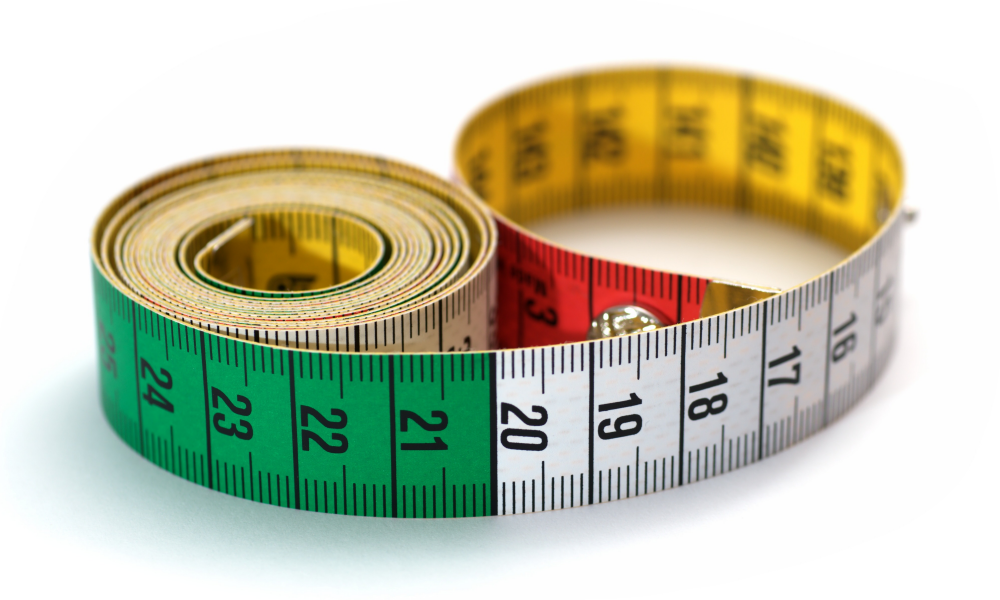
David R. Thompson
Jet Propulsion Laboratory, California Institute of Technology

Metric Learning

Sponsored by the Jet Propulsion Laboratory and by the Keck Institute for Space Studies, California Institute of Technology

Objectives

1. Understand the Mahalanobis Distance
2. Apply metric learning to a classification problem



Wikipedia

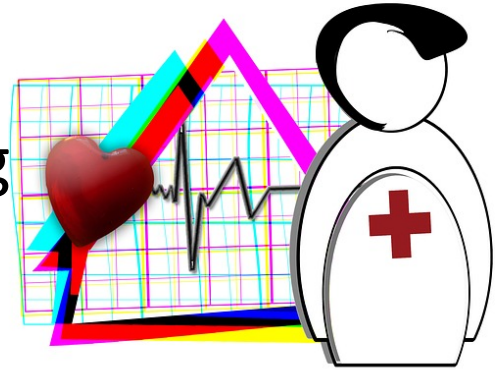
Motivation: predict heart attacks

Individual	Gender	Age	Weight (kg)	Income ($)
1	1	43	82	56,000
2	0	74	63	105,000
3	0	68	66	75,000
4	1	76	68	60,000

Problem: attributes are scaled differently, and may be correlated

Traditional solution: Preprocessing (normalization, standardization)

Can we do better with class information?



