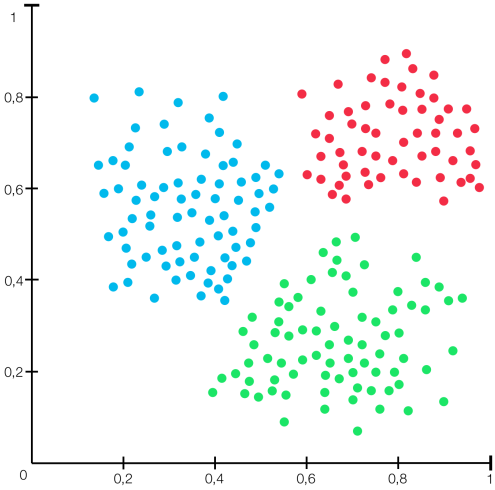
Statistical Data Analysis

Dr. Jana de Wiljes

04.01.2022

Universität Potsdam

Clustering



K-means clustering

Input:

- Number of Clusters K
- Set of points $\{x_1, \dots, x_M\}$ in vector space that need to be classified

Output:

- Sets $\mathcal{M}_k$ of the clusters

1. Initialize the centre of the cluster $\theta_1, \dots, \theta_K \in \mathbb{R}^n$ randomly
2. Repeat till a stopping criterion is fulfilled {
   for all $k = 1 : K$
   $\quad \mathcal{M}_k := \{\ \}$
   for all $m = 1 : M$
   $\quad j = \arg\min_h ||\theta_h - x_m||_2^2$
   $\quad \mathcal{M}_j = \mathcal{M}_j \cup \{x_m\}$
   for all $k = 1 : K$
   $\quad \theta_k = \frac{1}{|\mathcal{M}_k|} \sum_{x_m \in \mathcal{M}_k} x_m$
3. **return** $\theta_1, \dots, \theta_K$

Initialisation

- Random Partition Method
- Forgy Initialization
- kmeans++
  1. choose θ_1 uniformly at random from set of points
  2. Choose new center θ_i with probability

     $$\frac{D(x_m)^2}{\sum_{x_l} D(x_l)^2} \tag{1}$$

     where $D(x_m)$ denotes the shortest distance from data point x_m to the closest center we have already chosen
  3. Repeat Step 2 until we have all K centers

Disadvantages

- true number of clusters K unknow (requires tuning)
- K-means algorithm dependents on the chosen initial values
- Clustering data of varying sizes and density
- Centroids can be dragged by outliers

Using the Triangle Inequality to Accelerate k-Means

Using the Triangle Inequality to Accelerate k-Means

Using the Triangle Inequality to Accelerate k-Means

Using the Triangle Inequality to Accelerate k-Means

Algorithm:

1. Initialize the centre of the cluster $\theta_1, \ldots, \theta_K \in \mathbb{R}^n$ randomly
2. Set lower bounds to $l(x_m, \theta_i) = 0$ for all θ_i and x_m
3. Assign each x_m to its closest initial center $\theta(x_m) = \arg\min_h ||\theta_h - x_m||_2^2$ (avoid redundant calculations using Lemma 1)
4. Each time $||\theta_h - x_m||_2^2$ is computed, set $l(x_m, \theta_h) = ||\theta_h - x_m||_2^2$
5. Assign upper bounds $u(x_m) = min_i||\theta_i - x_m||_2^2$
6. Repeat till a stopping criterion is fulfilled $\{$

    6.1 **for all** θ_i and θ_j, compute $||\theta_i - \theta_j||_2^2$. **For all** centers θ_i, compute $s(\theta_i) = \frac{1}{2}\min_j ||\theta_i - \theta_j||_2^2$

    6.2 Identify all points x_m such that $u(x_m) \leq s(\theta(x_m))$.

    6.3 **for all** centers θ_i **for all** remaining points x_m check

    - $\theta_i \neq \theta(x_m)$ and
    - $u(x_m) > l(x_m, \theta_i)$ and
    - $u(x_m) > \frac{1}{2}||\theta(x_m) - \theta_i||_2^2$

    If conditions $r(x_{m)=\text{true}}$ are true compute $\|x_m - \theta(x_m)\|$ and assign $r(x_m) = \text{false}$. Otherwise $\|x_m - \theta(x_m)\|_2^2 = u(x_m)$.

