Parsimonious Gaussian Process Models for the Classification of Multivariate Remote Sensing Images
Mathieu Fauvel, Charles Bouveyron, Stephane Girard

To cite this version:

HAL Id: hal-01062378
https://hal.archives-ouvertes.fr/hal-01062378
Submitted on 9 Sep 2014

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
ABSTRACT
A family of parsimonious Gaussian process models is presented. They allow to construct a Gaussian mixture model in a kernel feature space by assuming that the data of each class live in a specific subspace. The proposed models are used to build a kernel Markov random field (pGPMRF), which is applied to classify the pixels of a real multivariate remotely sensed image. In terms of classification accuracy, some of the proposed models perform equivalently to a SVM but they perform better than another kernel Gaussian mixture model previously defined in the literature. The pGPMRF provides the best classification accuracy thanks to the spatial regularization.

Index Terms— Kernel, remote sensing images, Gaussian process, parsimony, hyperspectral.

1. INTRODUCTION
In a multivariate remote sensing images, a pixel is represented by a vector $\mathbf{x} \in \mathbb{R}^d$ for which each component is a measurement corresponding to specific wavelengths [1]. The classification of such images requires algorithms that are robust to the number $d$ of spectral wavelengths and that are able to include additional spatial information in the classification process [2].

Kernel methods, such as SVM, have shown good abilities in classifying images with a large number of spectral bands [3]. The use of a kernel function that defines a measure of similarity between two samples, here two pixel-vectors, make them robust to the spectral dimension. However, including spatial information in the classification process is not easy, and the resulting algorithms usually involve a separate step for the extraction/inclusion of the spatial information [2].

On the contrary, conventional statistical method such as Markov random fields (MRF) model the spatial relationship between adjacent pixels in a proper way by using a local energy function

$$U(y_i|x_i,N_i) = -\ln (p(x_i|y_i)) + \rho \mathcal{E}(N_i)$$

where $y_i$ is the label, $p(x_i|y_i)$ the conditional probability of having $x_i$ given $y_i$, $\mathcal{E}$ is an energy term that characterizes the local context of the pixel, $N_i$ represents the neighborhood of $x_i$ in the spatial domain and $\rho$ is a positive parameter. The statistical modeling in the spectral domain usually suffers from the increase of the dimension. With the conventional Gaussian assumption, $p(x_i|y_i = c) \sim \mathcal{N}(\mu_c, \Sigma_c)$, the number of parameters of the model scales with the square of the number of spectral variables, which makes difficult reliable estimations.

Several approaches have been proposed to combine MRF and kernel methods. The main problem is to compute properly the conditional probability in the MRF energy function. Indeed with kernel functions, the samples $x$ are implicitly mapped to $\phi(x)$ that live on a feature space. For the commonly used Gaussian kernel function, the dimension of the feature space is infinite, and probability functions cannot be defined. Several strategies were proposed to overcome this difficulty. In [4], it was proposed to use SVM to estimate the conditional probability $p(\phi(x_i)|y_i)$ in the MRF energy function. However, the estimated probability is not a true probability, but a scaled version of the SVM output [5]. In [6], the authors defined theoretically the conditional probability in the kernel feature space, but they assumed that the covariance matrices were common to the different classes in the feature space. A similar assumption regarding the covariance matrix was done in [7]. Although the performances of the above mentioned method were good, such assumption about the covariance matrix could limit the effectiveness of the methods.

In this paper, a family of parsimonious Gaussian process models is proposed to construct $p(\phi(x_i)|y_i)$ in eq. (1). These models allow to build from a finite set of training samples, a Gaussian mixture model in the kernel feature space, even in the infinite dimensional case. They assume that the data of each class live in a specific subspace of the kernel feature space [8].

The remainder of the paper is organized as follows. Section 2 presents the family of parsimonious Gaussian process models. Section 3 focuses on the experimental results on one real hyperspectral images. Finally, conclusions and perspectives are discussed in Section 4.

