Information Extraction Principles and Methods for Multispectral and Hyperspectral Image Data

David Landgrebe
School of Electrical & Computer Engineering
Purdue University, West Lafayette, IN 47907-1285
Voice: 765-494-3486 Fax: 765-494-3358
landgreb@ecn.purdue.edu

Introduction

In the U.S., the era of space-based multispectral remote sensing of land areas began soon after the launch of the first Earth satellite in 1957 and the creation of NASA in 1959. Satellites to observe the weather became an initial focus of Earth-looking satellites, with the first U.S. satellite designed for that purpose launched in 1960. Research on land remote sensing began soon thereafter and was initially focused on the Earth's renewable and non-renewable resources, and especially on food and fiber production. The need envisioned was very much use-driven, as compared to a purely scientific interest. The primary objective of the early research was to create a practical but especially, an economical technology to supply usable and needed information about the Earth's resources for both application and scientific interests. Vertical views of the Earth's land surface provide a unique vantagepoint, one different than ordinary human experience. The world does not look the same from altitude looking down. But it is not so much the simple uniqueness of this vantagepoint that was the attractive feature as the fact that from altitude one can see more, thus suggesting the economic value of the synoptic view and the ability to cover large areas quickly and inexpensively. On the other hand, this raises the question of dealing with large quantities of data.

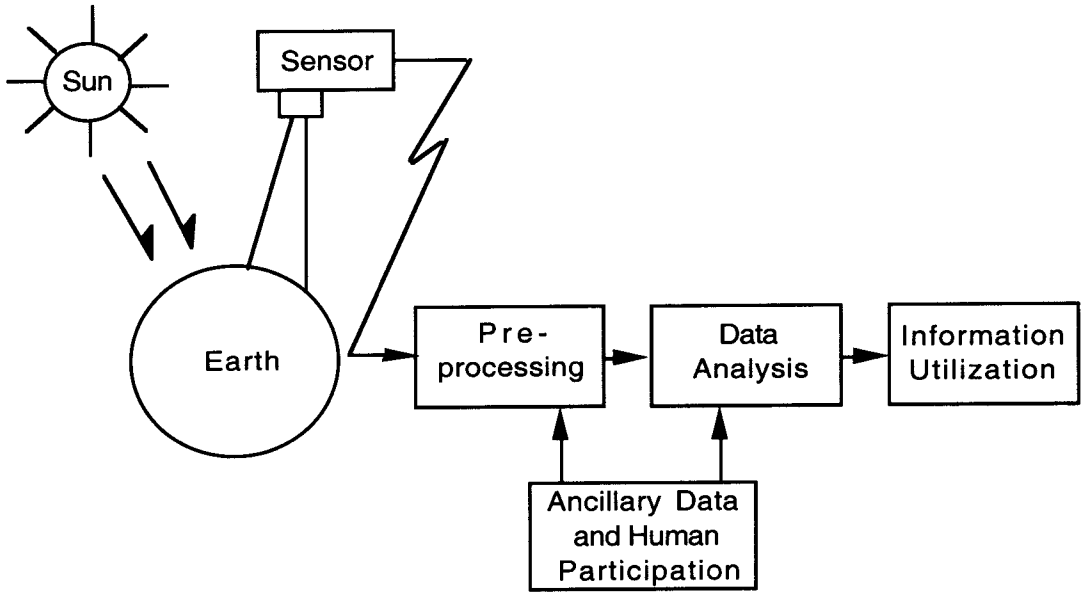
This motivation and line of thinking led to considering ways to collect information via a space platform from image data with the lowest spatial resolution usable. The spatial resolution of a spaceborne sensor is one of the more expensive parameters. Higher spatial resolution leads not only to larger quantities of data for a given area as resolution is increased, but to larger (thus heavier) sensor systems, increased precision requirements on spacecraft guidance and control, wider bandwidth downlinks, and the like. This is what led to the concept of using spectral measurements of a pixel to identify what ground cover the pixel represents, rather than to use spatial variations (imagery) and more conventional image processing techniques, methods which generally require higher spatial resolution.

For similar reasons, methods were pursued which marry the unique capabilities of both human and computer. Rather than seeking a fully "automatic" system, it appeared to be much wiser to attempt to construct a data analysis scheme that takes advantage of keen perceptive and associative powers of the human in conjunction with the objective quantitative abilities of computers. Fully manual systems in the form of air photo interpretation had been used for many years. Though a great deal of work has gone into fully machine implemented systems, the problem of devising completely automatic schemes has proved to be quite daunting with only limited success even to the current time. The most functional schemes that have become available in a practical sense can perhaps best be described as human assisted machine processing schemes. Where the focus has been placed on "image enhancement," such systems might be labeled machine assisted human schemes. In the following we will focus on the human assisted machine approach. Indeed, there is a down side to even seeking a fully automatic system. As compared to a machine, the human has exceptional

This work was originally prepared for and appears in **Information Processing for Remote Sensing**, edited by C. H. Chen, published by the World Scientific Publishing Co., Inc., 1060 Main Street, River Edge, NJ 07661, USA (Spring, 1999). In addition to general works such as this one, this book contains sections on Pattern Recognition, SAR Image Processing and Segmentation, Parameter Extraction, Neural Network and Fuzzy Logic Methods, Change Detection, Knowledge-based Methods and Data Fusion, Image Processing Algorithms including wavelet analysis techniques, Image Compression, and Discrimination of Buried Objects.

perceptive powers and abilities to abstract and to generalize. A fully automatic system would have to forego these capabilities, and thus is likely to be more limited and less robust in its abilities and perhaps less economical as well.

The following figure shows in broad concept form, an overview of the complete information system. It is important to have such an overview of the entire system in mind, the illumination, the Earth surface, factors related to it observation, and the data and its processing, when planning the analysis of a data set, rather than simply considering the data and its processing in isolation of the other elements. What is the case in other parts of the system has a significant impact on how the processing and analysis should proceed.



A conceptual overview of a remote sensing based information system.

For reasons of simplicity, we will focus the discussion on passive systems in the optical portion of the electromagnetic spectrum. Thus the sun is used as the illumination source of the scene on the Earth. The sensor system then measures the reflected (or emitted) energy from the areas of interest. The data thus collected is next transmitted to the Earth for processing, information extraction and utilization. As seen, then, a key element is the merging of ancillary data with the data stream and the use of human perception and guidance in the processing of the data.

It is logical to think of this system in three parts. The first will be referred to as the scene. It consists of that part of the system in front of the sensor, including the sun, the Earth's surface, and the atmosphere. This part of the system has two distinguishing characteristics:

1. It is that part of the system which is not under human control, not at the time of system design and not later when the system is being operated, and,

2. It is by far the most complex and dynamic part of the overall system.

Both of these have a very considerable impact on how one must address the problem of analysis of data from such a system. Obviously, since this first part of the system is not under human control, it cannot be designed or optimized; the best one can do is learn as much about the scene that must

be dealt with as possible. In addition, then, the complexity of the scene and its dynamism dictates what types of approaches to scene models and what data analysis approaches might be useful. It is very easy to underestimate this complexity and dynamism of the scene and to undertake too simplistic an approach to data analysis, thus limiting the robustness of the procedure and the accuracy and detail of the information that can result. For example, the spectral response of a given type of vegetation may be expected to change significantly from day to day due to growth and maturity factors, prior weather conditions, and the like. It may be expected to change even from minute to minute and from one part of a scene to another due to wind effects, the sun angle-view angle combination, and other variables. In general, one cannot count on a vegetative canopy having a stable "spectral signature."

The second part of the system is the sensor portion. This portion of the system is characterized by the fact that, though it is under human design control, it is usually not under the control of the analyst at the time of data acquisition. Thus, the analyst must pretty much accept what is given, in terms of the parameters of the data produced, i.e., the spectral and spatial resolution, the quantization precision, the sensor field of view and look angle and the like. Though the system designer may select these characteristics in the process of optimizing the system for a certain class of uses, the individual user usually does not have a choice of them for the particular application, site, and time of season in mind.

It is the third part of the system, all of that after arrival of the data at the processing point, over which the analyst has the greatest control. Thus it is here that choices can be made with regard to algorithm selection and operation to optimize performance to the specific data set and use. In the remainder of this chapter, we will explore the factors that go into making these choices.

The Three Views Of Data

Extending the above line of thinking, how one thinks about the data in a multispectral data set, generally speaking, may be from any of three different points of view. We will explore these briefly, in terms of a data representation scheme or a representation space.

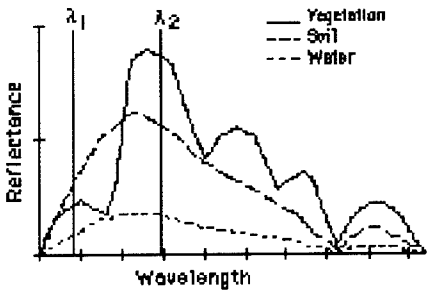
1. **Image Space.** Perhaps the first thought about how to view a new data set is to think of it as an image. This is a quite natural first thought in that the human vision system is a quite wide channel into the human brain; it is one that is thus very attractive in a human sense. The concept here is to display the data samples in relation to one another in a geometric, or more properly, geographic sense, thus providing a "picture" of the ground scene for the human viewer. How these pixels relate to one another can be information-bearing. However, given the basis for the acquisition of information by multispectral means described above, such a data presentation does not carry a large proportion of the information that is obtainable from such multiband data. One may only view the data in one (BW) or three (Color) bands at a time. In addition, between band relationships are not very apparent. It is, on the other hand, very useful in providing an overview of the data, and it can often make apparent to the analyst certain kinds of faults in the data.

