BAYESIAN LOGISTIC REGRESSION WITH SPARSE GENERAL REPRESENTATION PRIOR FOR MULTISPECTRAL IMAGE CLASSIFICATION

Juan G. Serra[1]*, Pablo Ruiz*[1]*, Rafael Molina*[1] *and Aggelos K. Katsaggelos*[2]

[1] Dpto. de Ciencia de la Computación e I.A., Universidad de Granada.
[2] Dpt. of Electrical Engineering and Computer Science, Northwestern University.
*e-mail:jgserra@decsai.ugr.es

ABSTRACT

In this work we address the multispectral image classification problem from a Bayesian perspective. We develop an algorithm which utilizes the logistic regression function as the observation model in a probabilistic framework, Super-Gaussian (SG) priors which promote sparsity on the adaptive coefficients, and Variational inference to obtain estimates of all the model unknowns. The proposed algorithm is validated on both synthetic and real experiments and compared with other state-of-the-art methods, such as Support Vector Machine and Gaussian Processes, demonstrating its improved performance.

Index Terms— Image Classification, Bayes Methods, Inference Algorithms.

1. INTRODUCTION

Remote sensing images are of great interest in numerous applications such as map drawing, delimitation of parcels, studies on hydrology and forest or agriculture [1–3]. Many of these applications involve the classification of pixels in an image into a number of classes. In the multispectral image classification problem, an expert provides the labels of a small set of pixels to train a classifier, which estimates the labels for the remaining pixels automatically.

Kernel based classification methods such as Support Vector Machine (SVM) or Gaussian Processes (GP) currently represent the state of the art for this task. For a two-class classification problem, SVM finds the maximum margin decision boundary between both classes [4, 5]. This method depends on a set of parameters which must be estimated using cross-validation. This has two drawbacks: first, the parameter estimation does not lead to a reliable estimation when the training set is small, and second, cross-validation implies a grid search, and therefore, a high computational cost. On the other hand, GP are also capable of estimating non-linear decision boundaries using kernel functions and inferring the required parameters automatically [6–8]. However, during the training phase, it is necessary to invert an $N \times N$ matrix (where N is the number of samples), which is very inefficient for large datasets. Some approaches such as GP Latent Variable Models [9] tackle this problem; however, they are still inefficient in applications with limited training time.

These reasons make SVM and GP unsuited for developing real-time applications, so simpler classifiers such as Logistic Regression (LR) are preferred for this purpose. See, for instance, the implementation for hyperspectral image classification in [10], or crash risk prediction model in [11].

This work has been supported in part by the Ministerio de Economía y Competitividad under contract TIN2013-43880-R and the Department of Energy grant DE-NA0002520.

Logistic Regression models the probability of belonging to a class with a sigmoidal function of a linear combination of the features pondered by a set of weights, called adaptive coefficients, that is,

$$\mathrm{p}(y(\mathbf{x})|\mathbf{w}) = \left(\frac{1}{1+e^{-\mathbf{w}^{\mathrm{T}}\mathbf{x}}}\right)^{y} \left(\frac{1}{1+e^{\mathbf{w}^{\mathrm{T}}\mathbf{x}}}\right)^{1-y} \quad (1)$$

where $\mathbf{x} \in \mathbb{R}^M$, is a feature vector whose first component we assumed equal to 1 to consider the bias of the model, $y(\mathbf{x}) \in \{0, 1\}$ is its corresponding label, and $\mathbf{w} \in \mathbb{R}^M$ is the vector of adaptive coefficients to be estimated.

A penalty function on $\mathbf{w}$ is frenquently considered. From a Bayesian perspective [12, 13], this is equivalent to adding a prior distribution on $\mathbf{w}$. Perhaps, the most widely used penalty term is a quadratic function (ridge regression) [14], which forces the coefficients to be close to zero. To impose higher sparsity, other authors use the ℓ_p pseudonorms with $0 < p \leq 1$. For instance, in [15] ℓ_1 regularization, also called LASSO, is used for fMRI data classification, and in [16] a Bayesian framework is used for softmax classification, with an ℓ_p pseudonorm-based prior to classify multispectral images.

Notice that when an adaptive coefficient is 0, the corresponding feature in the samples becomes irrelevant for classification, therefore, if sparsity is imposed on the adaptive coefficients, the non-zero entries of $\mathbf{w}$ represent the useful features for classification [17].

