Covariance Matrix Estimation and Classification with Limited Training Data[1]

Joseph P. Hoffbeck
AT&T Bell Laboratories
67 Whippany Road
Whippany, NJ 07981
joe@fuwutai.att.com

David A. Landgrebe*
School of Electrical & Computer Engineering
Purdue University
West Lafayette, IN 47907-1285
landgreb@ecn.purdue.edu

Reprinted from IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 18, No. 7, pp. 763-767, July 1996.

1 Work leading to the paper was funded in part by NASA Grants NAGW-925 and NAGW-3924.

* Corresponding author

Covariance Matrix Estimation and Classification with Limited Training Data

Joseph P. Hoffbeck
AT&T Bell Laboratories

David A. Landgrebe
School of Electrical Engineering
Purdue University

Abstract

A new covariance matrix estimator useful for designing classifiers with limited training data is developed. In experiments, this estimator achieved higher classification accuracy than the sample covariance matrix and common covariance matrix estimates. In about half of the experiments, it achieved higher accuracy than regularized discriminant analysis, but required much less computation.

I. Introduction

When classifying data with the Gaussian maximum likelihood classifier, the mean vector and covariance matrix of each class usually are not known and must be estimated from training samples. For p-dimensional data, the sample covariance matrix estimate is singular, and therefore unusable, if fewer than p+1 training samples from each class are available, and it is a poor estimate of the true covariance matrix unless many more than p+1 samples are available. In some applications, such as remote sensing, there are often a large number of features available, but the number of training samples is limited due to the difficulty and expense in labeling them. Since inaccurate estimates of the covariance matrix lead to lowered classification accuracy, having too few training samples can be a major impediment in using the Gaussian maximum likelihood classifier to classify high dimensional data. When the number of training samples is limited, estimating the mean vector for each class, but using a common covariance matrix estimate for all the classes, can sometimes lead to higher accuracy because it reduces the number of parameters to be estimated.

The proposed covariance matrix estimator examines mixtures of the sample covariance matrix, common covariance matrix, diagonal sample covariance matrix, and diagonal common covariance matrix. Whereas the maximum likelihood estimator maximizes the joint likelihood of all the training samples, the proposed covariance matrix estimator selects the mixture that maximizes the likelihood of training samples not included in the covariance matrix estimation.

The estimator is defined in this paper, and an efficient implementation that incorporates an approximation is derived. The results of several experiments are presented that compare the estimator, with and without the approximation, to the sample covariance matrix estimate, common covariance matrix, Euclidean distance, and regularized discriminant analysis (RDA). With the approximation, the proposed estimator usually led to higher classification accuracy than the sample estimate, common covariance matrix, and Euclidean distance. In about half of the experiments, it led to higher accuracy than RDA, but required much less computation. Without the approximation, the proposed estimator led to even higher accuracy in some cases, but required significantly more computation.

II. Gaussian Maximum Likelihood Classification

The decision rule in a Gaussian maximum likelihood classifier is to label the (p by 1) vector x as class j if the likelihood of class j is the greatest among the classes:

$$\text{Choose } \omega_j \text{ if } \arg\max_i \left[f(x|m_i, \Sigma_i) \right] = j \qquad (1)$$

where $f(x|m_i, \Sigma_i) = \frac{1}{\sqrt{(2\pi)^p |\Sigma|}} \exp\left\{ \frac{-1}{2} (x - m_i)^T \Sigma_i^{-1} (x - m_i) \right\}$, m_i is the mean vector of class i, and Σ_i is the covariance matrix. Usually the true values of the mean and covariance matrix are not known and must be estimated from training samples. The mean is typically estimated by the sample mean $m_i = \frac{1}{N_i} \sum_{j=1}^{N_i} x_{i,j}$, where $x_{i,j}$ is sample j from class i, and N_i is the number of training samples from class i. The covariance matrix is typically estimated by the sample covariance matrix

$\Sigma_i = \frac{1}{N_i - 1} \sum_{j=1}^{N_i} (x_{i,j} - m_i)(x_{i,j} - m_i)^T$, or the maximum likelihood covariance matrix estimate $\Sigma_i^{ML} = \frac{1}{N_i} \sum_{j=1}^{N_i} (x_{i,j} - m_i)(x_{i,j} - m_i)^T$.

