(t)}$ of class centroids for $\mathcal{C}^{(t)}$:

$$z_i^{(t)} := \overline{x}_{C_i^{(t)}} = \frac{1}{|C_i^{(t)}|} \sum_{\ell \in C_i} x_\ell$$

Problem A:

$$g(\mathcal{C}^{(t)}, \mathcal{Z}) \longrightarrow \min_{\mathcal{Z}}$$

(II) Determine the min-dist partition $\mathcal{C}^{(t+1)}$ for $\mathcal{Z}^{(t)}$:

$$C_i^{(t+1)} := \{\ell \in \mathcal{O} \mid ||x_\ell - z_i^{(t)}|| = \min_j ||x_\ell - z_j^{(t)}||\}$$

Problem B:

$$g(\mathcal{C}, \mathcal{Z}^{(t)}) \longrightarrow \min_{\mathcal{C}}$$

Stopping:

Iterate until stationarity, i.e., $g(\mathcal{C}^{(t)}) = g(\mathcal{C}^{(t+1)})$

$$\boxed{g(\mathcal{C}, \mathcal{Z}) \;:=\; \sum_{i=1}^{k} \sum_{\ell \in C_i} ||x_\ell - z_i||^2 \quad \longrightarrow \quad \min_{\mathcal{C}, \mathcal{Z}}}$$

Remarks: This two-parameter form contains a continuous ($\mathcal{Z}$) and a discrete ($\mathcal{C}$) variable.
The k-means algorithm is a relaxation algorithm (in the mathematical sense).

Theorem:

The k-means algorithm

$$\begin{aligned} \mathcal{Z}^{(t)} &:= \mathcal{Z}(\mathcal{C}^{(t)}) \\ \mathcal{C}^{(t+1)} &:= \mathcal{C}(\mathcal{Z}^{(t)}) \qquad\qquad t = 0, 1, 2, \ldots \end{aligned}$$

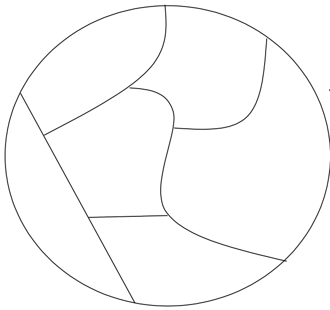
produces m-partitions $\mathcal{C}^{(t)}$ and prototype systems $\mathcal{Z}^{(t)}$
with steadily decreasing criterion values:

$$g(\mathcal{C}^{(t)}) \equiv g(\mathcal{C}^{(t)}, \mathcal{Z}^{(t)}) \;\geq\; g(\mathcal{C}^{(t+1)}, \mathcal{Z}^{(t)}) \;\geq\; g(\mathcal{C}^{(t+1)}, \mathcal{Z}^{(t+1)}) \equiv g(\mathcal{C}^{(t+1)})$$

Continuous version of the SSQ criterion:

Given: A random vector X in $\mathbb{R}^p$ with known distribution P, density $f(x)$

Problem: Find an 'optimal' partition $\mathcal{B} = (B_1, ..., B_k)$ of $\mathbb{R}^p$
with k Borel sets (classes) $B_i \subset \mathbb{R}^p$, $i = 1, ..., k$
characterized by class prototypes: $\mathcal{Z} = (z_1, ..., z_k)$



- **Continuous version of SSQ criterion:**

$$\boxed{G(\mathcal{B}) := \sum_{i=1}^{k} \int_{B_i} ||x - E[X|X \in B_i]||^2 \, dP(x) \quad \to \quad \min_{\mathcal{B}}}$$

with class centroids (expectations) $z_1^* = E[X|X \in B_1], ..., z_k^* = E[X|X \in B_k]$.

- **Two-parameter form:**

$$\boxed{G(\mathcal{B}, \mathcal{Z}) := \sum_{i=1}^{k} \int_{B_i} ||x - z_i||^2 \, dP(x) \quad \to \quad \min_{\mathcal{B}, \mathcal{Z}}}$$

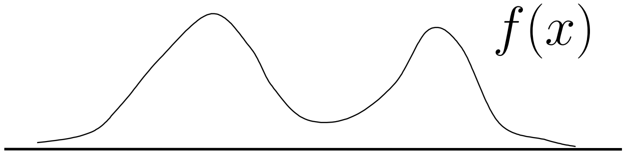
$\Longrightarrow$ **Continuous version of the k-means algorithm**

