An Adaptive Classifier Design for High-Dimensional Data Analysis with a Limited Training Data Set[1]

Qiong Jackson and David Landgrebe
School of Electrical &Computer Engineering
Purdue University

Abstract-In this paper, we propose a self-learning and self-improving adaptive classifier to mitigate the problem of small training sample size that can severely affect the recognition accuracy of classifiers when the dimensionality of the multispectral data is high. This proposed adaptive classifier utilizes classified samples (referred as semi-labeled samples) in addition to original training samples iteratively. In order to control the influence of semi-labeled samples, the proposed method gives full weight to the training samples and reduced weight to semi-labeled samples. We show that by using additional semi-labeled samples that are available without extra cost, the additional class label information may be extracted and utilized to enhance statistics estimation and hence improve the classifier performance, and therefore the Hughes phenomenon (peak phenomenon) may be mitigated. Experimental results show this proposed adaptive classifier can improve the classification accuracy as well as representation of estimated statistics significantly.

Index Terms-Adaptive iterative classifier, high-dimensional Data, limited training data set, labeled samples, semi-labeled samples

I. INTRODUCTION

In remote sensing applications, increased spectral resolution brought about by the current sensor technology has offered new potentials and challenges to data analysts. On one hand, the availability of a large number of spectral bands makes it possible to identify more detailed classes with higher accuracy than would be possible with the data from earlier sensors. On the other hand, a large number of classes of interest and a large number of spectral bands available require a large number of training samples, which unfortunately are expensive or tedious to acquire. As a result, the class statistics must be

[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.

estimated from the limited training sample set. When the ratio of the number of training samples to the number of spectral features is small, the parameter estimates become highly variable, causing classification performance to deteriorate with increasing dimensionality. This phenomenon where with finite training samples, classifier performance rises with dimensionality at first and then declines, was studied in detail by Hughes [1], and is later referred to as the Hughes phenomenon.

An additional problem that usually exists in remote sensing applications is the unrepresentative training sample problem. Since usually training samples are selected from spatially adjacent regions, they may not be good representatives of the samples of the entire class, which is likely distributed over the entire scene. This problem further aggravates the difficulties in analyzing remote sensing data.

To mitigate the small training sample problem, a self-learning and self-improving adaptive classifier is proposed in this paper. This adaptive classifier enhances statistics estimation and hence improves classification accuracy iteratively by utilizing the classified samples (referred as semi-labeled samples), in addition to the original training samples, in subsequent statistics estimation. In this iterative process, samples are initially classified based on the estimated statistics using the original training samples only. Then the classified results are subsequently used with the original training samples to update class statistics, and the samples are reclassified by the updated statistics. This process is repeated until convergence is reached.

The proposed adaptive classifier potentially has the following benefits:

1) The large number of semi-labeled samples can enhance the statistics estimates, decreasing the estimation error and therefore reduce the effect of the small sample size problem, because the semi-labeled samples in effect enlarge the training sample size.
2) The estimated statistics are more representative of the true class distribution, because samples used to estimate statistics are from a larger portion of the entire data set.
3) This classifier is adaptive in the sense that it can improve the accuracy by using the information extracted from its output. With proper conditions, a positive feedback system can be formed, whereby better statistics estimation leads to higher classification accuracy, and in return, higher classification accuracy results in even better parameter estimation.
4) In a way, this approach augments automation of the classifier. It is possible that to start with a small number of training samples (minimum input from the analyst) this classifier may be able to continuously extract useful information from the data, adjusting itself accordingly, and eventually evolve automatically to an optimal classifier which produces optimal classification accuracy with a given data set. Hence the analyst's effort can be greatly reduced.

Since the semi-labeled samples can be fed back before or after any feature extraction is performed, it offers flexibility of implementation, that is, depending on the requirement

of accuracy and the computation load, the semi-labeled samples can be used in more than one way.

There are five sections in this paper. In the first section, the information available for estimating the statistics of a mixture of two normal distributions is examined for training samples and semi-labeled samples in terms of Fisher information matrices. In the second section the effect of semi-labeled samples on the probability of error is investigated.

In the third section, a self-learning and self-improving adaptive classifier is presented where both training and semi-labeled samples are used. In order to control the influence from semi-labeled samples, the proposed method gives full weight to the training samples and reduced weight to semi-labeled samples.

In the fourth section, experiments on the proposed adaptive classifier using simulated and real data set are presented. With a large number of semi-labeled samples available, the usual case in remote sensing applications, experimental results show this proposed adaptive classifier can improve the classification accuracy significantly. Final remarks are presented in the last section.

II. INFORMATION OF TWO NORMAL DISTRIBUTIONS

In this section, the information available for estimating the parameters of a mixture of two normal distributions is examined in terms of the Fisher information matrix, denoted by I_s. According to the Crame-Rao inequality [2], if $\hat{\theta}$ is any absolutely unbiased estimate of θ based on the measure data z, then the covariance of the error in the estimate is bounded below by the inverse of the Fisher information matrix, assuming it exists. Furthermore, if $\hat{\theta}$ is asymptotically (a large sample size) unbiased and efficient (for example, maximum likelihood estimates always possess these properties [2]), then $\mathrm{cov}(\hat{\theta}) \approx I_s^{-1}$. Loosely speaking, with more information available, then the determinant and trace of the inverse of the Fisher information matrix become smaller, and correspondingly, the covariance of an unbiased estimate is smaller too. In other words, the estimate becomes more stable.

Consider a classification problem involving two multivariate classes which can be represented as Gaussian distributions with probability density functions (pdf's) $f_i(x|\mu_i, \Sigma_i), i = 1,2$, where μ_i, and Σ_i denote the mean vector and covariance matrix of class *i*. The prior probabilities associated with the two classes are designated by P_1 and P_2. We consider the following case: n independent unlabeled observations (X_1, X_2,…., X_n) are drawn from the mixture of these two classes, and are subsequently classified as class one (C_1) and class two (C_2) based on the Bayes decision rule which assigns an observation to the class with the highest a posteriori probability for minimizing the total classification error:

$$\begin{aligned} X \quad \Omega_1 &\Leftrightarrow P_1 f_1(x) \geq P_2 f_2(x) \\ X \quad \Omega_2 &\Leftrightarrow P_1 f_1(x) < P_2 f_2(x) \end{aligned} \tag{1}$$

where Ω_1 and Ω_2 are two sub-spaces corresponding to class one and class two respectively. Suppose n_1 samples are correctly classified, and n_2 samples are misclassified, i.e., $n_1 + n_2 = n$. Denoting I_{sl} as the Fisher information matrix for this case, using the definition of Fisher information matrix [2], then we have:

$$\begin{aligned} I_{sl} &= nE\left\{[\frac{\partial}{\partial\theta}\log f(x,\theta)][\frac{\partial}{\partial\theta}\log f(x,\theta)]^T\right\} \\ &= n_1P_1E\left\{[\frac{\partial}{\partial\theta}\log f(x,\theta)][\frac{\partial}{\partial\theta}\log f(x,\theta)]^T \middle| x \quad \Omega_1, x \text{ is } C_1\right\} \\ &+ n_1P_2E\left\{[\frac{\partial}{\partial\theta}\log f(x,\theta)][\frac{\partial}{\partial\theta}\log f(x,\theta)]^T \middle| x \quad \Omega_2, x \text{ is } C_2\right\} \\ &+ n_2P_1E\left\{[\frac{\partial}{\partial\theta}\log f(x,\theta)][\frac{\partial}{\partial\theta}\log f(x,\theta)]^T \middle| x \quad \Omega_2, x \text{ is } C_1\right\} \\ &+ n_2P_2E\left\{[\frac{\partial}{\partial\theta}\log f(x,\theta)][\frac{\partial}{\partial\theta}\log f(x,\theta)]^T \middle| x \quad \Omega_1, x \text{ is } C_2\right\} \end{aligned} \tag{2}$$

Without loss of generality, consider the canonical form where $\mu_1=0$, and $\mu_2=[\Delta\ 0\ldots0]^T$, and $\Sigma_1=\Sigma_2=I_d$, $\Delta>0$, Δ^2 is the Mahalanobis distance between the two classes, and I_d is a $d \times d$ identity matrix (d is the dimension of the feature space). Since the error rate of probability is the subject of our study in the next section and is invariant under nonsingular linear transformation, the canonical form can be used here without loss of generality. Any other two-class problem for which $\Sigma_1=\Sigma_2$ can be transformed into the above form through a linear transformation [3]. Using these conditions, Eq. (2) can be simplified as follows (the detailed derivation is shown at appendix A):

$$I_{sl} = n\begin{bmatrix} P_1k_1 & & & \\ & P_1k_2I_{d-1} & & \\ & & P_2k_3 & \\ & & & P_2k_4I_{d-1} \end{bmatrix} \tag{3}$$

where

$$k_1 = r_c\alpha_1 + (1-r_c)(1-\alpha_1)$$

$$k_2 = r_c\beta_1 + (1-r_c)(1-\beta_1)$$

$$k_3 = r_c\alpha_2 + (1-r_c)(1-\alpha_2)$$

$$k_4 = r_c\beta_2 + (1-r_c)(1-\beta_2)$$

$$r_c = \frac{n_1}{n}$$

$$\alpha_1 = \Phi(t) - t\phi(t)$$

$$\beta_1 = \Phi(t)$$

$$\alpha_2 = \Phi(\Delta - t) - (t - \Delta)\phi(t - \Delta)$$

$$\beta_2 = \Phi(\Delta - t)$$

$$\Phi(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-\frac{x^2}{2}} dx$$

$$\phi(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}$$

Here $\Phi(t)$ and $\phi(t)$ are the cumulative distribution function (cdf) and probability density function (pdf) of the standard normal distribution respectively, and r_c is the classification accuracy. From equation (3) we can derive the following interesting results:

1) If two classes are quite separated, i.e., $\Delta >> 1$, then $t >> 1$ and hence $\Phi(t) \approx 1$ and $t\phi(t) \approx 0$, $\alpha_1 \approx \alpha_2 \approx \beta_1 \approx \beta_2 \approx 1$. In this case, equation (3) can be simplified as:

$$I_{sl} \approx n_1 \begin{bmatrix} P_1 I_d & 0 \\ 0 & P_2 I_d \end{bmatrix} \le n \begin{bmatrix} P_1 I_d & 0 \\ 0 & P_2 I_d \end{bmatrix} \qquad (4)$$

where the above inequality is a matrix inequality indicating that the right hand side minus the left hand side is a positive semi-definite matrix. Notice that the right hand side of the above inequality is the Fisher information matrix for estimating θ if the n randomly drawn samples have been labeled. In particular, let I_s be the information matrix for this case. One can write:

$$\begin{aligned} I_s &= n\{P_1 E\left[\frac{\partial}{\partial\theta}\log f_1(x)\right]\left[\frac{\partial}{\partial\theta}\log f_1(x)\right]^T \mid x \text{ is } C_1\} \\ &+ P_2 E\left[\frac{\partial}{\partial\theta}\log f_2(x)\right]\left[\frac{\partial}{\partial\theta}\log f_2(x)\right]^T \mid x \text{ is } C_2\}\} \qquad (5) \\ &= n\begin{bmatrix} P_1 I_d & 0 \\ 0 & P_2 I_d \end{bmatrix} \end{aligned}$$

Therefore, inequality (4) reveals the conceptually appealing fact that the information contained in n classified observations based on the Bayes decision rule is less than or equal to that of n labeled observations. The missing information in this case using only semi-labeled samples (referred as semi-supervised learning) is due to the mis-assigned labels. From now on we refer to the right hand side of (4) as the "supervised bound" for I_{sl}. Usually, classification accuracy achieved by Bayes rule with known class condition probability density functions goes up with the separation of classes. Therefore, if two classes are quite separated, we have $n_1 >> n_2$ or $n_1 \approx n$, leading to $I_{sl} \approx I_s$, which implies more information can be gained from more correctly classified samples.

