An Adaptive Method for Combined Covariance Estimation and Classification

Qiong Jackson and David Landgrebe
School of Electrical & Computer Engineering
Purdue University, West Lafayette IN USA
qiong@ecn.purdue.edu and landgreb@ecn.purdue.edu

An Adaptive Method for Combined Covariance Estimation and Classification[1]

Qiong Jackson and David Landgrebe
School of Electrical & Computer Engineering
Purdue University, West Lafayette IN USA
qiong@ecn.purdue.edu and landgreb@ecn.purdue.edu

***Abstract*-**In this paper a family of adaptive covariance estimators is proposed to mitigate the problem of limited training samples for application to hyperspectral data analysis in quadratic maximum likelihood classification. These estimators are the combination of adaptive classification procedures and regularized covariance estimators. In these proposed estimators, the semi-labeled samples (whose labels are determined by a decision rule) are incorporated in the process of determining the optimal regularized parameters and estimating those supportive covariance matrices that formulate final regularized covariance estimators. In all experiments with simulated and real remote sensing data, these proposed combined covariance estimators achieved significant improvement on statistics estimation and classification accuracy over conventional regularized covariance estimators and an adaptive Maximum Likelihood classifier (MLC). The degree of improvement increases with dimensions, especially for ill-posed or very ill-posed problems where the total number of training samples is smaller than the number of dimensions.

Index Terms-Adaptive iterative classification procedure, regularized covariance estimation, high-dimensional data, semi-labeled samples, hyperspectral data

I. INTRODUCTION

In quadratic maximum likelihood classification, each true class mean vector and covariance matrix are usually unknown and must be estimated by the sample mean and sample covariance matrix based on training samples. When the training sample size is quite small relative to the dimensionality, the sample covariance matrix becomes highly variable and consequently, this greatly decreases the classifier performance. In particular,

---

[1] The work described in this paper was sponsored in part by the U.S. Army Research Office under Grant Number DAAH04-96-1-0444.

when the number of training samples is less than the dimensionality, the sample covariance matrix becomes singular and hence quadratic classifiers cannot be used. This is unfortunate in the analysis of high dimensional data because in the high-dimensional feature space, different classes sharing the same expected values can become separable with very little error, provided that their covariance matrices are sufficiently distinct [1]. Thus, the second-order statistics can assume a preponderant role in remote sensing image data classification, possibly allowing for the separation of classes spectrally close to each other and therefore not separable well by analysis methods that do not take into account second-order statistics (covariance matrices). This poses limitations on the number of dimensions (or features) that can be used in remote sensing applications where training samples are usually small compared to the number of dimensions available. This is especially true for the analysis of hyperspectral data.

One way to deal with this is to employ a linear classifier that is obtained by replacing sample covariance matrices for all classes by their average, S_w. Even if each sample covariance matrix differs greatly, using the average can sometime lead to better performance for small training sets because S_w reduces the number of parameters to be estimated and decreases the variance. This has been verified by several studies [2][3][4]. Even though a linear classifier may perform better than a quadratic classifier in the small training set size case, the choice between these two is a quite sensitive matter. Several more flexible regularized methods have been proposed in which a sample covariance estimate is replaced by partially pooled covariance matrices of various forms, and a varying degree of regularization is applied to control the number of parameters to be estimated and consequently improve the classifier performance based on training samples.

In [5], a regularized procedure referred as " regularized discriminate analysis" (RDA) is proposed, which is a two-dimensional optimization over four-way mixtures as shown in the following:

$$\hat{\Sigma}_i(\lambda,\gamma) = (1-\gamma)\hat{\Sigma}_i(\lambda) + \gamma\left(\frac{tr\left(\hat{\Sigma}_i(\lambda)\right)}{p}\right)I \qquad 0 \le \gamma \le 1 \qquad (1)$$

The pair of regularized parameters (λ,γ) is selected by cross-validating on the total number of misclassifications based on available training samples [5].