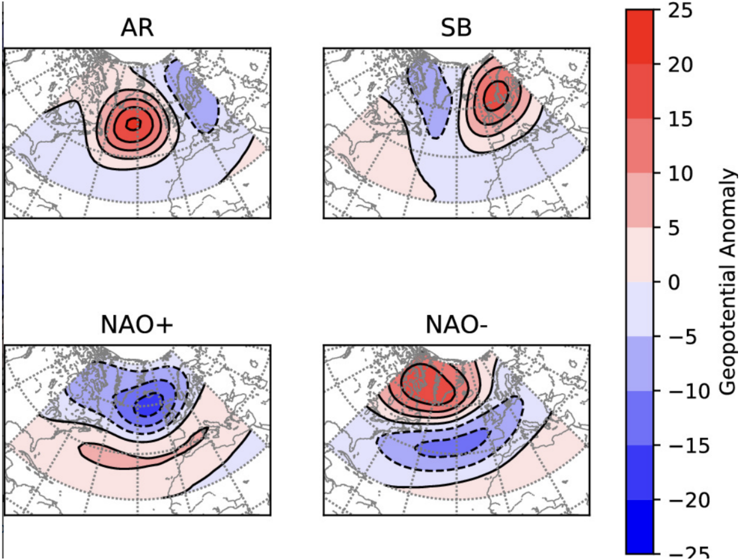
    6.4 if $\|x_m - \theta(x_m)\|_2^2 > l(x_m, \theta_i)$ or $\|x_m - \theta(x_m)\} > \frac{1}{2}||\theta(x_m) - \theta_i||_2^2$ then

    - compute $\|(x_m - \theta_i)\|_2^2$
    - if $\|(x_m - \theta_i)\|_2^2 < \|(x_m - \theta(x_m))\|_2^2$ then assign $\theta(x_m) = \theta_i$

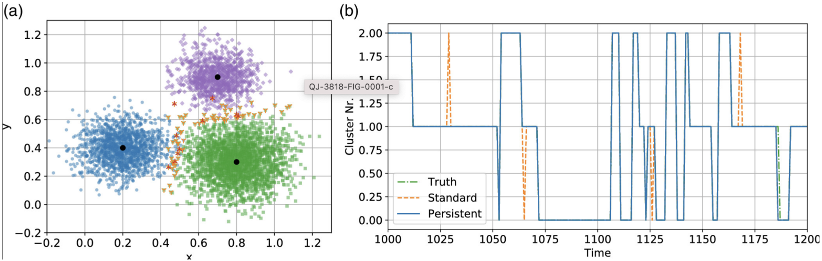
7. **for all** centers θ_i, let $m(\theta_i)$ be the mean of the points assigned to θ_i
8. **for all** points x_m and **for all** centers θ_i assign $l(x_m, \theta_i) = \max\{l(x_m, \theta_i) - \|\theta_i - m(\theta_i)\|_2^2, 0\}$
9. **for all** points x_m, assign $u(x_m) = u(x_m) + \|m(\theta(x_m)) - \theta(x_m)\|$ and $r(x_m) = \text{true}$
10. replace each center θ_i with $m(\theta_i)$
11. **return** $\theta_1, \ldots, \theta_K$

