Supervised Classification in High Dimensional Space: Geometrical, Statistical and Asymptotical Properties of Multivariate Data[1]

Luis Jimenez and David Landgrebe[2]
Dept. Of ECE, PO Box 5000
University Of Puerto Rico
Mayaguez PR
00681
jimenez@exodo.upr.clu.edu

School of Elect. & Comp. Eng.
Purdue University,
West Lafayette, IN
47907-1285
landgreb@ecn.purdue.edu

 This paper appeared in the IEEE Transactions on Systems, Man, and Cybernetics, Volume 28 Part C Number 1, pp. 39-54, Feb. 1998.

Abstract

The recent development of more sophisticated remote sensing systems enables the measurement of radiation in many more spectral intervals than previous possible. An example of this technology is the AVIRIS system, which collects image data in 220 bands. The increased dimensionality of such hyperspectral data greatly enhances the data information content but provides a challenge to the current techniques for analyzing such data.

Human experience in three dimensional space tends to mislead one's intuition of geometrical and statistical properties in high dimensional space, properties which must guide our choices in the data analysis process. Using Euclidean and Cartesian geometry, in this paper high dimensional space properties are investigated and their implication for high dimensional data and its analysis is studied in order to illuminate the differences between conventional spaces and hyperdimensional space.

I. Introduction

The complexity of dimensionality has been known for more than three decades, and its impact varies from one field to another. In combinatorial optimization over many dimensions, it is seen as an exponential growth of the computational effort with the number of dimensions. In statistics, it manifests itself as a problem with parameter or density estimation due to the paucity of data. The negative effect of this paucity results from some geometrical, statistical and asymptotical properties of high dimensional feature space. These characteristics exhibit surprising behavior of data in higher dimensions.

1 Work reported herein was funded in part by NASA Grant NAGW-3924.
2 Corresponding author.

There are many assumptions that we make about characteristics of lower dimensional spaces based on our experience in three dimensional Euclidean space. There is a conceptual barrier that makes it difficult to have proper intuition of the properties of high dimensional space and its consequences in high dimensional data behavior. Most of the assumptions that are important for statistical purposes we tend to relate to our three dimensional space intuition, for example, as to where the concentration of volume is of such figures as cubes, spheres, and ellipsoids or where the data concentration is in known density function families such as normal and uniform. Other important perceptions that are relevant for statistical analysis are, for example, how the diagonals relate to the coordinates, the number of labeled samples required for supervised classification, the assumption of normality in data, and the importance of mean and covariance difference in the process of discrimination among different statistical classes. In the next section some characteristics of high dimensional space will be studied, and their impact in supervised classification data analysis will be discussed. Most of these properties do not fit our experience in three dimensional Euclidean space as mentioned before.

II. Geometrical, Statistical and Asymptotical Properties

In this section we illustrate some unusual or unexpected hyperspace characteristics including a proof and discussion. These illustrations are intended to show that higher dimensional space is quite different from the three dimensional space with which we are familiar.

As dimensionality increases:

A. *The volume of a hypercube concentrates in the corners* [15].

It has been shown [9] that the volume of the hypersphere of radius r and dimension d is given by the equation:

$$V_s(r) = \text{volume of a hypersphere} = \frac{2r^d}{d}\frac{\pi^{\frac{d}{2}}}{\Gamma\left(\frac{d}{2}\right)} \qquad (1)$$

and that the volume of a hypercube in $[-r, r]^d$ is given by the equation:

$$V_c(r) = \text{volume of a hypercube} = (2r)^d \qquad (2)$$

The fraction of the volume of a hypersphere inscribed in a hypercube is:

$$f_{d1} = \frac{V_s(r)}{V_C(r)} = \frac{\pi^{d/2}}{d2^{d-1}\Gamma\left(d/2\right)} \qquad (3)$$

where d is the number of dimensions. We see in Figure 1 how f_{d1} decreases as the dimensionality increases.

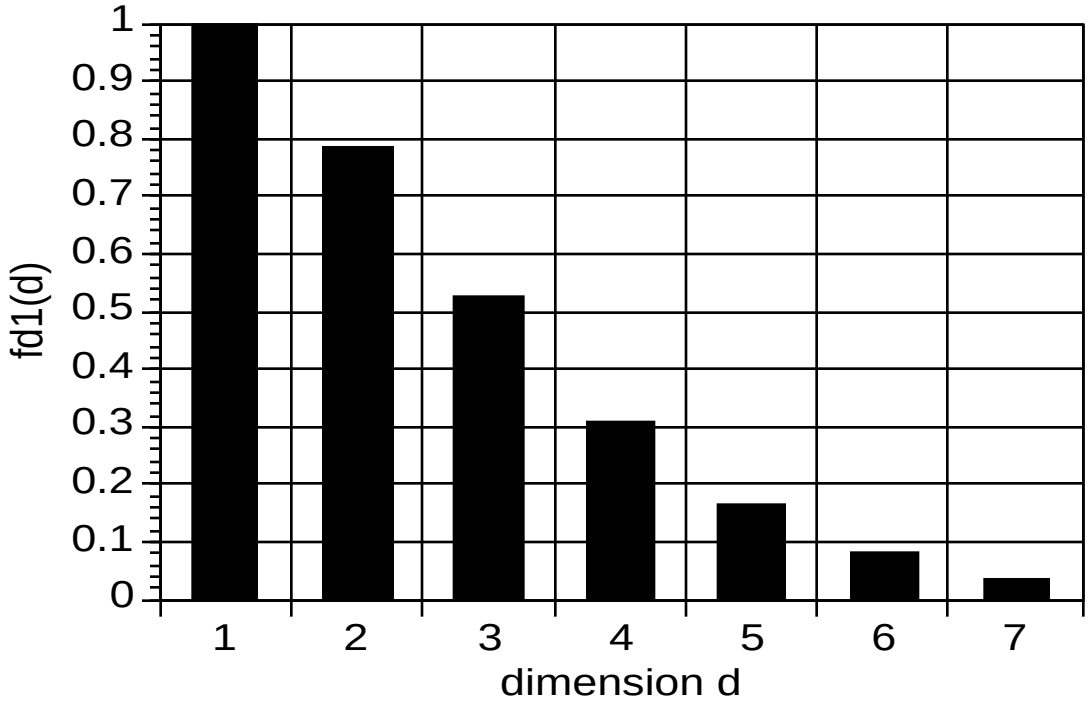


Figure 1. Fractional volume of a hypersphere inscribed in a hypercube as a function of dimensionality.

Note that $\lim_{d\to\infty} f_{d1} = 0$ which implies that the volume of the hypercube is increasingly concentrated in the corners as d increases.

B. *The volume of a hypersphere concentrates in an outside shell* [15, 16].

The fraction of the volume in a shell defined by a sphere of radius r-ε inscribed inside a sphere of radius r is:

$$f_{d2} = \frac{V_d(r) - V_d(r-\varepsilon)}{V_d(r)} = \frac{r^d - (r-\varepsilon)^d}{r^d} = 1 - \left(1 - \frac{\varepsilon}{r}\right)^d$$

In Figure 2 we can observe, for the case $\varepsilon = r/5$, how as the dimension increases the volume concentrates in the outside shell.

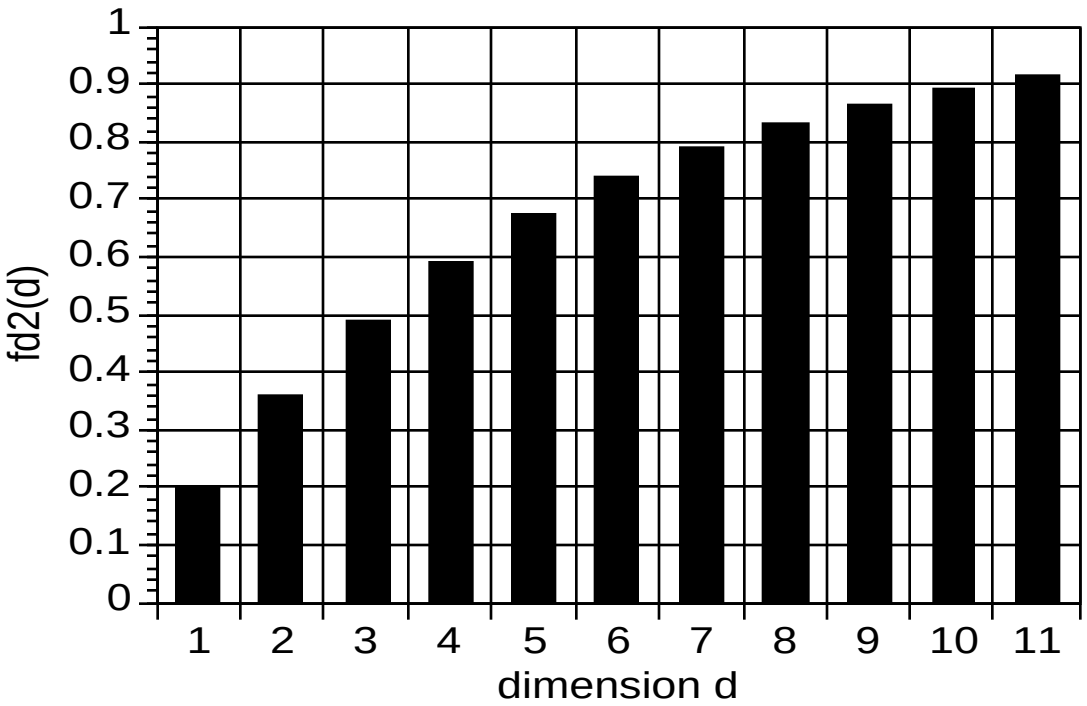


Figure 2. Volume of a hypersphere contained in the outside shell as a function of dimensionality for $\varepsilon = r/5$.

Note that $\lim_{d\to\infty} f_{d2} = 1, \quad \varepsilon > 0$, implying that most of the volume of a hypersphere is concentrated in an outside shell.

