Parsimonious Gaussian process models for the spectral-spatial classification of hyperspectral remote sensing images

Seminar MIAT

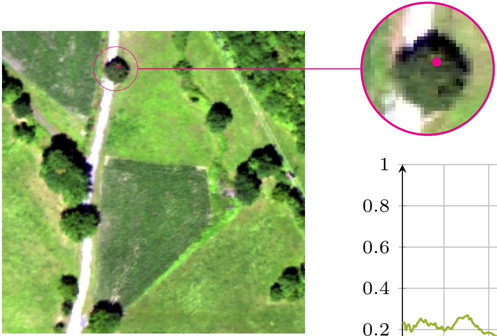
M. Fauvel [1], C. Bouveyron [2] and S. Girard [3]

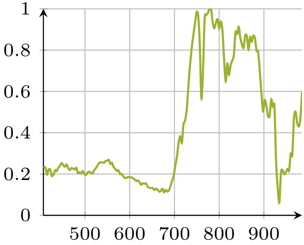
[1] UMR 1201 DYNAFOR INRA & Institut National Polytechnique de Toulouse
[2] Laboratoire MAP5, UMR CNRS 8145, Université Paris Descartes & Sorbonne Paris Cité
[3] Equipe MISTIS, INRIA Grenoble Rhône-Alpes & LJK

Outline

Nature of remote sensing images

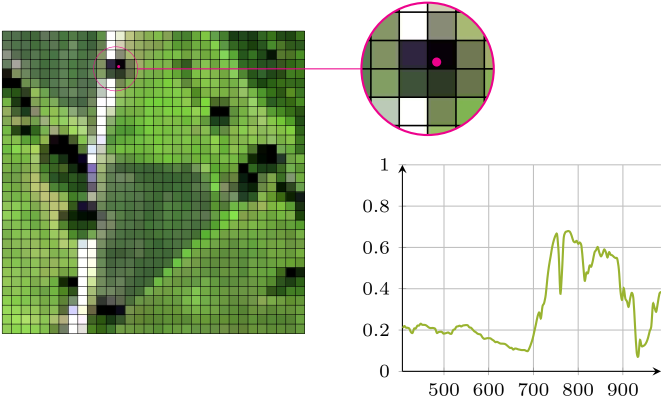
A remote sensing image is a sampling of a spatial, spectral and temporel process





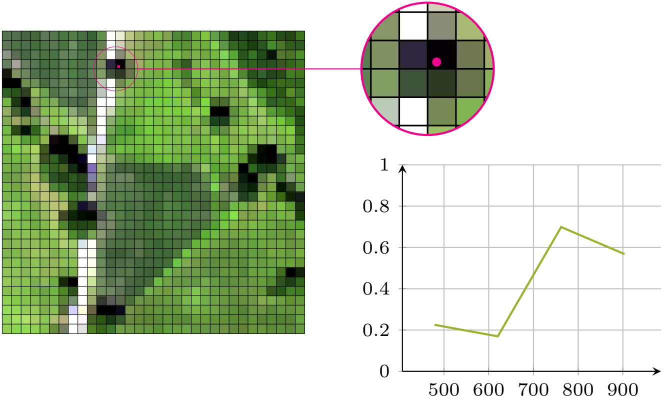
Nature of remote sensing images

A remote sensing image is a sampling of a spatial, spectral and temporel process



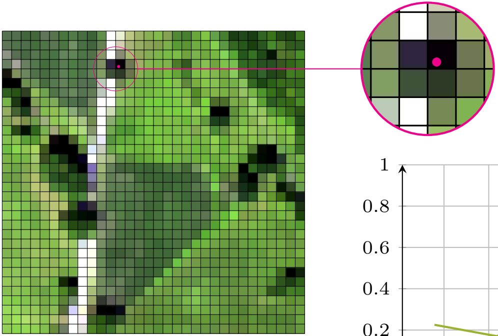
Nature of remote sensing images

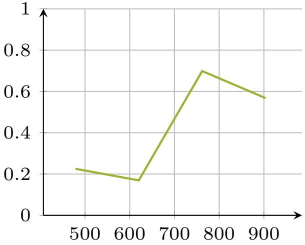
A remote sensing image is a sampling of a spatial, spectral and temporel process