2. CLASSIFICATION WITH PARSIMONIOUS GAUSSIAN PROCESS MODELS
2.1. Gaussian process in the kernel feature space

Let $\mathcal{S} = \{(x_i, y_i)\}_{i=1}^n$ be a set of training samples, where $x_i \in J$, \(J \subseteq \mathbb{R}^d\), is a pixel and $y_i \in \{1, \ldots, C\}$ its class, and $C$ the number of classes. For short, in the following $-\ln(p(\phi(x_i)|y_i))$ will be referred to $\Omega(\phi(x_i), y_i)$.

In this work, the conventional Gaussian kernel function

$$k(x_i, x_j) = \exp\left(\frac{-||x_i - x_j||^2}{2\sigma^2}\right), \quad \sigma > 0, \quad (2)$$

is used. Its associated feature space is $\mathcal{F}$ and the mapping function is $\phi : \mathbb{R}^d \rightarrow \mathcal{F}$. From the Mercer theorem, we have $k(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle_\mathcal{F}$ and the kernel evaluation can be written as the series converges absolutely and uniformly for almost all $(x_i, x_j)$ [9]:

$$k(x_i, x_j) = \sum_{m=1}^{d_F} c_m v_m(x_i) v_m(x_j)$$  

(3)
where $d_F = \dim(F)$, $\{e_m, m = 1, 2, \ldots\}$ is the sequence of positive eigenvalues in decreasing order of the integral operator $T_k$, $(T_k f)(z) = \int f(z, x) f(x) \, dx$, associated to $k$ and $\{v_m : \mathbb{R}^d \rightarrow \mathbb{R}, m = 1, 2, \ldots\}$ is the sequence of corresponding normalized eigenfunctions. From eq.(3), $\phi$ can be defined as

$$\phi : x \mapsto \sqrt{e_m} v_m(x), \ m = 1, 2, \ldots$$

(4)

For the Gaussian kernel $d_F = +\infty$ [9]. Therefore the conventional multivariate normal distribution cannot be defined.

To overcome this, let us assume that $\phi(x)$, conditionally on $y = c$, is a Gaussian process with mean $\mu_c$ and covariance function $\Sigma_c$. Hence, for all $r \geq 1$, random vectors on $\mathbb{R}^r$ defined by $[\phi(x)_1, \ldots, \phi(x)_r]$ are, conditionally on $y = c$, a multivariate normal vectors. Therefore, it is possible to write for $y_i = c$

$$\Omega(\phi(x)_i, y_i) = \sum_{j=1}^{r_c} \left[ \frac{(\phi(x)_i - \mu_c, q_{cj})^2}{2\lambda_{cj}} + \frac{\ln(\lambda_{cj})}{2} \right] + \gamma$$

(5)

where $\lambda_{cj}$ is the $j$-th eigenvalue of $\Sigma_c$ in decreasing order, $q_{cj}$ its associated eigenvector and $\gamma$ a constant term that does not depend on $c$. If the Gaussian process is not degenerated (i.e., $\lambda_{cj} \neq 0, \forall j$), $r$ has to be large to get a good approximation of the Gaussian process. Unfortunately, only a part of the above equation can be computed from a finite training sample set:

$$\Omega(\phi(x)_i, y_i) = \sum_{j=1}^{r_c} \left[ \frac{(\phi(x)_i - \mu_c, q_{cj})^2}{2\lambda_{cj}} + \frac{\ln(\lambda_{cj})}{2} \right]$$

$$+ \sum_{j=r_c+1}^{r} \left[ \frac{(\phi(x)_i - \mu_c, q_{cj})^2}{2\lambda_{cj}} + \frac{\ln(\lambda_{cj})}{2} \right]$$

(6)

where $r_c = \min(n_c, r)$ and $n_c$ is the number of training samples of class $c$.

2.2. Parsimonious Gaussian process

To make the above computational problem tractable, it is proposed to use a parsimonious Gaussian process model in the feature space for each class.