C. *The volume of a hyperellipsoid concentrates in an outside shell.*

Here the previous result will be generalized to a hyperellipsoid. Let the equation of a hyperellipsoid in d dimensions be written as:

$$\frac{X_1^2}{\lambda_1^2} + \frac{X_2^2}{\lambda_2^2} + \cdots + \frac{X_d^2}{\lambda_d^2} = 1$$

The volume is calculated by the equation [9]:

$$V_e(\lambda_i) = \frac{2\prod_{i=1}^{d}\lambda_i}{d}\,\frac{\pi^{\frac{d}{2}}}{\Gamma\left(\frac{d}{2}\right)}$$

The volume of a hyperellipsoid defined by the equation:

$$\frac{X_1^2}{(\lambda_1-\delta_1)^2} + \frac{X_2^2}{(\lambda_2-\delta_2)^2} + \cdots + \frac{X_d^2}{(\lambda_d-\delta_d)^2} = 1$$

where $0 \le \delta_i < \lambda_i, \quad i$, is calculated by:

$$V_e(\lambda_i - \delta_i) = \frac{2\prod_{i=1}^{d}(\lambda_i - \delta_i)}{d} \frac{\pi^{\frac{d}{2}}}{\Gamma\left(\frac{d}{2}\right)}$$

The fraction of the volume of $V_e(\lambda_i - \delta_i)$ inscribed in the volume $V_e(\lambda_i)$ is:

$$f_{d3} = \frac{\prod_{i=1}^{d}(\lambda_i - \delta_i)}{\prod_{i=1}^{d}\lambda_i} = \prod_{i=1}^{d}\left(1 - \frac{\delta_i}{\lambda_i}\right)$$

Let $\gamma_{min} = \min\left(\frac{\delta_i}{\lambda_i}\right)$, then

$$f_{d3} = \prod_{i=1}^{d}\left(1 - \frac{\delta_i}{\lambda_i}\right) \leq \prod_{i=1}^{d}(1 - \gamma_{min}) = (1 - \gamma_{min})^d$$

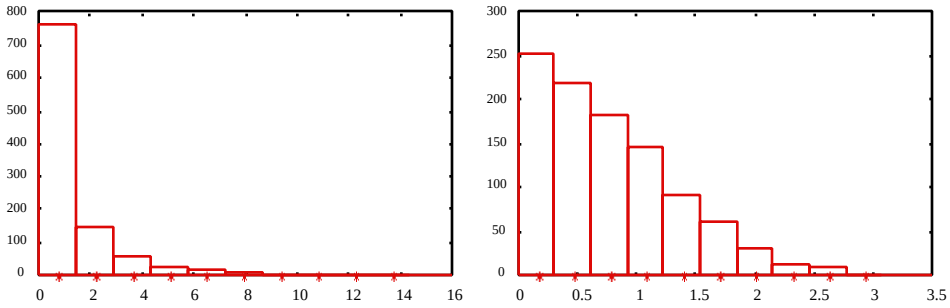
Using the fact that $f_{d3} \geq 0$, it is concluded that $\lim_{d \to \infty} f_{d3} = 0$.

The characteristics previously mentioned have two important consequences for high dimensional data that appear immediately. The first one is that high dimensional space is mostly empty, which implies that multivariate data in R^d is usually in a lower dimensional structure. As a consequence high dimensional data can be projected to a lower dimensional subspace without losing significant information in terms of separability among the different statistical classes. The second consequence of the foregoing, is that normally distributed data will have a tendency to concentrate in the tails; similarly, uniformly distributed data will be more likely to be collected in the corners, making density estimation more difficult. Local neighborhoods are almost surely empty, requiring the bandwidth of estimation to be large and producing the effect of losing detailed density estimation.

Support for this tendency can be found in the statistical behavior of normally and uniformly distributed multivariate data at high dimensionality. It is expected that as the dimensionality increases the data will concentrate in an outside shell. As the number of dimensions increases that shell will increase its distance from the origin as well.

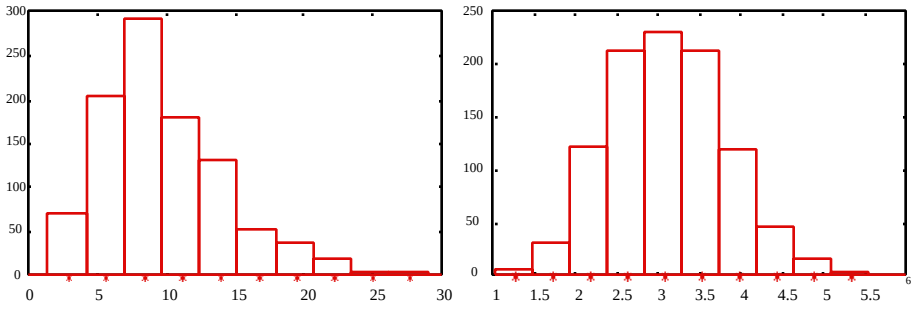
To show this specific multivariate data behavior, an experiment was developed. Multivariate normal and uniform distributed data were generated. The normal and uniform variables are independent identically distributed samples from the distributions N(0,1) and U(-1,1), respectively. Figures 3 and 4 illustrate the histograms of random variables, the distance from the zero coordinate and its square, that are functions of normal or uniform vectors for different number of dimensions.

Normal, dimensions = 1



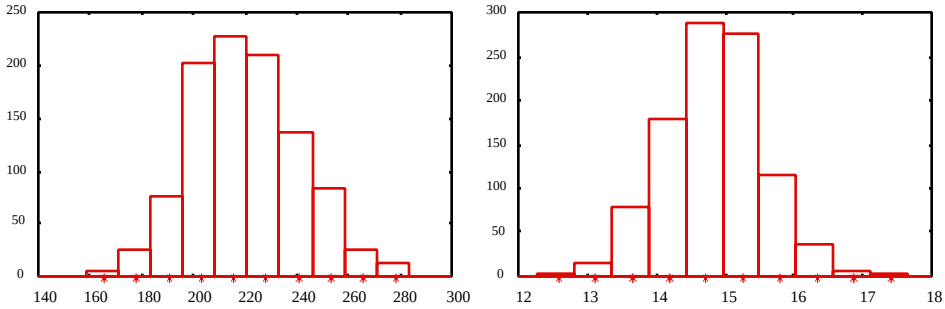
$\sum_i x_i^2, \mu = 1.0168, \sigma = 1.4017$ $\sqrt{\sum_i x_i^2}, \mu = .7736, \sigma = .5737$

Normal, dimensions = 10



$\sum_i x_i^2, \mu = 9.8697, \sigma = 4.5328$ $\sqrt{\sum_i x_i^2}, \mu = 3.1026, \sigma = 0.7042$

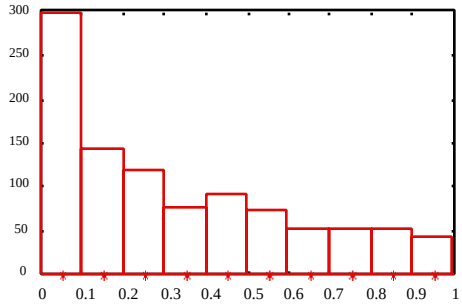
Normal, dimensions = 220

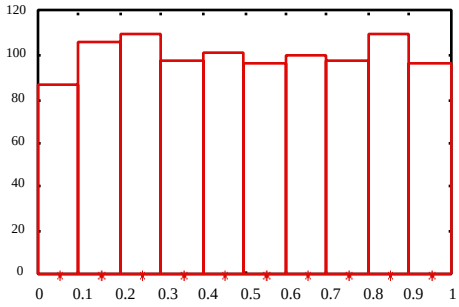


$\sum_i x_i^2, \mu = 220.3732, \sigma = 20.7862$ $\sqrt{\sum_i x_i^2}, \mu = 14.8497, \sigma = 0.7129$

Figure 3. Histograms of functions of Normally distributed random variables.

Uniform, dimensions = 1

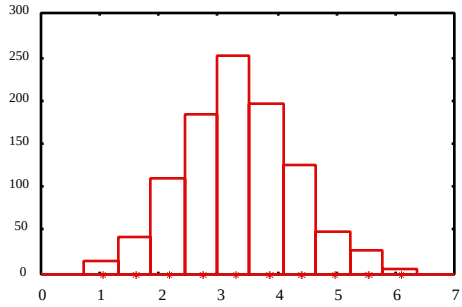


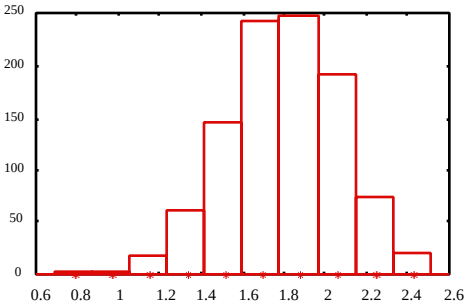


$\sum_i x_i^2, \mu = 0.3277, \sigma = 0.2883$

$\sqrt{\sum_i x_i^2}, \mu = 0.5041, \sigma = 0.2887$

Uniform, dimensions = 10

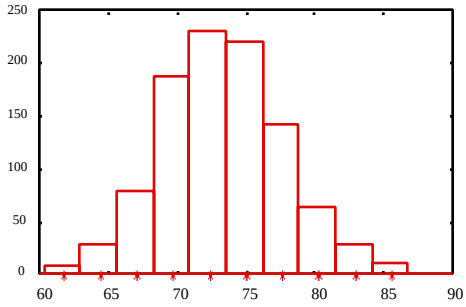


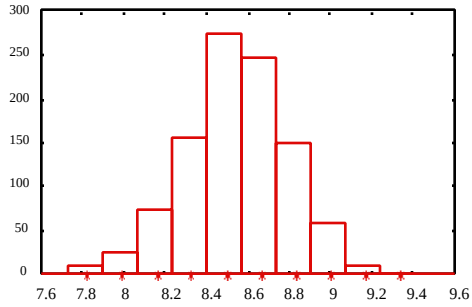


$\sum_i x_i^2, \mu = 3.3444, \sigma = 0.9390$

$\sqrt{\sum_i x_i^2}, \mu = 1.8010, \sigma = 0.2678$

Uniform, dimensions = 220





$\sum_i x_i^2, \mu = 73.3698, \sigma = 4.3854$

$\sqrt{\sum_i x_i^2}, \mu = 8.5488, \sigma = 0.2505$

Figure 4. Histograms of functions of Uniformly distributed random variables.

These experiments show how the means and the standard deviations are functions of the number of dimensions. As the dimensionality increases the data concentrates in an outside shell. The mean and standard deviation of two random variables

$$r = \sqrt{\sum_{i=1}^{d} x_i^2} \quad \text{and} \quad R = \sum_{i=1}^{d} x_i^2$$

are computed. These variables are the distance and the square of the Euclidean distance of the random vectors. The values of the parameters and the histograms of the random variables are shown in Figure 3 and 4 for normal and uniform distribution of the data. As the dimensionality increases the distance from the zero coordinate of both random variables increases as well. These results show that the data have a tendency to concentrate in an outside shell and how the shell's distance from the zero coordinate increases with the increment of the number of dimensions.

Note that $R = \sum_{i=1}^{d} x_i^2$ has a chi-square distribution with d degrees of freedom when the x_i's are samples from the N(0,1) distribution. The mean and variance of R are: E(R) = d, Var(R) = 2d [14, pp. 62-64]. This conclusion supports the previous thesis.

Under these circumstances it would be difficult to implement any density estimation procedure and to obtain accurate results. Generally nonparametric approaches will have even greater problems with high dimensional data.