The sample mean and the maximum likelihood covariance matrix estimate have the property that they maximize the joint likelihood of the training samples, which are assumed to be statistically independent (see, for example, [1]): $(m_i, \Sigma_i) = \arg\max_{m,\Sigma} \prod_{j=1}^{N_i} f(x_{i,j} | m, \Sigma)$.

The classification rule that results from substituting the maximum likelihood estimates for the mean and covariance matrix into Equation (1) as if they were the true mean and covariance matrix, achieves optimal classification accuracy only asymptotically as the number of training samples increases toward infinity. This classification scheme is not optimal when the training sample is small [2].

When the training set is small, the sample covariance matrix estimate is usually highly elliptical and can vary drastically from the true covariance matrix. In fact when the number of training samples is less than p+1, the sample covariance matrix is always singular regardless of the true value of the covariance matrix.

For limited training data, the common covariance matrix estimate ($S = \frac{1}{L} \sum_{i=1}^{L} \Sigma_i$) can lead to higher accuracy than the sample estimate even when the true covariance matrices are quite different [3]. It is useful, then, to determine whether the sample covariance matrix estimate or the common covariance matrix estimate would be appropriate in a given situation, and this is accomplished by the proposed estimator.

III. Covariance Matrix Estimation

A. Definition of the Covariance Matrix Estimator

Depending on the true class statistics, different covariance matrix estimators are optimal. For example, if the classes all have the same covariance matrix, the common covariance matrix estimate will lead to higher classification accuracy than the sample covariance matrix. Even if the covariance matrix of each class differs greatly, the common covariance matrix estimate can lead to higher classification if the number of training samples is small. Which estimate is best in a given situation depends in a complex fashion on the true statistics of the classes, the number of features, and the number of training samples.

In addition to the sample covariance matrix and common covariance matrix estimates, the proposed estimator examines the diagonal sample covariance matrix, the diagonal common covariance matrix, and some pair-wise mixtures of these estimates to determine which would be most appropriate. The proposed estimator has the following form:

$$C_i(\alpha_i) = \alpha_{i1}\,\mathrm{diag}(\Sigma_i) + \alpha_{i2}\Sigma_i + \alpha_{i3}S + \alpha_{i4}\,\mathrm{diag}(S) \qquad (2)$$

where Σ_i is the sample covariance matrix, the common covariance matrix is defined by the average sample covariance matrix $S = \frac{1}{L}\sum_{i=1}^{L}\Sigma_i$, and L is the number of classes. The elements of the mixing parameter $\alpha_i = [\alpha_{i1}, \alpha_{i2}, \alpha_{i3}, \alpha_{i4}]^T$ are required to sum to unity: $\sum_{j=1}^{4}\alpha_{ij} = 1$. Furthermore, in order to reduce the required computation, we only consider mixtures between the diagonal sample covariance matrix and the sample covariance matrix ($\alpha_{i3}, \alpha_{i4} = 0$), between the sample covariance matrix and the common covariance matrix ($\alpha_{i1}, \alpha_{i4} = 0$), and between the common covariance matrix and the diagonal common covariance matrix ($\alpha_{i1}, \alpha_{i2} = 0$). Next we examine how an appropriate value of α_i can be estimated.

B. Selecting an Appropriate Mixture

The value of the mixing parameter α_i is selected so that a best fit to the training samples is achieved, in the sense that the average likelihood of omitted samples is maximized. The technique is to remove a sample, estimate the mean and covariance matrix from the remaining samples, then compute the likelihood of the sample which was left out, given the mean and covariance matrix estimates. Each sample is removed in turn, and the average log likelihood is computed over all the left out

samples. Mixtures for several different values of α_i are examined, and the value of α_i that maximizes the average log likelihood is selected.

