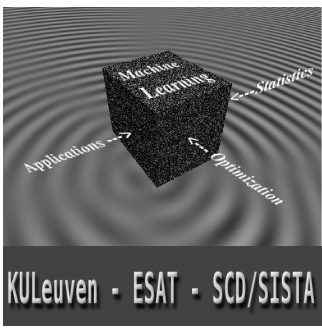


Clustering shrinkage, L_0 and Staircases

K. PELCKMANS, J.A.K. SUYKENS, B. DE MOOR

NIPS workshop on theoretical foundations of clustering
December 2005

KULeuven - Department of Electrical Engineering - SCD/SISTA
Kasteelpark Arenberg 10, 3001 Heverlee (Leuven), Belgium
Kristiaan.Pelckmans@esat.kuleuven.ac.be



Optimization view to Clustering

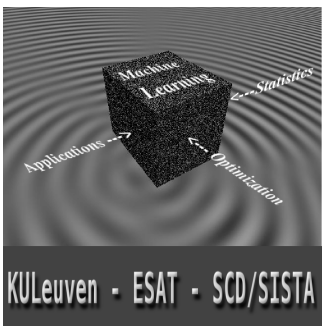
Empirical Convex Clustering Shrinkage:

- Dataset $\{x_i\}_{i=1}^N \subset \mathbb{R}^D$
- N centroids: $\{M_i\}_{i=1}^N \subset \mathbb{R}^D$

Empirical CCS



Theoretical CCS



Optimization view to Clustering

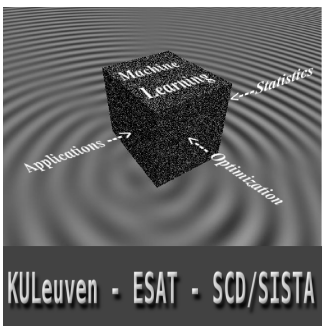
Empirical Convex Clustering Shrinkage:

- Dataset $\{x_i\}_{i=1}^N \subset \mathbb{R}^D$
- N centroids: $\{M_i\}_{i=1}^N \subset \mathbb{R}^D$
- Given $\gamma \geq 0$
- Distance measure $\|\cdot\|$
- Convex complexity measure $\ell : \mathbb{R}^D \rightarrow \mathbb{R}^+$

Empirical CCS



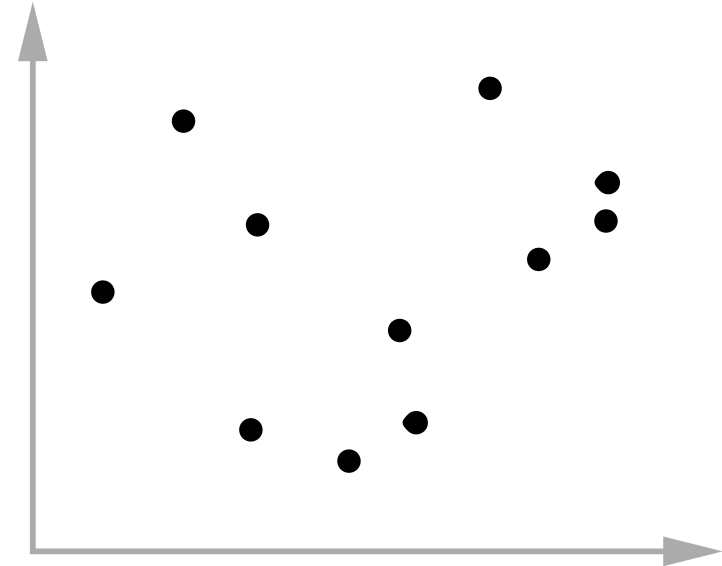
Theoretical CCS



Optimization view to Clustering

Empirical Convex Clustering Shrinkage:

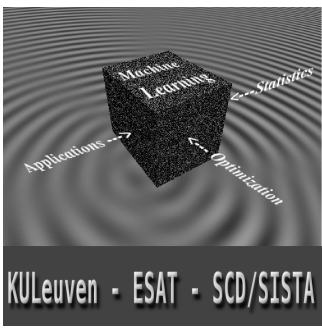
- Dataset $\{x_i\}_{i=1}^N \subset \mathbb{R}^D$
- N centroids: $\{M_i\}_{i=1}^N \subset \mathbb{R}^D$
- Given $\gamma \geq 0$
- Distance measure $\|\cdot\|$
- Convex complexity measure $\ell : \mathbb{R}^D \rightarrow \mathbb{R}^+$



