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

STATLEARN 2016

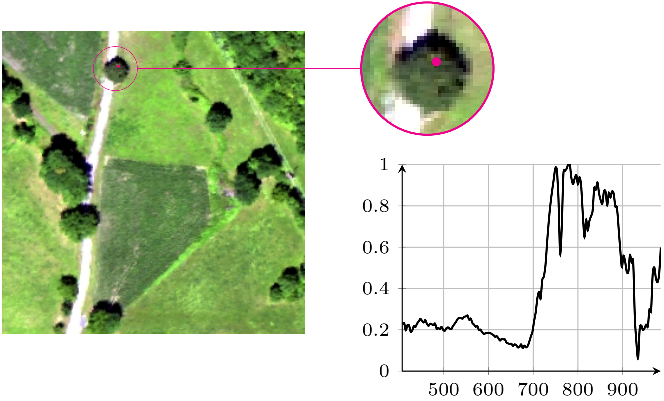
M. Fauvel [1], C. Bouveyron [2] and S. Girard [3]

[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

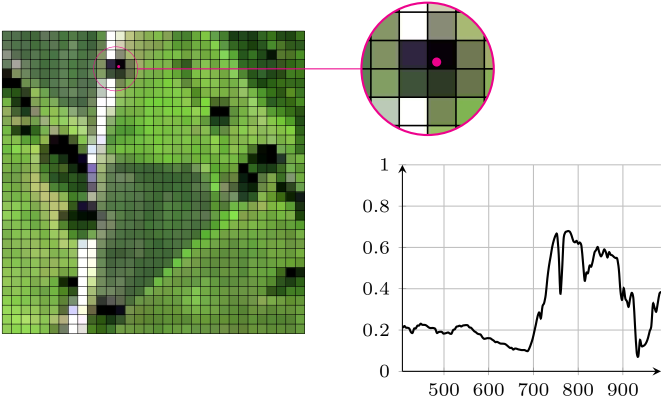
Nature of remote sensing images

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



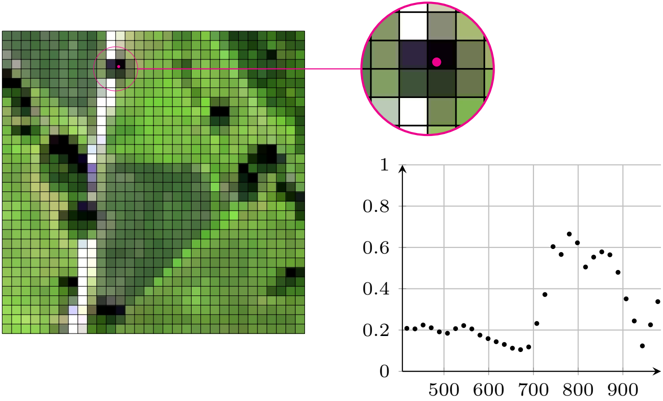
Nature of remote sensing images

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



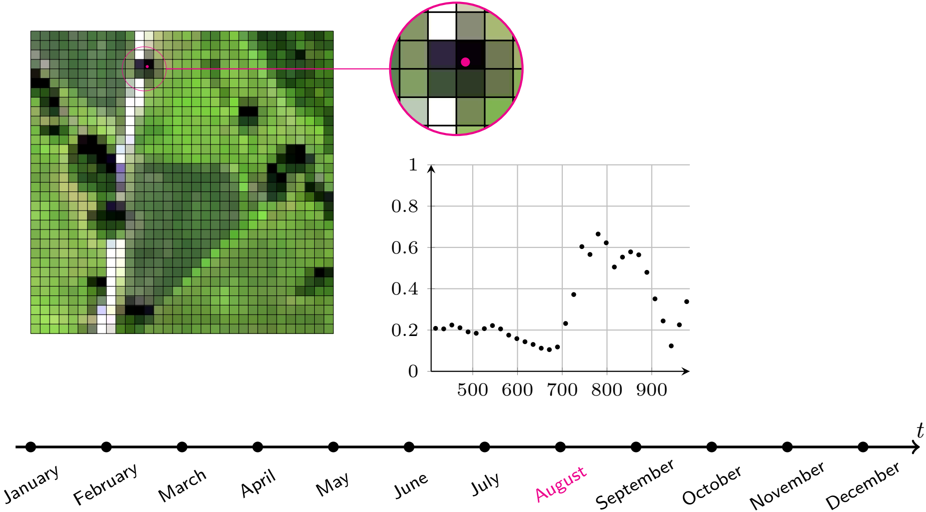
Nature of remote sensing images

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



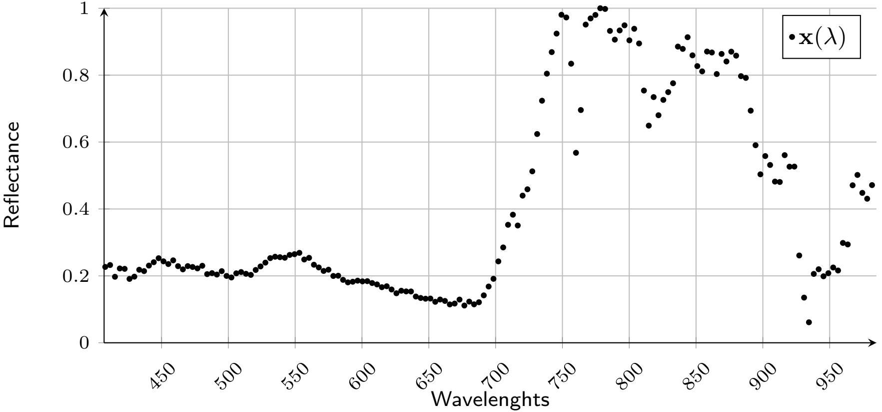
Nature of remote sensing images