   A key use of imagery in multispectral processing is to serve as a means for the analyst to associate multispectral data points (pixels) with specific locations (points) in the ground scene. In the analysis process it is very useful, therefore, in the labeling of pixels in the data set as

training samples, i.e., examples of the classes that the analyst wishes to identify. Fundamentally, the analysis process consists of bringing together the wishes of the analyst in terms of what classes are desired with the scene spectral properties as expressed in the data. One cannot expect a satisfactory final product unless the analyst can carefully and completely define which spectral properties are intended to belong to which classes. The training samples are the means for doing this. Thus the means for allowing the analyst to accomplish this is the key to successful analysis. As will be seen, this fundamental step becomes even more crucial for hyperspectral data.

The spatial relationships available in an image expression of multispectral data has also been found effective in a limited way as an adjunct to spectral relationships in extracting information from the data.

2. **Spectral Space.** The emergence of the multispectral concept began to focus attention on how the response measured in an individual pixel varies as a function of wavelength as an information-bearing aspect, and indeed, perhaps the key such aspect. The idea is that, if response vs. wavelength effectively conveys needed information by which to identify the contents of an individual pixel, this provides a fundamental simplicity that is important from a processing economics point of view. In this circumstance, pixels can be processed one at a time, a much simpler arrangement than would be needed using so-called picture processing or image processing schemes. It is inherently more suited to the computer and quantitative representation of data. Further, compared to image processing schemes, being able to label each pixel individually results in a higher resolution result than, for example, a label associated with a neighborhood of pixels, as would be the case for conventional image processing schemes.

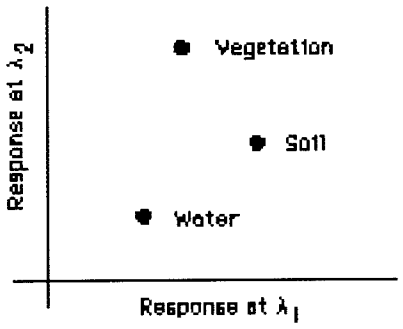


Response as a function of wavelength has the very useful characteristic that it provides the analyst with spectral information that is often directly interpretable. Especially when a high degree of spectral detail is present, characteristics of a given pixel response can be related to physical properties of the contents of the pixel area. For example, one can easily tell whether a pixel contains vegetation, soil, or water. In the case of high-resolution spectra, one may even be able to identify a particular molecule based upon the location of specific absorption bands, in a manner similar to that used by chemical spectroscopists in the laboratory. Thus, for the analyst, a display of spectral response can provide a direct link to physical properties. For this reason, fundamental scientists tend to initiate their thinking about multispectral data from the point of view of spectral space. In limited cases, such spectral curves may be used directly in machine implemented spectral matching schemes.

But viewing a graph of response vs. wavelength for an individual pixel does not provide the whole story so far as the relationship between spectral response and information available from the scene is concerned. The spectral response of any given Earth surface cover type tends to vary in a characteristic way. The spectral response of corn field pixels at a given time, for example, is not uniform for all the pixels of the field, but varies in a characteristic way about some mean value. This is due to the relationships between the size and mixture of leaves, stalks, soil background, the physiology of the plants, etc, leading to different mixtures of illuminated and shadowed surfaces and the like. This variation is, to a useful degree, diagnostic of the plant species and thus useful in discriminating between corn and the other plant species that may be in the scene. But such variation, though present in the spectral responses from the field, is not easily discernable from a presentation of a series of plots of response vs.

wavelength for the class. For this purpose, the third form of data expression, the feature space proves more useful, especially in relation to machine processing.

3. **Feature Space.** If one samples the spectral response at two different wavelengths, λ_1 and λ_2 as shown above, the values resulting can be plotted as shown at right, thus creating a 2-dimensional display. If one samples at more values of λ, for example, 10 values of λ, the point representing each spectral response would then be a point in 10-dimensional space. This turns out to be an especially useful way of representing spectral responses. If one samples the spectrum at enough wavelengths, so that one could reconstruct the spectral curve from the samples, the information the spectral response contains is preserved, and it is now represented as a vector. Though one cannot show such a point graphically, a computer can more easily deal it with than with a graph.



Further, this is a mathematical representation of what a multispectral sensor does, i.e. sampling the spectral response in each of N spectral bands. The result is then an N-dimensional vector containing all the available spectral information about that pixel. The advantage of this type of representation is that it is a quantitative way of representing not only the numerical values of individual pixels, but also how the values for a given material may vary about their central or mean value. As indicated above, this turns out to often be quite diagnostic of the material. We shall deal more fully with this fact later.

Each of these three data spaces, then, has its advantages and limitations. Image space shows the relationship of spectral response to its geographic position, and it provides a way to associate each pixel with a location on the ground. It also provides some additional information useful in analysis. Spectral space often enables one to relate a given spectral response to the type of material it results from. Feature space provides a representation especially convenient for machine processing, for example, by use of a pattern recognition algorithm. It is this latter method of data analysis, a very common one for multispectral analysis, we will explore next, using it as a means for exploring how information is contained in spectral data, and how it may be extracted.

Analysis Algorithms And the Relationship With Ancillary Data

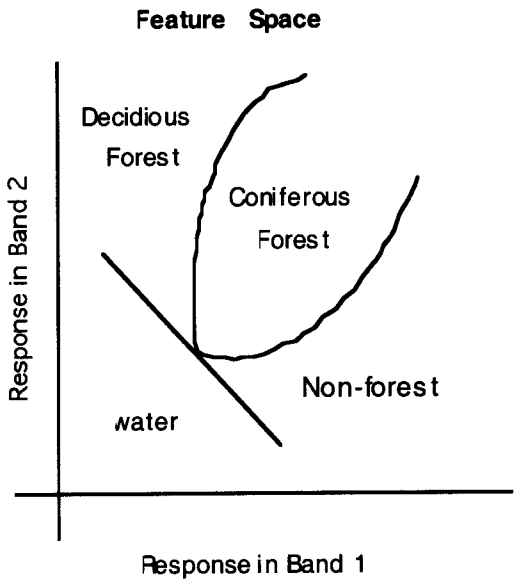
We next consider the process of analysis or extracting information from the data. The term data analysis can mean many things in different applications. A common one is to make a thematic map of the scene by associating a class label with each pixel of the scene. For simplicity, we will focus on that objective. In terms of feature space, one might think of the analysis process as that of delineating the region of the feature space associated with each class of interest. For example, one might somehow determine which region of the space contains spectral values associated with wheat, which part contains forest pixels, which contains urban pixels and the like. In terms of the thematic map concept, this means partitioning up the feature space such that each possible location in the space has a unique class label associated with it.

A very common means for doing this is via a pattern recognition algorithm using a *discriminant function.* Assume that the digital values of the response in each of the N spectral bands of a pixel form a vector designated **X**. Then assume that, for the M classes that exist in the data set, a set of M functions $\{g_1(\mathbf{X}), g_2(\mathbf{X}), \ldots g_M(\mathbf{X})\}$ can be found such that $g_i(\mathbf{X})$ is larger than the others when the pixel in question contains class i. Then the classification rule to be machine implemented can be defined as follows.

Let ω_i denote the i^{th} class. Decide **X** is a member of class ω_i if and only if

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

This procedure, then, defines how every pixel can be assigned to a class by means of a computer. The computer algorithm has only to evaluate each of the M $g_i(\mathbf{X})$ functions for each pixel and call out the class for which the g-function is a maximum.



The next question is how to find the set of M discriminant functions in any given case. There have been a number of ways to do this defined in the pattern recognition literature. A currently popular one is through the use of a *neural network.* One begins with a pre-labeled set of samples, called training samples or design samples, which are representative examples of each of the classes one wishes to identify. Then an iterative scheme may be used, as follows.

1. Choose a parametric form for the discriminant function, e.g.
   $g_1(\mathbf{X}) = a_{11}x_1 + a_{12}x_2 + b_1$
   $g_2(\mathbf{X}) = a_{21}x_2 + a_{22}x_2 + b_2$
2. Initially set the a's and b's arbitrarily, e.g. to +1 and -1
3. Sequence through the training samples, calculating the g's and noting the implied decision for each. When it is correct, do nothing, but when it is incorrect, augment (reward?) the a's and b's of the correct class discriminant, and diminish (punish?) those of the incorrect classes. For example, if **X** is in ω_1, but $g_2 > g_1$ then let

$$a'_{11} = a_{11} + \alpha x_1 \qquad a'_{21} = a_{21} - \alpha x_2$$
$$a'_{12} = a_{12} + \alpha x_2 \qquad a'_{22} = a_{22} - \alpha x_2$$
$$b'_1 = b_1 + \alpha \qquad b'_2 = b_2 - \alpha$$

4. Continue iterating through the training samples until the number of incorrect classifications is zero or adequately small.