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 i/k indicates the quantity is computed without using sample k from class i. The sample covariance matrix of class i, without sample k, is

$$\Sigma_{i/k} = \frac{1}{N_i - 2} \sum_{\substack{j=1 \\ j \neq k}}^{N_i} \left(x_{i,j} - m_{i/k}\right)\left(x_{i,j} - m_{i/k}\right)^T \qquad (3)$$

and the common covariance matrix, 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 proposed covariance matrix estimate for class i, without sample k, can then be computed as follows.

$$C_{i/k}(\alpha_i) = \alpha_{i1} \operatorname{diag}(\Sigma_{i/k}) + \alpha_{i2} \Sigma_{i/k} + \alpha_{i3} S_{i/k} + \alpha_{i4} \operatorname{diag}(S_{i/k})$$

Next the average log likelihood of the left-out samples, which we call the leave-one-out likelihood (LOOL), is computed as follows.

$$\mathrm{LOOL}_i(\alpha_i) = \frac{1}{N_i} \sum_{k=1}^{N_i} \ln\left[f\left(x_{i,k} \mid m_{i/k}, C_{i/k}(\alpha_i)\right)\right]$$

This computation is repeated for several values of α_i, and the value with the highest average log likelihood is selected. Once the appropriate value of α_i has been estimated, the proposed covariance matrix estimate is computed using all the training samples (Equation (2)) and can be substituted into the maximum likelihood classifier (Equation (1)). Since the LOOL index depends on training samples from only one class, a separate value of the mixing parameter α_i is computed for each class.

Since evaluation of the Gaussian density function requires the inverse of the covariance matrix, an estimate of the covariance matrix is only useful in classification if it is non-singular (i.e. invertible). The sample covariance matrix estimate is singular if there are fewer than p+1 samples available. Since a diagonal matrix is non-singular if its diagonal elements are all non-zero, the proposed estimate is non-singular as long as the sample covariance matrix has non-zero diagonal elements, which is the usual case if there is more than one sample. The division in Equation (3), however, requires

at least three samples in each class. The proposed estimate, then, will usually be non-singular with as few as three training samples per class, regardless of the dimension of the data.

C. Efficient Implementation of the Mixture Between the Sample Covariance Matrix and the Common Covariance Matrix

If implemented directly, the computation of the proposed estimate would require computing the inverse and determinant of the (p by p) matrix $C_{i/k}(\alpha_i)$ for each training sample, which would be quite computationally expensive. Fortunately, a significant reduction in the required computation can be achieved by writing the matrix in a form that allows the determinant and inverse to be computed efficiently. Consider the mixture between the sample covariance matrix and the common covariance matrix ($\alpha_{i1}, \alpha_{i4} = 0$). The sample covariance matrix estimate of class i without sample k can be written as follows [4]:

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

where $v = x_{i,k} - m_i$. The common covariance matrix estimate without sample k from class i can be written as follows.

$$S_{i/k} = \frac{1}{L} \sum_{\substack{j=1 \\ j \neq i}}^{L} \Sigma_j + \frac{1}{L} \Sigma_{i/k} = S + \frac{1}{L(N_i - 2)} \Sigma_i - \frac{N_i}{L(N_i - 2)(N_i - 1)} vv^T$$

Then the proposed estimate becomes:

$$C_{i/k}(\alpha_i) = \alpha_{i2} \Sigma_{i/k} + \alpha_{i3} S_{i/k} = G - k_1 vv^T$$

$$G = \left[\frac{\alpha_{i2}(N_i - 1)}{(N_i - 2)} + \frac{\alpha_{i3}}{L(N_i - 2)} \right] \Sigma_i + \alpha_{i3} S$$

$$k_1 = \left[\frac{\alpha_{i2} N_i}{(N_i - 2)(N_i - 1)} + \frac{\alpha_{i3} N_i}{L(N_i - 2)(N_i - 1)} \right]$$

Then $C_{i/k}(\alpha_i)^{-1}$ can be computed efficiently using the Sherman-Morrison-Woodbury formula [5] as follows.

$$C_{i/k}(\alpha_i)^{-1} = G^{-1} + \frac{k_1 G^{-1} vv^T G^{-1}}{1 - k_1 v^T G^{-1} v}$$