Empirical CCS



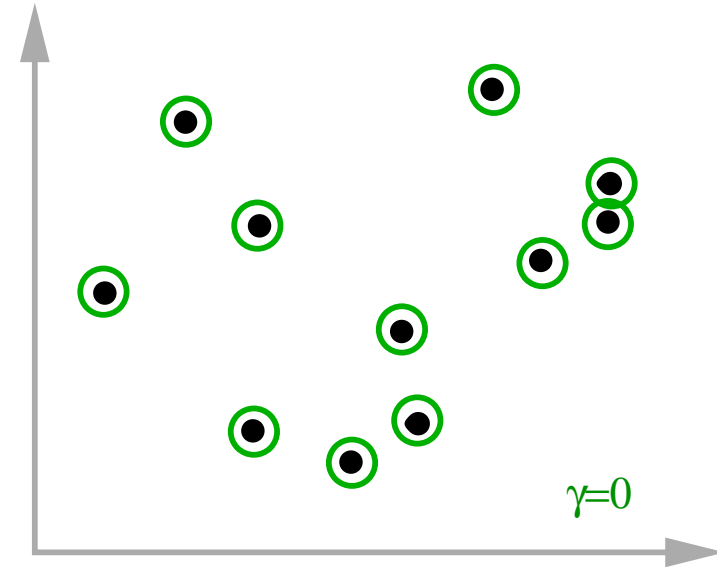
Theoretical CCS



Optimization view to Clustering

Empirical Convex Clustering Shrinkage:

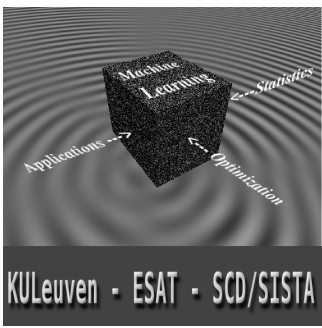
- Dataset $\{x_i\}_{i=1}^N \subset \mathbb{R}^D$
- N centroids: $\{M_i\}_{i=1}^N \subset \mathbb{R}^D$
- Given $\gamma \geq 0$
- Distance measure $\|\cdot\|$
- Convex complexity measure $\ell : \mathbb{R}^D \rightarrow \mathbb{R}^+$



Empirical CCS



Theoretical CCS



Optimization view to Clustering

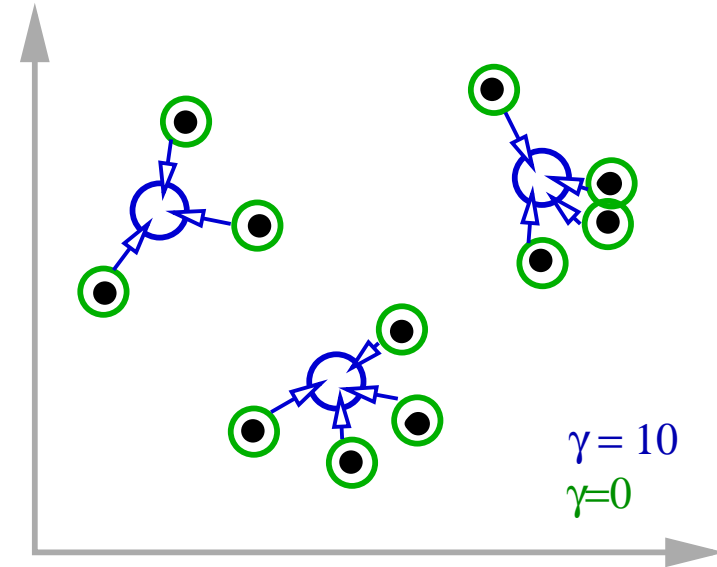
Empirical Convex Clustering Shrinkage:

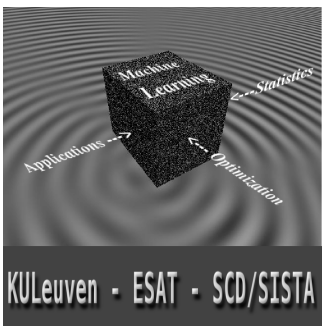
- Dataset $\{x_i\}_{i=1}^N \subset \mathbb{R}^D$
- N centroids: $\{M_i\}_{i=1}^N \subset \mathbb{R}^D$
- Given $\gamma \geq 0$
- Distance measure $\|\cdot\|$
- Convex complexity measure $\ell : \mathbb{R}^D \rightarrow \mathbb{R}^+$

Empirical CCS



Theoretical CCS





Optimization view to Clustering

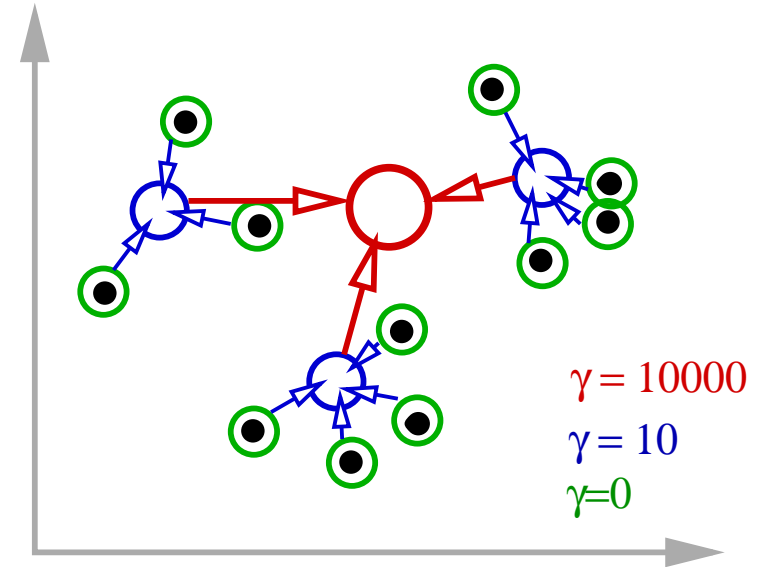
Empirical Convex Clustering Shrinkage:

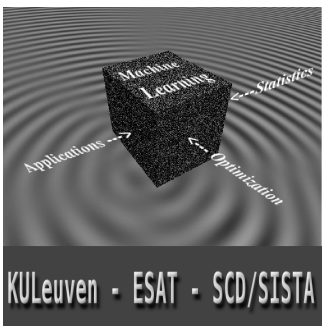
- Dataset $\{x_i\}_{i=1}^N \subset \mathbb{R}^D$
- N centroids: $\{M_i\}_{i=1}^N \subset \mathbb{R}^D$
- Given $\gamma \geq 0$
- Distance measure $\|\cdot\|$
- Convex complexity measure $\ell : \mathbb{R}^D \rightarrow \mathbb{R}^+$

Empirical CCS



Theoretical CCS





Optimization view to Clustering

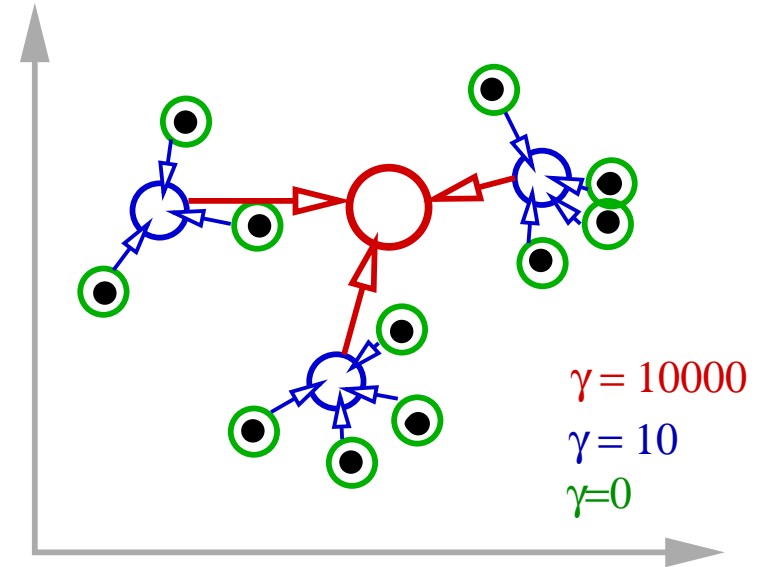
Empirical Convex Clustering Shrinkage:

- Dataset $\{x_i\}_{i=1}^N \subset \mathbb{R}^D$
- N centroids: $\{M_i\}_{i=1}^N \subset \mathbb{R}^D$
- Given $\gamma \geq 0$
- Distance measure $\|\cdot\|$
- Convex complexity measure $\ell : \mathbb{R}^D \rightarrow \mathbb{R}^+$

Empirical CCS

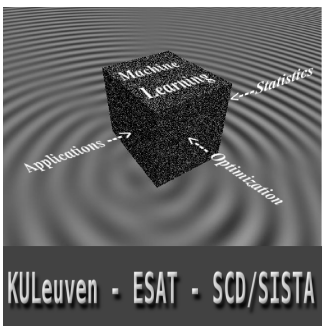


Theoretical CCS



Convex Programming Problem:

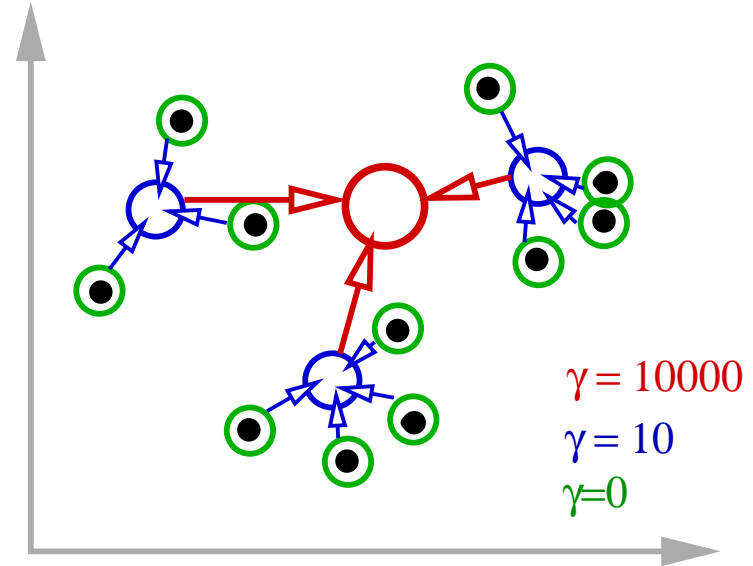
$$\min_{M_i} \mathcal{J}_\gamma(M_i) = \frac{1}{2} \sum_{i=1}^N \|x_i - M_i\|_p$$



