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

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 ($p\mathcal{GP}$MRF), 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 $p\mathcal{GP}$MRF 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

$$U(y_i|\mathbf{x}_i, \mathcal{N}_i) = -\ln\big(p(\mathbf{x}_i|y_i)\big) + \rho\mathcal{E}(\mathcal{N}_i) \tag{1}$$

where y_i is the label, $p(\mathbf{x}_i|y_i)$ the conditional probability of having $\mathbf{x}_i$ given y_i, $\mathcal{E}$ is an energy term that characterizes the local context of the pixel, $\mathcal{N}_i$ represents the neighborhood of $\mathbf{x}_i$ in the spatial domain and ρ is a positive parameter. The statistical modeling in the spectral domain usually suffers from the increase of the dimension. With the conventional Gaussian assumption, $p(\mathbf{x}_i|y_i = c) \sim \mathcal{N}(\mu_c, \Sigma_c)$, 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 $\mathbf{x}$ are implicitly mapped to $\phi(\mathbf{x})$ 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 $p\big(\phi(\mathbf{x}_i)|y_i\big)$ 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 compute $p\big(\phi(\mathbf{x}_i)|y_i\big)$ 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 $\mathcal{S} = \big\{(\mathbf{x}_i, y_i)\big\}_{i=1}^n$ 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

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

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

$$k(\mathbf{x}_i, \mathbf{x}_j) = \sum_{m=1}^{d_{\mathcal{F}}} e_m \mathbf{v}_m(\mathbf{x}_i)\mathbf{v}_m(\mathbf{x}_j) \tag{3}$$

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

where $d_{\mathcal{F}} = \dim(\mathcal{F})$, $\{e_m,\ m = 1, 2, \ldots\}$ is the sequence of positive eigenvalues in decreasing order of the integral operator T_k, $(T_k f)(\mathbf{z}) = \int_J k(\mathbf{z}, \mathbf{x}) f(\mathbf{x}) d\mathbf{x}$, associated to k and $\{\mathbf{v}_m : \mathbb{R}^d \to \mathbb{R},\ m = 1, 2, \ldots\}$ is the sequence of corresponding normalized eigenfunctions. From eq.(3), ϕ can be defined as

$$\phi : \mathbf{x} \mapsto \sqrt{e_m}\mathbf{v}_m(\mathbf{x}),\ m = 1, 2, \ldots \tag{4}$$

For the Gaussian kernel $d_{\mathcal{F}} = +\infty$ [9]. Therefore 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 $y_i = c$

$$\Omega(\phi(\mathbf{x}_i), y_i) = \sum_{j=1}^{r} \left[\frac{\langle \phi(\mathbf{x}_i) - \boldsymbol{\mu}_c, \mathbf{q}_{cj} \rangle^2}{2\lambda_{cj}} + \frac{\ln(\lambda_{cj})}{2} \right] + \gamma \tag{5}$$

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. 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:

$$\begin{aligned} \Omega(\phi(\mathbf{x}_i), y_i) = & \underbrace{\sum_{j=1}^{r_c} \left[\frac{\langle \phi(\mathbf{x}_i) - \boldsymbol{\mu}_c, \mathbf{q}_{cj} \rangle^2}{2\lambda_{cj}} + \frac{\ln(\lambda_{cj})}{2} \right]}_{\text{computable quantity}} \\ + & \underbrace{\sum_{j=r_c+1}^{r} \left[\frac{\langle \phi(\mathbf{x}_i) - \boldsymbol{\mu}_c, \mathbf{q}_{cj} \rangle^2}{2\lambda_{cj}} + \frac{\ln(\lambda_{cj})}{2} \right]}_{\text{non computable quantity}} \end{aligned} \tag{6}$$

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 $\phi(\mathbf{x})$ 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 $c = 1, \ldots, C$.*

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 $\lambda_{cr} \approx 0$. 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 λ. This model is referred to by $p\mathcal{GP}_0$.

Table 1. List of the sub-models of the parsimonious Gaussian process model. $\mathcal{F}_c$ refers to the signal subspace of the considered class.

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

From this model, it is possible to derive several sub-models. Table 1 lists the different models that can be built from $p\mathcal{GP}_0$. For the model $p\mathcal{GP}_1$, 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 $p\mathcal{GP}_2$, the variance of $\mathcal{F}_c$ is assumed to be equal for all eigenvectors, i.e., $\lambda_{cj} = \lambda_c,\ \forall j \in \{1, \ldots, p_c\}$. For the model $p\mathcal{GP}_4$, it is assumed that the intrinsic dimension is common to every class and the variance is common between them, i.e., $\lambda_{cj} = \lambda_{c'j},\ \forall j \in \{1, \ldots, p\}$ and $c, c' \in \{1, \ldots, C\}$. In term of parsimony, $p\mathcal{GP}_0$ is the least parsimonious model while $p\mathcal{GP}_8$ is the most parsimonious of the proposed models.