Neural network implementations of the discriminant function concept can be of varying degrees of complexity. For example, one may add additional terms to the illustrative functions $g_1(\mathbf{X})$ and $g_2(\mathbf{X})$, as defined above thus making them nonlinear. In doing so, one increases the number of parameters that must be correctly adjusted, thus increasing the generality of the classifier, but also the complexity and duration of the training process. Neural network implementations are characterized by the fact that they require no foreknowledge of the nature of the classes, since they rely totally on an empirical use of the training samples. On the other hand, they cannot make use of any foreknowledge that one might have, and they usually require a large amount of computation time in the iterative process of accomplishing the training. Further, neural network implementations

do not lend themselves to analytical evaluation very easily. It is more difficult to predict the performance, for example, and therefore to adjust the configuration and parameters to an optimum.

A second common approach to determining a set of discriminant functions utilizes a statistical approach. The training samples, instead of being used in an empirical calculation as above, are used to evaluate a probabilistic model for each class. That is, they can be used to estimate the probability density function associated with each class. Recall that the value of a probability density function at any point indicates the relative likelihood of that point. Thus, by using class probability density functions as discriminant functions, one is deciding in favor of the most likely class for each pixel.

More formally, let $p(\mathbf{X}|\omega_i)$ be the (N-dimensional) probability density function for class i, and $p(\omega_i)$ be the probability that class i occurs in the data set. Then, the decision rule becomes:

Decide $\mathbf{X}$ is in class ω_i if and only if

$$p(\mathbf{X}|\omega_i)p(\omega_i) \geq p(\mathbf{X}|\omega_j)p(\omega_j) \text{ for all } j = 1,2,...,m$$

This decision rule is known as the Bayes Rule, and it can be shown that it provides the minimum probability of error for the density functions used.

Any of a wide variety of probabilistic models can be used, in either parametric or non-parametric form. Often the density for the classes can be assumed to be normally or Gaussianly distributed. In this case, the class probability density function becomes,

$$p(\mathbf{X}|\omega_i) = (2\pi)^{-N/2} |\Sigma_i|^{-1/2} \exp\{-1/2 (\mathbf{X} - \overline{\mathbf{X}}_i)^T \Sigma_i^{-1} (\mathbf{X} - \overline{\mathbf{X}}_i) \}$$

where $\overline{\mathbf{X}}_i$ is the class mean value and Σ_i is its covariance matrix. In this case, one has only to use the training samples to estimate the class mean vectors and covariance matrices, a very short calculation.

Furthermore, if the Gaussian assumption is applicable, as it often is, due in part to the Central Limit Theorem, significant simplification of the process can be made. If $p(\mathbf{X}|\omega_i)p(\omega_i) \geq p(\mathbf{X}|\omega_j)p(\omega_j)$ for all $j = 1,2,...,M$, then it is also true that

$$\ln p(\mathbf{X}|\omega_i)p(\omega_i) \geq \ln p(\mathbf{X}|\omega_j)p(\omega_j) \text{ for all } j = 1,2,...,M.$$

Thus one may take the following as an equivalent discriminant function, but one that requires substantially less computation time. (Note in this expression that we have dropped the factor involving 2π since it would be common to all class discriminant functions and thus does not contribute to the discrimination.)

$$g_i(\mathbf{X}) = \ln p(\omega_i) - (1/2)\ln|\Sigma_i| - (1/2)(\mathbf{X}-\overline{\mathbf{X}}_i)^T\Sigma_i^{-1}(\mathbf{X}-\overline{\mathbf{X}}_i)$$

or

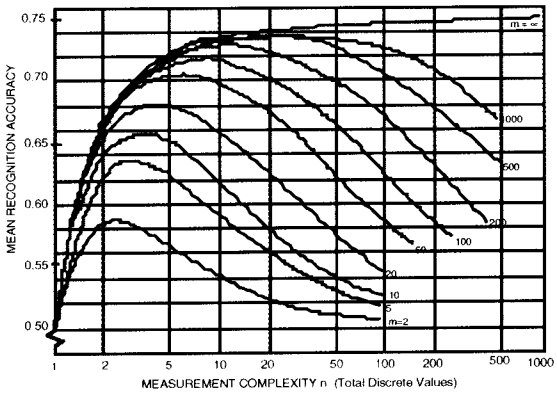
$$2g_i(\mathbf{X}) = \ln \frac{p^2(\omega_i)}{|\Sigma_i|} - (\mathbf{X}-\overline{\mathbf{X}}_i)^T\Sigma_i^{-1}(\mathbf{X}-\overline{\mathbf{X}}_i)$$

Note also that the first term on the right must only be computed once per class, and only the last term must be computed for each pixel to be classified. The calculation that must be implemented is thus quite straightforward and simple.

Thus, to achieve optimal performance, one must (a) have good estimates of the class mean vectors and covariance matrices, after having (b) chosen an appropriate probability model and a proper set of classes. These conditions turn out to be the key ones to achieving good results. We shall next explore these two conditions more fully.

(a) On the importance of accurate class statistics. One of the defining circumstances of the remote sensing problem is the fact that training samples are usually not as numerous as would be desirable. As it turns out, this factor has a strong relationship to the number of spectral bands contained in the measurement and the signal-to-noise ratio of the sensor. That is to say that the number of training samples needed to adequately define the classes quantitatively, regardless of what discriminant function implementation is used, grows very rapidly with the number of spectral bands to be used. To understand how this influences the performance, we begin by drawing attention to the following long-standing theoretical result. [1]

The study in mind investigated in a very generalized fashion the relationship between the accuracy to be expected in a classification to the complexity of measurement used and the number of training samples used. One of the results of this study is given in the figure below. The variable of the vertical axis of this figure is the mean recognition accuracy obtainable from a pattern classifier, averaged over all possible pattern classifiers. Thus the result shown is very general. This is plotted as a function of measurement complexity on the horizontal axis. Here, measurement complexity is a measure of how complex and detailed a measurement is taken. In the case of digital multispectral data, it is related to the number of bins or brightness-level values, k, recorded in each band, raised to the p^{th} power, where p is the number of spectral bands, i.e., the number of possible discrete locations in N-dimensional feature space. For example, for Landsat Thematic Mapper data with its 7 bands of 8 bit data, if all 8 bits of all 7 bands were active, this number would be $(2^8)^7 \approx 7 \times 10^{16}$. The more bands one uses and the more brightness levels in each band, the greater the measurement complexity.



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

The result shown is for the two-class case, and in this figure, the two classes are assumed equally likely. The parameter, m, is the number of training samples used. A perhaps unexpected phenomenon is observed here in that the curve has a maximum. This suggests that for a fixed number of training samples there is an optimal measurement complexity. Too many spectral bands or too many brightness levels per spectral band are undesirable from the standpoint of expected classification accuracy.

While perhaps at first surprising, the occurrence of this phenomenon is predictable as follows. One would expect the separability between classes to increase with increasing numbers of bands, but ultimately, the rate of increase would be expected to slow. Indeed, for the case of $m \rightarrow \infty$ above, the probably accuracy rises rapidly at first, but eventually becomes asymptotic to 0.75, a

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

probability half between 0.5 (chance performance) and 100%. For a fixed, finite numbers of training samples, one would expect the accuracy with which one could estimate the class distribution would decrease as the measurement complexity grows. For example, in the case of Gaussian statistics, the number of parameters to be estimated in the covariance matrix grows rapidly with dimensionality, and the preciseness needed would grow with the increased detail as the number of digital bins grows. Thus there are two counterbalancing effects, one increasing with increasing measurement complexity and the other decreasing. The shape of the above curve is explained, then, by the fact that first the former effect dominates but eventually the latter does so, thus a maximum occurs.

It is significant to note that the value of the maximum in this curve moves upward and to the right as m is increased. The practical implication of this is that one can expect to be able to increase accuracy by using increased numbers of bands and/or signal-to-noise ratio, but to achieve it, increased numbers of training samples, implying increased precision in the estimation of class distributions, will be needed. This observation becomes increasingly important as one moves from lower dimensional data to hyperspectral data with its many 10's to several hundreds of bands. We shall return to this point later.

(b) On defining classes and their probability models. In addition to adequate numbers of training samples, the other key factor in successful analysis is the matter of the definition of classes. There are three conditions for optimal class definition, as follows.

Optimal class definition requires that the classes defined must be

- *Exhaustive.* There must be a logical class to assign each pixel in the scene to.
- *Separable.* The classes must be separable to an adequate degree in terms of the spectral features available.
- *Of informational value.* The classes must be ones that meet the users needs.

A few comments about each of these conditions are in order.

Exhaustive. First, relative to the requirement that the class list be exhaustive, it is a basic engineering reality that relative determinations can be made more accurately than can absolute ones. Just as one can measure the distance between two objects more precisely than one can measure the absolute location of the objects, or one can measure the time between two events more precisely than one can measure the absolute time of day of the two events, one has a better chance of assigning a pixel to the correct class if one can consider all possible classes, selecting the best alternative, than simply trying to identify the class of the pixel without taking into account the other possibilities. Thus, one speaks of the classifier being a *relative classifier* rather than an *absolute classifier.* To have a relative classification scheme, one must have an exhaustive list of possibilities, where here exhaustive implies a list of all classes that occur in the specific data set to be analyzed.

