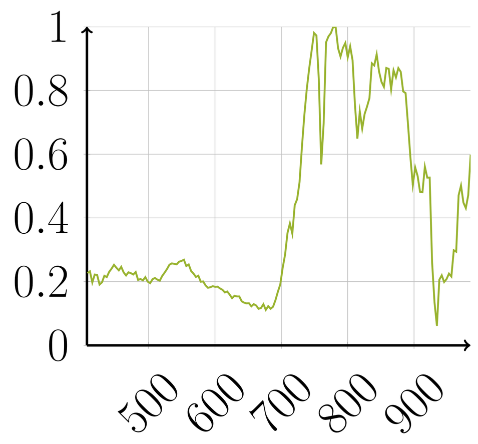
Parsimonious Gaussian process models for the classification of multivariate remote sensing images

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

Context

Classification of multispectral imagery



- A pixel is represented by $\mathbf{x} \in \mathbb{R}^d$.
- Spatial-spectral classification.
- Kernel methods: including spatial information in the classification process is not easy.
- Statistical methods: MRF, but hard when d is large

Combine statistical methods (GMM-MRF) and kernel methods.

Classification with parsimonious Gaussian process models

Gaussian process in the kernel feature space

Let $\mathcal{S} = \big\{(\mathbf{x}_i, y_i)\big\}_{i=1}^n$ be a set of training samples, where $\mathbf{x}_i \in J$, $J \subset \mathbb{R}^d$, is a pixel and $y_i \in \{1, \ldots, C\}$ its class, and C the number of classes. For short, in the following $-\ln\Big(p\big(\phi(\mathbf{x}_i)|y_i\big)\Big)$ will be referred to $\Omega(\phi(\mathbf{x}_i), y_i)$.
In this work, the conventional Gaussian kernel function is used:

$$k(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2_{\mathbb{R}^d}}{2\sigma^2}\right), \quad \sigma > 0, \tag{1}$$

Its associated feature space is $\mathcal{F}$ and the mapping function is $\phi : \mathbb{R}^d \to \mathcal{F}$. We have $d_{\mathcal{F}} = +\infty$ and the conventional multivariate normal distribution cannot be defined.
To overcome this, let us assume that $\phi(\mathbf{x})$, conditionally on $y = c$, is a Gaussian process with mean $\boldsymbol{\mu}_c$ and covariance function $\boldsymbol{\Sigma}_c$. Hence, for all $r \geq 1$, random vectors on $\mathbb{R}^r$ defined by $[\phi(\mathbf{x})_1, \ldots, \phi(\mathbf{x})_r]$ are, conditionally on $y = c$, a multivariate normal vectors. Therefore, it is possible to write for $y_i = c$

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

where λ_{cj} is the j^{th} eigenvalue of $\boldsymbol{\Sigma}_c$ in decreasing order, $\mathbf{q}_{cj}$ its associated eigenvector and γ a constant term that does not depend on c.

Parsimonious Gaussian process

Definition 1 (Parsimonious Gaussian process) *A parsimonious Gaussian process is a Gaussian process $\phi(\mathbf{x})$ for which, conditionally to $y = c$, the eigen-decomposition of its covariance operator $\boldsymbol{\Sigma}_c$ is such that:*

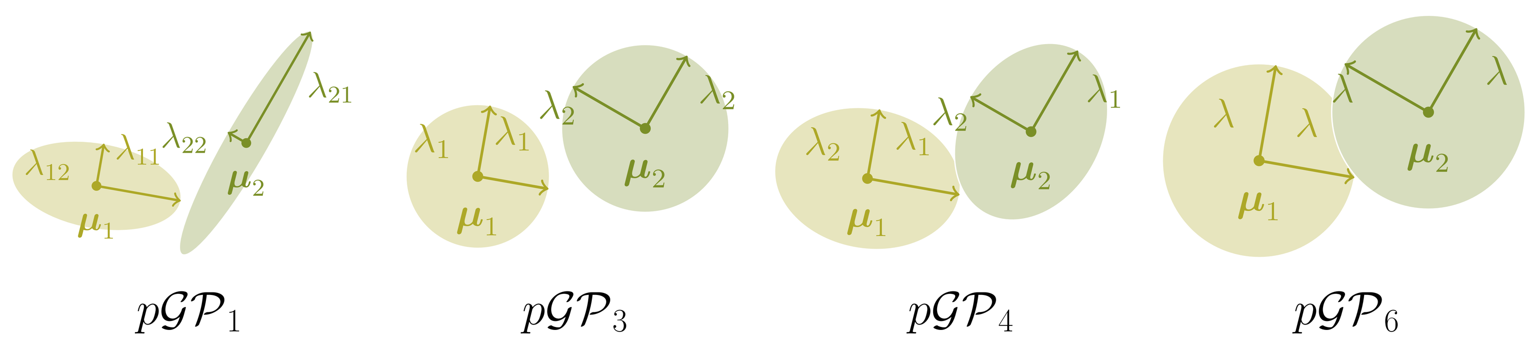
A1. It exists a dimension $r < +\infty$ such that $\lambda_{cj} = 0$ for $j \geq r$ and for all $c = 1, \ldots, C$.