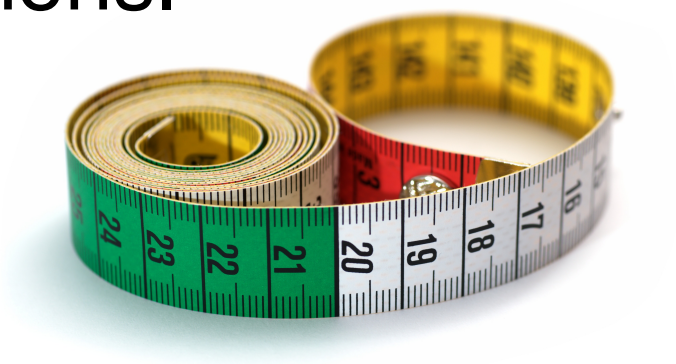
Pixabay.com

Getting serious about distances

Measuring distance is a critical aspect of pattern recognition

Until now, we've been letting the intrinsic data values define distances

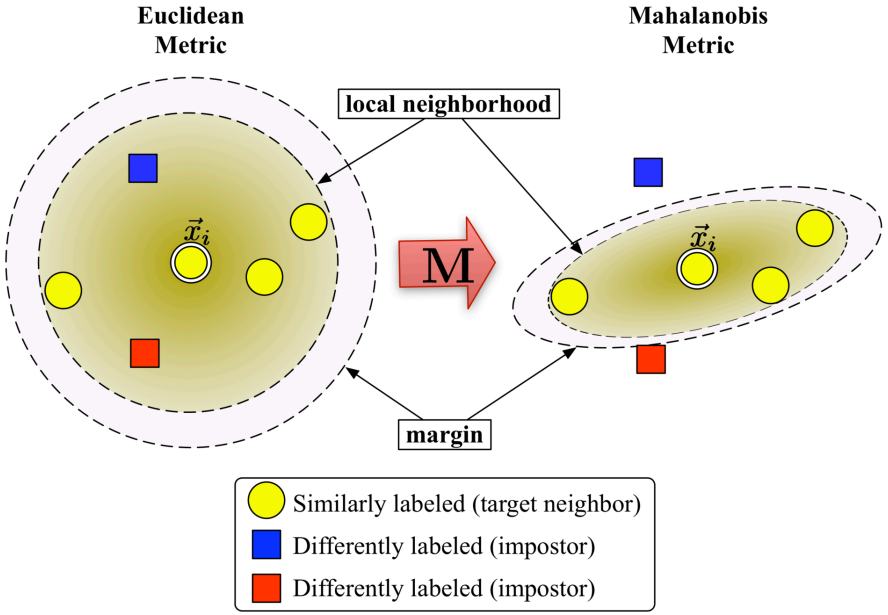
Recall Euclidean distances are less meaningful in high dimensions.



Wikipedia

Distances based on the data

One can use **non-isotropic** distances that reflect structure in class relationships.



wikipedia

Mahalanobis distance metric

- Squared distance weighted by the inverse covariance matrix

$$d(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)\ \Sigma_{\mathbf{X}}^{-1}\ (\mathbf{x}_i - \mathbf{x}_j)}$$

(An ellipsoidal distance function)

Mahalanobis distance metric

- Squared distance weighted by the inverse covariance matrix

$$d(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)\ \Sigma_{\mathbf{X}}^{-1}\ (\mathbf{x}_i - \mathbf{x}_j)}$$

- Equivalent to a linear pre-transformation of the data. Using $\Sigma_{\mathbf{X}}^{-1} = \mathbf{V}^T\mathbf{V}$ we have:

$$\begin{aligned} d(\mathbf{x}_i, \mathbf{x}_j) &= \sqrt{(\mathbf{x}_i - \mathbf{x}_j)^T\ \Sigma_{\mathbf{X}}^{-1}\ (\mathbf{x}_i - \mathbf{x}_j)} \\ &= \sqrt{(\mathbf{x}_i - \mathbf{x}_j)^T\ \mathbf{V}^T\mathbf{V}\ (\mathbf{x}_i - \mathbf{x}_j)} \\ &= \sqrt{(\mathbf{V}\mathbf{x}_i - \mathbf{V}\mathbf{x}_j)^T(\mathbf{V}\mathbf{x}_i - \mathbf{V}\mathbf{x}_j)} \end{aligned}$$