Definition 1 (Parsimonious Gaussian process) A parsimonious Gaussian process is a Gaussian process $\phi(x)$ for which, conditionally to $y$, the eigen-decomposition of its covariance operator $\Sigma_c$ 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 $c = 1, \ldots, C$.

The assumption A1 is motivated by the quick decay of the eigenvalues for a Gaussian kernel [10]. Hence, it is possible to find $r < +\infty$ such as $\lambda_r \approx 0$. The assumption A2 expresses that the data of each class live in a specific subspace of size $p_c$, the signal subspace, of the feature space. The variance in the signal subspace for the class $c$ is modeled by the parameters $\lambda_{c1}, \ldots, \lambda_{cp_c}$ and the variance in the noise subspace, common to all the classes, is modeled by $\lambda$. This model is referred to by $pGP_c$.

From this model, it is possible to derive several sub-models. Table 1 lists the different models that can be built from $pGP_c$. For the model $pGP_1$, it is additionally assumed that the data of each class share the same intrinsic dimension, i.e., $p_c = p, \forall c \in \{1, \ldots, C\}$. In the model $pGP_2$, the variance of $F_c$ is assumed to be equal for all eigenvectors, i.e., $\lambda_{cj} = \lambda_c, \forall j \in \{1, \ldots, p\}$. For the model $pGP_3$, it is assumed that the intrinsic dimension is common to every class and the variance is common between them, i.e., $\lambda_{cj} = \lambda_{cj'}, \forall j \in \{1, \ldots, p\}$ and $c, c' \in \{1, \ldots, C\}$. In term of parsimony, $pGP_0$ is the least parsimonious model while $pGP_8$ is the most parsimonious of the proposed models.

In the following, only the model $pGP_0$ is discussed. Similar results can be obtained for the other models.

Proposition 1 Letting $p_M = \max(p_1, \ldots, p_C)$, eq. (5) can be written for $pGP_0$ as

$$\Omega(\phi(x)_i, y_i) = \sum_{j=1}^{p_c} \left( \frac{1}{\lambda_{cj} - \lambda} (\phi(x)_i - \mu_c, q_{cj}) \right)^2$$

$$+ \frac{1}{2} ||\phi(x) - \mu_c||^2 + \sum_{j=1}^{p_c} \ln(\lambda_{cj})$$

(7)

where $\gamma'$ is a constant term that does not depend on the index $c$ of the class.

The computation of eq. (7) is now possible since $p_c = n_c, \forall c \in \{1, \ldots, C\}$. In the following, it is shown that the estimation of the parameters and the computation of eq. (7) can be done using only the kernel evaluation, as in standard kernel methods.

2.3. Model inference

Let us define the centered Gaussian kernel function according to class $c$ as:

$$\tilde{k}_c(x_i, x_j) = k(x_i, x_j) + \frac{1}{n_c} \sum_{i_l, i_l' = 1}^{n_c} k(x_i, x_{i_l}, x_{i_l'})$$

$$- \frac{1}{n_c} \sum_{l=1}^{n_c} \sum_{y_i = c} k(x_i, x_i) + k(x_j, x_{i_l})$$

(8)

The associated normalized kernel matrix $K_c$ of size $n_c \times n_c$ is defined by

$$[K_c]_{i, i'} = \frac{\tilde{k}_c(x_i, x_{i'})}{n_c}.$$
\[
\Omega(\phi(x_i), y_0) = \frac{1}{2n_c} \sum_{c=1}^{n_c} \frac{1}{\lambda_{c,j}} \left( \frac{1}{\hat{\lambda}_j} - \frac{1}{\lambda_{c,j}} \right) \left( \sum_{j=1}^{n_c} \beta_{c,j} \hat{c}(x_i, x_i) \right)^2 + \frac{1}{2\lambda} \hat{c}(x_i, x_i) + \sum_{j=1}^{n_c} \ln(\hat{\lambda}_j) + \frac{\hat{p}_M - \hat{p}_c}{2} \ln(\hat{\lambda})
\]

(10)

**Table 2.** Information classes for the *University Area* data set and classification accuracy. SVM refers to the support vectors machine, GMM refers to the Gaussian Mixture Model, KGMM refers to the kernel GMM proposed in [7] and OA refers to overall accuracy.