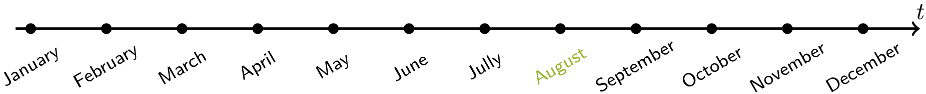


Nature of remote sensing images

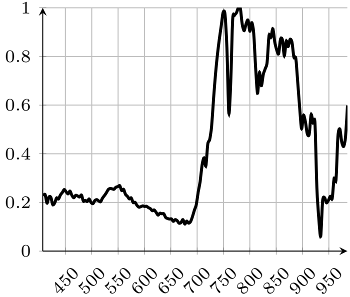
A remote sensing image is a sampling of a spatial, spectral and temporel process







Hyperspectral Imagery 1/3



- Pixels are represented by random vector $\mathbf{x} \in \mathbb{R}^d$ with d large, associated to a random variable x that represents the class/label.
- Classification: predict the membership y of $\mathbf{x}$, $y = f(\mathbf{x})$.

Hyperspectral Imagery 2/3

Instrument	Range (nm)	# Bands	Bandwidth (nm)	Spatial resolution (m)
AVIRIS	400-2500	224	10	20/1-4
HYDICE	400-2500	210	10	1-4
ROSIS-03	400-900	115	4	1
Hyspec	400-2500	427	3	1
HyMAP	400-2500	126	10-20	5
CASI	380-1050	288	2.4	1-2
HYPERION	400-2500	200	10	30

Hyperspectral Imagery 3/3

Definition of more classes with finer resolution:

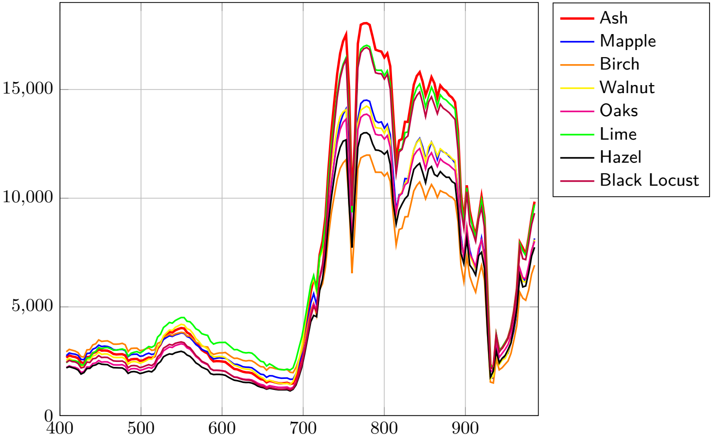
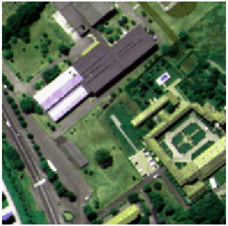
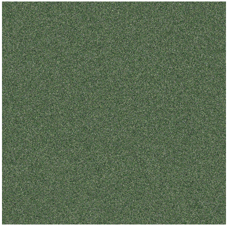


Image classification in high dimensional space

- High number of measurements but limited number of training samples.
- Curse of dimensionality: Statistical, geometrical and computational issues. Conventional method failed [Jimenez and Landgrebe, 1998].
- Kernel methods have shown great potential in many situations.
- Pixelwise classification not adapted [Fauvel et al., 2013].





- Need to incorporate spatial information in the classification process: **additional complexity**.

Kernel methods VS Parametric methods

1. Kernel methods [Camps-Valls and Bruzzone, 2009]:
   - Good abilities for classification,
   - Spatial information included through kernel function or additional features.

$$k_s(\mathbf{x}_i, \mathbf{x}_j) = \sum_{\substack{m \sim i \\ n \sim j}} k(\mathbf{x}_m, \mathbf{x}_n)$$

2. Parametric methods [Solberg et al., 1996]:
   - Markov Random Field: able to model spatial relationship between pixels,
   - Problem of the estimation of the spectral energy term.
3. *Parametric kernel methods*: probabilistic models in the kernel feature space.
   - Allow to get probability membership, with robust classifier
   - Allow to use the MRF modelization

Kernel methods and MRF

- Maximum *a posteriori*: $\max_Y (Y|\mathbf{X})$
- When Y is MRF: $P(Y|\mathbf{X}) \propto \exp(-U(Y|\mathbf{X}))$
  where $U(Y|\mathbf{X}) = \sum_{i=1}^{n} U(y_i|\mathbf{x}_i, \mathcal{N}_i)$ with

$$U(y_i|\mathbf{x}_i, \mathcal{N}_i) = \Omega(\mathbf{x}_i, y_i) + \rho\, \mathcal{E}(y_i, \mathcal{N}_i)$$

