A history of the k-means algorithm

Hans-Hermann Bock, RWTH Aachen, Allemagne

1. Clustering with SSQ and the basic k-means algorithm
   1.1 Discrete case
   1.2 Continuous version

2. SSQ clustering for stratified survey sampling
   Dalenius (1950/51)

3. Historical k-means approaches
   Steinhaus (1956), Lloyd (1957), Forgy/Jancey (1965/66)
   MacQueen’s sequential k-means algorithm (1965/67)

4. Generalized k-means algorithms
   Maranzana’s transportation problem (1963)
   Generalized versions, e.g., by Diday et al. (1973 - ...)

5. Convexity-based criteria and k-tangent algorithm

6. Final remarks

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: \( \mathcal{O} = \{1, \ldots, n\} \) set of \( n \) objects
\[ x_1, \ldots, x_n \in \mathbb{R}^p \] \( n \) data vectors

Problem: Determine a partition \( \mathcal{C} = (C_1, \ldots, C_k) \) of \( \mathcal{O} \)
with \( k \) classes \( C_i \subset \mathcal{O}, i = 1, \ldots, k \)
characterized by class prototypes: \( \mathcal{Z} = (z_1, \ldots, z_k) \)

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

\[
g(\mathcal{C}) := \sum_{i=1}^{k} \sum_{\ell \in C_i} ||x_\ell - \overline{x}_{C_i}||^2 \quad \rightarrow \quad \min \mathcal{C} \]

with class centroids (class means) \( z^*_1 = \overline{x}_{C_1}, \ldots, z^*_k = \overline{x}_{C_k} \).

Two-parameter form:

\[
g(\mathcal{C}, \mathcal{Z}) := \sum_{i=1}^{k} \sum_{\ell \in C_i} ||x_\ell - z_i||^2 \quad \rightarrow \quad \min \mathcal{C,\mathcal{Z}} \]

Remark: \( g(\mathcal{C}) \equiv g(\mathcal{C}, \mathcal{Z}^*) \)
The well-known $k$-means algorithm

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

$t = 0$:

Start from an arbitrary initial partition $C^{(0)} = (C^{(0)}_1, \ldots, C^{(0)}_k)$ of $O$

$t \rightarrow t + 1$:

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

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

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

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

Stopping:

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

Problem A:

$$g(C^{(t)}, Z) \rightarrow \min_Z$$

Problem B:

$$g(C, Z^{(t)}) \rightarrow \min_C$$
\[ g(C, Z) := \sum_{i=1}^{k} \sum_{\ell \in C_i} ||x_\ell - z_i||^2 \rightarrow \min_{C,Z} \]

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

**Theorem:**

The \(k\)-means algorithm

\[
\begin{align*}
Z(t) &:= Z(C(t)) \\
C(t+1) &:= C(Z(t)) & t = 0, 1, 2, \ldots
\end{align*}
\]

produces \(m\)-partitions \(C(t)\) and prototype systems \(Z(t)\) with steadily decreasing criterion values:

\[
g(C(t)) \equiv g(C(t), Z(t)) \geq g(C(t+1), Z(t)) \geq g(C(t+1), Z(t+1)) \equiv g(C(t+1))
\]
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 \( \mathcal{B} = (B_1, \ldots, B_k) \) of \( \mathbb{R}^p \) with \( k \) Borel sets (classes) \( B_i \subset \mathbb{R}^p \), \( i = 1, \ldots, k \) characterized by class prototypes: \( \mathcal{Z} = (z_1, \ldots, z_k) \)

- Continuous version of SSQ criterion:

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

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

- Two-parameter form:

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

\( \implies \) 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 $f(x)$

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

Problem: Estimate unknown expected income $\mu$ by using $n$ samples (persons)

- **Strategy I: Simple random sampling**

  Sample $n$ persons, observed income values $x_1, \ldots, x_n$

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

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

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

• **Strategy II: Stratified sampling**

Partitioning \( \mathcal{I} \) into \( k \) classes (strata): \( B_1, \ldots, B_k \)

Class probabilities: \( p_1, \ldots, p_k \)

Sampling from stratum \( B_i \): \( Y_i \sim X \mid X \in B_i \)

\[ \mu_i := E[Y_i] = E[X \mid X \in B_i] \]

\[ \sigma_i^2 := Var(Y_i) = Var(X \mid X \in B_i) \]

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

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

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

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

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

• **Strategy III: Proportional stratified sampling**

Use sample sizes proportional to class frequencies: \( n_i = n \cdot p_i \)
• **Strategy III:** Proportional stratified sampling
  
  Use sample sizes proportional to class frequencies: $n_i = n \cdot p_i$

  $\implies$ Resulting variance:

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

  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):

$\mathcal{X} \subset \mathbb{R}^p$ a solid (mechanics; similarly: anthropology, industry)

with mass distribution density $f(x)$

Problem:

Dissecting $\mathcal{X}$ into $k$ parts $B_1, \ldots, B_k$

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) $z_1, \ldots, z_k \in \mathbb{R}^p$
  - we determine the (index of the) code vector that is closest to $x$:
    $$ i(x) = \arg\min_{j=1,\ldots,k} \{ \| x - z_j \|^2 \} $$
  - transmit only the index $i(x)$
  - and decode the message $x$ by the code vector $\hat{x} := z_{i(x)}$.

