On Information Extraction Principles for Hyperspectral Data
A White Paper

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

Preface

Means for optimally analyzing hyperspectral data has been the topic of a study of ours since 1986[1]. The point of departure for this study has been that of signal theory and the signal processing principles that have grown primarily from the communication sciences area over the last half century. The basic approach has been to seek a more fundamental understanding of high dimensional signal spaces in the context of the remote sensing problem, and then to use that knowledge to extend the methods of conventional multispectral analysis to the hyperspectral domain in an optimal or near optimal fashion. The purpose of this white paper is to outline what has been learned so far in this effort.

The introduction of hyperspectral sensors which produce much more complex data than those previously should provide much enhanced abilities to extract useful information from the data stream they produce. However, it is also the case that this more complex data requires more complex and sophisticated data analysis procedures if their full potential is to be achieved. Much of what has been learned about the necessary procedures is not particularly intuitive, and indeed, in many cases is counter-intuitive. In what follows, we shall attempt not only to illuminate some of these counter-intuitive aspects, but to make them clear and therefore acceptable.

1 Work leading to the material presented here was funded in part by NASA Grants NAGW-925(1986-94), NAG5-3975 (1994-97), and ongoing Grant NAG5-3975.

A System Overview

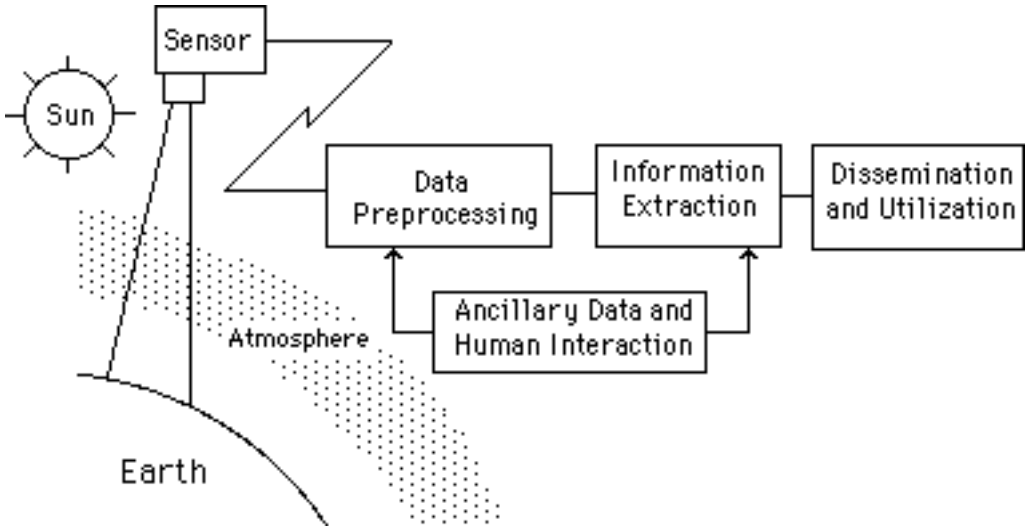


Figure 1. A system overview.

It is important to see the data analysis process not in isolation, but as a part of the whole process of the scene, sensing, and deriving the desired information. For this reason, we begin with a broad system overview of multispectral remote sensing system[2]. See Figure 1. The entire system consists of three distinctly different parts. These are:

1. The scene.
2. The sensor system.
3. The processing system.

The scene refers to that part of the system which is in front of the sensor. It includes not only the Earth's surface but also the sun for passive optical systems and the atmosphere through which the energy passes both on the way to the Earth's surface from the sun and on the return passage back to the sensor. The distinguishing characteristics of this part of the system are that a) there is no human control over it, either on the part of the system designer before construction or of the system operator after, and b) it is by far the most complex part of the entire system. Thus, in devising an optimal data analysis procedure, one must adequately account for this complexity, since it cannot be otherwise changed or controlled. Indeed, the complexity of the scene and the dynamic nature of it is so dominant, that, except for the extraction of relatively simple information, supervision of classifiers must be redone for every new data set collected.

The second part, the sensor system, functions to gather the main body (but not all) of the data about the scene. Its design parameters must be selected so that the scene and its complexity will be adequately represented by the data for purposes of extracting the needed information.

---

2 P. H. Swain, S. M. Davis (Eds.), *Remote Sensing: The Quantitative Approach*, Chapt. 7, McGraw-Hill, 1978.

All of the remainder of the system, occurring after the sensor system in the data stream, we will refer to as the processing system.

Key System Parameters

In using an optimum systems approach, an important element is to have an adequate model of the system that is to be optimized[3,4]. Such quantitative models for multispectral remote sensing systems have been devised. While it would seem inappropriate to include them in this overview discussion, it is useful to provide a bit more detail on the overall system than above, this time focused more on the information content of a data stream. For this purpose, we next list the key parameters of such an information system. They are,

1. The spatial sampling scheme. How is the scene to be sampled spatially?
2. The spectral sampling scheme. How are the pixels to be sampled spectrally?
3. The signal-to-noise ratio, S/N. What is the relation of the information-bearing aspects of the sensed data to the non-information-bearing aspects.
4. The ancillary information available. How will the classification process be supervised?
5. The informational classes desired and their interrelationships. What information is desired as output and how complex is it.

An important characteristic of this list is that all five members of this list are interrelated. For example, the first three are obviously related through the principle of conservation of energy. The amount of energy emanating from the surface per unit area and per unit wavelength is finite. If one asks for very fine spatial resolution and at the same time a very narrow spectral band, then there is very little energy per pixel to overcome the noise generated in the sensor detector. The interrelationship of the last two in this list with the first three is perhaps a little less obvious. This relationship will be made clear shortly.

As a practical matter, how well an analysis process can work depends also on the analyst, his/her expectations, initial assumptions, and point of view. We will comment only briefly on these to demonstrate there impact on the analysis process.

Initial Assumptions About Multispectral Data Analysis

Some key questions to illuminate this issue are the following.

A. How does one visualize or view multispectral data?
B. What does the data really look like in N-dimensional feature space?
C. What is the explanation for the scatter of the data in N-dimensional space?

---

3 John Kerekes and David Landgrebe "Modeling, Simulation, and Analysis of Optical Remote Sensing Systems," (PhD Thesis) Technical Report TR-EE 89-49, Purdue School of Electrical Engineering, August 1989.

4 John P. Kerekes and David A. Landgrebe, "Simulation of Optical Remote Sensing Systems," *IEEE Transactions on Geoscience and Remote Sensing,* Vol. 27, No. 6, pp. 762-771, November 1989.

D. What relationship between the data complexity and the type of analysis algorithm is most appropriate?

We will briefly examine each of these.

The matter of how the variations are represented mathematically and conceptually is an important first step in defining how the analysis process should proceed. There have been three principal ways in which multispectral data is represented quantitatively and visualized. See Figure 2 below.

- In image form, i. e., pixels displayed in geometric relationship to one another,
- As spectra, i. e., variations within pixels as a function of wavelength,
- In feature space, i. e., pixels displayed as points in an N-dimensional space.

We will refer to these three as image space, spectral space and feature space, and next summarize some of the ramifications of these three perspectives.

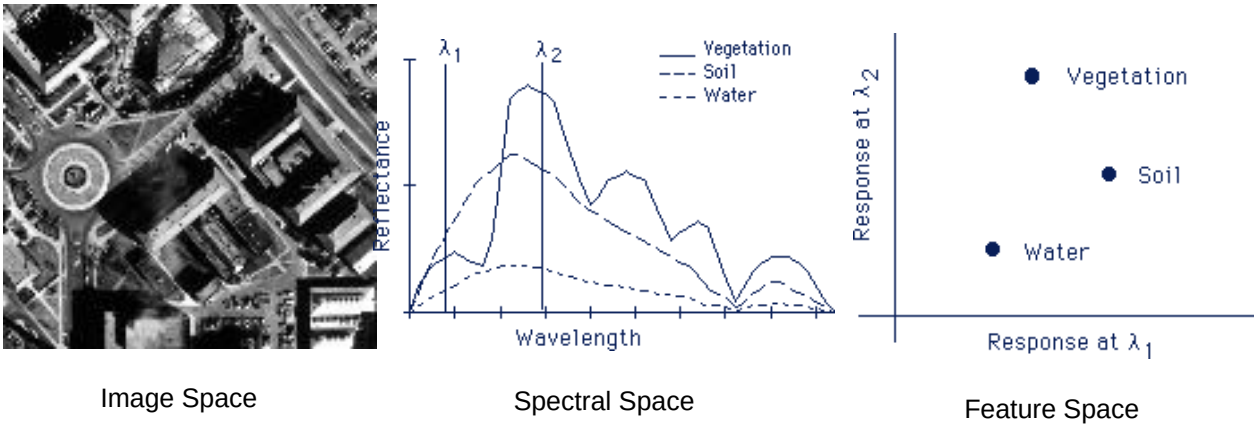


Figure 2. The forms for representing multispectral data.

Image Space. Though the image form is perhaps the first form one thinks of when first considering remote sensing as a source of information, its principal value has been somewhat ancillary to the central question of deriving thematic information from the data. Data in image form serve as the human/data interface in that image space helps the user to make the connection between individual pixel areas and the surface cover class they represent. It also provides for supporting area mensuration activities usually associated with use of remote sensing techniques. Thus, it becomes very important as to how accurately the true geometry of the scene is portrayed in the data. However, it is the latter two of the three means for representing data that have been the point of departure for most multispectral data analysis techniques.

Spectral Space. Many analysis algorithms which appear in the literature begin with a representation of a response function as a function of wavelength. Early in the work, the term "spectral matching" was often used, implying that the approach was to compare an unknown spectrum with a series of pre-labeled spectra to determine a match, and thereby to identify the unknown. This line of thinking has led, at various

times, to attempts to construct a "signature bank," a dictionary of candidate spectra whose identity had been pre-established.

A second example of the use of spectral space is the "imaging spectrometer" concept, whereby identifiable features within a spectral response function, such as absorption bands due to resonances at the molecular level, can be used to identify a material associated with a given spectrum. This approach, arising from the concepts of chemical spectroscopy, which has long been used in the laboratory for molecular identification, is perhaps one of the most fundamentally cause/effect based approaches to multispectral analysis.

Feature Space. The third basis for data representation also begins with a spectral focus, i.e., that energy or reflectance vs. wavelength contains the desired information, but it is less related to pictures or graphs. It began by noting that the function of the sensor system inherently samples the continuous function of emitted and reflected energy vs. wavelength and converts it to a set of measurements associated with a pixel which constitute a vector, i.e., a point in an N-dimensional vector space. This conversion of the information from a *continuous* function of wavelength to a *discrete point* in a vector space is not only inherent in the operation of a multispectral sensor, it is very convenient if the data are to be analyzed by a machine-implemented algorithm. It, too, is quite fundamentally based, being one of the most basic concepts of signal theory. Further, it is a convenient form if a more general form of feature extraction is to precede the analysis step, itself. As will be seen below, of the three data representations, the feature space provides the most powerful one from the standpoint of information extraction.