A prior distribution is considered sparse when it is Super-Gaussian [18], i.e., it has heavier tails than the Gaussian distribution, it is more peaked, and it has excess kurtosis. As detailed in [19], these distributions are sparse because most of the distribution mass is located around zero (hence strongly favoring zero values), but the probability of occurrence of large values is higher compared to the Gaussian distribution. SG priors have been used in other image processing problems such as blind image deconvolution [19, 20], or face recognition [21].

In this work we address multispectral image classification utilizing SG priors which provide a general formulation for sparse priors, some of them such as ℓ_1 or ℓ_p have been previously studied in the literature, but the SG formulation allow us to use new priors such as log or the bottom-up approximation. We also use a complete Bayesian framework for modeling and variational inference, which generalizes other inferences methods proposed in the literature, such as Maximum a Posteriori (MAP).

The paper is organized as follows. In section 2 we model the multispectral image classification problem from a Bayesian perspective using Super-Gaussian priors. In section 3 variational inference is used to derive an algorithm to estimate all the model unknowns. In section 4 we evaluate the proposed method on both synthetic and real experiments and compare it with SVM and GP. Finally, section 5 concludes the paper.

2. BAYESIAN MODELING

To perform Bayesian inference we assume that we already have the classification labels $y_j = y(\mathbf{x}_j)$ associated with the feature samples $\mathbf{x}_j$, $j = 1, \ldots, N$. Then we write

$$\mathrm{p}(\mathbf{y}|\mathbf{w}) = \prod_{j=1}^{N} \mathrm{p}(y_j|\mathbf{w}), \tag{2}$$

where $\mathbf{y}$ is an $N \times 1$ vector with its jth component y_j and $\mathrm{p}(y_j|\mathbf{w})$ has been defined in Eq. (1).

We use the following prior model for $\mathbf{w}$

$$\mathrm{p}(\mathbf{w}) \propto \exp\left(-\sum_{i=1}^{M} \rho(w_i)\right), \tag{3}$$

where $\mathbf{w} = (w_1, \ldots, w_M)^{\mathrm{T}}$ and $\rho(\cdot)$ is a sparsity promoting function.

We consider even functions $\rho(\cdot)$ associated with SG distributions, see [19]. The function $\rho(\sqrt{s})$ has to be increasing and concave for $s \in (0, \infty)$ [18]. This condition is equivalent to $\rho'(s)/s$ being decreasing on $(0, \infty)$, that is, for $s_1 \geq s_2 \geq 0$, $\rho'(s_1)/s_1 \leq \rho'(s_2)/s_2$. If this condition is satisfied, then $\rho(\cdot)$ can be represented as (using [22, Ch. 12]),

$$\rho\,(s) = \inf_{\xi > 0} \frac{1}{2}\,\xi\,s^2 \; - \rho^*\left(\frac{1}{2}\xi\right), \tag{4}$$

where $\rho^*\,(\xi)$ is the concave conjugate of $\rho(\sqrt{s})$ and ξ is a variational parameter. The relationship dual to (4) is given by [22],

$$\rho^*\left(\frac{1}{2}\xi\right) = \inf_{s} \frac{1}{2}\,\xi\,s^2 - \rho(s)\,. \tag{5}$$

Equality in (4) is obtained at the optimal values of ξ, which are computed from the dual representation (5) by taking the derivative with respect to s and setting it to zero, which gives $\xi = \rho'(s)/s$. In Fig. 1 we plot some examples of the ρ function which will be used in the experimental section. The exact form of these functions is shown in Table 1.

Finally, we have the following global model

$$\mathrm{p}(\mathbf{w}, \mathbf{y}) = \mathrm{p}(\mathbf{y}|\mathbf{w})\mathrm{p}(\mathbf{w}). \tag{6}$$

Table 1. Super-Gaussian priors.