Separable. Clearly, one must have a list of classes that are separable to an adequate degree, for this is the whole point of the process, the division of the data set into classes of interest. Thus in the analysis procedure, one must find the most optimal procedure one can to discriminate successfully between the classes. This statement has implications not only on the algorithms used in the analysis, but on the way the classes are defined and training samples drawn in the first place. More will be said about this point as procedures for practical analysis are discussed.

Of Informational Value. This is the point at which the user's requirement is expressed for what output from the analysis is desired. For any given multispectral data set, there are many different types of information that might be desired. For example, over an area with an incomplete canopy of vegetation, one might want to derive a soil map. On the other hand, one might wish to ignore the soil variations as background variation and attempt to obtain a vegetation species map. Various

other possibilities might exist. It is thus in the definition of classes that the user's specific interests in the analysis result are expressed.

These three conditions on the list of classes must be met simultaneously. Note that the exhaustive condition and separability are properties of the data set, while the user imposes the informational value condition. It is the bringing together of these circumstances, those imposed by the data with those imposed by the user's desires that is the challenge to the analyst. It is further noted that the classes are defined by the training samples selected. That is to say that the definition of classes is a quantitative and objective one, not a semantic one. One has not really defined a class one might wish to call "forest" until one has labeled the training samples to be associated with that class name, thus documenting quantitatively what is meant (and what is not meant) by the word "forest."

There is one additional aspect of class definition to be mentioned. An equivalent statement to the conditions for class definition above is that a well trained classifier must have successfully modeled the distribution of the entire data set, but it must be done in such a way that the different classes of interest to the user are as distinct from one another as possible. What is desired in mathematical terms is to have the density function of the entire data set modeled as a mixture of class densities, i.e.,

$$p(\mathbf{x}|\theta) = \sum_{i=1}^{M} \alpha_i p_i(\mathbf{x}|\Phi_i)$$

where $\mathbf{x}$ is a 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 which is the probability of class i, and M is the number of classes. The parameter sets θ and Φ_i are to be discussed next in the context of what limitations are appropriate to the form of these densities.

The training process consists of defining a list of classes, labeling training samples for each class, so as to satisfy the three conditions. Generally at that point, the analyst is focused on the training samples and how exhaustive, separable, and of informational value they are, without any real way of knowing if they are representative of the whole data set. That is to say that the analyst is focused on the right side of the equation above, but there is not much in the process that insures that the right side will indeed equal the left side. A pattern recognition technician would express this in terms of wanting to know if the training will generalize well to samples of a given class not used in training.

This question of generalization can be dealt with by carrying out an iterative calculation based upon both the training samples and a systematic sampling of all the pixels in the scene which will adjust or "enhance" the statistics so that, while still being defined by the training samples, the collection of class conditional statistics better fit the entire data set[2]. This amounts to a hybrid, supervised/unsupervised training scheme.

This process has several possible benefits.

(1) The process tends to make the training set more robust, providing an improved fit to the entire data set, thus providing improved generalization to data other than the training samples.

---

[2] 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.

(2) The process tends to mitigate the Hughes phenomena. Enhancing the statistics by such a scheme in effect, tends to increase the size of the training set and thus tends to move the peak accuracy vs. number of features to a higher value at a higher dimensionality, thus allowing one to obtain greater accuracy with a limited training set.

(3) An estimate is obtained for the prior probabilities of the classes, the α's in the equation above, as a result of the use of the unlabeled samples, something that cannot be done with the training samples alone. In some cases, where only how much of an area contains a given class is desired, and not a map of where it occurs, this could be the final desired result.

To carry out the process, for each class S_j, assume there are N_j training samples available. Denote these samples by z_{jk} where j=1,...,J indicates the class of origin and k=1,...,N_j is the index of each particular sample. The training samples are assumed to come from a particular class without any reference to the exact component within that class. In addition to the training samples, assume N unlabeled samples, denoted by x_k, k=1,...,N, are also available from the mixture.

The process to be followed is referred to as the EM (expectation maximization) algorithm. The procedure is to maximize the log likelihood to obtain maximum likelihood estimates of the parameters involved. The log likelihood expression to be maximized can be written in the following form.

$$L(\theta) = \sum_{k=1}^{N} \log p(x_k|\theta) + \sum_{j=1}^{J} \sum_{k=1}^{N_j} \log \left(\frac{1}{\sum_{t \in S_j} \alpha_t} \sum_{l \in S_j} \alpha_l p_l(z_{jk}|\phi_l) \right)$$

The first term in this function is the likelihood of the unlabeled samples with respect to the mixture density. The second term indicates the likelihood of the training samples with respect to their corresponding classes of origin. The EM equations for obtaining the ML estimates are the following:

$$\alpha_i^+ = \frac{\sum_{k=1}^{N} P^c(i|x_k) + \sum_{k=1}^{N_j} P_j^c(i|z_{jk})}{N\left(1 + \frac{N_j}{\sum_{r \in S_j} \sum_{k=1}^{N} P^c(r|x_k)}\right)}$$

$$\mu_i^+ = \frac{\sum_{k=1}^{N} P^c(i|x_k)x_k + \sum_{k=1}^{N_j} P_j^c(i|z_{jk})z_{jk}}{\sum_{k=1}^{N} P^c(i|x_k) + \sum_{k=1}^{N_j} P_j^c(i|z_{jk})}$$

$$\Sigma_i^+ = \frac{\sum_{k=1}^{N} P^c(i|x_k)(x_k - \mu_i^+)(x_k - \mu_i^+)^T + \sum_{k=1}^{N_j} P_j^c(i|z_{jk})(z_{jk} - \mu_i^+)(z_{jk} - \mu_i^+)^T}{\sum_{k=1}^{N} P^c(i|x_k) + \sum_{k=1}^{N_j} P_j^c(i|z_{jk})}$$

The equations are applied iteratively with respect to the training and unlabeled samples where "c" and "+" refer to the current and next values of the respective parameters, $i \in S_j$, and $P^c(.|.)$ and $P^c_j(.|.)$ are the current values of the posterior probabilities:

$$P^c(i|x_k) = \frac{\alpha_i^c f_i(x_k|\mu_i^c, \Sigma_i^c)}{f(x_k|\theta^c)} \qquad P_j^c(i|z_{jk}) = \frac{\alpha_i^c f_i(z_{jk}|\mu_i^c, \Sigma_i^c)}{\sum_{i \in j} \alpha_i^c f_i(z_{jk}|\mu_i^c, \Sigma_i^c)}$$

Thus as the iteration proceeds, successively revised values for the mean, covariance, and weighting coefficient of each component of each class are arrived at which steadily approach the values for a maximum of the expected likelihood value for the mixture density.

Characteristics Of Higher Dimensional Space.

Multispectral data and analysis methods have been under study for more than three decades, and yet they have not found wide use. One must ask why, what holds back this technology, which obviously has such wide applicability. There are several parts to the answer of this question.

One clearly is the availability of data. So far, satellite systems have been limited to two or three at a time. Cloud cover significantly impacts the optical part of the spectrum. Given this low number of sensors on orbit and orbits providing a repeat cycle of coverage over any given spot of typically once every 15 days or so, a user cannot realistically expect to have data collected over a given site even within a few days of when it is desired. Data availability has not been available on demand, but rather its availability for a given use is more of a chance occurrence. This situation is likely to remain until an adequate fleet of sensors is on orbit and operating so as to insure that data availability is being driven by user demand rather than other factors such as orbital mechanics or atmospheric condition.

A second factor limiting the use of this technology is the cost of the data. Land remote sensing is seen in a different light than atmospheric remote sensing for some reason. Rather than operational land remote sensing data sources being seen as a government function as it is for weather satellite data, it is apparently seen as a private sector function. Data from experimental systems such as Landsat has been quite expensive, placing it largely out of reach of the broad spectrum of research and application uses. This has placed the technology in a catch-22 situation. The price is high because the volume is low, and the volume is low, at least in part, because the price is high. Again, until the volume builds to a reasonable level, data is likely to be too expensive for most uses.

However, it is the third reason for the limited use to this time that is to be addressed at greater length here, because it is a technologically-based limitation rather than a government policy one, and because advancing technology is in the process of removing this third limitation. The limitation in mind is that due to the small number of spectral bands that have been available. The research in the 1960's that led to the Landsat system was done using an aircraft system that had from 12 to 18

spectral bands of 8 bit data. However, this degree of spectral detail was beyond what was technically feasible for Landsat 1, and a four band, 6 bit system resulted. When in 1975 it was time to devise a second generation sensor, the bar was raised to initially 6 and finally a 7 band, 8 bit system, certainly an improvement over four, but still a significant limitation. This limited measurement complexity placed many applications of the technology in a borderline area or simply out of reach. It has not been until recent years that more complex data, often under the label of hyperspectral data, began to be studied and planned. Such an advance greatly broadens the problems that can be realistically addressed, but it also complicates the matter of how to analyze this more complex data. It is this latter point that will be addressed next. The initial focus is on the nature of high dimensional data and how it differs from more conventional data.