Next, consider how multispectral data typically appears in feature space. We will use a particularly simple situation to illustrate this. The graph below shows a scatter plot of two bands of Landsat Thematic Mapper data for an agricultural area. The area involved contains a small number of agricultural fields containing different species of agricultural crops. One sees from this graph that, even though agricultural crop responses are separable by appropriate means, this is not apparent from the scatter plot. The different crop responses do not manifest themselves as relatively distinct clusters. Rather, the data distributes itself more or less in a continuum over this space. This is typical of multispectral data, and indicates that the characteristics that allow discrimination between classes are more subtle than such straightforward examination would permit.

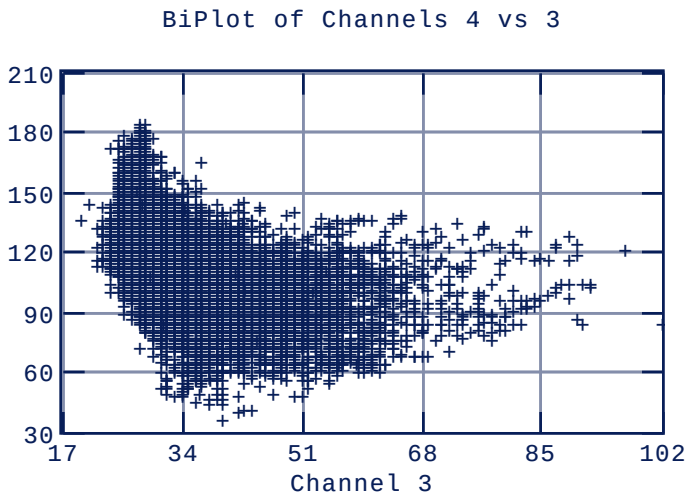


Figure 3. Scatter plot of TM Channel 4 (0.76-0.90 µm) vs. Channel 3 (0.63-0.69 µm) for an agricultural area containing a small number of crop types.

It also makes clear why entirely unsupervised classification schemes are not adequate for multispectral discrimination purposes. Without guidance from the user as to what classes are desired to be identified, an unsupervised scheme will partition the space in an unpredictable way. Further, we note that what appears to be random scatter is not "noise," meaning harmful or even useless variability. This scatter is indeed information-bearing, if appropriate means are used to model it.

Another key characteristic which is fundamental to the engineering task of optimally designing a data analysis system is the basis for the mathematical representation of the data. A number of approaches have been considered for multispectral data over the years. The following are some examples.

- Deterministic Approaches
- Stochastic Models
- Fuzzy Set Theory
- Dempster-Shafer Theory of Evidence
- Robust Methods, Theory of Capacities, Interval Valued Probabilities
- Chaos Theory and Fractal Geometry
- AI Techniques, Neural Networks

All of these approaches have been examined to varying degrees, and each has certain facets which are attractive. Deterministic approaches, for example, tend to be the most intuitive. This is important in a multidisciplinary field such as remote sensing, where different workers have different backgrounds. However, deterministic methods tend not to be as powerful, and may have other disadvantages such as being more sensitive to noise than is necessary.

Having investigated each, we have based our work on the stochastic or random process approach[5,6]. This approach has the advantage of rigor and power, and, due to its maturity, has a large stable of tools that prove of pivotal usefulness in the work.

On the Significance of Second order Statistics

Use of a stochastic process approach for modeling the spectral response of a ground scene requires determining the necessary parameters for each given data set. Using a parametric model for such modeling thus reduces the problem to that of accurately determining the mean vector and the covariance matrix in N-dimensional feature space for each class of ground cover to be identified. Because of the central importance of this point, we shall illustrate this fact with several brief illustrative arguments.

1. First, as previously indicated, one of the advantages of the stochastic process approach is the wealth of mathematical tools available using this method. For example, it is frequently the case that one would like to calculate the degree of separability of two spectral classes in order to project the accuracy it is possible to achieve in discriminating between them. There are available in the literature a number of "statistical distance" measures for this purpose. They measure the statistical distance between two distributions of points in N-dimensional space. One with particularly good characteristics for this purpose is the Bhattacharyya Distance. In parametric form it is expressed as follows.

$$B = \frac{1}{8}[\boldsymbol{\mu}_1-\boldsymbol{\mu}_2]^T\left[\frac{\Sigma_1+\Sigma_2}{2}\right]^{-1}[\boldsymbol{\mu}_1-\boldsymbol{\mu}_2] + \frac{1}{2}\,\mathrm{Ln}\,\frac{\left|\frac{1}{2}[\Sigma_1+\Sigma_2]\right|}{\sqrt{|\Sigma_1|\,|\Sigma_2|}} \qquad (1)$$

where $\boldsymbol{\mu}_i$ is the mean vector for class i and Σ_i is the corresponding class covariance matrix. This distance measure bears a nearly linear, nearly one-to-one relationship with classification accuracy. Examining this equation, one sees that the first term on the right indicates the part of the net class separability due to the difference in mean values of the two classes, while the second term indicates the portion of the total separability due to the class covariances. This makes clear from a quantitative point of view what the relationship is between first order variations (the first term on the right) and second order variations (the second term on the right) is. This illustrates, for example, that two classes can have the same mean value, in which case the first term is zero, and still be quite separable. Note that methods which are deterministically based only make use of separability measured by the first term.

2. A second way of seeing the importance of the second order variations in a more graphical fashion is via the following example spectral data[7]. Shown in Figure 4 is a plot in spectral space of data from two classes of vegetation. These data were measured in the laboratory under well controlled circumstances so that the data

5 Cooper, G. R. & C. D. McGillem, *Probabilistic Methods of Signal and System Analysis,* Second Edition, Holt, Rinehart & Winston, 1986, Chapter 7.

6 Papoulis, A., *Probability, Random Variables, and Stochastic Processes,* Second Edition, McGraw-Hill 1984.

7 P. H. Swain, S. M. Davis (Eds.), *Remote Sensing: The Quantitative Approach*, p. 14 ff, McGraw-Hill, 1978.

spread at each wavelength is not noise, but is due to the natural variability of reflectance present from the leaves of such vegetation.

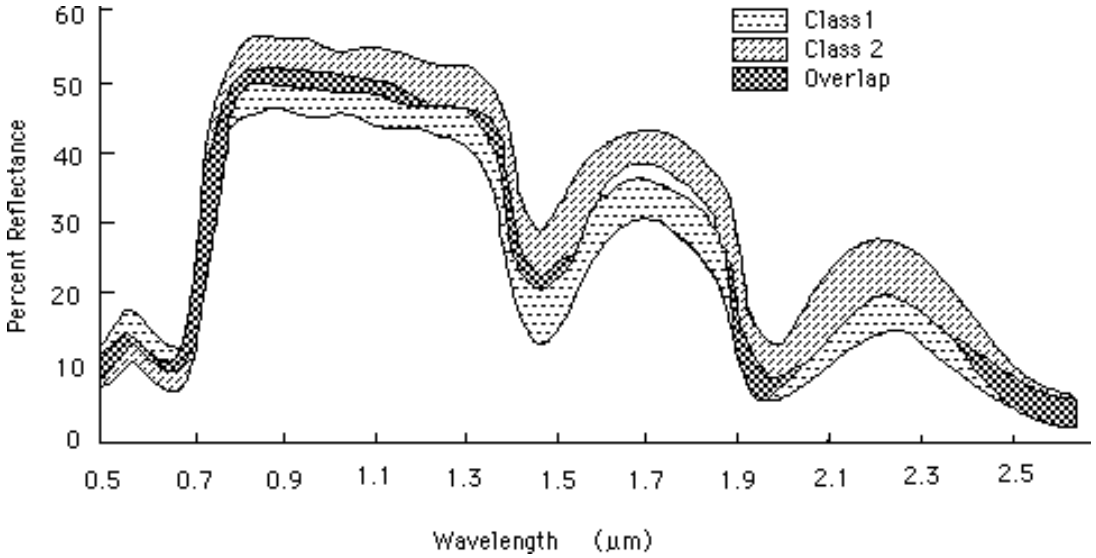


Figure 4. Spectral responses for a typical pair of classes, showing the interval at each wavelength into which they fall.

It would appear from this spectral space view that the two classes are separable only in the region around 1.7 µm. However, even in the region around 0.7 µm where there is maximum overlap of the two classes, they are separable classes if a method based upon both the first and the second order effects is utilized.

To illuminate this further, in Figure 5 is shown the actual data values plotted in spectral space for bands at 0.67 and 0.69 µm. It is clear from this presentation of the data that the reflectances do heavily overlap in these two bands.

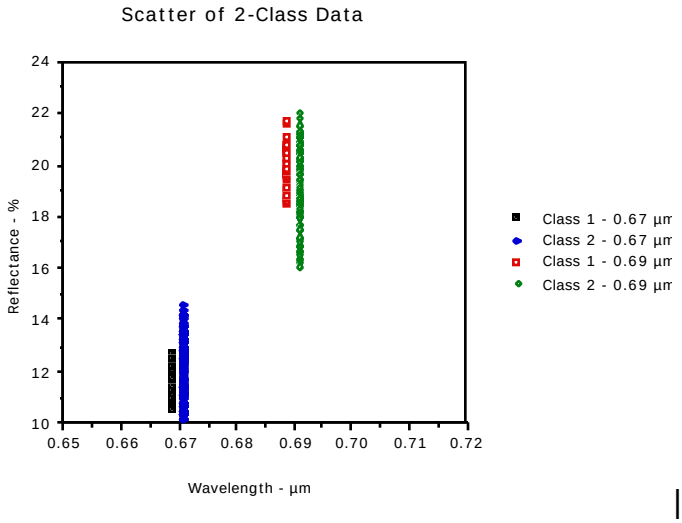


Figure 5. Data points for two vegetation classes in two spectral bands.

However, by plotting the same data in feature space as shown in Figure 6. it can be seen that the data are highly separable even via a simple linear classifier. Analyzing this situation as to what allows this separability, one sees first that the data for both

classes, being distributed in a 45° direction, are highly correlated in these two bands. This correlation plus only a small difference in the class mean values makes the two classes separable. Note that the correlation in this case is seen as providing information about the shape of the class distributions rather than seeing it merely as indicating redundancy. It is this class shape information, as indicated by the correlation, a second order statistic, taken together with the class mean values, a first order statistic, that determines the degree of separability.