2. Continuous SSQ clustering for stratified sampling

Dalenius (1950), Dalenius/Gurney (1951)

Given: A random variable (income) X in $\mathbb{R}$ with density $f(x)$

$$\mu := E[X], \qquad \sigma^2 := Var(X)$$



Problem: Estimate unknown expected income μ by using n samples (persons)

- **Strategy I: Simple random sampling**

  Sample n persons, observed income values $x_1, ..., x_n$

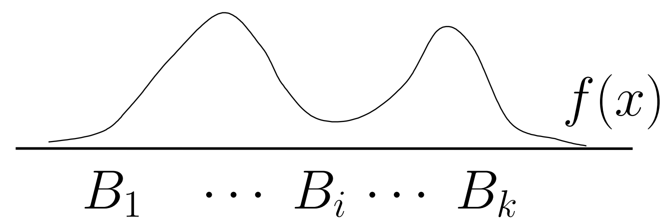
  Estimator: $\hat{\mu} := \overline{x} = \frac{1}{n} \sum_{j=1}^{n} x_j$

  Performance: $E[\hat{\mu}] = \mu$ unbiasedness

  $Var(\hat{\mu}) = \sigma^2/n.$

- **Strategy II: Stratified sampling**

Partitioning $I\!R$ into k classes (strata): $B_1, ..., B_k$
Class probabilities: $p_1, ..., p_k$



Sampling from stratum B_i: $\quad Y_i \sim X|X \in B_i$

$$\mu_i := E[Y_i] = E[X|X \in B_i]$$
$$\sigma_i^2 := Var(Y_i) = Var(X|X \in B_i)$$

Sampling: n_i samples from B_i: $y_{i1}, ..., y_{in_i}$ $\qquad (\sum_{i=1}^k n_i = n)$

$$\hat{\mu}_i := \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij}$$

Estimator: $\quad \hat{\hat{\mu}} := \sum_{i=1}^k p_i \cdot \hat{\mu}_i$

Performance: $\quad E[\hat{\hat{\mu}}] = \mu$ (unbiasedness)

$$Var(\hat{\hat{\mu}}) = \sum_{i=1}^k \frac{p_i^2}{n_i} \cdot \sigma_i^2 = \sum_{i=1}^k \frac{p_i}{n_i} \int_{B_i} (x - \mu_i)^2 dP(x) \ \leq \ \sigma^2/n$$

- **Strategy III: Proportional stratified sampling**

Use sample sizes proportional to class frequencies: $n_i = n \cdot p_i$ $\qquad \Longrightarrow$

- **Strategy III: Proportional stratified sampling**

  Use sample sizes proportional to class frequencies: $n_i = n \cdot p_i$

  $\Longrightarrow$ Resulting variance:

$$Var(\hat{\hat{\mu}}) \;=\; \frac{1}{n}\sum_{i=1}^{k}\int_{B_i}(x-\mu_i)^2 dP(x) \;=\; \frac{1}{n}\cdot G(\mathcal{B}) \;\rightarrow\; \min_{\mathcal{B}}$$

Implication:

Optimum stratification $\equiv$ Optimum SSQ clustering

Remark: Dalenius did **not** use the k-means algorithm for determining $\mathcal{B}$!

3. Les origines: historical k-means approaches

- **Steinhaus (1956):**

$\mathcal{X}$

$\mathcal{X} \subset I\!R^p$ a solid (mechanics; similarly: anthropology, industry)
with mass distribution density $f(x)$

Problem:

Dissecting $\mathcal{X}$ into k parts $B_1, ..., B_k$
such that sum of class-specific inertias is minimized:

$$G(\mathcal{B}) := \sum_{i=1}^{k} \int_{B_i} ||x - E[X|X \in B_i]||^2 \, f(x)dx \;\; \rightarrow \;\; \min_{\mathcal{B}}$$

Steinhaus proposes: Continuous version of k-means algorithm

Steinhaus discusses:
– Existence of a solution
– Uniqueness of the solution
– Asymptotics for $k \rightarrow \infty$

- **Lloyd (1957):**

Quantization in information transmission: Pulse-code modulation

Problem: Transmitting a p-dimensional random signal X with density f(x)

Method:

Instead of transmitting the original message (value) x

– we select k different fixed points (code vectors) $z_1, ..., z_k \in \mathbb{R}^p$

– we determine the (index of the) code vector that is closest to x:

$$i(x) = argmin_{j=1,...,k}\{||x - z_j||^2\}$$

– transmit only the index $i(x)$

– and decode the message x by the code vector $\hat{x} := z_{i(x)}$.