Optimization view to Clustering

Empirical Convex Clustering Shrinkage:

- Dataset $\{x_i\}_{i=1}^N \subset \mathbb{R}^D$
- N centroids: $\{M_i\}_{i=1}^N \subset \mathbb{R}^D$
- Given $\gamma \geq 0$
- Distance measure $\|\cdot\|$
- Convex complexity measure $\ell : \mathbb{R}^D \rightarrow \mathbb{R}^+$



Empirical CCS

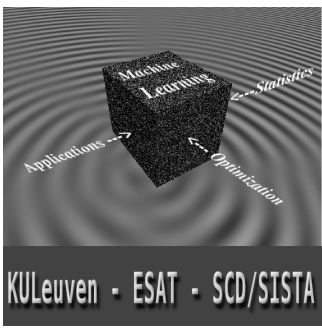


Theoretical CCS

Convex Programming Problem:

$$\min_{M_i} \mathcal{J}_\gamma(M_i) = \frac{1}{2} \sum_{i=1}^N \|x_i - M_i\|_p + \gamma \sum_{i < j} \ell(M_i - M_j)$$

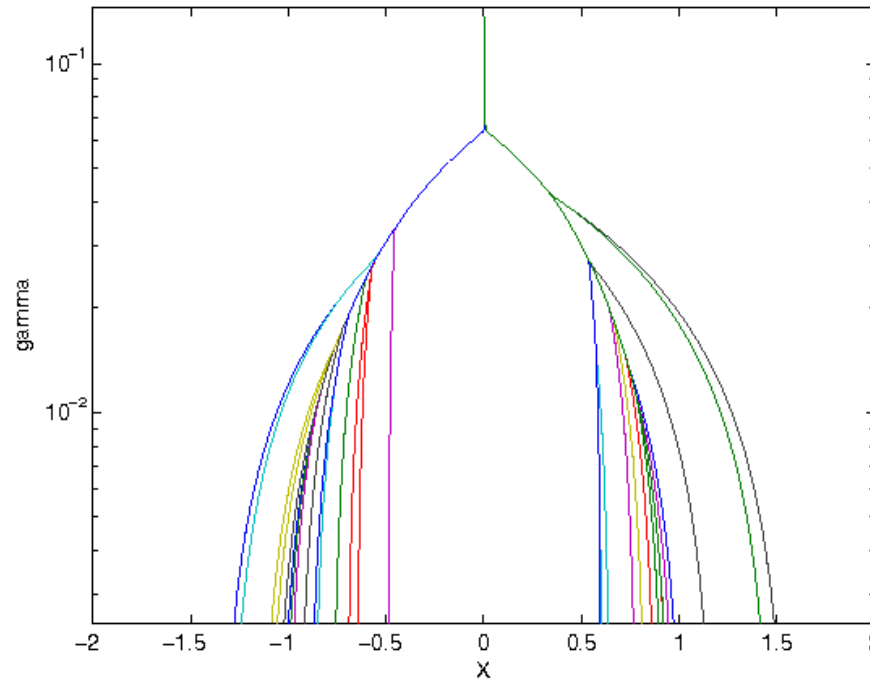
→ Pelckmans *et al.*, Convex Clustering Shrinkage, PASCAL workshop 2005



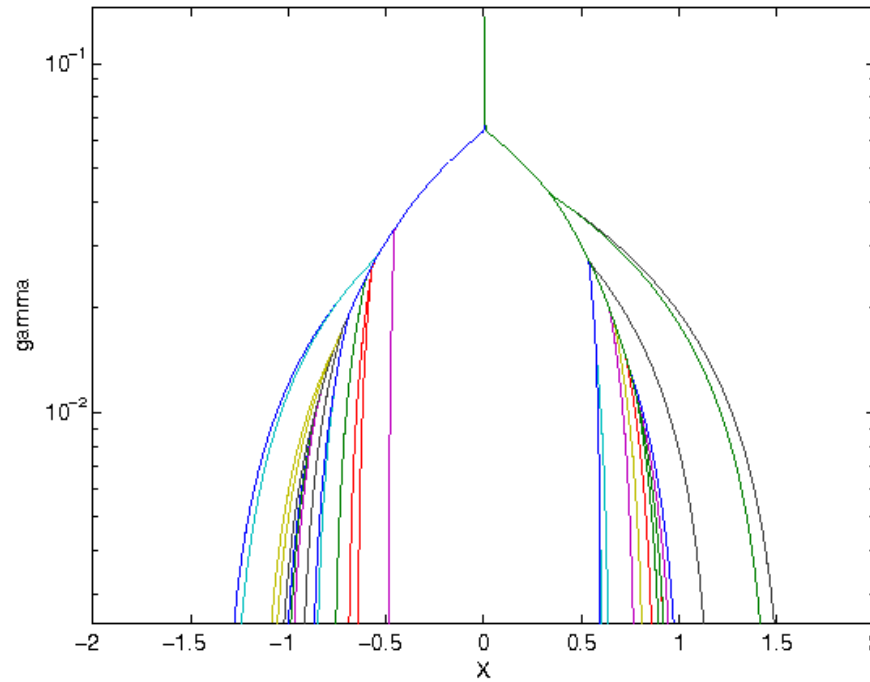
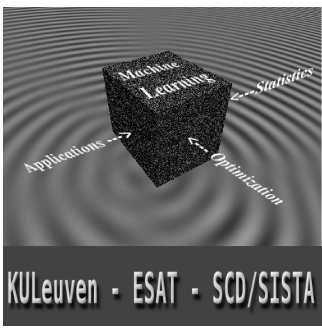
Empirical CCS