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



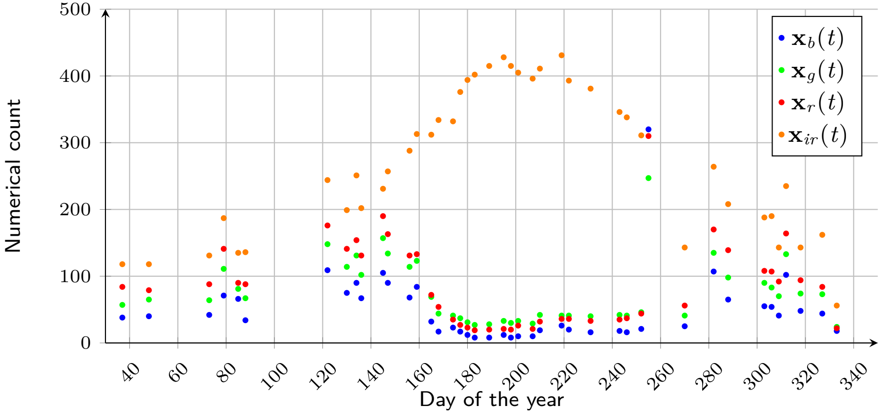
Hyperspectral

High number of spectral measurements and one temporal measurement: $\mathbf{x} \in \mathbb{R}^{180}$



Hypertemporal Image

High number of temporal measurements for few spectral measurements: $\mathbf{x} \in \mathbb{R}^{4\times 46}$



Some numbers

- Hyperspectral sensors

Instrument	Range (nm)	# Bands	Spatial resolution (m)
AVIRIS	400-2500	224	1-4
ROSIS-03	400-900	115	1
Hyspec	400-2500	427	1
HyMAP	400-2500	126	5
CASI	380-1050	288	1-2

- Hypertemporal missions

Mission	Revisit time (days)	# Bands	Spatial resolution (m)
Sentinel-2	5	13	10-20-60
MODIS	1	7	500
Landsat-8	16	8	30