In the following, only the model $p\mathcal{GP}_0$ 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 $p\mathcal{GP}_0$ as*

$$\begin{aligned} \Omega(\phi(\mathbf{x}_i), y_i) = & \sum_{j=1}^{p_c} \left(\frac{1}{\lambda_{cj}} - \frac{1}{\lambda} \right) \frac{\langle \phi(\mathbf{x}_i) - \boldsymbol{\mu}_c, \mathbf{q}_{cj} \rangle^2}{2} \\ + & \frac{1}{2\lambda} \|\phi(\mathbf{x}) - \boldsymbol{\mu}_c\|^2 + \sum_{j=1}^{p_c} \frac{\ln(\lambda_{cj})}{2} \\ + & (p_M - p_c) \frac{\ln(\lambda)}{2} + \gamma' \end{aligned} \tag{7}$$

where γ' 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:

$$\begin{aligned} \bar{k}_c(\mathbf{x}_i, \mathbf{x}_j) = & k(\mathbf{x}_i, \mathbf{x}_j) + \frac{1}{n_c^2} \sum_{\substack{l,l'=1 \\ y_l, y_{l'}=c}}^{n_c} k(\mathbf{x}_l, \mathbf{x}_{l'}) \\ - & \frac{1}{n_c} \sum_{\substack{l=1 \\ y_l=c}}^{n_c} \big(k(\mathbf{x}_i, \mathbf{x}_l) + k(\mathbf{x}_j, \mathbf{x}_l) \big). \end{aligned} \tag{8}$$

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

$$(\overline{\mathbf{K}}_c)_{l,l'} = \frac{\bar{k}_c(\mathbf{x}_l, \mathbf{x}_{l'})}{n_c}. \tag{9}$$

With these notations, the following result holds for $p\mathcal{GP}_0$.

$$\Omega(\phi(\mathbf{x}_i), y_i) = \frac{1}{2n_c}\sum_{j=1}^{\hat{p}_c}\frac{1}{\hat{\lambda}_{cj}}\left(\frac{1}{\hat{\lambda}_{cj}} - \frac{1}{\hat{\lambda}}\right)\left(\sum_{\substack{l=1\\ y_l=c}}^{n_c}\beta_{cjl}\bar{k}_c(\mathbf{x}_i,\mathbf{x}_l)\right)^2 + \frac{1}{2\hat{\lambda}}\bar{k}_c(\mathbf{x}_i,\mathbf{x}_i) + \sum_{j=1}^{\hat{p}_c}\frac{\ln(\hat{\lambda}_{cj})}{2} + (\hat{p}_M - \hat{p}_c)\frac{\ln(\hat{\lambda})}{2} \quad (10)$$

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.

(a). Information classes

Class	Samples
Asphalt	6631
Meadow	18649
Gravel	2099
Tree	3064
Metal Sheet	1345
Bare Soil	5029
Bitumen	1330
Brick	3682
Shadow	1947
Total	42776

(b). Classification accuracy

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

Proposition 2 *For $c = 1, \ldots, C$ and the model $p\mathcal{GP}_0$, eq. (7) can be computed with eq. (10), 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*

$$\hat{\lambda} = \frac{1}{\sum_{c=1}^{C}\hat{\pi}_c(r_c - \hat{p}_c)}\sum_{c=1}^{C}\hat{\pi}\big(\text{trace}(\overline{\mathbf{K}}_c) - \sum_{j=1}^{\hat{p}_c}\hat{\lambda}_{cj}\big) \quad (11)$$

and $\hat{\pi}_c = n_c/n$. See [8] for the proof.

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:

$$\frac{\sum_{j=1}^{\hat{p}_c}\hat{\lambda}_{cj}}{\sum_{j=1}^{n_c}\hat{\lambda}_{cj}} > t_h. \quad (12)$$

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

3. EXPERIMENTAL RESULTS

In this section, results obtained on one real data set are presented. 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, see Figure 1.(a). Nine classes have been defined by a photo-interpret for a total of 42776 referenced pixels, see Table 2.(a). 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. The mean result in terms of overall accuracy (percentage of correctly classified pixels) are reported.

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 λ_r refers to the ridge regularization term in KGMM.

Method	Parameter	Range
$p\mathcal{GP}$	σ^2	$\{2^{-3}, 2^{-2}, \ldots, 2^6\}$
	t_h	$\{0.900, 0.911, \ldots, 0.999\}$
	$\hat{p}$	$\{5, 10, \ldots, 45\}$
SVM	σ^2	$\{2^{-3}, 2^{-2}, \ldots, 2^4\}$
	C_p	$\{10^{-2}, \ldots, 10^4\}$
KGMM	σ^2	$\{2^0, 2^1, \ldots, 2^8\}$
	τ	$\{10^{-13}, \ldots, 10^{-6}\}$

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 $p\mathcal{GP}_1$ 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