Theoretical CCS



- $\gamma = 0: M_i = X_i$
- $\gamma \rightarrow +\infty: M_1 = \dots = M_N = \bar{X}$
- $\ell = |\cdot|_1$
- Ranging γ , increasing number of sparse differences

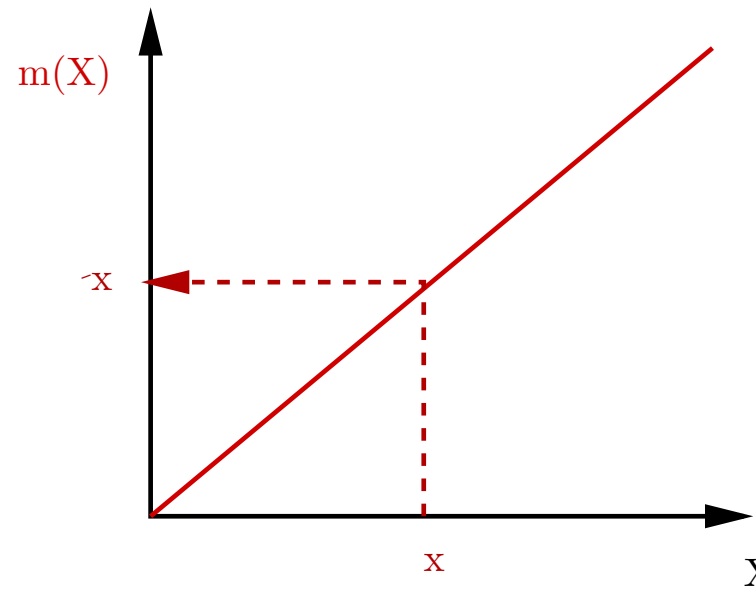


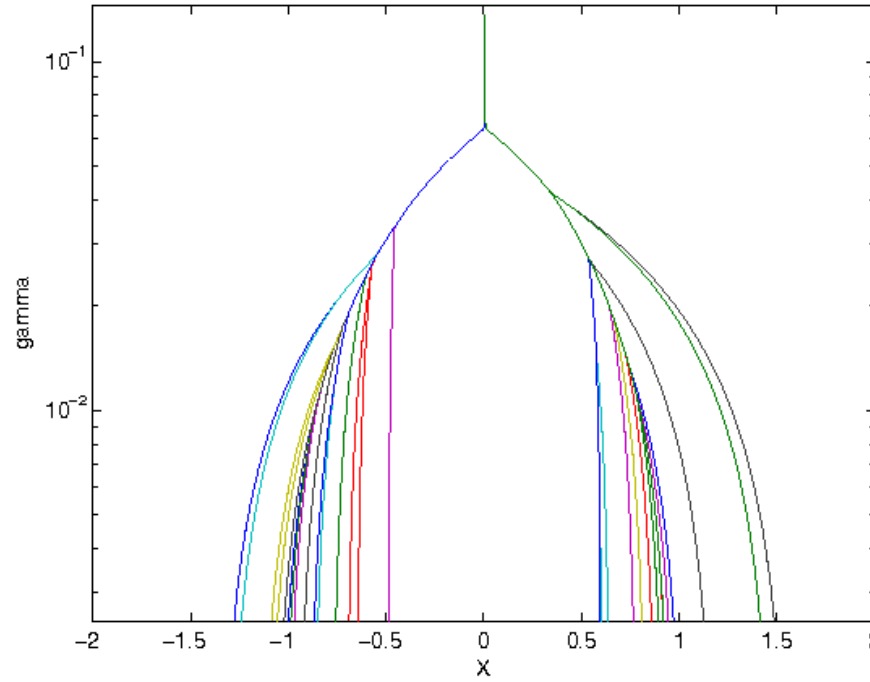
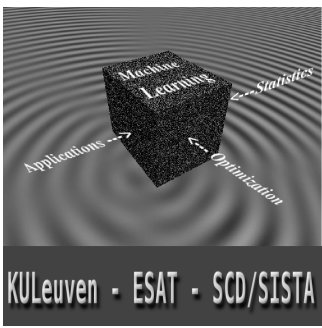
- $\gamma = 0: M_i = X_i$
- $\gamma \rightarrow +\infty: M_1 = \dots = M_N = \bar{X}$
- $\ell = |\cdot|_1$
- Ranging γ , increasing number of sparse differences