Thematic applications

- Land cover/use at the national scale


- Monitor ecosystems & biodiversity
- Disaster management

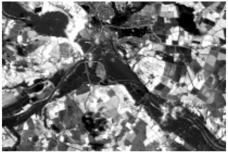
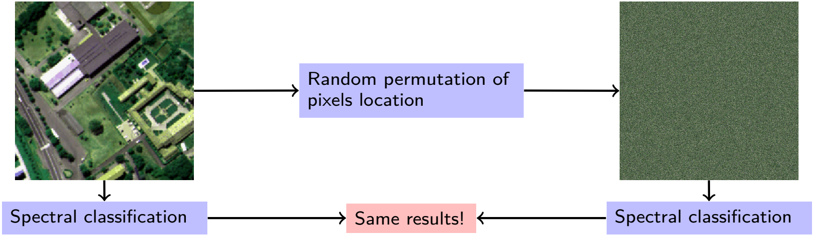



Image classification in high dimensional space

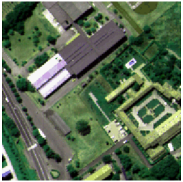
- High number of measurements but limited number of training samples: $d \approx n$.
- Curse of dimensionality: Statistical, geometrical and computational issues. Conventional methods failed [Jimenez and Landgrebe, 1998].
- Pixelwise classification is not adapted [Fauvel et al., 2013]: invariant to pixels location, sensible to *mixels*!



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

Image classification in high dimensional space

- High number of measurements but limited number of training samples: $d \approx n$.
- Curse of dimensionality: Statistical, geometrical and computational issues. Conventional methods failed [Jimenez and Landgrebe, 1998].
- Pixelwise classification is not adapted [Fauvel et al., 2013]: invariant to pixels location, sensible to *mixels*!



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

Spatial-spectral classification in remote sensing

Extract spatial features

- Using some image processing filters
  - Morphological Profile, adaptive neighborhood [Fauvel et al., 2013]
  - Local statistical moments [Camps-Valls et al., 2006]
  - Wavelets [Mercier and Girard-Ardhuin, 2006]
  - Texture, Gabor ...
- Then combine the spectral and the spatial information
  - Stack vector of all variables or few extracted variables (PCA ...) & kernel classifier
  - Fusion of separate classifiers ...
  - Composite kernels [Camps-Valls et al., 2006]

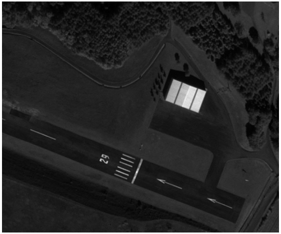
✓ Easy to implement.

✓ Easy to plug different spatial features.

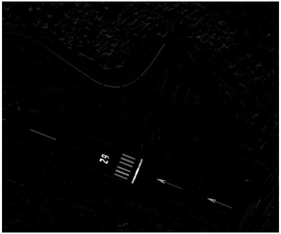
× Additional variables: statistical issues & computational issues.

× Most of the image processing tools are defined for $\mathbb{R}^1$-valued pixel.

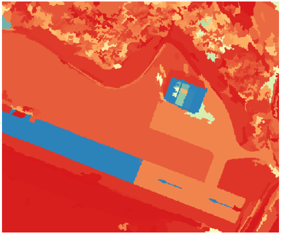
Spatial-spectral classification in remote sensing



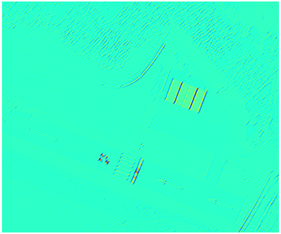
Original image



Top-hat



Local Stand. Deviat.

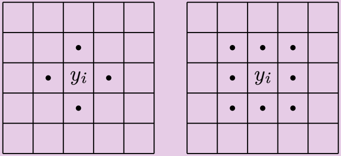


Gabor features

Spatial-spectral classification in remote sensing

Model the spatial dependencies

- Markov Random Field: local neighborhood is used in the decision $p(y_i = c|\mathbf{x}_i, \mathcal{N}_i)$



First-order   Second order

  - ✓ Allow for a fine modelization of inter-pixel dependencies.
  - × Estimation of the spectral energy term (through GMM).

- Mean kernel [Gurram and Kwon, 2013] $K(\mathbf{x}_i, \mathbf{x}_j) = \gamma^{-2} \sum_{m\in\mathcal{N}_i, n\in\mathcal{N}_j} k(\mathbf{x}_m, \mathbf{x}_n)$

  - ✓ Good performances in terms of classification accuracy.
  - × Rough modelization of the inter-pixel dependencies.
  - × Computing time.

