Metric Learning as Convex Combinations of Local Models with Generalization Guarantees

Valentina Zantedeschi[1], Rémi Emonet[1], Marc Sebban[1]

[1] Univ Lyon, UJM-Saint-Etienne, CNRS, Institut d Optique Graduate School, Laboratoire Hubert Curien UMR 5516, F-42023, SAINT-ETIENNE, France
{valentina.zantedeschi,remi.emonet,marc.sebban}@univ-st-etienne.fr

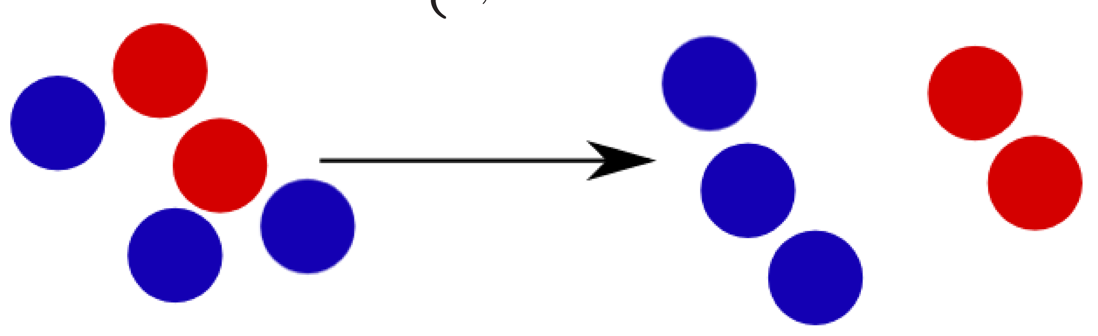
Metric Learning

Aim: Learn a better representation of the data that reflects their underlying geometry. This often leads to the optimization of a matrix M:

$$\text{argmin}_M \sum_{i,j} h_{ij} d_M(x_i, x_j) + \lambda \|M\|_F^2$$

$$s.t. \sum_{i,j} (1 - h_{ij}) d_M(x_i, x_j) \geq 1$$

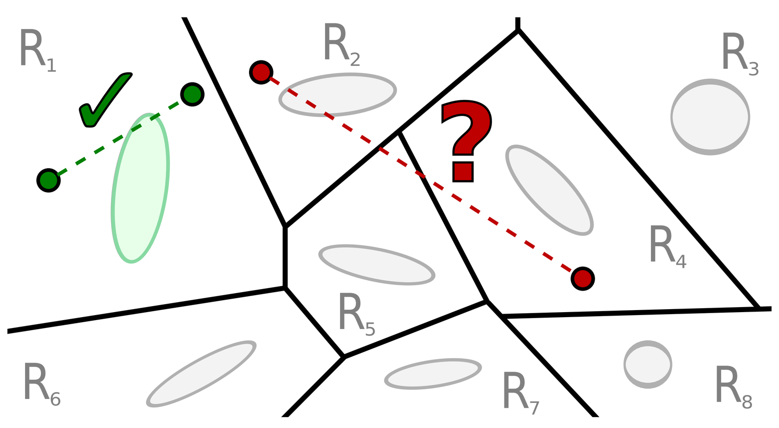
$$\text{where } h_{ij} = \begin{cases} 1, & \text{if } x_i, x_j \text{ are similar} \\ 0, & \text{otherwise} \end{cases}$$



$d_M(.)$ is a metric that measures the distance or the similarity between a pair of instances. For instance, it can be instantiated as a Mahalanobis distance $d_M(x_1, x_2) = \sqrt{(x_1 - x_2)^T M (x_1 - x_2)}$ or a bilinear similarity $d_M(x_1, x_2) = x_1^T M x_2$.

Local Metric Learning

In order to deal with non-linearities and multi-modalities, the instance space $U \subset \mathbb{R}^d$ is decomposed in K clusters or regions ($\{R_z\}_{z=1}^K$) and, on each cluster, a local model $s_z : U^2 \to \mathbb{R}$ is learned to compare instances belonging to that specific cluster.