Prior	$\rho(s)$	$\rho'(s)/s$
Gaussian	$s^2/2$	1
log	$log\lvert s\rvert$	$1/\lvert s\rvert^2$
ℓ_p	$\lvert s\rvert^p/p$	$1/\lvert s\rvert^{p-2}$
exp	$-\sigma\exp(-s^2/2\sigma)$	$\exp(-s^2/2\sigma)$
n-exp	$-\sigma\exp(-\lvert s\rvert^n/n\sigma)$	$\lvert s\rvert^{n-2}\exp(-\lvert s\rvert^n/n\sigma)$
BU	$-$	$1/(1+\exp(\mu s^2))$
MoG	$-\log\sum_k a_k\mathcal{N}_s(0,\sigma_k)$	$\frac{\sum_k a_k/\sigma_k^2\mathcal{N}_s(0,\sigma_k)}{\sum_k a_k\mathcal{N}_s(0,\sigma_k)}$

3. VARIATIONAL BAYESIAN INFERENCE

The Bayesian paradigm dictates that inference on $\mathbf{w}$ should be based on $\mathrm{p}(\mathbf{w}|\mathbf{y})$. However, $\mathrm{p}(\mathbf{w}|\mathbf{y})$ cannot be found in closed form. Therefore, we apply variational methods to approximate this posterior distribution by $\mathrm{q}(\mathbf{w})$. The variational criterion used to find

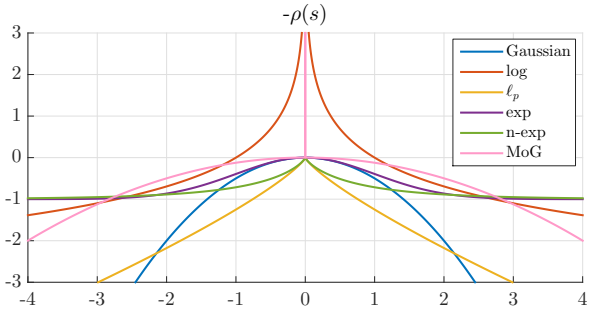

Fig. 1. Set of considered $\rho(s)$ functions from Table 1.

$\mathrm{q}(\mathbf{w})$ is the minimization of the Kullback-Leibler (KL) divergence, given by

$$\begin{aligned}\mathrm{KL}(\mathrm{q}(\mathbf{w})\|\mathrm{p}(\mathbf{w}|\mathbf{y})) &= \int \mathrm{q}(\mathbf{w}) \log\left(\frac{\mathrm{q}(\mathbf{w})}{\mathrm{p}(\mathbf{w}|\mathbf{y})}\right) d\mathbf{w} \\ &= C + \int \mathrm{q}(\mathbf{w}) \log\left(\frac{\mathrm{q}(\mathbf{w})}{\mathrm{p}(\mathbf{w},\mathbf{y})}\right) d\mathbf{w}. \end{aligned} \tag{7}$$

Unfortunately, due to the form of the prior and the observation models defined in (3) and (2) respectively, the integral above cannot be calculated. To solve this problem we find a lower bound for the distribution $\mathrm{p}(\mathbf{w}, \mathbf{y})$ with a function which renders the calculation of $\mathrm{KL}(\mathrm{q}(\mathbf{w}) \parallel \mathrm{p}(\mathbf{w}|\mathbf{y}))$ possible when $\mathrm{p}(\mathbf{w}, \mathbf{y})$ is replaced by such a function.

For the prior in Eq. (3) we have

$$\mathrm{p}(\mathbf{w}) \geq C' \exp\left[-\sum_{i=1}^{M}\left(\frac{1}{2}\xi_i w_i^2 - \rho^*(\frac{1}{2}\xi_i)\right)\right] = \mathbf{M}(\mathbf{w}, \boldsymbol{\xi}). \tag{8}$$

where $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_M)$, and C' is a constant.

In order to obtain a lower bound on $\mathrm{p}(\mathbf{y}|\mathbf{w})$ we follow [23] and apply the variational bound to the sigmoid function. We have

$$\log \mathrm{p}(\mathbf{y}|\mathbf{w}) = \sum_{j=1}^{N} \log \mathrm{p}(\mathbf{y}_j|\mathbf{w}) \geq \log \mathbf{H}(\mathbf{w}, \boldsymbol{\gamma}, \mathbf{y}), \tag{9}$$

where

$$\begin{aligned}\log \mathbf{H}(\mathbf{w}, \boldsymbol{\gamma}, \mathbf{y}) = \sum_{j=1}^{N} \Big\{ & y_j \mathbf{w}^{\mathrm{T}}\mathbf{x}_j - \lambda(\gamma_j)((\mathbf{w}^{\mathrm{T}}\mathbf{x}_j)^2 - \gamma_j^2) \\ & - \frac{\mathbf{w}^{\mathrm{T}}\mathbf{x}_j - \gamma_j}{2} - \log(1 + e^{\gamma_j}) \Big\}, \end{aligned} \tag{10}$$

where $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_N)$ and $\gamma_j \in \mathbb{R}^+$ with $\lambda(\gamma_j) = \frac{1}{2\gamma_j}(\frac{1}{1+e^{-\gamma_j}} - \frac{1}{2})$.