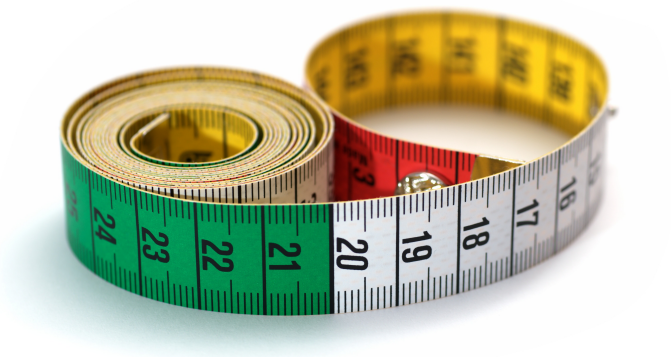
Euclidean distance for linearly-projected data

Metric learning

Most modern research treats metric learning as a convex optimization, learning a Mahalanobis metric while minimizing constraint violations (cannot link, must link)

Applications include

- Computer vision
- Information retrieval
- Bioinformatics



A few metric learning examples

Page	Name	Year	Source Code	Supervision	Form of Metric	Scalability w.r.t. n	Scalability w.r.t. d	Optimum	Dimension Reduction	Regularizer	Additional Information
11	MMC	2002	Yes	Weak	Linear	★☆☆	☆☆☆	Global	No	None	—
11	S&J	2003	No	Weak	Linear	★★☆	★★★	Global	No	Frobenius norm	—
12	NCA	2004	Yes	Full	Linear	★☆☆	★★☆	Local	Yes	None	For k-NN
12	MCML	2005	Yes	Full	Linear	★☆☆	☆☆☆	Global	No	None	For k-NN
13	LMNN	2005	Yes	Full	Linear	★★☆	★☆☆	Global	No	None	For k-NN
13	RCA	2003	Yes	Weak	Linear	★★☆	★★☆	Global	No	None	—
14	ITML	2007	Yes	Weak	Linear	★☆☆	★★☆	Global	No	LogDet	Online version
15	SDML	2009	No	Weak	Linear	★☆☆	★★☆	Global	No	LogDet+L_1	$n \ll d$
15	POLA	2004	No	Weak	Linear	★★★	★☆☆	Global	No	None	Online
15	LEGO	2008	No	Weak	Linear	★★★	★★☆	Global	No	LogDet	Online
16	RDML	2009	No	Weak	Linear	★★★	★★☆	Global	No	Frobenius norm	Online
16	MDML	2012	No	Weak	Linear	★★★	★☆☆	Global	Yes	Nuclear norm	Online
16	mt-LMNN	2010	Yes	Full	Linear	★★☆	☆☆☆	Global	No	Frobenius norm	Multi-task
17	MLCS	2011	No	Weak	Linear	★☆☆	★★☆	Local	Yes	N/A	Multi-task
17	GPML	2012	No	Weak	Linear	★☆☆	★★☆	Global	Yes	von Neumann	Multi-task
18	TML	2010	Yes	Weak	Linear	★★☆	★★☆	Global	No	Frobenius norm	Transfer learning
19	LPML	2006	No	Weak	Linear	★★☆	★★☆	Global	Yes	L_1 norm	—
19	SML	2009	No	Weak	Linear	★☆☆	☆☆☆	Global	Yes	$L_{2,1}$ norm	—
19	BoostMetric	2009	Yes	Weak	Linear	★☆☆	★★☆	Global	Yes	None	—