However, local metric approaches:

1. are sensible to overfitting,
2. are no suited to compare points of different regions,
3. loose continuity in the metric space.

C2LM: Convex Combinations of Local Models

Let $S = \{s_z(.)\}_{z=1}^K$ be a set of (learned) metric functions defined on the regions $\{R_z\}_{z=1}^K$ of the instance space U. $\forall (R_i, R_j) = R_{ij}$ let $W_{ij} \in \mathbb{R}^K$ be the vector of contributions of each local model while estimating the similarity between $x_1 \in R_i$ and $x_2 \in R_j$. The **C2LM** optimization problem is defined as:

$$\text{argmin}_W \ \hat{R}^l + \lambda_1 D(W) + \lambda_2 S(W)$$

$$s.t. \ \forall i, j = 1, ..., K : \sum_{z=1}^K W_{ijz} = 1 \text{ and } W_{ij} \geq 0$$

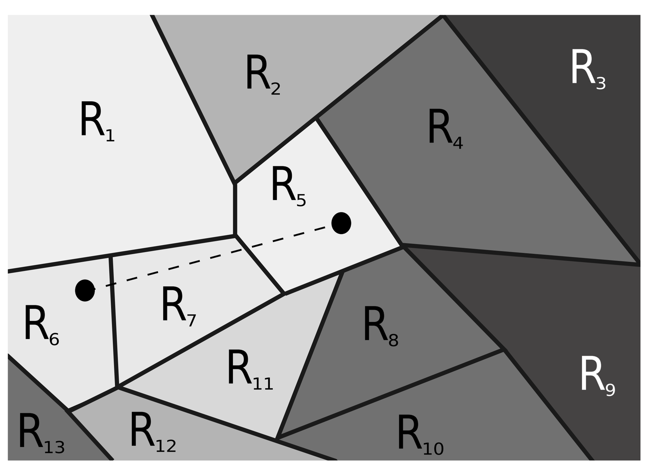
where

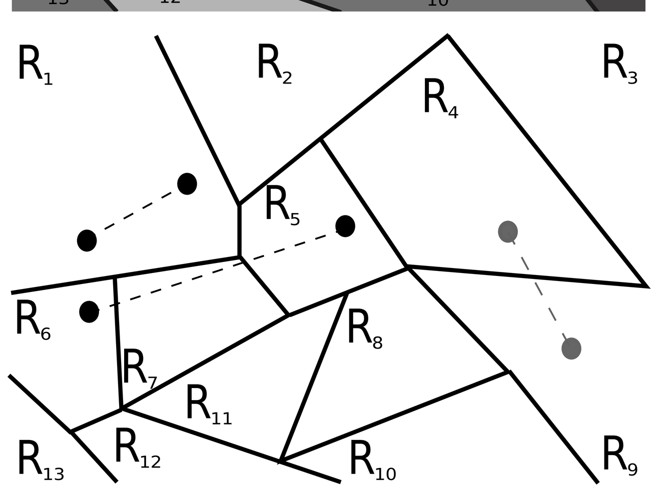
$$\hat{R}^l = \frac{1}{n} \sum_{i=1}^K \sum_{j=1}^i \sum_{p \in R_{ij}} \left| \sum_{z=1}^K W_{ijz} s_z(x_1, x_2) - y(x_1, x_2) \right|$$

is the mean loss over all training pairs $p = (x_1, x_2, y(x_1, x_2)) \in \mathcal{Z} = U^2 \times \mathbb{R}$, and