Using the lower bounds in Eqs. (8) and (9), the joint distribution is bounded below by

$$\mathrm{p}(\mathbf{w}, \mathbf{y}) \geq \mathbf{M}(\mathbf{w}, \boldsymbol{\xi})\mathbf{H}(\mathbf{w}, \boldsymbol{\gamma}, \mathbf{y}) = \mathbf{F}(\boldsymbol{\xi}, \boldsymbol{\gamma}, \mathbf{w}, \mathbf{y})\,. \tag{11}$$

We replace $\mathrm{p}(\mathbf{w}, \mathbf{y})$ by this lower bound in Eq. (7). Then we have that $\mathrm{q}(\mathbf{w})$ is a Gaussian distribution with mean and covariance matrix given by

$$\Sigma_{\mathbf{w}}^{-1} = \Xi + 2\sum_{j=1}^{N} \lambda(\gamma_j)\mathbf{x}_j\mathbf{x}_j^{\mathrm{T}}, \tag{12}$$

$$<\mathbf{w}> = \Sigma_{\mathbf{w}} \sum_{j=1}^{N} (y_j - \frac{1}{2})\mathbf{x}_j\,,$$

with $\Xi = \mathrm{diag}(\xi_i), i = 1, \ldots, M$.

The update of the auxiliary vectors $\boldsymbol{\xi}$ and $\boldsymbol{\gamma}$ is given by

$$\hat{\boldsymbol{\xi}} = \arg\max_{\boldsymbol{\xi}} < \log \mathbf{F}(\boldsymbol{\xi}, \boldsymbol{\gamma}, \mathbf{w}, \mathbf{y}) >_{\mathrm{q}(\mathbf{w})} \quad (13)$$

$$\hat{\boldsymbol{\gamma}} = \arg\max_{\boldsymbol{\gamma}} < \log \mathbf{F}(\boldsymbol{\xi}, \boldsymbol{\gamma}, \mathbf{w}, \mathbf{y}) >_{\mathrm{q}(\mathbf{w})} \quad (14)$$

which produces

$$\hat{\xi}_i = \frac{\rho'(\sqrt{<w_i^2>})}{\sqrt{<w_i^2>}} \quad (15)$$

where $<w_i^2> = <w_i>^2 + \Sigma_{\mathbf{w}}(i,i)$ and

$$\hat{\gamma}_j = \sqrt{<(\mathbf{w}^{\mathrm{T}}\mathbf{x}_j)^2>} = \sqrt{(<\mathbf{w}>^{\mathrm{T}}\mathbf{x}_j)^2 + \mathbf{x}_j^{\mathrm{T}}\Sigma_{\mathbf{w}}\mathbf{x}_j}\,. \quad (16)$$

The above inference leads to a learning procedure which is summarized in Algorithm 1.

Notice that to estimate $\mathrm{q}(\mathbf{w})$, $\boldsymbol{\xi}$ and $\boldsymbol{\gamma}$ we do not need to know the explicit form of ρ. This leads to the "bottom up" (*BU*) approach [19, 20], where instead of explicitly defining ρ, we replace $\rho'(s)/s$ in the estimation of $\boldsymbol{\xi}$, by any function decreasing on $(0, \infty)$.

At convergence, Algorithm 1 estimates all the parameters, including the distribution of the adaptive coefficients w_i. The point estimates of the adaptive coefficient vector are given by $<\mathbf{w}>$ in Eq. (12). Given a new sample $\mathbf{x}^*$ (whose first component is equal to 1), we utilize as predictive distribution of the classes

$$\mathrm{p}(\mathcal{C}_1|\mathbf{x}^*) = \frac{1}{1 + e^{-<\mathbf{w}>^T\mathbf{x}^*}},$$
$$\mathrm{p}(\mathcal{C}_0|\mathbf{x}^*) = 1 - \mathrm{p}(\mathcal{C}_1|\mathbf{x}^*). \quad (17)$$

Algorithm 1 Learning Procedure

Require: $\gamma_j^0 = 1, j = 1, \ldots, N,\ \ \xi_i^0 = 1, i = 1, \ldots, M.$

1: **repeat**
2: Calculate $\mathrm{q}^{n+1}(\mathbf{w})$ using Eq. (12).
3: Update ξ_i^{n+1} using Eq. (15).
4: Update γ_j^{n+1} using Eq. (16).
5: **until** convergence

4. EXPERIMENTAL RESULTS

We now present a series of synthetic and real image classification experiments that empirically show the benefits of using SG sparse-promoting priors.