D. *The diagonals are nearly orthogonal to all coordinate axis* [15, pp. 27-31] [16].

The cosine of the angle between any diagonal vector and a Euclidean coordinate axis is:

$$\cos(\theta_d) = \pm \frac{1}{\sqrt{d}} \ ,$$

Figure 5 illustrates how the angle between the diagonal and the coordinates, θ_d, approaches 90^o with increases in dimensionality.

Note that $\lim_{d \to \infty} \cos(\theta_d) = 0$, which implies that in high dimensional space the diagonals have a tendency to become orthogonal to the Euclidean coordinates.

This result is important because the projection of any cluster onto any diagonal, e.g., by averaging features, could destroy information contained in multispectral data. In order to explain this, let $\mathbf{a_{diag}}$ be any diagonal in a d dimensional space. Let $\mathbf{ac_i}$ be the ith coordinate of that space. Any point in the space can be represented by the form:

$$\mathbf{P} = \sum_{i=1}^{d} \alpha_i \mathbf{ac}_i$$

The projection of $\mathbf{P}$ over $\mathbf{a_{diag}}$, $\mathbf{P_{diag}}$ is:

$$\mathbf{P}_{\mathbf{diag}} = \left(\mathbf{P}^T \mathbf{a}_{\mathbf{diag}}\right)\mathbf{a}_{\mathbf{diag}} = \sum_{i=1}^{d} \alpha_i \left(\mathbf{ac}_i^T \mathbf{a}_{\mathbf{d}}\right)\mathbf{a}_{\mathbf{d}}$$

But as d increases $\mathbf{ac}_i^T \mathbf{a}_{\mathbf{diag}} \approx 0$ which implies that $\mathbf{P}_{\mathbf{diag}} \approx \mathbf{0}$. As a consequence $\mathbf{P_{diag}}$ is being projected to the zero coordinate, losing information about its location in the d dimensional space.

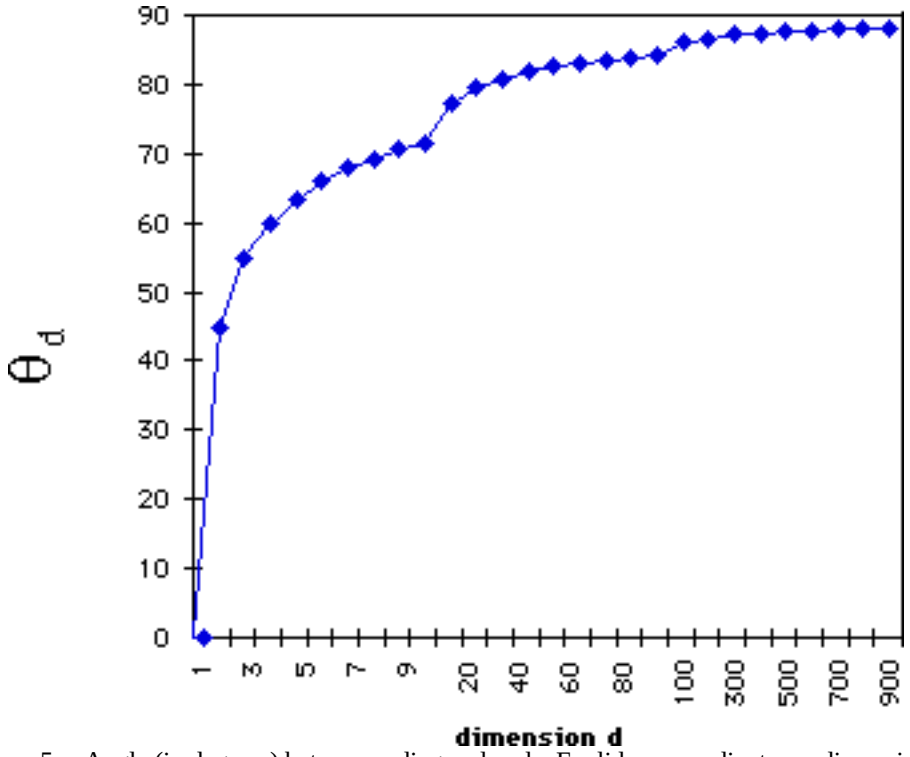


Figure 5. Angle (in degrees) between a diagonal and a Euclidean coordinate vs. dimensionality.

E. *The required number of labeled samples for supervised classification increases as a function of dimensionality.*

Fukunaga [2] proves that the required number of training samples is linearly related to the dimensionality for a linear classifier and to the square of the dimensionality for a quadratic classifier. That fact is very relevant, especially since experiments have demonstrated that there are circumstances where second order statistics are more relevant than first order statistics in discriminating among classes in high dimensional data [13]. In terms of nonparametric classifiers the situation is even more severe. It has been estimated that as the number of dimensions increases, the sample size needs to increase exponentially in order to have an effective estimate of multivariate densities [15, pp. 208-212] [5].

It is reasonable to expect that high dimensional data contains more information in the sense of a capability to detect more classes with more accuracy. At the same time the above characteristics tell us that current techniques, which are usually based on computations at full dimensionality, may not deliver this advantage unless the available labeled data is substantial. This was proven by [4] who proved that with a limited number of training samples there is a penalty in classification accuracy as the number of features increases beyond some point.

F. *For most high dimensional data sets', low linear projections have the tendency to be normal, or a combination of normal distributions, as the dimension increases.*

That is a significant characteristic of high dimensional data that is quite relevant to its analysis. It has been proved [1, 3] that as the dimensionality tends to infinity, lower dimensional linear projections will approach a normality model with probability approaching one (see Figure 6). Normality in this case implies a normal or a combination of normal distributions.

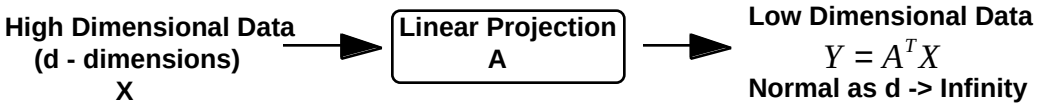


Figure 6. The tendency of lower dimensional projections to be normal.

Several experiments will illustrate this with simulated and real data. The procedure in these experiments is to project the data from a high dimensional space to a one dimensional subspace. We examine the behavior of the projected data as the number of dimensions in the original high dimensional space increases from one to ten and finally to one hundred. The method of projecting the data is to multiply it with a normal vector with random angles from the coordinates. A histogram is used to observe the data distribution. A normal density function is plotted with the histogram to compare the results to normal.

Figure 7 shows the case of generated data from a uniform distribution. As the number of dimensions increases in the original space the projected data's histogram has a tendency to be normal. Figure 8 shows the results of the same experiment with real AVIRIS data with one soybeans class. Note that the results are similar to the generated data.

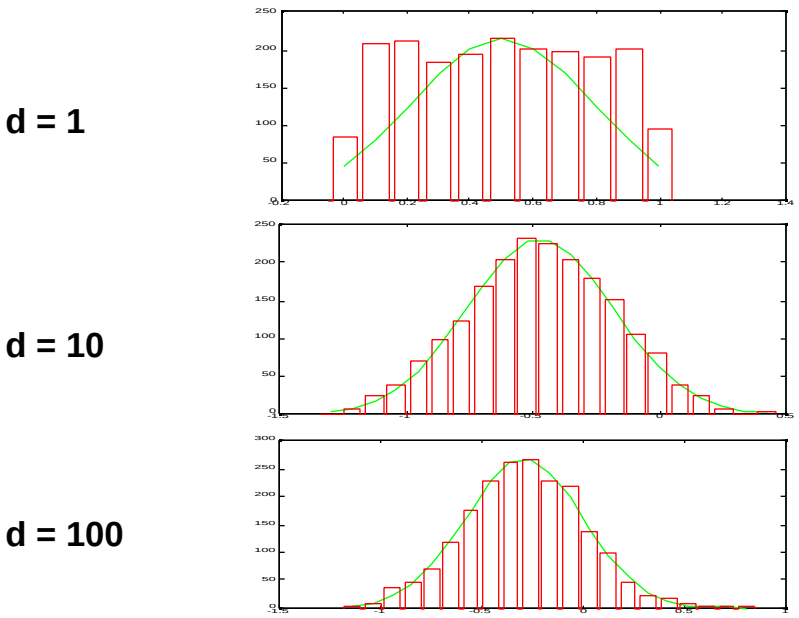


Figure 7. Generated data: One class with uniform distribution.

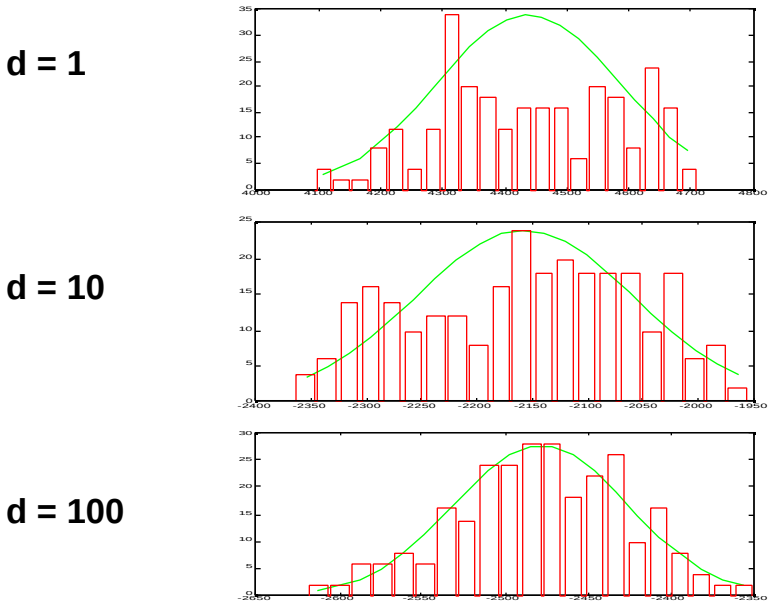


Figure 8. AVIRIS Multispectral data: One class, soybeans.

These results tempt us to expect that the data can be assumed to be a combination of normal distributions in the projected subspace without any problem. Other experiments show that a combination of normal distributions where each one represents a different statistical class could collapse into one normal distribution. That will imply loss of information. Figures 9 and 10 show the result of repeating the experiments for a two class problem. Both show the risk of damaging data projecting it into one normal distribution loosing separability and information. In the case of Figure 10 we have real AVIRIS data with a corn and a soybeans class.