In [6], a covariance estimator is proposed which has the following form:

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

where S is the average of all sample covariance matrices. The regularized parameter α_i is determined by maximizing the average leave-one-out log likelihood of each class:

$$LOOL_i = \frac{1}{N}\sum_{k=1}^{N_i} \ln[f(x_{i,k} \mid m_{i/k}, \hat{\Sigma}_{i/k}(\alpha_i)] \tag{3}$$

where N_i is the number of training samples in the class i.

In [7], a covariance estimator is developed which virtually is the combination of RDA, LOOC, and the empirical Bayesian approach. There are two forms of this new covariance estimation depending on the form of covariance matrices used. When the ridge estimator is adopted, the proposed estimator is called as (bLOOC1) and has the following form:

$$\hat{\Sigma}_i(\alpha_i) = \begin{cases} (1-\alpha_i)\dfrac{tr(S_i)}{p}I + \alpha_i S_i & 0 \le \alpha_i \le 1 \\ (2-\alpha_i)S_i + (\alpha_i - 1)S_p^*(\mathrm{t}) & 1 \le \alpha_i < 2 \\ (3-\alpha_i)S + (\alpha_i - 2)\dfrac{tr(S)}{p}I & 2 < \alpha_i \le 3 \end{cases} \tag{4}$$

where the pooled covariance matrices S_p^* are determined under a Bayesian context and can be represented as:

$$S_p^*(t) = \left[\sum_{i=1}^{L} \frac{f_i}{f_i + t - p - 1}\right]^{-1} \sum_{i=1}^{L} \frac{f_i S_i}{f_i + t - p - 1} \tag{5}$$

When the mixture of covariance and covariance-diagonal covariance matrices is used, the proposed estimator is referred as (bLOOC2) and is defined as the following

$$\hat{\Sigma}_i(\alpha_i) = \begin{cases} (1-\alpha_i)diag(S_i) + \alpha_i S_i & 0 \le \alpha_i \le 1 \\ (2-\alpha_i)S_i + (\alpha_i - 1)S_p^*(\mathrm{t}) & 1 \le \alpha_i < 2 \\ (3-\alpha_i)S + (\alpha_i - 2)diag(S) & 2 \le \alpha_i \le 3 \end{cases} \tag{6}$$

The regularized parameters α_i are determined by maximizing average leave-one-out log likelihood.

As an extension of RDA, LOOC, the Empirical Bayesian covariance estimators, bLOOC1 and bLOOC2 have appealing benefits possessed by these methods. For example, like LOOC, bLOOC1 and bLOOC2 are quite flexible on the training sample size. They can deal with a broad range of limited training sample sizes, from well-posed (the number of training sample in each class N_i >> the number of dimensions), to poorly posed ($N_i \sim p$), and ill-posed problems (the number of all training samples $N < p$, where $N = \sum_{i=1}^{L} N_i$), and the regularized parameters are customized for each class.

However, bLOOC1 and bLOOC2 suffer from drawbacks inherited from RDA, LOOC and the Empirical Bayesian covariance estimators. First of all, they have a major disadvantage of having no direct relation with classification accuracy. Most important of all, even though instability of covariance estimates posed by limited training samples can be reduced using a covariance mixture in the aforementioned approaches, the degree of improvement is certainly limited. This is true because supportive matrices and regularized parameters used in the covariance mixture are all based on limited training samples only. In particular, when the training sample sets are so small, for instance ill-posed or very ill-posed, the estimated covariance matrices can be over-tuned to accommodate training samples only and they may not be good representatives of statistics for the entire data.