20	DML-p	2012	No	Weak	Linear	★☆☆	★☆☆	Global	No	None	—
20	RML	2010	No	Weak	Linear	★★☆	☆☆☆	Global	No	Frobenius norm	Noisy constraints
21	MLR	2010	Yes	Full	Linear	★★☆	☆☆☆	Global	Yes	Nuclear norm	For ranking
22	SiLA	2008	No	Full	Linear	★★☆	★★☆	N/A	No	None	Online
22	gCosLA	2009	No	Weak	Linear	★★★	☆☆☆	Global	No	None	Online
23	OASIS	2009	Yes	Weak	Linear	★★★	★★☆	Global	No	Frobenius norm	Online
23	SLLC	2012	No	Full	Linear	★★☆	★★☆	Global	No	Frobenius norm	For linear classif.
24	RSL	2013	No	Full	Linear	★☆☆	★★☆	Local	No	Frobenius norm	Rectangular matrix
25	LSMD	2005	No	Weak	Nonlinear	★☆☆	★★☆	Local	Yes	None	—
25	NNCA	2007	No	Full	Nonlinear	★☆☆	★★☆	Local	Yes	Recons. error	—
26	SVML	2012	No	Full	Nonlinear	★☆☆	★★☆	Local	Yes	Frobenius norm	For SVM
26	GB-LMNN	2012	No	Full	Nonlinear	★★☆	★★☆	Local	Yes	None	—
26	HDML	2012	Yes	Weak	Nonlinear	★★☆	★★☆	Local	Yes	L_2 norm	Hamming distance
27	M^2-LMNN	2008	Yes	Full	Local	★★☆	★☆☆	Global	No	None	—
28	GLML	2010	No	Full	Local	★★★	★★☆	Global	No	Diagonal	Generative
28	Bk-means	2009	No	Weak	Local	★☆☆	★★★	Global	No	RKHS norm	Bregman dist.
29	PLML	2012	Yes	Weak	Local	★★☆	☆☆☆	Global	No	Manifold+Frob	—
29	RFD	2012	Yes	Weak	Local	★★☆	★★★	N/A	No	None	Random forests
30	χ^2-LMNN	2012	No	Full	Nonlinear	★★☆	★★☆	Local	Yes	None	Histogram data
31	GML	2011	No	Weak	Linear	★☆☆	★★☆	Local	No	None	Histogram data
31	EMDL	2012	No	Weak	Linear	★☆☆	★★☆	Local	No	Frobenius norm	Histogram data
34	LRML	2008	Yes	Semi	Linear	★☆☆	☆☆☆	Global	No	Laplacian	—
35	M-DML	2009	No	Semi	Linear	★☆☆	☆☆☆	Local	No	Laplacian	Auxiliary metrics
35	SERAPH	2012	Yes	Semi	Linear	★☆☆	☆☆☆	Local	Yes	Trace+entropy	Probabilistic
36	CDML	2011	No	Semi	N/A	N/A	N/A	N/A	N/A	N/A	Domain adaptation
36	DAML	2011	No	Semi	Nonlinear	★☆☆	☆☆☆	Global	No	MMD	Domain adaptation