- Spectral term: $-\log[p(\mathbf{x}_i|y_i)]$
  - SVM outputs [Farag et al., 2005, Tarabalka et al., 2010, Moser and Serpico, 2013]
  - Kernel-probabilistic model [Dundar and Landgrebe, 2004]
- Spatial term
  - Potts model: $\mathcal{E}(y_i, \mathcal{N}_i) = \sum_{j \in \mathcal{N}_i} [1 - \delta(y_i, y_j)]$

y_1	y_2	y_3
y_4	y_i	y_5
y_6	y_7	y_8

(Kernel) Gaussian mixture models

- Quadratic decision rule in the input space

$$D_c(\mathbf{x}_i) = (\mathbf{x}_i - \boldsymbol{\mu}_c)^\top \boldsymbol{\Sigma}_c^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_c) + \log(\det(\boldsymbol{\Sigma}_c)) - 2\ln(\pi_c)$$

- Quadratic decision rule in the feature space [Dundar and Landgrebe, 2004]:

$$D_c\big(\phi(\mathbf{x}_i)\big) = \bar{\phi}_c(\mathbf{x}_i)^\top \mathbf{K}_c^{-1}\bar{\phi}_c(\mathbf{x}_i) + \log(\det(\mathbf{K}_c)) - 2\ln(\pi_c)$$

- Problem: $\mathbf{K}$ is badly conditioned (and non-invertible).
- Unlike SVM, there is no regularization for $\mathbf{K}_c^{-1}$ and $\log(\det(\mathbf{K}_c))$ in the estimation process.
- So it needs to be included in the model.

(Kernel) Gaussian mixture models

- Quadratic decision rule in the input space

$$D_c(\mathbf{x}_i) = (\mathbf{x}_i - \boldsymbol{\mu}_c)^\top \boldsymbol{\Sigma}_c^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_c) + \log(\det(\boldsymbol{\Sigma}_c)) - 2\ln(\pi_c)$$

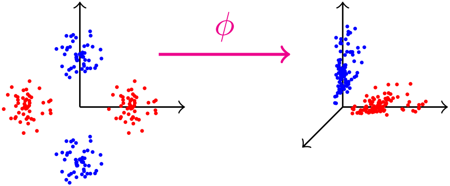
- Quadratic decision rule in the feature space [Dundar and Landgrebe, 2004]:

$$D_c\big(\phi(\mathbf{x}_i)\big) = \bar{\phi}_c(\mathbf{x}_i)^\top \mathbf{K}_c^{-1}\bar{\phi}_c(\mathbf{x}_i) + \log(\det(\mathbf{K}_c)) - 2\ln(\pi_c)$$

- Problem: $\mathbf{K}$ is badly conditioned (and non-invertible).
- Unlike SVM, there is no regularization for $\mathbf{K}_c^{-1}$ and $\log(\det(\mathbf{K}_c))$ in the estimation process.
- So it needs to be included in the model.

Enforce parsimony in the model

Kernel induced feature space



- Gaussian kernel: $k(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\gamma\|\mathbf{x}_i - \mathbf{x}_j\|^2_{\mathbb{R}^d}\right)$
- From Mercer theorem: $k(\mathbf{x}_i, \mathbf{x}_j) = \langle\phi(\mathbf{x}_i), \phi(\mathbf{x}_j)\rangle_{\mathcal{F}}$ which can be written

$$k(\mathbf{x}_i, \mathbf{x}_j) = \sum_{m=1}^{d_{\mathcal{F}}} \lambda_m \mathbf{q}_m(\mathbf{x}_i)\mathbf{q}_m(\mathbf{x}_j)$$

  where $d_{\mathcal{F}} = \dim(\mathcal{F})$.
- $\phi : \mathbf{x} \mapsto [\ldots, \sqrt{\lambda_m}\mathbf{q}_m(\mathbf{x}), \ldots],\ m = 1, 2, \ldots, d_{\mathcal{F}}$
- For the Gaussian kernel, $d_{\mathcal{F}} = +\infty$

Gaussian process

- Let us assume that $\phi(\mathbf{x})$, conditionally on $y = c$, is a Gaussian process with mean $\boldsymbol{\mu}_c$ and covariance function $\boldsymbol{\Sigma}_c$.
- The projection of $\phi(\mathbf{x})$ on the eigenfunction $\mathbf{q}_{cj}$ is noted $\phi(\mathbf{x})_j$:

$$\langle \phi(\mathbf{x}), \mathbf{q}_{cj} \rangle = \int_J \phi(\mathbf{x})(t)\mathbf{q}_{cj}(t)\,dt.$$