On the other hand, in [8] we proposed an adaptive iterative quadratic maximum likelihood classification procedure where the limited training samples problem is alleviated by using additional semi-labeled samples to enhance statistics (mean vectors and covariance matrices) estimation. As illustrated in Fig. 1, essentially it is formed by adding a feedback loop (highlighted by dark arrows) to a conventional ML classifier. The classifier starts with the initial classification where only training samples are used to estimate the statistics. After the initial classification, a class label is assigned to the unlabeled samples according to the ML decision rule. Subsequently unlabeled samples become semi-labeled samples, because class label information is partially obtained. At the following iteration, semi-labeled samples together with the training samples are used

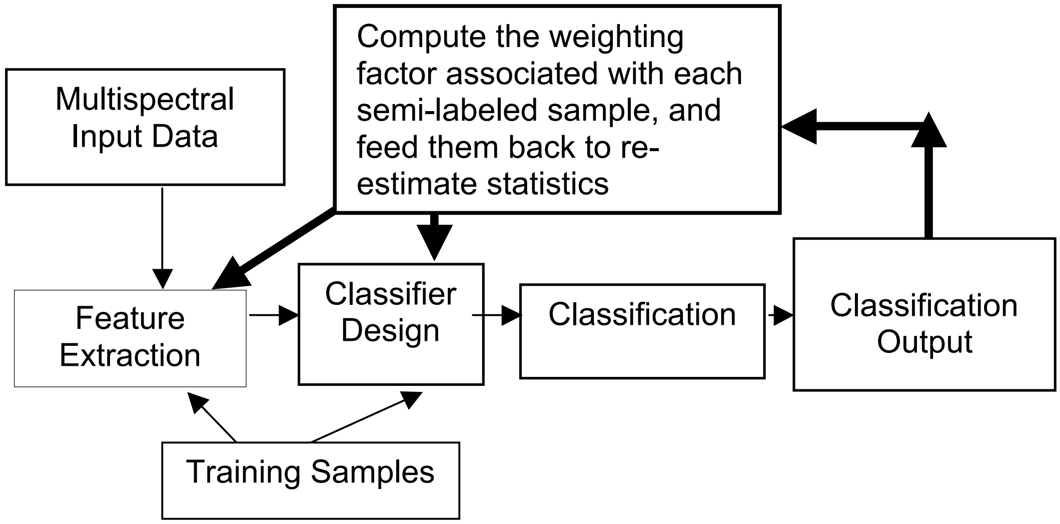

Fig. 1 An adaptive classification procedure.

to re-estimate the statistics. To control the influence of each semi-labeled sample for statistics estimation, a full weight is assigned to a training sample, and reduced weight is assigned to a semi-labeled samples. We have shown that in an adaptive classifier started with a reasonably good initial accuracy achieved by using training samples only, a positive feedback process can be established where semi-labeled samples can provide additional useful class label information and, when they are used, the estimation of statistics can be enhanced and the classification accuracy can be improved. In return, the class label information from semi-labeled samples can be further enhanced in the later stages when better statistics estimation and higher classification accuracy are achieved. However, when the number of dimensions is very high (up to a few hundreds), the number of parameters in the covariance matrix estimation process increases dramatically (approximate to the square of the dimensions). In such cases, using additional semi-labeled samples alone may not be sufficient to reduce the variance of covariance estimation.

In this paper, a method of combining the adaptive quadratic classifier and regularized covariance estimation is proposed to alleviate the extremely small training sample problem in the analysis of hyperspectral data in general, and in particular to deal with ill-posed and very ill-posed problems. Depending on the method of selecting support covariance matrices and the regularization parameters, a group of new adaptive covariance estimators are then introduced. The regularized parameters and supportive covariance matrices used in a covariance mixture are determined based on both training samples and semi-labeled samples, and they are repeatedly updated until the highest classification accuracy is reached.

Extensive experiments are performed using simulated data and real, aircraft-acquired hyperspectral data. With simulated data, the experimental results indicate the proposed adaptive covariance estimators can achieve equivalent classification performance with a small training sample size to that obtained using large training sample size. With hyperspectral data, the proposed adaptive covariance estimators can improve the classification performance dramatically with limited training samples. In all experiments the proposed methods outperform the conventional covariance estimators (LOOC and BLOOC) and the adaptive quadratic classifier [8].