For both experiments the SG penalty functions in Table 1 were used. For the "bottom-up" approach, we selected a function used in face recognition [21] which obtained good results. The proposed methods are compared with standard LR, and two state-of-the-art methods: SVM [5] and GP [7].

We now present an enumeration of the tested parameters: *Gaussian*, *log*, standard LR and GP (with Gaussian kernel) require no parameters; ℓ_p with $p = \{0.02, 0.25, 0.5, 0.75, 1\}$; *exp* with $\sigma = \{0.5, 1, 2, 3\}$; *n-exp* loops over the same σ values with $n = \{0.02, 0.2, 0.5\}$; *BU* with $\mu = \{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1\}$; *MoG*[1] with $a = \{0.2, 0.5, 0.8\}$ and $\sigma = 2$; for SVM we use the Gaussian kernel with length-scale $ls = \{0.04, 0.1, 0.2, 1, 2, 10\}$.

[1]The MoG model we selected consists of a mixture of two Gaussians. The first one with the parameters exposed in the main text and a second with a very small standard deviation, $\sigma = 10^{-8}$, and weight $1 - a$. This mixture simulates a Spike and Slab model [24].

All algorithms were implemented in MATLAB©, and run on an Intel i7-4702MQ @ 2.20 GHz processor.

4.1. Synthetic data

We create a synthetic multispectral image that consists of a collection of 5 bands, only two of which contain relevant information for classification. The image is divided in 2 vertical rectangles, one for each class, whose pixels are realizations from independent Gaussian distributions with means $\mu_1 = (0.8, 0.5, 0.2, 0.5, 0.5)^{\mathrm{T}}$ and $\mu_2 = (0.2, 0.5, 0.8, 0.5, 0.5)^{\mathrm{T}}$, respectively. Assuming independence between the bands, we use a small standard deviation $\sigma = 0.2$ to generate relevant bands (1, 3), and higher deviation $\sigma = 0.5$ for the no relevant ones (2, 4, 5).

The experiment is repeated 10 times with 10 different training sets, each with 20 samples, 10 from each class. As accuracy measure, we use the overall accuracy (OA) over a fixed and balanced test set of 500 samples.

Figure 2 shows the mean OA over the 10 realizations of each experiment for the parameter which gave the best performance; for example, the yellow bar, SVM, corresponds to the mean OA with parameter $ls = 0.2$. We can draw three important conclusions: first, standard logistic regression offers, as expected, the worst performance among the tested algorithms; secondly, the *gaussian* is surpassed by other more sparse-promoting priors, such as *log*, ℓ_p with $p = 1$, and *n-exp* ($n = 0.5$, $\sigma = 2$); finally, the figure provides evidence of the advantages of the proposed method against SVM and GP which perform relatively poorly due to the presence of uninformative bands.

Regarding the estimated weights $\mathbf{w}$, Fig. 3 shows that the SG priors successfully discard the noise bands, assigning high values to w_2 and w_4, the red and blue bands, but very low values to the other coefficients (again w_1 represents the bias). Notice that although the adaptive coefficient vector for LR is sparser than for *BU* or *MoG*, it obtains lower OA. This is due to the fact that LR does not consider any regularization and since the training set is small, LR is overfitting the data.

Eventually, the mean computational time was between 4 ms for *BU*, and *MoG*, and 151 ms for *n-exp*. The computational time for SVM and GP was 5 ms and 22 ms, respectively. Although SVM and GP seem to be faster than the proposed methods, it is due to the small number of samples in the training set.

4.2. Real data

For the second experiment, the proposed method was evaluated on two real multispectral images. They show the cities of Rome and Naples and were acquired by ERS2 SAR and Landsat sensors in 1995 in the context of the Urban Expansion Monitoring project [25]. The goal in this classification problem is to distinguish between urban (class $\mathcal{C}_1$) and non-urban (class $\mathcal{C}_0$) pixels. Both images are composed of 11 bands, where bands 1 and 2 represent the SAR backscattering intensities captured with a gap of 35 days; band 3 is the correlation between them, which is called the interferometric coherence; bands 4-10 are the 7 multispectral bands of Landsat TM sensor; and band 11 is a filtered version of the interferometric coherence, obtained with the method proposed in [25] which has been demonstrated to be very useful to distinguish between urban and non-urban pixels.

