Regularized Covariance Estimators for Hyperspectral Data Classification and Its Application to Feature Extraction

Bor-Chen Kuo, *Member, IEEE*
Department of Mathematic Education
National Taichung Teachers College
Taichung, Taiwan 403
Email: kbc@mail.ntctc.edu.tw

David A. Landgrebe, *Life Fellow, IEEE*
School of Electrical and Computer Engineering
Purdue University, West Lafayette, Indiana 47907-1285
Email: landgreb@ecn.purdue.edu

Copyright © 2002 IEEE. Reprinted Transactions of the International Geoscience and Remote Sensing Symposium, Toronto Canada, June 2002.

This material is posted here with permission of the IEEE. Internal or personal use of this material is permitted. However, permission to reprint/republish this material for advertising or promotional purposes or for creating new collective works for resale or redistribution must be obtained from the IEEE by sending a blank email message to pubs-permissions@ieee.org.

By choosing to view this document, you agree to all provisions of the copyright laws protecting it.

Regularized Covariance Estimators for Hyperspectral Data Classification and Its Application to Feature Extraction

Bor-Chen Kuo, *Member, IEEE*
Department of Mathematic Education
National Taichung Teachers College
Taichung, Taiwan 403
Email: kbc@mail.ntctc.edu.tw

David A. Landgrebe, *Life Fellow, IEEE*
School of Electrical and Computer Engineering
Purdue University, West Lafayette, Indiana 47907-1285
Email: landgreb@ecn.purdue.edu

***Abstract*-The main purpose of this work is to find an improved regularized covariance estimator of each class with the advantages of LOOC, and BLOOC, which are useful for high dimensional pattern recognition problems. The searching ranges of LOOC and BLOOC are between the linear combinations of three pair covariance estimators. The first proposed covariance estimator (Mixed-LOOC1) extended the searching range and is a general case of LOOC and BLOOC. By observing that the optimal value of leave-one-out likelihood function of LOOC usually occurs at near the end point of the parameter domain, the second covariance estimator (Mixed-LOOC2), which needs less computation, was proposed. Using the proposed covariance estimator to improve the linear feature extraction methods when the multivariate data is singular or nearly so is demonstrated.**

I. Introduction

Regularized Discriminant Analysis (RDA; [1]), Leave-One-Out Covariance Estimator (LOOC; [2]) and Bayesian Leave-One-Out Covariance Estimator (BLOOC; [3]) are proposed for solving the singular or nearly singular condition in high dimensional classification problem. From [4], the performance of LOOC is better than that of RDA, and BLOOC is only better than LOOC in a few cases. Based on this observation, a new regularized covariance estimator with the advantages of both LOOC and BLOOC is needed. In the first section, the models of LOOC and BLOOC will be discussed then in section 2 the hybrid models of LOOC and BLOOC are proposed.

The model of LOOC is

$$\hat{\Sigma}_i(\alpha_i) = \begin{cases} (1-\alpha_i)\,diag(S_i) + \alpha_i S_i & 0 \le \alpha_i \le 1 \\ (2-\alpha_i)S_i + (\alpha_i - 1)S & 1 \le \alpha_i \le 2 \\ (3-\alpha_i)S + (\alpha_i - 2)diag(S) & 2 \le \alpha_i \le 3 \end{cases}$$

where S_i is the ML covariance estimator of class i, and S is the common (pooled) covariance.

