Regularized Mahalanobis Kernel for the Classification of Hyperspectral Images

M. Fauvel\textsuperscript{1}, A. Villa\textsuperscript{2,3}, J. Chanussot\textsuperscript{2} and J. A. Benediktsson\textsuperscript{3}

\textsuperscript{1} DYNAFOR, INRA & ENSAT, INPT, Université de Toulouse - France
\textsuperscript{2} GIPSA-Lab, Grenoble Institute of Technology - France
\textsuperscript{3} University of Iceland, Reykjavik - Iceland

Atelier Astrostatistique, Grenoble 2011
Outline

High dimensional spaces

Regularized Mahalanobis kernel
  Subspace models
  Mahalanobis kernel
  SVM and Radius margin bound maximization

Experiments

Conclusions and perspectives
High dimensional spaces

Regularized Mahalanobis kernel
  Subspace models
  Mahalanobis kernel
  SVM and Radius margin bound maximization

Experiments

Conclusions and perspectives
High dimensional data

- High number of measurements but limited number of samples.

\[ x_i \in \mathcal{X}^d \text{ with } d \gg 100, \ i \in \{1, \ldots, n\} \text{ and } n \approx d \]

- Hyperspectral images: each pixel has thousands of spectral variables
  - \( \mathcal{X} \) can be sparse
  - \( \mathcal{X} \) can have different SNR
  - Why:
    - A phenomenon depends on a lot of spectral variables
    - We don’t know which variables will be useful
    - Quality and quantity of information!
Some properties of HD spaces 1/3

- The volume of an hypersphere tend to zero when the dimension grows
  No closed neighbors

- The volume of an hypersphere concentrates in an outside shell
  Normally distributed data concentrates in the tails

- The volume of an hypersphere is negligible compare to the volume of an hypercube
  Uniformly distributed data concentrates in the corners

\[ V_s(d), \quad r = 1 \]

\[ (1 - \frac{\varepsilon}{r})^d, \quad \varepsilon = 0.1r \]

\[ \frac{V_s(d)}{V_c(d)} \]
Some properties of HD spaces 2/3

- pdf to have a sample \( \|x\| = t \ (x \sim \mathcal{N}(0, I)) \)

\[
f(t) = \frac{dt^{d-1} \exp(-t^2/2)}{d^{(d/2)} \Gamma(d/2 + 1)}, \quad \text{maximum for } t^* = (d - 1)^{0.5}
\]

- Some simulations: \( n = 5000, \|x\| \).

\[
x \sim U([-1, 1]), \ d = 10
\]

\[
x \sim U([-1, 1]), \ d = 200
\]

\[
x \sim \mathcal{N}(0, I), \ d = 10
\]

\[
x \sim \mathcal{N}(0, I), \ d = 200
\]
Some properties of HD spaces 3/3

- **Concentration of measure phenomenon**: if $x$ random vector with i.i.d. variables

$$\frac{d_M(x) - d_m(x)}{d_m(x)} \xrightarrow{p} 0$$

for all Minkowski norm: $||x|| = \left( \sum_{i=1}^d |x_i|^l \right)^{1/l}$, $l = 1, 2 \ldots$

- **Empty space phenomenon**: most of the space is empty

  A curse but also a blessing!
Implication for classification algorithms 1/2

- **Generative methods**
  - Hughes phenomenon: For a fixed training set, there exits an optimal dimension
  - Statistical estimation very difficult: Emptiness + number of parameters
  - Gaussian mixture models
    - Number of parameters $\propto d^2$ by class
    - $\Sigma^{-1}$ ill-posed
  - Non-parametric models
    - Number of samples to approximate a Gaussian law $\propto 10^{0.6d}$

- **Discriminative methods**
  - Number of points to uniformly sample a unit hypercube: $10^d$
  - Methods based on nearest neighbors fail:
    - k-nn
    - Adjacency matrix (e.g. laplacian graph)
    - Local kernel machines
  - More generally, *methods based on Euclidean distance fail*
Implication for classification algorithms 2/2

- **Emptiness phenomenon**: the classes are more separable!

\[ x_1 \sim \mathcal{N}(0, \text{I}) \text{ and } x_2 \sim \mathcal{N}(\varepsilon, \text{I}) \]