Example: pattern recognition for atmospheric circulation regimes

Regime

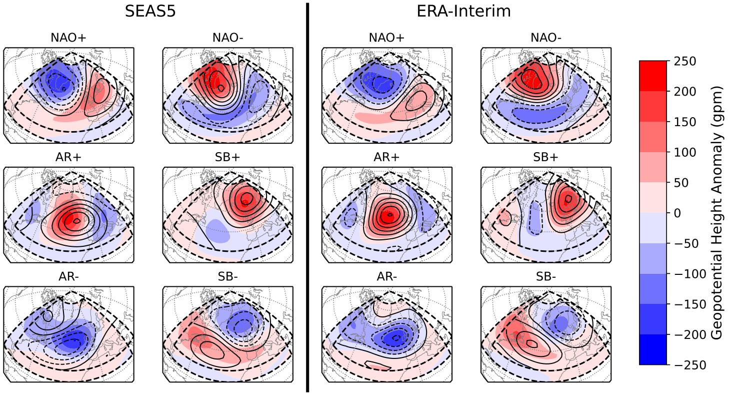


Time persistency constraint

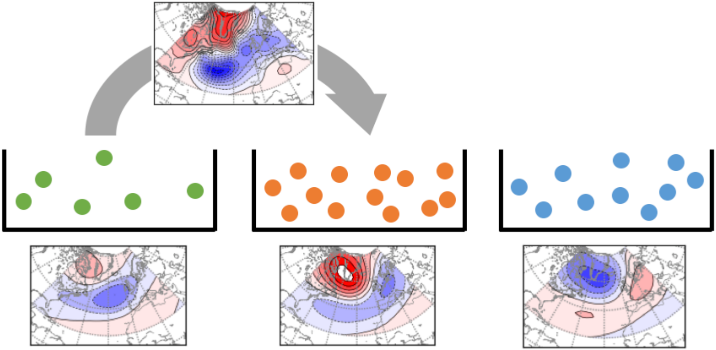


$$\sum_{t=1}^{T-1} |\gamma_k(t+1) - \gamma_k(t)| \leq N_C \quad \forall k$$

k-means clustering for different domains



k-means clustering for different domains



Optimisation problem

$$\mathbf{L}(\Theta, \Gamma) = \sum_{t=0}^{T} \sum_{n=1}^{N} \sum_{i=1}^{k} \gamma_i(t, n) \|x_{t,n} - \theta_i\|^2$$

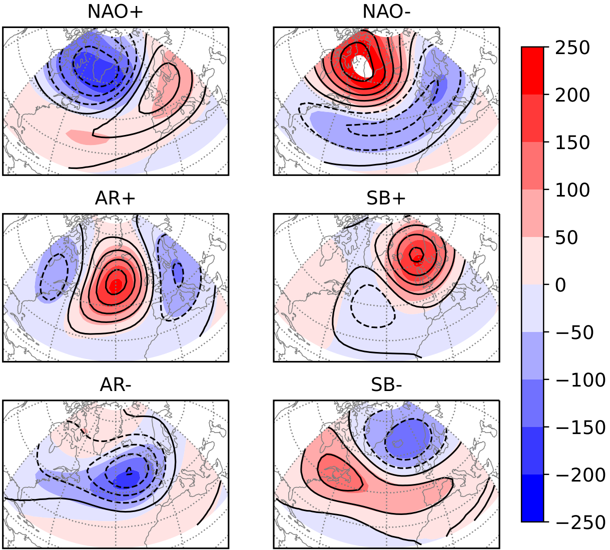
with

$$\sum_{i=1}^{k} \gamma_i(t, n) = 1, \qquad \forall t \in [0, T], \quad \forall n \in [1, N].$$

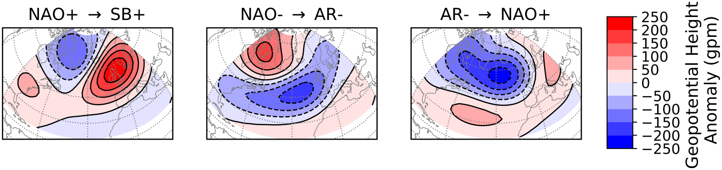
and

$$\sum_{i=1}^{k} \sum_{n_1, n_2} |\gamma_i(t, n_1) - \gamma_i(t, n_2)| \leq \phi \cdot C_{\text{eq}}, \qquad \forall t \in [0, T],$$