The mixing parameter α_i is determined by maximizing the average leave-one-out log likelihood of each class:

$$LOOL_i = \frac{1}{N_i}\sum_{k=1}^{N_i} \ln[f(x_k \mid m_{i/k}, C_{i/k}(\alpha_i))]$$

where $C_{i/k}(\alpha_i) = \begin{cases} (1-\alpha_i)diag(\Sigma_{i/k}) + \alpha_i \Sigma_{i/k} & 0 \le \alpha_i \le 1 \\ (2-\alpha_i)\Sigma_{i/k} + (\alpha_i - 1)S_{i/k} & 1 < \alpha_i \le 2 \\ (3-\alpha_i)S_{i/k} + (\alpha_i - 2)diag(S_{i/k}) & 2 < \alpha_i \le 3 \end{cases}$

The mean of class i, without sample k, is

$$m_{i/k} = \frac{1}{N_i - 1}\sum_{\substack{j=1 \\ j\neq k}}^{N_i} x_{i,j}$$

where the notation */k* indicates the quantity is computed without sample k. The sample covariance of class i, without sample k, is

$$\Sigma_{i/k} = \frac{1}{N_i - 2}\sum_{\substack{j=1 \\ j\neq k}}^{N_i} (x_{i,j} - m_{i/k})(x_{i,j} - m_{i/k})^T$$

and the common covariance, without sample k from class i, is

$$S_{i/k} = \left(\frac{1}{L}\sum_{\substack{j=1 \\ j\neq i}}^{L} \Sigma_j\right) + \frac{1}{L}\Sigma_{i/k}$$

The model of BLOOC is

$$\hat{\Sigma}_i(\alpha_i) = \begin{cases} (1-\alpha_i)\dfrac{tr(S_i)}{p} I + \alpha_i S_i & 0 \le \alpha_i \le 1 \\ (2-\alpha_i)S_i + (\alpha_i - 1)S_p^*(t) & 1 \le \alpha_i < 2 \\ (3-\alpha_i)S + (\alpha_i - 2)\dfrac{tr(S)}{p} I & 2 < \alpha_i \le 3 \end{cases}$$

The pooled covariance matrices are determined under a Bayesian context and can be represented as:

$$S_p^*(t) = \left[\sum_{i=1}^{L} \frac{f_i}{f_i + t\cdot p - 1}\right]^{-1}\left[\sum_{i=1}^{L} \frac{f_i S_i}{f_i + t\cdot p - 1}\right], \quad t = \frac{(\alpha_i - 1)f_i - \alpha_i(p+1)}{2 - \alpha_i}$$

and $f_i = N_i - 1$, which represents the degree of freedom in Wishart distributions.

The first difference between LOOC and BLOOC is that LOOC uses the diagonal entries of covariance matrices but BLOOC, like RDA, uses the trace of covariance matrices. Second, in LOOC, the maximum likelihood common

covariance estimator is used, but, in BLOOC, the maximum a posterior common covariance estimator (S_p^*) is added. From [5], S_p^* tends to mitigate the outlier problem, and so does BLOOC. The choosing mixing parameter method of BLOOC is the same as that of LOOC.

II. MIXED-LOOCS

LOOC and BLOOC are the linear combination of two of the three matrices, and in some situations, BLOOC is better than LOOC, elsewhere LOOC is better. The difference between LOOC and BLOOC is in those matrices that are used to formulate the regularized covariance estimator. So we know that only using some of the six matrices will not get good results in all situations. The basic idea of Mixed-LOOC is to use all six matrices to gain the advantages of both LOOC and BLOOC. Hence the first proposed regularized covariance estimator, Mixed-LOOC1, is

$$\hat{\Sigma}_i(a_i,b_i,c_i,d_i,e_i,f_i) = a_i\frac{tr(S_i)}{p}I + b_i diag(S_i) + c_i S_i + d_i\frac{tr(S)}{p}I + e_i diag(S) + f_i S$$

where $a_i + b_i + c_i + d_i + e_i + f_i = 1$ and $i = 1,2,...,L$

L : number of classes

p : number of dimensions

The mixing parameters are determined by maximizing the average leave-one-out log likelihood of each class:

$$LOOL_i = \frac{1}{N_i}\sum_{k=1}^{N_i}\ln[f(x_k\,|m_{i/k},\hat{\Sigma}_{i/k}(\Omega_i))],\ \text{where } \Omega_i = (a_i,b_i,c_i,d_i,e_i,f_i)$$

Since using Mixed-LOOC1 is computationally intensive, finding a more simplified estimator will be more practical. Reference [4] shows that given two known matrices, the ML (not Leave-One-Out) estimate of mixture parameters in LOOC and BLOOC are at the end points (λ_i =0, 1, 2, or 3), and when the ML covariance estimator is singular, the optimal choice of LOOC parameter under LOOL criteria is around the boundary points.