---

Gaussian mixture, **Minimum distance** and Linear-SVM
Existing solutions

- **Simple models:**
  - Linear models
  - Gaussian models: $\Sigma$ diagonal, equal for each class

- **Dimension reduction:** $x \rightarrow \phi(x)$
  - Statistical approach: PCA, FDA, ICA
  - Local distance: Laplacian eigenmaps, LLE, CCA

- **Kernel methods:** expect local kernels (evaluation of a new sample depends on its neighbors in the training set)

- **Regularization:** Tikhonov $\Sigma^{-1} \rightarrow (\Sigma + \lambda I)^{-1}$

- **Subspace models:** Each class is located in a specific subspace: $\Sigma$ is constrained
  - Probabilistic PCA
  - High Dimensional Discriminant Analysis (HDDA) models
Proposed approach

**Subspace models and kernel methods**

- Use **emptiness** property to construct the kernel

- **How:**
  - Mahalanobis distance for class $c$:
    \[
    d_{\Sigma_c}(x, z) = \sqrt{(x - z)^t \Sigma_c^{-1} (x - z)}
    \]
  - Gaussian Radial kernel:
    \[
    k_g(x, z) = \exp \left( - \frac{d(x, z)^2}{2\sigma^2} \right)
    \]
  - Mahalanobis kernel:
    \[
    k_m(x, z|c) = \exp \left( - \frac{(x - z)^t \Sigma_c^{-1} (x - z)}{2\sigma^2} \right)
    \]
High dimensional spaces

Regularized Mahalanobis kernel
- Subspace models
- Mahalanobis kernel
- SVM and Radius margin bound maximization

Experiments

Conclusions and perspectives
Kernel methods

- **Kernel function**: It computes the similarity between two samples. It is equivalent to a dot product in some feature space:

\[ k(x, z) = \langle \phi(x), \phi(z) \rangle_H, \quad \phi : \mathbb{R}^d \mapsto \mathcal{H} \]

- **Kernel methods**: The kernel is at the basis of the processing.

\[ f(x) = \sum_{i=1}^{n} \alpha_i k(x, x_i) + b \]

- **Some kernels**:
  - Linear: \( k(x, z) = \langle x, z \rangle \)
  - Polynomial: \( k(x, z) = (\langle x, z \rangle + q)^p \)
The HDDA model 1/3

- Family of **parsimonious models** for HD data [Bouveyron et al., 2007]

- **Cluster assumption:** each class \( c \) lives in a specific subspace

- Covariance matrix of class \( c \):

\[
\Sigma_c = Q_c \Lambda_c Q_c^t = \sum_{i=1}^{d} \lambda_{ci} q_{ci} q_{ci}^t
\]

- HDDA: \( \text{diag}(\Lambda_c) = \begin{bmatrix} \lambda_{c1} & \ldots & \lambda_{cp_c} \\ b_c & \ldots & b_c \end{bmatrix} \) with \( p_c \ll d \)

- Covariance matrix of class \( c \) under HDDA:

\[
\Sigma_c = \sum_{i=1}^{p_c} \lambda_{ci} q_{ci} q_{ci}^t + b_c \sum_{i=p_c+1}^{d} q_{ci} q_{ci}^t
\]

- \( A_c \) is the signal subspace and \( \bar{A}_c \) is the noise subspace (\( \mathbb{R}^d = A_c \oplus \bar{A}_c \))
The HDDA model 2/3

- In $\mathbb{R}^3$:

![Diagram showing the HDDA model in 3D space]

- The inverse can be computed explicitly:

\[
\Sigma_c^{-1} = \sum_{i=1}^{p_c} \frac{1}{\lambda_{ci}} q_{ci} q_{ci}^t + \frac{1}{b_c} \sum_{i=p_c+1}^{d} q_{ci} q_{ci}^t
\]

- Using $I = \sum_{i=1}^{d} q_{ci} q_{ci}^t$,

\[
\Sigma_c^{-1} = \sum_{i=1}^{p_c} \left( \frac{1}{\lambda_{ci}} - \frac{1}{b_c} \right) q_{ci} q_{ci}^t + \frac{1}{b_c} I
\]
So what?

