Robust Geometric Metric Learning

A. Collas[1], A. Breloy[2], G. Ginolhac[3], C. Ren[1], J.-P. Ovarlez[1,4]

[1]SONDRA, CentraleSupélec, University Paris-Saclay
[2]LEME, University Paris Nanterre
[3]LISTIC, University Savoie Mont Blanc
[4]DEMR, ONERA, University Paris-Saclay

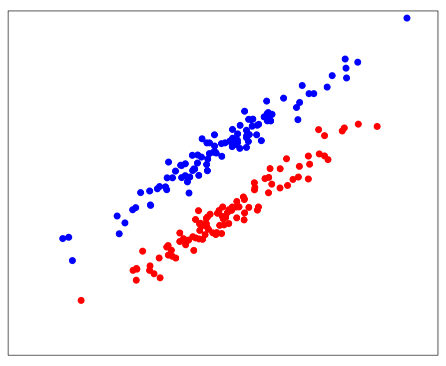
Github: github.com/antoinecollas/robust_metric_learning

Metric learning

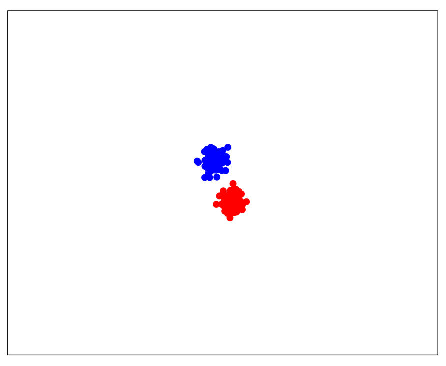
Supervised regime with K classes: $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$. Find a *Mahalanobis* distance

$$d_{\boldsymbol{A}}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sqrt{(\boldsymbol{x}_i - \boldsymbol{x}_j)^T \boldsymbol{A}^{-1} (\boldsymbol{x}_i - \boldsymbol{x}_j)}$$

relevant for classification problems.
$\boldsymbol{A} \in \mathcal{S}_p^{++}$ the set of $p \times p$ symmetric positive definite matrices.



$\{\boldsymbol{x}_i\}$



$\{\boldsymbol{A}^{-\frac{1}{2}} \boldsymbol{x}_i\}$

State of the art & covariance estimation

Set S: n_S pairs $(\boldsymbol{x}_l, \boldsymbol{x}_q)$ with $y_l = y_q$.
Set D: n_D pairs $(\boldsymbol{x}_l, \boldsymbol{x}_q)$ with $y_l \neq y_q$.

Information-Theoretic Metric Learning (ITML): [2]
Given $\boldsymbol{A}_0 \in \mathcal{S}_p^{++}$, and $u, v > 0$

$$\underset{\boldsymbol{A} \in \mathcal{S}_p^{++}}{\text{minimize}} \quad \text{Tr}(\boldsymbol{A}^{-1}\boldsymbol{A}_0) + \log|\boldsymbol{A}|$$

subject to $d^2_{\boldsymbol{A}}(\boldsymbol{x}_l, \boldsymbol{x}_q) \leq u, \quad (\boldsymbol{x}_l, \boldsymbol{x}_q) \in S$
$d^2_{\boldsymbol{A}}(\boldsymbol{x}_l, \boldsymbol{x}_q) \geq v, \quad (\boldsymbol{x}_l, \boldsymbol{x}_q) \in D$

$\boldsymbol{A}_0 = \frac{1}{m}\sum_{i=1}^m \boldsymbol{x}_i \boldsymbol{x}_i^T \implies$ *minimization of the Gaussian negative log-likelihood under constraints.*

Geometric Mean Metric Learning (GMML): [6]

$$\underset{\boldsymbol{A} \in \mathcal{S}_p^{++}}{\text{minimize}} \quad \frac{1}{n_S} \sum_{(\boldsymbol{x}_l, \boldsymbol{x}_q) \in S} d^2_{\boldsymbol{A}}(\boldsymbol{x}_l, \boldsymbol{x}_q) + \frac{1}{n_D} \sum_{(\boldsymbol{x}_l, \boldsymbol{x}_q) \in D} d^2_{\boldsymbol{A}^{-1}}(\boldsymbol{x}_l, \boldsymbol{x}_q)$$