The Mixed-LOOC2 is proposed as the following form:

$$\hat{\Sigma}_i(\lambda_i) = \lambda_i A + (1-\lambda_i)B$$

where $A = \frac{tr(S_i)}{p}I$, $diag(S_i)$, S_i, $\frac{tr(S)}{p}I$, $diag(S)$, or S

and B= S_i or *diag(S)* and λ_i is close to 1. B= S_i or *diag(S)* was chosen because if a class sample size is large, S_i will be a better choice. If total training sample size is less than the dimensionality then the common (pooled) covariance S is singular but has much less estimation error than S_i. For reducing estimation error and avoiding singularity, *diag(S)* will be a good choice. The selection criteria is the log leave-one-out likelihood function:

$$LOOL_i = \frac{1}{N_i}\sum_{k=1}^{N_i}\ln[f(x_k\,|m_{i/k},\hat{\Sigma}_{i/k}(\Omega_i))]$$

III. EXPERIMENT DESIGN AND RESULTS

In the all experiments, the grid method is used to estimate the mixture parameters of LOOC and Mixed-LOOC1. The range of the parameter λ in LOOC is from 0 to 3 and the grids are λ = [0, 0.25, 0.5,…, 2.75, 3]. There are six parameters in Mixed-LOOC1 and the ranges of them are from 0 to 1. The grids of Mixed-LOOC1 are [0, 0.25, 0.5, 0.75, 1]. For Mixed-LOOC2, the parameter λ is set to 0.05. In the simulation experiments, performances of all three covariance estimators are compared. Based on computational consideration, only the performances of LOOC and Mixed-LOOC2 are compared for the real data experiments.

Experiments 1 to 12 are based on simulated data sets. Experiments 1 to 6 and experiments 7 to 12 are generated from the same normal distributions respectively. The mean vectors and covariance matrices of experiments 1 to 6 (and 7 to 12) are the same as those six experiments in [1]. The only difference between these two set experiments is that experiment 1 to 6 are with equal training sample sizes in each class but experiments 7 to 12 are with different sample sizes in each class. Training and testing sample sizes of these experiments are in Table 2. There are three different dimensionalities, p=10, 30, 60, in every experiment. At each situation, 10 random training and testing data sets are generated for computing the accuracies of algorithms, and the standard deviations of the accuracies.

Table 1 The Design of Sample Size

	Experiments 1 ~ 6			Experiments 7 ~ 12		
Sample Size	Class 1	Class 2	Class 3	Class 1	Class 2	Class 3
Training	10	10	10	30	10	5
Testing	200	200	200	600	200	100

There are four different real data sets, the Cuprite site, which is an area of geologic interest, Jasper Ridge, an ecological site, Indian Pine, an agricultural/forestry site, and DC Mall, an urban site, in experiment 13 to 16 respectively. All real data sets have 191 bands. There are 8, 6, 6, and 7 classes used in the Cuprite Site, Jasper Ridge Site, Indian Pine Site, and DC Mall, respectively. There are 20 training samples in each class. At each experiment, 10 training and testing data sets are selected for computing the accuracies of algorithms, and the standard deviations of the accuracies.

The simulated data results are displayed in Table 2(a), 2(b), and 2(c). The real data results are displayed in Table 2(d). The shadowed parts indicate that the differences of performances of LOOC and Mixed-LOOC2 are larger than the standard deviation of Mixed-LOOC2. If the difference is smaller than the standard deviation, we assume that the

performances of these methods have no significant difference. All the experiments with significant differences indicate that Mixed-LOOC outperformed LOOC. Significant differences most often occurred in experiments 2, 7, and 8. Those are the situations in which BLOOC has better performances than LOOC. Since the Mixed-LOOCs are the union version of LOOC and BLOOC, based on these findings, we conclude that the Mixed-LOOCs have advantages over LOOC and BLOOC.