- **Less parameters** have to be estimated ($d = 100$ and $p_c = 10$)
  - Full $\Sigma$: $d(d + 3)/2$ parameters $\rightarrow$ 5150
  - HDDA: $d(p_c + 1) + 2 - p_c(p_c - 1)/2$ parameters $\rightarrow$ 1057

- **Better than PCA**
  - $\mathbf{x}$ and $\mathbf{z}$ may be artificially closed in $\mathcal{A}_c$
  - An accurate estimation of $p_c$ is necessary

### Estimation: From the sample covariance matrix

$$\hat{\Sigma}_c = \frac{1}{n_c} \sum_{i=1}^{n_c} (\mathbf{x}_i - \bar{\mathbf{x}}_c)(\mathbf{x}_i - \bar{\mathbf{x}}_c)^t, \quad \mathbf{x}_i \in c$$

- $\{\hat{\lambda}_{ci}\}_{i=1}^{p_c}$ are estimated by the first $p_c$ eigenvalues of $\hat{\Sigma}_c$
- $\{\hat{\mathbf{q}}_{ci}\}_{i=1}^{p_c}$ are estimated by the first $p_c$ eigenvectors of $\hat{\Sigma}_c$
- $\hat{b}_c$ is estimated by $\left(\text{trace}(\hat{\Sigma}_c) - \sum_{i=1}^{\hat{p}_c} \hat{\lambda}_{ci}\right)/(d - \hat{p}_c)$
- $\hat{p}_c$ is estimated with the scree test of Catell
Mahalanobis kernel 1/2

- $\{\hat{\lambda}_{ci}\}_{i=1}^{p_c}$ and $\hat{b}_c$ are switched to kernel hyperparameters $\{\sigma_i\}_{i=1}^{p_c+1}$

- The kernel:

$$k_m(x, z|c) = \exp \left(-\frac{1}{2} \left( \sum_{i=1}^{\hat{p}_c} \frac{(x - z)^t \hat{q}_{ci} \hat{q}_{ci}^t(x - z)}{\sigma_i^2} + \frac{\|x - z\|^2}{\sigma_{\hat{p}_c+1}^2} \right) \right)$$

- Another formulation: product of Gaussian kernels

$$k_m(x, z|c) = k_g(x, z) \times \prod_{i=1}^{\hat{p}_c} k_g(\hat{q}_{ci}^t x, \hat{q}_{ci}^t z)$$

- The Mahalanobis kernel constructs with the HDDA model is a mixture of a Gaussian kernel on the original data and a Gaussian kernel on the $p_c$ first principal components of the considered class
$k_m(0, \mathbf{x} | c)$ with $0 = [0, 0]$ and $\mathbf{x} \in [-1, 1]^2$

- $\Sigma_c = [0.6 - 0.2; -0.2 0.6]$ and $p_c = 1$
- Red contour line $\rightarrow k_m = 0.75$
- (a): Gaussian kernel
- (b): Mahalanobis kernel with $\sigma_1^2 = \sigma_2^2 = 0.5$
- (c): Mahalanobis kernel with $\sigma_1^2 = 1.5$ and $\sigma_2^2 = 0.5$
L2-Support Vectors Machines 1/2

- Supervised method: \( S = \{(x_i, y_i)\}_{i=1}^{n}, x_i \in \mathbb{R}^{d} \) and \( y_i \in \{-1, 1\} \)

\[
h(z) = \text{sign}(f(z)) \text{ with } f(z) = \sum_{i=1}^{n} \alpha_i k(z, x_i) + b
\]

- Hyperparameters \((\{\alpha_i\}_{i=1}^{n}, b)\) learn by solving:

\[
\min_{\alpha, b} \left[ \frac{1}{C} \|f\|^2 + \sum_{i=1}^{n} L(y_i, f(x_i))^2 \right]
\]

\(\|f\|^2 = \sum_{i,j=1}^{n} \alpha_i \alpha_j k(x_i, x_j)\)

\(L(y_i, f(x_i))^2 = \max(0, 1 - y_if(x_i))^2\)

Regularized Mahalanobis Kernel, M. Fauvel, DYNAFOR - INRA 19/34
Equivalently: with \( \tilde{k}(x_i, x_j) = k(x_i, x_j) + C^{-1} \delta_{ij} \)

\[
\max_{\alpha} g(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \tilde{k}(x_i, x_j)
\]

subject to \( 0 \leq \alpha_i \) and \( \sum_{i=1}^{n} \alpha_i y_i = 0 \)

Toy examples:

\( C = 100 \)

\( C = 0.01 \)
In our setting $p = [\sigma_1^2, \ldots, \sigma_{\hat{p}+1}^2, C]$.