The quadratic term in the Gaussian density function can be written as follows:

$$d_{i/k} = \left(x_{i,k} - m_{i/k}\right)^T C_{i/k}\left(\alpha_i\right)^{-1}\left(x_{i,k} - m_{i/k}\right)$$

$$= \left(\frac{N_i}{N_i - 1}\right)^2 v^T \left[G^{-1} + \frac{k_1 G^{-1} v v^T G^{-1}}{1 - k_1 v^T G^{-1} v}\right] v$$

$$= \left(\frac{N_i}{N_i - 1}\right)^2 \left[\frac{d}{1 - k_1 d}\right]$$

where $d = v^T G^{-1} v$. The determinant can be computed as shown below [4].

$$\left|C_{i/k}\left(\alpha_i\right)\right| = \left|G - k_1 v v^T\right| = |G|\left(1 - k_1 d\right)$$

Finally, the log likelihood function can be computed efficiently as follows.

$$\ln\left[f\left(x_{i,k} \mid m_{i/k}, C_{i/k}\left(\alpha_i\right)\right)\right] = -\frac{p}{2}\ln(2\pi) - \frac{1}{2}\ln\left[|G|\left(1 - k_1 d\right)\right] - \frac{1}{2}\left(\frac{N_i}{N_i - 1}\right)^2 \left[\frac{d}{1 - k_1 d}\right] \quad \alpha_{i1}, \alpha_{i4} = 0$$

Instead of inverting a (p by p) matrix and finding its determinant for every training sample in the class, it is only necessary to compute the inverse and the determinant of matrix G once, and then only a relatively simple computation ($d = v^T G^{-1} v$) is required for each sample.

D. Approximation for the Mixture Between the Diagonal Sample Covariance Matrix and the Sample Covariance Matrix

Unfortunately, there does not seem to be a similar method to avoid inverting a large matrix for each sample in the mixture between the diagonal sample covariance matrix and the sample covariance matrix ($\alpha_{i3}, \alpha_{i4} = 0$). However, if one is willing to accept the approximation that the diagonal covariance matrix changes little when a single sample is removed ($\mathrm{diag}\left(\Sigma_i\right) \approx \mathrm{diag}\left(\Sigma_{i/k}\right)$), a significant reduction in computation can be realized. Experiments presented below confirm the validity of this assumption when a modest number of training samples are available.

$$C_{i/k}\left(\alpha_i\right) \approx \alpha_{i1}\mathrm{diag}\left(\Sigma_i\right) + \alpha_{i2}\Sigma_{i/k} \approx G_2 - k_2 v v^T \qquad (4)$$

$$G_2 = \alpha_{i1}\mathrm{diag}\left(\Sigma_i\right) + \frac{\alpha_{i2}\left(N_i - 1\right)}{\left(N_i - 2\right)}\Sigma_i$$

$$k_2 = \frac{\alpha_{i2} N_i}{\left(N_i - 2\right)\left(N_i - 1\right)}$$

$$d_2 = v^T G_2^{-1} v$$

The log likelihood function can then be computed as follows.

$$\ln\left[f\left(x_{i,k}\,|m_{i/k},C_{i/k}(\alpha_i)\right)\right] \approx -\frac{p}{2}\ln(2\pi) - \frac{1}{2}\ln\left[|G_2|(1-k_2d_2)\right] - \frac{1}{2}\left(\frac{N_i}{N_i-1}\right)^2\left[\frac{d_2}{1-k_2d_2}\right] \quad \alpha_{i3},\alpha_{i4} = 0$$

E. Approximation for the Mixture Between the Common Covariance Matrix and the Diagonal Common Covariance Matrix

The computation of the mixture between the common covariance matrix and the diagonal common covariance matrix ($\alpha_{i1},\alpha_{i2}=0$) can be simplified similarly by assuming the diagonal common covariance matrix changes little when a single sample is removed ($\mathrm{diag}(S) \approx \mathrm{diag}(S_{i/k})$). Experiments presented below confirm the validity of this assumption for moderate sample sizes.