SVM & MRF

- Maximum *a posteriori*: $\max_Y P(Y|\mathbf{X})$
- 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: $\mathcal{E}(y_i, \mathcal{N}_i) = \sum_{j \in \mathcal{N}_i} [1 - \delta(y_i, y_j)]$

SVM & MRF

- Maximum *a posteriori*: $\max_Y P(Y|\mathbf{X})$
- 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: $\mathcal{E}(y_i, \mathcal{N}_i) = \sum_{j \in \mathcal{N}_i} [1 - \delta(y_i, y_j)]$

SVM & MRF

- Maximum *a posteriori*: $\max_Y P(Y|\mathbf{X})$
- 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: $\mathcal{E}(y_i, \mathcal{N}_i) = \sum_{j \in \mathcal{N}_i} [1 - \delta(y_i, y_j)]$

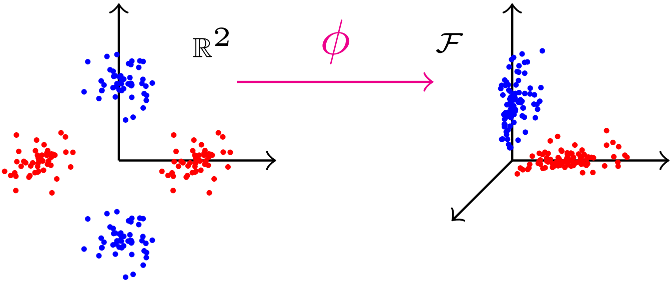
SVM & MRF

- Maximum *a posteriori*: $\max_Y P(Y|\mathbf{X})$
- 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

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

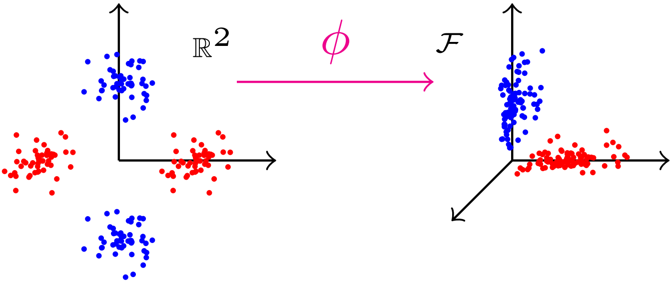
- Spectral term: $-\log[p(\mathbf{x}_i|y_i)]$
  - SVM outputs [Farag et al., 2005, Tarabalka et al., 2010, Moser and Serpico, 2013]
  - Kernel-probabilistic model [Dundar and Landgrebe, 2004]
- Spatial term
  - Potts model: $\mathcal{E}(y_i, \mathcal{N}_i) = \sum_{j \in \mathcal{N}_i} [1 - \delta(y_i,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

Kernel induced feature space



- From Mercer theorem: $k(\mathbf{x}_i, \mathbf{x}_j) = \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle_{\mathcal{F}}$
- Gaussian kernel: $k(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|^2_{\mathbb{R}^d}\right)$ with $d_{\mathcal{F}} = +\infty$
- Build a probabilistic model in $\mathcal{F}$ is not directly possible:

Kernel induced feature space



- From Mercer theorem: $k(\mathbf{x}_i, \mathbf{x}_j) = \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle_{\mathcal{F}}$
- Gaussian kernel: $k(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|^2_{\mathbb{R}^d}\right)$ with $d_{\mathcal{F}} = +\infty$
- Build a probabilistic model in $\mathcal{F}$ is not directly possible: Work with subspace models

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 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 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: $r_c = \min(n_c, r)$

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

- We propose to enforce parsimony in the model to make them computable.

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 $c = 1, \ldots, C$.