<table>
<thead>
<tr>
<th>Class</th>
<th>Samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asphalt</td>
<td>6631</td>
</tr>
<tr>
<td>Meadow</td>
<td>18649</td>
</tr>
<tr>
<td>Gravel</td>
<td>2099</td>
</tr>
<tr>
<td>Tree</td>
<td>3064</td>
</tr>
<tr>
<td>Metal Sheet</td>
<td>1345</td>
</tr>
<tr>
<td>Bare Soil</td>
<td>5029</td>
</tr>
<tr>
<td>Bitumen</td>
<td>1330</td>
</tr>
<tr>
<td>Brick</td>
<td>3682</td>
</tr>
<tr>
<td>Shadow</td>
<td>1947</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>42776</td>
</tr>
</tbody>
</table>

(a) Information classes

<table>
<thead>
<tr>
<th>Class</th>
<th>(b) Classification accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Method</td>
</tr>
<tr>
<td>Asphalt</td>
<td>p(\mathcal{GP}_0)</td>
</tr>
<tr>
<td>Meadow</td>
<td>(p\mathcal{GP}_1)</td>
</tr>
<tr>
<td>Gravel</td>
<td>(p\mathcal{GP}_2)</td>
</tr>
<tr>
<td>Tree</td>
<td>(p\mathcal{GP}_3)</td>
</tr>
<tr>
<td>Metal Sheet</td>
<td>(p\mathcal{GP}_4)</td>
</tr>
<tr>
<td>Bare Soil</td>
<td>(p\mathcal{GP}_5)</td>
</tr>
<tr>
<td>Bitumen</td>
<td>(p\mathcal{GP}_6)</td>
</tr>
<tr>
<td>Brick</td>
<td>SVM</td>
</tr>
<tr>
<td>Shadow</td>
<td>GMM</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>KGMM</td>
</tr>
</tbody>
</table>

\(p\mathcal{GP}\)MRF 91.2

**Table 3.** Grid search setting for the cross validation. \(t_h\) corresponds to the threshold value on the cumulative variance, \(C_p\) is the regularization term for the SVM and \(\lambda_c\) refers to the ridge regularization term in KGMM.

<table>
<thead>
<tr>
<th>Method</th>
<th>Parameter</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>(\sigma^2)</td>
<td>({2^{-3}, 2^{-2}, \ldots, 2^0})</td>
</tr>
<tr>
<td>SVM</td>
<td>(\lambda_c)</td>
<td>({0.900, 0.911, \ldots, 0.999})</td>
</tr>
<tr>
<td>SVM</td>
<td>(\lambda_c)</td>
<td>({5, 10, 45})</td>
</tr>
<tr>
<td>SVM</td>
<td>(C_p)</td>
<td>({10^{-2}, 10^{-4}})</td>
</tr>
<tr>
<td>KGMM</td>
<td>(\sigma^2)</td>
<td>({2^0, 2^1, \ldots, 2^8})</td>
</tr>
<tr>
<td>KGMM</td>
<td>(\tau)</td>
<td>({10^{-13}, 10^{-9}})</td>
</tr>
</tbody>
</table>

To assess the statistical significance of the observed differences in terms of classification accuracy, a Wilcoxon rank-sum test has been applied over the 50 repetitions. It tests if the data from two populations are samples from distribution with equal medians. In the experiments, it is used to test whether the 50 classification accuracy are significantly different (not equal medians) or not (equal medians).

Two sets of experiments have been conducted. First, the proposed models have been compared to three others models: the support vectors machines (SVM), a conventional Gaussian mixture model (GMM) and a Kernel GMM defined in [7]. These models use only the spectral information from the remote sensing image. Second, the model \(p\mathcal{GP}\) has been used to build a \(p\mathcal{GP}\)MRF classifier that uses both the spatial and the spectral information.