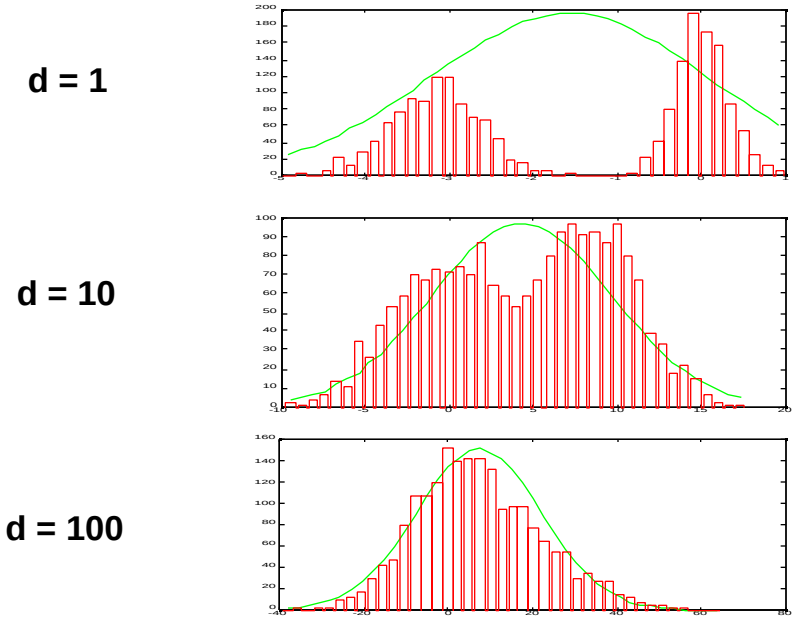


Figure 9. Generated data: Two classes with normal distributions.

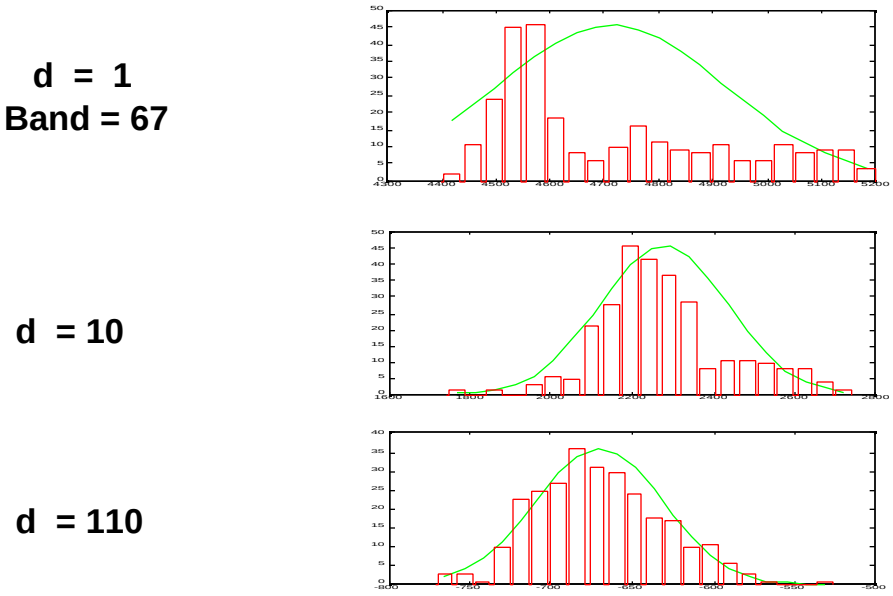


Figure 10. AVIRIS Multispectral data: Two classes, corn and soybeans.

In all the cases above we can see the advantage of developing an algorithm that will estimate the projection directions that separate the explicitly defined classes, doing the computations in a lower dimensional space. The vectors that it computes will separate the classes, and at the same time, the explicitly defined classes will behave asymptotically more like a normal distribution. The assumption of normality will be better grounded in the projected subspace than in full dimensionality.

III. Asymptotical first and second order statistics properties

Lee and Landgrebe [13] performed an experiment where they classified some high dimensional data in order to see the relative role that first and second order statistics played. Here a more general basis will be given for the role of the first and second order statistics in hyperspectral data where adjacent bands could be correlated in any way. The results will be based on the asymptotic behavior of high dimensional data. This will aid in the understanding of the conditions required for the predominance of either first order or second order statistics in the discrimination among the statistical classes in high dimensional space.

It is reasonable to assume that, as the number of features increases, the potential information content in multispectral data increases as well. In supervised classification that increment of information is translated to the number of classes and their separability. We will use Bhattacharyya distance here as the measure of separability. It provides a bound of classification accuracy taking into account first and second order statistics. Bhattacharyya distance is the sum of two components, one based primarily on mean differences and the other based on covariance differences.

The Bhattacharyya distance under the assumption of normality is computed by the equation:

$$\mu = \frac{1}{8}\left(M_2 - M_1\right)^T \left[\frac{\Sigma_1 + \Sigma_2}{2}\right]^{-1} \left(M_2 - M_1\right) + \frac{1}{2}\ln \frac{\left|\frac{\Sigma_1 + \Sigma_2}{2}\right|}{\sqrt{|\Sigma_1||\Sigma_2|}}$$

We will denote $\mu = \mu_M + \mu_C$ where μ_M, the first term on the right, is the mean difference component and μ_C, the second term on the right, is the covariance difference component.

In order to see how Bhattacharyya distance and its mean and covariance components can aid in the understanding of the role of first and second order statistics, two experiments are presented. The first one has conditions where second order statistics are more relevant in discriminating among the classes. The second experiment has conditions for the predominance of first order statistics.

Experiment 1

In this experiment data are generated for two classes. Both classes belong to normal distributions with different means and covariances. Each class has 500 points. Their respective parameters are:

$$M_1 = [0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0]^T$$

$$M_2 = [1.5 \quad 1 \quad .5 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0]^T$$

$$\Sigma_1 = \begin{bmatrix} 1 & & & & & & & & & 0 \\ & 1 & & & & & & & & \\ & & 1 & & & & & & & \\ & & & 1 & & & & & & \\ & & & & 1 & & & & & \\ & & & & & 1 & & & & \\ & & & & & & 1 & & & \\ & & & & & & & 1 & & \\ & & & & & & & & 1 & \\ 0 & & & & & & & & & 1 \end{bmatrix}, \quad \Sigma_2 = \begin{bmatrix} 1.5 & & & & & & & & & 0 \\ & 1.9 & & & & & & & & \\ & & 3 & & & & & & & \\ & & & 3 & & & & & & \\ & & & & 3 & & & & & \\ & & & & & 3 & & & & \\ & & & & & & 3 & & & \\ & & & & & & & 3 & & \\ & & & & & & & & 3 & \\ 0 & & & & & & & & & 3 \end{bmatrix}$$

The data is classified by three classifiers, the ML classifier, the ML (ML Cov) classifier constrained to use only covariance difference, and the minimum distance classifier (Min Dist). This enables us to have similar conditions to Lee and Landgrebe's experiment. The results is shown in Figure 11.

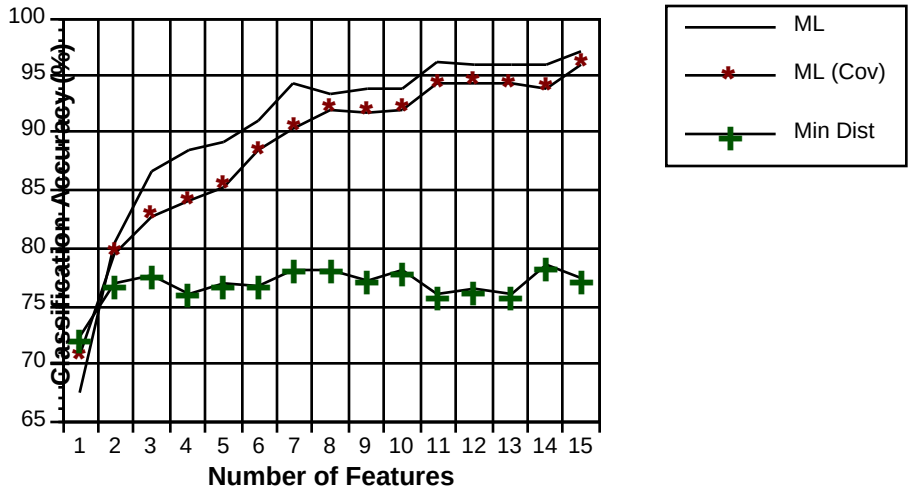


Figure 11. Performance comparison of Gaussian ML, Gaussian ML with zero mean data, and Minimum Distance classifier. Two generated classes.

The results resemble Lee and Landgrebe's results. In order to have an understanding of the roles played by first and second order statistics the mean (Bhatt Mean) and covariance (Bhatt Cov) components of Bhattacharyya distance and its sum were computed and are shown in Figure 12. Their ratio of Bhatt Mean / Bhatt Cov was calculated and shown in Figure 13.

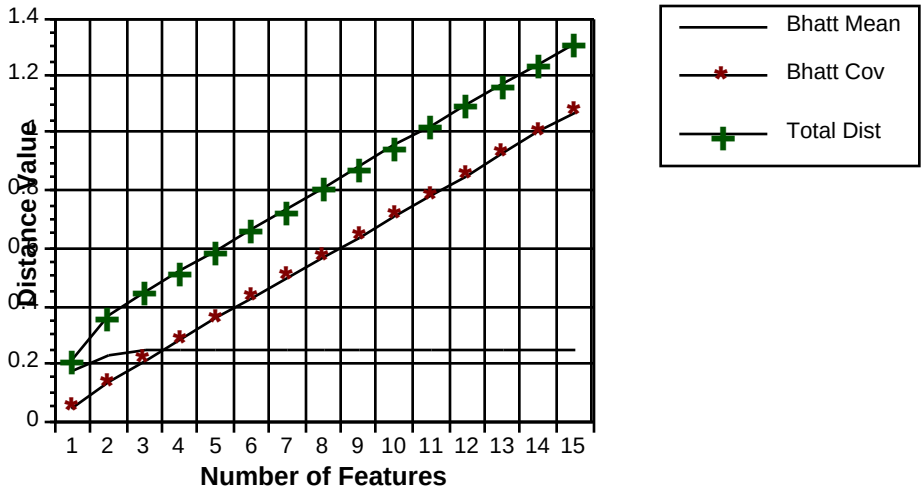


Figure 12. Bhattacharyya distance and its mean and covariance components.

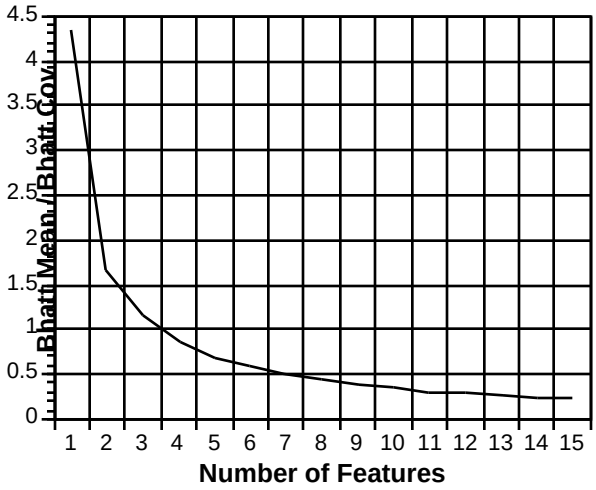


Figure 13. Ratio of Bhattacharyya distance mean component over the covariance component.