The methods described above were evaluated on each image independently. For each image we select 500 samples for training (250

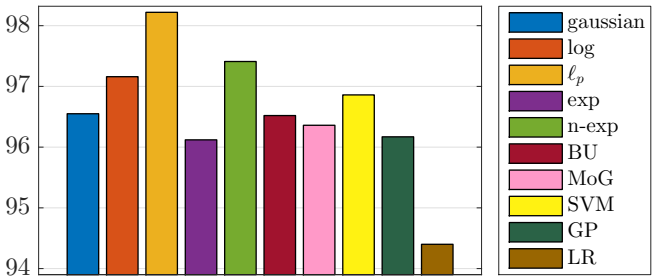

Fig. 2. Overall accuracy for different classification methods on synthetic experiment.

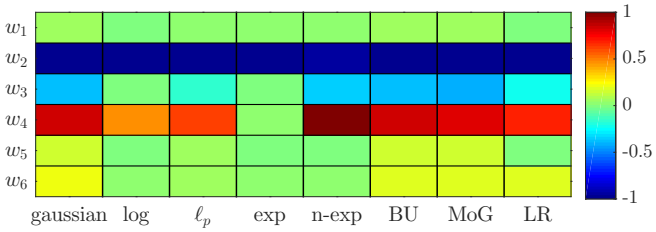

Fig. 3. Adaptive coefficients for different methods in a realization of the synthetic experiment.

for each class) and 10000 test samples (5000 for each class). To obtain unbiased results the experiment is repeated 10 times with different training sets.

In Fig. 4 we show the mean OA over 10 realizations on both images, for each method (optimal parameter). On "Rome95" the obtained mean OAs were 95.73%, 96.57%, 96.53%, 96.24%, 96.25%, 96.22% and 96.27%, for *Gaussian*, *log*, ℓ_p, *exp*, *nexp*, *BU* and *MoG*, respectively. In this case, *log* outperforms the other proposed methods. Standard LR obtains a 95.09% mean OA. However, although all LR methods surpassed 95%, in this case, SVM and GP performed better, obtaining 97.09% and 96.88% mean OA, respectively. This is due to the SVM and GP decision boundaries which are non-linear, and therefore, provide a better separation between classes.

For the "Naples95" image the obtained mean OAs were 96.03%, 96.83%, 97.00%, 96.49%, 96.38%, 97.00% and 96.52%, for *Gaussian*, *log*, ℓ_p, *exp*, *nexp*, *BU* and *MoG*, respectively, whereas standard LR, SVM and GP obtained 94.86%, 96.73% and 96.85% mean OA, respectively. In this case, *log*, ℓ_p and *BU*, performed better than the state-of-the-arts methods. The reason for this is simple: some bands whose coefficients were set to zero by the proposed algorithms may contain confusing and irrelevant information for a classifier, which results in lower performance for SVM and GP, which consider all bands. Therefore, in some cases, algorithms which can disregard irrelevant information can be more useful than methods that produce non-linear decision boundaries.

Regarding the most relevant bands for classification, Fig. 5 shows the adaptive coefficients for the proposed methods and standard Logistic Regression for realization 2 of the experiment with the "Naples95" image. We can see that standard LR, *Gaussian* and *MoG* obtain the least sparse adaptive coefficient vectors, which coincides with the lowest OA obtained by LR methods. We also see that all methods share 3 non-zero bands: the correlation between SAR bands, the Landsat TM band corresponding to the blue band and the filtered coherence.

The highest computational time was reached by *log* (170 ms);

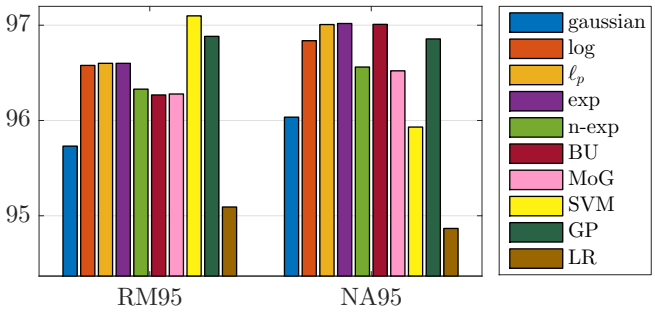

Fig. 4. Overall accuracy for different classification methods on "Rome95" and "Naples95" images.