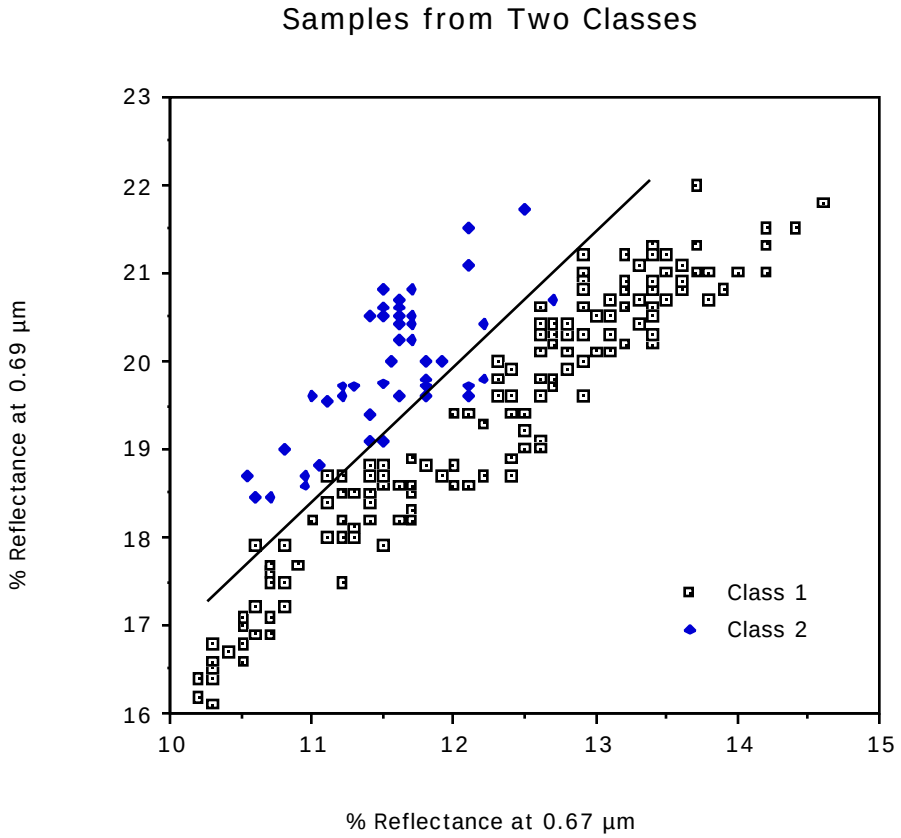


Figure 6. Data points for two vegetation classes in two dimensional feature space.

At this point we introduce another parameter in the analysis process, that of the degree of complexity of the analysis algorithm to be used. Extending the idea just introduced, it is common for spectral data of scenes to have a substantial degree of correlation between bands, and, as just seen, rather than seeing this as indicating useless redundancy, it is to be viewed as providing shape information about the class distribution. The hypothetical example of Figure 7 illustrates how this relates to classifier complexity[8]. Assume the two oval shaped areas of the figure indicate areas of concentration for two classes. If one were to use a minimum distance to means classifier, which uses only class mean information, the linear decision boundary marked would be the location and orientation assumed by the classifier. A slightly more complex Fisher Linear Discriminant classifier (see below) provides a slightly

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

shifted decision boundary as shown. The shaded areas would represent the pixels that would be classified incorrectly in this case.

If, on the other hand, one utilized a standard maximum likelihood Gaussian classifier, which utilizes both first and second order statistics, the curved decision boundary marked would be the result with much improved error performance. We shall provide greater detail with regard to the matter of algorithm complexity shortly.

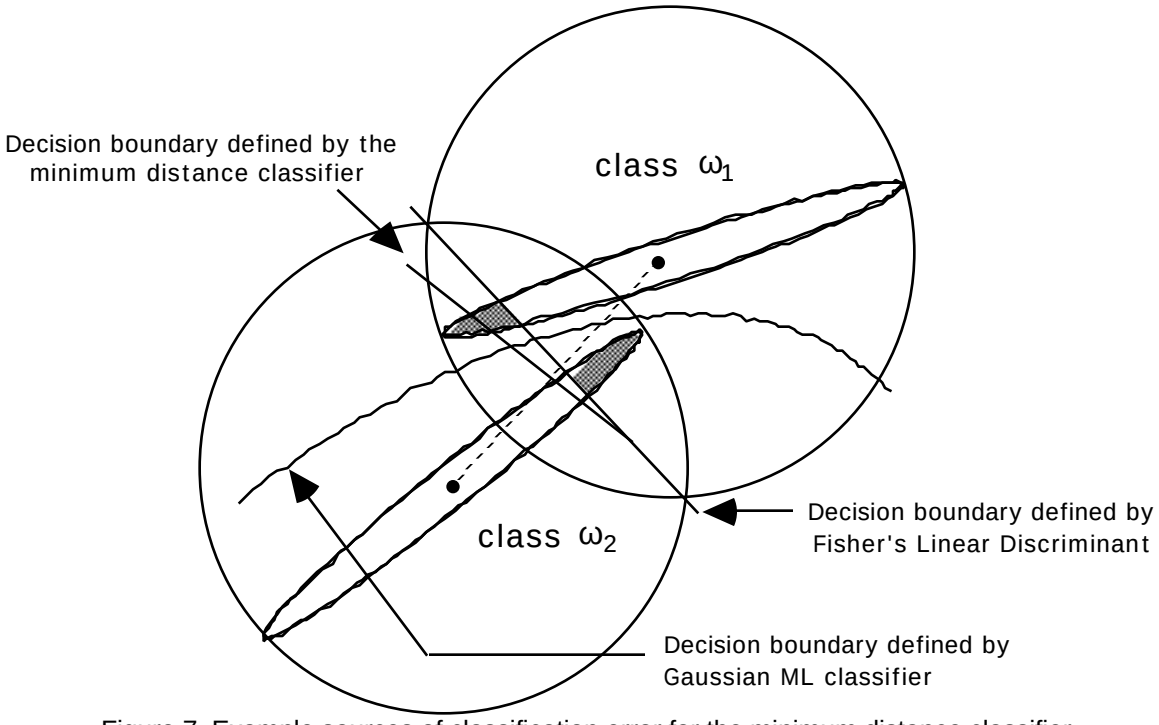


Figure 7. Example sources of classification error for the minimum distance classifier and maximum likelihood Gaussian classifiers.

3. The above example illustrates a concept in two dimensions, but what about higher dimensional data? This becomes more difficult to display intuitively, because one cannot draw in high dimensions. However, consider the following[9].

9 David Landgrebe, "Multispectral Data Analysis: A Signal Theory Perspective," 43 pages, ©1994 by David Landgrebe, Downloadable from http://dynamo.ecn.purdue.edu/~biehl/MultiSpec/documentation.html

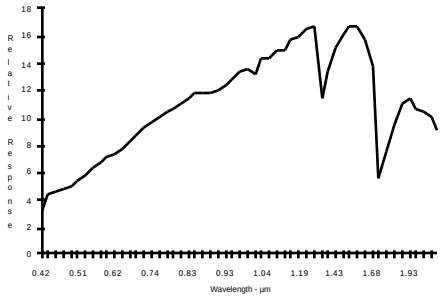


Figure 8. A typical spectral reflectance curve for soil.

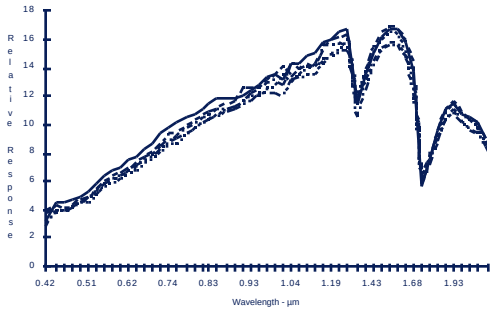


Figure 9. Spectral response showing second order variations.

Shown in Figure 8 is a single spectral reflectance curve for a certain soil type in spectral space. Except for the two absorption bands, it appears rather featureless, and indeed, no second order effects can be seen from a single deterministic curve. Shown in Figure 9 is the result in spectral space of making five measurements on samples of this soil type. Some variation is now apparent, although its structure cannot be discerned in this spectral space presentation. However, Figure 10 shows the result of plotting the data of Figure 9 after have subtracted out the mean of the five samples, thus separating the first order statistic, the mean value, out so that structure in the second order variation can be more clearly observed. It is seen that the samples have a high degree of correlation in the region up to about 0.9 µm, in that a sample that is above the mean at 0.5 µm tends to remain above the mean up to 0.9 µm, while one that is below tends to remain below. Above 0.9 µm this structure changes significantly. If this structure should be diagnostic of this soil type in that no other material in the same scene would have this same characteristic, then it would indicate a capability to discriminate this soil type in that scene.

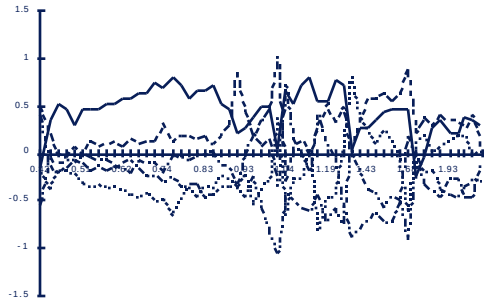


Figure 10. Second order variations about the mean for the samples of Figure 9

Though it is not possible to show a graph in the needed high dimensional feature space, it is possible to determine quantitatively its separability from other classes, for example, by using Bhattacharyya Distance, equation (1) above. Further, an additional visualization tool for high dimensional data has been devised and is referred to as "statistics image" [10]

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

4. An example classification from a recent paper will further illuminate the matter[11]. For this experiment, a multispectral data set with a large number of spectral bands was analyzed using standard pattern recognition techniques. The data were classified using first a single spectral feature, then two, and continuing on with greater and greater numbers of features. Three different classification schemes were used, (a) a standard maximum likelihood Gaussian scheme, in which both the means and the covariance matrices, i.e., both first and second order variations, were used, (b) the same except with the mean values of all classes adjusted to be the same, so that the classes differed only in their covariances, and (c) using a minimum distance to means scheme such that mean differences are used, but covariances are ignored. It is seen from the results shown in Figure 11 below that case (a) produced clearly the best result, as would be expected.