[Bellet, Habrard, Seban 2014]

Relationship with linear dimensionality reduction

If the sample covariance is low rank, or we use only a subspace of the result, this has a dimension-reducing effect

$$
\begin{aligned}
d(\mathbf{x}_i, \mathbf{x}_j) &= \sqrt{(\mathbf{x}_i - \mathbf{x}_j)^T \ \Sigma_{\mathbf{X}}^{-1} \ (\mathbf{x}_i - \mathbf{x}_j)} \\
&= \sqrt{(\mathbf{x}_i - \mathbf{x}_j)^T \ \mathbf{V}^T\mathbf{V} \ (\mathbf{x}_i - \mathbf{x}_j)} \\
&= \sqrt{(\mathbf{V}\mathbf{x}_i - \mathbf{V}\mathbf{x}_j)^T (\mathbf{V}\mathbf{x}_i - \mathbf{V}\mathbf{x}_j)}
\end{aligned}
$$

Multiclass Discriminant Analysis

- Start with labeled classes c_1, c_2, … c_k
- Learns a Mahalanobis distance metric that best separates these classes from each other.
- Provides a linear projection of dimensionality up to k-1
- 1-D case is equivalent to Fisher's classical Linear Discriminant analysis

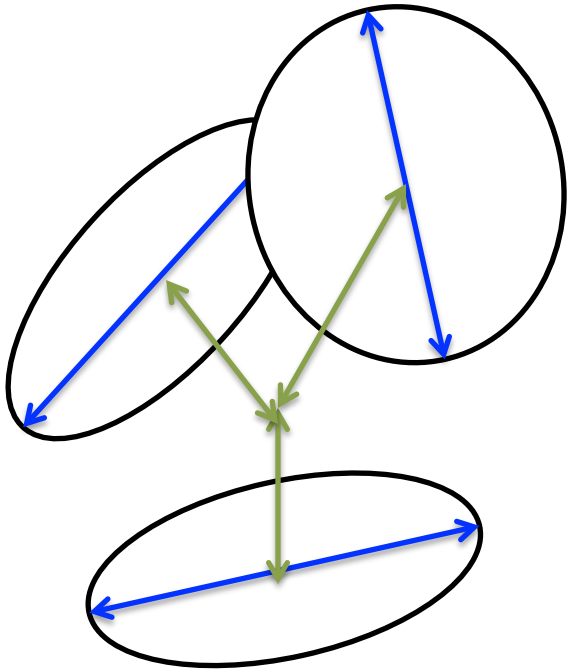
Definitions

Class mean: $\mu_i = \frac{1}{|C_i|} \sum_{\mathbf{x} \in C_i} \mathbf{x}$

Within class scatter matrix:

$$S_i = \sum_{\mathbf{x} \in C_i} (\mathbf{x} - \mu_i)(\mathbf{x} - \mu_i)^T$$

$$\Sigma_w = \sum_{i=1}^{k} S_i = \sum_{i} \sum_{\mathbf{x} \in C_i} (\mathbf{x} - \mu_i)(\mathbf{x} - \mu_i)^T$$



Between class scatter matrix:

$$\Sigma_b = \sum_{i=1}^{k} |C_i|(\mu_i - \mu)(\mu_i - \mu)^T$$

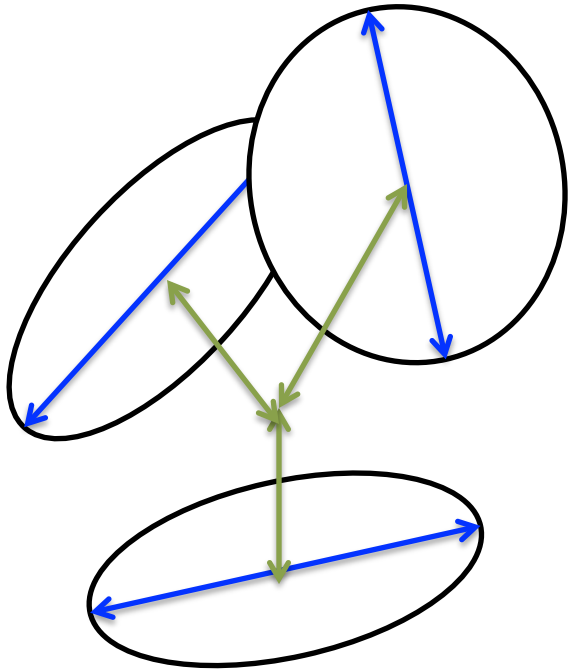
Total mean

To improve class separation..