- A1 is motivated by the quick decay of the eigenvalues of Gaussian kernels [Braun et al., 2008].
- A2 expresses that the data of each class lives in a specific subspace of size p_c.
- Refers in the following to as $p\mathcal{GP}_0$

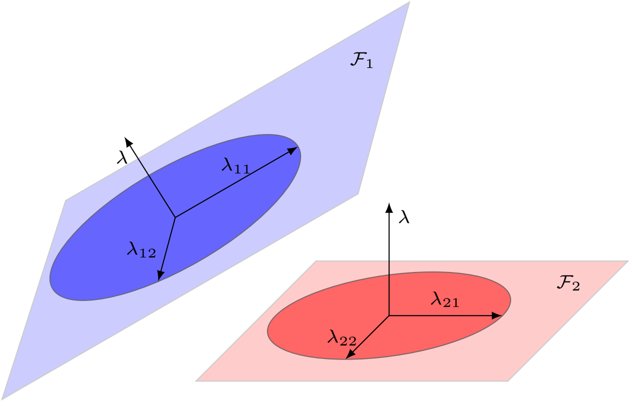


Figure: Visual illustration of parsimonious model. Dimension of $\mathcal{F}_c$ is common to both classes, they have specific variance inside $\mathcal{F}_c$ and they have common noise level.

$p\mathcal{GP}$ 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

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.

- $\boldsymbol{\mu}_c$ is the mean function of $\mathcal{GP}$ conditionally to $y = c$.
- $(\lambda_{cj},\mathbf{q}_{cj})$ are the first eigenvalues/eigenfunctions of the covariance function of $\mathcal{GP}$ conditionally to $y = c$.
- p_c is the size of $\mathcal{F}_c$

Proofs are given in [Bouveyron et al., 2014].

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 β_{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:

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

- Use of the property that the eigenfunction of the covariance function is a linear combination of $\phi(\mathbf{x}_i)-\boldsymbol{\mu}_c$
- $\langle \phi(\mathbf{x}_i)-\boldsymbol{\mu}_c, \phi(\mathbf{x}_j)-\boldsymbol{\mu}_c\rangle = \bar{k}_c(\mathbf{x}_i,\mathbf{x}_j)$

Estimation of the hyperparameters

- The parsimonious Gaussian process model and two hyperparameters have to be fitted:
  - $(n)p\mathcal{GP}_{0\ldots6}$
  - The kernel parameter γ
  - The size of signal subspace p_c
- Done by k-fold cross validation.
- Given one kernel hyperparameter value, all the others p_cs values can be tested at not cost since model parameters are derived from the eigendecomposition of $\mathbf{K}_c$.

Existing models

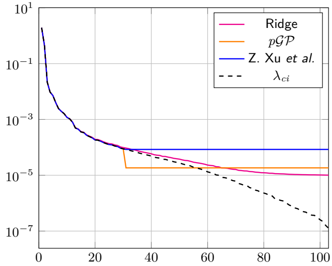
- [Dundar and Landgrebe, 2004]
  Equal covariance matrix assumption and ridge regularization. Complexity: $\mathcal{O}(n^3)$.
  Similar to $p\mathcal{GP}_4$ with equal eigenvectors.
- [Pekalska and Haasdonk, 2009]
  Ridge regularization, per class. Complexity: $\mathcal{O}(n_c^3)$.
- [Xu et al., 2009]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

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 $np\mathcal{GP}_1$.

✗ They all use eigenvectors associated to (very) small eigenvalues.

✓ $p\mathcal{GP}$ models use only eigenvectors associated to the largest eigenvalues.