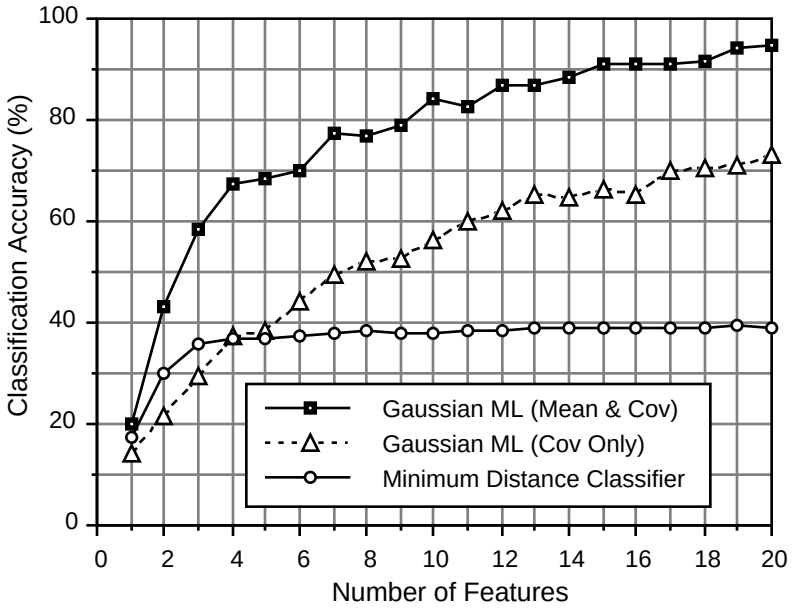


Figure. 11. Performance comparison of the Gaussian ML classifier, the Gaussian ML classifier with zero mean data, and the minimum distance classifier.

In comparing the latter two, though, it is seen that, at first in low dimensional space, the classifier using mean differences performed best. However, as the number of features was increased, this performance soon saturated, and improved no further. On the other hand, while the classifier of case (b) which used only second order effects, was at first the poorest, it soon outperformed the one of case (c) and its performance continued to improve as greater and greater numbers of features were used. Thus it is seen that second order effects, in this case represented by the class covariances, are not particularly significant at low dimensionality, but they become so as the number of features grows, to the point that they become much more significant than the mean differences between classes at any dimensionality. It is, of course, also possible to

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

show other example classifications where the mean vector dominates over the covariance[12].

These four examples show the added value of second order variations over first order ones alone, and the usefulness of an N-dimensional feature space view point. Even though one cannot draw in more than three dimensions, and thus "visualize" what is taking place, mathematical tools such as Bhattacharyya Distance are available to quantify what is the case in such spaces.

However, the potential advantage of second order effects can be easily lost if increased precision in determining the class distributions is not achieved. This is what is dealt with in the following section.

Ancillary Information and Classifier Supervision.

From the vantage point of the above, it is clear that analysis methods which utilize both first and second order statistics can provide superior performance compared to those which utilize only first order effects. However, in many cases, this is not what is observed in practice. The explanation for this becomes apparent from the following additional aspects of signal theory.

With regard to the ability to discriminate between a pair of classes, an illuminating theoretical result appeared in the literature some years ago[13]. In this paper, the result shown in Figure 12 was derived. The ordinate for the curves in this figure is the mean recognition accuracy for the two class case, averaged over the ensemble of classifiers. The abscissa is measurement complexity, which in the case of multispectral data, is directly related to the number of bands and the number of gray values per bands. The parameter for the different curves of the graph is the prior probability of one of the two classes. Looking specifically at the case for the prior probability of one half, one sees that the curve increases with measurement complexity, rapidly at first, but then more slowly. However the curve does not have a maximum, implying that it continues to increase.

---

12 Jimenez, Luis, and David Landgrebe, “Supervised Classification in High Dimensional Space: Geometrical, Statistical, and Asymptotical Properties of Multivariate Data,” *IEEE Transactions on System, Man, and Cybernetics*, To appear January, 1998. Downloadable from http://dynamo.ecn.purdue.edu/~biehl/MultiSpec/documentation.html

13 G. F. Hughes, "On The Mean Accuracy Of Statistical Pattern Recognizers," IEEE Trans. Infor. Theory, Vol. IT-14, No. 1, pp. 55-63, 1968

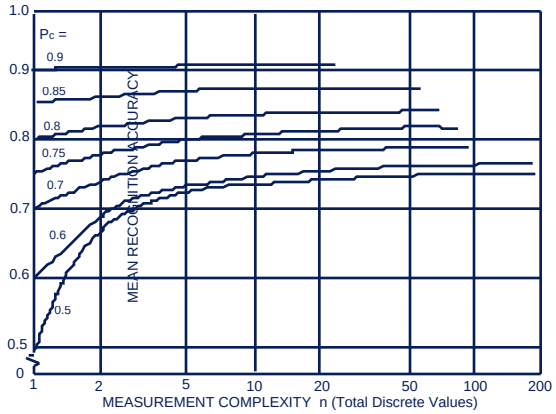


Figure 12. Mean Recognition Accuracy vs. Measurement Complexity for the infinite training case.

The graph of Figure 12 is for the case where there were an infinite number of training samples available, implying completely precise knowledge of the class distributions. Dr. Hughes also derived the result for finite training data. The result is shown in Figure 13 for the case of equally likely classes. Here the parameter for the various curves is m, the number of training samples. It is seen in this case that each curve (except for the $m \rightarrow \infty$ case) does have a maximum, indicating that there is a best measurement complexity. It depends upon how many training samples one has, and thus how precise is the estimate of the class distributions.

It is important to note that the maximum of the curves moves upward and to the right as m increases, indicating that one can expect, on the average, to see improved performance as one increases the measurement complexity, but to achieve it, one will need increased precision in estimating the class distributions.

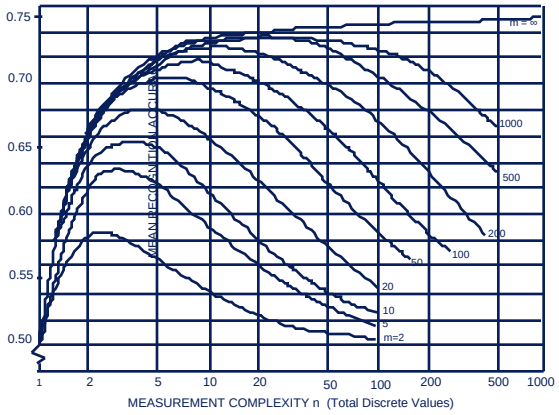


Figure 13. Mean Recognition Accuracy vs. Measurement Complexity for the finite training case.

This result, that shows that more spectral bands is not always better, and indeed, that it in fact becomes worse, was rather controversial when it was first introduced. The origin of this phenomenon can be made more understandable by the following simple drawings illustrating the basic concepts involved.

If one were to sketch the expected relationship between class separability and dimensionality, it should look something like Figure 14 (A), i.e. similar to Figure 12. Further, if one were to sketch the conceptual relationship between the accuracy of statistics estimation and dimensionality, it should be as in Figure 14 (B). That is, for a fixed number of training samples, as one increases the dimensionality, one would expect the accuracy of estimation to decline. For example, 100 samples may be enough to obtain a reasonably accurate estimate of the elements of a 5 dimensional mean vector and covariance matrix, but it would not be enough for 500 dimensional one. Further, if one increases the number of training samples, $N_1 \rightarrow N_2$, one would expect the curve to shift to the right.

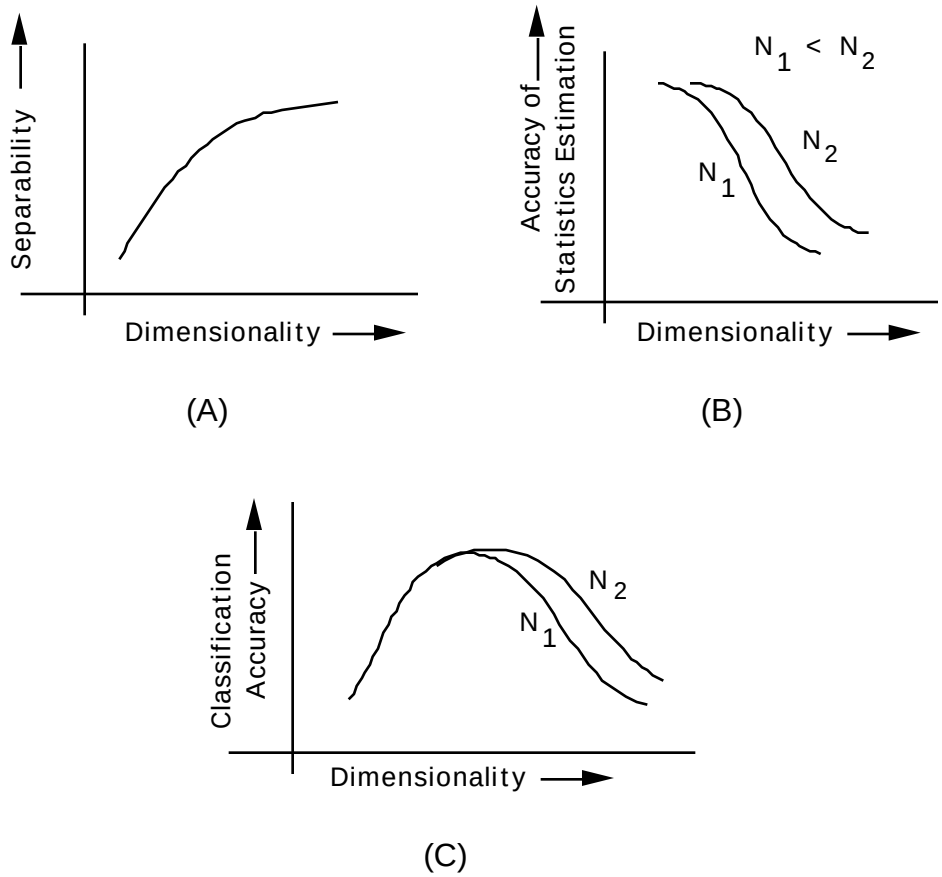


Figure 14. Effects which result in the Hughes Phenomenon.

Taken together, these two factors would produce an overall relationship as shown in Figure 14 (C), a relationship not unlike that of Figure 13.

On Classifier Complexity

The Hughes result above again reflects not only on the need for precise class distribution determination, but indirectly to a relationship between dimensionality and the complexity of the classifier algorithm to be used. There is now a large array of

different types of classifier algorithms that appear in the literature. Often it is difficult to discern from their description the extent to which they make use of the various information-bearing attributes of the multispectral or hyperspectral data. Thus it seems useful to have a generic list of classifier algorithms given in an ascending hierarchical order according to the portion of the spectral attributes that they utilize. Then, as new or specialized algorithms are encountered, they can be compared with the hierarchy to better understand what portion of the possible signal attributes they utilize. In doing so, we shall assume that the data are in feature space vector form, and will not take into account attributes other than spectral ones.