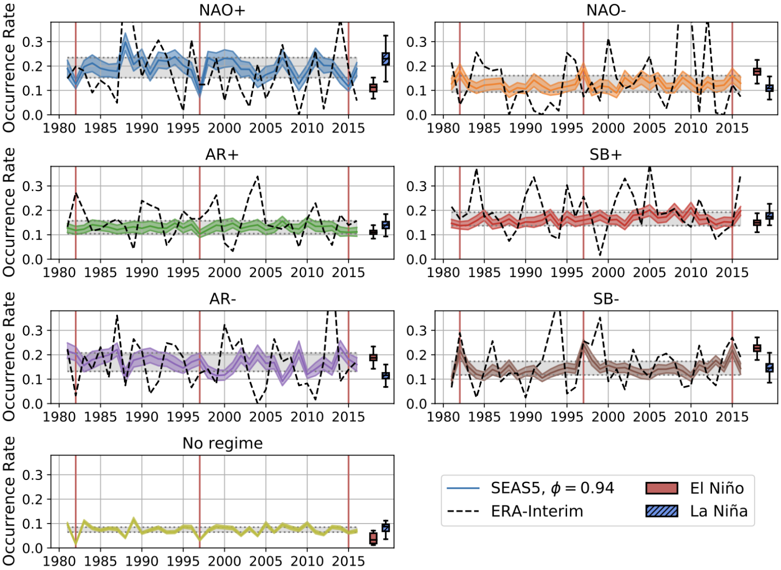
Ensemble persistency constraint



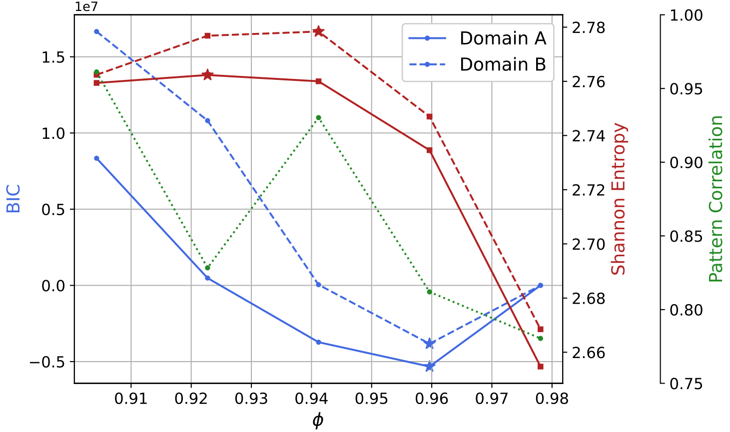
Ensemble persistency constraint



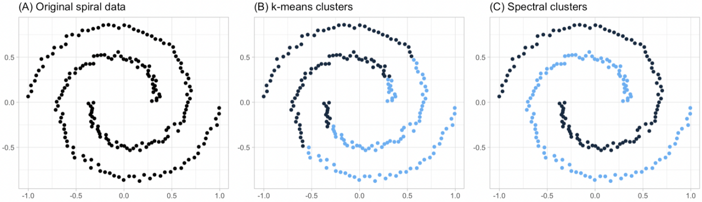
Occurrence rates



Optimal ϕ


1e7
Domain A
Domain B
BIC
Shannon Entropy
Pattern Correlation
ϕ
1.5
1.0
0.5
0.0
−0.5
0.91
0.92
0.93
0.94
0.95
0.96
0.97
0.98
2.78
2.76
2.74
2.72
2.70
2.68
2.66
1.00
0.95
0.90
0.85
0.80
0.75

K-Means vs Spectral Clustering



Eigenvalues and Eigenvectors

Definition

Let V be a K-Vector space, $f\colon V \to V$ an Endomorphismus, $\lambda \in K$. The scalar λ is called **Eigenvalue** of f, if there is a vector $v \in V, v \neq 0$, so that

$$f(v) = \lambda \cdot v.$$

The vector v is called **Eigenvector** of f an Eigenvalue λ.

Note: An Eigenvalue λ can be $0 \in K$, but an Eigenvector is always $\neq 0$.

Theorem

Theorem

Let V be a K-vector space, $n = \dim V < \infty$ and $f \colon V \to V$ an Endomorphismus. The following two are equivalent:

1. *V has a basis of Eigenvectors of f.*
2. *There is a Basis $\mathcal{B}$ of V, so that*

$$M_{\mathcal{B}}^{\mathcal{B}}(f) = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix} \text{ with } \lambda_i \in K.$$

Definition

Let $A \in K^{n\times n}$ and $\lambda \in K$ abitrary. Then

$$\mathrm{Eig}(A, \lambda) := \{v \in K^n \mid Av = \lambda v\}$$

is called the **Eigenspace** of A with respect to λ.

$$\chi_A(t) := \det(A - tE) \in K[t]$$

is called the **charakteristisches Polynom** of A.

Remark: For a matrix $A \in K^{n\times n}$ the following holds:

$$\lambda \in K \text{ is an Eigenvalue of } A \Leftrightarrow \mathrm{Eig}(A, \lambda) \neq 0.$$

Theorem

Let $A \in K^{n \times n}$ and $\lambda \in K$. Then:

$$\lambda \text{ is an Eigenvalue of } A \Leftrightarrow \lambda \text{ is a root of } \chi_A(t).$$

Multiplicity

Definition

Let $P(t) \in K[t]$ be a Polynom. $P(t)$ can be decomposed over K in **Linear factors** if and only if there are $\lambda_1, \ldots, \lambda_n \in K, c \in K$, so that

$$P(t) = c \cdot (t - \lambda_1) \cdots (t - \lambda_n) = c \cdot \prod_{j=1}^{r} (t - \lambda_j')^{m_j},$$

where $m_j \in \mathbb{N}$ and $\lambda_1', \ldots, \lambda_r' \in \{\lambda_1, \ldots, \lambda_n\}$ are pairwise different. m_j is called the **Multiplicity** of the root λ_j'. It holds that

$$\sum_{j=1}^{r} m_j = n.$$

Example

Example

Example