II. ADAPTIVE COVARIANCE ESTIMATORS

A new method is developed in the section that combines an adaptive classifier with various regularized covariance estimation methods, i.e., LOOC, bLOOC1 and bLOOC2. As an adaptive classifier, this method is an iterative approach, i.e., initially the regularized covariance matrices are determined by using training samples only (in other words, this method starts with the conventional covariance estimators, either LOOC or BLOOC), and then they are continuously updated using currently updated semi-labeled samples in addition to training samples until a convergence is reached where the classification outputs change very little. The scheme is shown in Fig. 2, where $Y_{ij} = (y_{i1},\ldots,y_{im_i})$ are the training samples for the i^{th} class, whose pdf is $f_i(y \mid \square_i)$, and $X_{ij} = (x^c_{i1},\ldots,x^c_{in_i})$ are the current semi-labeled samples that have been classified to belong to the i^{th} class.

Depending on the covariance estimator with which the adaptive classifier is

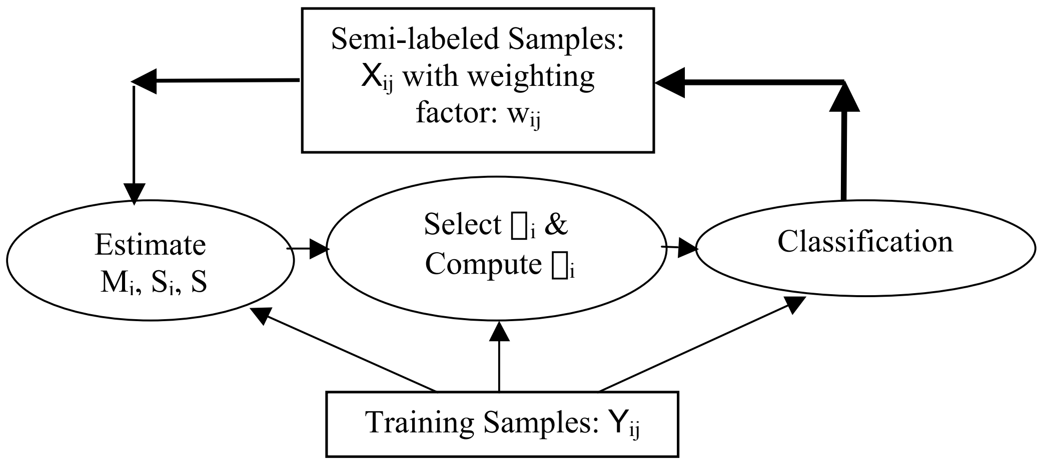

Fig. 2. Flow chart of an adaptive covariance estimator procedure.

combined, the proposed estimators have various forms. Adaptive LOOC is the combination of the adaptive classifier with LOOC, and adaptive BLOOC is the combination of the adaptive classifier with BLOOC. The optimal value of the regularized parameter α_i is determined by maximizing the leave-one-out mixed likelihood,

$$LOOL_i = \frac{1}{m_i + \sum_{k=1}^{n_i} w_{ik}} \left[\sum_{k=1}^{m_i} \ln(f(y_{ik} \mid \mu_{i/k}, C_{i/k}(\alpha_i))) + \sum_{k=1}^{n_i} w_{ik} \ln(f(x_{ik} \mid \mu_{i/k}, C_{i/k}(\alpha_i))) \right]$$

Here w_{ik} is the weighting factor associated with the k^{th} semi-labeled sample from class i, which is related to the likelihood of this sample belonging to i class. Since class label information of semi-labeled samples is updated at each iteration, any improvement of classification accuracy is reflected in the labels of semi-labeled samples. Hence, when semi-labeled samples are used to determine the optimal value of α_i, this criterion is more directly related to the classification accuracy, alleviating the disadvantage of this criterion aforementioned.