A standard way of describing a classification rule is to use the *discriminant function*. For the m-class case, where the pixel to be classified is specified as $\mathbf{X}$, a vector in feature space, assume we have m functions of $\mathbf{X}$, $\{g_1(\mathbf{X}), g_2(\mathbf{X}), \ldots g_m(\mathbf{X})\}$ such that $g_i(\mathbf{X})$ is larger than all others whenever $\mathbf{X}$ is from class i. These functions $g_i(\mathbf{X})$ are referred to as discriminant functions. Then the classification rule becomes

Let ω_i denote the i^{th} class. Then decide $\mathbf{X}$ is in class ω_i if and only if

$$g_i(\mathbf{X}) \geq g_j(\mathbf{X}) \quad \text{for all } j = 1,2, \ldots m.$$

For those classifiers in the list which specifically involve the parameter estimators, we shall specify them in terms of the appropriate discriminate function.

1. Ad hoc and deterministic algorithms.

The nature of variations in spectral response which are usable for discrimination purposes is quite varied. They may extend all the way from the general shape of the response function spread across many bands to very localized variations in one or a small number of narrow spectral intervals. Many algorithms have appeared in the literature which are designed to take advantage of specific characteristics on an ad hoc basis. Example algorithms of this type extend from simple parallelepiped algorithms or spectral matching schemes based upon least squares difference between an unknown pixel response that has been adjusted to reflectance and a known spectral response from a field spectral data base, on to an imaging spectroscopy scheme based upon one or more known molecular absorption features.

Such algorithms are sometimes motivated by a desire to take advantage of perceivable cause/effect relationships. These algorithms are usually of a nature that the class is defined by a single spectral curve, i.e., a single point in feature space. When this is the case, by that fact, they cannot utilize second order class information.

In the following parametric methods, $\mathbf{X}$ is the observed (vector-valued) pixel, $\boldsymbol{\mu}_i$ is the mean vector for class i and Σ_i is the corresponding class covariance matrix. It is assumed there are m classes, and, for simplicity for present purposes, the classes are assumed equally likely. In this case, given the additional assumptions specific to each case, these schemes are Bayes optimal in the sense that they will provide minimum error for the class statistics given. They are suboptimal only to the extent that the various assumptions are not, in fact, met, and that the finite training sets do not completely precisely determine the class statistics.

2. Minimum Distance to Means

$$g_i(\mathbf{X}) = (\mathbf{X} - \boldsymbol{\mu}_i)^T(\mathbf{X} - \boldsymbol{\mu}_i) \tag{2}$$

choose class i if $g_i(\mathbf{X}) \le g_j(\mathbf{X})$ for all j = 1,2, . . . m.

In this case, pixels are assigned to whichever class has the smallest Euclidean distance to its mean. The classes are, by default, assumed to have common covariances which are equal to the identity matrix. This is equivalent to assuming the classes all have unit variance in all features and the features are all uncorrlated to one another. The decision boundary in feature space will be linear and located equidistant between the class means and orthogonal to a line joining their means. See Figure 7.

3. Fisher's Linear Discriminant

$$g_i(\mathbf{X}) = (\mathbf{X} - \boldsymbol{\mu}_i)^T\Sigma^{-1}(\mathbf{X} - \boldsymbol{\mu}_i) \tag{3}$$

choose class i if $g_i(\mathbf{X}) \le g_j(\mathbf{X})$ for all j = 1,2, . . . m.

In this case, the classes are assumed to have a common covariance specified by Σ. This is equivalent to assuming the classes do not have the same variance in all features, the features are not necessarily uncorrlated, but both classes have the same variance and correlation structure. In this case the decision boundary in feature space will be linear, but its location between the class mean values will depend upon Σ.

4. Quadratic (Gaussian) Classifier

$$g_i(\mathbf{X}) = - (1/2)\ln|\Sigma_i| - (1/2)(\mathbf{X}-\boldsymbol{\mu}_i)^T\Sigma_i^{-1}(\mathbf{X}-\boldsymbol{\mu}_i) \tag{4}$$

choose class i if $g_i(\mathbf{X}) \ge g_j(\mathbf{X})$ for all j = 1,2, . . . m.

In this case, the classes are not assumed to have the same covariance, each being specified by Σ_i. The decision boundary in feature space will be a second order hypercurve (or several segments of second order hypercurves if more than one subclass per class is assumed), and its form and location between the class mean values will depend upon the Σ_i's.

5. Nonparametric Methods

$$g_i(\mathbf{X}) = \frac{1}{N_i}\sum_{j=1}^{N_i} K\left(\frac{\mathbf{X}-\mathbf{X}_{ji}}{\lambda}\right) \tag{5}$$

choose class i if $g_i(\mathbf{X}) \ge g_j(\mathbf{X})$ for all j = 1,2, . . . m.

Nonparametric classifiers take on many forms, and their key attractive feature is their generality. As represented above, $K(\cdot)$ is a kernel function which can take on many

forms. The entire discriminate function has N_i terms, each of which may contain one or more arbitrarily selected parameters. Thus, the characteristic which gives a nonparametric scheme its generality is this often large number of features. However, every detailed aspect of the class density must be determined by this process, and this can quickly get out of hand. For example, while Fukunaga[14] proves that in a given circumstance, 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, in a nonparametric case, 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[15,16]. It is for this reason that nonparametric schemes, including the currently popular neural network methods, are less attractive for the remote sensing circumstance, than they might at first appear. In addition, for neural network methods, which use iterative training, the large amount of computation required in the training process detract from their practical value, since training must be redone for every data set. As a result, we will focus on parametric methods hereafter.

We again note that methods which use multiple samples for training have a substantial advantage over deterministic methods that utilize only a single spectrum to define a class. The latter tend to require very high signal-to-noise ratios, where as those based upon multiple sample training sets tend to be more immune to the effects of noise.

The means for quantitatively describing a class distribution from a finite number of training samples commonly comes down to estimating the elements of the class mean vector and covariance matrix, as has been seen. Sound practice dictates that the number of training samples must be large compared to the number of elements in these matrices. When the number of training samples is limited, as it nearly always is in remote sensing, and the dimensionality of the data becomes large, the needed relationship between the training set size and the number of matrix elements that must be estimated quickly becomes strained even in the parametric case. This is especially true with regard to the covariance matrix, whose element population grows very rapidly with dimensionality. For example, the following table illustrates the number of elements in the various covariance matrix forms which must be estimated for the case of 5 classes and several different numbers of features, p.

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

15 Scott, D. W. "Multivariate Density Estimation." John Wiley & Sons, pp. 208-212, 1992.

16 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.

No. of Features p	Class Covar. $5\{p^2 - [(p-1)^2 + (p-1)]/2\}$	Diagonal Class Common Covar. $5p$	Common Covar. $\{p^2 - [(p-1)^2 + (p-1)]/2\}$	Diagonal Common Covar. p
5	75	25	15	5
10	275	50	55	10
20	1050	100	210	20
50	6375	250	1275	50
200	100,500	1000	20,100	200

Table 1. Number of elements in various covariance matrix forms to be estimated. A case for 5 classes is assumed.

The relationship between training set size and dimensionality has been examined quantitatively[17], and it has been found that, as the dimensionality goes up (or the number of samples available goes down), it may be advantageous to reduce the number of elements that must be estimated by reducing the algorithm complexity, i.e., by deciding between using individual class covariance matrices, a common covariance matrix, and a diagonal common covariance matrix. This allows for a more precise estimation of the parameters needed. The tradeoff of gaining precision by reducing complexity when the training sets are limited, can result in improved accuracy of classification. It has been codified into a scheme referred to as LOOC (Leave One Out Covariance) estimation which is relatively transparent to the user. The scheme is as follows. The quantity to be estimated is $C_i(\alpha_i)$, where,

$$C_i(\alpha_i) = \begin{cases} (1-\alpha_i)\,\mathrm{diag}(\Sigma_i) + \alpha_i \Sigma_i & 0 \le \alpha_i \le 1 \\ (2-\alpha_i)\Sigma_i + (\alpha_i - 1)S & 1 < \alpha_i \le 2 \\ (3-\alpha_i)S + (\alpha_i - 2)\,\mathrm{diag}(S) & 2 < \alpha_i \le 3 \end{cases} \quad (6)$$

Σ_i is the sample covariance matrix, estimated for class i from the training samples. The common covariance is defined by the average sample covariance matrix $S = \frac{1}{L}\sum_{i=1}^{L} \Sigma_i$ where a total of L classes are assumed. The variable α_i is a mixing parameter that determines which estimate or mixture of estimates is selected. If $\alpha_i = 0$, the diagonal sample covariance is used. If $\alpha_i = 1$, the estimator returns the sample covariance estimate. If $\alpha_i = 2$, the common covariance is selected, and if $\alpha_i = 3$ the diagonal common covariance results. Other values of α_i lead to mixtures of two estimates. Projected accuracy is estimated a priori by the well-known leave-one-out method using the available training samples.

Another way of looking at the analysis process is that it compares an unknown sample to known data or information. Deterministic methods, for example, have this

17 Hoffbeck, Joseph P. and David A. Landgrebe, "Covariance Matrix Estimation and Classification with Limited Training Data," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 18, no. 7, pp. 763-767, July 1996.

comparison taking place at the pixel to pixel or spectrum to spectrum level. The next step up in the effective utilization of reference data is a scheme which compares a pixel to a class distribution. This is the function of procedures 2 through 5 above, and is the most common form of pixel classifiers. One can take the process one step further by using a scheme which results in distribution to distribution comparison. This scheme, sometimes called a sample classifier, requires that one have as the unknown a set of pixels which are all assumed to be members of the same class. Classifiers which utilize spatial as well as spectral information can be arranged to operate in this way. The ECHO classifier[18,19] is an example of this case. It proceeds by first segmenting the scene on a multivariant basis into statistically homogeneous objects using spatial information, then classifying the objects using a distribution to distribution comparison. Using the same class descriptions as a pixel classifier, it nearly always achieves higher accuracy and usually does so with less computation time.

In addition to these methods, additional aspects of classifier design have been investigated, including more complex decision logic[20,21] and ways to speed the classification computation[22,23]. With the rapid increase of computational processor speeds in recent years, processing speed has turned out not to be the pressing problem it once was, and until the more pressing problems of the analysis process are solved, complex decision logic potentials can also reasonably be postponed. Thus these aspects are being pursued at a lower priority.

One additional aspect of classifier design which appears to have significant utility has also been investigated. It has been shown[24,25] that by adding unlabeled samples to the classifier design process, better estimates for the discriminant functions can be obtained. This has resulted in an algorithm referred to as "statistics enhancement." The algorithm iterates between the labeled (training) samples and unlabeled (all other) samples from the data set to modify the class statistics so that a better fit to the overall data distribution is obtained. In this way, the ability of the classifier to generalize beyond its training samples is improved. In mathematical terms, what is desired is to

---

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

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

20 B. Kim and D. Landgrebe, "Hierarchical Classifier Design in High Dimensional Numerous Class Cases," *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 29, No. 4, July 1991, pp. 518-528.