2) At the worst case where half of the samples are correctly classified and the remaining half are misclassified, i.e., $n_1 = n_2 = n/2$, I_{sl} can be written as:

$$I_{sl} = \frac{n}{2}\begin{bmatrix} P_1 I_d & 0 \\ 0 & P_2 I_d \end{bmatrix} = \frac{1}{2} I_s \qquad (6)$$

This indicates that at least 50% of class label information is generated after classification.

In summary, for the canonical two component normal mixtures with unknown means, after the classification is performed based on the Bayes decision rule, the Fisher information matrix I_{sl} is bounded as follows:

$$\frac{n}{2}\begin{bmatrix} P_1 I_d & 0 \\ 0 & P_2 I_d \end{bmatrix} \le I_{sl} \le n\begin{bmatrix} P_1 I_d & 0 \\ 0 & P_2 I_d \end{bmatrix} \qquad (7)$$

Under suitable regularity conditions the inverse of the Fisher information matrix (I^{-1}) is the asymptotic (large sample) variance-covariance matrix for the maximum likelihood estimates [2]. For the equal prior probability case (P_1=P_2=0.5), by inverting the bounds in Eq. (7), the asymptotic covariance of the ML (Maximum Likelihood) estimate of $\theta = [\mu_1^T, \mu_2^T]^T$ can be bounded from above and below. Notice that for any two positive definite matrices A and B, if A≥B, then $B^{-1} \ge A^{-1}$[4]. Denoting $\hat{\theta}$ as the ML estimate of θ obtained by using semi-labeled samples, then $\mathrm{cov}(\hat{\theta})$ is bounded as follows:

$$\mathrm{cov}(\hat{\theta}) \le \frac{1}{2n}\begin{bmatrix} \frac{1}{P_1} I_d & \\ & \frac{1}{P_2} I_d \end{bmatrix} \qquad (8a)$$

and

$$\mathrm{cov}(\hat{\theta}) \geq \frac{1}{n}\begin{bmatrix} \frac{1}{P_1 k_1} & \frac{1}{P_1 k_2} I_{d-1} \\ \frac{1}{P_2 k_3} & \frac{1}{P_2 k_4} I_{d-1} \end{bmatrix} \qquad (8b)$$

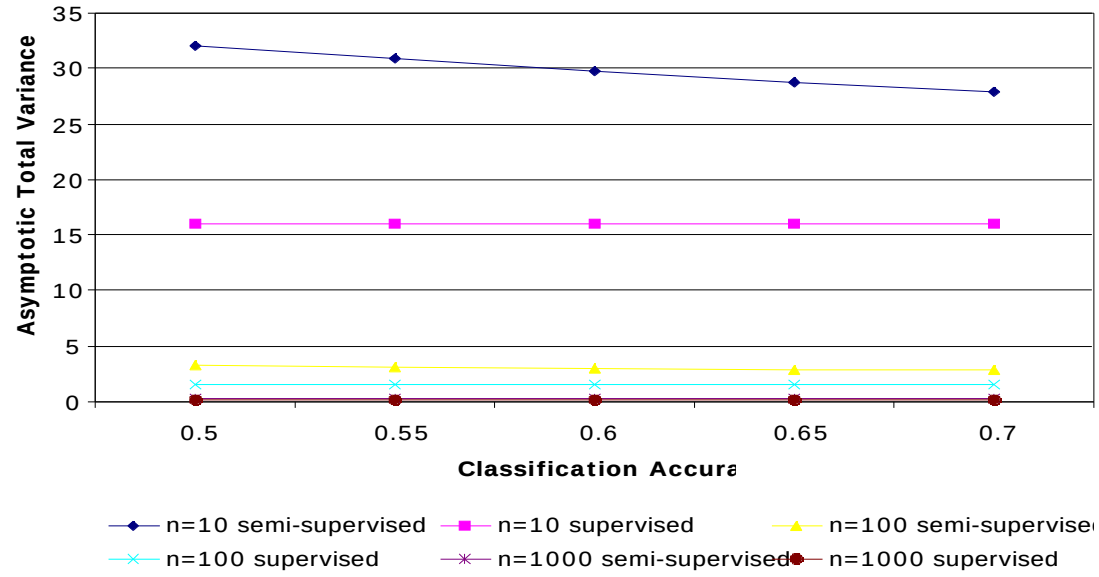
Using || and *tr* to denote the determinant and trace operators respectively then $|I^{-1}|$ and $tr(I^{-1})$ represent the asymptotic generalized and total variance [5]. Using Eq. (7) we can obtain the trace and determinant of $(I_{sl})^{-1}$:

$$\begin{aligned} tr((I_{sl})^{-1}) &= \frac{1}{n}\left(\frac{1}{P_1 k_1} + \frac{d-1}{P_1 k_2} + \frac{1}{P_2 k_3} + \frac{d-1}{P_2 k_4}\right) \\ &= \frac{1}{n}\left[\left(\frac{1}{P_1 k_1} + \frac{1}{P_2 k_3}\right) + (d-1)\left(\frac{1}{P_1 k_2} + \frac{1}{P_2 k_4}\right)\right] \end{aligned} \qquad (9a)$$

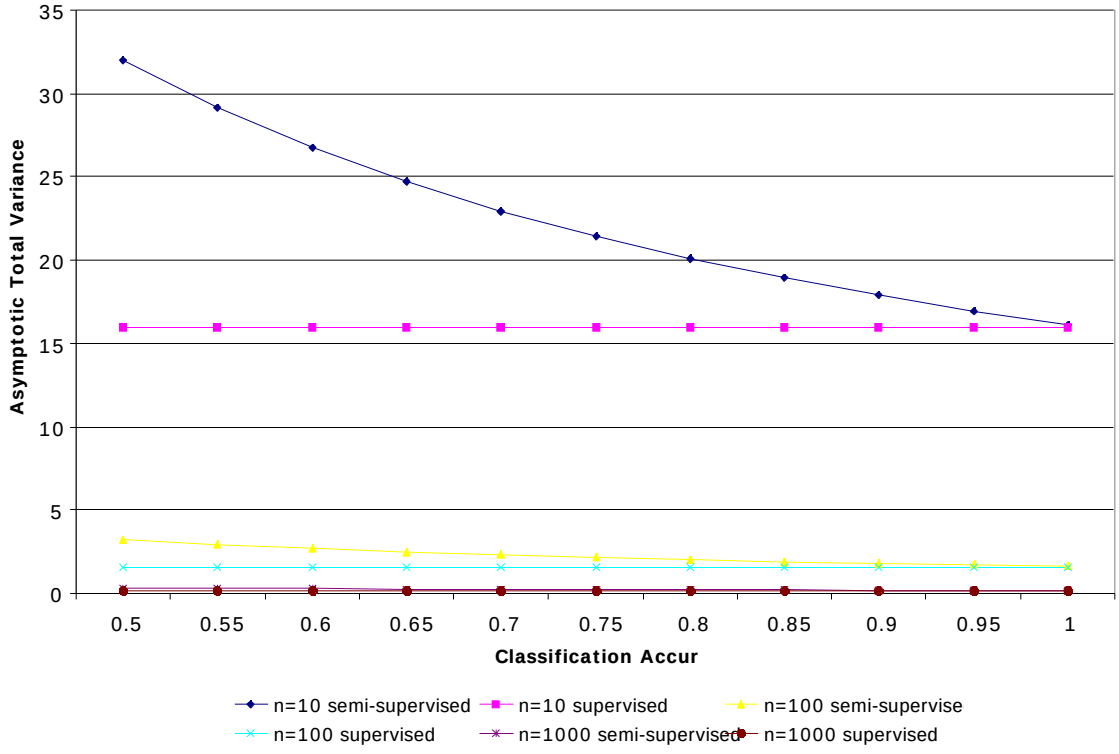
and

$$|(I_{sl})^{-1}| = \left(\frac{1}{P_1 P_2 n}\right)^d \left(\frac{1}{k_1 k_3}\right)\left(\frac{1}{k_2 k_4}\right)^{d-1} \qquad (9b)$$

Figure (1a) to (1b) illustrate the variation of asymptotic total variance with the accuracy, the number of samples, separations for semi-supervised learning (only semi-labeled samples are used) and supervised learning (only labeled samples are used). Note that accuracy achieved by Bayes rule is approximately 69% for Δ=1, and 99% Δ=5 with equal prior probabilities [3]. From these figures it is seen that 1) asymptotic total variance decreases with increase of classification accuracy. It drops faster when two classes are more separated; 2) Asymptotic total variance increases with increase of dimensionality, but decreases dramatically with increase of the number of samples; 3) The difference of asymptotic total variance using labeled and semi-labeled samples reduces with classification accuracy and separability of two classes.



(a) delta=1, d=40



(b) delta=5, d=40

Figure 1. Asymptotic total variance using semi-labeled.

The above results imply that when semi-labeled samples are used, 1) the improvement of classification accuracy may reduce the total variance and hence enhance the estimation of statistics, and in return, the enhanced statistics can further improve the classification accuracy. This implies when semi-labeled samples are used to integrate statistics estimation with classification, a positive feedback can be established where statistics estimation and classification enhance each other and eventually a close to optimal classification accuracy can be reached with a given data set. 2) The large number of semi-labeled samples may significantly reduce the total variance and therefore mitigate the effect of small training sample size problem. 3) Semi-labeled samples can provide comparable class label information when two classes are quite separable and classification accuracy is high.

III. BOUND ON PROBABILITY OF ERROR

A. Semi-Supervised Learning

In the equal covariance case ($\Sigma_1 = \Sigma_2 = \Sigma$), the optimal classifier is linear:

$$h(x) = (\mu_2 - \mu_1)\Sigma^{-1} + \frac{1}{2}(\mu_1^T\Sigma^{-1}\mu_1 - \mu_2^T\Sigma^{-1}\mu_2) + \log\frac{P_2}{P_1} \begin{cases} < 0 & class1 \\ > 0 & class2 \end{cases} \quad (10)$$

When the true parameter values are used to evaluate h(x), the above linear classifier minimizes probability of error, which is referred as the Bayes probability of error. If the parameters are replaced by their estimates in h(x), the error rises. The probability of error is therefore a convex function of the parameters in the neighborhood of the true parameter values [2]. The expected probability of error using estimated parameters can be written as [3]:

$$\begin{aligned} E\{\hat{err}\} &\approx err^* + \frac{1}{2}tr\left[\frac{\partial^2 err}{\partial\theta^2}\Big|_{\theta=\theta^*}\mathrm{cov}(\hat{\theta})\right] \\ &= err^* + \frac{1}{2}tr\left[\int_{-\infty}^{\infty}\frac{1}{j\omega}\left[\frac{\partial^2}{\partial\theta^2}e^{j\omega h(x)}\right]_{\theta=\theta^*}\left[P_1 f_1(x) - P_2 f_2(x)\right]dx\,d\omega\,\mathrm{cov}(\hat{\theta})\right] \\ &= err^* + \frac{1}{2\pi}\int_{-\infty}^{\infty}\frac{1}{2}tr\left[\left[\frac{\partial^2 h(x)}{\partial\theta^2} + j\omega\frac{\partial h(x)}{\partial\theta}\frac{\partial h^T(x)}{\partial\theta}\right]_{\theta=\theta^*}\mathrm{cov}(\hat{\theta})\right] \\ &\quad \times e^{j\omega h(x)}\left[P_1 f_1(x) - P_2 f_2(x)\right]dx\,d\omega \end{aligned} \quad (11)$$