The direct implementation of the leave-one-out likelihood function for each class with n_i training samples and m_i semi-labeled samples would require the computation of (n_i+m_i) matrix inverses and determinants at each value of α_i. Fortunately, a more efficient implementation can be derived using the rank-one down of the covariance matrix [9]. In addition, the computation of optimality can be further simplified if one assumes that

$$diag(S) \approx diag(S_{i/k})$$

in the adaptive LOOC or adaptive bLOOC2 estimators and the approximation of

$$\frac{tr(S_{i/k})}{p} I \approx \frac{tr(S)}{p} I$$

in the adaptive bLOOC1 estimator. For notational purposes, in the following sections and experiments, the adaptive LOOC, bLOOC2, and bLOOC1 without approximation is denoted as ALOOC-exact (Adaptive Leave One Out Covariance Estimation), AbLOOC2-exact, and AbLOOC1-exact (Adaptive Bayesian Leave One Out Covariance Estimation), respectively, whereas the implementation with approximation is designated as ALOOC, AbLOOC2, and AbLOOC1, respectively.

III. EXPERIMENTAL RESULTS

A. Experiments with Simulated Data

In this section, the experimental results from computer-generated data are presented. Six proposed covariance estimates, namely, ALOOC, ALOOC–Exact, AbLOOC1, AbLOOC1-Exact, AbLOOC2, and AbLOOC2-Exact are used. Six experiments presented in [10] were performed, and two of them are presented in this paper. Results of the other experiments are contained in [9]. The dimensions p are chosen to be 10, 40, and 60, which represents poorly posed (p=10), ill-posed (p=40) and very ill-posed (p=60) problems, respectively. To benchmark the performance of these covariance estimators, all labeled samples (1000/class) are used as training samples, and the accuracy is called supervised learning. Also, the pseudoinverse of a covariance matrix is also used to replace the inverse of a covariance matrix in the quadratic decision rule when it is singular. Notice here all accuracies are obtained by using an independent generated data set (10,000/class) as a testing data set. Hence they are called hold out accuracies [1].

In experiment 1, all three classes have the identity covariance matrix. The mean of the first class is at the origin. The mean of the second class is taken to be 3.0 in the first variable and zero in the others, and the mean of the third class is 3.0 in the second variable and zero in the rest. The mean accuracy is plotted in Figure 3.