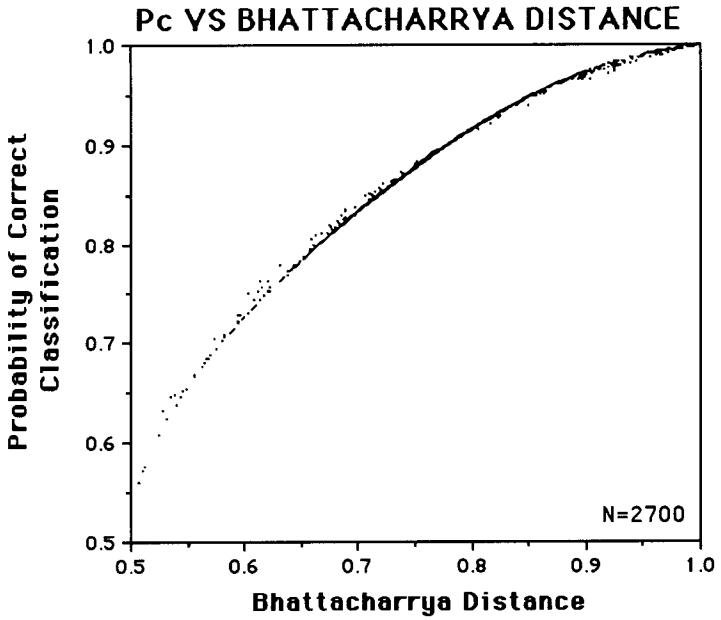
Much of the material in previous sections has been offered in the context of a feature space in familiar two or three dimensional geometry. However, hyperspectral data has many more than three bands, and thus the feature space of interest has much higher dimensionality. One must inquire as to whether one's ordinary intuitive perception developed from three-dimensional geometry still apply in higher dimensional space. The answer is that, in general, it does not, and this fact substantially influences what is appropriate in the analysis process.

As an example of this, consider the case of predicting the accuracy of a classification from the training samples of the classes defined. One of the most common ways of accomplishing such a prediction is by use of a statistical distance measure. Such a distance measure is Bhattacharyya distance. For the 2-class case of Gaussian data, the definition of Bhattacharyya distance is,

$$B = \frac{1}{8}\,[\mu_1-\mu_2]^T\left[\frac{\Sigma_1+\Sigma_2}{2}\right]^{-1}[\mu_1-\mu_2] + \frac{1}{2}\,\mathrm{Ln}\,\frac{\left|\frac{1}{2}[\Sigma_1+\Sigma_2]\right|}{\sqrt{|\Sigma_1||\Sigma_2|}}$$

Where μ_i and Σ_i are, respectively, the mean vector and the covariance matrix of class i.

In the multispectral remote sensing context, Bhattacharyya distance has shown itself to be a good predictor of classification accuracy. Though there is not a closed form, one-to-one relationship between Bhattacharyya distance and classification accuracy, the following graph shows the result of a Monte Carlo test of this relationship for the two-class, two-feature case.



It is seen that in this case, the relationship is nearly one-to-one, and nearly linear.

Examining the equation defining it, one sees that of the two terms on the right, the first measures the separation of the classes due to the difference in class means. The second does not depend at all on the difference in means, but measures the portion of the separation due to the difference in covariance matrices. In low dimensional space, where geometric visualization is possible, the mean vector defines the location of a distribution in the feature space while the covariance matrix provides information about its shape. For example, a covariance matrix with significantly sized off-diagonal components indicating significant correlation between bands would tend geometrically to be long and narrow, while a covariance matrix with only small off-diagonal components, and thus low correlation between bands, would tend to be more circular in shape. Now, one implication of this is that two classes may lie precisely on top of one another, in the sense of having exactly the same mean values, and yet they may be separable. Indeed, if the dimensionality is high enough, they may be quite separable. It turns out that this is especially true as the dimensionality is increased. Why might this be so?

What is needed is a more in-depth understanding of such unintuitive characteristics of high dimensional feature spaces. We will review a selection of these unusual or unexpected characteristics[3], because they point the way to some practical procedures for data analysis that might not be otherwise apparent.

A. As dimensionality increases the volume of a hypercube concentrates in the corners.

The volume of the hypersphere of radius r and dimension d is known to be given by the equation:

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

[3] 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*, Volume 28 Part C Number 1, pp. 39-54, Feb. 1998.

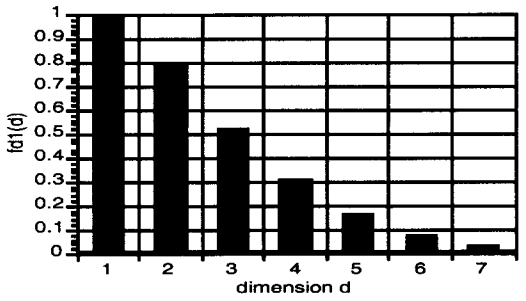
The volume of a hypercube in the interval [-r, r] and of dimension d is given by the equation:

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

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

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

where d is the number of dimensions. This ratio is plotted as a function of d in the following figure. It shows how f_{d1} decreases as the dimensionality increases.



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

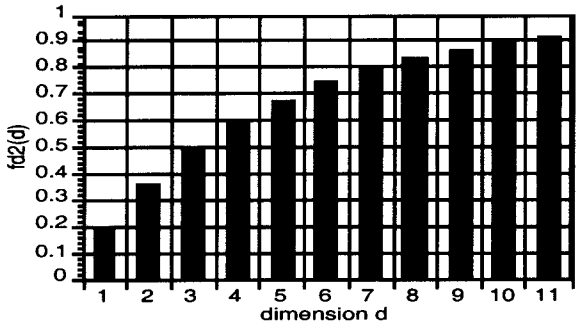
Note that $\lim_{d\to\infty} f_{d1} = 0$, which implies that the volume of the hypercube is increasingly concentrated in the corners outside of the hypersphere as d increases.

B. *As dimensionality increases the volume of a hypersphere concentrates in an outside shell.*

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

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

The following figure shows, for the case $\varepsilon = r/5$, how as the dimension increases the volume concentrates in the outside shell.



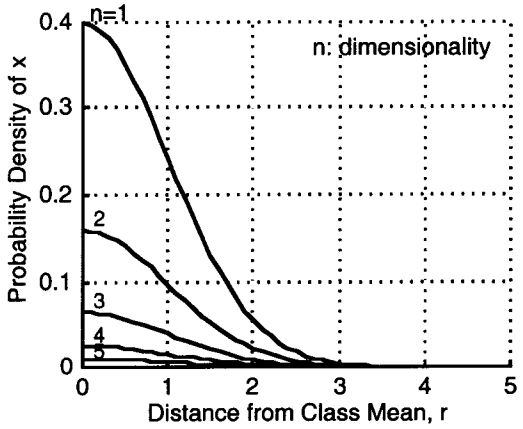
Volume of a hypersphere contained in the outside shell as a function of dimensionality for $\varepsilon = r/5$.

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

These characteristics have two important consequences that bear upon practical methods for data analysis. First, higher dimensional space is mostly empty, which implies that the multivariate data in any given case is usually in a lower dimensional structure. This implies that a high dimensional data set 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. A quantitative demonstration of these characteristics is given in [4,5].

The tendency for Gaussian data to concentrate in the tails seems like a paradox, since it is clear from the Gaussian density function that the "most likely" values are near the mean, not in the tails? This paradox can be explained as follows[6]. First note what happens to the magnitude of a zero mean Gaussian density function as the dimensionality increases. This is shown in the following graph.



It is seen that, while the shape of the curve remains bell-shaped, its magnitude becomes smaller with increasing dimensionality, as it must, since the overall volume must remain one, and, of course, it decreases exponentially as r increases.

---

4 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*, Volume 28 Part C Number 1, pp. 39-54, Feb. 1998.

5 Luis O. Jimenez and David Landgrebe, "High Dimensional Feature Reduction Via Projection Pursuit," PhD thesis and School of Electrical & Computer Engineering Technical Report TR-ECE 96-5, April 1996.

6 This explanation was provided by graduate student Pi-fuei Hsieh.

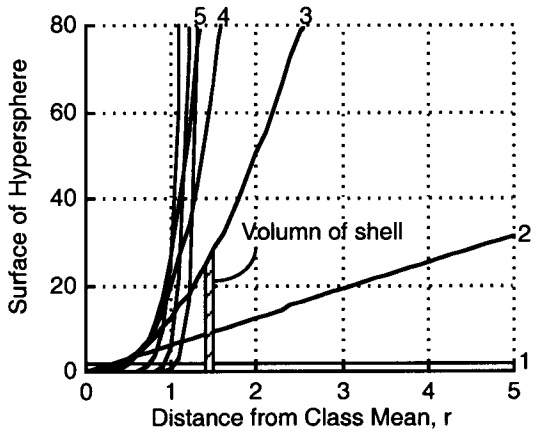
Next, consider how the volume density in the space changes as dimensionality increases. The volume of a hypersphere of radius r as a function of dimensionality was given above as,

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

Therefore, the volume in a differential shell as a function of radius r is

$$\frac{dV}{dr} = \frac{2\pi^{d/2}}{\Gamma(d/2)} r^{(d-1)}$$

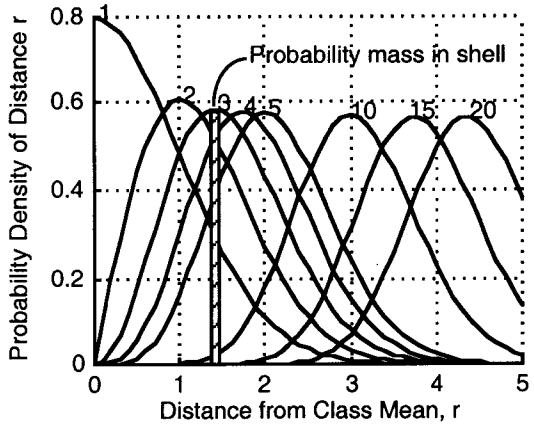
A plot of this for several values of d is given in the following graph.