- The random vector $[\phi(\mathbf{x})_1, \ldots, \phi(\mathbf{x})_r] \in \mathbb{R}^r$ is, conditionally on $y = c$, a multivariate normal vector.
- Gaussian mixture model (Quadratic Discriminant) decision rules:

$$D_c\big(\phi(\mathbf{x}_i)\big) = \sum_{j=1}^{r} \left[\frac{\langle \phi(\mathbf{x}_i) - \boldsymbol{\mu}_c, \mathbf{q}_{cj} \rangle^2}{\lambda_{cj}} + \ln(\lambda_{cj}) \right] - 2\ln(\pi_c)$$

Gaussian process

- Let us assume that $\phi(\mathbf{x})$, conditionally on $y = c$, is a Gaussian process with mean $\boldsymbol{\mu}_c$ and covariance function $\boldsymbol{\Sigma}_c$.
- The projection of $\phi(\mathbf{x})$ on the eigenfunction $\mathbf{q}_{cj}$ is noted $\phi(\mathbf{x})_j$:

$$\langle \phi(\mathbf{x}), \mathbf{q}_{cj} \rangle = \int_J \phi(\mathbf{x})(t)\mathbf{q}_{cj}(t)\,dt.$$

- The random vector $[\phi(\mathbf{x})_1, \ldots, \phi(\mathbf{x})_r] \in \mathbb{R}^r$ is, conditionally on $y = c$, a multivariate normal vector.
- Gaussian mixture model (Quadratic Discriminant) decision rules: $r_c = \min(n_c, r)$

$$D_c(\phi(\mathbf{x}_i)) = \sum_{j=1}^{r_c} \left[\frac{\langle \phi(\mathbf{x}_i) - \boldsymbol{\mu}_c, \mathbf{q}_{cj} \rangle^2}{\lambda_{cj}} + \ln(\lambda_{cj}) \right] - 2\ln(\pi_c) + \sum_{j=r_c+1}^{r} \left[\frac{\langle \phi(\mathbf{x}_i) - \boldsymbol{\mu}_c, \mathbf{q}_{cj} \rangle^2}{\lambda_{cj}} + \ln(\lambda_{cj}) \right]$$

Definitions

Definition (Parsimonious Gaussian process with common noise)

$p\mathcal{GP}$ is a Gaussian process $\phi(\mathbf{x})$ for which, conditionally to $y = c$, the eigen-decomposition of its covariance operator $\mathbf{\Sigma}_c$ is such that

A1. It exists a dimension $r < +\infty$ such that $\lambda_{cj} = 0$ for $j \geq r$ and for all $c = 1, \ldots, C$.

A2. It exists a dimension $p_c < \min(r, n_c)$ such that $\lambda_{cj} = \lambda$ for $p_c < j < r$ and for all $c = 1, \ldots, C$.

Definition (Parsimonious Gaussian process with class specific noise)

A3. It exists a dimension $r_c < r$ such that $\lambda_{cj} = 0$ for all $j > r_c$ and for all $c = 1, \ldots, C$. When $r = +\infty$, it is assumed that $r_c = n_c - 1$.

A4. It exists a dimension $p_c < r_c$ such that $\lambda_{cj} = \lambda_c$ for $j > p_c$ and $j \leq r_c$, and for all $c = 1, \ldots, C$.

- A1 and A3 are motivated by the quick decay of the eigenvalues of Gaussian kernels.
- A2 and A4 express that the data of each class lives in a specific subspace of size p_c.

$p\mathcal{GP}$ models: List of sub-models

Model	Variance inside $\mathcal{F}_c$	$\mathbf{q}_{cj}$	p_c
	Variance outside $\mathcal{F}_c$: Common		
$p\mathcal{GP}_0$	Free	Free	Free
$p\mathcal{GP}_1$	Free	Free	Common
$p\mathcal{GP}_2$	Common within groups	Free	Free
$p\mathcal{GP}_3$	Common within groups	Free	Common
$p\mathcal{GP}_4$	Common between groups	Free	Common
$p\mathcal{GP}_5$	Common within and between groups	Free	Free
$p\mathcal{GP}_6$	Common within and between groups	Free	Common
	Variance outside $\mathcal{F}_c$: Free		
$np\mathcal{GP}_0$	Free	Free	Free
$np\mathcal{GP}_1$	Free	Free	Common
$np\mathcal{GP}_2$	Common within groups	Free	Free
$np\mathcal{GP}_3$	Common within groups	Free	Common
$np\mathcal{GP}_4$	Common between groups	Free	Common