21 S. Rasoul Safavian and David Landgrebe, "A Survey of Decision Tree Classifier Methodology," *IEEE Transactions on Systems, Man, and Cybernetics*, Vol. 21, No. 3, May/June 1991, pp. 660-674.

22 Chulhee Lee and David A. Landgrebe, "Fast Likelihood Classification," *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 29, No. 4, July 1991, pp. 509-517.

23 Byeungwoo Jeon and David A. Landgrebe, "Fast Parzen Density Estimation Using Clustering-Based Branch and Bound," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 16, No. 9, pp. 950-954, September 1994.

24 Behzad M. Shahshahani and David A. Landgrebe, "The Effect of Unlabeled Samples in Reducing the Small Sample Size Problem and Mitigating the Hughes Phenomenon," *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 32, No. 5, pp. 1087-1095, September 1994.

25 Behzad M. Shahshahani, "Classification of Multi-Spectral Data By Joint Supervised-Unsupervised Learning," PhD Thesis and School of Electrical Engineering Technical Report TR-EE-94-1, January, 1994.

have the density function of the entire data set modeled as a mixture of class densities, i.e.,

$$p(\mathbf{x}|\theta) = \sum_{i=1}^{m} \alpha_i p_i(\mathbf{x}|\varphi_i) \tag{7}$$

where **x** is the measured feature (vector) value, p is the probability density function describing the entire data set to be analyzed, θ symbolically represents the parameters of this probability density function, p_i is the density function of class i desired by the user with its parameters being represented by φ_i, α_i is the weighting coefficient or probability of class i, and m is the number of classes. Basically, the training classes define the p_i's, while the data to be classified defines $p(\mathbf{x}|\theta)$. What is needed then is to bring the two sides of the equation to equality. An iterative scheme adjusting the φ_i's and determining α_i's is used to accomplish this. The process thus improves the generalization capabilities of the classifier, i.e., improves the accuracy performance on samples in the scene other than the training samples.

Geometrical, Statistical and Asymptotical Properties of High Dimensional Spaces

The previous sections of this paper are primarily in the context of conventional multispectral data. In this section[26], we will describe some of the unique or unusual aspects of hyperspectral data, in order to illuminate some of the circumstances which must be accounted for in dealing with hyperspectral data in an optimal fashion.

For a high dimensional space, as dimensionality increases:

A. The volume of a hypercube concentrates in the corners[27]

It has been shown[28] that the volume of a hypersphere of radius r and dimension d is given by the equation:

$$V_s(r) = \text{volume of a hypersphere} = \frac{2r^d}{d}\frac{\pi^{\frac{d}{2}}}{\Gamma\left(\frac{d}{2}\right)} \tag{8}$$

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

$$V_c(r) = \text{volume of a hypercube} = (2r)^d \tag{9}$$

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

---

26 Material in this section is taken from Luis O. Jimenez, "High Dimensional Feature Reduction Via Projection Pursuit," PhD Thesis and School of Electrical & Computer Engineering Technical Report TR-ECE 96-5, April 1996. See also reference [10].

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

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

$$f_{d1} = \frac{V_s(r)}{V_C(r)} = \frac{\pi^{d/2}}{d2^{d-1}\Gamma\left(\frac{d}{2}\right)} \tag{10}$$

where d is the number of dimensions. We see in Figure 15 how f_{d1} decreases as the dimensionality increases.

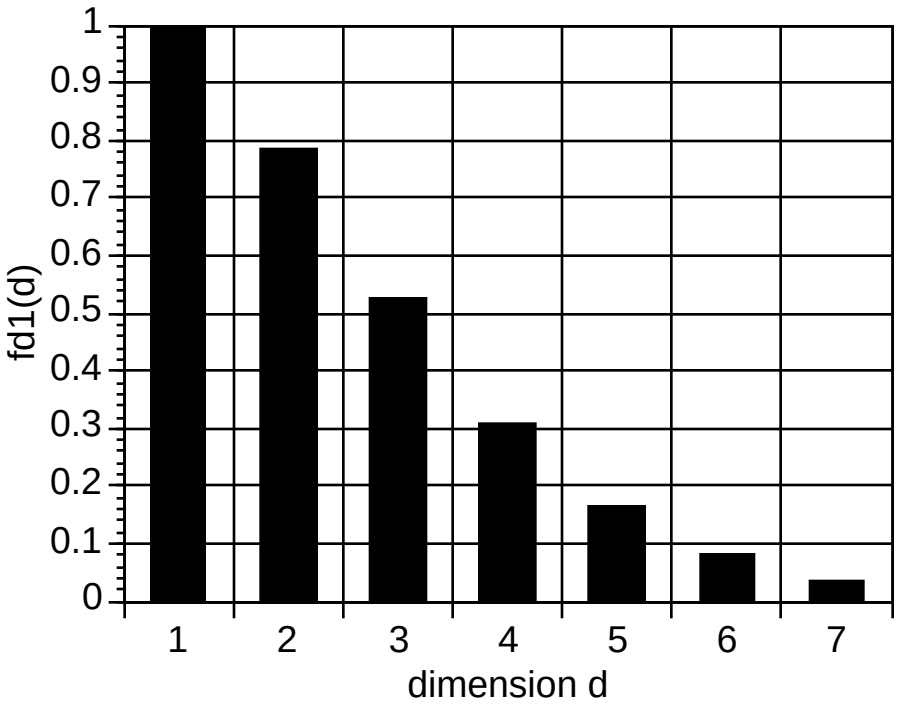


Figure 15. Fractional volume of a hypersphere inscribed in a hypercube as a function of dimensionality.

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[29,30]

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

$$f_{d2} = \frac{V_d(r) - V_d(r-\varepsilon)}{V_d(r)} = \frac{r^d - (r-\varepsilon)^d}{r^d} = 1 - \left(1 - \frac{\varepsilon}{r}\right)^d \tag{11}$$

In Figure 16 observe, for the case ε = r/5, how as the dimension increases the volume concentrates in the outside shell.

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

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

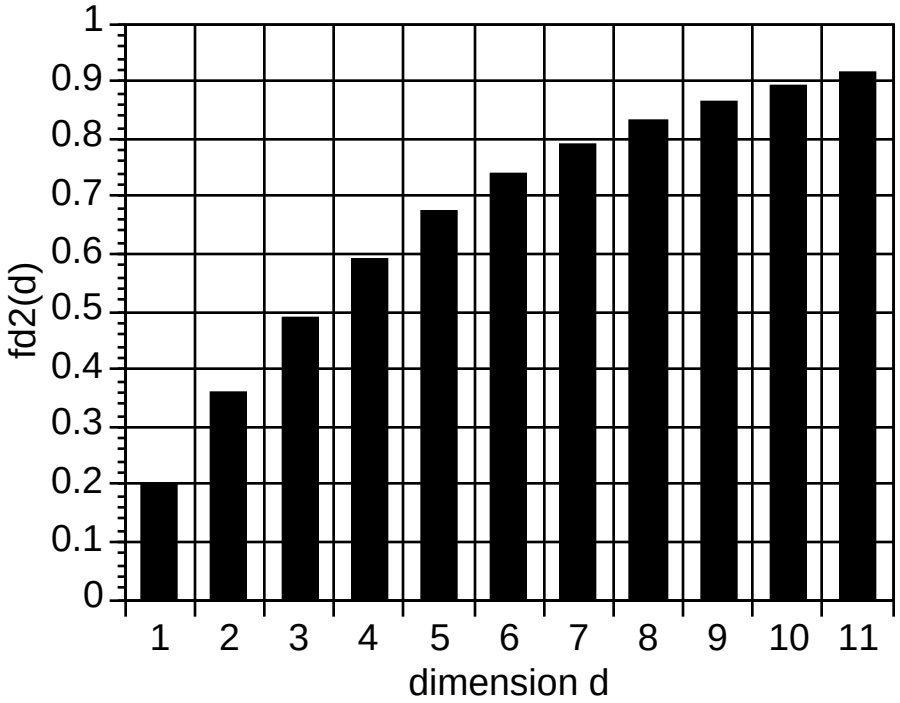


Figure 16. Volume of a hypersphere contained in the outside shell as a function of dimensionality for ε = r/5.

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:

$$\frac{X_1^2}{\lambda_1^2} + \frac{X_2^2}{\lambda_2^2} + \cdots + \frac{X_d^2}{\lambda_d^2} = 1 \tag{12}$$

The volume is calculated by the equation[31]:

$$V_e(\lambda_i) = \frac{2\prod_{i=1}^{d}\lambda_i}{d}\,\frac{\pi^{\frac{d}{2}}}{\Gamma\left(\frac{d}{2}\right)} \tag{13}$$

The volume of a hyperellipsoid defined by the equation:

$$\frac{X_1^2}{(\lambda_1 - \delta_1)^2} + \frac{X_2^2}{(\lambda_2 - \delta_2)^2} + \cdots + \frac{X_d^2}{(\lambda_d - \delta_d)^2} = 1 \tag{14}$$

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

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

$$V_e(\lambda_i - \delta_i) = \frac{2\prod_{i=1}^{d}(\lambda_i - \delta_i)}{d} \frac{\pi^{\frac{d}{2}}}{\Gamma\left(\frac{d}{2}\right)} \tag{15}$$

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

$$f_{d3} = \frac{\prod_{i=1}^{d}(\lambda_i - \delta_i)}{\prod_{i=1}^{d}\lambda_i} = \prod_{i=1}^{d}\left(1 - \frac{\delta_i}{\lambda_i}\right) \tag{16}$$

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

$$f_{d3} = \prod_{i=1}^{d}\left(1 - \frac{\delta_i}{\lambda_i}\right) \le \prod_{i=1}^{d}(1 - \gamma_{min}) = (1 - \gamma_{min})^d \tag{17}$$

Using the fact that $f_{d3} \ge 0$, 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 a high dimensional feature space 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,

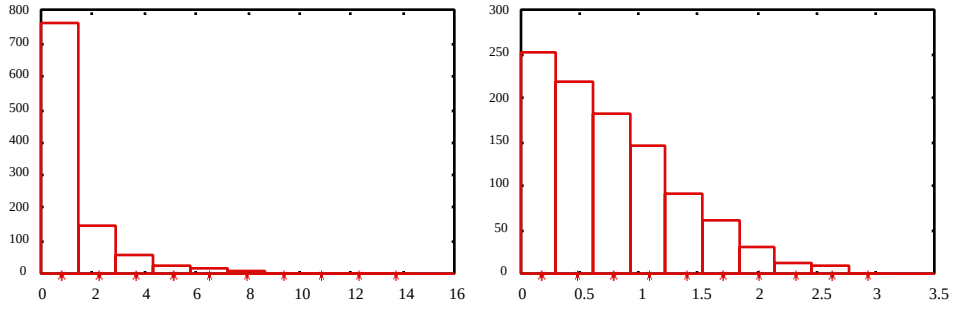
- Uniformly distributed data will be more likely to be collected in the corners,

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.