Thus the volume available in a differential shell at radius r increases very rapidly with r as d becomes larger. Then the probability mass as a function of radius r, the combination of these two, may be shown to be,

$$f_r(r) = \frac{r^{d-1}e^{-\frac{r^2}{2}}}{2^{\frac{d}{2}-1}\Gamma\left(\frac{d}{2}\right)}$$

This function is plotted in the following graph. It may be shown that the peak of this function occurs at $\sqrt{d-1}$.



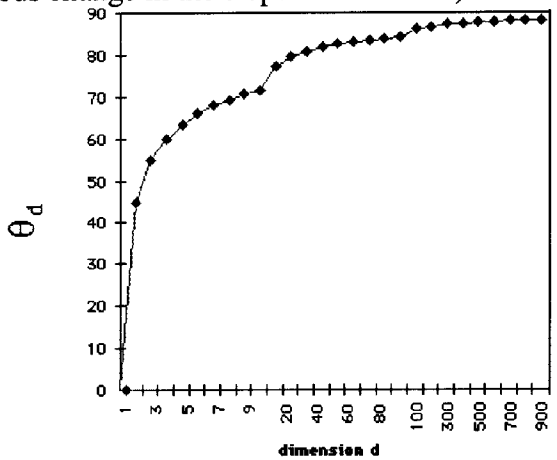
Because the volume of a differential shell increases much more rapidly with r than the density function decreases, the net effect is as shown in the graph. Thus, it is seen that the peak of the probability mass moves away from the mean as the dimensionality increases, indicating that "most of the data becomes concentrated in the tails of the density" even though it is Gaussianly distributed.

C. *As the dimensionality increases the diagonals are nearly orthogonal to all coordinate axis .*

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

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

The following figure illustrates how the angle between the diagonal and the coordinates, θ_d, approaches 90^o with increases in dimensionality. (Note the nonuniform scale on the x axis accounts for the discontinuous change in the slope of the curve.)



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

Note that $\lim_{d\to\infty} \cos(\theta_\delta) = 0$, 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.

D. *For most high dimensional data sets, lower dimensional 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 that as the dimensionality tends to infinity, lower dimensional linear projections will approach a normality model with probability approaching one. Normality in this case implies a normal or a combination of normal distributions. This property increases the viability and justification for a Gaussian class model when the data have been projected to a lower dimensional space.

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

There is also a relationship between dimensionality and the number of training samples on the one hand and classifier complexity on the other. Fukunaga[7] 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[8]. 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.

As previously noted, in remote sensing, the number of training samples available by which the user defines the classes of interest is almost always limited. This limitation grows in importance as the number of features available is increased. Thus, though a larger number of features makes possible more accurate and detailed classifications, the price is that the need for greater precision in the estimation of class statistics grows as well. As seen in the above, when data is gathered in a large number of bands, the information-bearing structure is nearly always present in a subspace of the feature space. This suggests the value of algorithms that can determine which subspace contains that structure for the problem at hand, thus showing a way to reduce the dimensionality without significant loss of information. Such "feature extraction" algorithms are the subject of the next section below.

However, there is another way that this effect due to limited training sets can be mitigated. It is found that when, due to the Hughes effect described above, the accuracy is below optimality due to limited training, a less complex classifier algorithm may provide increased classification accuracy. The classification rule that results from using the class conditional maximum likelihood estimates for the mean and covariance in the discriminant function as if they were the true mean and covariance achieves optimal classification accuracy only asymptotically as the number of training samples increases toward infinity. This classification scheme is not optimal when the training sample is finite[9]. When the training set is small, the sample estimate of the covariance is usually highly elliptical and can vary drastically from the true covariance. In fact, for p features, when the number of training samples is less than p+1, the sample covariance is always singular.

For limited training data, the common covariance estimate, obtained by assuming all classes have the same covariance matrix, can lead to higher accuracy than the class conditional sample estimate,

---

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

8 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.

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

even when the true covariance matrices are quite different[10]. This leads to several possible assumptions that could turn out to be advantageous. To illustrate the possibilities,[11] suppose one has a 5 class problem and 6 dimensional data. The various possibilities for coefficients that must be estimated are illustrated by the following diagrams of covariance matrices.

```
Class 1        Class 2        Class 3        Class 4        Class 5
b              b              b              b              b
a b            a b            a b            a b            a b
a a b          a a b          a a b          a a b          a a b
a a a b        a a a b        a a a b        a a a b        a a a b
a a a a b      a a a a b      a a a a b      a a a a b      a a a a b
a a a a a b    a a a a a b    a a a a a b    a a a a a b    a a a a a b

                              Common Covar.
                              d
                              c d
                              c c d
                              c c c d
                              c c c c d
                              c c c c c d
```

Now for limited numbers of training samples, the greater the numbers of coefficients one must estimate, the lower the accuracy of the estimates. If one were to attempt a normal estimate of individual class covariances, then one must estimate coefficients in positions marked a and b above. If, on the other hand, it appeared advantageous to ignore the correlation between channels, then one would only need to estimate coefficients marked b. If one was willing to assume that all classes have the same covariance, then one would only need to estimate the smaller number of coefficients marked c and d above. And finally, if in addition, one was willing to ignore the correlation between channels of this common covariance, one would only need to estimate the coefficients marked d above.

For p dimensional data, the number of coefficients in a class covariance function is (p+1)p/2. The following table illustrates the number of coefficients in the various covariance matrix forms which must be estimated for the case of 5 classes and several different numbers of features, p.

No. of Features p	Class Covar. (a & b above) 5{ p+1)p/2}	Diagonal Class Common Covar. (b above) 5p	Common Covar. (c & d above) { p+1)p/2	Diagonal Common Covar. (d above) 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

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

[11] Joseph P. Hoffbeck, "Classification of High Dimensional Multispectral Data," PhD Thesis, Purdue School of Electrical and Computer Engineering, May 1995, TR-EE 95-14. See also, Hoffbeck, Joseph P. and David A. Landgrebe, "Covariance MatriavailableEstimation and Classification with Limited Training Data," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 18, no. 7, pp. 763-767, July 1996.

Thus, if for example, one has 100 training samples for each of 5 classes and 20 features, then, for the individual class covariance case, one might expect an accuracy problem attempting to estimate the 1050 coefficients with only 500 total training pixels. It is useful in any given case, then, to determine whether the sample estimate or the common covariance estimate would be more appropriate in a given situation. This illustrates the manner in which a properly chosen estimator could improve classifier performance.

An estimator referred to as LOOC, found useful for this situation[12], examines the sample covariance and the common covariance estimates, as well as their diagonal forms, to determine which would be most appropriate. Furthermore, it examines the following pair-wise mixtures of estimators:

- sample covariance-diagonal sample covariance,
- sample covariance-common covariance, and
- common covariance-diagonal common covariance.

The proposed estimator has the following form:

$$\mathbf{C}_i(\alpha_i) = \begin{cases} (1-\alpha_i)\mathrm{diag}(\Sigma_i) + \alpha_i\Sigma_i & 0 \le \alpha_i \le 1 \\ (2-\alpha_i)\Sigma_i + (\alpha_i - 1)\mathbf{S} & 1 \le \alpha_i \le 2 \\ (3-\alpha_i)\mathbf{S} + (\alpha_i - 2)\mathrm{diag}(\mathbf{S}) & 2 \le \alpha_i \le 3 \end{cases}$$

where $\mathbf{S} = \frac{1}{L}\sum_{i=1}^{L}\Sigma_i$ is the common covariance matrix, computed by averaging the class covariances across all classes. This, in effect, assumes all classes have the same covariance matrix. The variable α_i is a mixing parameter that determines which mixture 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. Utilization of this procedure to determine the best set of assumptions to use in the face of limited training sets can improve the accuracy, which is suspected of being suboptimal due to the Hughes effect. Further, though the estimated class covariance matrices become singular when the number of training samples available falls below one more than the number of features, use of LOOC allows for cases where the number of training samples becomes as small as three before the $C_i(\alpha_i)$ becomes singular, thus allowing for some use of second order variations to be made for even smaller training sets.

Some Feature Extraction Schemes.

As has been seen in the above,

- As the dimensionality of data is increased, its greater ability to permit discrimination between detailed classes is compromised by the limitation on the number of training samples typically available, leading to less precise quantitative descriptions of the classes of interest.