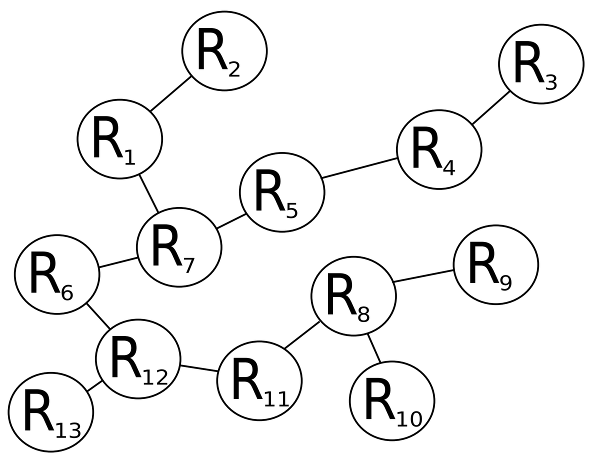
$$D(W) = \sum_{i=1}^K \sum_{j=1}^i \|E_{ij}^T W_{ij}\|_F^2$$

$$S(W) = \sum_{i=1}^K \sum_{j=1}^i \sum_{i'=1}^K \sum_{j'=1}^{i'} K_{ij i' j'} \|W_{ij} - W_{i'j'}\|_2^2$$





are two manifold regularizers, λ_1 and λ_2 are the corresponding regularization parameters. $D(W)$ takes into account the prior influence of each local model and $S(W)$ constrains the vectors defined on close pairs of regions to be similar.

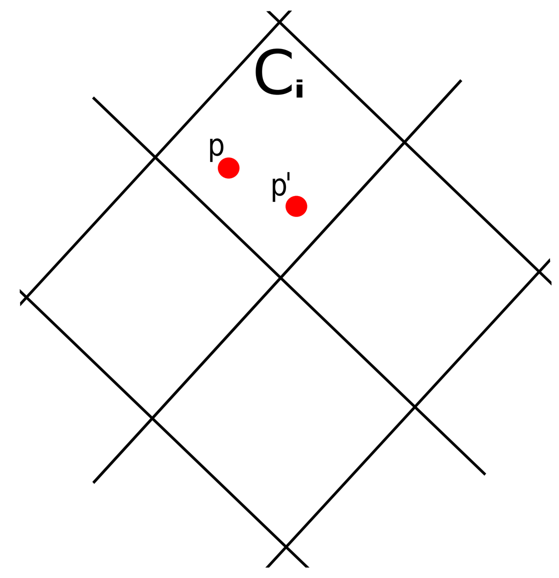


For estimating both regularization terms, we need to define a distance function between regions. For instance the number of edges of the shortest path connecting two regions in the Minimum Spanning Tree of the complete graph of region centroids.

Robustness and Generalization Bound

Algorithmic Robustness An algorithm A is said $(H, \epsilon(.))$-robust, for $H \in \mathbb{N}$ and $\epsilon : \mathcal{Z}^n \to \mathbb{R}$ if $\mathcal{Z}$ can be partitioned into H disjoint subsets, denoted by $\{C_i\}_{i=1}^H$, such that the following holds for all sets $P \in \mathcal{Z}^n$:

$$\forall p \in P, \forall p' \in \mathcal{Z}, \forall i = 1, ..., H$$
$$\text{if } p, p' \in C_i \text{ then } |l(p) - l(p')| \leq \epsilon(P)$$



with $l()$ some loss function used in the algorithm.

Robustness of C2LM

If $\forall z = 1, ..., K$, $s_z(.)$ is θ_z-lipschitz w.r.t. the norm $\|.\|_2$, **C2LM** is $(H, \theta\sqrt{2}\gamma_1 + \gamma_2)$-robust, with $\theta = \max_{z=1..K} \theta_z$.

H is the covering number of $\mathcal{Z}$: $\forall p, p' \in C_i$, $\|x_1 - x_1'\|_2 \leq \gamma_1$, $\|x_2 - x_2'\|_2 \leq \gamma_1$ and $|y(x_1, x_2) - y(x_1', x_2')| \leq \gamma_2$.