Estimate of the generalization error: Radius-margin bound (upper bound of LOO)

$$T(p) := R^2 \tilde{g}$$

$\tilde{g}$ depends on $(\tilde{\alpha}, p)$ and $\tilde{\alpha}$ depends on $p$. But, since $\tilde{g}$ depends on $\alpha$ via an optimization problem, the gradient of $\alpha$ w.r.t. $p$ does not enter into the computation of $\tilde{g}$.

$$\tilde{g}(p) = \max_{\alpha} g(p, \alpha) = g(p, \tilde{\alpha}(p))$$

$$\nabla \tilde{g} = \left( \frac{\partial g}{\partial p}, \frac{\partial g}{\partial \tilde{\alpha}} \right)$$

$$= \left( \frac{\partial g}{\partial p}, \left. \frac{\partial g}{\partial \alpha} \right|_{\alpha=\tilde{\alpha}} \frac{\partial \alpha}{\partial p} \right)$$

$$= \left( \frac{\partial g}{\partial p}, 0 \right)$$

Gradient descent on the radius margin bound: $\nabla T = \frac{\partial R^2}{\partial p} g + R^2 \frac{\partial g}{\partial p}$

Training: $\min \max$ problem (non-convex)
**Toy example:** \( \{ \mathbf{x} \mid \text{var}(x_1) \ll \text{var}(x_2) \} \)
Multiclass: one classifier per class (but $\text{SVM}_{c_i \ vs \ c_j} \neq \text{SVM}_{c_j \ vs \ c_i}$)

Complexity:
- HDDA: $\frac{2d^3}{3}$ or $p^2d$, computation of the eigenvalues/eigenvectors
- SVM: $\approx dn^3$, CQP solver
- Gradient step: $\approx (p + 1)n^2$
High dimensional spaces

Regularized Mahalanobis kernel
  Subspace models
  Mahalanobis kernel
  SVM and Radius margin bound maximization

Experiments

Conclusions and perspectives
Simulated data 1/3

- Experimental setup: Mixture of Gaussian following HDDA model

\[ \mathbf{x} = \sum_{i=1}^{c} \alpha_i \mathbf{s}_i + \mathbf{b}, \quad y = j \text{ such as } \alpha_j = \max_i \alpha_i \text{ and } \mathbf{s}_i \sim \text{HDDA} \]

- \( d = 413, \ p = 10, \ n = 1000, \ n_t = 1500 \) and \( \text{SNR} = 1 \)
- Mean values were extracted from spectral library
- Number of classes \( N_c = 2, 3 \) and 4
- 50 tries
Simulated data 1/3

- Experimental setup: Mixture of Gaussian following HDDA model

\[ \mathbf{x} = \sum_{i=1}^{c} \alpha_i \mathbf{s}_i + \mathbf{b}, \quad y = j \text{ such as } \alpha_j = \max_i \alpha_i \text{ and } \mathbf{s}_i \sim \text{HDDA} \]

- \( d = 413, \ p = 10, \ n = 1000, \ n_t = 1500 \) and \( SNR = 1 \)
- Mean values were extracted from spectral library
- Number of classes \( N_c = 2, 3 \) and 4
- 50 tries

\[
\begin{array}{cccc}
& \text{Gaussian} & \text{PCA} & \text{HDDA} \\
N_c = 2 & 0.92 & 0.94 & 0.96 \\
N_c = 3 & 0.65 & 0.7 & 0.75 \\
N_c = 4 & 0.75 & 0.8 & 0.85 \\
\end{array}
\]

Regularized Mahalanobis Kernel, M. Fauvel, DYNAFOR - INRA
The model has 5 parameters (Sylvain Douté): the grain size of water and CO$_2$ ice, the proportion of water, CO$_2$ ice and dust.