$$C_{i/k}(\alpha_i) \approx \alpha_{i3}S_{i/k} + \alpha_{i4}\mathrm{diag}(S) \approx G_3 - k_3vv^T \tag{5}$$

$$G_3 = \alpha_{i3}\left[S + \frac{1}{L(N_i-2)}\Sigma_i\right] + \alpha_{i4}\mathrm{diag}(S)$$

$$k_3 = \frac{\alpha_{i3}N_i}{L(N_i-2)(N_i-1)}$$

$$d_3 = v^TG_3^{-1}v$$

The log likelihood function can then be computed as follows.

$$\ln\left[f\left(x_{i,k}\,|m_{i/k},C_{i/k}(\alpha_i)\right)\right] \approx -\frac{p}{2}\ln(2\pi) - \frac{1}{2}\ln\left[|G_3|(1-k_3d_3)\right] - \frac{1}{2}\left(\frac{N_i}{N_i-1}\right)^2\left[\frac{d_3}{1-k_3d_3}\right] \quad \alpha_{i1},\alpha_{i2} = 0$$

For convenience, we will designate the estimator resulting from the approximations in Equations (4) and (5) as the leave-one-out covariance matrix estimate (LOOC), and the estimator without these approximations as the LOOC-Exact estimate.

F. Comparison of LOOC and Regularized Discriminant Analysis

Regularized discriminant analysis (RDA) is another covariance matrix estimation method useful for designing classifiers with limited training data. It is a two-dimensional optimization over mixtures of the sample covariance matrix, common covariance matrix, and the identity matrix times a scalar. RDA takes the following form (assuming equal weighting of training samples) [3]:

$$C_i^{RDA}(\lambda,\gamma) = (1-\gamma)(1-\lambda)\frac{S_i^{RDA}}{W_i(\lambda)} + (1-\gamma)\lambda\frac{S^{RDA}}{W_i(\lambda)} + \frac{\gamma}{p}\,\mathrm{trace}\left[\frac{(1-\lambda)S_i^{RDA} + \lambda S^{RDA}}{W_i(\lambda)}\right] I$$

where $S_i^{RDA} = \sum_{j=1}^{N_i}(x_{i,j} - m_i)(x_{i,j} - m_i)^T$, $S^{RDA} = \sum_{i=1}^{L} S_i^{RDA}$, and $W_i(\lambda) = (1-\lambda)N_i + \lambda\sum_{i=1}^{L} N_i$.

The two mixing parameters λ and γ, which are restricted to the range 0 to 1, are selected to maximize the leave-one-out classification accuracy. Since this index depends on the covariance matrix estimates of the other classes, the same values of the mixing parameters are used for all classes.

LOOC and RDA are similar in that they both consider mixtures of covariance matrix estimates, but they differ in the mixtures they consider and the index used to select the best mixture. Both LOOC and RDA employ the sample covariance matrix and common covariance matrix estimates, but LOOC also considers diagonal forms of these matrices, whereas RDA considers the identity matrix multiplied by a scalar. In LOOC the search is restricted to pair-wise mixtures, whereas RDA considers general mixtures. The index maximized in LOOC is the leave-one-out likelihood which allows a separate mixing parameter to be computed for each class. RDA, on the other hand, maximizes the leave-out-out classification accuracy, and is restricted to using the same value of the mixing parameters for all the classes.

LOOC requires much less computation than RDA. For each point on the optimization grid, LOOC requires only one density function be evaluated for each training sample, whereas RDA requires the density function of every class be evaluated. Thus, if there are L classes, RDA requires the evaluation of L times as many density functions. Also, since LOOC requires what is effectively only a one-dimensional optimization and RDA involves a two-dimensional optimization, many more optimization points must be visited with RDA, especially if the optimization is to be done over a fine grid. Finally, RDA requires the computation of the eigenvalues and eigenvectors for a (p by p) matrix for each value of λ, which is not required by LOOC. The LOOC-Exact method, however, requires more computation than RDA.

LOOC is scale invariant, but RDA is not. Thus, scaling features individually by a non-zero constant, which is commonly done before quantizing sensor output to digital values, has no effect on the classification accuracy with the LOOC method, but may affect the accuracy of the RDA method. Unlike LOOC, however, RDA is rotationally

invariant. Subjecting the data to a orthonormal rotation may affect LOOC, but not RDA. Neither LOOC nor RDA is affected by shifting the data by a constant offset.

