MAHALANOBIS KERNEL FOR THE CLASSIFICATION OF HYPERSPECTRAL IMAGES

M. Fauvel[⋆]*, A. Villa*[†,⋄] *, J. Chanussot*[†] *and J. A. Benediktsson*[⋄]

[⋆]MISTIS, INRIA Rhône Alpes & Laboratoire Jean Kuntzmann, Grenoble, France
[†]GIPSA-lab, Grenoble Institute of Technology, France
[⋄]Faculty of Electrical and Computer Engineering, University of Iceland, Iceland

ABSTRACT

The definition of the Mahalanobis kernel for the classification of hyperspectral remote sensing images is addressed. Class specific covariance matrices are regularized by a probabilistic model which is based on the data living in a subspace spanned by the p first principal components. The inverse of the covariance matrix is computed in a closed form and is used in the kernel to compute the distance between two spectra. Each principal direction is normalized by a hyperparameter tuned, according to an upper error bound, during the training of an SVM classifier. Results on real data sets empirically demonstrate that the proposed kernel leads to an increase of the classification accuracy by comparison to standard kernels.

Index Terms— Mahalanobis kernel, probabilistic principal component analysis, support vector machine, hyperspectral images, classification.

1. INTRODUCTION

Kernel methods have received a considerable attention during the last decade [1]. Their performances for classification, regression or feature extraction make them popular in the remote sensing community [2]. The core of kernel learning algorithms is the *kernel function*. It measures the similarity between two spectra $\mathbf{x}$ and $\mathbf{y}$ in a d-dimensional vector space, and enable the transformation of a linear algorithm into a non-linear one [3]. Over the different kernels used in remote sensing, the Gaussian kernel:

$$k(\mathbf{x},\mathbf{y}) = \exp\left(-\frac{1}{2}\sum_{i=1}^{d}\frac{(x_i-y_i)^2}{\gamma^2}\right). \tag{1}$$

is widely used because it usually gives good results and has only one hyperparameter (γ) to be tuned. By using kernels, an implicit mapping of the spectra from the *input space* to the *feature space* is done, the feature space being associated to the kernel [1].

Under some weak conditions, the feature space induced by the kernel is a Riemannian manifold [4, 5]. The metric tensor is

$$g_{ij}(\mathbf{x}) = \left.\frac{\partial^2 k(\mathbf{x},\mathbf{y})}{\partial x_i \partial y_j}\right|_{\mathbf{y}=\mathbf{x}} \tag{2}$$

which is for the Gaussian kernel: $g_{ij}(\mathbf{x}) = \gamma^{-2}\boldsymbol{\delta}_{ij}$ with $\boldsymbol{\delta}_{ij} = 1$ if $i = j$ and 0 otherwise. This metric stretches or compresses the Euclidean distance between $\mathbf{x}$ and $\mathbf{y}$ by a factor γ^{-2} and the implicit statistical model associated to the input data $\mathbf{x}$ is the normal law with a diagonal covariance matrix and identical elements [6, Chap. 4]: Each variable[1] has the same variance and there is no covariance between variables (which is not true in practice). The model also assumes that each variable is equally relevant for the given task, *e.g.*, classification or regression. A more advanced model is to consider that the data follow a normal law with a diagonal covariance matrix, but with no identical diagonal elements: Each variable has its own variance, but still no covariance. It is then possible to tune the relevance of each variable separately. It was shown in [7] that this model improves the classification accuracy of remote sensing images, but it also increases the computational load. The more general model, full covariance matrix, leads to the well known Mahalanobis kernel (MK) [8]:

$$k(\mathbf{x},\mathbf{y}) = \exp\left(-\frac{1}{2\gamma^2}\big(\mathbf{x}-\mathbf{y}\big)^t\mathbf{Q}\big(\mathbf{x}-\mathbf{y}\big)\right). \tag{3}$$