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 for a Gaussian kernel. *A2* expresses that the data of each class live in a specific subspace $\mathcal{F}_c$ of size p_c.

Sub-models

Model	Variance inside $\mathcal{F}_c$	$\mathbf{q}_{cj}$	p_c
$p\mathcal{GP}_0$	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



Model inference

Centered Gaussian kernel function according to class c as:

$$\bar{k}_c(\mathbf{x}_i, \mathbf{x}_j) = k(\mathbf{x}_i, \mathbf{x}_j) + \frac{1}{n_c^2}\sum_{\substack{l,l'=1\\ y_l,y_{l'}=c}}^{n_c} k(\mathbf{x}_l, \mathbf{x}_{l'}) - \frac{1}{n_c}\sum_{\substack{l=1\\ y_l=c}}^{n_c}\big(k(\mathbf{x}_i, \mathbf{x}_l) + k(\mathbf{x}_j, \mathbf{x}_l)\big). \tag{3}$$

The associated normalized kernel matrix $\overline{\mathbf{K}}_c$ of size $n_c \times n_c$ is defined by

$$(\overline{\mathbf{K}}_c)_{l,l'} = \frac{\bar{k}_c(\mathbf{x}_l, \mathbf{x}_{l'})}{n_c}. \tag{4}$$

Proposition 1 *For $p_M = \max(p_1, \ldots, p_C)$, $c = 1, \ldots, C$ and the model $p\mathcal{GP}_0$, eq. (2) can be computed with*

$$\Omega(\phi(\mathbf{x}_i), y_i) = \frac{1}{2n_c}\sum_{j=1}^{\hat{p}_c}\frac{1}{\hat{\lambda}_{cj}}\left(\frac{1}{\hat{\lambda}_{cj}} - \frac{1}{\hat{\lambda}}\right)\Bigg(\sum_{\substack{l=1\\ y_l=c}}^{n_c}\beta_{cjl}\bar{k}_c(\mathbf{x}_i, \mathbf{x}_l)\Bigg)^2 + \frac{1}{2\hat{\lambda}}\bar{k}_c(\mathbf{x}_i, \mathbf{x}_i) + \sum_{j=1}^{\hat{p}_c}\frac{\ln(\hat{\lambda}_{cj})}{2} + (\hat{p}_M - \hat{p}_c)\frac{\ln(\hat{\lambda})}{2} \tag{5}$$

where β_{cjl} is the l^{th} component of the normalized eigenvector $\boldsymbol{\beta}_{cj}$ associated to j^{th} largest eigenvalue $\hat{\lambda}_{cj}$ of $\overline{\mathbf{K}}_c$ and

$$\hat{\lambda} = \frac{1}{\sum_{c=1}^{C}\hat{\pi}_c(r_c - \hat{p}_c)}\sum_{c=1}^{C}\hat{\pi}\Big(\text{trace}(\overline{\mathbf{K}}_c) - \sum_{j=1}^{\hat{p}_c}\hat{\lambda}_{cj}\Big) \tag{6}$$

and $\hat{\pi}_c = n_c/n$.

The estimation of p_c is done by looking at the cumulative variance for the sub-models $p\mathcal{GP}_{0,2,5}$. In practice, p_c is estimated such as the percentage of the cumulative variance is higher than a given threshold t_h:

$$\frac{\sum_{j=1}^{\hat{p}_c}\hat{\lambda}_{cj}}{\sum_{j=1}^{n_c}\hat{\lambda}_{cj}} > t_h. \tag{7}$$

For the other sub-models, $\hat{p}$ is a fixed parameter given by the user.

Experimental results

The data set is the *University Area* of Pavia, Italy, acquired with the ROSIS-03 sensor. The image has 103 spectral bands ($d = 103$) and is 610×340 pixels. 50 pixels for each class have been randomly selected from the samples for the training set, and the remaining set of pixels has been used for validation. The process has been repeated 50 times, each time a new training set has been generated and the variables have been scaled between -1 and 1.

Method	$p\mathcal{GP}_0$	$p\mathcal{GP}_1$	$p\mathcal{GP}_2$	$p\mathcal{GP}_3$	$p\mathcal{GP}_4$	$p\mathcal{GP}_5$	$p\mathcal{GP}_6$	SVM	GMM	KGMM	$p\mathcal{GP}$MRF
OA	**83.5**	**84.2**	62.7	69.6	73.4	61.1	69.9	**84.5**	77.7	80.4	91.2

A conventionnal Potts model is used to construct a Markov Random Field (MRF) for which the conditional probability is computed with $p\mathcal{GP}_1$. For the optimization, a Metropolis algorithm is used.

Part of Pavia image — Thematic map obtained with $p\mathcal{GP}_1$ — Thematic map obtained with $p\mathcal{GP}_1$MRF

Conclusions and perspectives

Conclusions:
- Family Kernel GMMs has been proposed
- Good classification accuracies w.r.t SVM
- Extension to MRF classifier

Perspectives
- Influence of the training set size
- Combination of kernel
- Advanced MRF models