For the canonical form where μ_1=0, and μ_2=[Δ 0…0]^T, and Σ_1=Σ_2=I_d, Δ>0, we have:

$$\left[\frac{\partial h(x)}{\partial \theta}\ \frac{\partial h^T(x)}{\partial \theta}\right]_{\theta=\theta^*} = \begin{bmatrix} xx^T & -x(x-\mu_2)^T \\ -(x-\mu_2)x^T & (x-\mu_2)(x-\mu_2)^T \end{bmatrix} \quad (12a)$$

and $$\left.\frac{\partial^2 h(x)}{\partial \theta^2}\right|_{\theta=\theta^*} = \begin{bmatrix} I_d & \\ & -I_d \end{bmatrix} \quad (12b)$$

The integrals in (11) can be computed by the method provided in [3]. Replacing $\mathrm{cov}(\hat{\theta})$ in (8) by its upper and lower bounds described in Eq. (12a) through Eq. (12b) leads to the following inequalities for the bias of $\hat{err}$:

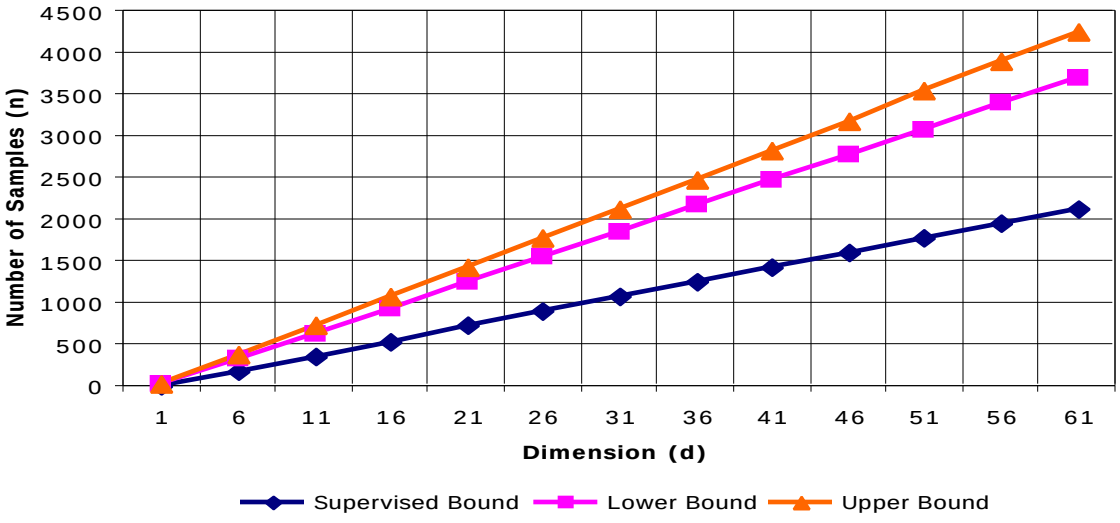
$$bias(\hat{err}) \geq \frac{1}{n\sqrt{2\pi}\Delta} e^{-\frac{\Delta^2}{8}} \left[\frac{\Delta^2}{4} + d - 1\right] \text{(supervised lower bound)} \quad (13a)$$

$$bias(\hat{err}) \geq \frac{1}{n\sqrt{2\pi}\Delta} e^{-\frac{\Delta^2}{8}} \left[\frac{\Delta^2}{8}\left(\frac{1}{k_3}+\frac{1}{k_1}\right) + \frac{(d-1)}{2}\left(\frac{1}{k_2}+\frac{1}{k_4}\right)\right] \quad (13b)$$

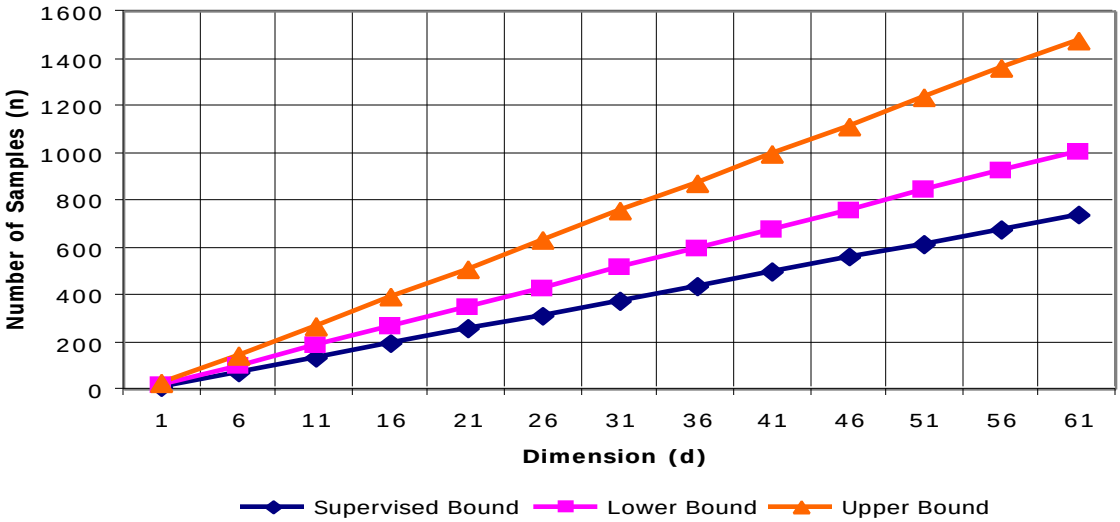
$$bias(\hat{err}) \leq \frac{2}{n\sqrt{2\pi}\Delta} e^{-\frac{\Delta^2}{8}} \left[\frac{\Delta^2}{4} + d - 1\right] \quad (13c)$$

Here the supervised lower bound is applied for supervised learning where n samples are labeled. It is possible to show that the variance of $\hat{err}$ is $O(\frac{1}{n^2})$ [5] and is therefore negligible.

Figure (2a) and (2b) show the bounds on the number of semi-labeled samples required to maintain the bias of classification error to less than 1% when dimensionality varies. Figure (3) shows the upper and lower bounds of the bias of the probability of error (in percent) versus Δ (Square root of the Mahalanobis distance), when P_1=P_2, d=4, and n=1000. Notice that as Δ goes up the semi-supervised curves gets closer to the supervised lower bound indicating when classes are far away from each other, semi-supervised learning can achieve comparable performance to supervised learning.



(a) delta=1



(b) delta=2

Figure 2. Number of training samples for supervised learning and semi-labeled samples for semi-supervised learning required having bias (error) <1%.

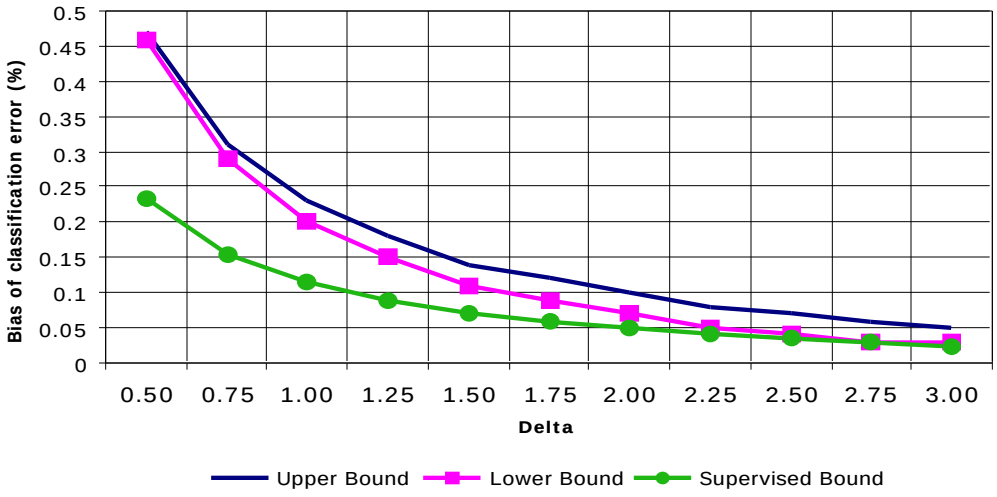


Figure 3. Bounds on the bias of the classification error for semi-supervised learning. (P_1=P_2, d=4, and n=1000)

B. Combined Supervised and Semi-Supervised Learning

In practical applications, usually both training and semi-labeled samples are available. Assuming that the training and semi-labeled samples are statistically independent, one can write the Fisher information matrix corresponding to the combined supervised and semi-supervised learning as the sum of the Fisher information matrices corresponding to the training and semi-labeled samples. This implies that if both training samples and semi-labeled samples are used simultaneously to estimate the parameters of the decision rule, better performance with lower bias and variance can be achieved than when using training samples alone [6]. By using the bounds obtained for the Fisher information matrix corresponding to the semi-labeled samples in equation (7), similar bounds can be obtained for the combined supervised and semi-supervised learning case. These bounds can then be utilized to determine the upper and lower bounds for bias of classification error as is done in the previous section for the semi-supervised case.

Assume that in addition to the n semi-labeled samples, n_{1t} labeled samples from class 1 and n_{2t} labeled samples from class 2 are also available for training the classifier. If the estimate of the parameter set $\theta = [\mu_1^T \; \mu_1^T]^T$ obtained by using all of these samples in the decision rule (10), the bias of the classification error, for the case P_1=P_2, is bounded as:

$$bias(\hat{err}) \geq \left(\frac{1}{n_{1t} + n/2} + \frac{1}{n_{2t} + n/2} \right) \frac{1}{4\sqrt{2\pi}\Delta} e^{-\frac{1}{8}\Delta^2} \left[\frac{\Delta^2}{4} + d - 1 \right] \quad (14a)$$

(supervised lower bound)

$$bias(\hat{err}) \geq \frac{1}{4\sqrt{2\pi}\Delta} e^{-\frac{\Delta^2}{8}} \left[\frac{\Delta^2}{4} \left(\frac{1}{n_{1t} + \frac{n}{2} k_1} + \frac{1}{n_{2t} + \frac{n}{2} k_3} \right) + (d-1) \left(\frac{1}{n_{1t} + \frac{n}{2} k_2} + \frac{1}{n_{2t} + \frac{n}{2} k_4} \right) \right] \quad (14b)$$

$$\geq \left(\frac{1}{n_{1t} + n/2} + \frac{1}{n_{2t} + n/2} \right) \frac{1}{4\sqrt{2\pi}\Delta} e^{-\frac{1}{8}\Delta^2} \left[\frac{\Delta^2}{4} + d - 1 \right]$$

$$bias(\hat{err}) \leq \left(\frac{1}{n_{1t} + n/4} + \frac{1}{n_{2t} + n/4} \right) \frac{1}{4\sqrt{2\pi}\Delta} e^{-\frac{1}{8}\Delta^2} \left[\frac{\Delta^2}{4} + d - 1 \right] \quad (14c)$$

The variance of $\hat{err}$ is again negligible since it is inversely proportional to the square of the number of training samples.

Figure (4) shows the bounds of the bias of the probability of error versus Δ when $P_1=P_2$, d=4, n=100, and $n_{1t}=n_{2t}=10$. The no-semi-labeled curve in this figure refers to the case when only labeled samples are used. It is seen that by using additional semi-labeled samples, the bias of the classification error is substantially reduced. The amount of the reduction depends on the separation between two classes as characterized by Δ.