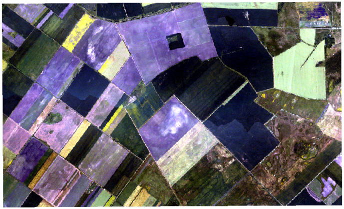


Protocol

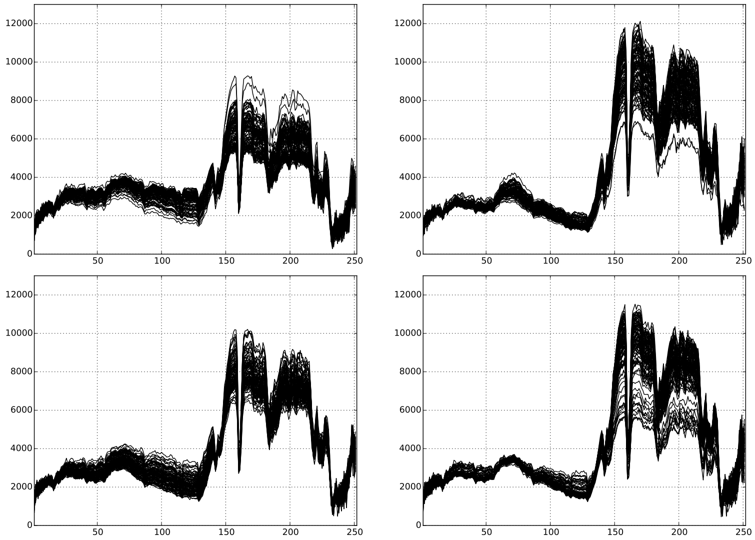
- [Fauvel et al., 2015]
- 50 training pixels for each class have been randomly selected from the samples.
- The remaining set of pixels has been used for validation to compute the correct classification rate.
- Repeated 20 times.
- Variables have been scaled between 0 and 1.
- Competitive methods
  -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

data sets

- *university of pavia*: uint16, 103 spectral bands, 9 classes and 42,776 referenced pixels.
- *kennedy space center*: uint16, 224 spectral bands, 13 classes and 4,561 referenced pixels.
- *heves*: uint16, 252 spectral bands, 16 classes and 360,953 pixels.



Samples



Classification accuracy and processing time

	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

Data set

Formosat-2 SITS: uint8, 4 spectral bands, 43 dates for 2006 and 13 woody classes.