Closed form solution (Riemannian interpolation):

$$\boldsymbol{A}^{-1} = \boldsymbol{S}^{-1} \#_t \boldsymbol{D} = \boldsymbol{S}^{-\frac{1}{2}} \left(\boldsymbol{S}^{\frac{1}{2}} \boldsymbol{D} \boldsymbol{S}^{\frac{1}{2}}\right)^t \boldsymbol{S}^{-\frac{1}{2}} \text{ with } t \in [0, 1]$$

$$\boldsymbol{S} = \frac{1}{n_S} \sum_{(\boldsymbol{x}_l, \boldsymbol{x}_q) \in S} (\boldsymbol{x}_l - \boldsymbol{x}_q)(\boldsymbol{x}_l - \boldsymbol{x}_q)^T \quad \text{and} \quad \boldsymbol{D} = \frac{1}{n_D} \sum_{(\boldsymbol{x}_l, \boldsymbol{x}_q) \in D} (\boldsymbol{x}_l - \boldsymbol{x}_q)(\boldsymbol{x}_l - \boldsymbol{x}_q)^T.$$

In practice, works well for t small, *i.e.* $\boldsymbol{A} \approx \boldsymbol{S}$.

Assumption: Data points of each class are realizations of independent random vectors with class-dependent first and second order moments

$$\boldsymbol{x}_{kl} \stackrel{d}{=} \boldsymbol{\mu}_k + \boldsymbol{\Sigma}_k^{\frac{1}{2}} \boldsymbol{u}_{kl}$$

with $\boldsymbol{\mu}_k \in \mathbb{R}^p$, $\boldsymbol{\Sigma}_k \in \mathcal{S}_p^{++}$, $\mathbb{E}[\boldsymbol{u}_{kl}] = \boldsymbol{0}$ and $\mathbb{E}[\boldsymbol{u}_{kl}\boldsymbol{u}_{kq}^T] = \boldsymbol{I}_p$ if $kl = kq$, $\boldsymbol{0}_p$ otherwise.

$$\implies \mathbb{E}[\boldsymbol{S}] = 2 \sum_{k=1}^K \pi_k \boldsymbol{\Sigma}_k \approx \mathbb{E}[\boldsymbol{A}]$$

where $\{\pi_k\}$ are the classes proportions.

Robust Geometric Metric Learning (RGML)

$$\underset{(\boldsymbol{A}, \{\boldsymbol{A}_k\}) \in \left(\mathcal{S}_p^{++}\right)^{K+1}}{\text{minimize}} \quad \underbrace{\sum_{k=1}^K \pi_k \mathcal{L}_k(\boldsymbol{A}_k)}_{\text{negative log-likelihood}} + \lambda \underbrace{\sum_{k=1}^K \pi_k d^2_{\mathcal{S}_p^{++}}(\boldsymbol{A}, \boldsymbol{A}_k)}_{\text{cost function to compute the center of mass of } \{\boldsymbol{A}_k\}}$$

where $\lambda > 0$ and $d_{\mathcal{S}_p^{++}}$ is the Riemannian distance on $\mathcal{S}_p^{++}$

$$d^2_{\mathcal{S}_p^{++}}(\boldsymbol{A}, \boldsymbol{A}_k) = \left\| \log\left(\boldsymbol{A}^{-\frac{1}{2}} \boldsymbol{A}_k \boldsymbol{A}^{-\frac{1}{2}}\right) \right\|_F^2.$$

Gaussian negative log-likelihood & Tyler cost function

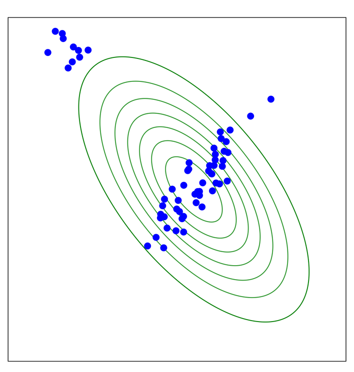
Set S_k: n_k pairs $(\boldsymbol{x}_l, \boldsymbol{x}_q)$ with $y_l = y_q = k$.