Empirical CCS



Theoretical CCS



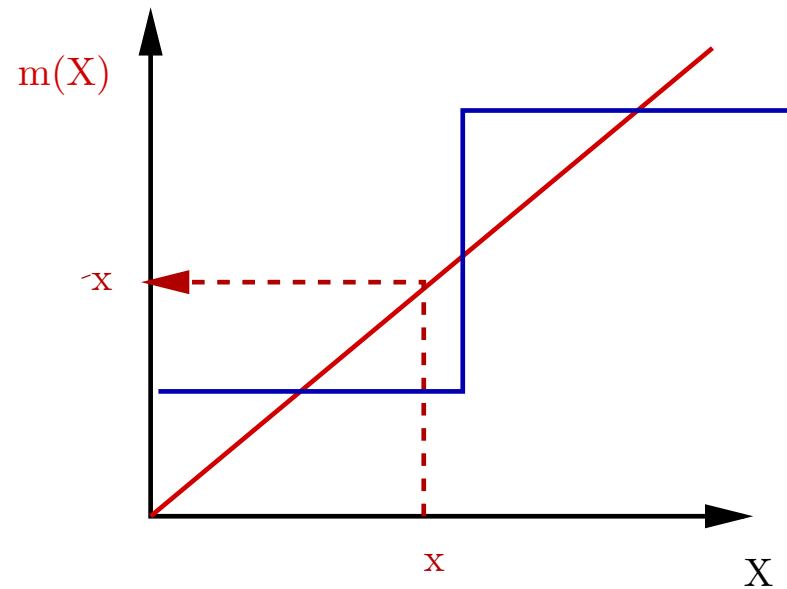


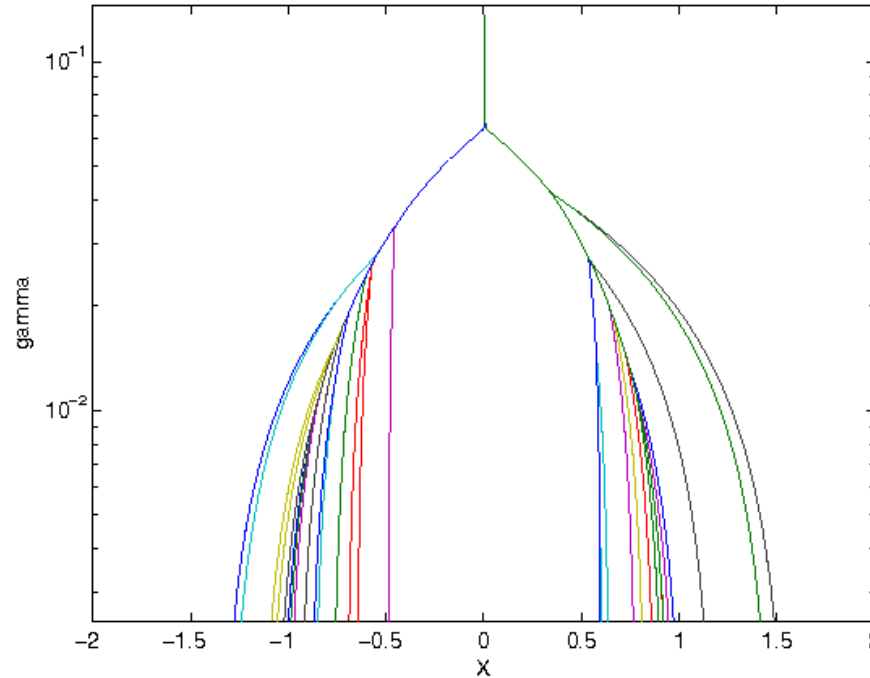
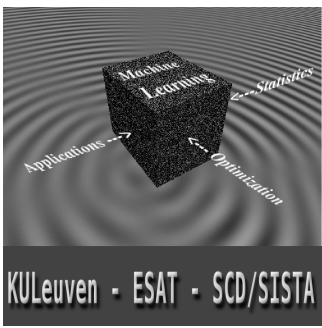
- $\gamma = 0: M_i = X_i$
- $\gamma \rightarrow +\infty: M_1 = \dots = M_N = \bar{X}$
- $\ell = |\cdot|_1$
- Ranging γ , increasing number of sparse differences