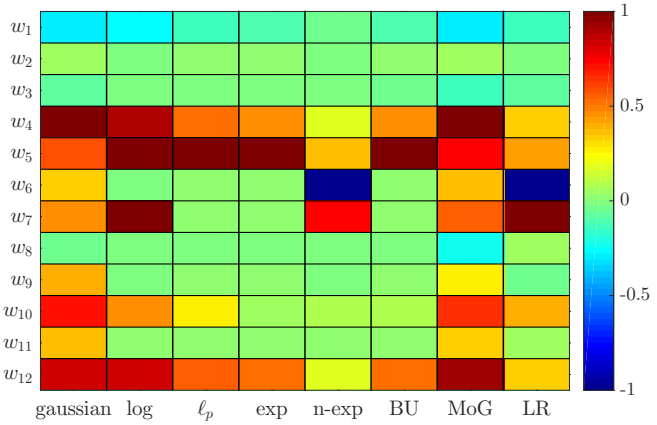

Fig. 5. Adaptive coefficients for different methods in a realization of "Naples95".

however, the remaining methods were between 10 ms (*gaussian*) and 19 ms (*TV*). The computational times for SVM and GP were 93 ms and 12×10^3 ms, respectively.

From this experiment we can conclude that the proposed methods perform well in multispectral classification problems. Moreover, they can identify the most relevant bands for classification and discard useless information. We also observed that depending on the input image, other classifiers such as SVM or GP can perform better, but at the expense of higher computational cost, which makes the proposed methods a good choice for developing real-time applications.

5. CONCLUSIONS

In this work we addressed the multispectral image classification problem from a Bayesian perspective. The likelihood function models a Logistic Regression classifier, where the adaptive coefficients must be estimated. Following the Bayesian paradigm we used a general representation of sparse distributions known as Super-Gaussians. These prior models promote sparsity, which allows us to identify the non-relevant features for classification. Using variational inference we developed an iterative algorithm which estimates the adaptive coefficients. These coefficients were used to calculate the probability of a sample belonging to a class. In the experimental section, the proposed method was validated on both, synthetic and real experiments. Furthermore, it was compared with state-of-the-art methods such as SVM and GP, achieving comparable performance with lower computational cost.

REFERENCES

[1] S. Liang, *Quantitative Remote Sensing of Land Surfaces*, Wiley-Interscience, Hoboken, N.J, 1st edition, Dec. 2003.

[2] G. Camps-Valls, D. Tuia, L. Gómez-Chova, S. Jiménez, and J. Malo, "Remote Sensing Image Processing," *Synthesis Lectures on Image, Video, and Multimedia Processing*, vol. 5, no. 1, pp. 1–192, 2011.

[3] T. Lillesand, R.W. Kiefer, and J. Chipman, *Remote Sensing and Image Interpretation*, John Wiley & Sons, New York, 7th edition, 2015.

[4] B. Scholkopf, A.J. Smola, and B. Schalkopf, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*, Mit Pr, Cambridge, Mass, new edition, 2001.

[5] G. Camps-Valls and L. Bruzzone, *Kernel Methods for Remote Sensing Data Analysis*, John Wiley & Sons, Chichester, U.K, 1st edition, 2009.

[6] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*, Mit Pr, Cambridge, Mass, new edition, Nov. 2005.

[7] P. Ruiz, J. Mateos, G. Camps-Valls, R. Molina, and A. K. Katsaggelos, "Bayesian active remote sensing image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 52, no. 4, pp. 2186–2196, April 2014.

[8] N. Rohani, P. Ruiz, E. Besler, R. Molina, and A. K. Katsaggelos, "Variational gaussian process for sensor fusion," in *European Signal Processing Conference (EUSIPCO 2015)*. Nice (France), September 2015.

[9] M. K. Titsias and N. D. Lawrence, "Bayesian gaussian process latent variable model.," in *AISTATS*, Y. W. Teh and D. M. Titterington, Eds. 2010, vol. 9 of *JMLR Proceedings*, pp. 844–851, JMLR.org.