Expected transmission (approximation) error:

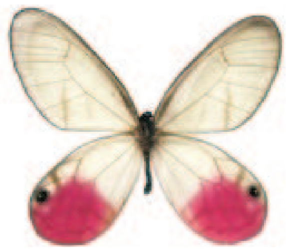
$$\gamma(z_1, ..., z_k) \;:=\; \int_{\mathbb{R}^p} \min_{j=1,...,k}\{\;||x - z_j||^2\;\}f(x)dx \;=\; G(\;\mathcal{B}(\mathcal{Z}), \mathcal{Z})$$

where $\mathcal{B}(\mathcal{Z})$ is the minimum-distance partition of $\mathbb{R}^p$ generated by $\mathcal{Z} = \{z_1, ..., z_m\}$.

Lloyd's Method I: Continuous version of k-means (in $\mathbb{R}^1$)

- **Forgy (1965), Jancey (1966):**

Taxonomy of genus Phyllota Benth. (Papillionaceae)

$x_1, ..., x_n$ are feature vectors characterizing n butterflies

Forgy's lecture proposes the discrete k-means algorithm

(implying the SSQ clustering criterion only implicitly!)

A strange story:

– only indirect communications by Jancey, Anderberg, MacQueen

– nevertheless often cited in the data analysis literature

Terminology:

k-means: – iterated minimum-distance partitioning (Bock 1974)
- nuées dynamiques (Diday et al. 1974)
- dynamic clusters method (Diday et al. 1973)
- nearest centroid sorting (Anderberg 1974)
- HMEANS (Späth 1975)

However: MacQueen (1967) has coined the term 'k-means algorithm' for a sequential version:

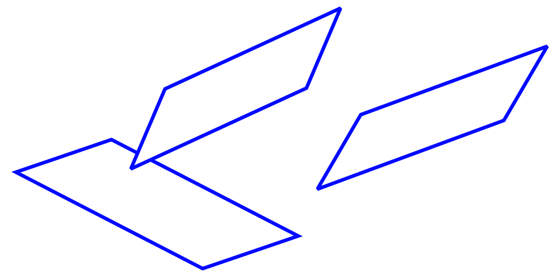
- Processing the data points x_s in a sequential order: s=1,2,...
- Using the first k data points as 'singleton' classes ($=$ centroids)
- Assigning a new data point x_{s+1} to the closest class centroid from step s
- Updating the corresponding class centroid after the assignment

Various authors use 'k-means' in this latter (and similar) sense (Chernoff 1970, Sokal 1975)

4. La Belle Epoque: Generalized k-means algorithms

for clustering criteria of the type:

$$\boxed{g(\mathcal{C}, \mathcal{Z}) := \sum_{i=1}^{m} \sum_{k \in C_i} d(k, z_i) \quad \rightarrow \quad \min_{\mathcal{C}, \mathcal{Z}}}$$



where $\mathcal{Z} = (z_1, ..., z_m)$ is a system of 'class prototypes'

and $d(k, z_i)$ = dissimilarity between

– the object k (the data point x_k) and

– the class C_i (the class prototype z_i)

Great flexibility in the choice of d and the structure of prototypes z_i:

– Other metrics than Euclidean metric

– Other definitions of a 'class prototype' (subsets of objects, hyperplanes,...)

– Probabilistic clustering models (centroids $\leftrightarrow$ m.l. estimation)

– New data types: similarity/dissimilarity matrices, symbolic data, ...

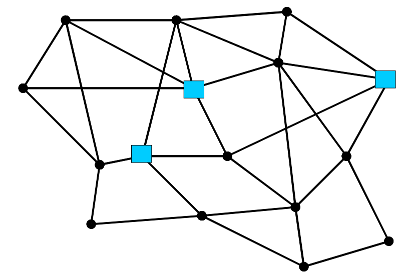
– Fuzzy clustering

• Maranzana (1963): k-means in a graph-theoretical setting

Situation: Industrial network with n factories: $\mathcal{O} = \{1, ..., n\}$

Pairwise distances $d(\ell, t)$,

e.g., minimum road distance, transportation costs



Problem: Transporting commodities from the factories

to k suitable warehouses as follows:

– Partition $\mathcal{O}$ into k classes $C_1, ..., C_k$

– Select, for each class C_i, one factory $z_i \in \mathcal{O}$ as 'class-specific warehouse' (products from a factory $\ell \in C_i$ are transported to z_i for storing)

– Minimize the transportation costs:

$$\boxed{g(\mathcal{C}, \mathcal{Z}) \quad := \quad \sum_{i=1}^{k} \sum_{\ell \in C_i} d(\ell, z_i) \to \min_{\mathcal{C}, \mathcal{Z}} \qquad \text{with } z_i \in C_i \text{ for } i = 1, ..., m}$$

$\Rightarrow$ **k-means-type algorithm**: Determining the 'class prototypes' z_i by:

$$\boxed{Q(C_i, z) := \sum_{\ell \in C_i} d(\ell, z) \ \to \ \min_{z \in C_i}}$$

Kaufman/Rousseeuw (1987): medoid of C_i, partitioning around medoids

• **Diday (1971,...), Bock (1968,...), Govaert (1974), Charles (1977),...:**

$$\boxed{g(\mathcal{C}, \mathcal{Z}) \;:=\; \sum_{i=1}^{m} \sum_{k \in C_i} d(k, z_i) \quad \longrightarrow \quad \min_{\mathcal{C}, \mathcal{Z}}}$$

?? **??** **??**

– Kernel clustering: prototype z_i = a subset of C_i with $|z_i| = 4$, say

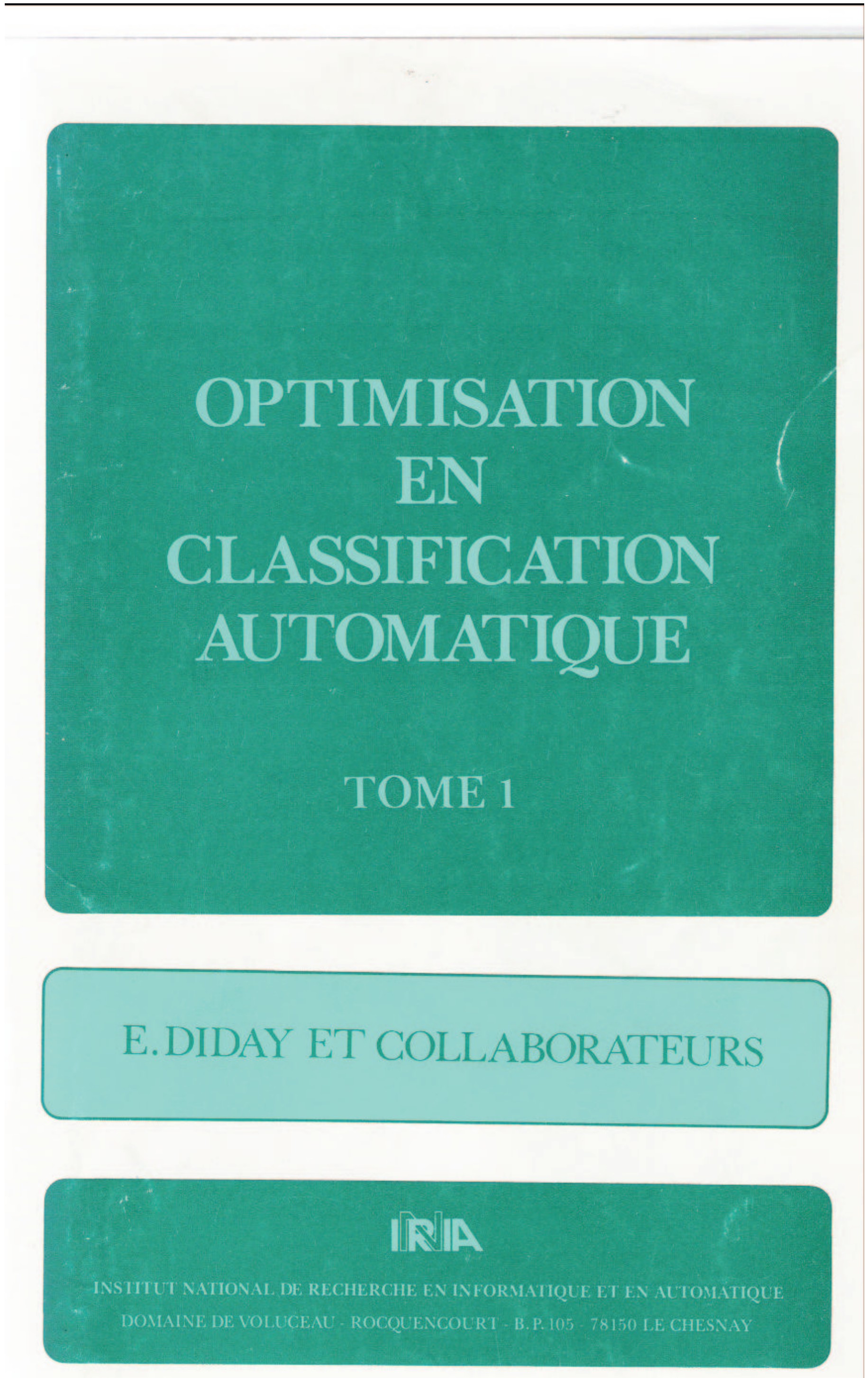
– Determinantal criterion: $d(x_\ell, z_i) = ||x_\ell - z_i||_Q^2$ with $det(Q) = 1$