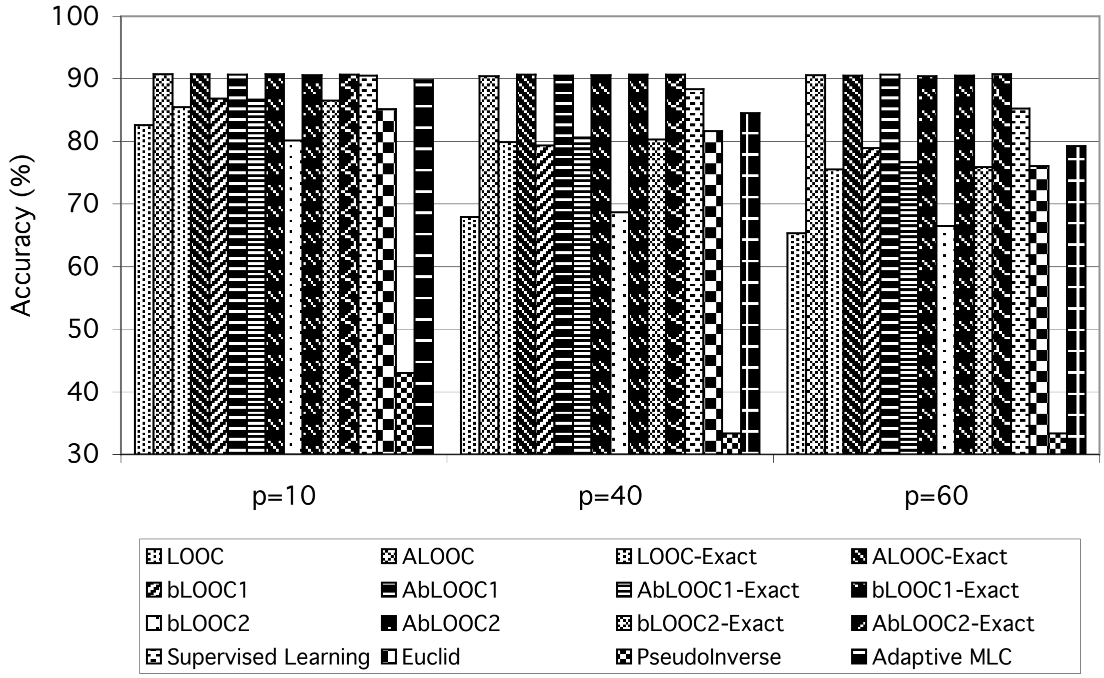

Fig. 3 Mean accuracy for experiment 1.

The following results may be observed: 1) at all dimensions, the adaptive covariance estimators outperform the conventional covariance estimators, the adaptive MLC [8], and Euclidean distance classifier where only mean vectors are used on the decision rule, and the improvement increases with dimensions. 2) At higher dimensions (p=40 and 60), the adaptive covariance estimators achieves higher accuracy than the one supervised sample covariance estimators where a large number of training samples were used. 3) Even though the performance of the conventional covariance estimators LOOC and BLOOC declines dramatically with increasing dimensionality due to the small training sample problem, or Hughes phenomenon [11], the performance of the adaptive covariance estimators varies very little. The mean accuracies obtained by these approaches are very close to the optimal value. This indicates the Hughes phenomenon has been greatly alleviated. 4) Furthermore, the standard deviation (not shown here, but available in [9]) is reduced by about 10-50 fold, which indicates the final estimated statistics are more representative of the true ones. 5) All adaptive covariance estimators yield similar final classification accuracies, indicating their performance is comparable. This means that computational time can be reduced greatly by using the approximated version of these approaches. 6) The pseudoinverse approach has the worst performance among all methods.

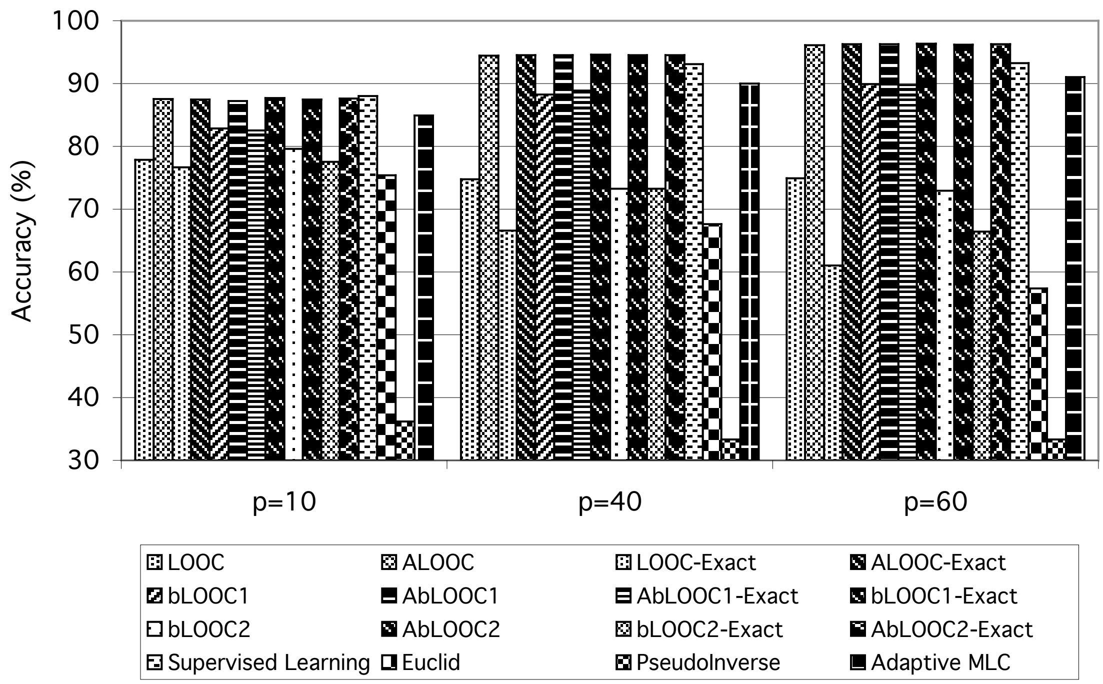

Fig. 4 Mean accuracy for experiment 2.