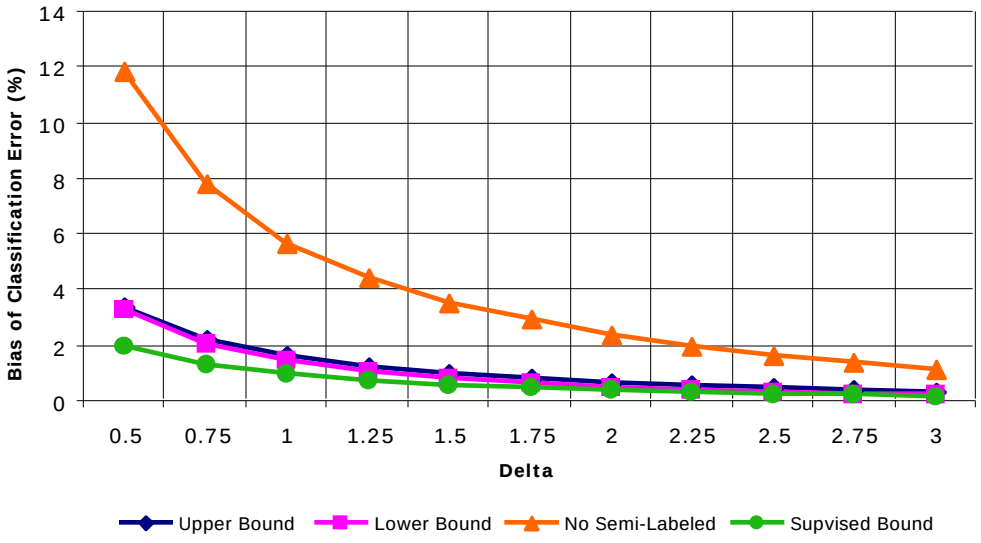


Figure 4. Bounds on the bias of the classification error for combined learning

In conclusion, semi-supervised learning can achieve comparable performance to supervised learning when the classes are relatively separated. When the classes are highly overlapped, a large number of semi-labeled samples are necessary for designing a classifier that matches the performance of the one designed by supervised learning. When both training and semi-labeled samples are available, the combined supervised and semi-

supervised learning that uses these two kinds of samples can outperform supervised learning. This result is significant for the remote sensing applications where the number of training samples is usually limited compared to the dimensionality of data obtained by high spectral resolution sensors, while a large amount of semi-labeled samples are available after the classification is performed without additional effort. In such cases, utilizing semi-labeled samples may mitigate the Hughes phenomenon [1]. If we know which samples have been correctly classified and use them accordingly to re-estimate statistics in addition to original training samples, the estimated statistics should be more precise because the actual training samples have been enlarged. Since usually we have no knowledge of classification accuracy for each individual sample, the key is to design a scheme that is able to apply a control factor that is related to the likelihood of a semi-labeled sample to a class. In the next section, an adaptive classifier is designed to achieve this goal.

IV.DESIGN OF AN ADAPTIVE CLASSIFIER

If we assume every sample in the data set is unique, i.e. it belongs only to one class, we would expect it should only contribute to statistics of the only class to which it belongs. In the EM algorithm [8] and its application in remote sensing [6], each unlabeled sample has a certain amount of membership for each class and correspondingly has weighted contribution to the statistics of every class. Even though this is reasonable at this point because the sample labels are completely unknown, the contribution of the sample to the class to which it does not belong is definitely undesired. This negative influence may be significant enough to cause the estimated statistic to deviate from the true one, especially when a large number of unlabeled samples are used, or there exists a class whose statistics are quite different from the rest of classes. For example, if the class proportion is quite unbalanced, i.e., a few classes are quite dominant in the given data set, then the large number of unlabeled samples used may be mostly from these dominant classes. With small numbers of training samples, the estimated statistics of minority classes will be overwhelmed by the unlabeled samples and consequently may deviate from the true one. This phenomenon has been observed in practice, and it has been noticed that better classification accuracy could be achieved by using approximately the same number of unlabeled samples as the number of training samples, which is small. This is unfortunate because more information can be obtained and utilized with additional unlabeled samples [6][7].

In this section, an adaptive classifier based on the Maximum Likelihood (ML) rule is proposed to enhance the statistics estimation by using semi-labeled samples in addition to training samples. In this new classifier, the partial information of the class label obtained in the process of classification is utilized in such a way that each semi-labeled sample only affects the statistics of the extract class into which it has been partitioned. Furthermore this classifier assigns full weight to training samples, but automatically gives reduced weight to semi-labeled samples. Therefore, it utilizes the additional class label information provided by correctly classified semi-labeled samples and at the same time limits the undesired influence from misclassified samples. Before we describe the proposed adaptive classifier, we first provide a brief review of Expectation Maximization (EM) algorithm.

The EM algorithm is an iterative method for numerically approximating the maximum likelihood (ML) estimates of the parameters in a mixture model. Under the mixture model, the distribution of an observation x R^P is given as:

$$f(x|\Phi) = \sum_{i=1}^{L} \alpha_i f_i(x|\phi_i)$$

where $\alpha_1, \ldots, \alpha_L$ are the class prior probabilities and thus the mixing proportions, f_i is the component density parameterized by ϕ_i and L is the total number of components. The mixture density f is then parameterized by $\Phi = (\alpha_1,....,\alpha_L, \phi_1,...\phi_L)$.

Assume that $y = (y_1,...,y_{mi})$ are the m_i training samples from class i. Also, there are L classes and a total of n unlabeled samples denoted by $x = (x_1,...,x_n)$. The parameter set Φ then contains all the prior probabilities, mean vectors and covariance matrices. Assume that $\phi_1,...,\phi_L$ are mutually independent. The EM algorithm can then be expressed as the following iterative equation [8]:

E-step:

$$\tau_{ij}^c = \tau_i(x_j|\phi_i^c) = \alpha_i^c f_i(x_j|\phi_i^c) / \sum_{i=1}^{L} \alpha_i^c f_i(x_j|\phi_i^c) \qquad (15)$$

where τ_{ij}^c is the posterior probability that x_j belongs to class i.

M-step:

$$\alpha_i^+ = \sum_{j=1}^{n} \tau_{ij}^c / n \qquad (16a)$$

$$\phi_i^+ \quad \underset{\phi_i \in \Omega}{\arg\max} \left(\sum_{k=1}^{m_i} \ln(f_i(y_k|\phi_i)) + \sum_{k=1}^{n} \tau_{ik} \ln(f_i(x_k|\phi_i)) \right) \qquad (16b)$$

Equation (16b) indicates that the optimal ϕ_i maximizes the weighted summation of the log likelihood of training samples and unlabeled samples. For every training sample, the weighting factor is one, and for every unlabeled sample, the weighting factor is the posterior probability. If L classes can be represented as Gaussian distributions, Eq. (16a) and (16b) yield:

$$\mu_i^+ = \frac{\sum_{j=1}^{m_i} y_{ij} + \sum_{j=1}^{n} \tau_{ij} x_j}{m_i + \sum_{j=1}^{n} \tau_{ij}} = \frac{m_i}{m_i + \sum_{j=1}^{n} \tau_{ij}} \frac{\sum_{j=1}^{m_i} y_{ij}}{m_i} + \frac{\sum_{j=1}^{n} \tau_{ij}}{m_i + \sum_{j=1}^{n} \tau_{ij}} \frac{\sum_{j=1}^{n} \tau_{ij} x_j}{\sum_{j=1}^{n} \tau_{ij}} \tag{17a}$$

$$\Sigma_i^+ = \frac{\sum_{j=1}^{m_i} (y_{ij} - \mu_i^+)(y_{ij} - \mu_i^+)^T + \sum_{j=1}^{n} \tau_{ij} (x_j - \mu_i^+)(x_j - \mu_i^+)^T}{m_i + \sum_{j=1}^{n} \tau_{ij}} = \frac{m_i}{m_i + \sum_{j=1}^{n} \tau_{ij}} \frac{\sum_{j=1}^{m_i} (y_{ij} - \mu_i^+)(y_{ij} - \mu_i^+)^T}{m_i} + \frac{\sum_{j=1}^{n} \tau_{ij}}{m_i + \sum_{j=1}^{n} \tau_{ij}} \frac{\sum_{j=1}^{n} \tau_{ij} (x_j - \mu_i^+)(x_j - \mu_i^+)^T}{\sum_{j=1}^{n} \tau_{ij}} \tag{17b}$$

In [6][7], the EM algorithm has been studied and applied to remote sensing data. It was shown that by assuming a mixture model and using both training samples and unlabeled samples in obtaining the statistics estimates, the classification performance can be improved, and the Hughes phenomenon can then be delayed to a higher dimensionality and hence more features can be applied to achieve better performance. In addition, the parameter estimates represent the true class distribution more completely.

As indicated by Eq. (15) through Eq. (17b), in the EM algorithm each unlabeled sample contributes to the statistics of all classes selected, and the amount of contribution is weighted by the sample's posterior probability. This is reasonable because at this stage the class label information of an unlabeled sample is completely missing. However, if we assume each sample has a unique class label, apparently the influence from one of the unlabeled samples k of the j^{th} class to the i^{th} class statistics ($i \neq j$) is undesired, specifically, if i^{th} and j^{th} are quite different, and it is possible sample k has a large posterior probability for i^{th} class. This negative influence may be significant enough to cause the estimated statistics to deviate from the true ones. As a result, the iteration may converge to erroneous solutions. This situation can become very severe when a large number of unlabeled samples are used. For example, if the class proportion is quite

unbalanced, i.e., there are a few classes that are quite dominant in the given data set, then the large number of unlabeled samples used may be mostly from these dominant classes.

An alternative way is to replace unlabeled samples by semi-labeled samples, which contain partial information of class origin obtained by a decision rule in the classification process. With the additional information of class labels, one can limit the effect of a semi-labeled sample to one class to which it has been assigned with the highest likelihood. In addition, by using semi-labeled samples, parameter estimation and classification can be integrated in an iterative way such that they enhance each other consistently. In this process, every bit of improvement from classification is fed back to the process of parameter estimation and hence leads to better statistic estimation, and in return a better classification accuracy can be achieved. In other words, a self-learning and self-adapting process can then be established. This is advantageous for the analysis of high-dimensional data with limited training samples. In high dimensional space, in general, samples are more separable, and higher classification accuracy can be achieved if class statistics can be estimated more precisely. In the following section, an adaptive classifier will be proposed using both training samples and semi-labeled samples to obtain close to optimal statistics estimation and classification iteratively.

The proposed adaptive classifier is an iterative method to numerically find close to optimal performance for given data by integrating parameter estimation with classification. Denote $y = (y_{i1},..., y_{im_i})$ as the training samples for the i^{th} class, whose pdf is $f_i(x|\varphi_i)$, and $x = (x_{i1},..., x_{in_i})$ are the semi-labeled samples that have been classified to the i^{th} class. Among these semi-labeled samples, there are two types of samples, the correctly classified samples and misclassified samples. Correctly classified samples can play a role as equivalent to training samples and enhance statistics estimation. On the other hand, misclassified samples introduce undesired effects as information noise to the estimated statistics. Ideally, one would like to just use those semi-labeled samples that have been correctly classified. However, information about the classification accuracy for individual sample is not available at this point. Therefore, in order to control the effect from semi-labeled samples, a weighting factor is applied to represent this influence.