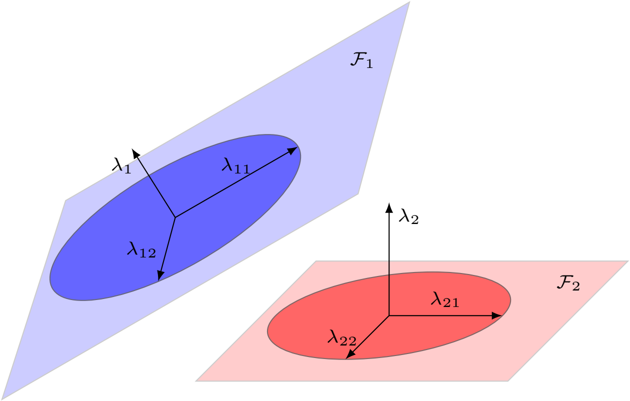


Figure: Visual illustration of model $np\mathcal{GP}_1$. Dimension of $\mathcal{F}_c$ is common to both classes, they have specific variance inside $\mathcal{F}_c$ and they have specific noise level.

Decision rules for $p\mathcal{GP}_0$

Proposition

For $p\mathcal{GP}_0$, the decision rule can be written:

$$D_c\big(\phi(\mathbf{x}_i)\big) = \sum_{j=1}^{p_c} \frac{\lambda - \lambda_{cj}}{\lambda_{cj}\lambda} \langle \phi(\mathbf{x}_i) - \boldsymbol{\mu}_c, \mathbf{q}_{cj} \rangle^2 - 2\ln(\pi_c) + \frac{\|\phi(\mathbf{x}) - \boldsymbol{\mu}_c\|^2}{\lambda} + \sum_{j=1}^{p_c} \ln(\lambda_{cj}) + (p_M - p_c)\ln(\lambda) + \gamma$$

where γ is a constant term that does not depend on the index c of the class.

- Proofs are given in [Bouveyron et al., 2014].
- Decompose the sum: $\sum_{j=1}^{p_c} \lambda_{cj} + \sum_{j=p_c+1}^{r} \lambda$
- Use the property: $\sum_{j=1}^{r} \langle \phi(\mathbf{x}) - \boldsymbol{\mu}_c, \mathbf{q}_{cj} \rangle^2 = \|\phi(\mathbf{x}) - \boldsymbol{\mu}_c\|^2$

Estimation of the parameters

- Centered Gaussian kernel function according to class c:

$$\bar{k}_c(\mathbf{x}_i, \mathbf{x}_j) = k(\mathbf{x}_i, \mathbf{x}_j) + \frac{1}{n_c^2} \sum_{\substack{l,l'=1 \\ y_l, y_l'=c}}^{n_c} k(\mathbf{x}_l, \mathbf{x}_{l'}) - \frac{1}{n_c} \sum_{\substack{l=1 \\ y_l=c}}^{n_c} \Big(k(\mathbf{x}_i, \mathbf{x}_l) + k(\mathbf{x}_j, \mathbf{x}_l) \Big).$$

and $\overline{\mathbf{K}}_c$ of size $n_c \times n_c$: $(\overline{\mathbf{K}}_c)_{l,l'} = \dfrac{\bar{k}_c(\mathbf{x}_l, \mathbf{x}_{l'})}{n_c}$.

- $\hat{\lambda}_{cj}$ is the j^{th} largest eigenvalue of $\overline{\mathbf{K}}_c$, and $\boldsymbol{\beta}_{cj}$ is its associated normalized eigenvector.
- $\hat{\lambda} = \dfrac{1}{\sum_{c=1}^{C} \hat{\pi}_c (r_c - \hat{p}_c)} \sum_{c=1}^{C} \hat{\pi} \Big(\text{trace}(\overline{\mathbf{K}}_c) - \sum_{j=1}^{\hat{p}_c} \hat{\lambda}_{cj} \Big).$
- $\hat{\pi}_c = n_c / n.$
- $\hat{p}_c$: percentage of cumulative variance.