In experiment 2, all three classes have different spherical covariance matrices and different mean vectors. The covariance of class one, two, and three are I, 2I, and 3I, respectively, where I represents the identity matrix. The mean vectors are the same as those in experiment 1. The mean classification accuracy for each estimator is graphed in Fig. 4. Here similar results can be seen as those from experiment 1. In addition, note that contrary to the first experiment, for this data set, the separability of classes due to the second order statistics increases with dimensions. This suggests the potential for dramatic improvement of accuracy as long as the class statistics can be estimated precisely in the high dimension space. Except for the methods bLOOC1 and bLOOC-exact, the performance of the other four conventional covariance estimators and the Euclidean classifier deteriorate drastically when the number of dimensional increases. However, the final accuracy from all adaptive covariance estimators increases with dimensions, and values of the final accuracy are quite close and much higher than those obtained by the conventional covariance estimators. This demonstrates the preponderant role of the second order statistics and the importance of precisely estimating them in the analysis of high dimensional data.

B. Experiments with real Hyperspectral data

In the following experiments, three hyperspectral data sets that represent different scenes, i.e., geological, ecological and agricultural, are used. The detailed information about these data sets is shown in Table 1, and classification results are illustrated in Table 2. It took substantial effort to label the large number of samples for these data sets. For example, those in Exp. 3 were gathered by visual comparison of the remotely sensed spectra to a library of laboratory reflectance spectra [10]. Those in Exp. 4 were also visually identified; while those in Exp. 5 were from an available ground truth map. All these processes are quite time consuming (more than a few hours for visually identification) and tedious. By comparison, to collect the small number of training samples is a far easier task.

Table 1. Data information for experiment 3, 4, and 5

	Sites	Bands	Labeled Samples	Total Training samples
Exp. 3	Cuprite, Navada. (Geological)	191	2744	28
Exp. 4	Jasper Ridge, California. (Ecological)	193	3207	16
Exp. 5	Indian Pine, Indiana. (Agricultural)	191	2521	25

In all these experiments, the total number of training samples is much smaller than the dimensions; they all represent very-ill posed problems. It is seen that similar to the results from simulated data, the adaptive covariance estimators outperform substantially the conventional covariance estimators and Euclidean classifiers. For experiment 3 and 4, some adaptive covariance estimators achieve the accuracy very close to the resubstitution one, for example, ALOOC-exact and AbLOOC2 in experiment 3, and ALOOC-exact and AbLOOC2-Exact in experiment 4. Again, the Pseudoinverse method has the poorest performance.

Table 2. Mean Classification Accuracy (%) Experiment 3, 4 and 5

Accuracy	Overall Mean Accuracy (%)		
	Experiment 3	Experiment 4	Experiment 5
LOOC	74.6(17.2)	91.6(3.1)	52.4(3.7)
ALOOC	90.2(7.9)	97.4(1.2)	67.8(5.1)
Difference	**15.2(-9.3)**	**5.7(-1.8)**	**15.4(1.4)**
LOOC-Exact	82.1(5.8)	92.8(2.2)	67.2(4.1)
ALOOC-Exact	93.2(2.6)	97.2(1.7)	74.1(4.7)
Difference	**11.14(-3.18)**	**4.39(-0.5)**	**6.87(0.61)**
bLOOC2	78.8(3.7)	88.1(5.5)	52.9(9.1)
AbLOOC2	94.1(3.7)	95.7(4.6)	70.8(7.5)
Difference	**15.3(-0.0)**	**7.61(-0.9)**	**17.8(-1.6)**
bLOOC2 -Exact	80.2(4.1)	90.7(9.9)	64.5(4.5)
AbLOOC2 -Exact	91.5(5.3)	98.2(1.1)	72.6(6.1)
Difference	**11.35(1.17)**	**7.47(-8.81)**	**8.0(1.6)**
Resubstitution Accuracy	94.0 (2.7)	98.1(3.5)	94.0 (2.3)
Euclid	40.8(5.6)	95.4(1.5)	65.5 (6.5)
PseudoInverse	26.57 (1.0)	66.57(3.2)	20.63(0.5)
Adaptive MLC	80.2(10.0)	90.4(2.3)	66.4(6.3)