Both figures show that there is a relationship between second order statistics predominance and Bhatt Cov relevance. As the number of dimensions increases the ratio Bhatt Mean / Bhatt Cov decreases significantly and the ML Cov classifier becomes more effective than Min Dist. That shows that if as the dimensionality increases the ratio Bhatt Mean / Bhatt Cov decreases then second order statistics are more relevant in high dimensional data even when that could not be the case in low dimensionality.

Experiment 2

This experiment is similar to the previous one. The difference is in the fact that first order statistics are predominant in this case. The parameters of the two statistical classes are:

$$M_1 = [0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0]^T$$

$$M_2 = [1.5 \quad 1 \quad .5 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0]^T$$

$$\Sigma_1 = \begin{bmatrix} 1 & & & & & & & & & 0 \\ & 1 & & & & & & & & \\ & & 1 & & & & & & & \\ & & & 1 & & & & & & \\ & & & & 1 & & & & & \\ & & & & & 1 & & & & \\ & & & & & & 1 & & & \\ & & & & & & & 1 & & \\ & & & & & & & & 1 & \\ 0 & & & & & & & & & 1 \end{bmatrix}, \quad \Sigma_2 = \begin{bmatrix} 2.5 & & & & & & & & & 0 \\ & 2 & & & & & & & & \\ & & 1 & & & & & & & \\ & & & 1 & & & & & & \\ & & & & 1 & & & & & \\ & & & & & 1 & & & & \\ & & & & & & 1 & & & \\ & & & & & & & 1 & & \\ & & & & & & & & 1 & \\ 0 & & & & & & & & & 1 \end{bmatrix}$$

The classification results are shown in Figure 14. Observe that Min Dist classifier becomes more accurate than Min Cov after six dimensions.

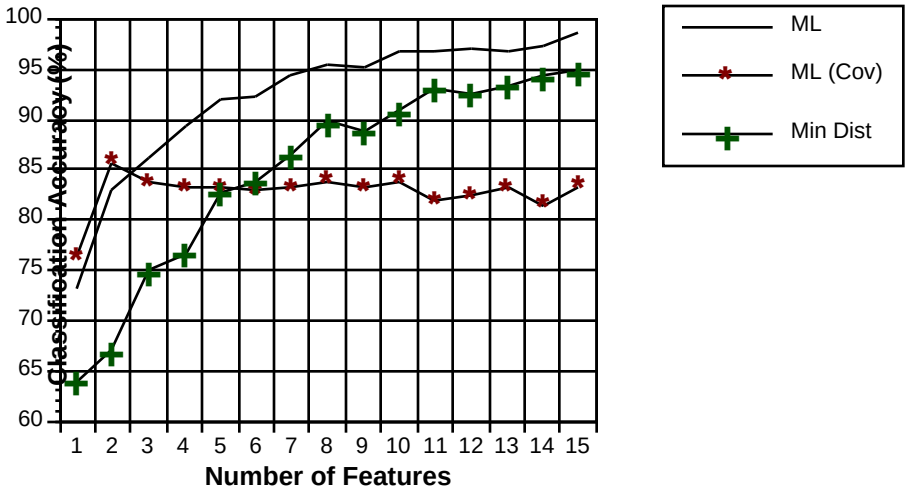


Figure 14. Performance comparison of Gaussian ML, Gaussian ML with zero mean data, and Minimum Distance classifier for two generated classes.

The mean (Bhatt Mean) and covariance (Bhatt Cov) components of Bhattacharyya distance and their sum were computed and are shown in Figure 15. Their ratio of Bhatt Cov / Bhatt Mean was calculated and shown in Figure 16. As the number of dimensions increases the ratio Bhatt Cov / Bhatt Mean decreases showing that first order statistics are more relevant in the classification of data.

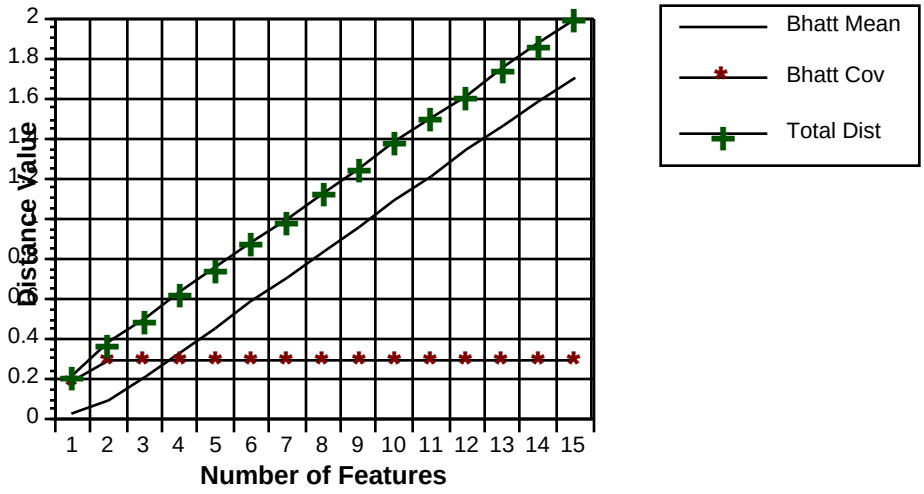


Figure 15. Bhattacharyya distance and its mean and covariance components.

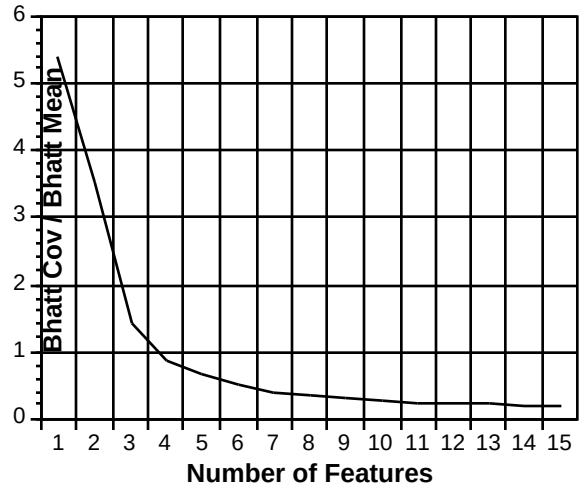


Figure 16. Ratio of Bhattacharyya distance covariance component over the mean component.

The previous results show how the predominance of the mean or covariance Bhattacharyya distance components relates directly with first or second order statistics relevance in terms of classification accuracy. In the present work both components will be computed analytically and used to calculate upper bounds that will be functions of the number of dimensions. These bounds will be calculated for the case where the mean difference plays a predominant role and for the case where the covariance difference became predominant. Then the limits of the number of dimensions increment will be taken enabling one to understand the behavior of high dimensional data under such circumstances. That is the reason for dividing all the calculations into two cases: covariance predominance and mean predominance.

Case 1: Covariance difference as the dominant role in statistical class separability.

Assume a two class problem where without loss of generality the first and second order statistics are:

$$\Sigma_1 = \begin{bmatrix} \sigma^2 & & 0 \\ & \ddots & \\ 0 & & \sigma^2 \end{bmatrix} \text{ and } \Sigma_2 = \begin{bmatrix} \alpha_1\sigma^2 & & 0 \\ & \ddots & \\ 0 & & \alpha_d\sigma^2 \end{bmatrix},$$

$$(M_2 - M_1) = \left[\varepsilon_1 \quad \cdots \quad \varepsilon_k \quad \hat{\varepsilon}_{k+1} \quad \cdots \quad \hat{\varepsilon}_d\right]^T$$

Observe that every two covariance matrices can be simultaneously diagonalized to obtain the form of previous covariance matrices [2]. That will enable us to have less complicated calculations without losing generality.

Under the conditions that:

(a) $\alpha_i \quad (\alpha_{\min}, \alpha_{\max})$, where $\alpha_{\min} > 0$, and at least there exist an α_i such that $\alpha_i \neq 1$.

(b) $\varepsilon_{\max} = \max_{i \quad (k+1,d)} \left(\left|\hat{\varepsilon}_i\right|\right)$ be such that $\varepsilon_{\max} \approx 0$.

(c) $k = f(d) \quad \lim_{d \to \infty} \frac{k}{d} = 0$, (as an example $\lambda > 0, d = k^{(1+\lambda)}$)

(d) $\varepsilon_i^2 \quad (E_{\min}, E_{\max}), \quad i \quad (1,k)$ and $E_{\max} < \infty$ (to see the validity of this last assumption, see Appendix B).

Then as d increases, the covariance contribution will dominate the Bhattacharyya distance.

Proof:

The means contribution to the Bhattacharyya distance can be written as (see Appendix A)

$$\mu_M = \frac{1}{8}\sum_{i=1}^{k}\frac{\varepsilon_i^2}{\bar{\sigma}_i^2} + \frac{1}{8}\sum_{i=k+1}^{d}\frac{\hat{\varepsilon}_i^2}{\bar{\sigma}_i^2} \le \frac{1}{8}\sum_{i=1}^{k}\frac{\varepsilon_i^2}{\bar{\sigma}_i^2} + \frac{1}{8}\sum_{i=k+1}^{d}\frac{\varepsilon_{\max}^2}{\bar{\sigma}_i^2} \le \hat{\mu}_M$$

Observe that $\alpha_{\min}$ minimizes $(1+\alpha_i), \quad i$. Then

$$\hat{\mu}_M = \frac{k}{4\sigma^2(1+\alpha_{\min})}\left[\frac{1}{k}\sum_{i=1}^{k}\varepsilon_i^2\right] + \frac{d-k}{4\sigma^2(1+\alpha_{\min})}\varepsilon_{\max}^2$$

Note that $\frac{1}{k}\sum_{i=1}^{k}\varepsilon_i^2 \le \frac{1}{k}\sum_{i=1}^{k}E_{\max} = E_{\max}$ with the consequence that

$$\hat{\mu}_M \le \mu_{M\max} = \frac{k}{4\sigma^2(1+\alpha_{\min})}E_{\max} + \frac{d-k}{4\sigma^2(1+\alpha_{\min})}\varepsilon_{\max}^2$$

The covariances contribution to the Bhattacharyya distance can be written as (see Appendix A):

$$\mu_C = \frac{1}{2}\sum_{i=1}^{d}\ln\left(\frac{\sigma^2+\alpha_i\sigma^2}{2\sigma^2\sqrt{\alpha_i}}\right) = \frac{1}{2}\sum_{i=1}^{d}\ln\left(\frac{1+\alpha_i}{2\sqrt{\alpha_i}}\right)$$