– Adaptive distance clustering: $d(x_\ell, z_i) = ||x_\ell - z_i||_{Q_i}^2$ with $det(Q_i) = 1$

– Principal component clustering: Prototypes z_i are class-specific hyperplanes

– Regression clustering: Prototypes z_i are class-specific regression hyperplanes

– Projection pursuit clustering: Prototypes $z_1, ..., z_k$ on the same low-dim. hyperplane


OPTIMISATION
EN
CLASSIFICATION
AUTOMATIQUE
TOME 1
E. DIDAY ET COLLABORATEURS
INRIA
INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE
DOMAINE DE VOLUCEAU - ROCQUENCOURT - B.P. 105 - 78150 LE CHESNAY

• **Diday & Schroeder (1974 ff.), Sclove (1977):**

Classification maximum likelihood, fixed-partition model, model-based clustering:

Model: $X_1, ..., X_n$ independent random vectors, density family $f(\bullet; z)$

Exists a k-partition $\mathcal{C} = (C_1, ..., C_k)$ of $\mathcal{O} = \{1, ..., n\}$

Exist k class-specific parameter vectors $z_1, ..., z_k$

such that

$$\boxed{X_\ell \quad \sim \quad f(\bullet; z_i) \qquad \text{for all } \ell \in C_i}$$

Maximum likelihood estimation of $\mathcal{C}$ and $\mathcal{Z} = (z_1, ..., z_k)$:

$$\Longrightarrow \boxed{g(\mathcal{C}, \mathcal{Z}) \quad := \quad \sum_{i=1}^{k} \sum_{\ell \in C_i} [-\log f(x_\ell; z_i)] \quad \rightarrow \quad \min_{\mathcal{C}, \mathcal{Z}}}$$

A two-parameter clustering criterion !

$\Longrightarrow$ A **generalized k-means algorithm** alternating

- class-specific m.l. estimation of parameters z_i
- minimum-distance (maximum likelihood) assignment of all data points

5. Les temps modernes: Convexity-based criteria and k-tangent algorithm

$$g(\mathcal{C}) \;:=\; \sum_{i=1}^{k} \sum_{\ell \in C_i} ||x_\ell - \overline{x}_{C_i}||^2 \;=\; \sum_{\ell=1}^{n} ||x_\ell||^2 - \underbrace{\sum_{i=1}^{k} |C_i| \cdot ||\overline{x}_{C_i}||^2} \;\rightarrow\; \min_{\mathcal{C}}$$

Equivalent, with the convex function $\phi(x) := ||x||^2$:

$$\boxed{G_n(\mathcal{C}) \;:=\; \frac{1}{n} \sum_{i=1}^{k} |C_i| \cdot ||\overline{x}_{C_i}||^2 \;=\; \sum_{i=1}^{k} \frac{|C_i|}{n} \cdot \phi(\overline{x}_{C_i}) \quad \rightarrow \quad \max_{\mathcal{C}}}$$

Continuous analogue for random vector $X \sim P$ in $\mathbb{R}^p$:

$$\boxed{G(\mathcal{B}) \;:=\; \sum_{i=1}^{k} P(X \in B_i) \cdot \phi(E[X | X \in B_i]) \;\rightarrow\; \max_{\mathcal{B}}}$$