Several definitions of $\mathbf{Q}$ exist for the problem of classification: It can be either the inverse of the covariance matrix Σ of the total training samples [8] or the covariance matrix of the considered class [9], *e.g.*, for m classes problem, if the classifier separates the class c against all the others, $\mathbf{Q}$ is Σ_c^{-1}. Generally, $\mathbf{Q}$ can be any positive definite matrix. The metric induced is $g_{ij}(\mathbf{x}) = \gamma^{-2}\mathbf{Q}_{ij}$, which stretches or compresses the variance of the data along their principal directions (see Section 3). Although the MK better handles the data characteristics than the conventional Gaussian kernel in small/moderate dimensional space, it is difficult to use in high dimensional spaces, such as hyperspectral remote sensing images. As a matter of fact, the estimation of the covariance matrix is ill-conditioned, making its inversion impossible or at least unstable. Moreover, all the principal directions are

This work has been in part supported by the European Community's Marie Curie Research Training Networks Programme under contract MRTN-CT-2006-035927, Hyperspectral Imaging Network (HYPER-I-NET).

[1] In this study, the variables are the different components of the spectra.

not equally relevant for classification: A subset of them corresponds to the signal while the remaining dimensions correspond to the noise.

In this article, it is proposed to regularize $\mathbf{Q}$ in a suitable way and to tune the weight of the principal directions according to their relevance in classification. The regularization strategy is detailed in Section 2. The link between probabilistic PCA and the proposed regularization is also discussed. The Mahalanobis kernel is then constructed with the regularized inverse covariance matrix in Section 3. Finally, experiments on real hyperspectral data sets are reported in Section 4.

2. REGULARIZATION OF COVARIANCE MATRICES

From the spectral theorem, a covariance matrix can be written as:

$$\Sigma = \mathbf{V}\Delta\mathbf{V}^t \tag{4}$$

where Δ is the diagonal matrix of eigenvalues and $\mathbf{V}$ is an orthonormal matrix of corresponding eigenvectors. Its inverse is

$$\Sigma^{-1} = \mathbf{V}\Delta^{-1}\mathbf{V}^t. \tag{5}$$

Ill-conditioning is related to a high condition number $\kappa(\Sigma)$, *i.e.*, the ratio between the largest δ_1 and the smallest δ_d eigenvalue [10]. In general, in hyperspectral imagery, Σ is not full rank because of the high correlation between two adjacent bands. One consequence of this is a high condition number. Another consequence is that not all principal directions carry the relevant signal and thus it is possible to discard principal directions corresponding to zero (or close to) eigenvalues.

Following [11, 12], it is proposed in this article to use the PCA-Ridge regularization based approach: Noting $\mathbf{I}_d^p$ the diagonal matrix with the p first elements equal to 1 and the remaining equal to 0 and defining $\Omega = \mathbf{V}\mathbf{I}_d^p\mathbf{V}^t$, the ill-posed problem (5) is changed to $\big(\Omega\Sigma + \tau\mathbf{I}_d\big)^{-1}\Omega$ which is, after trivial simplifications, equal to $\mathbf{V}\Lambda(\tau,p)\mathbf{V}^t$ with

$$\Lambda(\tau,p) = \operatorname{diag}\left[\frac{1}{\delta_1+\tau},\ \dots\ ,\ \frac{1}{\delta_p+\tau},\ 0,\ \dots\ ,\ 0\right]. \tag{6}$$

Usually, the regularization parameter τ is set to a small value: $\delta_i+\tau \approx \delta_i$ if δ_i is high enough. Therefore, principal directions corresponding to high eigenvalues are slightly regularized and those corresponding to small eigenvalues are largely regularized. The remaining $(d-p)$ principal directions are discarded.