Let γ be the argument that minimizes $\frac{1+\alpha_i}{2\sqrt{\alpha_i}}, \quad i$, subject to the constrain that $\gamma \neq 1$. That argument must exist, based on the fact that $\alpha_i \quad (\alpha_{\min}, \alpha_{\max})$, where $\alpha_{\min} > 0$ and that $\quad i \quad \alpha_i \neq 1$. Then

$$\mu_C \ge \mu_{C\min} = \frac{d}{2}\ln\left(\frac{1+\gamma}{2\sqrt{\gamma}}\right).$$

Define a bound as $\Phi(\alpha_i, \varepsilon_i, d) = \frac{\mu_M}{\mu_C} \le \frac{\mu_{M\max}}{\mu_{C\min}} = \Phi_{\max}(d)$ where:

$$\Phi_{\max}(d) = \frac{\frac{1}{4\sigma^2(1+\alpha_{\min})}\left[kE_{\max} + (d-k)\varepsilon_{\max}^2\right]}{\frac{d}{2}\ln\left(\frac{1+\gamma}{2\sqrt{\gamma}}\right)}$$

The quantity $\Phi_{\max}(d)$ is an upper bound of $\Phi(\alpha_i, \varepsilon_i, d)$ and it can be rewritten as

$$\Phi_{\max}(d) = \frac{\frac{k}{d}E_{\max} + \frac{d-k}{d}\varepsilon_{\max}^2}{2\sigma^2(1+\alpha_{\min})\ln\left(\frac{1+\gamma}{2\sqrt{\gamma}}\right)}$$

Finally taking the limit of d

$$\lim_{d\to\infty}\Phi_{\max}(d) = \frac{\varepsilon_{\max}^2}{2\sigma^2(1+\alpha_{\min})\ln\left(\frac{1+\gamma}{2\sqrt{\gamma}}\right)}$$

By the assumption that $\varepsilon_{\max} \approx 0$, then $\lim_{d\to\infty}\Phi_{\max}(d) \approx 0$. As a consequence $\lim_{d\to\infty}\Phi(\alpha_i, \varepsilon_i, d) \approx 0$

In conclusion, second order statistics and the hyperellipsoids shapes will play a more important role in discriminating among the classes than the means and the hyperellipsoids positions relative to one another.

Discussion

This proof only requires that $\alpha_{\max} - \alpha_{\min} > 0$ (differences in variances). It does not depend on how much this difference should be. The quantity $\max|\varepsilon_i|$ can be as large as the physical devices permit. Also it only requires that $k = f(d)$ $\lim_{d\to\infty}(k/d) = 0$, but it does not constrain the rate. In other words, in low dimensional data the differences in covariance can be small and $k \approx d$ and in terms of the mean such difference can be very large. In that case first order statistics will be more relevant in providing information than second order statistics in such low dimensional subspaces. But if as the dimension increases, the rate at which covariance information (even a small amount of information in low dimensional subspace) grows faster (nothing is said about how much faster) than the rate at which mean information grows (even large amounts of differences) then there will be a point where the total covariances information plays a more important role in discriminating among the classes than the means information.

Case 2: Mean differences as dominant in statistical class separability.

Assume a two class problem, where without loss of generality, the first and second order statistics are:

$$\Sigma_1 = \begin{bmatrix} \sigma^2 & & 0 \\ & \ddots & \\ 0 & & \sigma^2 \end{bmatrix} \text{ and } \Sigma_2 = \begin{bmatrix} \alpha_1\sigma^2 & & 0 \\ & \ddots & \\ 0 & & \alpha_d\sigma^2 \end{bmatrix},$$

$$(M_2 - M_1) = \begin{bmatrix} \varepsilon_1 & \cdots & \varepsilon_d \end{bmatrix}^T$$

Under the assumptions that:

(a) $\alpha_i \quad (\alpha_{\min}, \alpha_{\max})$ where $0 < \alpha_{\min} < \alpha_{\max} < \infty, \quad i \quad (1,k)$.

(b) $\hat{\alpha}_i \quad (1-\delta, 1+\delta), \quad i \quad (k+1,d)$ where $\delta \approx 0$.

(c) $\varepsilon_i^2 \geq E_{\min} > 0, \quad i \quad (1,d)$.

(d) $\lim_{d \to \infty} (k/d) = 0$, (as an example $\lambda > 0, d = k^{(1+\lambda)}$).

As d increases, the means differences will dominate the Bhattacharyya distance.

Proof:

The means contribution to the Bhattacharyya distance can be written as (see Appendix A)

$$\mu_M = \frac{1}{4} \sum_{i=1}^{k} \frac{\varepsilon_i^2}{(1+\alpha_i)\sigma^2} + \sum_{i=k+1}^{d} \frac{\varepsilon_i^2}{(1+\hat{\alpha}_i)\sigma^2} = \frac{k}{4\sigma^2} \left[\frac{1}{k} \sum_{i=1}^{k} \frac{\varepsilon_i^2}{(1+\alpha_i)} \right] + \frac{d-k}{4\sigma^2} \left[\frac{1}{d-k} \sum_{i=k+1}^{d} \frac{\varepsilon_i^2}{(1+\hat{\alpha}_i)} \right]$$

Note that the maximum of $(1+\hat{\alpha}_i) = (2+\delta)$ and that the maximum of $(1+\alpha_i) = (1+\alpha_{\max})$. As a consequence

$$\mu_M \geq \hat{\mu}_M = \frac{k}{4\sigma^2(1+\alpha_{\max})} \left[\frac{1}{k} \sum_{i=1}^{k} \varepsilon_i^2 \right] + \frac{d-k}{4\sigma^2(2+\delta)} \left[\frac{1}{d-k} \sum_{i=k+1}^{d} \varepsilon_i^2 \right]$$

Observe that $E_{\min} \leq \frac{1}{m} \sum_{i=1}^{m} \varepsilon_i^2, \quad m$. This implies that:

$$\hat{\mu}_M \geq \mu_{M\min} = \frac{E_{\min}}{4\sigma^2} \left[\frac{k}{1+\alpha_{\max}} + \frac{d-k}{2+\delta} \right]$$

The covariance's contribution to the Bhattacharyya distance can be written as (see Appendix A):

$$\mu_C = \frac{1}{2} \sum_{i=1}^{k} \ln\left(\frac{1+\alpha_i}{2\sqrt{\alpha_i}} \right) + \frac{1}{2} \sum_{i=k+1}^{d} \ln\left(\frac{1+\hat{\alpha}_i}{2\sqrt{\hat{\alpha}_i}} \right)$$

Let α be the argument that maximizes $(1+\alpha_i) / \left(2\sqrt{\alpha_i}\right), \quad i \quad (1,k)$. Let $\hat{\alpha}$ be the argument that maximize $(1+\hat{\alpha}_i) / \left(2\sqrt{\hat{\alpha}_i}\right), \quad i \quad (k+1,d)$, where $\hat{\alpha} \quad (1-\delta, 1+\delta)$. Then

$$\mu_C \leq \mu_{C\max} = \frac{k}{2} \ln\left(\frac{1+\alpha}{2\sqrt{\alpha}} \right) + \frac{d-k}{2} \ln\left(\frac{1+\hat{\alpha}}{2\sqrt{\hat{\alpha}}} \right)$$

Define a bound $P(\alpha_i, \varepsilon_i, d) = \frac{\mu_C}{\mu_M} \leq \frac{\mu_{C\max}}{\mu_{M\min}} = P_{\max}(d)$

Substituting equations, the upper bound $P_{\max}(d)$ will be calculated as:

$$P_{\max}(d) = \frac{\frac{k}{2}\ln\left(\frac{1+\alpha}{2\sqrt{\alpha}}\right) + \frac{d-k}{2}\ln\left(\frac{1+\hat{\alpha}}{2\sqrt{\hat{\alpha}}}\right)}{\frac{E_{\min}}{4\sigma^2}\left[\frac{k}{1+\alpha_{\max}} + \frac{d-k}{2+\delta}\right]} = \frac{\frac{2\sigma^2}{E_{\min}}\left[\frac{k}{d-k}\ln\left(\frac{1+\alpha}{2\sqrt{\alpha}}\right) + \ln\left(\frac{1+\hat{\alpha}}{2\sqrt{\hat{\alpha}}}\right)\right]}{\left[\frac{k}{d-k}\frac{1}{1+\alpha_{\max}} + \frac{1}{2+\delta}\right]}$$

Taking the limit as d tends to infinity:

$$\lim_{d\to\infty} P_{\max}(d) = \frac{2\sigma^2(2+\delta)}{E_{\min}}\ln\left(\frac{1+\hat{\alpha}}{2\sqrt{\hat{\alpha}}}\right)$$

Observe that because $\delta \approx 0$ then $\hat{\alpha} \approx 1$ and $\lim_{d\to\infty} P_{\max}(d) \approx 0$. As a consequence $\lim_{d\to\infty} P(\alpha_i, \varepsilon_i, d) \approx 0$.

In conclusion then, for the conditions specified in this case, first order statistics and the hyperellipsoids positions relative to one another will play a more important role than second order statistics and the hyperellipsoid shape.

Discussion

This proof only requires that $\varepsilon_i^2 \geq E_{\min} > 0, \quad i \quad (1,d)$. It does not require a limitation on how large $E_{\min}$ should be. $\alpha_{\max}$ could be as large as the physical devises will allow. Also it requires that $\lim_{d\to\infty}(k/d) = 0$, but it does not constrain how the limit should approach zero. Even if in low dimensional data, where $k \approx d$, the covariance difference is very large and dominates over the means, if as the dimensionality increases, the rate at which means differences (even small differences) grows faster than the covariance one, then there will be a point where the total mean differences will provide more information for classes discrimination than covariances differences.

IV. High dimensional characteristics implications for supervised classification

Based on the characteristics of high dimensional data that the volume of hypercubes have a tendency to concentrates in the corners, and in a hyperellipsoid in an outside shell, it is apparent that high dimensional space is mostly empty, and multivariate data is usually in a lower dimensional structure. As a consequence it is possible to reduce the dimensionality without losing significant information and separability. Due to the difficulties of density estimation in nonparametric approaches, a parametric version of data analysis algorithms maybe expected to provide better performance where only limited numbers of labeled samples are available to provide the needed a priori information.

The increased number of labeled samples required for supervised classification as the dimensionality increases presents a problem to current feature extraction algorithms where computation is done at full dimensionality, e.g. Principal Components, Discriminant Analysis and Decision Boundary Feature Extraction [12]. A new method is required that, instead of doing the computation at full dimensionality, computes in a lower dimensional subspace. Performing the computation in a lower dimensional subspace that is a result of a linear projection from the original

high dimensional space will make the assumption of normality better grounded in reality, giving a better parameter estimation, and better classification accuracy.

A preprocessing method of high dimensional data based on such characteristics has been developed based on a technique called Projection Pursuit. The preprocessing method is called Parametric Projection Pursuit [6, 7].