With this in mind, an adaptive classifier is designed, which obtains close to optimal performance by maximizing the weighted log likelihood of training samples and semi-labeled samples. Similar to the EM algorithm, it is an iterative approach and achieves the optimal statistics estimation and classification by starting with initial estimate φ^0 and classification based on training samples only and repeating the following steps at each iteration using training samples and semi-labeled samples:

1) Computing Weighting Factors:

$$w_{ij}^c = \frac{f_i(x_{ij} \mid \phi_i^c)}{\sum_{k=1}^{L} f_k(x_{ij} \mid \phi_k^c)} \qquad (18a)$$

2) Maximizing the mixed log likelihood:

$$\phi_i^+ = \underset{\phi_i \ \Omega}{\arg\max}\left(\sum_{k=1}^{m_i} \ln(f_i(y_k \mid \phi_i)) + \sum_{k=1}^{n_i} w_{ik}^c \ln(f_i(x_{ik} \mid \phi_i))\right) \tag{18b}$$

3) Performing classification based on the maximum likelihood (ML) classification rule:

$$x \quad i \Leftrightarrow i = \underset{1 \le i \le L}{\arg\max}(\ln(f_i(x \mid \phi_i^+))) \tag{18c}$$

Here the superscript "c" and "+" designate the current and next value respectively. If all L classes are Gaussian distributed, Eq. (18b) yields:

$$\mu_i^+ = \frac{\sum_{j=1}^{m_i} y_{ij} + \sum_{j=1}^{n_i} w_{ij}^c x_{ij}}{m_i + \sum_{j=1}^{n_i} w_{ij}^c} = \frac{m_i}{m_i + \sum_{j=1}^{n_i} w_{ij}^c} \frac{\sum_{j=1}^{m_i} y_{ij}}{m_i} + \frac{\sum_{j=1}^{n_i} w_{ij}^c}{m_i + \sum_{j=1}^{n_i} w_{ij}^c} \frac{\sum_{j=1}^{n_i} w_{ij}^c x_{ij}}{\sum_{j=1}^{n_i} w_{ij}^c} \tag{19a}$$

$$\Sigma_i^+ = \frac{\sum_{j=1}^{m_i} (y_{ij} - \mu_i^+)(y_{ij} - \mu_i^+)^T + \sum_{j=1}^{n_i} w_{ij}^c (x_{ij} - \mu_i^+)(x_{ij} - \mu_i^+)^T}{m_i + \sum_{j=1}^{n_i} w_{ij}^c} = \frac{m_i}{m_i + \sum_{j=1}^{n_i} w_{ij}^c} \frac{\sum_{j=1}^{m_i} (y_{ij} - \mu_i^+)(y_{ij} - \mu_i^+)^T}{m_i} + \frac{\sum_{j=1}^{n_i} w_{ij}^c}{m_i + \sum_{j=1}^{n_i} w_{ij}^c} \frac{\sum_{j=1}^{n_i} w_{ij}^c (x_{ij} - \mu_i^+)(x_{ij} - \mu_i^+)^T}{\sum_{j=1}^{n_i} w_{ij}^c} \tag{19b}$$

and Eq. (18c) yields:

$$x \quad i \Leftrightarrow i = \arg\min_{1 \le i \le L} d_i(x)$$

where d_i is a discriminant function [2] given by:

$$d_i(x) = (x - \mu_i^+)(\Sigma_i^+)^{-1}(x - \mu_i^+)^T + \ln|\Sigma_i^+|$$

Note that in a manner similar to the EM algorithm, the mean vectors and covariance matrices are weighted mixtures of ML estimates from training samples and semi-labeled samples, and the weight for each sample is related to the relative likelihood, which is less than one. But in this proposed adaptive classifier, unique membership is assumed and each semi-labeled sample only has contribution to the same class to which is classified. In addition, in this iterative process, the membership of each training sample remains the same. However, the membership of each semi-labeled sample is being updated at every iteration through the whole procedure.

V. EXPERIMENTAL RESULTS

In the following experiments, we test the performance of the proposed adaptive classifier using both simulated and real multispectral data. The first two experiments use simulated data of dimensionality of 6, 20, and 40. The third uses 12 dimensional real data.

In experiment 1 and 2, there are three simulated classes with Gaussian distributions. Three sets of labeled samples are generated independently. In the first set, there are 1000 samples for each class, and 10 samples are selected randomly from 1000 samples and subsequently used for training; the other 990 samples are then classified and become semi-labeled samples, which are used to estimate statistics at the following iteration. In the second data set, there are 10,000 random samples for each class and they are used for testing the performance of the classifier. The third data set is generated to benchmark the performance of the proposed adaptive classifier. In this data set, there are 1000 random samples for each class, and then all of them are used for designing a classifier, which is then tested by using 10,000 test samples from the second data set. The convergence criterion is that the relative difference of classification accuracy between two consecutive iterations is less than 0.01%. Each experiment is repeated ten times, and the mean classification accuracy and standard deviation are then estimated.

A. Experiment 1: Equal Spherical Covariance

1) d=6: In this experiment, the covariance matrix of all three classes is the identity matrix, but each class had a slightly different mean vector. The mean of the first class is at the origin; the mean of the second class is 3.0 in the first variable and zero in the other variables. The dimension is d=6. The mean classification accuracy versus iteration number is graphed in Fig. (5a).

Here SC represents the mean classification accuracy and standard deviation of the data where a sample covariance estimate is used as the initial estimate from training samples, and the mixed sample covariance shown in Eq. (19b) is used for the later estimation. The SC_Test represents the results for the testing data. LOOC represents the results where a mixed covariance estimator, LOOC, is used to estimate covariance matrices [9], and, similar to SC case, the mixed sample covariance shown in Eq. (19b) is then used for the following covariance estimation. LOOC_Test represents the results of the testing data.

The results show that with additional semi-labeled samples, the mean accuracy of data and testing data increases steadily with iterations until it reaches convergence. Note that in this data set, in the supervised learning process the mean classification accuracy for training data (resubstution accuracy [3]) is 91.01% with a standard deviation 0.66%, and for testing (hold out accuracy [3]) it is 90.67% with a standard deviation 0.15%. The Bayes accuracy (optimal) is bounded between these two. Therefore, we believe the final convergence solution is optimal within a range of standard deviation. Also, it is observed that the difference of the mean accuracy between data and test data are within a standard deviation. Further, the standard deviation is reduced with iterations. The final one is reduced by about five folds. Additional results not shown here indicate that the estimated statistics become more and more representative to the true ones and more robust. This, then, is a self-improving adaptive classifier where statistics estimation and classification enhance each other.

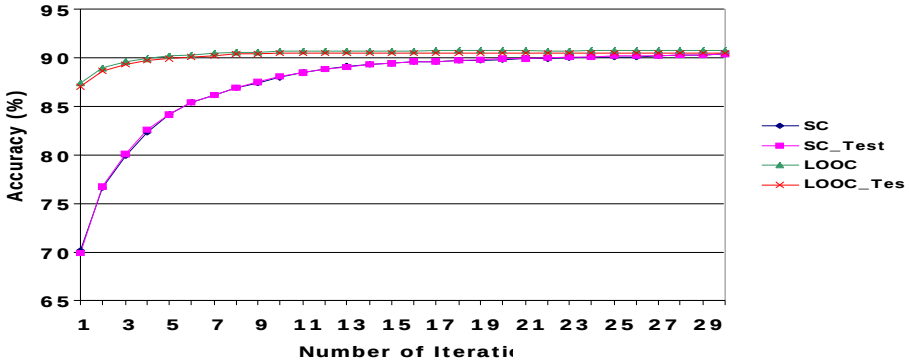
Also, it is seen that without LOOC, the initial accuracy is lower, and as a result convergence is attained more slowly but the final accuracy is very close to that with LOOC. This further indicates that eventually semi-labeled samples can compensate for the deterioration of classifier performance caused by lack of training samples.

2) d=20: In this experiment, the synthetic data from the experiment 1a is used with the exception that the dimensionality is raised from 6 to 20. Hence, the number of dimension is now greater than the number of class training samples but smaller than the total number of training samples. This case represents a poorly posed problem where the dimension size is greater than the training sample size. Mean classification accuracy is plotted in Fig. 5b. Since the number of dimension is greater than the class training sample size, the sample covariance matrix becomes singular and uninvertible. The covariance estimator LOOC must be used for the initial iteration. In this experiment, for supervised learning, the mean accuracy for data is 91.51% (std. dev. 0.59%) and for test data is 90.12 (std. dev. 0.12%).

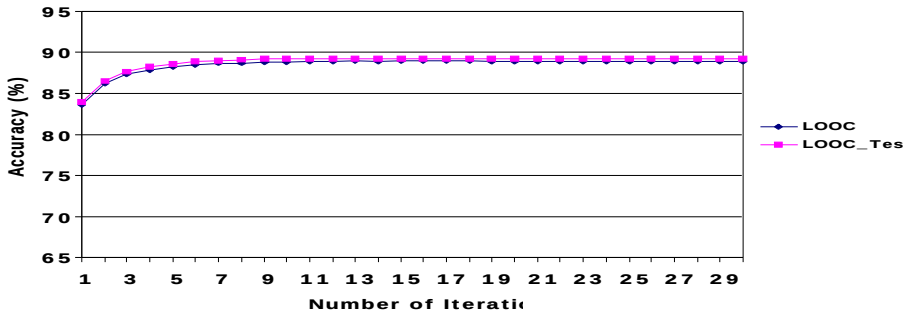
Comparing with experiment one, even though the initial classification accuracy reduces about 3% relatively, the classification accuracy still steadily increases and final classification accuracy is only about 2% lower. These results indicate that even with the poorly posed problem, this proposed adaptive classifier is still able to perform well.

3) d=40: Again, in this experiment the synthetic data from the experiment 1a is used with the exception that the dimension is increased to 40. Hence, the number of

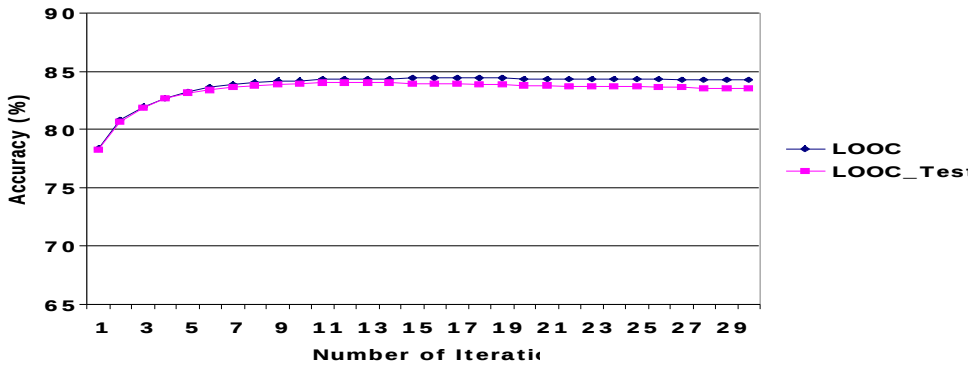
dimensions is much greater than the number of class training samples and even greater than the total number of training samples. This case represents an ill-posed problem where the number of dimensions exceeds the total number of training samples, and the number of parameters (2000) is twice the number of samples available. Mean classification accuracy is plotted in Fig. 5c. Again, since the number of dimension is greater than the class training sample size, the sample covariance matrix is singular and uninvertible. The covariance estimator LOOC is again used for the initial iteration. In this experiment, for supervised learning, the mean accuracy for data is 93.46% (std. dev. 0.57%) and for test data is 88.33 (std. dev. 0.28%).



(a) d=6



(b) d=20



(c) d=40

Figure 5. Mean Accuracy for Experiment 1.

Compared to the results of LOOC in experiment one, even though the initial classification accuracy is reduced about 10% relatively, the classification accuracy for the data still steadily increases. Final classification is about 7% less, and the standard deviation reduces with iterations as well. For testing data, the classification accuracy converges more slowly, and the final value is a little lower than previous accuracy. But overall these results show that this proposed adaptive classifier still is able to perform relatively well even for an ill-posed problem.