Computable decision rule

Proposition

The decision rule can be computed as:

$$D_c\big(\phi(\mathbf{x}_i)\big) = \frac{1}{n_c}\sum_{j=1}^{\hat{p}_c}\frac{\hat{\lambda}-\hat{\lambda}_{cj}}{\hat{\lambda}_{cj}^2\hat{\lambda}}\left(\sum_{\substack{l=1\\y_l=c}}^{n_c}\beta_{cjl}\bar{k}_c(\mathbf{x}_i,\mathbf{x}_l)\right)^2 + \frac{\bar{k}_c(\mathbf{x}_i,\mathbf{x}_i)}{\hat{\lambda}} + \sum_{j=1}^{\hat{p}_c}\ln(\hat{\lambda}_{cj}) + (\hat{p}_M-\hat{p}_c)\ln(\hat{\lambda}) - 2\ln(\hat{\pi}_c)$$

- Proofs are given in [Bouveyron et al., 2014].
- Use of the property that the eigenfunction of the covariance function is a linear combination of $\phi(\mathbf{x}_i)-\boldsymbol{\mu}_c$
- $\langle\phi(\mathbf{x}_i)-\boldsymbol{\mu}_c,\phi(\mathbf{x}_j)-\boldsymbol{\mu}_c\rangle = \bar{k}_c(\mathbf{x}_i,\mathbf{x}_j)$

Numerical considerations

- The proposed model allow a *safe* computation of $\mathbf{K}_c^{-1}$ and $\log\Big(\det(\mathbf{K}_c)\Big)$ that appears in the kernel quadratic decision rule.
- Only the p_c first eigenvector/eigenvalue are used
- Eigenvectors corresponding to small eigenvalues are not used
- If p_cs are not too large, $\log(\hat{\lambda})$ is stable.

Numerical considerations

- The proposed model allow a *safe* computation of $\mathbf{K}_c^{-1}$ and $\log\big(\det(\mathbf{K}_c)\big)$ that appears in the kernel quadratic decision rule.
- Only the p_c first eigenvector/eigenvalue are used
- Eigenvectors corresponding to small eigenvalues are not used
- If p_cs are not too large, $\log(\hat{\lambda})$ is stable.

Proof: $\mathbf{K}_c$ is *pdf* so it can be decomposed into $\mathbf{Q}_c\mathbf{\Lambda}_c\mathbf{Q}_c^\top = \sum_{j=1}^{r}\lambda_{cj}\mathbf{q}_{cj}\mathbf{q}_{cj}^\top$

$$\begin{aligned}\mathbf{K}_c^{-1} &= \mathbf{Q}_c\mathbf{\Lambda}_c^{-1}\mathbf{Q}_c^\top = \sum_{j=1}^{r}\lambda_{cj}^{-1}\mathbf{q}_{cj}\mathbf{q}_{cj}^\top = \sum_{j=1}^{p_c}\lambda_{cj}^{-1}\mathbf{q}_{cj}\mathbf{q}_{cj}^\top + \lambda^{-1}\sum_{j=p_c+1}^{r}\mathbf{q}_{cj}\mathbf{q}_{cj}^\top \\ &= \sum_{j=1}^{p_c}\lambda_{cj}^{-1}\mathbf{q}_{cj}\mathbf{q}_{cj}^\top + \lambda^{-1}\left(\mathbf{I}_{n_c} - \sum_{j=1}^{p_c}\mathbf{q}_{cj}\mathbf{q}_{cj}^\top\right) = \sum_{j=1}^{p_c}\frac{\lambda-\lambda_{cj}}{\lambda\lambda_{cj}}\mathbf{q}_{cj}\mathbf{q}_{cj}^\top + \lambda^{-1}\mathbf{I}_{n_c}\end{aligned}$$

Numerical considerations

- The proposed model allow a *safe* computation of $\mathbf{K}_c^{-1}$ and $\log\big(\det(\mathbf{K}_c)\big)$ that appears in the kernel quadratic decision rule.
- Only the p_c first eigenvector/eigenvalue are used
- Eigenvectors corresponding to small eigenvalues are not used
- If p_cs are not too large, $\log(\hat{\lambda})$ is stable.