Finally, with the regularized estimate of the covariance matrix, $\|\mathbf{x}-\mathbf{y}\|^2_{\Sigma^{-1}}$can be rewritten as:

$$\begin{aligned}\|\mathbf{x}-\mathbf{y}\|^2_{\Sigma^{-1}} &= \big(\mathbf{x}-\mathbf{y}\big)^t\mathbf{V}\Lambda(\tau,p)\mathbf{V}^t\big(\mathbf{x}-\mathbf{y}\big)\\ &= \left\|\big[\mathbf{V}\Lambda^{0.5}(\tau,p)\big]^t\big(\mathbf{x}-\mathbf{y}\big)\right\|^2\\ &= \left\|\mathbf{A}^t\big(\mathbf{x}-\mathbf{y}\big)\right\|^2\end{aligned} \tag{7}$$

with

$$\mathbf{A} = \left[\frac{\mathbf{v}_1}{(\delta_1+\tau)^{0.5}}\middle|\cdots\middle|\frac{\mathbf{v}_p}{(\delta_p+\tau)^{0.5}}\right] \tag{8}$$

the projection operator on the vector space $\mathcal{A}$ spanned by the p first regularized principal directions (the projection on the $(d-p)$ last principal directions are always null) and $\|\cdot\|^2$ the Euclidean norm in $\mathbb{R}^p$. $\mathcal{A}$ represents the class specific subspace[2] and $\bar{\mathcal{A}}$ the noise subspace: $\mathbb{R}^d = \mathcal{A}+\bar{\mathcal{A}}$. With such a regularization, it is supposed that all the necessary information for the discrimination of class c is included in $\mathcal{A}_c$.

Table 1. Number p of principal components kept for the computation of Σ^{-1} for each class c.

c	1	2	3	4	5	6	7	8	9
p	20	21	14	26	13	17	12	19	14

This model is closely related to *probabilistic principal component analysis* (PPCA) [13]. In PPCA, it is assumed that the observed d variables $\mathbf{x}$ are a linear combination of p unobserved variables $\mathbf{s}$ (p≪d):

$$\mathbf{x} = \mathbf{W}\mathbf{s} + \boldsymbol{\mu} + \boldsymbol{\varepsilon} \tag{9}$$

with $\mathbf{s} \sim \mathcal{N}(0,\mathbf{I})$ and $\boldsymbol{\varepsilon} \sim \mathcal{N}(0,\epsilon^2\mathbf{I})$. It follows that $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu},\mathbf{W}\mathbf{W}^t+\epsilon^2\mathbf{I})$. With PPCA, it is clear that all the information/signal in $\mathbf{x}$ is contained in $\mathbf{s}$. It can be proved that $\mathbf{s}$ lives in a subspace spanned by the p first eigenvectors of the covariance matrix of $\mathbf{x}$ [13]. This subspace correspond to $\mathcal{A}$, that was previously defined.

Using [13], $\mathcal{A}$ is estimated using the p first eigenvalues and corresponding eigenvectors computed on the empirical sample estimate of $\mathbf{\Sigma}$:

$$\hat{\mathbf{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i-\bar{\mathbf{x}})(\mathbf{x}_i-\bar{\mathbf{x}})^t \tag{10}$$

with n the number of training samples and $\bar{\mathbf{x}}$ being the sample mean of $\mathbf{x}$.

It results that the inverse of the covariance matrix can be computed in a closed form from the empirical covariance matrix. The parameter p controls the model: the size of $\mathcal{A}$ where the data originally lived. Therefore, it is important to tune p correctly, since if p is set to a small value relevant information is lost, while with a too big value some noise is included in the computation. In remote sensing applications, it is conventionally selected by considering a certain percentage of the cumulative variance (total eigenvalues sum). A more robust approach is to use statistical model selection criteria [14]: *Bayesian Information Criteria* (BIC) or *Akaike Information Criteria* (AIC) ... After some trials, BIC was selected in this work (AIC gave too high p). Table 1 reports the size of the subspace for each class obtained with the BIC (see Section 3 for a description of the data set). BIC is computed as:

$$\mathrm{BIC}(p) = -2l + (d-1)(p-1)\log(n) \tag{11}$$

where l is the log-likelihood associated to the PPCA model for the considered p. The optimal p minimizes the BIC.