B. Experiment 2: Unequal Spherical Covariance Matrices

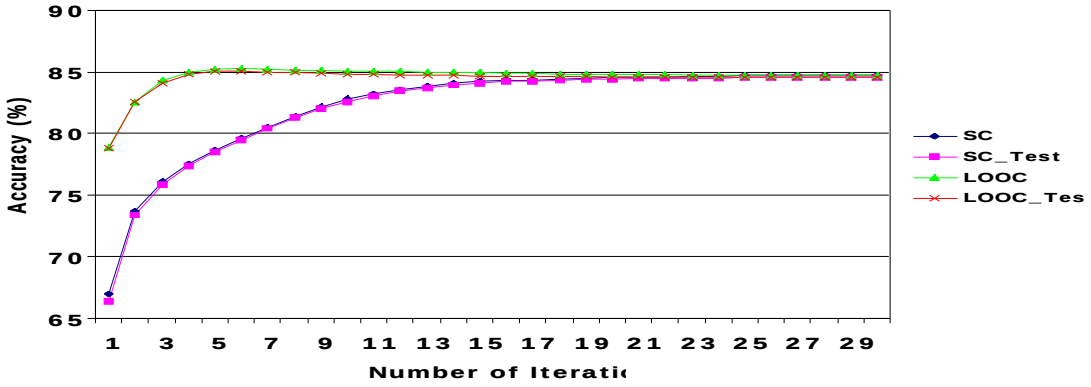
1) d=6: In this experiment, the three classes have unequal mean vectors and spherical covariance matrices. The mean vectors are the same as those in the experiment one. The covariance matrices of class one, two and three are I, 2I and 3I respectively. In this case, these three classes overlap more and are more difficult to discriminate than the equal covariance case. Mean accuracy is plotted in Figure 6a. It is observed that these results are similar to those in experiment 1a. In this experiment, for supervised learning, the mean accuracy for data is 88.68% (std. dev. 0.75%) and for test data is 85.99 (std. dev. 0.20%).

2) d=20: In this experiment, the simulated data in Experiment 2a is used with exception that the dimension is twenty, which is greater than the number of training samples. This is thus again a poorly posed problem. Mean accuracy is plotted in Figure 6b. In this experiment, for supervised learning, the mean accuracy for data is 92.48% (std. 0.56%) and for test data is 90.98 (std. 0.13%).

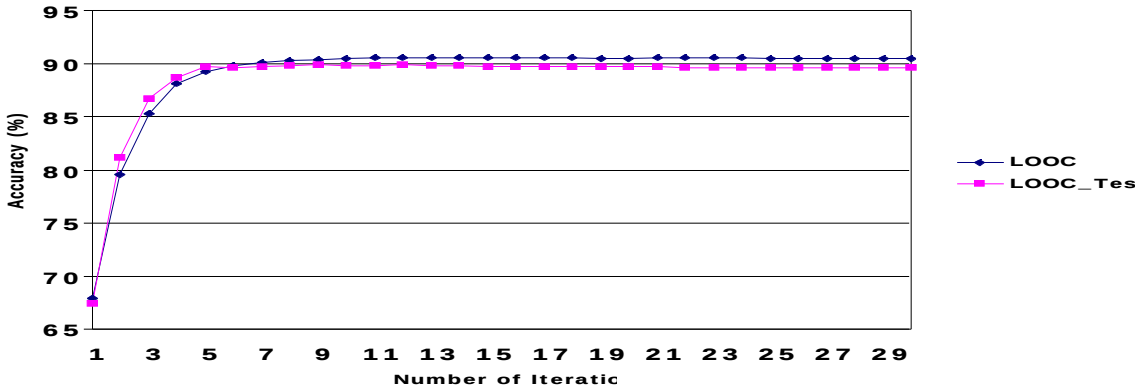
It is worth noting that even though the initial classification mean accuracy is reduced by 7% relatively, the final increases by 5%. This shows the appealing fact that with semi-labeled samples the proposed adaptive classifier is able to utilize the increment of separability provided by additional dimensions, and then improve the classification accuracy. In other words, Hughes phenomenon is mitigated.

3) d=40: In this experiment, the simulated data in Experiment 2a is used with exception that the dimension forty. Mean accuracy is plotted in Figure 6c. In this experiment, for supervised learning, the mean accuracy for data is 96.27% (std. 0.40%) and for test data is 93.07 (std. 0.14%).

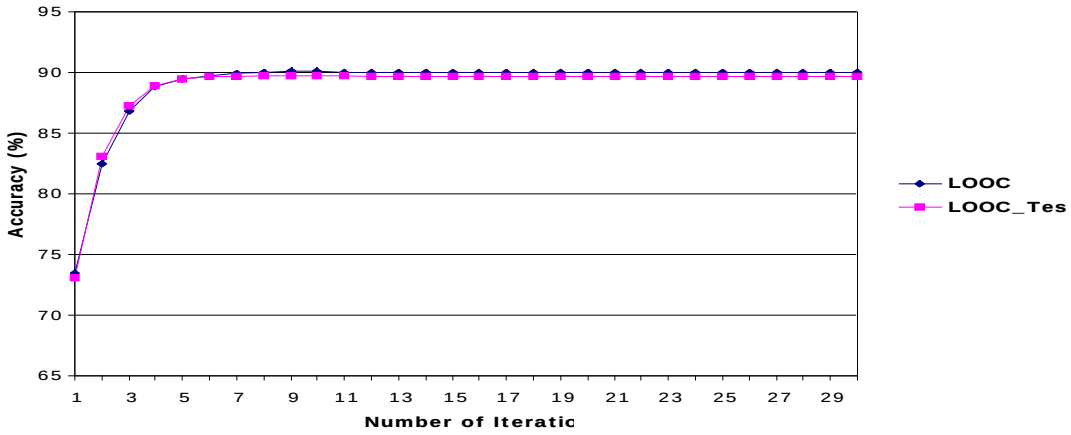
With such a high ratio of the number of dimensions to the number of samples, it is seen due to the Hughes phenomenon, the accuracy with only ten training samples is greatly reduced, about 10%. However, with additional semi-labeled samples being fed back to statistics estimation, the classification accuracy is able to clime up and quickly converges to a value which is just slightly lower than the optimal with diminishing standard deviation.



(a) d=6



(b) d=20



(c) d=40

Figure 6. Mean Accuracy for Experiment 2.

C. Experiment 3: Flight line C1

This experiment is conducted using real samples from a data designated Flightline C1 (FLC1), which is 12-band multispectral data taken over Tippecanoe County, Indiana by the M7 scanner [10] in June, 1966. The number of training samples and testing samples in each class is listed in Table 1. The training sample size was deliberately chosen to be very small, representing a poorly-posed problem where the number of training samples for each class are comparable to dimensions. Since the testing data in this experiment is very large, and in particular for some of classes with small number of samples almost entire samples of these class are included in the testing data. For this reason, the testing samples and majority of training samples are independent, and there are small overlap on the testing data and training data. Also, for the same reason, test samples which are not training samples are used as semi-labeled samples and are used to update the class statistics. Otherwise, there may not be sufficient semi-labeled samples to modify the class statistics for some minority classes. The classification results are plotted in Fig. 7, based on available ground truth for the area, a test field map is provided in Fig. 8a, and thematic map for the initial and final classifications are shown in Fig. 8b and 8c. It is seen from Fig. 7, the classification accuracy increases and converges quickly, and the final accuracy is slightly lower than 94.7%, the resubstution classification accuracy which is obtained by using all testing samples as training samples. Also, comparing Fig. 8b with Fig. 8c, the speckle error has been greatly reduced.

Table 1: Training and testing samples for Flight line C1

Class Names	No. of Testing samples	No. of training samples
Alfalfa	3,375	12
Br Soil	1,230	8
Corn	10,625	16
Oats	5,781	8
Red Clover	12,147	12
Rye	2,385	4
Soybeans	25,174	16
Water	18	4
Wheat-1	7,827	12
Wheat-2	2,091	16
Total	70,653	104

To illustrate how this proposed classifier improves itself iteratively by reducing the class statistics estimation error, the close up snapshots of the classified map for two crops are presented in Fig. (9) and Fig. (10). Figure (9) is of the rye field a little below the middle of the flightline (Figure 8). As shown in Fig. (9a), the rye training field of 4 pixels was selected in it. As illustrated in figure (9b), due to poorly estimated statistics using limited training samples only, the majority of pixels have been misclassified as something else other than rye. However, at the second iteration when semi-labeled

samples are added to enhance the statistics, there are more pixels around the training field classified as rye. This trend continues and at the last iteration, a majority of pixels in the field are eventually correctly classified as rye. In fact, some of the pixels in this rye field are not actually rye.

The second close up example involves the field of oats within a doughnut shaped wheat field just above the middle of the flightline. There are no training fields for oats in this field, and instead oats training is located elsewhere in the flightline. As expected, at the first iteration, on the test field for oats only very few pixels are correctly classified as oats. However, at the second iteration, more pixels around those pixels that have been previously classified as oats have been identified as oats. As this process continues, more and more pixels on this test field for oats have been correctly identified as oats. In figure (10f), at the fifth iteration a group of pixels of the shape of a strip across the oats field has been misclassified as wheat, this is not an error of omission for the class oats. Instead, this area is really a sod water way unplowed by the farmer. Since there are no training samples for this class of ground cover, this result further indicates that the proposed adaptive classifier adjusted itself to the next nearest class based on the information provided by the semi-labeled samples.

To show how representative the estimated parameters are, the probability map [11] associated with the classification is obtained. The probability map is determined by color coding the Mahalanobis distance of each pixel for the class to which it is classified. Blue pixels are ones that classified with low conditional probabilities. The color/likelihood scale indicates increasing likelihood from blue to yellow to red with red pixels being the ones that are classified with the highest likelihood. Figure 11 shows the probability map for the rye field of Figure 9. It is seen from this figure that when only the initial supervised learning is used the only bright spots are near the training fields. In other words, the rest of the data are not represented well. By adding semi-labeled samples to the estimation process, more representative estimates are obtained, and thus the probability maps indicate increased likelihood by the brighter, red color.