IV. CONCLUSION

A new family of adaptive covariance estimators is presented which is produced by combining an adaptive classification process with various regularized covariance estimators, i.e., LOOC, bLOOC1 and bLOOC2. They are proposed as a means to mitigate small training sample problems in general, and in particular, for the poorly or ill-posed problem where for high dimension data the number of training samples is comparable to the number of features or where the sum of all training samples is even smaller than the number of features. A set of experiments on simulated data and real hyperspectral data are performed and reported. Additional such experiments are described in [9].

For simulated and real data, the proposed adaptive covariance estimators offer similar performance. They all outperform conventional regularized covariance estimators, and the Euclidean classifier. In addition, the improvement of performance increases with dimensionality. They also appear more robust against variations in training sets as indicated by the decreased standard deviation among the repeated test trials for most of experiments.

In conclusion, the proposed adaptive covariance estimators have the advantage of both an adaptive classifier and a regularized covariance estimator and are able to produce higher classification accuracy than either used alone. This method is also robust in the sense that from all experiments performed where training samples are randomly selected, the mean classification accuracy has been improved and for most of them the standard deviation of multiple trials has been reduced.

REFERENCES

[1] K. Fukunaga, " Introduction to Statistical Pattern Recognition", second edition, Academic Press, 1990.

[2] S.P.Lin and M.D. Perlman, " A Monte Carlo comparison of four estimators of a covariance matrix," Multivariate analysis—VI : Proceedings of the Sixth International Symposium on Multivariate Analysis, P.R. Krishnaiah, ed., Amsterdam: Elsevier Science Pub. Co., 1985, pp. 411-429.

[3] P.W. Wahl and R.A. Kronmall, " Discriminant Functions when Covariances are Equal and Sample Sizes are Moderate," *Biometrics*, vol. 33, pp. 479-484, 1977.

[4] S. Marks and O.J. Dunn, " Discriminant Functions when the Covariance Matrices are unequal," *Journal of the American Statistical Association*, vol., 69, pp. 555-559,1974.

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

[6] 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*.

[7] Saldju Tadjudin and David Landgrebe, "Covariance Estimation With Limited Training Samples," *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 37, No. 4, pp. 2113-2118, July 1999*.

[8] Qiong Jackson and D. Landgrebe, "An Adaptive Classifier Design for High-Dimensional Data Analysis with a Limited Training Data Set," *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 39, No. 12, pp. 2664-2679, December 2001*.

[9] Qiong Jackson and David Landgrebe, Design Of An Adaptive Classification Procedure For The Analysis Of High-Dimensional Data With Limited Training Samples, PhD Thesis and School of Electrical & Computer Engineering Technical Report TR-ECE 01-5, December 2001*.

[10] J.P.Hoffbeck and D.A. Landgrebe, Classification of High Dimensional Multispectral Data, Purdue University, West Lafayette, IN., TR-EE 95-14, pp.43-71, May, 1995*.

[11] G. F. Hughes, " On the mean accuracy of statistical pattern recognition", *IEEE Trans. Information Theory*, Vol. IT-14, No. 1, pp 55-63, 1968.

* Available for download from http://dynamo.ecn.purdue.edu/~landgreb/publications.html