[2] In the following the subscript c indicates the corresponding class.

3. MAHALANOBIS KERNEL

In this section, the proposed Mahalanobis kernel is detailed. In $\mathcal{A}_c$, the variables from the class c are uncorrelated. It is therefore suitable to tune the relevance of each variable for the classification problem by introducing a diagonal matrix Γ of hyperparameters ($\Gamma_{ii} = 1/\gamma_i^2$) [7]:

$$k_c(\mathbf{x}, \mathbf{y}) \quad = \quad \exp\left(-\frac{1}{2}\big(\mathbf{x}-\mathbf{y}\big)^t \mathbf{A}_c \Gamma \mathbf{A}_c^t \big(\mathbf{x}-\mathbf{y}\big)\right). \quad (12)$$

The metric tensor is now

$$g_{ij}(\mathbf{x}) = \sum_{q=1}^{p} \frac{\mathbf{v}_q[i]\mathbf{v}_q[j]^t}{(\delta_q + \tau)\gamma_q^2} \qquad (13)$$

with $\mathbf{v}_q[i]$ the i^{th} element of eigenvector $\mathbf{v}_q$ associated to eigenvalue δ_q, τ the regularization parameter, p the number of remaining principal directions and γ_q the hyperparameter, which will be tuned during the training process. With this kernel, the distance between two spectra is stretched or compressed along their p first principal components. The variation along the $(d-p)$ last principal directions is assumed to be caused by the noise and they are not considered in the computation.

The above formulation has several advantages over (3):

1. The condition number of the matrix is equal to $(\delta_1 + \tau)/(\delta_p + \tau)$, which is controlled by the two parameters p and τ. For instance, with the class 1, the initial $\kappa(\hat{\Sigma})$ is approximately 3×10^6 while with the proposed approach it is approximately 4×10^3.
2. It is known that the principal directions are not optimal for classification since they do not maximize any discrimination criterion. However, they still span a subspace where there are some variation in the data. By controlling, with Γ, which directions are relevant (or discriminative) for the classification, it is possible to go further in the classification process: The feature space is modified during the training process to ensure a better discrimination between samples.
3. Γ provides some information about the relevance of each principal direction.

4. EXPERIMENTS

In this section, results obtained on real data sets are presented. The data is the *University Area* of Pavia, Italy, acquired with the ROSIS-03 sensor. The image has 103 bands and is 610×340 pixels. Nine classes have been defined. More details about the data sets can be found in [15].

For the classification, a SVM with gradient based approach to tune the hyperparameters was used [16] and one versus all multiclass strategy was employed. The proposed kernel has been compared with the conventional Gaussian kernel, the Mahalanobis kernel where a small ridge regularization has been done to prevent instability, and the proposed regularized Mahalanobis kernel. The covariance matrix Σ_c for class c was estimated with the available training samples. For the experiments, the regularization parameter p was selected with the cumulative variance (99% and 99.9%) and with the BIC criterion.

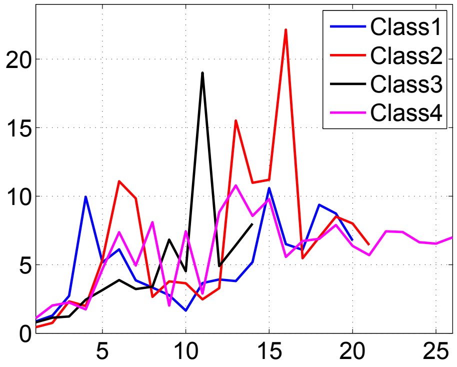

Fig. 1. Hyperparameters γ_q for the four first classes. Horizontal axis: Number of the principal component q. Vertical axis: value of the hyperparameters γ_q. The size of the specific subspace was selected with the BIC, see Table 1.