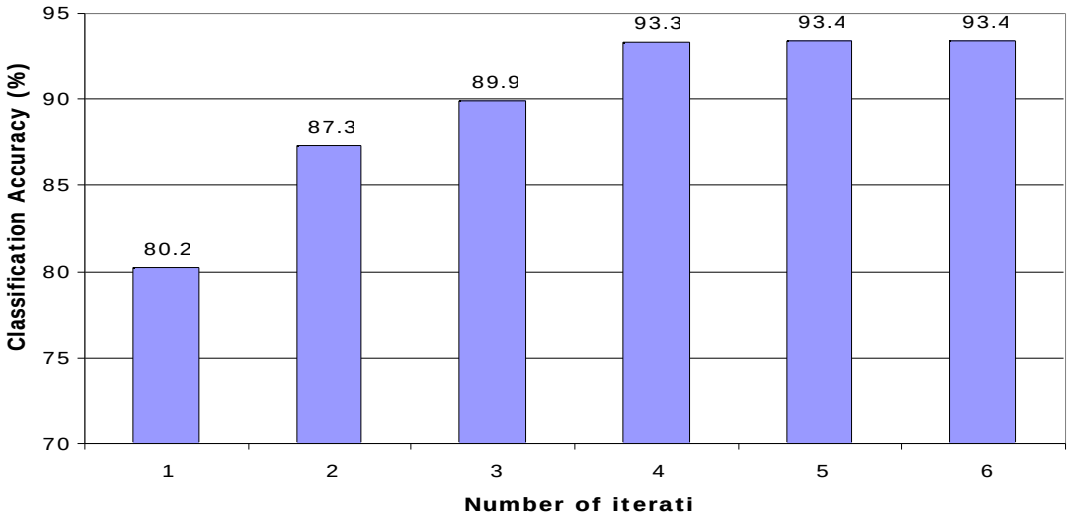
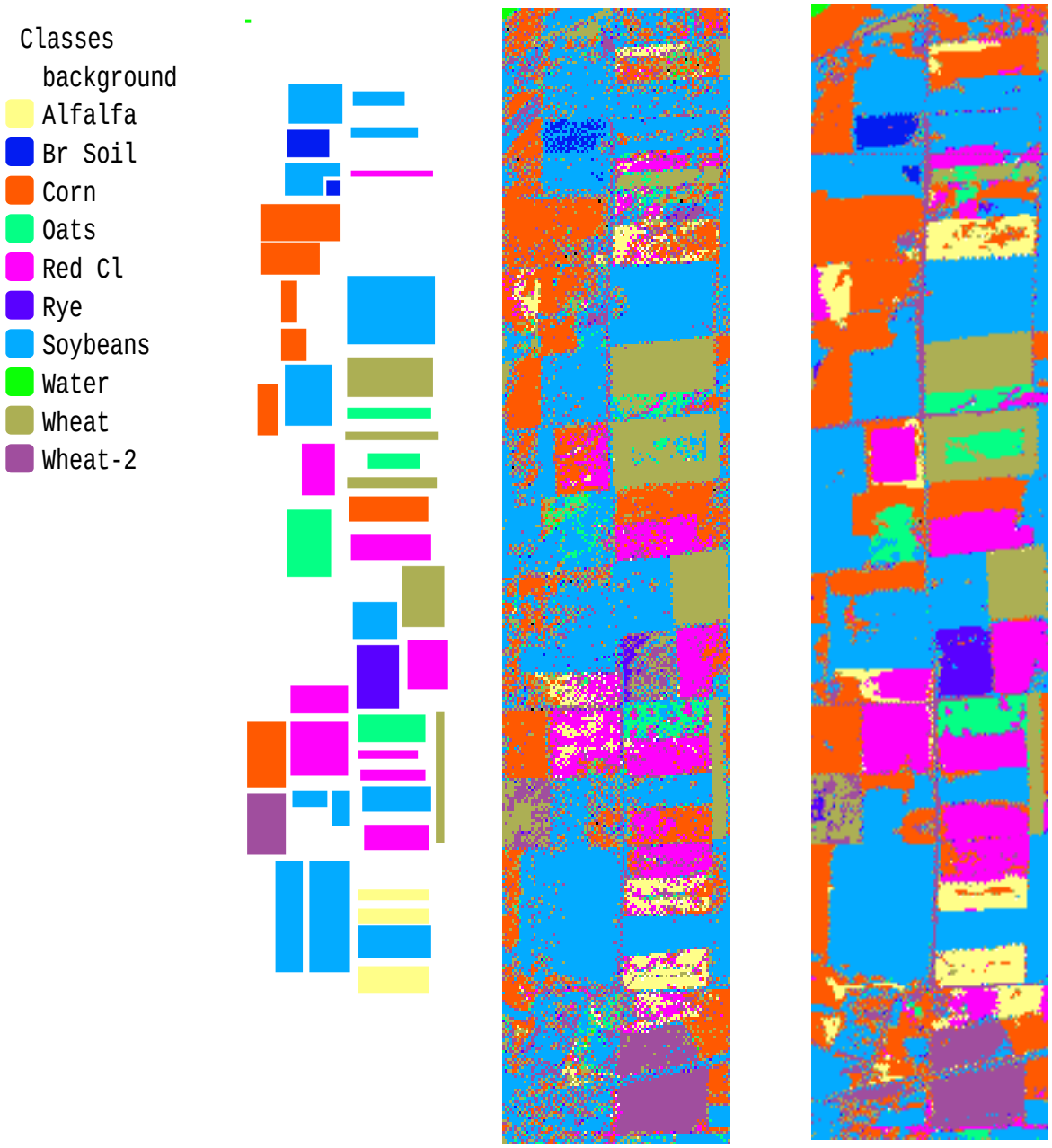
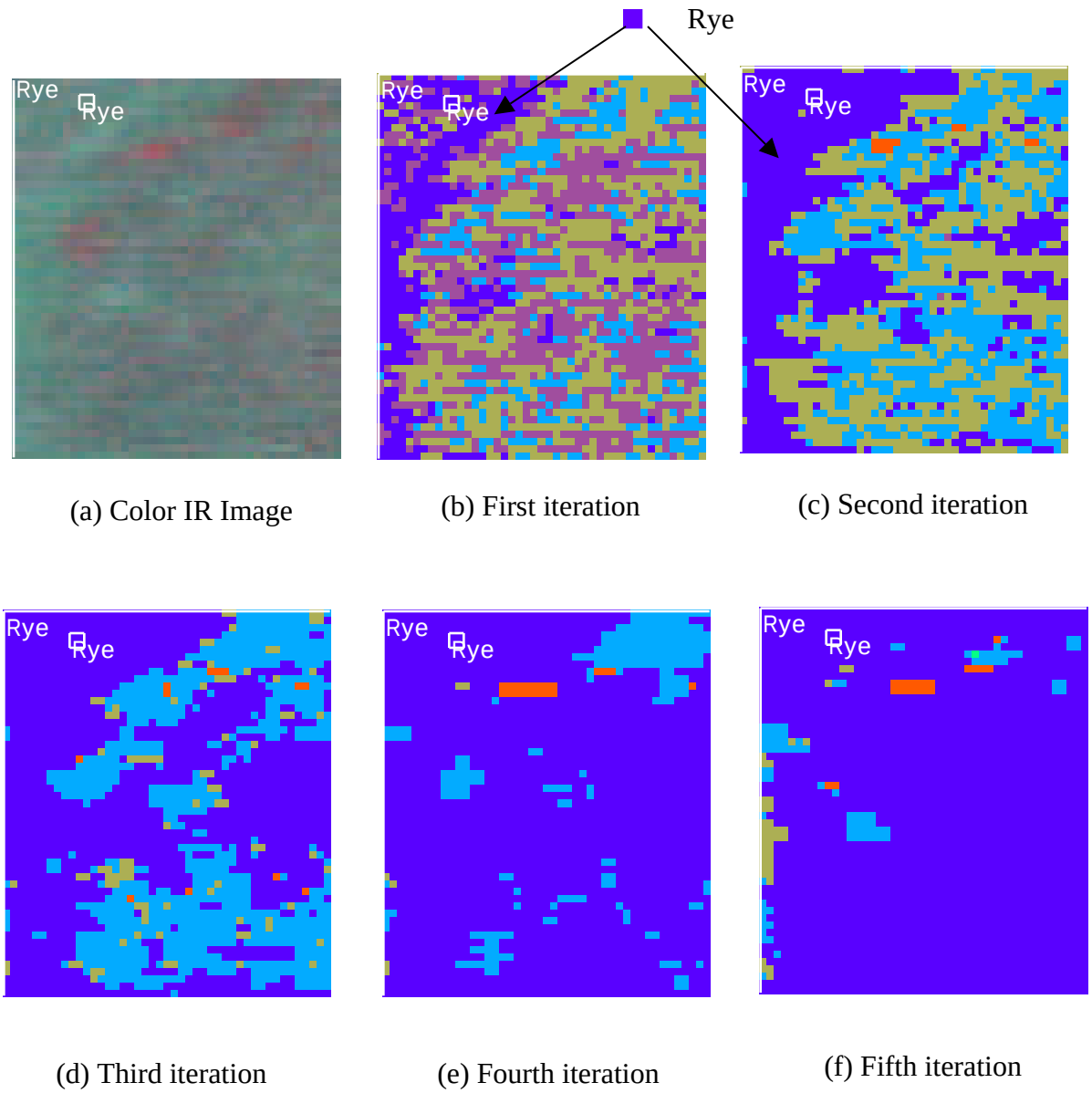


Figure 7. Classification Accuracy for Flight Line C1.



(a) Testing field (b) Initial Classification (c) Finial Classification

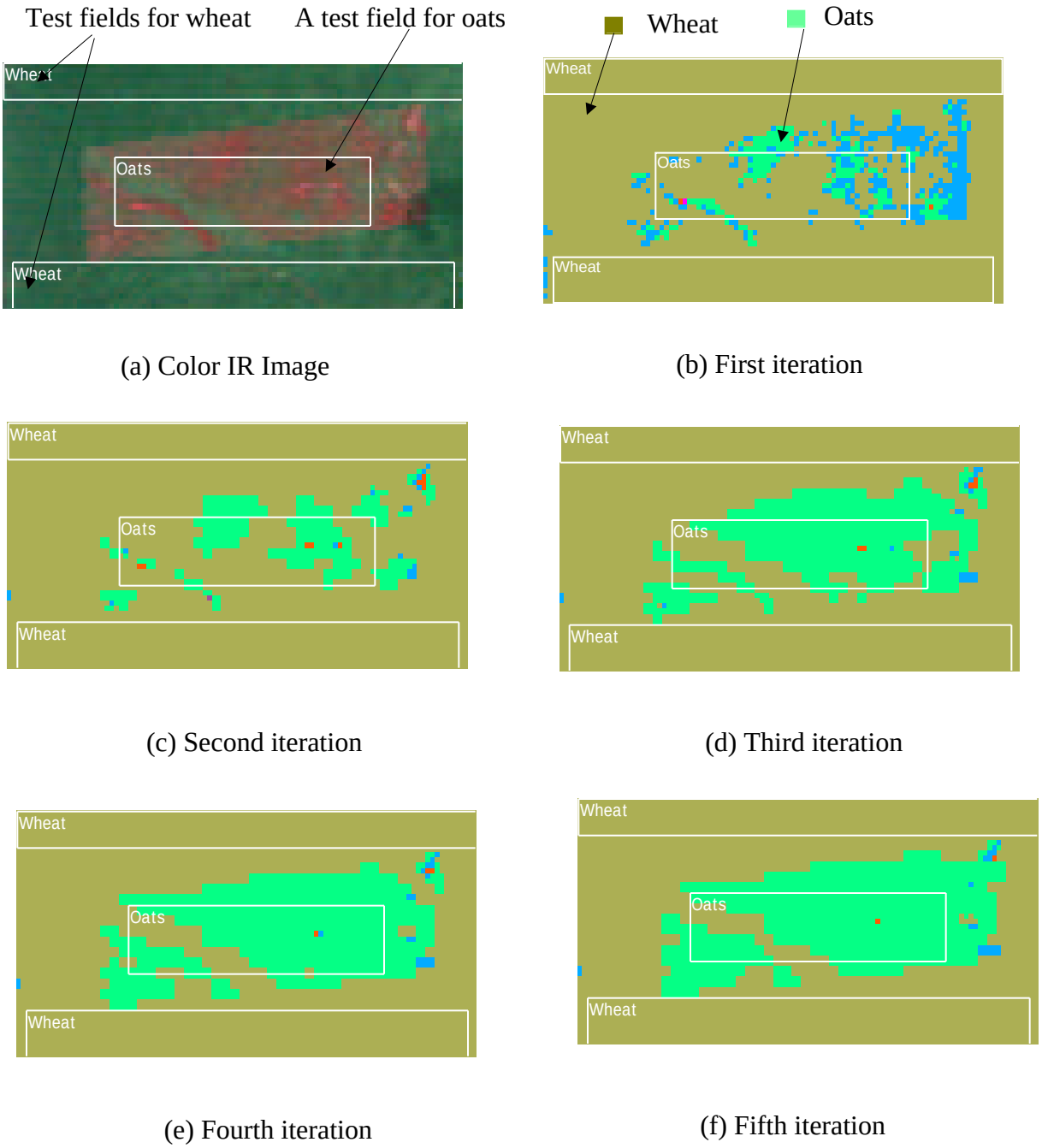
Figure 8. Test and Classification Map for Flight Line C1.