IV. Experimental Results

Experiments were conducted with both computer generated data and remote sensing data to compare the classification accuracy from LOOC, and LOOC-Exact to that obtained from the sample covariance matrix, common covariance matrix, RDA, and Euclidean distance (equivalent to assuming the covariance matrices are equal to the identity matrix).

A. Computer Generated Data

In the experiments with computer generated data, 15 independent random training samples were drawn from three different normal distributions, the mean and covariance matrix were estimated, and the classification accuracy was measured by classifying 100 independent test samples from each class. Three experiments with various distributions adapted from [3] and four different dimensions were performed. Each experiment was repeated 25 times, and the mean and standard deviation of the classification accuracy were recorded. The results of additional experiments were reported in [6].

In the experiments, the values of the mixing parameters were sampled over a very coarse grid. In LOOC and LOOC-Exact, α_i took the thirteen points listed in Table 1, and in RDA, the values of both λ and γ were (0.0, .25, .50, .75, 1.0), resulting in 25 data points.

Table 1. Values of α_i

α_{i1}	1.0	.75	.50	.25	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
α_{i2}	0.0	.25	.50	.75	1.0	.75	.50	.25	0.0	0.0	0.0	0.0	0.0
α_{i3}	0.0	0.0	0.0	0.0	0.0	.25	.50	.75	1.0	.75	.50	.25	0.0
α_{i4}	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	.25	.50	.75	1.0

In Experiment 1, the covariance matrices of all three classes were equal to the identity matrix, but each class had a different mean vector. The mean of the first class was at the origin, the mean of the second class was 3.0 in the first variable and zero in the other variables, and the mean of the third class was 3.0 in the second variable and zero in the other variables. Table 2 contains the mean accuracy of each classifier and the standard deviation in parentheses. The notation N/A in the Sample Cov row indicates that the sample covariance matrix was singular and could not be inverted in order to classify the test samples.

Table 2. Mean Classification Accuracy for Experiment 1

	p=6	p=10	p=20	p=40
Common Cov	88.1 (2.0)	86.0 (2.5)	76.8 (4.8)	51.2 (5.6)
Sample Cov	79.7 (4.6)	64.4 (6.3)	N/A	N/A
Euclidean	89.8 (1.9)	88.8 (2.3)	86.6 (2.5)	84.1 (2.2)
RDA	89.6 (2.0)	87.8 (2.6)	85.9 (2.7)	82.5 (3.4)
LOOC	87.9 (2.5)	86.1 (2.0)	80.9 (4.4)	76.5 (5.8)
LOOC-Exact	89.1 (2.2)	88.2 (2.4)	85.9 (2.6)	83.1 (3.3)

In Experiment 2, all three classes had identical, highly elliptical covariance matrices, and the primary difference in the mean vectors was in the variables with low variance. The covariance matrix for all three classes was a diagonal matrix whose diagonal elements were given by $\sigma_i = \left[9(i-1)/(p-1)+1\right]^2 \quad 1 \le i \le p$. The mean vector of the first class was at the origin, the elements of the mean vector of the second class were given by $\mu_{2,i} = 2.5\sqrt{\sigma_i / p}\left[(p-i)/(p/2-1)\right]$, and the mean of class three was defined by $\mu_{3,i} = (-1)^i \mu_{2,i}$. See Table 3 for the results.

Table 3. Mean Classification Accuracy for Experiment 2

	p=6	p=10	p=20	p=40
Common Cov	93.3 (2.3)	89.0 (1.9)	78.0 (4.4)	49.3 (6.1)
Sample Cov	88.0 (2.8)	70.5 (6.5)	N/A	N/A
Euclidean	75.8 (4.3)	71.7 (4.7)	64.5 (4.5)	57.0 (3.8)
RDA	92.9 (2.9)	87.8 (4.4)	75.9 (4.9)	61.3 (5.7)
LOOC	93.5 (2.1)	89.4 (2.3)	83.4 (3.3)	75.9 (3.5)
LOOC-Exact	94.2 (2.1)	91.5 (1.7)	87.2 (2.2)	82.9 (2.6)

In Experiment 3, the mean vector of all three classes was at the origin, but the class covariance matrices were quite different and highly elliptical. The diagonal elements of the covariance matrices for each class were defined by $\sigma_{1,i} = \left[9(i-1)/(p-1)+1\right]^2 \quad 1 \le i \le p$, $\sigma_{2,i} = \left[9(p-i)/(p-1)+1\right]^2$, and $\sigma_{3,i} = \left[9\left(i-(p-1)/2\right)/(p-1)\right]^2$. The results of the experiment are listed in Table 4.