Empirical CCS



Theoretical CCS



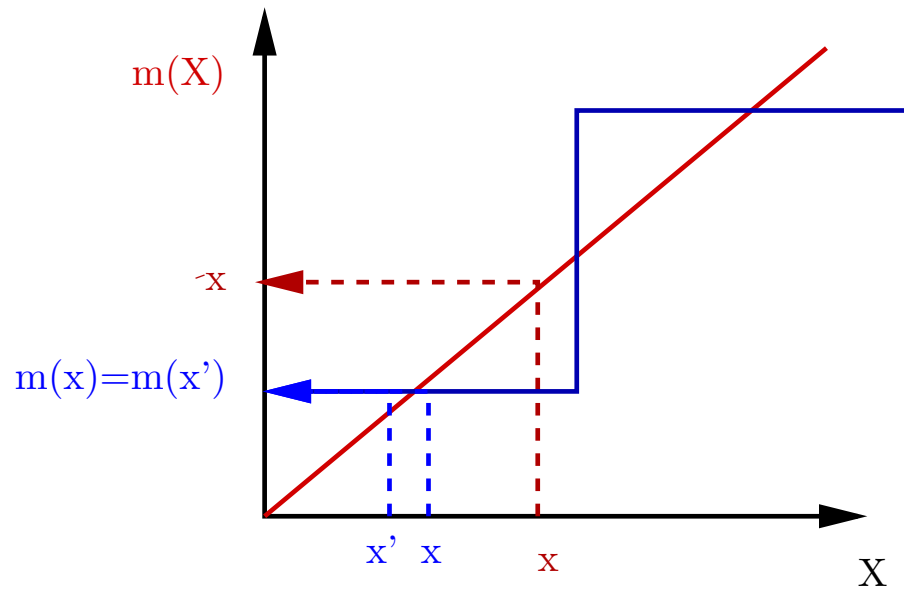


- $\gamma = 0: M_i = X_i$
- $\gamma \rightarrow +\infty: M_1 = \dots = M_N = \bar{X}$
- $\ell = |\cdot|_1$
- Ranging γ , increasing number of sparse differences

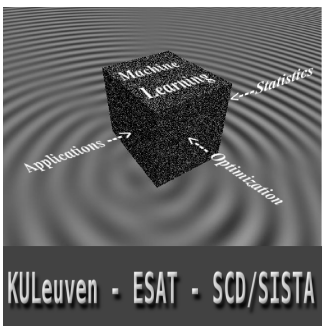
Empirical CCS



Theoretical CCS



Univariate $x_i \in \mathbb{R}$
 $M_i \rightarrow$ Discrete
 $m(x_i) \rightarrow$ Continuous



Clustering Shrinkage (Ct'd)

Modifications:

- 0-norm (count different pairs) → non-convex but interpretability!
- ϵ -neighborhood: $B(\epsilon)$ ball with measure $|B(\epsilon)|$

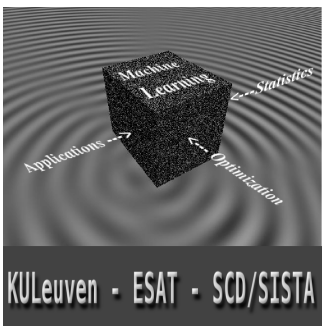
Empirical CCS



Theoretical CCS

$$\begin{aligned}
 \hat{m}_\epsilon &= \arg \min_{m: \mathbb{R}^D \rightarrow \mathbb{R}^D} \mathcal{J}_\gamma^{\epsilon, p}(m) \\
 &= \frac{1}{p} \sum_{i=1}^N \|m(x_i) - x_i\|_p + \frac{\gamma}{|B(\epsilon)|} \sum_{i=1}^N \sum_{\|x_i - x_j\| \leq \epsilon} I(\|m(x_i) - m(x_j)\| > 0),
 \end{aligned}
 \tag{1}$$

→ the second term measures the density of different assigned datapoints in a local neighborhood (cfr. histogram density estimator).



Clustering Shrinkage (Ct'd)

Definition 1. [Theoretical Shrinkage Clustering] Let $m : \mathbb{R} \rightarrow \mathbb{R}$ be such that $\lim_{\|\delta\| \rightarrow 0} \frac{m(x-\delta) - m(x+\delta)}{|B(\|\delta\|)|}$ exists almost everywhere. Let the cdf $P(x)$ underlying the dataset be known and assume its pdf $p(x)$ exists everywhere and is nonzero on a connected compact interval $C \subset \mathbb{R}$ with nonzero measure $|C| > 0$. We will study the following theoretical counterpart to (1)

$$\hat{m} = \arg \min_{m: \mathbb{R} \rightarrow \mathbb{R}} \mathcal{J}_\gamma^{p,0}(m) = \int_C \|m(x) - x\|_p dP(x) + \gamma \int_C \|m'(x)\|_0 dP(x), \quad (2)$$

where we define the latter term -denoted further as the zero-norm variation- formally as follows

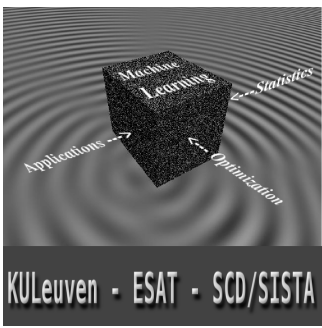
$$\|m'(x)\|_0 \triangleq \lim_{\epsilon \rightarrow 0} \left(\frac{I(m(B(x; \epsilon)) \neq \text{const})}{|B(x, \epsilon)|} \right), \quad (3)$$

with the characteristic function $I(m(B(x; \epsilon)) \neq \text{const})$ equals one if $\exists y \in B(x; \epsilon)$ such that $\|m(x) - m(y)\| > 0$ ($B(x, \epsilon)$ contains parts of different clusters), and equal to zero otherwise.