[10] Z. Wu, Q. Wang, A. Plaza, J. Li, L. Sun, and Z. Wei, "Real-Time Implementation of the Sparse Multinomial Logistic Regression for Hyperspectral Image Classification on GPUs," *IEEE Geoscience and Remote Sensing Letters*, vol. 12, no. 7, pp. 1456–1460, July 2015.

[11] C. Xu, W. Wang, P. Liu, and F. Zhang, "Development of a Real-Time Crash Risk Prediction Model Incorporating the Various Crash Mechanisms Across Different Traffic States," *Traffic Injury Prevention*, vol. 16, no. 1, pp. 28–35, 2015.

[12] C. Bishop, *Pattern Recognition and Machine Learning*, Springer, New York, 2006.

[13] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*, The MIT Press, Cambridge, MA, 1 edition edition, Aug. 2012.

[14] A. E. Hoerl and R. W. Kennard, "Ridge Regression: Biased Estimation for Nonorthogonal Problems," *Technometrics*, vol. 12, no. 1, pp. 55–67, Feb. 1970.

[15] S. Ryali, K. Supekar, D. A. Abrams, and V. Menon, "Sparse logistic regression for whole-brain classification of fMRI data," *NeuroImage*, 2010.

[16] P. Ruiz, N. Pérez de la Blanca, R. Molina, and A. K. Katsaggelos, "Bayesian classification and active learning using lp-priors. Application to image segmentation," in *Signal Processing Conference (EUSIPCO), 2014 Proceedings of the 22nd European*, Sept. 2014, pp. 1183–1187.

[17] R. Zakharov and P. Dupont, "Ensemble Logistic Regression for Feature Selection," in *Pattern Recognition in Bioinformatics*, M. Loog, L. Wessels, M.J.T. Reinders, and D. de Ridder, Eds., number 7036 in Lecture Notes in Computer Science, pp. 133–144. Springer Berlin Heidelberg, Nov. 2011, DOI: 10.1007/978-3-642-24855-9_12.

[18] J.A. Palmer, K. Kreutz-Delgado, and S. Makeig, "Strong Sub- and Super-Gaussianity," in *Latent Variable Analysis and Signal Separation*, V. Vigneron, V. Zarzoso, E. Moreau, R. Gribonval, and E. Vincent, Eds., vol. 6365 of *Lecture Notes in Computer Science*, pp. 303–310. Springer Berlin Heidelberg, 2010.

[19] S. D. Babacan, R. Molina, M. N. Do, and A. K. Katsaggelos, "Bayesian Blind Deconvolution with General Sparse Image Priors," in *Computer Vision ECCV 2012*, Andrew Fitzgibbon, Svetlana Lazebnik, Pietro Perona, Yoichi Sato, and Cordelia Schmid, Eds., number 7577 in Lecture Notes in Computer Science, pp. 341–355. Springer Berlin Heidelberg, 2012.

[20] P. Ruiz, X. Zhou, J. Mateos, R. Molina, and A. K. Katsaggelos, "Variational Bayesian Blind Image Deconvolution: A review," *Digital Signal Processing*, vol. 47, pp. 116 – 127, 2015, Special Issue in Honour of William J. (Bill) Fitzgerald.

[21] M. Yang, L. Zhang, J. Yang, and D. Zhang, "Regularized Robust Coding for Face Recognition," *IEEE Transactions on Image Processing*, vol. 22, no. 5, pp. 1753–1766, May 2013.

[22] R.T. Rockafellar, *Convex Analysis*, Princeton University Press, 1997.

[23] G. Bouchard, "Efficient bounds for the softmax function, applications to inference in hybrid models," in *NIPS'07 Workshop for Approximate Bayesian Inference in continuous/Hybrid Systems*, 2007.

[24] M. Lázaro-Gredilla and M. K. Titsias, "Spike and Slab Variational Inference for Multi-Task and Multiple Kernel Learning," in *Advances in Neural Information Processing Systems 24*, J. Shawe-Taylor, R.S. Zemel, P.L. Bartlett, F. Pereira, and K.Q. Weinberger, Eds., pp. 2339–2347. Curran Associates, Inc., 2011.

[25] L. Gómez-Chova, D. Fernández-Prieto, J. Calpe, E. Soria, J. Vila, and G. Camps-Valls, "Urban monitoring using multi-temporal SAR and multi-spectral data," *Pattern Recognition Letters*, vol. 27, no. 4, pp. 234–243, Mar. 2006.