Table 4. Mean Classification Accuracy for Experiment 3

	p=6	p=10	p=20	p=40
Common Cov	39.7 (4.1)	40.4 (4.1)	42.7 (3.3)	40.5 (4.5)
Sample Cov	85.4 (2.7)	83.3 (5.7)	N/A	N/A
Euclidean	38.8 (4.5)	40.6 (4.1)	43.8 (3.7)	45.0 (3.0)
RDA	83.6 (3.6)	86.1 (5.7)	90.6 (4.1)	93.0 (2.7)
LOOC	90.4 (1.7)	97.5 (0.9)	99.8 (0.3)	100.0 (0.1)
LOOC-Exact	90.4 (1.9)	97.5 (0.9)	99.8 (0.3)	100.0 (0.1)

B. Discussion

In all but one experiment with computer generated data (Experiment 1, p=6), LOOC led to higher accuracy than did the common covariance matrix estimate, and in all the experiments LOOC led to higher accuracy than the sample covariance matrix. In Experiments 2 and 3, where the true covariance matrices were not equal to the identity matrix, LOOC led to higher accuracy than the Euclidean distance. In Experiments 2 and 3, LOOC led to higher accuracy than RDA. The accuracy of LOOC was within 2.1% of the accuracy of LOOC-Exact except in the higher dimensions (p=10, and p=20) of Experiments 1 and 2, where the accuracy of LOOC was within 7% of the accuracy of LOOC-Exact.

The mixing values for LOOC-Exact were reasonable. In the experiments having identical covariance matrices (Experiments 1 and 2), the values of the mixing parameter α_i were close to $(0,0,0,1)^T$, which selects the diagonal common covariance matrix estimate. In Experiment 3, where each class had a different covariance matrix, the values of α_i were close to $(1,0,0,0)^T$, which selects the diagonal sample covariance matrix estimate.

The mixing values for LOOC were not as accurate. In all the experiments, the values of α_i were close to $(1,0,0,0)^T$, which selects the diagonal sample covariance matrix estimate. With only 15 training samples in each class, the approximation in Equation (4) biased the estimator toward the diagonal sample covariance matrix estimate. These values of α_i, though, still resulted in reasonable estimates, and the classification accuracy was within 7% of the LOOC-Exact method.

C. Experiments with Remote Sensing Data

The following experiments were performed on data taken in 1992 by the Airborne Visible/Infrared Imaging Spectrometer (AVIRIS). This instrument captures images of the earth's surface in 220 spectral bands covering the range 0.4 - 2.5 µm. In order to conduct the experiments, several samples (pixels) of various ground cover classes were identified in each scene. Then a small percentage of the samples were selected at random and used to estimate the mean and covariance matrix of each class. The remaining samples were classified to measure the classification accuracy. Each experiment was repeated 10 times, and the mean and standard deviation of the accuracy were recorded. Experiments were conducted with four different numbers of features. The features were selected evenly spaced across the spectrum, but did not include those wavelengths that are absorbed by water in the atmosphere.

The Cuprite, Nevada site, which is notable for its geology, has several exposed minerals. A total of 2744 samples (pixels) and 191 bands (0.40-1.34, 1.43-1.80, 1.96-2.46µm) were used in the experiment. The number of training samples in each class was 145, 14, 46, 77, 137, 50, 58, and 18, which represented 20% of the total number of available samples. The results of the experiment are presented in Table 5.