Gaussian negative log-likelihood:

$$\mathcal{L}_{G,k}(\boldsymbol{A}_k) = \frac{1}{n_k} \sum_{(\boldsymbol{x}_l, \boldsymbol{x}_q) \in S_k} (\boldsymbol{x}_l - \boldsymbol{x}_q)^T \boldsymbol{A}_k^{-1} (\boldsymbol{x}_l - \boldsymbol{x}_q) + \log|\boldsymbol{A}_k|$$

minimized for

$$\boldsymbol{A}_k = \frac{1}{n_k} \sum_{(\boldsymbol{x}_l, \boldsymbol{x}_q) \in S_k} (\boldsymbol{x}_l - \boldsymbol{x}_q)(\boldsymbol{x}_l - \boldsymbol{x}_q)^T$$

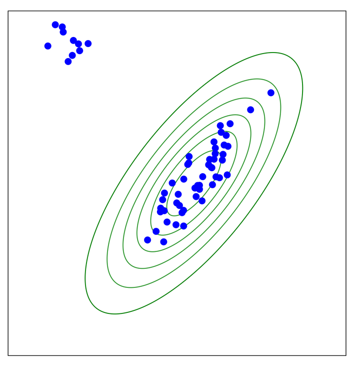


Tyler cost function: [5]

$$\mathcal{L}_{T,k}(\boldsymbol{A}_k) = \frac{p}{n_k} \sum_{(\boldsymbol{x}_l, \boldsymbol{x}_q) \in S_k} \log\left((\boldsymbol{x}_l - \boldsymbol{x}_q)^T \boldsymbol{A}_k^{-1} (\boldsymbol{x}_l - \boldsymbol{x}_q)\right) + \log|\boldsymbol{A}_k|$$

minimized for

$$\boldsymbol{A}_k = \frac{1}{n_k} \sum_{(\boldsymbol{x}_l, \boldsymbol{x}_q) \in S_k} \underbrace{\frac{p}{(\boldsymbol{x}_l - \boldsymbol{x}_q)^T \boldsymbol{A}_k^{-1} (\boldsymbol{x}_l - \boldsymbol{x}_q)}}_{\text{weight of } (\boldsymbol{x}_l - \boldsymbol{x}_q)} (\boldsymbol{x}_l - \boldsymbol{x}_q)(\boldsymbol{x}_l - \boldsymbol{x}_q)^T$$



Gaussian RGML & Tyler RGML

Gaussian RGML:

$$\underset{(\boldsymbol{A}, \{\boldsymbol{A}_k\}) \in \left(\mathcal{S}_p^{++}\right)^{K+1}}{\text{minimize}} \quad h_G(\boldsymbol{A}, \{\boldsymbol{A}_k\}) = \underbrace{\sum_{k=1}^K \pi_k \mathcal{L}_{G,k}(\boldsymbol{A}_k)}_{\text{Gaussian negative log-likelihood}} + \lambda \sum_{k=1}^K \pi_k d^2_{\mathcal{S}_p^{++}}(\boldsymbol{A}, \boldsymbol{A}_k)$$

Tyler RGML:

$$\underset{(\boldsymbol{A}, \{\boldsymbol{A}_k\}) \in \left(\mathcal{SS}_p^{++}\right)^{K+1}}{\text{minimize}} \quad h_T(\boldsymbol{A}, \{\boldsymbol{A}_k\}) = \underbrace{\sum_{k=1}^K \pi_k \mathcal{L}_{T,k}(\boldsymbol{A}_k)}_{\text{Tyler cost function}} + \lambda \sum_{k=1}^K \pi_k d^2_{\mathcal{S}_p^{++}}(\boldsymbol{A}, \boldsymbol{A}_k)$$

where $\mathcal{SS}_p^{++} = \left\{\boldsymbol{\Sigma} \in \mathcal{S}_p^{++} : |\boldsymbol{\Sigma}| = 1\right\}$ (unit determinant)