**Expected transmission (approximation) error:**

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

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

**Lloyd’s Method I:** Continuous version of $k$-means (in $\mathbb{R}^1$)
• Forgy (1965), Jancey (1966):
Taxonomy of genus Phyllota Benth. (Papillionaceae)

\[ x_1, \ldots, x_n \] are feature vectors characterizing \( n \) butterflies

Forgy’s lecture proposes the \textit{discrete} \( k \)-\textit{means algorithm}

\( \text{(implying the SSQ clustering criterion only implicitly!)} \)

\textbf{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:

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

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: $\mathcal{O} = \{1, \ldots, n\}$

- 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 $C_1, \ldots, C_k$

- 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:

\[
g(C, \mathcal{Z}) := \sum_{i=1}^{k} \sum_{\ell \in C_i} d(\ell, z_i) \rightarrow \min_{C, \mathcal{Z}} \quad \text{with } z_i \in C_i \text{ for } i = 1, \ldots, m
\]

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

\[
Q(C_i, z) := \sum_{\ell \in C_i} d(\ell, z) \rightarrow \min_{z \in C_i}
\]

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

\[
g(C, Z) := \sum_{i=1}^{m} \sum_{k \in C_i} d(k, z_i) \rightarrow \min_{C,Z}
\]

- Kernel clustering: prototype \( z_i \) = a subset of \( C_i \) with \( |z_i| = 4 \), say

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

- Adaptive distance clustering: \( d(x_\ell, z_i) = \|x_\ell - z_i\|^2_{Q_i} \) 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 \( z_1, \ldots, z_k \) on the same low-dim. hyperplane
OPTIMISATION EN CLASSIFICATION AUTOMATIQUE

TOME 1

E. DIDAY ET COLLABORATEURS

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE
DOMAINE DE VOLUCEAU - ROCQUECOURT - B. P. 105 - 78150 LE CHESNAY

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

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

- Exists a $k$-partition $C = (C_1, \ldots, C_k)$ of $O = \{1, \ldots, n\}$
- Exist $k$ class-specific parameter vectors $z_1, \ldots, z_k$

\[
X_\ell \sim f(\bullet; z_i) \quad \text{for all } \ell \in C_i
\]

Maximum likelihood estimation of $C$ and $Z = (z_1, \ldots, z_k)$:

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

A two-parameter clustering criterion!

\[
\Rightarrow \text{A generalized } k\text{-means algorithm} \quad \text{alternating}
\]
- 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

\[
g(C) := \sum_{i=1}^{k} \sum_{\ell \in C_i} ||x_\ell - \bar{x}_{C_i}||^2 = \sum_{\ell=1}^{n} ||x_\ell||^2 - \sum_{i=1}^{k} |C_i| \cdot ||\bar{x}_{C_i}||^2 \to \min_C
\]

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

\[
G_n(C) := \frac{1}{n} \sum_{i=1}^{k} |C_i| \cdot ||\bar{x}_{C_i}||^2 = \sum_{i=1}^{k} \frac{|C_i|}{n} \cdot \phi(\bar{x}_{C_i}) \to \max_C
\]

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

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

- 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' $z \in \mathbb{R}^p$, the convex function $\phi(x)$ has a support (tangent) hyperplane

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

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

$$\phi(x) \geq t(x; z) \quad \text{for all } x \in \mathbb{R}^p$$
$$\phi(z) = t(z; z) \quad \text{for } x = z.$$
Original clustering problem:

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

Equivalent dual two-parameter problem:

Looking for \( k \) support points \( z_1, \ldots, z_m \in \mathbb{R}^p \) and corresponding tangents (hyperplanes)

\[ t(x; z_i) := \phi(z_i) + a_i^r(x - z_i) \]

such that

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

"Minimum volume problem"
Original clustering problem:

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

Equivalent dual two-parameter problem:

Looking for \( k \) support points \( z_1, \ldots, z_m \in \mathbb{R}^p \) and corresponding tangents (hyperplanes)

\[ t(x; z_i) := \phi(z_i) + a^{tr}_i(x - z_i) \]

such that

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

"Minimum volume problem"
Alternating minimization: \( k \)-tangent clustering algorithm

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

\[
\min_{\mathcal{Z}} \tilde{G}(\mathcal{B}, \mathcal{Z}) = \tilde{G}(\mathcal{B}, \mathcal{Z}^*)
\]

yields the system \( \mathcal{Z}^* = (\tilde{z}_1^*, ..., \tilde{z}_m^*) \) of class centroids \( \tilde{z}_i^* := E[X | X \in B_i] \).

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

\[
\min_{\mathcal{B}} \tilde{G}(\mathcal{B}, \mathcal{Z}) = \tilde{G}(\mathcal{B}^*, \mathcal{Z})
\]

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

\[
B_i^* := \{ x \in \mathbb{R}^p | t(x; z_i) = \max_{j=1, ..., m} t(x; z_j) \} \quad i = 1, ..., m
\]

comprizing all \( x \in \mathbb{R}^p \) where the \( i \)-th tangent hyperplane \( t(x; z_i) \) is maximum.
An application:
$P_1, P_2$ two probability distributions for $X \in \mathbb{IR}^p$ 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 $\mathbb{IR}^p$ such that the discrete distributions
$P_1(X \in B_1), ..., P_1(X \in B_k)$ and $P_2(X \in B_1), ..., P_2(X \in B_k)$
are as different as possible in the sense:

$\chi^2$ non-centrality parameter criterion:

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

Csziszar’s divergence criterion with a convex $\phi$:

$$G(\mathcal{B}) := \sum_{i=1}^{k} P_1(B_i) \cdot \phi(E[\lambda(X)|X \in B_i]) \to \max_{\mathcal{B}}$$
6. L’avenir

Congratulations to Edwin!
Best wishes for your future work!