[12] 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.

- Furthermore, high dimensional feature spaces are found to be largely empty with the significant information-bearing structure existing in a lower dimensional space. The appropriate subspace is case-dependent.

- It has also been found that distributions in data transformed to a subspace have a greater tendency to be Gaussian. A stronger justification for the Gaussian model mitigates the need for more complex models such as nonparametric ones, and thus the more difficult estimation problems they present.

All of these point to the value of finding and applying algorithms that can locate case-specific optimal subspaces for discriminating between a given set of classes thereby reducing the dimensionality without loss of information, thus improving classifier performance. Such algorithms are referred to as feature extraction algorithms. A number of such algorithms are found in the literature. Two specifically suited to the remote sensing context will be described here in order to illustrate salient features of such algorithms.

1. The basis for *Discriminant Analysis Feature Extraction* is the maximization of the ratio,

$$\frac{\sigma_B^2}{\sigma_W^2} = \frac{\text{Between classes variance}}{\text{Average within class variance}}$$

where, for the two class case, σ_W^2 is the average of σ_{W1}^2 and σ_{W2}^2. In matrix form the within-class scatter matrix Σ_W and the between-class scatter matrix Σ_B may be defined as[13],

$$\Sigma_W = \sum_i P(\omega_i)\Sigma_i \qquad \text{(within class scatter matrix)}$$

$$\Sigma_B = \sum_i P(\omega_i)(M_i - M_o)(M_i - M_o) \qquad \text{(between class scatter matrix)}$$

$$M_o = \sum_i P(\omega_i)M_i$$

Here $\mathbf{M}_i$, Σ_i , and $P(\omega_i)$ are the mean vector, the covariance matrix, and the prior probability of class ω_i, respectively. The criterion for optimization may defined as,

$$J_1 = \mathrm{tr}(\Sigma_W^{-1} \Sigma_B)$$

New feature vectors are selected to maximize the criterion.

The basic concept, then, is quite simple. The greater the variance of the distance between classes, normalized by the average within class variation, the better the feature subspace. The calculation required utilizes the training statistics directly and is an eigenvalue type of calculation, forming new features that are linear combinations of the original bands, and rank ordering them from the most to the least valuable is discrimination capability. It does have two significant limitations, however.

First, it becomes ineffective if the classes involved have very little difference in mean values. Recall that classes, especially high dimensional ones, can be quite separable based on their second order statistics, i.e., their covariance matrices, alone. From the above ratio, it is apparent that classes with small difference in their means, but substantial separation due to their covariances, would not, by this means, lead to effective subspaces. A second limitation is that the method is only guaranteed to provide reliable features up to one less than the number of classes. If one has a problem that does not have a large number of classes, one will not have a very large subspace to work from. Still, the calculation involved is quite fast and very useful subspaces can be found quickly in many cases.

[13] K. Fukunaga, *Introduction to Statistical Pattern Recognition,* Academic Press, 1972.

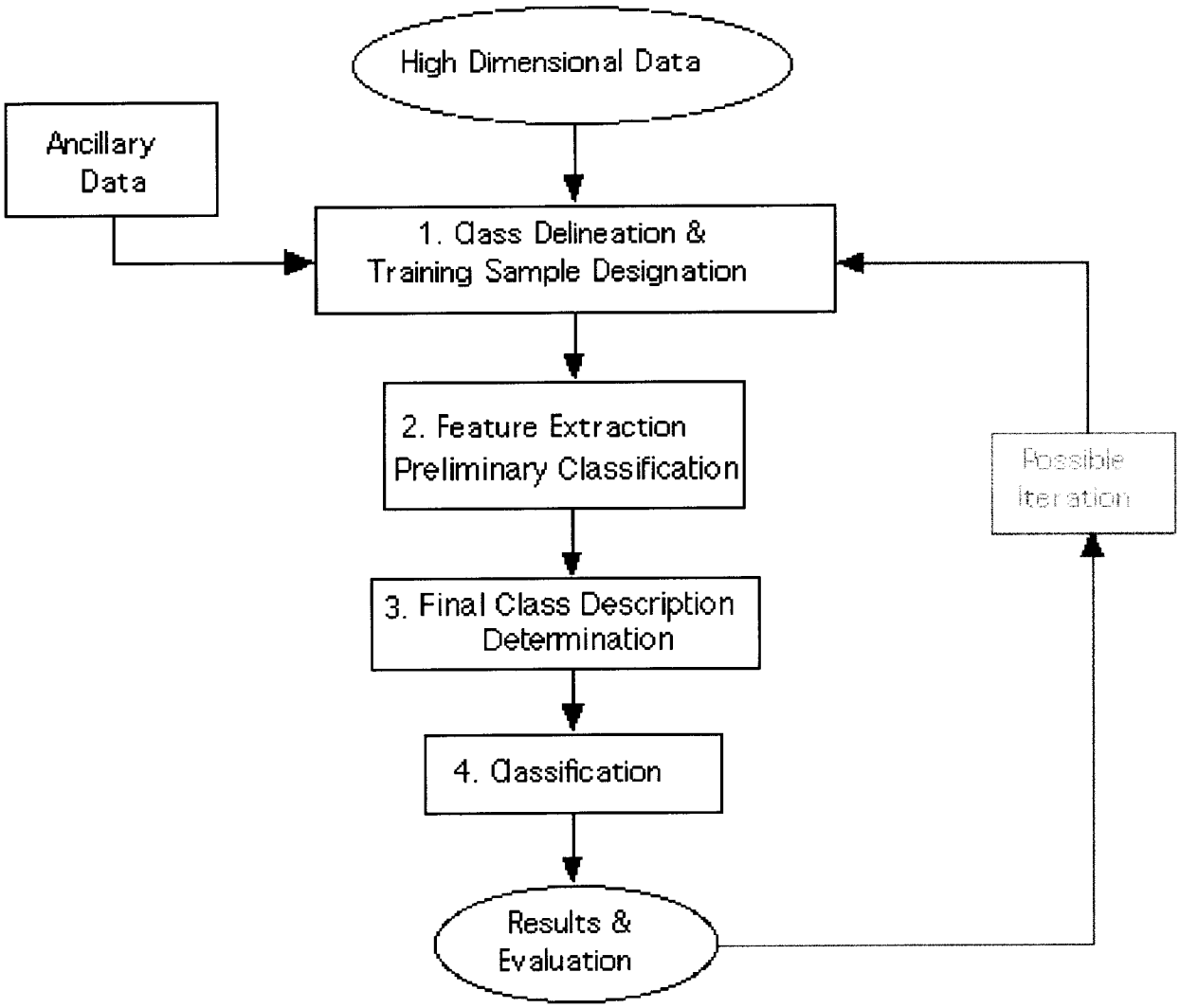
2. A second feature extraction algorithm, called *Decision Boundary Feature Extraction,*[14] does not have the limitations just cited. It utilizes the training samples directly, rather than statistics derived from them, to locate the decision boundary between the classes using a definition for discriminately informative and discriminately redundant features. Then, using the effective portion of that decision boundary, an intrinsic dimensionality for the problem is determined and a transformation defined which enables the calculation of the optimal features. The calculation produces not only the desired new features, as linear combinations of the original ones, but it provides eigenvalues that are a direct indication of how valuable each new feature will be. Since it works directly from the located decision boundary, it does not have the limitation of discriminant analysis feature extraction regarding class mean differences. However, it is quite a lengthy calculation, especially if the training set has many training samples. On the other hand, because it works directly with the training samples rather than statistics derived from them, it tends to be ineffective when the training set is small.

The value of high dimensional data on the one hand, but with the practical limitation of small training sets on the other, means that compromises must be made in devising and using feature extraction algorithms, as in all other parts of the analysis process. Because of the quite broad variation in the circumstances of data and user requirements, there is no single scheme that will be optimal in all cases. The intent in describing the strengths and weaknesses of the above two feature extraction algorithms, is to illustrate that the analyst must be able to knowledgeable select the best tool for the circumstances.

Procedures for Information Extraction Problems.

It is appropriate at this point to coalesce the concepts described above into an effective procedure for analyzing a data set. Though the specific steps needed varies with the scene, the data set and the classes desired, the following diagram provides a general outline.

[14] 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.



Nominal Sequence of Steps for the Analysis of A Hyperspectral Data Set

A brief description of each of the above steps is as follows.

1. *Class Delineation.* The analysis process begins with a general overview of the data set to be analyzed, often by viewing a 3-color simulated color infrared presentation of the data in image space. The intent is to create a list of classes which is suitably exhaustive, and which includes the classes of user interest. To the extent possible at this point and from such an image presentation, consideration should be given to picking classes in such a way as to provide for a set that are separable from a spectral standpoint.

   Training Sample Designation. Following, or as a part of listing the desired classes, the spectral description of the classes must be designated. How this is done varies widely with the particular data set and the information about the scene that is available to the analyst. Examples of some of the ways this might come about are as follows.