Parametric Projection Pursuit reduces the dimensionality of the data maintaining as much information as possible by optimizing a Projection Index that is a measure of separability. The projection index that is used is the minimum Bhattacharyya distance among the classes, taking in consideration first and second order characteristics. The calculation is performed in the lower dimensional subspace where the data is to be projected. Such preprocessing is used before a feature extraction algorithm and classification process, as shown in Figure 17.

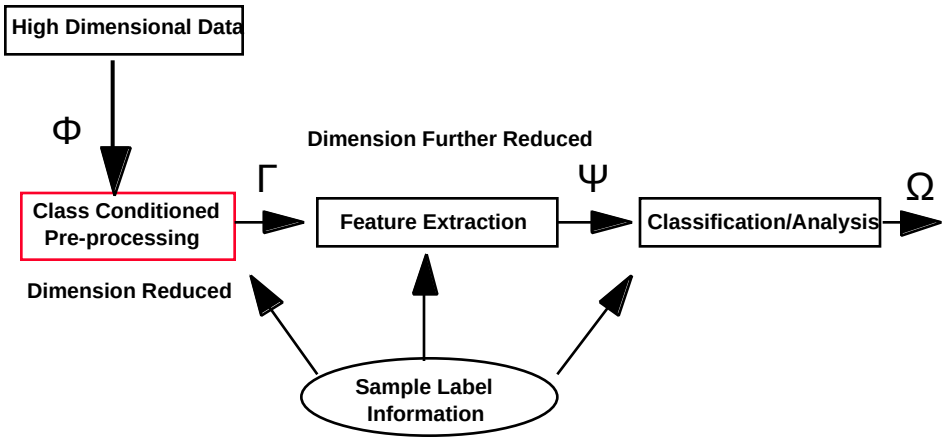


Figure 17. Classification of high dimensional data including reprocessing of high dimensional data.

In Figure 17 the different feature spaces have been named with Greek letters in order to avoid confusion. Φ is the original high dimensional space. Γ is the subspace resulting from a class-conditional linear projection from Φ using a preprocessing algorithm, e.g. Parametric Projection Pursuit. Ψ is the result of a feature extraction method. Ψ could be projected directly from Φ or, if preprocessing is used, it is projected from Γ. Finally Ω is a one dimensional space that is a result of classification of data from Ψ space. The intention is to use Parametric Projection Pursuit [6, 7, 8] in the role of Class Conditional Preprocessing, and a suitable class conditional feature extraction method such as Decision Boundary Feature Extraction [12] following this. Note that the three procedures, preprocessing, feature extraction and classification all use labeled samples as a priori information.

V. Experiment

In order to see the relevance of high dimensional geometrical and statistical properties for high dimensional data analysis purposes two experiments were designed. In both experiments a comparison is provided between high dimensional feature extraction and the method that uses a Parametric Projection Pursuit based preprocessing to reduce the dimensionality before a feature extraction method is used.

The multispectral data used in these experiments are a segment of AVIRIS data taken of NW Indiana's Indian Pine test site. From the original 220 spectral channels 200 were used, discarding the atmospheric absorption bands.

Experiment 1

In the present experiment, eight classes were defined. The total number of training samples is 1790 and the total number of test samples is 1630. The classification task for several classes in this and the next experiment are particularly difficult ones. The data were collected early in the growing season when the canopy of both corn and soybeans covered only about 5% of the area. There were three levels of tillage, no till in which there would be a great deal of residue on the soil surface from last year's crop, minimum till leaving a moderate amount of residue, and clean till for which there would be little or no residue. Add to this the normal amount of spectral variability due to the varying soil types present in the fields. Thus the 95% background would be highly variable, as compared to the relatively small difference in spectral response between corn and soybeans.

Table 1. Classes and Samples for the Eight Classes

Classes	Training Samples	Test Samples
Corn-min	229	232
Corn-notill	232	222
Soybean-notill	221	217
Soybean-min	236	262
Grass/Trees	227	216
Grass/Pasture	223	103
Woods	215	240
Hay-windrowed	207	138
Total	1790	1630

Four types of dimension reduction algorithms were used. The first is Decision Boundary Feature Extraction (DB 200-22) to reduce the dimensionality from 200 bands to 22 features. The second is Discriminant Analysis (DA 200-22) reducing the dimensionality again from 200 to 22. Both of these procedures perform a direct linear projection from Φ to Ψ. In the third and fourth methods Parametric Projection Pursuit was used to reduce the dimensionality from 200 to 22. These methods linearly project the data from Φ to Γ subspace. After that preprocessing method was used, a feature extraction algorithm follows in order to project the data once more from Γ to the Ψ subspace. Decision Boundary or Discriminant Analysis, was used (PPDB 22 and PPDA 22) with the advantages of doing the computation with the same number of training samples in less number of dimensions.

Four types of classifiers were used. The first one is ML classifier, the second is ML with 2% threshold. The third classifier is a spectral-spatial classifier named ECHO [10, 11] and the fourth is ECHO with a 2% threshold. In the second and the fourth, a threshold was applied to the standard classifiers whereby in case of true normal distributions of the data, 2% of the least likely points will be thresholded. These 2% thresholds provide one indication of how well the data fit the normal model. All of these classifiers performed a projection from Ψ to the resulted space Ω.

The results are shown in Figure 18. The methods that use Parametric Projection Pursuit as a preprocessing method, in order to use Γ as a stage between Φ and Ψ, performed better in terms of classification accuracy than directly using feature extraction at full dimensionality in Φ space (200 bands). That is because Parametric Projection Pursuit takes into consideration high dimensional characteristics. Using feature extraction methods at full dimensionality can harm the data and makes difficult the extraction of information.

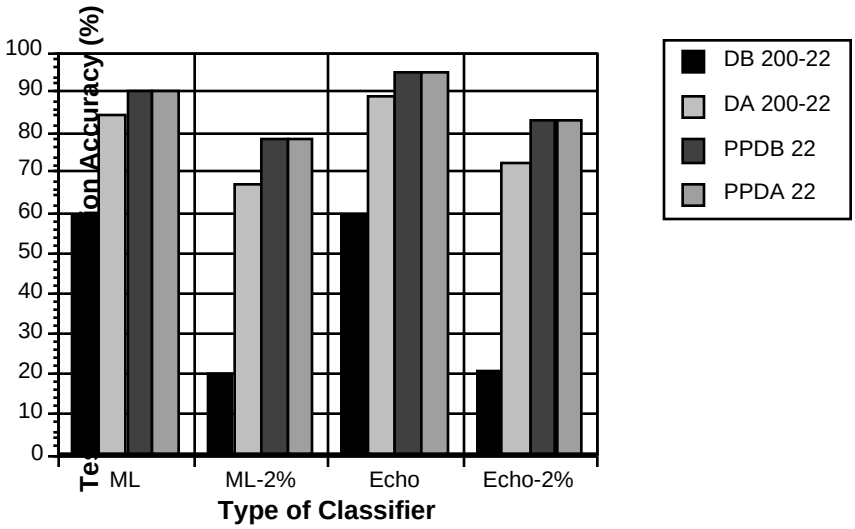


Figure 18. Test fields classification accuracy for four feature extraction methods and four classifiers.

Observe how significantly the performance of classifiers with 2% thresholds improves when using Parametric Projection Pursuit. The reason is that making the computation at low dimensional space, Γ, the assumption of normality has greater validity. In the case of having less samples and classes Discriminant Analysis will be significantly affected by the high dimensional geometrical and statistical characteristics. The next experiment will show this difficulty.

Experiment 2

In this experiment four classes were defined: corn, corn-notill, soybean-min, soybean-notill. The total number of training samples is 179 (less than the number of bands used) and the total number of test samples is 3501.

Table 2. Classes and Samples for the Four Classes

Classes	Training Samples	Test Samples
Corn-notill	52	620
Soybean-notill	44	737
Soybean-min	61	1910
Corn	22	234
Total	179	3501

Two types of dimensional reduction algorithms were used. The first is Discriminant Analysis (DA 200-3) that reduces the dimensionality from 200 to 3. It directly projects the data from Φ space to Ψ subspace. In the second method Parametric Projection Pursuit was used to reduce the dimensionality from 200 to 22. It projected the data from the Φ space to the Γ subspace. After that preprocessing method was used, Discriminant Analysis was used (PPDA 200-3) in order to linearly project the data from the Γ subspace to the Ψ subspace. As mentioned before, this has the advantage of doing the computation with the same number of training samples but at lower dimensionality. In both cases the best three features were used for classification purposes. The same four types of classifiers were used here as in the first experiment.

The results are shown in Figure 19. Parametric Projection Pursuit followed by Discriminant Analysis at lower dimensionality performed substantially better than using Discriminant Analysis at full dimensionality. The application of a threshold to Discriminant Analysis at full dimensionality

reduced its classification accuracy more severely than when a threshold was applied in the case where Projection Pursuit was first applied, followed by Discriminant Analysis at lower dimensionality. This is due to Parametric Projection Pursuit preprocessing being better fitted to the assumption of normality.

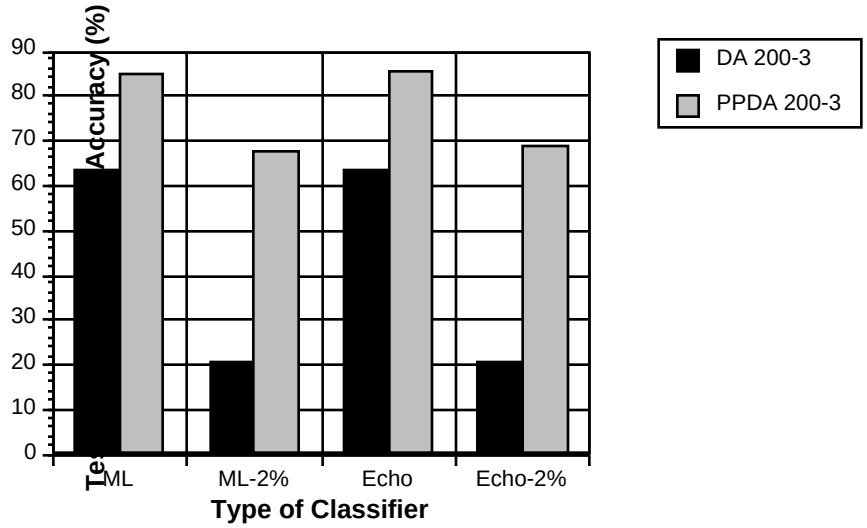


Figure 19. Test fields classification accuracy for two feature extraction methods and four classifiers.

We note in passing that Discriminant Analysis Feature Extraction provides predictably optimal results only for a number of features less than the number of classes. Thus if one has a problem in which the number of classes is not large, but the difficulty of separation requires a larger number of features than the number of classes to achieve satisfactory performance, Decision Boundary Feature Extraction would be expected to provide the better performance of the two.