Table 5. Mean Classification Accuracy for Cuprite Site

	p=10	p=50	p=100	p=191
Common Cov	92.4 (0.8)	95.5 (0.8)	96.1 (0.5)	96.0 (0.4)
Sample Cov	95.2 (0.8)	N/A	N/A	N/A
Euclidean	40.8 (1.2)	42.3 (1.5)	41.7 (0.9)	42.0 (1.2)
RDA	94.8 (0.6)	97.7 (0.4)	97.5 (0.4)	96.2 (1.1)
LOOC	95.8 (0.7)	98.1 (0.3)	97.4 (0.4)	95.2 (0.3)
LOOC-Exact	95.8 (0.7)	98.1 (0.3)	97.4 (0.4)	95.2 (0.3)

For the agricultural Indian Pine site, ground observations were used to identify a total of 2521 samples. Of the total number of available samples, 20% were used as training samples making the number of training samples in each class 104, 90, 74, 98, 77, and 60. See Table 6 for the results of the experiment.

Table 6. Mean Classification Accuracy for Indian Pine Site

	p=10	p=50	p=100	p=191
Common Cov	73.0 (0.6)	80.7 (0.7)	81.5 (0.8)	81.4 (1.0)
Sample Cov	80.5 (0.6)	68.7 (1.6)	N/A	N/A
Euclidean	65.5 (0.4)	66.5 (0.4)	66.6 (0.6)	66.9 (0.8)
RDA	80.5 (0.6)	83.8 (0.7)	82.7 (1.6)	82.6 (1.3)
LOOC	80.1 (0.6)	84.1 (0.8)	81.8 (1.2)	80.9 (0.8)
LOOC-Exact	80.1 (0.6)	84.1 (0.8)	81.8 (1.2)	80.9 (0.8)

D. Discussion

In 6 out of 8 experiments with remote sensing data, LOOC led to higher classification accuracy than the common covariance matrix estimate, and in all the experiments but one (Indian Pine Site, p=10), LOOC produced higher accuracy than the sample covariance matrix estimate. In all cases, LOOC led to higher classification accuracy than the Euclidean distance classifier. In 3 of the 8 experiments, LOOC led to higher classification accuracy than RDA. In all the experiments with remote sensing data, where there were more training samples than in the experiments with computer generated data, LOOC and LOOC-Exact selected the same values for the mixing parameters, and therefore returned precisely the same covariance matrix estimates. This result indicates the approximations in Equations (4) and (5) were valid for these cases.

The CPU time required to compute LOOC was much less than that for RDA (see Tables 7 and 8 for CPU times for our implementation on a 50MHz SPARC server 1000).

Table 7. CPU Time (in Seconds) for 3 Classes, 45 Total Training Samples

	p=6	p=10	p=20	p=40
LOOC	0.03	0.06	0.27	1.55
RDA	0.39	0.83	2.37	10.06

Table 8. CPU Time (in Seconds) for 8 Classes, 545 Total Training Samples

	p=10	p=50	p=100	p=191
LOOC	0.6	15.3	82.0	502.5
RDA	22.1	222.9	838.9	5875.8

V. Conclusion

A new covariance matrix estimator was presented which led to higher classification accuracy than the sample covariance matrix and the common covariance matrix estimators when the number of training samples was limited compared to the number of features. An efficient implementation of the estimator was derived that incorporated an approximation, and the approximation was found to be valid for modest sample sizes. In about half of the experiments, the new estimator led to higher classification accuracy than RDA, but required much less computation.

Acknowledgments: The authors would like to thank Jerome Friedman for supplying source code for the RDA algorithm and Wei-Liem Loh for an insightful discussion about covariance matrix estimation.

References

[1] H.W. Sorenson, *Parameter estimation: principles and problems*, New York, M. Dekker, 1980, pp. 183-184.

[2] T.W. Anderson, *An Introduction to Multivariate Statistical Analysis,* 2nd Ed., New York, John Wiley & Sons, 1984, p. 209.

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

[4] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, Second Ed., Boston, Academic Press, 1990, pp. 38-39, 225.

[5] G.H. Golub, and C.F. Van Loan, *Matrix Computations*, 2nd Ed., Baltimore, Johns Hopkins University Press, 1989, p. 51.

[6] J.P. Hoffbeck, *Classification of High Dimensional Multispectral Data*, Ph.D. Thesis, Purdue University, West Lafayette, IN, pp. 55-70.