We want a projection which shrinks the *within class* scatter

…while increasing the *between class* scatter

To measure volume, we use the determinant of the scatter matrices



Objective

- Find the projection V to maximize the ratio of the determinants:

$$\gamma(V) = \frac{|\mathbf{V}^T \Sigma_b \mathbf{V}|}{|\mathbf{V}^T \Sigma_w \mathbf{V}|}$$

← "Rayleigh coefficient"

Objective

- Find the projection V to maximize the ratio of the determinants:

$$\gamma(V) = \frac{|\mathbf{V}^T \Sigma_b \mathbf{V}|}{|\mathbf{V}^T \Sigma_w \mathbf{V}|}$$

← "Rayleigh coefficient"

- The solution is given by the following generalized eigenvalue problem:

$$\Sigma_b \mathbf{V} = \lambda \Sigma_w \mathbf{V}$$

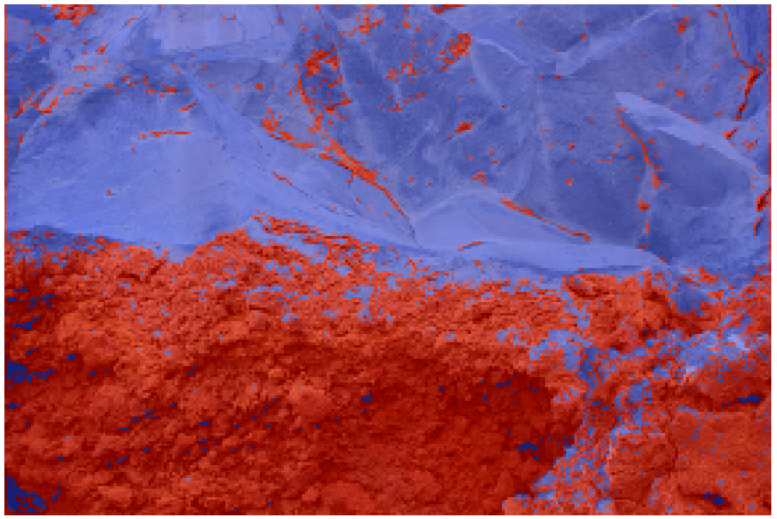
Example – metric learning to assist automated clustering

Geologic image analysis (Francis et al., iSAIRAS 2014).

Datapoints are a vector of pixel attributes including color, texture

Objective is to learn a projection that improves the quality of k-means image segmentation

Example: automated clustering

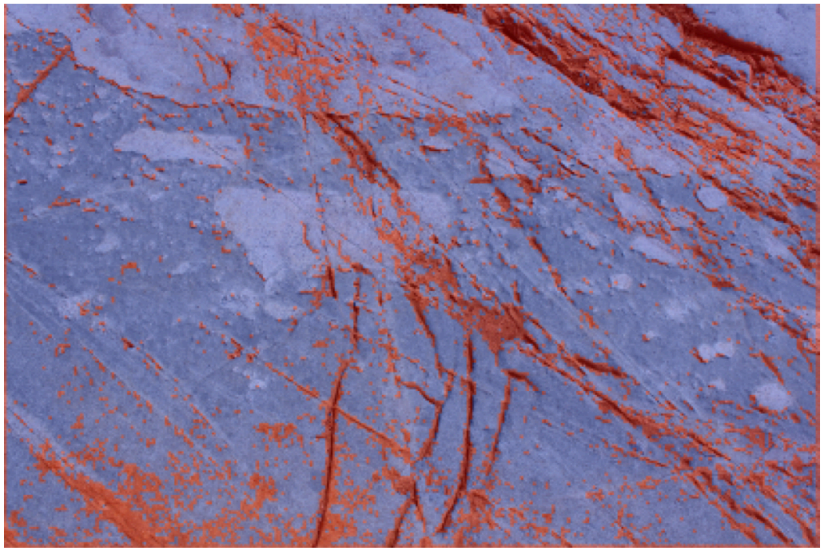


Geologic image analysis [Francis et al., iSAIRAS 2014)]