Classification results are reported in Table 2. The results must be considered as 9 binary classification problems: No fusion rule was applied to obtain the multiclass classification result.

Without any regularization, the conventional Mahalanobis kernel performs worst in terms of classification accuracy than the Gaussian kernel. With $\tau = 10^2$, it leads to a slight increase of the classification while the training time is drastically increased. The proposed kernel leads to an increased accuracy when p corresponds to 99.9% of the cumulative variance and when p is selected with the BIC. For the proposed kernel, the regularization parameter p has stronger influence on the classification accuracy than τ. We have set τ to zero with the BIC strategy and no difference in terms of classification have been found. By retaining more principal directions, up to a certain amount, the training process becomes longer with no increase of the accuracy. Then the accuracy decreases. The BIC strategy permits to select the right size of the class specific subspace and thus minimizes the total training time.

Figure 1 displays the hyperparameters γ_p found for the first four classes. It can be seen that the hyperparameters are class-dependent and, therefore, need to be tuned independently for each class.

5. CONCLUSIONS AND PERSPECTIVES

The classification of hyperspectral images with the Mahalanobis kernel was investigated in this article. A regularization of the covariance matrix was proposed, based on a probabilistic model. Using the BIC, the class specific subspace is defined for each class, leading to a closed form solution for the inverse of Σ. The distance between two spectra is now computed in the class specific subspace $\mathcal{A}$ rather than in $\mathbb{R}^d$. A set of hyperparameters is added in the kernel. They are tuned during the training of the SVM.

Experimental results on real data sets have shown that the proposed kernel can improve the classification accuracy when compared to a standard Gaussian kernel. Using BIC to select the size of the class specific subspace is more pertinent than considering the cumulative variance. Further experiments on other data sets are needed to better assess the

Table 2. Classification accuracies for the different kernels in percentage of correctly classified samples. Here S means that there is one hyperparameter and M means that there is one hyperparameter per variable. For the regularized kernel, the number in brackets represents p the number of selected principal directions. It corresponds, for the first two results, to 99% of the total variance and 99.9% for the following two results. The last column corresponds to p selected with the BIC. For each class, the number of training samples is indicated in brackets.

Kernel	Gaussian		Mahalanobis		Reg-Mahalanobis				
Γ	S	M	S	S	M	M	M	M	M
τ	-	-	0	10^2	0	10^2	0	10^2	0
Asphalt (548)	94.4	94.9	88.4	91.8	94.5 (7)	94.5 (7)	**95.9 (39)**	**95.9 (39)**	**96.0**
Meadow (540)	79.2	78.3	70.9	75.8	77.2 (4)	77.2 (4)	**81.9 (24)**	**81.9 (24)**	**82.0**
Gravel (392)	95.7	97.2	96.2	97.0	96.8 (5)	96.8 (5)	**97.5 (30)**	**97.5 (30)**	**97.6**
Tree (524)	93.8	94.4	96.5	97.8	94.3 (6)	94.3 (6)	**98.3 (23)**	**98.3 (23)**	**98.3**
Metal Sheet (265)	99.8	99.8	**99.9**	**99.9**	99.7 (2)	99.7 (2)	**99.9 (15)**	**99.9 (15)**	**99.9**
Bare Soil (532)	89.3	85.4	90.1	91.0	83.3 (4)	83.3 (4)	**92.4 (21)**	**92.4 (21)**	91.4
Bitumen (375)	97.8	98.5	98.9	**99.3**	98.9 (28)	98.9 (28)	99.1 (61)	99.1 (61)	98.7
Brick (514)	95.3	96.2	93.5	95.5	96.8 (8)	96.8 (8)	97.5 (41)	**97.5 (41)**	**97.5**
Shadow (231)	**99.9**	**99.9**	98.8	99.7	**99.9 (9)**	**99.9 (9)**	**99.9 (38)**	**99.9 (38)**	**99.9**
Average class accuracy	93.8	93.9	92.6	94.2	93.5	93.5	**95.9**	**95.9**	**95.7**

effectiveness of the proposed kernel.