Generalization Bound

As C2LM is $(H, \theta\sqrt{2}\gamma_1 + \gamma_2)$-robust and the training set P is obtained from n IID draws according to a multinomial random variable, for any $\delta > 0$ with probability at least $1 - \delta$, we have:

$$|R^l - \hat{R}^l| \leq \theta\sqrt{2}\gamma_1 + \gamma_2 + B\sqrt{\frac{2H \ln 2 + 2 \ln 1/\delta}{n}}$$

with B the upper bound of the used loss and n the number of pairs of the dataset.

Application: Perceptual Color Distance Learning

Human perception of color distance strongly depends on variations of visual conditions and on camera configuration.

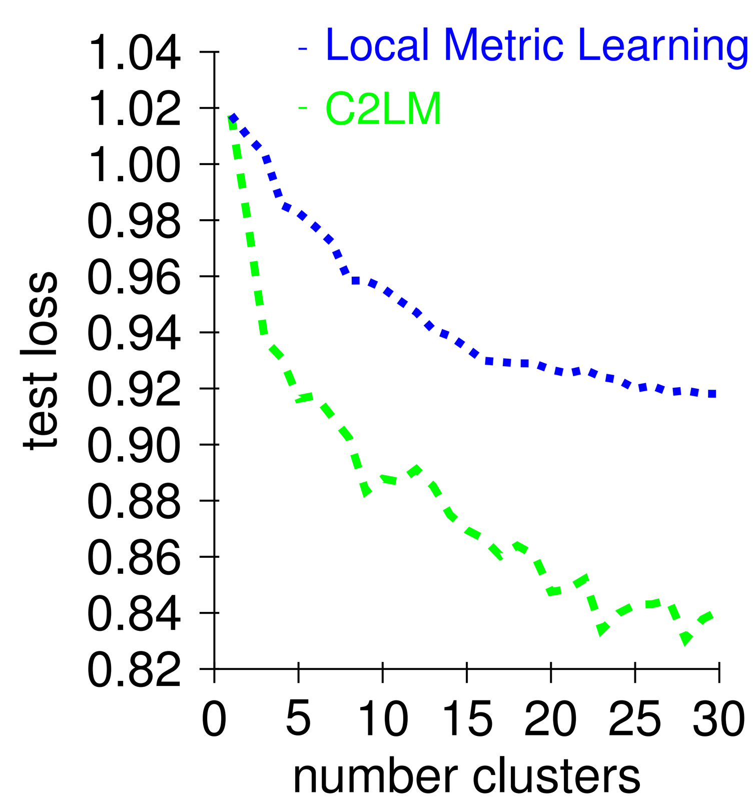
Experimental Setup:

1. Color patches are clustered using *k-means*.
2. A local model is learned on each region as a Mahalanobis-like distance.
3. **C2LM** is applied on the learned local models.

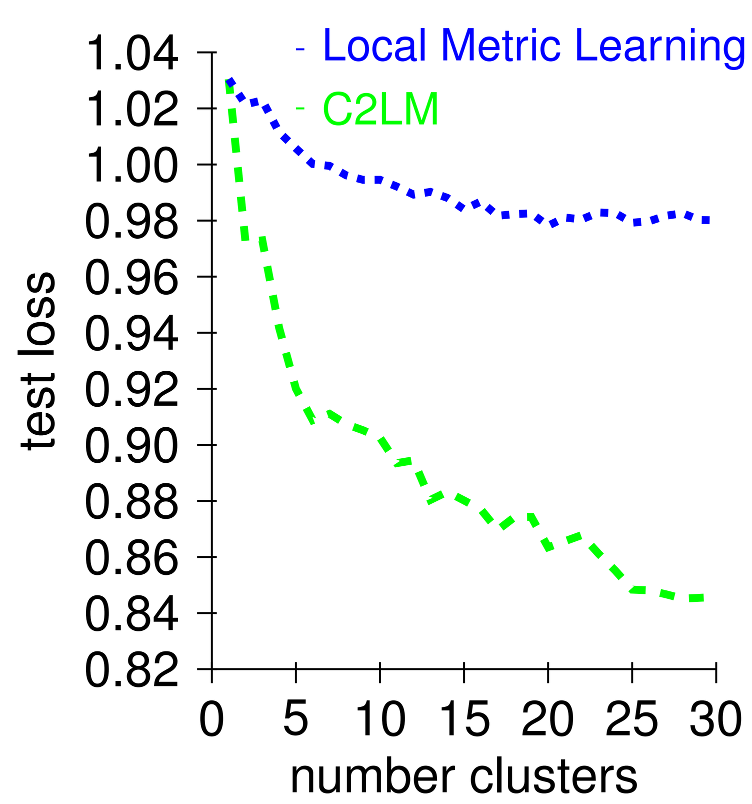
Dataset:

41800 pairs of color patches, taken under several viewing conditions and with 4 different cameras, with their reference perceptual distance ΔE_{00} computed using the CIEDE2000 color-difference formula based on CIELab space.

Results:

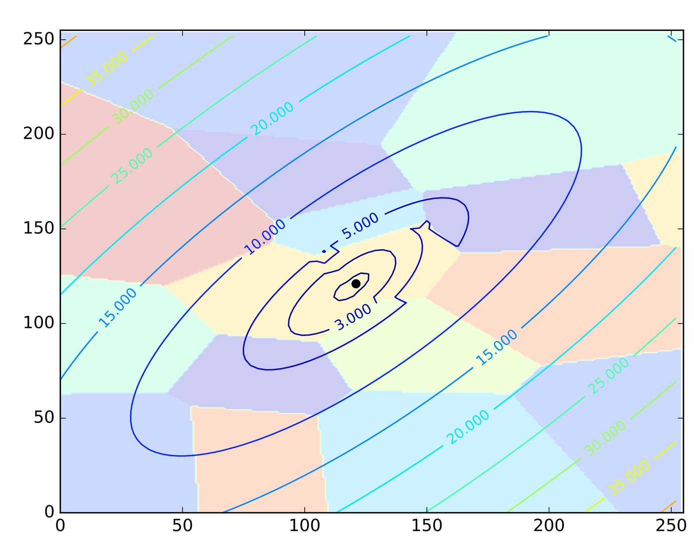


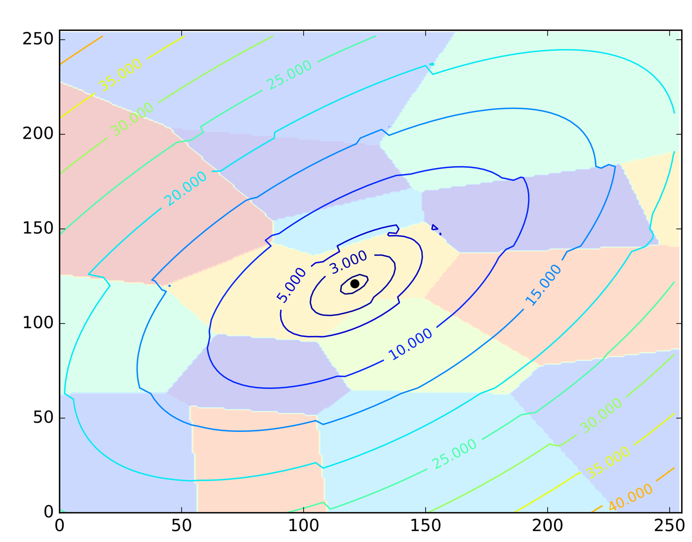
(a) Results on unseen colors



(b) Results on unseen cameras

C2LM allows us to learn smoother metrics than using local metric approaches.





2D projection of the contour lines of the metrics, drawn around an arbitrary point in the RGB space: (left) metric learned using a local metric approach; (right) metric learned with **C2LM**.