Table 2(a) The Accuracy of Simulated Data Sets (p=10)

Experiment	LOOC	Mixed-LOOC1	Mixed-LOOC2
1	0.8630 (0.0425)	0.8632 (0.0243)	0.8602 (0.0466)
2	0.7753 (0.0481)	0.8373 (0.0180)	0.8450 (0.0224)
3	0.8948 (0.0241)	0.8915 (0.0251)	0.8992 (0.0265)
4	0.8875 (0.0309)	0.8893 (0.0263)	0.8837 (0.0386)
5	0.9860 (0.0283)	0.9822 (0.0361)	0.9858 (0.0282)
6	0.9885 (0.0033)	0.9833 (0.0085)	0.9885 (0.0036)
7	0.8500 (0.0286)	0.8622 (0.0252)	0.8641 (0.0249)
8	0.8433 (0.0410)	0.8750 (0.0289)	0.8792 (0.0250)
9	0.9021 (0.0230)	0.9041 (0.0183)	0.9041 (0.0203)
10	0.8928 (0.0247)	0.8948 (0.0204)	0.8940 (0.0245)
11	0.9883 (0.0064)	0.9920 (0.0041)	0.9872 (0.0065)
12	0.9841 (0.0076)	0.9830 (0.0075)	0.9827 (0.0116)

Table 2(b) The Accuracy of Simulated Data Sets (p=30)

Experiment	LOOC	Mixed-LOOC1	Mixed-LOOC2
1	0.8317 (0.0227)	0.8285 (0.0196)	0.8267 (0.0213)
2	0.7263 (0.0510)	0.8700 (0.0205)	0.8813 (0.0204)
3	0.8162 (0.0220)	0.8142 (0.0223)	0.8152 (0.0237)
4	0.7978 (0.0619)	0.7955 (0.0609)	0.7972 (0.0612)
5	0.9993 (0.0014)	0.9975 (0.0037)	0.9993 (0.0014)
6	0.9990 (0.0021)	0.9945 (0.0087)	0.9992 (0.0016)
7	0.8239 (0.0345)	0.8469 (0.0154)	0.8504 (0.0171)
8	0.8718 (0.0311)	0.9210 (0.0130)	0.9189 (0.0118)
9	0.8228 (0.0274)	0.8343 (0.0206)	0.8241 (0.0268)
10	0.8326 (0.0162)	0.8370 (0.0186)	0.8313 (0.0156)
11	0.9976 (0.0021)	0.9994 (0.0008)	0.9984 (0.0018)
12	0.9953 (0.0059)	0.9991 (0.0007)	0.9978 (0.0047)

Table 2c) The Accuracy of Simulated Data Sets (p=60)

Experiment	LOOC	Mixed-LOOC1	Mixed-LOOC2
1	0.7378 (0.0540)	0.7607 (0.0259)	0.7605 (0.0287)
2	0.6578 (0.0631)	0.8792 (0.0213)	0.8882 (0.0175)
3	0.7632 (0.0265)	0.7615 (0.0235)	0.7583 (0.0281)
4	0.7483 (0.0324)	0.7473 (0.0308)	0.7435 (0.0288)
5	1.0000 (0.0000)	0.9998 (0.0005)	1.0000 (0.0000)
6	1.0000 (0.0000)	1.0000 (0.0000)	1.0000 (0.0000)
7	0.7820 (0.0327)	0.8098 (0.0229)	0.8120 (0.0192)
8	0.8876 (0.0219)	0.9401 (0.0075)	0.9400 (0.0073)
9	0.7947 (0.0216)	0.8024 (0.0150)	0.7958 (0.0203)
10	0.7802 (0.0302)	0.7932 (0.0277)	0.7837 (0.0275)
11	0.9988 (0.0021)	0.9997 (0.0011)	0.9997 (0.0011)
12	1.0000 (0.0000)	1.0000 (0.0000)	1.0000 (0.0000)