Current researches are now oriented to:

- The construction of polynomial kernel with the scalar product defined with the inverse of Σ:

$$k(\mathbf{x},\mathbf{y}) = \left(\mathbf{x}^t\Sigma^{-1}\mathbf{y}+1\right)^r. \tag{14}$$

- Investigate the influence of the noise in the kernel definition. In the proposed kernel, the distance between two spectra is considered in $\mathcal{A}$. However, important information may be lost: Two spectra can be close in $\mathcal{A}$ whereas they are far-off in $\bar{\mathcal{A}}$. In future research, it is important to assess the relevance of the information provided by $\bar{\mathcal{A}}$ and how it can be used, due to the low signal to noise ratio of $\bar{\mathcal{A}}$.

6. REFERENCES

[1] T. Hofmann, B. Schölkpof, and A. J. Smola, "Kernel methods in machine learning," *The annals of statistics*, vol. 36, no. 3, pp. 1171–1220, 2008.

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

[3] V. Vapnik, *The Nature of Statistical Learning Theory, Second Edition*, Springer, New York, 1999.

[4] B. Scholkopf, C. Burges, and A. Smola, *Geometry and Invariance in Kernel Based Methods* In *Advances in Kernel Methods - Support Vector Learning*, MIT Press, 1998.

[5] P. Williams, S. Li, J. Feng, and S. Wu, "A geometrical method to improve performance of the vector machine," *IEEE Transactions on Neural Networks*, vol. 18, no. 3, pp. 942–947, May 2007.

[6] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*, MIT Press, 2006.

[7] A. Villa, M. Fauvel, J. Chanussot, P. Gamba, and J. A. Benediktsson, "Gradient optimization for multiple kernel's parameters in support vector machines classification," in *Geoscience and Remote Sensing Symposium, 2008. IGARSS 2008. IEEE International*, July 2008.

[8] G. Camps-Valls, A. Rodrigo-Gonzalez, J. Muoz-Mari, L. Gomez-Chova, and J. Calpe-Maravilla, "Hyperspectral image classification with Mahalanobis relevance vector machines," in *Geoscience and Remote Sensing Symposium, 2007. IGARSS 2007. IEEE International*, July 2007, pp. 3802–3805.

[9] S. Abe, "Training of support vector machines with Mahalanobis kernels," in *Artificial Neural Networks: Formal Models and Their Applications - ICANN 2005*, Lecture Notes in Computer Science, pp. 571–576. Springer Berlin / Heidelberg, 2005.

[10] C. R. Vogel, *Computational Methods for Inverse Problems*, Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2002.

[11] C. Bernard-Michel, S. Douté, M. Fauvel, L. Gardes, and S. Girard, "Machine learning techniques for the inversion of planetary hyperspectrales images," in *Proc. of IEEE Int. Workshop on hyperspectral image and signal processing (WHISPERS-09)*, Grenoble, 2009.

[12] C. Bernard-Michel, L. Gardes, and S. Girard, "Gaussian regularized sliced inverse regression," *Statistics and Computing*, vol. 19, pp. 85–98, 2009.

[13] M. E. Tipping and C. M. Bishop, "Probabilistic principal component analysis," *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, vol. 61, no. 3, pp. 611–622, 1999.

[14] G. Schwarz, "Estimating the dimension of a model," *The annals of Statistics*, vol. 6, no. 2, pp. 461–464, 1978.

[15] M. Fauvel, *Spectral and spatial methods for the classification of urban remote sensing data*, Ph.D. thesis, Grenoble Institute of Technology and the University of Iceland, 2007.

[16] O. Chapelle, V. Vapnik, O. Bousquet, and S. Mukherjee, "Choosing multiple parameters for support vector machines," *Machine Learning*, vol. 46, no. 1-3, pp. 131–159, 2002.