(a) Color IR Image (b) First iteration (c) Second iteration

(d) Third iteration (e) Fourth iteration (f) Fifth iteration

Figure 9. Original image and classification map
for a rye test field at each iteration.



(a) Color IR Image

(b) First iteration

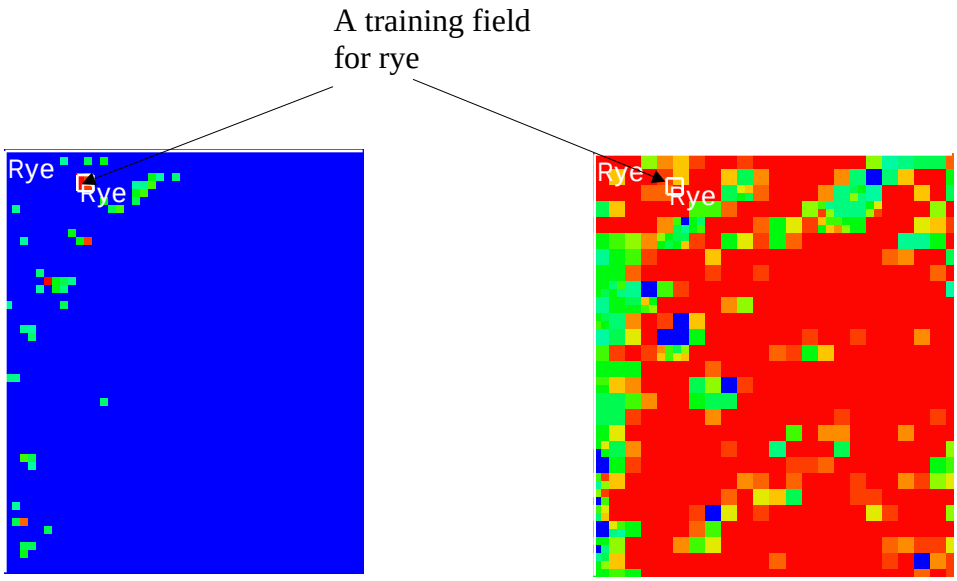
(c) Second iteration

(d) Third iteration

(e) Fourth iteration

(f) Fifth iteration

Figure 10. Original image and classification map of wheat and oats fields at each iteration.



(a) Initial iteration

(b) Final iteration

Figure 11. Portion of Probability map for
Flight Line C1.

VI. CONCLUSION

This paper is begun by investigating the information contained in semi-labeled samples of two Gaussian distributions in terms of the Fisher Information Matrix. Results show that higher classification accuracy can provide more useful class label information for statistical estimation, and so do the number of samples. The probability of error for semi-supervised learning and combined learning process is also investigated. Results indicate that when semi-labeled samples are fed back to the statistical estimation process, higher accuracy and more semi-labeled samples may enhance statistics significantly and consequently reduce the probability of error for the following classification.

Based on the above findings, a self-improving adaptive process is proposed which integrates statistical estimation and classification using semi-labeled samples. It may mitigate the Hughes phenomenon by iteratively utilizing the additional class label information extracted from classification process.

The experimental results further reveal several benefits of this classifier. First, all experiments show that the proposed adaptive classifier is able to raise classification accuracy steadily and eventually drive it close to the optimal value. Higher initial classification accuracy accelerates the rate of convergence but has little effect on the final classification.

Second, as is shown in experiment results 6a and 6b, when the separability increases with dimensionality, with semi-labeled samples, the peak performance is enhanced. In other words, the information in the new feature measurements can be used to further reduce the error. Without the semi-labeled samples, the peak performance occurs at a lower dimension after which no further improvement can be obtained from new feature measurements; instead performance deteriorates with dimensions.

Third, the estimated statistics are approaching the true ones with iterations. As is shown through all the experiments, the standard deviation is greatly reduced with iterations, indicating the estimated statistics are more and more robust. In particular, as shown in the last experiment with semi-labeled samples, most of samples are classified with high likelihood.

Despite the promising results, the proposed adaptive classifier has limitations. In particular, for a very ill-posed problem, where the number of dimensions is far greater than the number of training samples and the number of parameters is even greater than the number of all semi-labeled samples, the initial classification can be very bad. As a result a positive feedback could hardly be established and the proposed adaptive classifier may not converge. This necessitates the use of an adaptive covariance estimator, where semi-labeled samples are incorporated into the process to determine the optimal covariance mixture.

REFERENCES

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

[2] H.W. Sorenson, *Parameter Estimation: Principles and Problems*, New York: M. Dekker, 1980.

[3] K. Fukunaga, *Intro. Statistical Pattern Recognition*, 2nd. ed., New York: Academic, 1990

[4] F.A. Graybill, *Matrices With Applications in Statistics*, Belmont: Wadsworth Inc., 1983

[5] D.W. Hosmer, jr., "Information and Mixtures of two normal Distributions", *J. Statics. Comput. Simul.*, Vol. 6, pp. 137-148, 1997

[6] B.M. Shahshahani, "Classification of Multispectral Data by Joint Supervised-Unsupervised Learning," PhD Thesis and School of Electrical Engineering Technical Report TR-EE 94-1, January 1994, available from
http://dynamo.ecn.purdue.edu/~landgreb/publications.html

[7] B.M. Shahshahani and D.A. Landgrebe, "The Effect of Unlabeled Samples in Reducing the Small Sample Size Problem and Mitigating the Hughes Phenomenon", *IEEE Trans. On Geoscience and Remote Sensing*, Vol. 32, No. 5, pp 1087-1095, September 1994

[8] R.A. Redner, H.F. Walker, " Mixture Densities, Maximum Likelihood and the EM Algorithm," *SIAM Review*, Vol. 26, No. 2, pp 195-239, 1984

[9] 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, available from http://dynamo.ecn.purdue.edu/~landgreb/publications.html

[10] Swain, P.H. and S.M. Davis, eds., *Remote Sensing: The Quantitative Approach*, McGraw Hill, 1978, Chapter 2

[11] D. A. Landgrebe and L. Biehl, *An Introduction to MultiSpec*, School of Electrical Engineering, Purdue University, IN. 47907-1285, available for download from
http://dynamo.ecn.purdue.edu/~biehl/MultiSpec/documentation.html

Appendix A: Derivation of Fisher Information Matrix for Two Normal Distributions

Fisher information matrix expressed in Eq. (2) can be written as:

$$
\begin{aligned}
I_{sl} = {} & n_1 P_1 \int_{\Omega_1} [\frac{\partial}{\partial \theta} \log f(x,\theta)][\frac{\partial}{\partial \theta} \log f(x,\theta)]^T f_1(x|\mu_1, \Sigma_1) dx \\
& + n_1 P_2 \int_{\Omega_2} [\frac{\partial}{\partial \theta} \log f(x,\theta)][\frac{\partial}{\partial \theta} \log f(x,\theta)]^T f_2(x|\mu_2, \Sigma_2) dx \\
& + n_2 P_1 \int_{\Omega_2} [\frac{\partial}{\partial \theta} \log f(x,\theta)][\frac{\partial}{\partial \theta} \log f(x,\theta)]^T f_1(x|\mu_1, \Sigma_1) dx \\
& + n_2 P_2 \int_{\Omega_1} [\frac{\partial}{\partial \theta} \log f(x,\theta)][\frac{\partial}{\partial \theta} \log f(x,\theta)]^T f_2(x|\mu_2, \Sigma_2) dx
\end{aligned}
$$

Since the vector of unknown parameters is $\theta = [\mu_1^T, \mu_2^T]^T$, therefore:

$$
\frac{\partial}{\partial \theta} \log f_1(x) = \frac{1}{f_1(x)} \frac{\partial}{\partial \theta} f_1(x) = \frac{1}{f_1(x)} \begin{bmatrix} f_1(x)(x-\mu_1)\Sigma_1^{-1}(x-\mu_1)^T \\ 0 \end{bmatrix}
$$

$$
\frac{\partial}{\partial \theta} \log f_2(x) = \frac{1}{f_2(x)} \frac{\partial}{\partial \theta} f_2(x) = \frac{1}{f_2(x)} \begin{bmatrix} 0 \\ f_2(x)(x-\mu_2)\Sigma_2^{-2}(x-\mu_2)^T \end{bmatrix}
$$

With $\mu_1 = 0$ and $\Sigma_1 = \Sigma_2 = I_d$, the above can be simplified as:

$$
\frac{\partial}{\partial \theta} \log f_1(x) = \frac{1}{f_1(x)} \frac{\partial}{\partial \theta} f_1(x) = \frac{1}{f_1(x)} \begin{bmatrix} f_1(x)xx^T \\ 0 \end{bmatrix}
$$

$$
\frac{\partial}{\partial \theta} \log f_2(x) = \frac{1}{f_2(x)} \frac{\partial}{\partial \theta} f_2(x) = \frac{1}{f_2(x)} \begin{bmatrix} 0 \\ f_2(x)(x-\mu_2)(x-\mu_2)^T \end{bmatrix}
$$

Also, in the canonical case under consideration, the subspaces Ω_1 and Ω_2 can be determined as:

$$
\begin{aligned}
x \quad \Omega_1 &\Leftrightarrow x_1 \le t \\
x \quad \Omega_2 &\Leftrightarrow x_1 > t
\end{aligned}
$$

where

$$
t = \frac{1}{\Delta} \log(\frac{P_1}{P_2}) + \frac{1}{2} \Delta
$$

If we define:

$$I_1 = \int_{\Omega_1} [\frac{\partial}{\partial \theta} \log f_1(x)][\frac{\partial}{\partial \theta} \log f_1(x)]^T f_1(x|\mu_1, \Sigma_1) dx$$

$$I_2 = \int_{\Omega_2} [\frac{\partial}{\partial \theta} \log f_2(x)][\frac{\partial}{\partial \theta} \log f_2(x)]^T f_2(x|\mu_2, \Sigma_2) dx$$

$$I_3 = \int_{\Omega_2} [\frac{\partial}{\partial \theta} \log f_1(x)][\frac{\partial}{\partial \theta} \log f_1(x)]^T f_1(x|\mu_1, \Sigma_1) dx$$

$$I_4 = \int_{\Omega_1} [\frac{\partial}{\partial \theta} \log f_2(x)][\frac{\partial}{\partial \theta} \log f_2(x)]^T f_2(x|\mu_2, \Sigma_2) dx$$

then we have:

$$I_1 = \begin{bmatrix} \alpha_1 & 0 & 0 \\ 0 & \beta_1 I_{d-1} & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

$$I_2 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & \alpha_2 & 0 \\ 0 & 0 & \beta_2 I_{d-1} \end{bmatrix}$$

$$I_3 = \begin{bmatrix} 1-\alpha_1 & 0 & 0 \\ 0 & (1-\beta_1) I_{d-1} & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

$$I_4 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1-\alpha_2 & 0 \\ 0 & 0 & (1-\beta_2) I_{d-1} \end{bmatrix}$$

$$k_1 = r_c \alpha_1 + (1 - r_c)(1 - \alpha_1)$$

$$k_2 = r_c \beta_1 + (1 - r_c)(1 - \beta_1)$$

$$k_3 = r_c \alpha_2 + (1 - r_c)(1 - \alpha_2)$$

$$k_4 = r_c \beta_2 + (1 - r_c)(1 - \beta_2)$$

$$r_c = \frac{n_1}{n}$$

$$\alpha_1 = \Phi(t) - t\phi(t)$$

$$\beta_1 = \Phi(t)$$

$$\alpha_2 = \Phi(\Delta - t) - (t - \Delta)\phi(t - \Delta)$$

$$\beta_2 = \Phi(\Delta - t)$$

$$\Phi(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-\frac{x^2}{2}} dx$$

$$\phi(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}$$