Datapoints are pixel attributes including color, texture

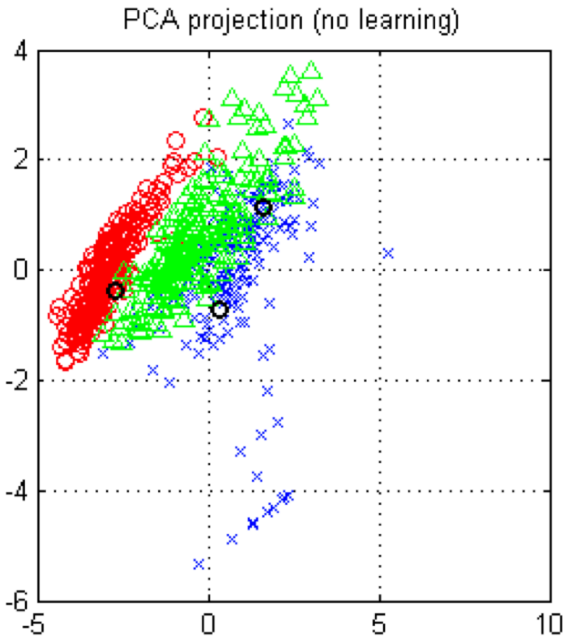
Objective is to find a projection that improves the quality of k-means image segmentation

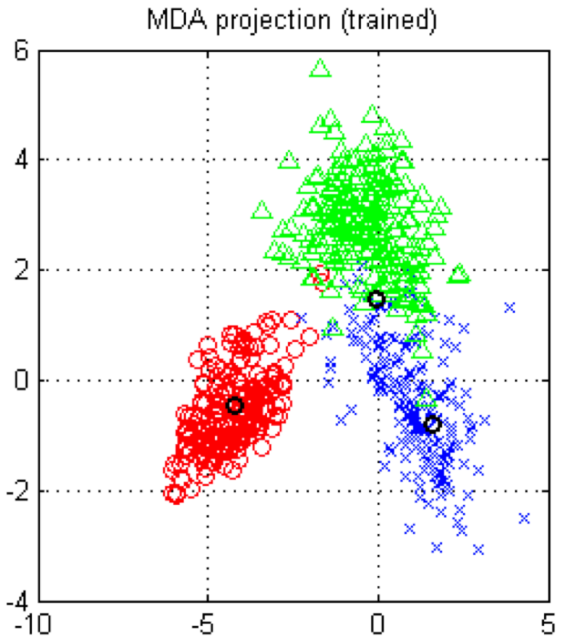
When clustering fails



Clustering raw pixel values does not always capture the geologic content of the scene

PCA vs. supervised projections

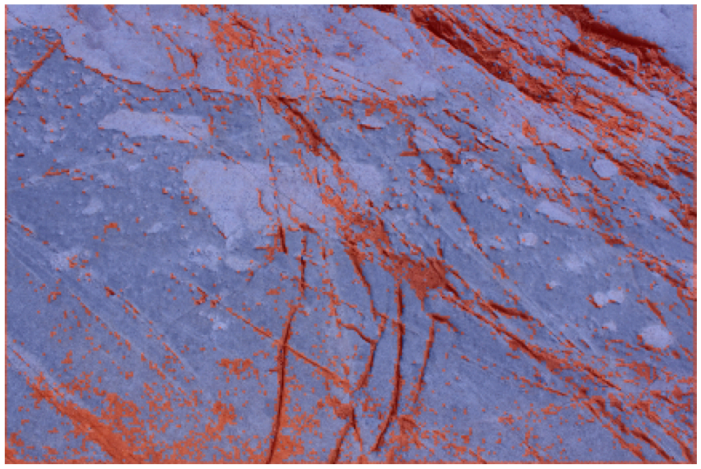




Multiclass discriminant analysis produces projections that reduce dimensionality while separating labeled training points from distinct classes

Effect on segmentation

Clustering in raw input space



Clustering in learned space

With supervised metric learning, k-means performs better and avoids confusion by incidental fracturing

Summary

- Mahalanobis distance metrics are equivalent to a linear transformation.
- Metric learning can outperform purely unsupervised dimensionality reduction by accounting for class structure.
- Multiclass Discriminant Analysis is a classical solution for projections of dimensionality up to k-1.