### 3.1. Comparison with others spectral classifier

In this section, the proposed models are compared to others classifiers in terms of classification accuracy. The Gaussian kernel was used for each kernel method. All the parameters have been selected by a five fold cross-validation. The ranges of the tested values are reported in Table 3. A small regularization term (\(\tau = 10^{-5}\)) has been added in the GMM for the inversion of the covariance matrix.

Results are reported in Table 2.(b). The three best results in terms of classification accuracy are obtained for \(p\mathcal{GP}_0\), \(p\mathcal{GP}_1\) and SVM, in boldface in the table. The differences between them are not significant according to the Wilcoxon test. All the other methods provide results that are significantly worst in terms of classification accuracy.

From the results, and for that experimental protocol (small training set and *University area* data set) the proposed models \(p\mathcal{GP}_0\) and \(p\mathcal{GP}_1\) provides classification accuracies similar to those obtained with SVM. They outperformed in terms of classification accuracy the KGMM proposed in [7] and the conventional GMM.

### 3.2. Classification with the \(p\mathcal{GP}\)MRF

In this section, the model \(p\mathcal{GP}_1\) is used to compute the conditional probability in eq.(1). For the spatial energy term in eq.(1), a conven-
tional Potts model is used [11]:

\[ \mathcal{E}(\mathcal{N}_i) = \sum_{x_j \in \mathcal{N}_i} \delta(y_i, y_j) \]  

(13)

where \( \delta \) is the delta function. A second order neighborhood is considered, i.e., \( \mathcal{N}_i \) is the set of 8 surrounding pixels of pixel \( x_i \). For the optimization, a Metropolis algorithm is used. It minimizes the global energy, \( U_G = \sum_i U(y_i | x_i, \mathcal{N}_i) \), by an iterative minimization of the local energy (1). For details, see [11]. \( \rho \) was set to 16.

The thematic maps obtained for one repetition is reported in the Figure 1 and the mean overall accuracy is reported in the Table 2. Using the MRF modeling leads to an mean improvement of 8.3% in terms of overall accuracy in comparison with the use of \( \rho GP \) alone. The Wilcoxon test shows that the improvement is significant. The Figure 1.(c) is much more homogeneous than Figure 1.(b), thanks to the spatial regularization operated by the \( \rho GP \)MRF.

4. CONCLUSIONS AND PERSPECTIVES

A family of parsimonious Gaussian process models was presented in this article. They make possible the computation of the Gaussian mixture model when the original samples are mapped into an infinite dimensional space, such as ones associated to the Gaussian kernel function. By assuming that the data of each class are located in a specific subspace of the kernel feature space, it was shown that all the computations can be expressed with kernel evaluation.

Experimental results exhibit, for a small size of training set and for the University area data set, that two models \( \rho GP_{0.1} \) provide similar performances with SVM in terms of classification accuracy, and outperformed another kernel GMM that were proposed in [7].

In order to take into account the spatial correlation in the image, a \( \rho GP \)MRF was build using the conditional probability provided by the model \( \rho GP \). The classification accuracy was increased by 8%. Hence the proposed models associated to the MRF model are appropriated for the classification of multivariate remote sensing image, in particular when the number of spectral bands is large.

Further analysis are required to better assess the performance of the proposed models. The effect of the size of the training set should be investigated. When \( n_c \) grows, the size of the associated feature spanned by each class is possibly larger. Therefore, models that are more parsimonious, e.g. \( \rho GP_{4} \), may provide better results than \( \rho GP_{0.1} \).

Regarding the \( \rho GP \)MRF, the Potts model is a very simple model and more sophisticated models can be used [11]. Furthermore, the selection of the parameter \( \rho \) in eq. (1) value can benefit of a dedicated optimization algorithm.

5. REFERENCES