Table 2(d) The Accuracy of Real Data Sets (p=191)

Real Data	LOOC	Mixed-LOOC2
Cuprite	0.7743 (0.1372)	0.9524 (0.0117)
Jasper Ridge	0.9864 (0.0042)	0.9849 (0.0019)
Indian Pine	0.7612 (0.0127)	0.7625 (0.0144)
DC Mall	0.7831 (0.0455)	0.7858 (0.0431)

IV. DAFE BASED ON MIXED-LOOC

Usually Discriminant Analysis Feature Extraction (DAFE) uses the ML covariance estimator of each class. When singular or nearly singular situations occur, ML covariance estimator could be replaced by regularized covariance estimator.

For convenience, denote DAFE based on ML estimators as DAFE and DAFE based on Mixed-LOOC2 as DAFE-Mix2, Gaussian classifier based on ML estimators as GC, and Gaussian classifier based on Mixed-LOOC2 estimators as GC-Mix2. Experiments 17 to 19 are for determining the performances of DAFE-Mix2. The classification process in experiment 17 is to use DAFE then GC, in experiment 18 use DAFE-Mix2 then GC, and in experiment 19 use DAFE-Mix2 then GC-Mix2. The class sample sizes of experiment 18 and 19 are the same as those of experiments 13 to 16 (N_i=20). Since using those sample sizes in DAFE will cause very poor results, we increase the sample size of each class in Cuprite, Jasper Ridge, Indian Pine, and DC Mall data sets up to 40. The number of features extracted from the original space is set to L-1. The results of those experiments are shown in Fig. 1.

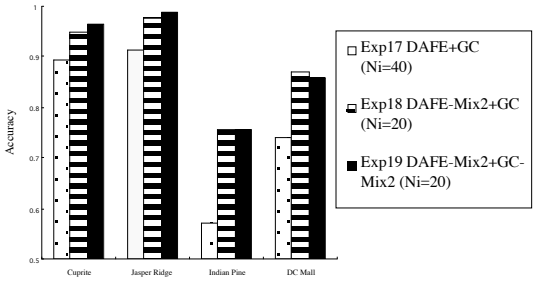

Fig.1. The Mean Accuracies of Experiments 17 to 19

V. COMMENTS

The singularity or near-singularity problem often occurs in the case of high dimensional classification. From the above discussion, we know that finding a suitable regularized covariance estimator is a way to mitigate this problem. Further, Mixed-LOOC2 has advantages over LOOC and BLOOC and needs less computation than those two. Usually DAFE cannot be used when the training sample size is less than dimensionality. The new procedure, DAFE-Mix2, overcomes this shortcoming, and can provide higher accuracy when the sample size is limited.

VI. REFERENCE

[1] J.H. Friedman, " Regularized Discriminant Analysis," *Journal of the American Statistical Association*, vol. 84, pp. 165-175, March 1989

[2] J. P. Hoffbeck and D.A. Landgrebe, " Covariance matrix estimation and classification with limited training data" IEEE Transactions on Pattern Analysis & Machine Intelligence, vol 18, No. 7, pp. 763-767, July 1996.

[3] S. Tadjudin and D.A. Landgrebe, Classification of High Dimensional Data with Limited Training Samples, Purdue University, West Lafayette, IN., TR-EE 98-8, April, 1998, pp35-82.

[4] B-C. Kuo, *Improved Statistics Estimation and Feature Extraction for Hyperspectral Data Classification*. Ph.D. thesis. Purdue University, December 2001

[5] W. Rayens and T. Greene, " Covariance pooling and stabilization for classification." *Computational Statistics and Data Analysis*, vol. 11, pp. 17-42, 1991