Proof: $\mathbf{K}_c$ is *pdf* so it can be decomposed into $\mathbf{Q}_c\mathbf{\Lambda}_c\mathbf{Q}_c^\top = \sum_{j=1}^{r}\lambda_{cj}\mathbf{q}_{cj}\mathbf{q}_{cj}^\top$

$$\begin{aligned}
\mathbf{K}_c^{-1} &= \mathbf{Q}_c\mathbf{\Lambda}_c^{-1}\mathbf{Q}_c^\top = \sum_{j=1}^{r}\lambda_{cj}^{-1}\mathbf{q}_{cj}\mathbf{q}_{cj}^\top = \sum_{j=1}^{p_c}\lambda_{cj}^{-1}\mathbf{q}_{cj}\mathbf{q}_{cj}^\top + \lambda^{-1}\sum_{j=p_c+1}^{r}\mathbf{q}_{cj}\mathbf{q}_{cj}^\top \\
&= \sum_{j=1}^{p_c}\lambda_{cj}^{-1}\mathbf{q}_{cj}\mathbf{q}_{cj}^\top + \lambda^{-1}\left(\mathbf{I}_{n_c} - \sum_{j=1}^{p_c}\mathbf{q}_{cj}\mathbf{q}_{cj}^\top\right) = \sum_{j=1}^{p_c}\frac{\lambda-\lambda_{cj}}{\lambda\lambda_{cj}}\mathbf{q}_{cj}\mathbf{q}_{cj}^\top + \lambda^{-1}\mathbf{I}_{n_c}
\end{aligned}$$

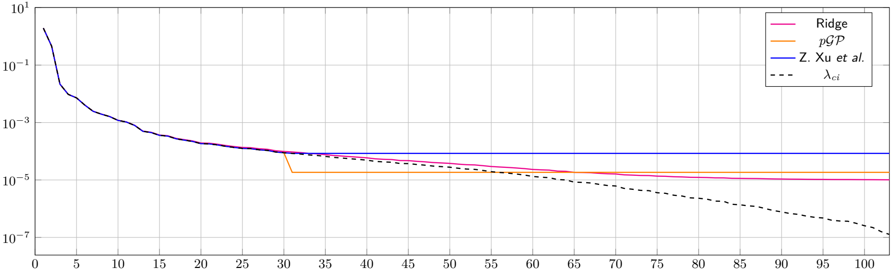
$$\log\big(\det(\mathbf{K}_c)\big) = \sum_{j=1}^{p_c}\log(\lambda_{cj}) + (r-p_c)\log(\lambda)$$

Existing models

- [Dundar and Landgrebe, 2004]
  Equal covariance matrix assumption and ridge regularization. Complexity: $\mathcal{O}(n^3)$.
  Similar to $p\mathcal{GP}_4$ with equal eigenvectors.
- [Pekalska and Haasdonk, 2009]
  Ridge regularization, per class. Complexity: $\mathcal{O}(n_c^3)$.
- [Xu et al., 2009]
  The last $n_c - p - 1$ eigenvalues are equal to λ_{cp}. Complexity: $\mathcal{O}(n_c^3)$.
  Similar to $p\mathcal{GP}_1$.

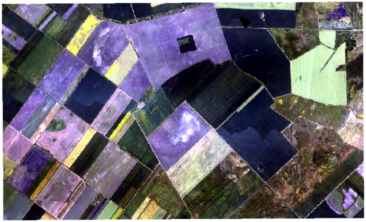
Existing models

- [Dundar and Landgrebe, 2004]
  Equal covariance matrix assumption and ridge regularization. Complexity: $\mathcal{O}(n^3)$. Similar to $p\mathcal{GP}_4$ with equal eigenvectors.
- [Pekalska and Haasdonk, 2009]
  Ridge regularization, per class. Complexity: $\mathcal{O}(n_c^3)$.
- [Xu et al., 2009]
  The last $n_c - p - 1$ eigenvalues are equal to λ_{cp}. Complexity: $\mathcal{O}(n_c^3)$. Similar to $p\mathcal{GP}_1$.



Data sets

- *University of Pavia*: 103 spectral bands, 9 classes and 42,776 referenced pixels.
- *Kennedy Space Center*: 224 spectral bands, 13 classes and 4,561 referenced pixels.
- *Heves*: 252 spectral bands, 16 classes and 360,953 pixels.



Protocol

- [Fauvel et al., 2015]
- 50 training pixels for each class have been randomly selected from the samples.
- The remaining set of pixels has been used for validation to compute the correct classification rate.
- Repeated 20 times.
- Variables have been scaled between 0 and 1.
- Competitive methods
  - SVM
  - RF
  - Kernel-DA (M. Dundar and D. A. Landgrebe, 2004)
- Hyperparameters learn by 5-CV.