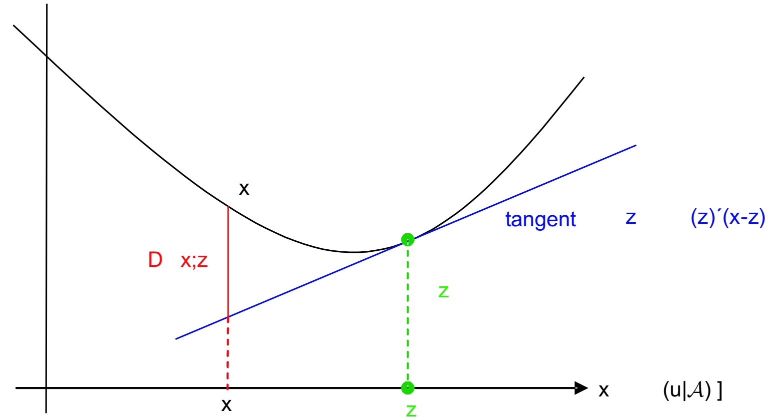
– Is this a relevant problem for practice?

– Is there an analogue to the k-means algorithm for SSQ?

– How to find an equivalent *two*-parameter criterion?

Reminder:

For each 'support point' $z \in R^p$, the convex function $\phi(x)$ has a support (tangent) hyperplane



$$t(x; z) := \phi(z) + a^{tr}(x - z)$$

with a slope vector $a = \nabla_x \phi(x)_{x=z} \in I\!R^p$ and

$$\begin{aligned} \phi(x) &\geq t(x; z) \qquad \text{for all } x \in I\!R^p \\ \phi(z) &= t(z; z) \qquad \text{for } x = z. \end{aligned}$$

Original clustering problem:

$$\boxed{G(\mathcal{B}) \;:=\; \sum_{i=1}^{k} P(X \in B_i) \cdot \phi(E[X|X \in B_i]) \to \max_{\mathcal{B}}}$$

Equivalent dual two-parameter problem:

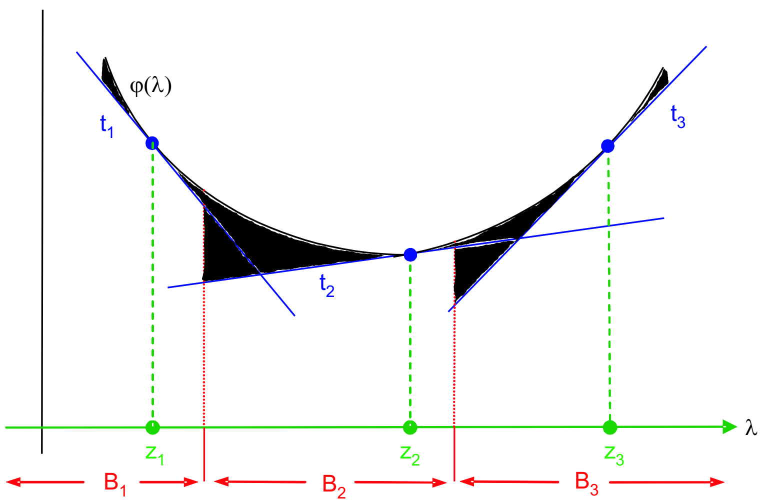
Looking for k support points $z_1, \ldots, z_m \in I\!R^p$

and corresponding tangents (hyperplanes)

$$t(x; z_i) := \phi(z_i) + a_i^{tr}(x - z_i)$$

such that

$$\boxed{\widetilde{G}(\mathcal{B}, \mathcal{Z}) := \sum_{i=1}^{k} \int_{B_i} [\phi(x) - t(x; z_i)]dP(x) \to \min_{\mathcal{B},\mathcal{Z}}}$$



"Minimum volume problem"

Original clustering problem:

$$\boxed{G(\mathcal{B}) := \sum_{i=1}^{k} P(X \in B_i) \cdot \phi(E[X|X \in B_i]) \to \max_{\mathcal{B}}}$$

Equivalent dual two-parameter problem:

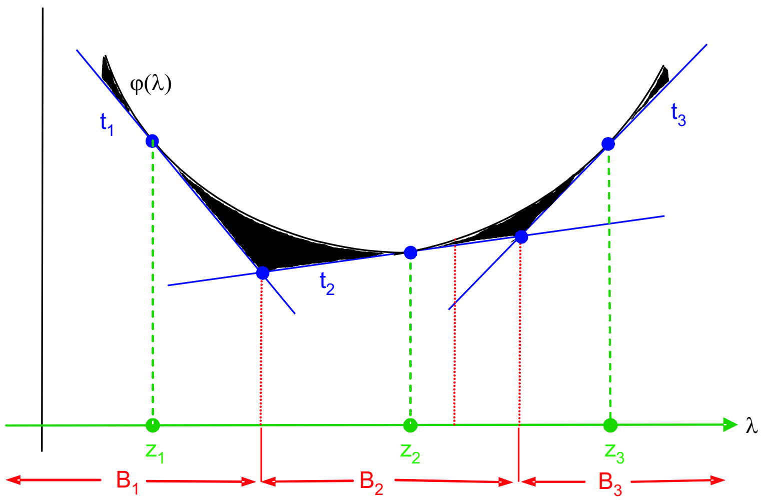
Looking for k support points $z_1, ..., z_m \in I\!R^p$

and corresponding tangents (hyperplanes)

$$t(x; z_i) := \phi(z_i) + a_i^{tr}(x - z_i)$$

such that

$$\boxed{\widetilde{G}(\mathcal{B}, \mathcal{Z}) := \sum_{i=1}^{k} \int_{B_i} [\phi(x) - t(x; z_i)] dP(x) \to \min_{\mathcal{B},\mathcal{Z}}}$$



"Minimum volume problem"

Alternating minimization: k-tangent clustering algorithm

(I) Partial minimization w.r.to the support point system $\mathcal{Z} = (z_1, ..., z_m)$:

$$\min_{\mathcal{Z}} \ \widetilde{G}(\mathcal{B}, \mathcal{Z}) = \tilde{G}(\mathcal{B}, \mathcal{Z}^*)$$

yields the system $\mathcal{Z}^* = (z_1^*, ..., z_m^*)$ of class centroids $z_i^* := E[X|X \in B_i]$.

(II) Partial minimization w.r.t. the partition $\mathcal{B} = (B_1, ..., B_m)$ of $\mathbb{R}^p$:

$$\min_{\mathcal{B}} \ \widetilde{G}(\mathcal{B}, \mathcal{Z}) = \tilde{G}(\mathcal{B}^*, \mathcal{Z})$$

yields the maximum-support-plane partition $\mathcal{B}^* = (B_1^*, ..., B_m^*)$ with classes

$$B_i^* := \{ \ x \in \mathbb{R}^p \mid t(x; z_i) \ = \ \max_{j=1,...,m} t(x; z_j) \ \} \qquad i = 1, ..., m$$

comprizing all $x \in \mathbb{R}^p$ where the <u>i-th tangent hyperplane $t(x; z_i)$ is maximum</u>.

An application:

P_1, P_2 two probability distributions for $X \in \mathbb{R}^p$ with densities $f_1(x), f_2(x)$,

likelihood ratio $\lambda(x) := f_2(x)/f_1(x)$

Discretization of X:

Look for a partition $\mathcal{B} = (B_1, ..., B_k)$ of $\mathbb{R}^p$ such that the discrete distributions

$$P_1(X \in B_1), ..., P_1(X \in B_k) \quad \text{and} \quad P_2(X \in B_1), ..., P_2(X \in B_k)$$

are as different as possible in the sense:

χ^2 non-centrality parameter criterion:

$$\boxed{\begin{aligned} G(\mathcal{B}) \; &:= \; \sum_{i=1}^{k} \frac{(P_1(B_i) - P_2(B_i))^2}{P_1(B_i)} = \sum_{i=1}^{k} P_1(B_i) \left(1 - \frac{P_2(B_i)}{P_1(B_i)}\right)^2 \\ &= \; \sum_{i=1}^{k} P_1(B_i) \cdot (1 - E[\lambda(X)|X \in B_i])^2 \; \rightarrow \; \max_{\mathcal{B}} \end{aligned}}$$

Csziszar's divergence criterion with a convex ϕ:

$$\boxed{G(\mathcal{B}) \; := \; \sum_{i=1}^{k} P_1(B_i) \cdot \phi(E[\lambda(X)|X \in B_i]) \; \rightarrow \; \max_{\mathcal{B}}}$$

6. L'avenir

Congratulations to Edwin !

Best wishes for your future work!