VI. Conclusion

We have shown that the characteristics of high dimensional space are quite different from those of the three dimensional space we are used to. This has significant implications in the context of supervised classification techniques. In terms of class parameter estimation, a large number of samples are required to make an adequately precise estimation, and the problem grows as the dimensionality increases. In a nonparametric approach, the number of samples required to satisfactorily estimate a class density is even greater. Both kinds of estimations confront the problem of high dimensional space characteristics.

The present work is in the context of multispectral remote sensing, where commonly, training sets by which to estimate class statistics are quite small. As a consequence, it is desirable to project the data to a lower dimensional space where the effects of high-dimensional geometric characteristics and the Hughes phenomena are reduced. Commonly used techniques such as Principal Components, Discriminant Analysis, and Decision Boundary Feature Extraction have the disadvantage of requiring computations at full dimensionality in which case the required number of labeled samples is very large. The procedures use estimated statistics thus are not necessarily accurate, leading to reduced classifier performance. Another problem is the assumption of normality. Nothing guarantees that at full dimensionality, that model fits well.

It has been shown that high dimensional spaces are mostly empty, indicating that the data structure involved exists primarily in a subspace. The problem is which subspace it is to be found in is situation-specific. Thus the goal is to reduce the dimensionality of the data to the right subspace without losing separability information. In this paper we have described a procedure to make the

computations in a lower dimensional space, i.e. in Γ instead of Φ, where the projected data produce a maximally separable structure and which, in turn, avoids the problem of dimensionality in the face of the limited number of training samples. Further, a linear projection to a lower dimensional subspace will make the assumption of normality in the Γ subspace more suitable than in the original Φ. In such a lower dimensional subspace any method used for feature extraction could be used before a final classification of data, often those that have the assumption of normality.

More details of the information presented in this paper are contained in [8].

This work is part of a longer effort to find effective ways to analyze high dimensional multispectral remote sensing data. As a perhaps unusual feature of this research program we are following the practice of making available new algorithms resulting from this work in an application program for personal computers. This application program, called MultiSpec©, is made available at no cost to persons interested via the world wide web. The URL for this site is http://dynamo.ecn.purdue.edu/~biehl/MultiSpec/. The projection pursuit algorithm described in this paper is already implemented in MultiSpec, along with Decision Boundary Feature Extraction, Discriminate Analysis Feature Extraction, and a number of other algorithms related to classification of multispectral and hyperspectral image data. Thus it is possible for interested users to try out these algorithms on their own desktop computers.

Appendix A

The Bhattacharyya distance is the sum of the contribution of the difference of the means and the difference of the covariances. $\mu = \mu_M + \mu_C$, where

$$\mu_M = \frac{1}{8}(M_2 - M_1)^T \bar{\Sigma}^{-1}(M_2 - M_1), \quad \bar{\Sigma} = \frac{\Sigma_1 + \Sigma_2}{2}$$

and

$$\mu_C = \frac{1}{2}\ln\left(\frac{|\bar{\Sigma}|}{\sqrt{|\Sigma_1||\Sigma_2|}}\right)$$

For the two class problem in a d-dimensional space assume, without generality, the following.

$$(M_2 - M_1) = \begin{bmatrix}\varepsilon_1 & \cdots & \varepsilon_d\end{bmatrix}^T \text{ and } \Sigma_1 = \begin{bmatrix}\sigma_{11}^2 & & 0 \\ & \ddots & \\ 0 & & \sigma_{1d}^2\end{bmatrix}, \ \Sigma_2 = \begin{bmatrix}\sigma_{21}^2 & & 0 \\ & \ddots & \\ 0 & & \sigma_{2d}^2\end{bmatrix}$$

then

$$\bar{\Sigma} = \begin{bmatrix}\sigma_1^2 & & 0 \\ & \ddots & \\ 0 & & \sigma_d^2\end{bmatrix} = \begin{bmatrix}\frac{\left(\sigma_{11}^2 + \sigma_{21}^2\right)}{2} & & 0 \\ & \ddots & \\ 0 & & \frac{\left(\sigma_{1d}^2 + \sigma_{2d}^2\right)}{2}\end{bmatrix}$$

For that case, the computation of the mean and covariances components of Bhattacharyya distance are:

$$\mu_M = \frac{1}{8}\sum_{i=1}^{d}\frac{\varepsilon_i^2}{\overline{\sigma}_i^2}$$

$$\mu_C = \frac{1}{2}\ln\left(\prod_{i=1}^{d}\frac{\frac{\sigma_{1i}^2+\sigma_{2i}^2}{2}}{\sigma_{1i}\sigma_{2i}}\right) = \frac{1}{2}\sum_{i=1}^{d}\ln\left(\frac{\frac{\sigma_{1i}^2+\sigma_{2i}^2}{2}}{\sigma_{1i}\sigma_{2i}}\right)$$

Appendix B

The amount of energy that real sensors receive and their bandwidth is finite. As a consequence we can model ε_i^2 as a random variable that is defined over the range ε_i^2 (E_{min}, E_{max}) such that $E_{max} < \infty$, i.

Under the assumption that the $E(\varepsilon_i^2)$ exist then:

$$E_{min} \le E\left(\varepsilon_i^2\right) \le E_{max}$$

$$Var(\varepsilon_i^2) = E(\varepsilon_i^4) - E^2(\varepsilon_i^2) \le E_{max}^2 - E_{min}^2$$

Both are finite quantities.

References

[1] Diaconis, P., Freedman, D. "Asymptotics of Graphical Projection Pursuit." The Annals of Statistics Vol. 12, No 3 (1984): pp. 793-815.

[2] Fukunaga, K. "Introduction to Statistical Pattern Recognition." San Diego, California, Academic Press, Inc., 1990.

[3] Hall, P., Li, K. "On Almost Linearity Of Low Dimensional Projections From High Dimensional Data." The Annals of Statistics, Vol. 21, No. 2 (1993): pp. 867-889.

[4] Hughes, G. F., "On the mean accuracy of statistical pattern recognizers," IEEE Transactions on Information Theory, Vol. IT-14, No. 1, January 1968.

[5] Hwang, J., Lay, S., Lippman, A., "Nonparametric Multivariate Density Estimation: A Comparative Study.", IEEE Transactions on Signal Processing, Vol. 42, No. 10, 1994, pp. 2795-2810.

[6] Jimenez, L., Landgrebe, D., "Projection Pursuit For High Dimensional Feature Reduction: Parallel And Sequential Approaches," presented at the International Geoscience and Remote Sensing Symposium (IGARSS'95), Florence Italy, July 10-14, 1995.

[7] Jimenez, L., Landgrebe, D., "Projection Pursuit in High Dimensional Data Reduction: Initial Conditions, Feature Selection and the Assumption of Normality", Presented at IEEE International Conference on Systems, Man and Cybernetics (SMC 95), Vancouver Canada, October 22-25, 1995.

[8] Jimenez, L., Landgrebe, D., "High Dimensional Feature Reduction Via Projection Pursuit," Technical Report TR-ECE 96-5, School of Electrical & Computer Engineering, April 1996, and Ph.D. Thesis, Purdue University, May, 1996.

[9] Kendall, M. G., A Course in the Geometry of n-dimensions, Hafner Publishing Co., 1961.

[10] Kettig, R. L. and D. A. Landgrebe, "Computer Classification of Remotely Sensed Multispectral Image Data by Extraction and Classification of Homogeneous Objects," *IEEE Transactions on Geoscience Electronics*, Volume GE-14, No. 1, pp. 19-26, January 1976.

[11] Landgrebe, D. A., "The Development of a Spectral-Spatial Classifier for Earth Observational Data," *Pattern Recognition*, Vol. 12, No. 3, pp. 165-175,1980.

[12] Lee, Chulhee and David A. Landgrebe, "Feature Extraction Based On Decision Boundaries," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 15, No. 4, April 1993, pp. 388-400.

[13] Lee, Chulhee and David A. Landgrebe, "Analyzing High Dimensional Multispectral Data," *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 31, No. 4, pp. 792-800, July, 1993.

[14] Scharf, L. L. "Statistical Signal Processing. Detection, Estimation, and Time Series Analysis." Massachusetts: Addison-Wesley, 1991.

[15] Scott, D. W. "Multivariate Density Estimation." New York: John Wiley & Sons, 1992.

[16] Wegman, E. J., "Hyperdimensional Data Analysis Using Parallel Coordinates," Journal of the American Statistical Association, Vol. 85, No. 411, 1990, pp. 664-675.

Luis O. Jimenez

Dr. Luis O. Jimenez received the BSEE from University of Puerto Rico at Mayaguez, in 1989. He received his MSEE from University of Maryland at College Park in 1991 and his PhD from Purdue University in 1996. Currently he is an Assistant Professor of Electrical and Computer Engineering at the University of Puerto Rico, Mayaguez Campus. His research interest include Pattern Recognition, Remote Sensing, Feature Extraction, Artificial Intelligence, and Image Processing.

Dr. Jimenez is a associate member of the IEEE. He is also member of the Tau Beta Pi and Phi Kappa Phi honor societies.

David A. Landgrebe

Dr. Landgrebe holds the BSEE, MSEE, and PhD degrees from Purdue University. He is presently Professor of Electrical and Computer Engineering at Purdue University. His area of specialty in research is communication science and signal processing, especially as applied to Earth observational remote sensing. His contributions over the last 25 years in that field have related to the proper design from a signal processing point of view of multispectral imaging sensors, suitable spectral and spectral/spatial analysis algorithms, methods for designing and training classifier algorithms, and overall systems analysis. He was one of the originators of the multispectral approach to Earth observational remote sensing in the 1960's, was instrumental in the inclusion of the MSS on board Landsat 1, 2, and 3, and hosted and chaired the NASA meeting at which the bands and other key parameters were selected for the Thematic Mapper. He has been a member of a number of NASA and NRC advisory committees for this area since the 1960's.

He was President of the IEEE Geoscience and Remote Sensing Society for 1986 and 1987 and a member of its Administrative Committee from 1979 to 1990. He received that Society's Outstanding Service Award in 1988. He is a co-author of the text, *Remote Sensing: The Quantitative Approach*, and a contributor to the book, *Remote Sensing of Environment*, and the *ASP Manual of Remote Sensing (1st edition)*. He has been a member of the editorial board of the journal, *Remote Sensing of Environment*, since its inception.

Dr. Landgrebe is a Life Fellow of the Institute of Electrical and Electronic Engineers, a Fellow of the American Society of Photogrammetry and Remote Sensing, and a member of the American Society for Engineering Education, as well as Eta Kappa Nu, Tau Beta Pi, and Sigma Xi honor societies. He received the NASA Exceptional Scientific Achievement Medal in 1973 for his work in the field of machine analysis methods for remotely sensed Earth observational data. He was the 1990 recipient of the William T. Pecora Award, presented by NASA and the U.S. Department of Interior, for contributions to the field of remote sensing. He was the 1992 recipient of the IEEE Geoscience and Remote Sensing Society's Distinguished Achievement Award.