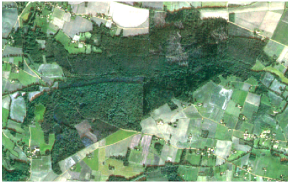
t_1

t_2

t_3

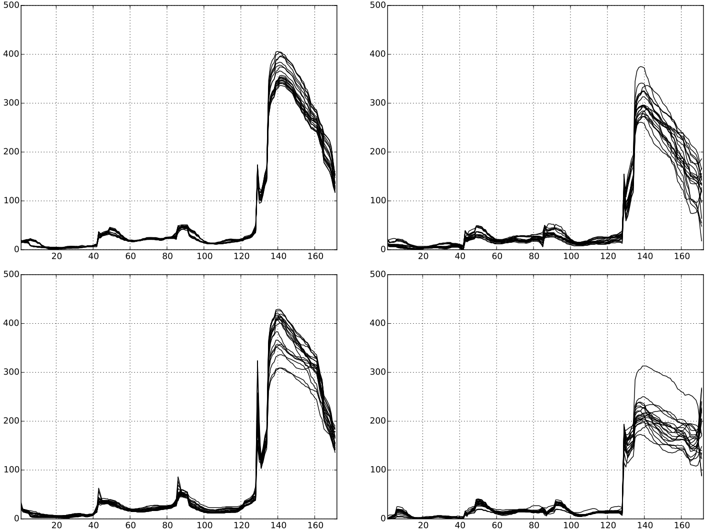
t_4

t_5



t_6

Samples

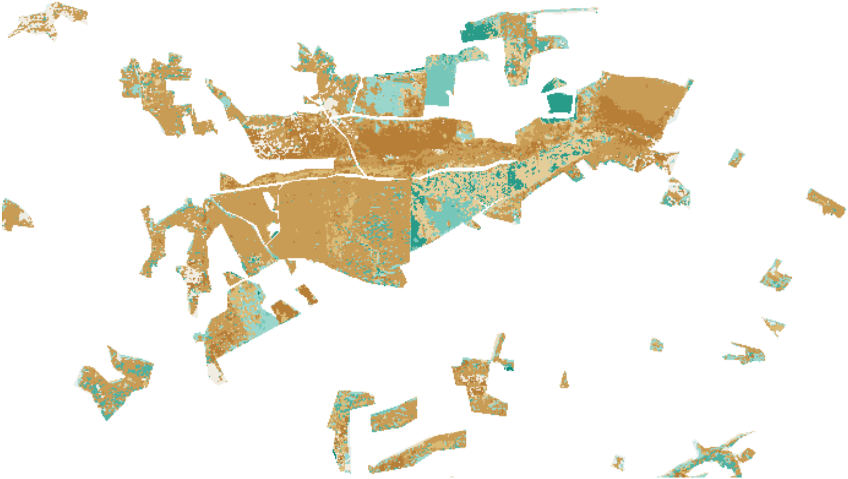


Classification accuracy and processing time

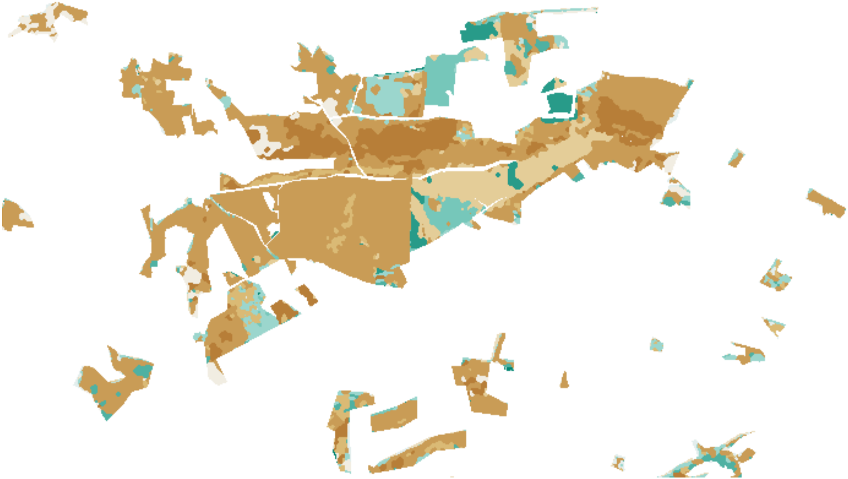
	Kappa Coefficient	Processing time (s)
$p\mathcal{GP}_0$	**0.950**	13
$p\mathcal{GP}_1$	**0.955**	9
$p\mathcal{GP}_2$	0.887	13
$p\mathcal{GP}_3$	0.887	9
$p\mathcal{GP}_4$	0.932	9
$p\mathcal{GP}_5$	0.846	13
$p\mathcal{GP}_6$	0.891	8
$np\mathcal{GP}_0$	0.941	12
$np\mathcal{GP}_1$	0.943	8
$np\mathcal{GP}_2$	0.883	13
$np\mathcal{GP}_3$	0.871	9
$np\mathcal{GP}_4$	0.871	9
KDA	0.942	69
RF	0.896	1
SVM	0.944	10

$p\mathcal{GP}$MRF

$p\mathcal{GP}$MRF



$p\mathcal{GP}$MRF



- Family of parsimonious Gaussian process models has been presented.
- Good performances w.r.t 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 [Bouveyron et al., 2014]
  - Binary data [Sylla et al., 2015]
  - Unsupervised learning [Bouveyron et al., 2014]

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.

[Braun et al., 2008] Braun, M. L., Buhmann, J. M., and Muller, K.-R. (2008).
On Relevant Dimensions in Kernel Feature Spaces.
Journal of Machine Learning Research, 9:1875–1908.

[Camps-Valls et al., 2006] Camps-Valls, G., Gomez-Chova, L., Muñoz-Marí, J., Vila-Francés, J., and Calpe-Maravilla, J. (2006).
Composite kernels for hyperspectral image classification.
Geoscience and Remote Sensing Letters, IEEE, 3(1):93–97.

[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.

References II

[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.

[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.

[Gurram and Kwon, 2013] Gurram, P. and Kwon, H. (2013).
Contextual svm using hilbert space embedding for hyperspectral classification.
Geoscience and Remote Sensing Letters, IEEE, 10(5):1031–1035.

[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.

[Mercier and Girard-Ardhuin, 2006] Mercier, G. and Girard-Ardhuin, F. (2006).
Partially supervised oil-slick detection by sar imagery using kernel expansion.
Geoscience and Remote Sensing, IEEE Transactions on, 44(10):2839–2846.

References III

[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.

[Sylla et al., 2015] Sylla, S. N., Girard, S., Diongue, A. K., Diallo, A., and Sokhna, C. (2015).
A classification method for binary predictors combining similarity measures and mixture models.
Dependence Modeling, 3:240–255.

[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.