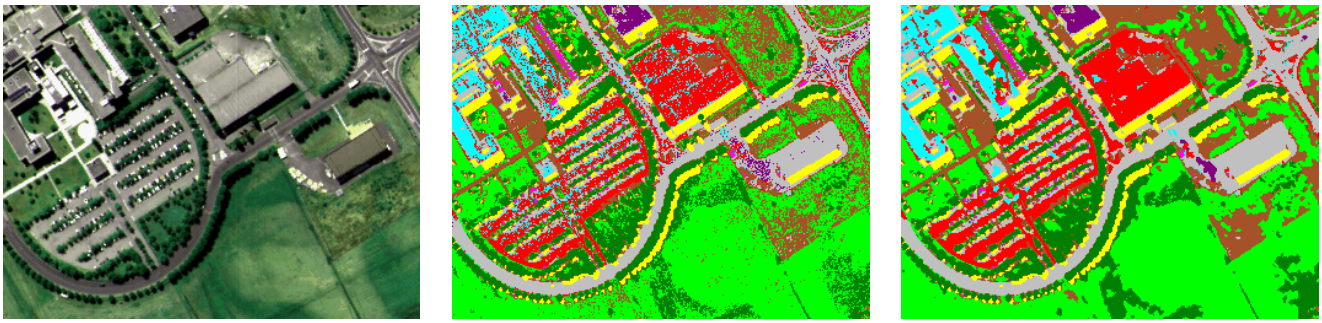
In this section, the proposed models are compared to other classifiers in terms of classification accuracy. The Gaussian kernel was used for each kernel method. All the parameters have been selected by a five fold cross-validation. The ranges of the tested values are reported in Table 3. A small regularization 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-



(a) (b) (c)

Fig. 1. (a) RGB color composition for the *University area*, thematic map obtained with (b) $p\mathcal{GP}_1$ and (c) the $p\mathcal{GP}$MRF.

tional Potts model is used [11]:

$$\mathcal{E}(\mathcal{N}_i) = \sum_{\mathbf{x}_j \in \mathcal{N}_i} \delta(y_i, y_j) \tag{13}$$

where δ is the delta function. A second order neighborhood is considered, i.e., $\mathcal{N}_i$ is the set of 8 surrounding pixels of pixel $\mathbf{x}_i$. For the optimization, a Metropolis algorithm is used. It minimizes the global energy, $U_G = \sum_i U(y_i|\mathbf{x}_i, \mathcal{N}_i)$, by an iterative minimization of the local energy (1). For details, see [11]. ρ 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 $p\mathcal{GP}_1$ 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 $p\mathcal{GP}$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 $p\mathcal{GP}_{0,1}$ 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 $p\mathcal{GP}$MRF was build using the conditional probability provided by the model $p\mathcal{GP}_1$. 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. $p\mathcal{GP}_4$, may provide better results than $p\mathcal{GP}_{0,1}$.

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

5. REFERENCES

[1] C. Chang, *Hyperspectral imaging. Techniques for spectral detection and classification*, Kluwer Academic, 2003.

[2] M. Fauvel, Y. Tarabalka, J. A. Benediktsson, J. Chanussot, and J. Tilton, "Advances in Spectral-Spatial Classification of Hyperspectral Images," *Proceedings of the IEEE*, vol. 101, no. 3, pp. 652–675, Mar. 2013.

[3] G. Camps-Valls and L. Bruzzone, Eds., *Kernel Methods for Remote Sensing Data Analysis*, Wiley, 2009.

[4] Y. Tarabalka, M. Fauvel, J. Chanussot, and J. A. Benediktsson, "SVM and MRF-Based Method for Accurate Classification of Hyperspectral Images," *IEEE Geoscience and Remote Sensing Letters*, vol. 7, no. 4, pp. 736–740, 2010.

[5] J. C. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," in *Advances in Large Margin Classifiers*. 1999, pp. 61–74, MIT Press.

[6] G. Moser and S.B. Serpico, "Combining support vector machines and markov random fields in an integrated framework for contextual image classification," *IEEE Trans. on Geoscience and Remote Sensing*, vol. 51, no. 5, pp. 2734–2752, 2013.

[7] M. Dundar and D. A. Landgrebe, "Toward an optimal supervised classifier for the analysis of hyperspectral data.," *IEEE Trans. Geoscience and Remote Sensing*, vol. 42, no. 1, pp. 271–277, 2004.

[8] C. Bouveyron, M. Fauvel, and S. Girard, "Kernel discriminant analysis and clustering with parsimonious Gaussian process models," http://hal.archives-ouvertes.fr/hal-00687304.

[9] B. Schölkopf, S. Mika, C. J. C. Burges, P. Knirsch, K.-R. Müller, G. Rätsch, and A. J. Smola, "Input space versus feature space in kernel-based methods.," *IEEE Transactions on Neural Networks*, vol. 10, no. 5, pp. 1000–1017, 1999.

[10] M. L. Braun, J. M. Buhmann, and K.-R. Müller, "On relevant dimensions in kernel feature spaces," *J. Mach. Learn. Res.*, vol. 9, pp. 1875–1908, June 2008.

[11] Stan Z. Li, *Markov Random Field Modeling in Image Analysis*, Springer Publishing Company, Incorporated, 3rd edition, 2009.