$\mathcal{S}_p^{++}$ and $\mathcal{SS}_p^{++}$ as Riemannian manifolds

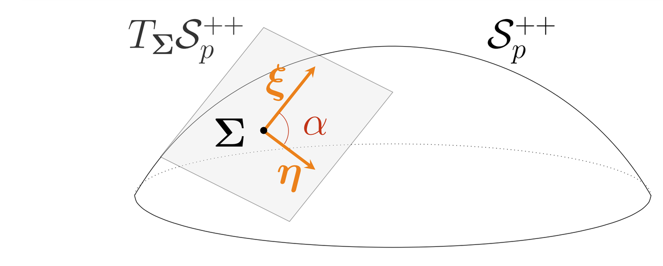


Figure 1. Representation of $\mathcal{S}_p^{++}$ as a Riemannian manifold with a point $\boldsymbol{\Sigma}$ and tangent vectors $\xi, \eta \in T_{\Sigma}\mathcal{S}_p^{++}$.

On $\mathcal{S}_p^{++} / \mathcal{SS}_p^{++}$: curvature induced by

- the Riemannian metric: $\langle \boldsymbol{\xi}, \boldsymbol{\eta} \rangle_{\boldsymbol{\Sigma}}^{\mathcal{S}_p^{++}} = \text{Tr}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\xi}\boldsymbol{\Sigma}^{-1}\boldsymbol{\eta})$.
- constraint on $\mathcal{SS}_p^{++}$: $|\boldsymbol{\Sigma}| = 1$.

Chosen Riemannian metric: $\forall \theta = (\boldsymbol{A}, \{\boldsymbol{A}_k\}), \forall \xi = (\boldsymbol{\xi}, \{\boldsymbol{\xi}_k\}), \eta = (\boldsymbol{\eta}, \{\boldsymbol{\eta}_k\})$

$$\langle \xi, \eta \rangle_\theta = \text{Tr}\left(\boldsymbol{A}^{-1}\boldsymbol{\xi}\boldsymbol{A}^{-1}\boldsymbol{\eta}\right) + \sum_{k=1}^K \text{Tr}\left(\boldsymbol{A}_k^{-1}\boldsymbol{\xi}_k\boldsymbol{A}_k^{-1}\boldsymbol{\eta}_k\right)$$

$\implies$ cost functions h_G and h_T are geodesically convex.

Riemannian gradient descents [1]

Given $\alpha > 0$ a step size

Iterations of Gaussian RGML:

$$\theta_{\ell+1} = \underbrace{R_{\theta_\ell}^{\left(\mathcal{S}_p^{++}\right)^{K+1}}}_{\text{retraction on } \left(\mathcal{S}_p^{++}\right)^{K+1}} \left(-\alpha \underbrace{\nabla^{\left(\mathcal{S}_p^{++}\right)^{K+1}} h_G(\theta_\ell)}_{\text{Riemannian gradient of } h_G} \right)$$

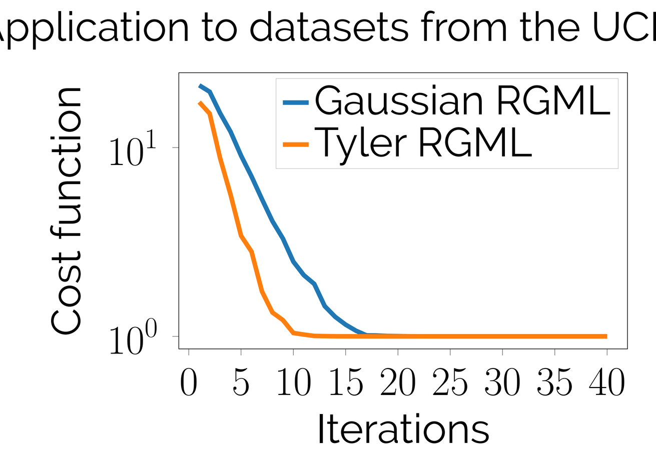
Iterations of Tyler RGML:

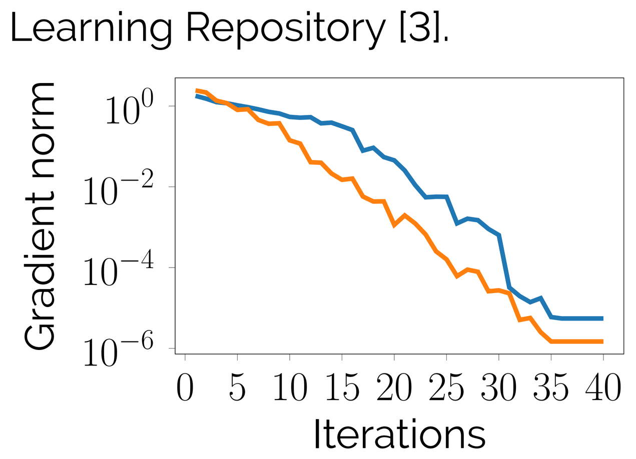
$$\theta_{\ell+1} = \underbrace{R_{\theta_\ell}^{\left(\mathcal{SS}_p^{++}\right)^{K+1}}}_{\text{retraction on } \left(\mathcal{SS}_p^{++}\right)^{K+1}} \left(-\alpha \underbrace{\nabla^{\left(\mathcal{SS}_p^{++}\right)^{K+1}} h_T(\theta_\ell)}_{\text{Riemannian gradient of } h_T} \right)$$

Retractions and Riemannian gradients are given in Section III of the paper.

Application

Application to datasets from the UCI Machine Learning Repository [3].





RGML + k-nearest neighbors

Method	Wine $p = 13$, $n = 178$, $K = 3$				Vehicle $p = 18$, $n = 846$, $K = 4$				Iris $p = 4$, $n = 150$, $K = 3$			
	Mislabeling rate				Mislabeling rate				Mislabeling rate			
	0%	5%	10%	15%	0%	5%	10%	15%	0%	5%	10%	15%
Euclidean	30.12	30.40	31.40	32.40	38.27	38.58	39.46	40.35	3.93	4.47	5.31	**6.70**
SCM	10.03	11.62	13.70	17.57	23.59	24.27	25.24	26.51	12.57	13.38	14.93	16.68
ITML - Identity	3.12	4.15	5.40	**7.74**	24.21	23.91	24.77	26.03	3.04	4.47	5.31	**6.70**
ITML - SCM	2.45	4.76	6.71	10.25	23.86	23.82	24.89	26.30	3.05	13.38	14.92	16.67
GMML	2.16	3.58	5.71	9.86	21.43	22.49	23.58	25.11	2.60	5.61	9.30	12.62
LMNN	4.27	6.47	7.83	9.86	20.96	24.23	26.28	28.89	3.53	9.59	11.19	12.22
Gaussian RGML	**2.07**	**2.93**	5.15	9.20	**19.76**	21.19	22.52	24.21	**2.47**	5.10	8.90	12.73
Tyler RGML	**2.12**	**2.90**	**4.51**	8.31	19.90	**20.96**	**22.11**	**23.58**	**2.48**	**2.96**	**4.65**	7.83

Table 1. Misclassification errors on 3 datasets: Wine, Vehicle and Iris. Mislabeling rate: percentage of labels randomly changed in the training set.

References

[1] P.-A. Absil et al. *Optimization Algorithms on Matrix Manifolds*. Princeton, NJ, USA: Princeton University Press, 2008.

[2] J. V. Davis et al. "Information-Theoretic Metric Learning". In: *Proceedings of the 24th International Conference on Machine Learning*. 2007.

[3] D. Dua and C. Graff. *UCI Machine Learning Repository*. 2017.

[4] E. Ollila et al. *Simultaneous penalized M-estimation of covariance matrices using geodesically convex optimization*. 2016.

[5] D. E. Tyler. "A Distribution-Free M-Estimator of Multivariate Scatter". In: *The Annals of Statistics* 15 (1987).

[6] P. Zadeh et al. "Geometric Mean Metric Learning". In: *Proceedings of The 33rd International Conference on Machine Learning*. Proceedings of Machine Learning Research. 2016.

antoine.collas@centralesupelec.fr