Classification accuracy

	Kappa coefficient			Processing time (s)		
	University	KSC	Heves	University	KSC	Heves
$p\mathcal{GP}_0$	0.768	0.920	0.664	18	31	148
$p\mathcal{GP}_1$	**0.793**	**0.922**	**0.671**	18	33	151
$p\mathcal{GP}_2$	0.617	0.844	0.588	18	31	148
$p\mathcal{GP}_3$	0.603	0.842	0.594	19	33	152
$p\mathcal{GP}_4$	0.661	0.870	0.595	19	34	152
$p\mathcal{GP}_5$	0.567	0.820	0.582	18	32	148
$p\mathcal{GP}_6$	0.610	0.845	0.583	19	34	152
$np\mathcal{GP}_0$	0.730	0.911	0.640	17	31	148
$np\mathcal{GP}_1$	**0.792**	0.921	**0.677**	18	33	151
$np\mathcal{GP}_2$	0.599	0.838	0.573	18	31	148
$np\mathcal{GP}_3$	0.578	0.817	0.585	19	33	152
$np\mathcal{GP}_4$	0.578	0.817	0.585	19	33	152
KDC	0.786	**0.924**	0.666	98	253	695
RF	0.646	0.853	0.585	**3**	**3**	**18**
SVM	**0.799**	**0.928**	0.658	10	28	171

$p\mathcal{GP}$MRF

- Family of parsimonious Gaussian process models.
- Good performances wrt SVM and KDA
- Faster computation than previous KDA.
- $(n)p\mathcal{GP}_1$ perform the best.
- MRF extension.
- `https://github.com/mfauvel/PGPDA`
- Extension:
  - Non numerical data
  - Binary data
  - Unsupervised learning

References I

[Bouveyron et al., 2014] Bouveyron, C., Fauvel, M., and Girard, S. (2014).
Kernel discriminant analysis and clustering with parsimonious gaussian process models.
Statistics and Computing, pages 1–20.

[Camps-Valls and Bruzzone, 2009] Camps-Valls, G. and Bruzzone, L., editors (2009).
Kernel Methods for Remote Sensing Data Analysis.
Wiley.

[Dundar and Landgrebe, 2004] Dundar, M. and Landgrebe, D. A. (2004).
Toward an optimal supervised classifier for the analysis of hyperspectral data.
IEEE Trans. Geoscience and Remote Sensing, 42(1):271–277.

[Farag et al., 2005] Farag, A., Mohamed, R., and El-Baz, A. (2005).
A unified framework for map estimation in remote sensing image segmentation.
IEEE Trans. on Geoscience and Remote Sensing, 43(7):1617–1634.

[Fauvel et al., 2015] Fauvel, M., Bouveyron, C., and Girard, S. (2015).
Parsimonious gaussian process models for the classification of hyperspectral remote sensing images.
Geoscience and Remote Sensing Letters, IEEE, 12(12):2423–2427.

References II

[Fauvel et al., 2013] Fauvel, M., Tarabalka, Y., Benediktsson, J. A., Chanussot, J., and Tilton, J. (2013).
Advances in Spectral-Spatial Classification of Hyperspectral Images.
Proceedings of the IEEE, 101(3):652–675.

[Jimenez and Landgrebe, 1998] Jimenez, L. and Landgrebe, D. (1998).
Supervised classification in high-dimensional space: geometrical, statistical, and asymptotical properties of multivariate data.
Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on, 28(1):39 –54.

[Moser and Serpico, 2013] Moser, G. and Serpico, S. (2013).
Combining support vector machines and markov random fields in an integrated framework for contextual image classification.
IEEE Trans. on Geoscience and Remote Sensing, 51(5):2734–2752.

[Pekalska and Haasdonk, 2009] Pekalska, E. and Haasdonk, B. (2009).
Kernel discriminant analysis for positive definite and indefinite kernels.
IEEE Trans. Pattern Anal. Mach. Intell., 31(6):1017–1032.

[Solberg et al., 1996] Solberg, A., Taxt, T., and Jain, A. (1996).
A markov random field model for classification of multisource satellite imagery.
Geoscience and Remote Sensing, IEEE Transactions on, 34(1):100–113.

References III

[Tarabalka et al., 2010] Tarabalka, Y., Fauvel, M., Chanussot, J., and Benediktsson, J. (2010).
Svm- and mrf-based method for accurate classification of hyperspectral images.
IEEE Geoscience and Remote Sensing Letters, 7(4):736–740.

[Xu et al., 2009] Xu, Z., Huang, K., Zhu, J., King, I., and Lyu, M. R. (2009).
A novel kernel-based maximum a posteriori classification method.
Neural Netw., 22(7):977–987.