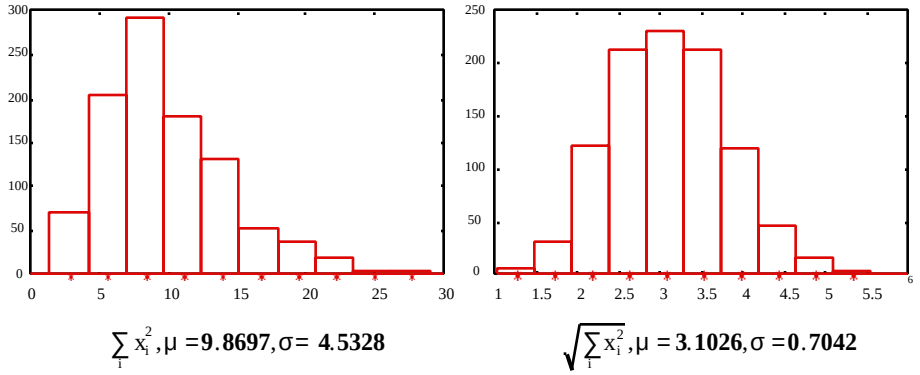
Support for this tendency can be found in the statistical behavior of normally and uniformly distributed multivariate data at high dimensionality. It is expected that as the dimensionality increases the data will concentrate in an outside shell. As the number of dimensions increases that shell will increase 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 17 and 18 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



Normal, dimensions = 10



Normal, dimensions = 220

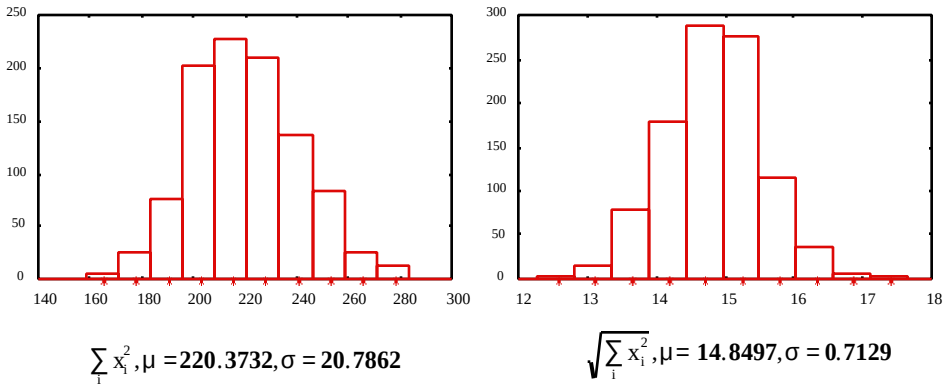
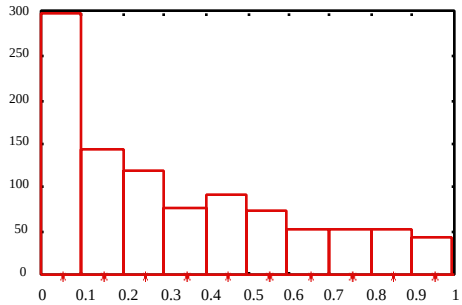
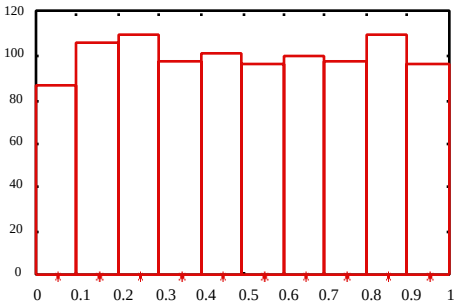


Figure 17. Histograms of functions of Normally distributed random variables.

Uniform, dimensions = 1

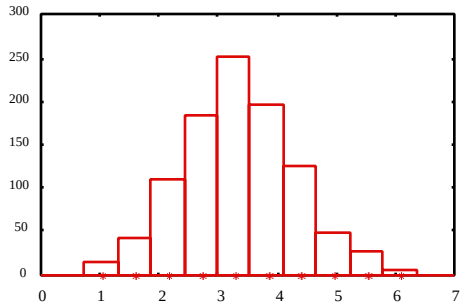


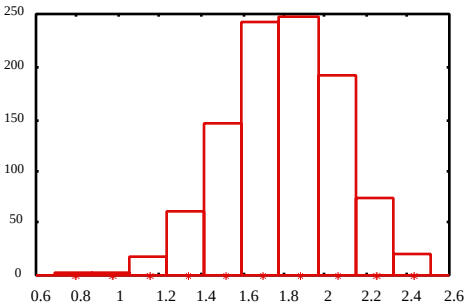


$\sum_i x_i^2, \mu = 0.3277, \sigma = 0.2883$

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

Uniform, dimensions = 10

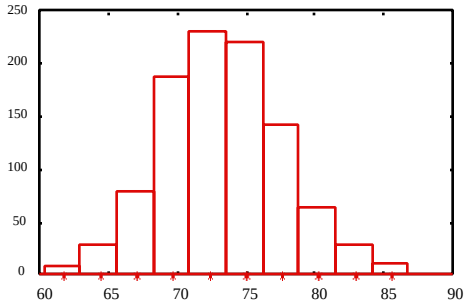


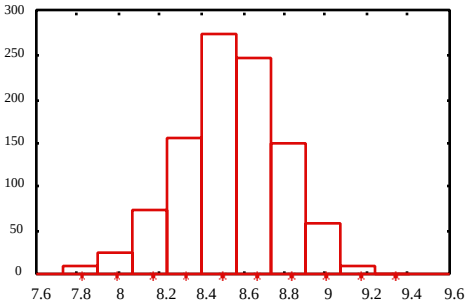


$\sum_i x_i^2, \mu = 3.3444, \sigma = 0.9390$

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

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 18. 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 16 and 17 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[32]: E(R) = d, Var(R) = 2d . This conclusion supports the previous thesis.

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

D. The diagonals are nearly orthogonal to all coordinate axes[33,34]

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

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

Figure 19 illustrates how the angle between the diagonal and the coordinates, θ_d, approaches 90^o with increases in dimensionality.

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

33 Scott, D. W. "Multivariate Density Estimation." John Wiley & Sons, pp. 27-31, 1992.

34 Wegman, E. J., "Hyperdimensional Data Analysis Using Parallel Coordinates," Journal of the American Statistical Association, Vol. 85, No. 411, 1990

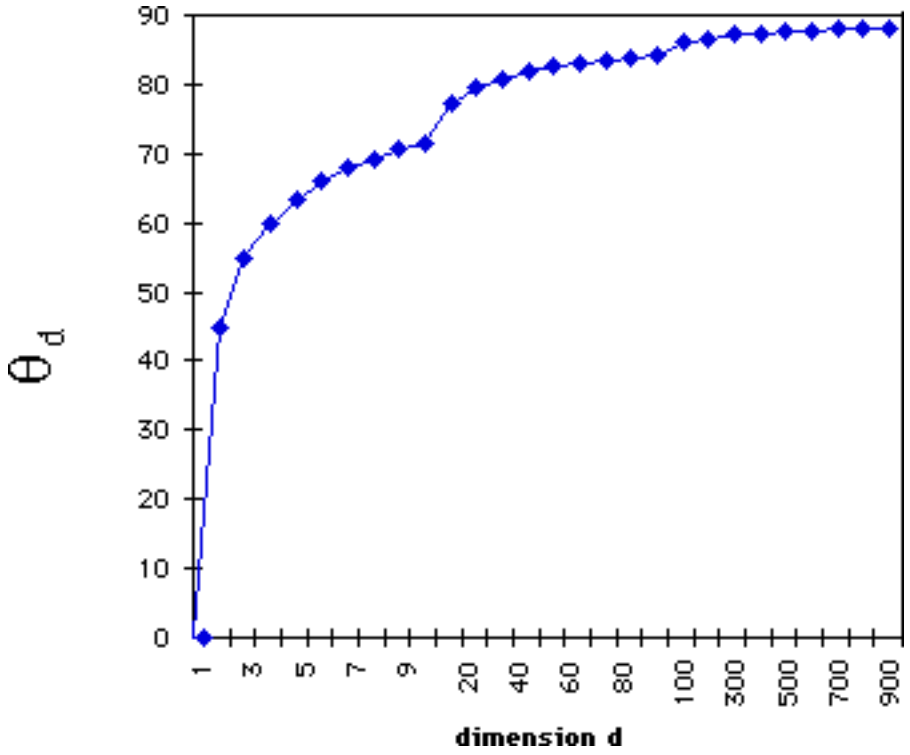


Figure 19. Angle (in degrees) between a diagonal and a Euclidean coordinate vs. dimensionality.

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:

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

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

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

As previously stated, Fukunaga[35], in a given circumstance, 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[36]. 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[37,38].

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. As a matter of fact, since the curves of Figure 12 are montonically increasing, ultimately one can expect 100% accuracy, on the average. 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 shown in Figure 13 where, 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[39,40] that, as the dimensionality tends to infinity, lower dimensional linear projections will approach a normal (Gaussian) distribution with probability approaching one (see Figure 20). Normality in this case implies a normal or a combination of normal distributions. This lends credence to using Gaussian classifiers after having reduced the dimensionality via feature extraction and indeed, to using class mean vectors and covariance matrices in evaluating the separability of classes. Properly used, parametric classifiers should provide good performance, and nonparametric schemes, with their higher demands for training data, should not be needed.

---

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

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

37 Scott, D. W. "Multivariate Density Estimation." John Wiley & Sons, pp. 208-212, 1992.

38 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.

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

40 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.

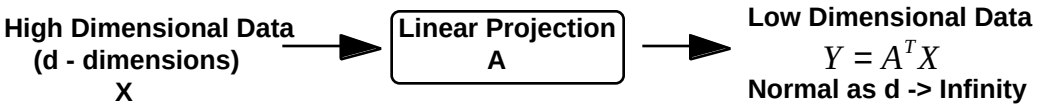


Figure 20. The tendency of lower dimensional projections to be normal.

Feature Extraction.