- $\mathbf{x} \in \mathbb{R}^{184}$ and $n = 31500$.

- Fives classes according to the grain size of water, $n = n_t = 15750$
Simulated data 3/3

- Estimated subspace size: \( s = 10^{-5} \)

<table>
<thead>
<tr>
<th>( c )</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \hat{p} )</td>
<td>15</td>
<td>14</td>
<td>12</td>
<td>13</td>
<td>14</td>
</tr>
</tbody>
</table>

- Classification accuracies:

<table>
<thead>
<tr>
<th>Kernel</th>
<th>Gaussian</th>
<th>PCA-Mahalanobis</th>
<th>HDDA-Mahalanobis</th>
</tr>
</thead>
<tbody>
<tr>
<td>( y = 50 )</td>
<td>99.7</td>
<td>99.7</td>
<td>99.8</td>
</tr>
<tr>
<td>( y = 150 )</td>
<td>97.6</td>
<td>98.2</td>
<td>98.3</td>
</tr>
<tr>
<td>( y = 250 )</td>
<td>94.7</td>
<td>96.0</td>
<td>96.1</td>
</tr>
<tr>
<td>( y = 350 )</td>
<td>89.4</td>
<td>93.4</td>
<td>93.4</td>
</tr>
<tr>
<td>( y = 450 )</td>
<td>95.0</td>
<td>95.3</td>
<td>95.4</td>
</tr>
<tr>
<td>OA</td>
<td>78.3</td>
<td>91.1</td>
<td>91.3</td>
</tr>
<tr>
<td>K</td>
<td>85.4</td>
<td>88.9</td>
<td>89.1</td>
</tr>
</tbody>
</table>

- McNemar(HDDA/PCA) \( \rightarrow 2.58 \)
Influence of the parameter $\hat{p}_c$

- OA vs $\hat{p}_c$ (class $y=350$):
Real data

- Data from the imaging spectrometer OMEGA (visible and infra red, 0.95-4.15, 184 wavelengths). Atmospherically corrected (S. Douté).
- Parameters learn with the simulated data.
- Colormap:
  - 0: no data
  - 1: $y = 50$
  - 2: $y = 150$
  - 3: $y = 250$
  - 4: $y = 350$
  - 5: $y = 450$
High dimensional spaces

Regularized Mahalanobis kernel
- Subspace models
- Mahalanobis kernel
- SVM and Radius margin bound maximization

Experiments

Conclusions and perspectives
Classification of hyperspectral images

A Mahalanobis kernel based on HDDA was proposed:
  ▶ Cluster assumption
  ▶ Multiple hyperparameters

Link with mixture kernels

SVM Classification framework

Good classification results on three data sets
  ▶ Better than the conventional RBF
  ▶ As good as PCA + RBF
Implementation: Optimization of the hyperparameters

- Estimation of $\hat{p}_c$

- Construction of others kernel:

$$k(x, z) = \left(x^t \Sigma^{-1} z + 1\right)^p$$

- Investigate mixture of kernels:

$$k_m(x, z | c) = \mu_o k_g(x, z) + \sum_{i=1}^{\hat{p}_c} \mu_i k_g(\hat{q}_c^t x, \hat{q}_c^t z)$$

- Discriminative subspaces (Fisher ... )
- Supervised - VS - Unsupervised
- Model transfert : From simulated data to real data
- Semi-supervised methods
- Face the strong non-linearity of the physical model (saturation of the parameters).
Regularized Mahalanobis Kernel
for the Classification of Hyperspectral Images

M. Fauvel\textsuperscript{1}, A. Villa\textsuperscript{2,3}, J. Chanussot\textsuperscript{2} and J. A. Benediktsson\textsuperscript{3}

\textsuperscript{1} DYNAFOR, INRA & ENSAT, INPT, Université de Toulouse - France
\textsuperscript{2} GIPSA-Lab, Grenoble Institute of Technology - France
\textsuperscript{3} University of Iceland, Reykjavik - Iceland

Atelier Astrostatistique, Grenoble 2011