Empirical CCS



Theoretical CCS



Clustering Shrinkage (Ct'd)

Theorem 1. [Univariate Staircase Representation] When $P(x)$ is a fixed, smooth and differentiable distribution function with pdf $p : \mathbb{R} \rightarrow \mathbb{R}^+$ which is nonzero on a compact interval $C \subset \mathbb{R}$, the minimizer to (2) takes the form of a staircase function uniquely defined on C with a finite number of positive steps (say $K < +\infty$) of size $a = (a_1, \dots, a_K)^T \in \mathbb{R}^K$ at the points $\mathcal{D}_{(K)} = \{x_{(k)}\}_{k=1}^K \subset C$

Empirical CCS



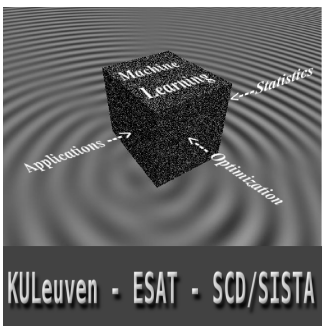
Theoretical CCS

$$\hat{m}(x; a, \mathcal{D}_{(K)}) = \sum_{k=1}^K a_k I(x > x_{(k)}) \quad \text{s.t.} \quad a_k \geq 0, x_{(k)} \in C \quad \forall k \quad (4)$$

Moreover, the optimization problem (2) is equivalent to the problem

$$\min_{a, \mathcal{D}_{(K)}} \mathcal{J}_K^p(a, \mathcal{D}_{(K)}) = \int_C \left\| \sum_{k=1}^K a_k I(x > x_{(k)}) - x \right\|_p p(x) dx + \sum_{k=1}^K p(x_{(k)}), \quad (5)$$

where $K \in \mathbb{N}$ relates to $\gamma \in \mathbb{R}^+$ in a way depending on \mathcal{D} .



Interpretations

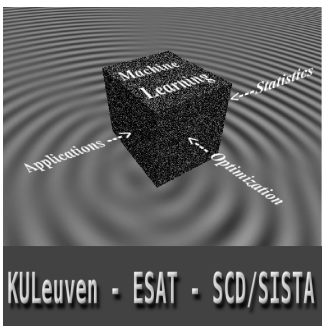
Unifying perspective:

- Vector Quantization (k -means)

Empirical CCS



Theoretical CCS



Interpretations

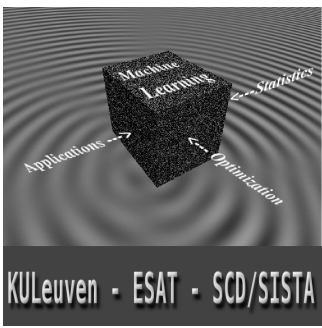
Unifying perspective:

- Vector Quantization (k -means)
- Bump-hunting and max-cut

Empirical CCS



Theoretical CCS



Interpretations

Unifying perspective:

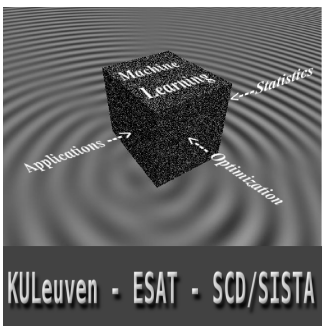
- Vector Quantization (k -means)
- Bump-hunting and max-cut
- Optimal coding: "finding a short code for X that preserves the maximum information about X itself."

$$L_2 \rightarrow \text{KL}$$

Empirical CCS



Theoretical CCS



Interpretations

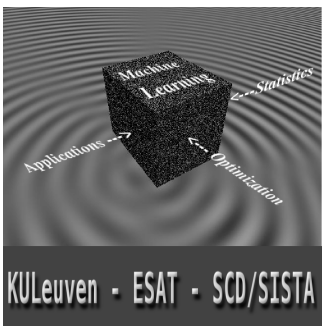
Unifying perspective:

- Vector Quantization (k -means)
- Bump-hunting and max-cut
- Optimal coding: "finding a short code for X that preserves the maximum information about X itself."
 $L_2 \rightarrow \text{KL}$
- Optimal bin placement

Empirical CCS



Theoretical CCS



Interpretations

Unifying perspective:

- Vector Quantization (k -means)
- Bump-hunting and max-cut
- Optimal coding: "finding a short code for X that preserves the maximum information about X itself."

$$L_2 \rightarrow \text{KL}$$

- Optimal bin placement

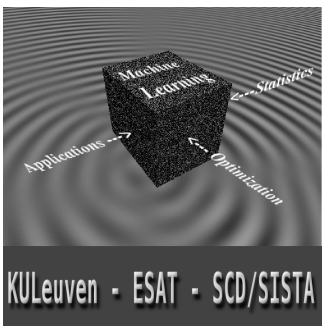
Main message:

- Optimization view to clustering

Empirical CCS



Theoretical CCS



Interpretations

Unifying perspective:

- Vector Quantization (k -means)
- Bump-hunting and max-cut
- Optimal coding: "finding a short code for X that preserves the maximum information about X itself."

$$L_2 \rightarrow \text{KL}$$

- Optimal bin placement

Main message:

- Optimization view to clustering
- Clustering \rightarrow study of the class of staircases (cfr. classification).

Empirical CCS



Theoretical CCS