The findings above point to the importance of finding the lowest dimensional effective subspace to use for classification purposes. Thus, feature extraction becomes an important tool in the analysis process for hyperspectral data. As a result, feature extraction methods already existing in the literature were studied relative to the high dimensional remote sensing context. The most suitable appeared to be Discriminate Analysis Feature Extraction (DAFE). The basic concept[41] for DAFE is to form a linear combination of the original features so as to maximize the ratio,

$$\frac{\sigma_A^2}{\sigma_W^2} = \frac{\text{between classes variance}}{\text{within classes variance}}$$

The calculation of the needed linear transformation is fast and straightforward. Even so, it has several significant shortcomings for this environment, among them being that it does not perform well for cases where there is little difference in class mean vectors. It also only generates reliable features up to one less than the number of classes for the given problem.

For use in problems where these shortcomings would be serious, Decision Boundary Feature Extraction (DBFE) was created[42,43,44]. DBFE also determines an optimum linear transformation to a new feature space. It uses training samples directly to determine discriminately informative and discriminately redundant features, and results in eigenfunctions which define the required transformation. The eigenvalues resulting are directly related to the usefulness of the corresponding features in discriminating among the given classes. Thus this transformation has the advantage of showing the analyst directly how many features must be used.

However, both DAFE and DBFE calculations begin with computation in the full dimensional space in order to find the optimal transformation to a lower dimensional space, thus these calculations may, too, suffer from small training set situations. To deal with this limitation, a class-conditional pre-processing algorithm was designed

41 Richards, John A, *Remote Sensing Digital Image Analysis, An Introduction*, Second Edition, Springer Verlag, 1993, pp 255 ff.

42 Chulhee Lee 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.

43 Chulhee Lee and David A. Landgrebe, "Decision Boundary Feature Selection for Non-Parametric Classification," *IEEE Transactions on System, Man, and Cybernetics*, Vol. 23, No. 2, March/April, 1993, pp. 433-444.

44 Chulhee Lee and David A. Landgrebe, "Decision Boundary Feature Extraction for Neural Networks," *IEEE Transactions on Neural Networks*, Vol. 8, No. 1, pp. 75-83, January 1997.

based upon a method known as projection pursuit[45,46]. This algorithm does the necessary calculations in the projected space, rather than the original, high dimensional space. Figure 21 shows the overall scheme. The data at point Φ might be 200 dimensional. Through projection pursuit, a subspace of perhaps 20 dimensions might be determined, and in this case, all calculations are done at a dimensionality of 20. This can then more optimally be followed by DAFE or DBFE to find a subspace of perhaps 10 dimensions in which to do the classification. In this way, maximal advantage can be taken of a training set of limited size.

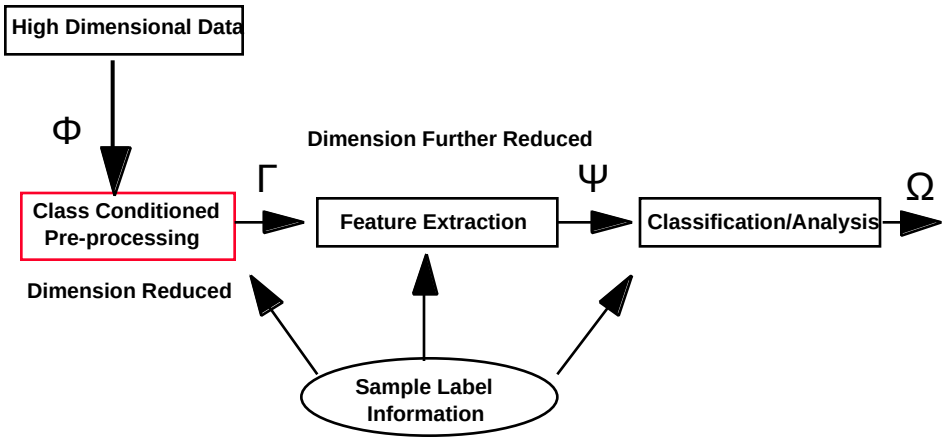


Figure 21. Classification of high dimensional data including reprocessing of high dimensional data.

Summary and Conclusions

What is sought are powerful, general analysis procedures that approach the optimum in information extraction capabilities and yet are within the reach of and practical for a broad range of Earth scientists and other remote sensing practitioners. The techniques must be 1) powerful in terms of accuracy and detail with respect to the classes which can be discriminated, 2) objective in their performance, 3) robust with regard to the breadth of discipline problems which can be successfully approached, and yet 4) must appear sound and practical to scientists with a broad set of discipline backgrounds. They must be derived with appropriate mathematical rigor, but in the end, they must meet the practical conditions of the randomness of the scene, noise introduced by the atmosphere, the scene, and the sensor, and the varied skills and expectations of the users.

Summarizing to this point, the key conclusions expressed above are,

45 Luis Jimenez and David Landgrebe, “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.

46 Luis Jimenez and David Landgrebe, “Projection Pursuit in High Dimensional Data Reduction: Initial Conditions, Feature Selection and the Assumption of Normality,” IEEE International Conference on Systems, Man, and Cybernetics, Vancouver, Canada, October 22-25, 1995.

- The complexity of the scene and the dynamic nature of it is so dominant, that, except for the extraction of relatively simple information, supervision of classifiers must be redone for every new data set collected.

- Of the three data space presentations discussed, the feature space is the most useful for analytical purposes, though the other two are helpful for visualization purposes.

- The typically rather uniformly distributed nature of data in feature space makes clear why entirely unsupervised classification schemes are not likely to produce satisfactory results for multispectral discrimination purposes.

- A stochastic or random process approach for data modeling has been chosen for reasons of its rigor and power, and the large stable of tools that prove of pivotal usefulness in the work.

- Both first order variations (e.g. mean values) and second order variations (e.g. covariance matrices) are found to be significant in the discrimination process. On a case specific basis, either is likely to provide the most significant contribution to the ability to discriminate between classes. Neither should be ignored without good justification.

- A significant relationship has been demonstrated between the number of spectral bands, the amount of ancillary data available for classifier supervision, and overall classifier accuracy achievable.

- A significant relationship has also been found between classifier complexity, the amount of ancillary data available for classifier supervision, and overall classifier accuracy achievable.

- Given these findings, a generic hierarchy of classifier algorithms has been given against which to judge more specialized algorithms for their likely performance robustness.

- A number of novel characteristics of high dimensional spaces are presented which bear upon the analysis of hyperspectral data. Among them are the facts that,
  - Higher dimensional spaces are mostly empty, because of the rapidity with which volume increases with dimensionality. This suggests the importance of feature extraction algorithms to find the lower dimensional space in which the most important discriminate structure exists.
  - Unlike three dimensional space, data in hyperspace tends to concentrate in the corners of a hypercube, in the outer shell of a hyperellipsoid, and thus in the corners of a uniform distribution and the tails of a Gaussian distribution. This increases the importance of having adequate numbers of training samples when estimating high dimensional density function parameters, and of using the lowest dimensionality which will provide best results.

- The diagonals in high dimensional are nearly orthogonal to all coordinate axis. This has implications relative to averaging features.
- For most high dimensional data sets, as the dimension increases, lower dimensional linear projections have a tendency to be normal, or a combination of normal distributions. Thus, the Gaussian assumption becomes better justified after feature extraction to a lower dimensional space.
- The required number of labeled samples for supervised classification increases as a function of dimensionality, and more so with increased generality of the classifier algorithm used.

As a result of these findings, feature extraction algorithms assume increased importance, and a two stage feature extraction process has been put forth in order to take maximal advantage of the dimensionality available when, as is usually the case in the remote sensing circumstance, the number of training samples is limited.

Some of these results raise other issues with regard to current and future analysis procedures. For example, it is seen that second order variations can be and often are more significant than first order ones in making discrimination between classes possible. "Data correction" procedures are now common in preparing data for analysis, but most, if not all, such procedures are directed at adjusting for first order effects only. Generally, the impact they might have on the second order variations in a data set have not been considered. The positive value they may have is taken on faith and has not generally been subject to study. There is some evidence[47] that they may not always have this assumed positive effect, and indeed, there effect could be detrimental in some cases. Other such issues of this nature need also to be addressed.

Substantial progress toward an optimal and robust hyperspectral data analysis procedure has been made based upon the findings reported in this paper, however, some key problems remain if such a procedure is to have significant widespread impact. Among these are the need to,

- Make the analysis process viable for smaller and smaller training sets, down to one spectrum for some classes, while still retaining optimal characteristics of both first and second order statistics to the extent possible. It is the case that no one likes the idea of needing to retrain a classifier for each data set. On the other hand, the dynamic nature of the Earth's surface requires it if many of the more complex and challenging information extraction tasks are to be completed successfully.

- Make the analysis process systematic, making as much of the complexity of it transparent to the user, so that it appears attractive and reasonable to the user community. The need is to take advantage of human knowledge and

47 Joe Hoffbeck and David A. Landgrebe, “Effect Of Radiance-To-Reflectance Transformation And Atmosphere Removal On Maximum Likelihood Classification Accuracy Of High-Dimensional Remote Sensing Data,” Proceedings of the International Geoscience and Remote Sensing Symposium (IGARSS'94), CD-ROM pp. 3289-3294, Pasadena, Calif, August 8-12,1994.

perception while at the same time not requiring analysts to be highly trained and experienced signal processing engineers.

As an example, the labeling of samples for training sets seems onerous to everyone, and unreasonable or unnecessary to some. It is certainly desirable to avoid doing this whenever possible. However, the more challenging information extraction problems simply require it. There often are ways to mitigate the problem which are situation specific in any given case. An example in a geologic survey case, making use of chemical spectroscopy characteristics has been given[48] as an illustration in one circumstance. Any way to make this part of the analysis process acceptable to the Earth scientist or practitioner would be an important contribution to the field. It is to these and related objectives that future research needs to be directed.

And finally, it is recognized that a key problem is to deliver the knowledge and algorithms derived during this research to the potential users. For this purpose, an application program for personal computers has been created with a basic multispectral data analysis capability and made available to the community without charge. Then as new algorithms emerge from the research, they are incorporated into the program and new versions of it issued. In this way, new algorithms, which may be quite complex to implement may be tried by users with a minimum of effort on their part. The program, called MultiSpec, together with substantial documentation is available for anyone interested to download from the world wide web. The URL for the web site is

http://dynamo.ecn.purdue.edu/~biehl/MultiSpec/

Some of the algorithms mentioned above which it now contains are, Discriminate Analysis Feature Extraction (DAFE), Decision Boundary Feature Extraction (DBFE), Statistics Enhancement, and Statistics Image. It also contains the spatial/spectral analysis algorithm created some years ago called ECHO, as mentioned earlier. This algorithm has proven to be easy to use, computationally efficient and effective in increasing classification accuracy, but it is not simple to implement, and this has no doubt inhibited its wider use.

48 Hoffbeck, Joseph P. and David A. Landgrebe, "Classification of Remote Sensing Images having High Spectral Resolution," *Remote Sensing of Environment*, Vol. 57, No. 3, pp. 119-126, September 1996.