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

[⋄]Dept. of Electrical and Computer Engineering, University of Iceland, Iceland

I. Introduction

Kernel methods have received a considerable attention this 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 is very popular:

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

It usually gives good results and has only one hyperparameter (γ) to be tuned.

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 by a factor γ^{-2} and the implicit model associated to the input data $\mathbf{x}$ is the normal law with a diagonal covariance matrix and identical elements: Each variable[1] has the same variance and there is no covariance between variables (which is not true in practice). It also assumes that each variable is equally relevant for the given task, *e.g.* classification. 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 [6] that it improves the classification accuracy, but it also increases the computational load. The more general model, full covariance matrix, leads to the well known Mahalanobis kernel (mk) [7]:

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

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 [7] or the covariance matrix of the considered class [8], *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 $g_{ij}(\mathbf{x}) = \gamma^{-2}\mathbf{Q}_{ij}$, it stretches or compresses the variance of the data along their principal directions. Although the mk better handles the data characteristics than conventional Gaussian kernel in small/moderate dimensional space, it is difficult to use in high dimensional space, such as hyperspectral remote sensing images. As a matter of fact, the estimation of the covariance matrix is ill-conditioned, making its inversion unstable. Moreover, all the principal directions are not equally relevant for the classification purpose: 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 the classification problem.

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

II. Regularization of the covariance matrix

From the spectral theorem, the 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 the 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 $\kappa(\Sigma)$: The ratio between the largest δ_1 and the smallest δ_d eigenvalue [9]. In general, in hyperspectral imagery, Σ is not full rank. One consequence 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 closed to) eigenvalues.

Following [10], [11], 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

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

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, (3) can be rewritten as:

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

with $\mathbf{A}$ 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. In the following the subscript c indicates the corresponding class.

III. 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 ($\Gamma_{ii} = 1/\gamma_i^2$) [6]:

$$k_c(\mathbf{x},\mathbf{y}) = \exp\left(-\frac{1}{2}\big(\mathbf{x}-\mathbf{y}\big)^t\mathbf{A}_c\Gamma\mathbf{A}_c^t\big(\mathbf{x}-\mathbf{y}\big)\right). \tag{8}$$

The metric tensor is now

$$g_{ij}(\mathbf{x}) = \sum_{q=1}^{p}\frac{\mathbf{v}_i(q)\mathbf{v}_j(q)^t}{(\delta_q+\tau)\gamma_q^2} \tag{9}$$

with $\mathbf{v}_i(q)$ 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.

This formulation has several advantages over (3): (i) The condition number of the matrix is equal to $\gamma_p^2(\delta_1+\tau)/\gamma_1^2(\delta_p+\tau)$, which is controlled by the two parameters p and τ; (ii) it is known that the principal directions are not optimal for a classification application since they do not maximize any discrimination criterion, but 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; (iii) Γ provides some information of the relevance of each principal direction.

IV. 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. 9 classes have been defined.

For the classification, a SVM with gradient based approach to tune the hyperparameters was used [12] 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.

Classification results are reported in Table I. The proposed kernel leads to an increased accuracy when p corresponds to 99.9% of the cumulative variance. From the table, if for the conventional Mahalanobis kernel τ has an influence on the process, with the proposed kernel the regularization parameter p is more critical. 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. Further research are conducted to find the optimal p for each class.

Table I

Classification accuracies for the different kernels in percentage of correctly classified samples. s means there is one hyperparameter and m means there is one hyperparameter per variable. For the regularized kernel, the number in brackets represents p the number of remaining principal direction. It corresponds, for the firsts two results, to 99% of the total variance and 99.9% for the last two results. 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
τ	-	-	0	10^2	0	10^2	0	10^2
Asphalt (548)	94.4	94.9	88.4	91.8	94.5 (7)	94.5 (7)	**95.9 (39)**	**95.9 (39)**
Meadow (540)	79.2	78.3	70.9	75.8	77.2 (4)	77.2 (4)	**81.9 (24)**	**81.9 (24)**
Gravel (392)	95.7	97.2	96.2	97.0	96.8 (5)	96.8 (5)	**97.5 (30)**	**97.5 (30)**
Tree (524)	93.8	94.4	96.5	97.8	94.3 (6)	94.3 (6)	**98.3 (23)**	**98.3 (23)**
Metal Sheet (265)	99.8	99.8	**99.9**	**99.9**	99.7 (2)	99.7 (2)	**99.9 (15)**	**99.9 (15)**
Bare Soil (532)	89.3	85.4	90.1	91.0	83.3 (4)	83.3 (4)	**92.4 (21)**	**92.4 (21)**
Bitumen (375)	97.8	98.5	98.9	**99.3**	98.9 (28)	98.9 (28)	99.1 (61)	99.1 (61)
Brick (514)	95.3	96.2	93.5	95.5	96.8 (8)	96.8 (8)	97.5 (41)	**97.5 (41)**
Shadow (231)	**99.9**	**99.9**	98.8	99.7	**99.9 (9)**	**99.9 (9)**	**99.9 (38)**	**99.9 (38)**
Average class accuracy	93.8	93.9	92.6	94.2	93.5	93.5	**95.9**	**95.9**

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] 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.

[7] 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.

[8] 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.

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

[10] 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.

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

[12] 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.