- Observations taken from a portion of the ground scene taken from the ground at the time of the data collection. See for example, [15] where this was done for a region-sized problem over an entire growing season to track a particular disease in a vegetative species.
- Observations from aerial photographs from which examples of each class can be labeled. See for example, [16] where again, this was done for a region-sized problem on a land use mapping problem.
- Conclusions that can be drawn directly from the image space, itself. See the example analysis below, an urban mapping problem where the spatial resolution was great enough to make objects of human interest recognizable in image space.
- Conclusions that can be drawn about individual pixels by observing a spectral space representation of a pixel. The use of "imaging spectroscopy" characteristics, where specific absorption bands of individual molecules are used to identify specific minerals, are an example of this. See [17] for an example.

2. *Feature Extraction and Preliminary Classification.* At this point one can expect to have training sets defined for each class, but they may be small. There would thus be value in eliminating features that are not effective for the particular set of classes at hand, so as to reduce the dimensionality without loss of information. A feature extraction algorithm would be used for this purpose, followed by a preliminary classification. From the preliminary classification, one can determine if the class list is suitably exhaustive, or if there have been classes of land cover of significant size that have been overlooked. One can also determine if the desired classes are adequately separable. If not, the classification can be used to increase the selection of training samples, so that a more precise and detailed set of quantitative class descriptions are determined.

3. *Final Class Description Determination.* With the now augmented training set, in terms of either additional classes having been defined or more samples labeled for the classes or both, any of several steps may be taken to achieve the final class descriptions in terms of class statistics.
   - It may be appropriate to re-apply a feature extraction algorithm, given the improved class descriptions. In this way, a more optimal subspace may be found.
   - The Statistics Enhancement algorithm may be applied. This algorithm is known to be sensitive to outliers, and thus would not be expected to perform well until it is known that the list of classes is indeed exhaustive, as classes not previously identified would function as outliers to the defined classes. The intended result of applying this algorithm at this point is to increase the accuracy performance of the following classification and to improve the generalization capabilities of the classifier from the training areas to the rest of the data set.
   - If the training set is still smaller than desirable relative to the number of features needed to achieve satisfactory performance, it might be appropriate to use the LOOC estimation scheme, which can function down to as few as three samples for a class.

---

15 MacDonald, R. B., M. E. Bauer, R. D. Allen, J. W. Clifton, J. D. Ericson, and D. A. Landgrebe, 1972, "Results of the 1971 Corn Blight Watch Experiments," Proceedings of the Eighth International Symposium on Remote Sensing of Environment, Vol. I. Environmental Research Institute of Michigan, Ann Arbor, Michigan, pp. 157-190.

16 Swain, P. H. and S. M. Davis, *Remote Sensing: The Quantitative Approach,* McGraw-Hill, 1978, pp. 309-314.

17 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.

4. Classification. A final classification of adequate quality should be possible at this point, and evaluation of it can proceed. However, if this is not the case, depending on the nature of the further improvement needed, a return to any of the above steps may be used.

Hyperspectral Analysis Example of Urban Data

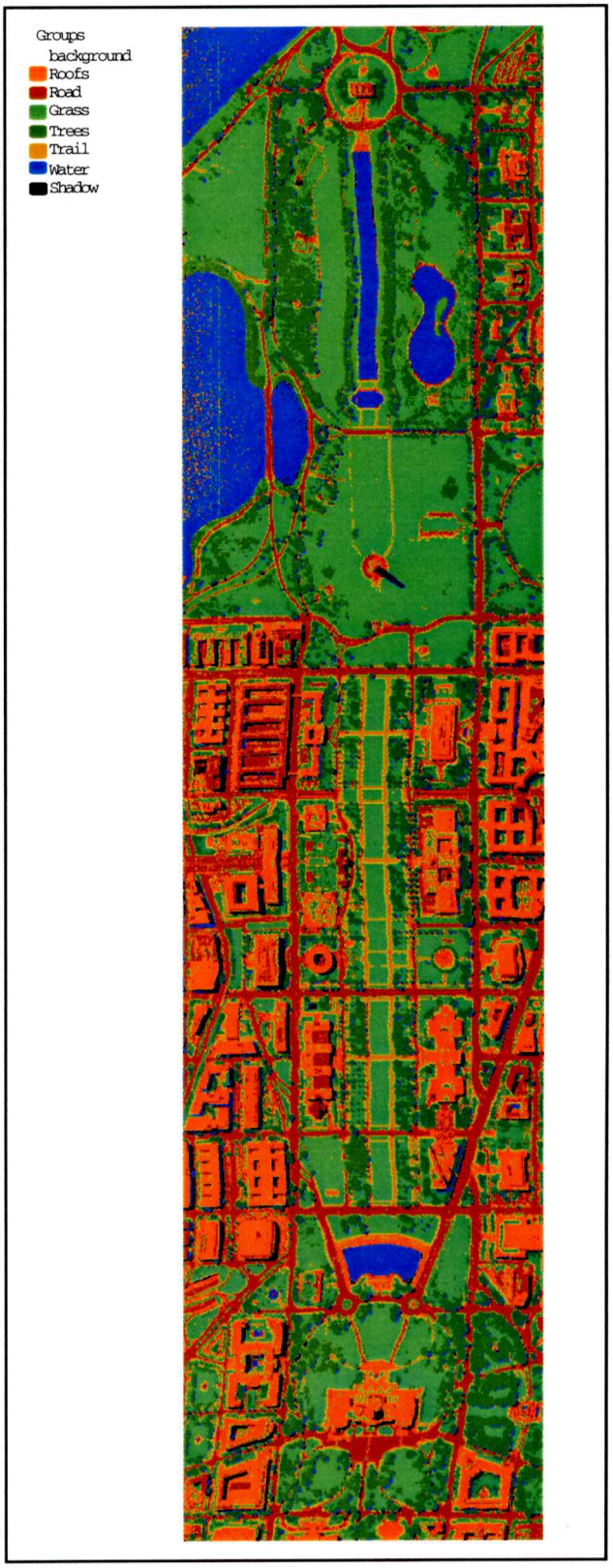
The following is offered to illustrate this procedure. The figure to the left shows a color IR presentation of an airborne hyperspectral data flightline over the Washington DC Mall. The sensor system used in this case is known as HYDICE. It collects data in 210 bands in the 0.4 to 2.4 μm region of the visible and infrared spectrum. This data set contains 1208 scan lines with 307 pixels in each scan line. The data set totals approximately 150 Megabytes. With data of this complexity, one might expect a rather complex analysis process. However, following the scheme above, it is possible to achieve a quite simple and inexpensive analysis. The steps used and the time for this analysis needed on a personal computer costing less than $3000 are listed in the following table and are briefly describe below.

Operation	CPU Time (sec.)	Analyst Time
Display Image	18	
Define Classes		< 20 min.
Feature Extraction	12	
Reformat	67	
Initial Classification	34	
Inspect and Add 2 Training Fields		≈ 5 min.
Final Classification	33	
Total	164 sec = 2.7 min.	≈ 25 min.

Define Classes. A software application program called MultiSpec, available to anyone at no cost from

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

is used. The first step is to present to the analyst a view of the data set in image form so that training samples, examples of each class desired in the final thematic map, can be marked. A simulated color infrared photograph form is convenient for this purpose; to do so, bands 60, 27, and 17 are used in MultiSpec for the red, green, and blue colors, respectively. The result is shown at left. The classes desired in this case are, from the user's standpoint, quite simple. They are rooftops, streets, grass, and trees. Inspection of the image space



presentation shows that additional classes of trails (graveled paths), water, and shadows will be needed. From a spectral class definition standpoint, however, the classes are not so simple. It is apparent from the image that there are many different materials used in the roofs, and they are in quite varying conditions and ages, resulting in a wide variety of spectral characteristics. This must be dealt with by defining several spectral subclasses for the class roofs. Further, some of the rooftops are constructed from materials, e.g., tar and gravel, which are quite similar to that used in streets and parking lots, thus making discrimination between rooftops and streets more challenging. As a result, care must be exercised in the following steps.

Feature Extraction. After designating an initial set of training areas, a feature extraction algorithm is applied to determine a feature subspace that is optimal for discriminating between the specific classes defined. Discriminant Analysis Feature Extraction (DAFE) is selected for this purpose, as a result of it being fast and not overly sensitive to small training sets. Also, recall that its weakness, not functioning well with regard to classes that have a small difference in mean values does not appear to be a limiting condition at this point in the analysis. The result of the DAFE calculation is a linear combination of the original 210 bands to form 210 new features that automatically occur in descending order of their value for producing an effective discrimination. From the MultiSpec output, it is seen that the first nine of these new features will be adequate for successfully discriminating between the classes.

Reformatting. The new features defined above are used to create a 9 band data set consisting of the first nine of the new features, thus reducing the dimensionality of the data set from 210 to 9.

Initial Classification. Having defined the classes and the features, next an initial classification is carried out. An algorithm in MultiSpec called ECHO[18,19] (Extraction and Classification of Homogeneous Objects) is used. This algorithm is a maximum likelihood classifier

[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 Geoscience Electronics*, 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.

that first segments the scene into spectrally homogeneous objects. It then classifies the objects utilizing both first and second order statistics, thus taking advantage of spatial characteristics of the scene, and doing so in a multivariate sense.

Finalize Training. An inspection of the initial classification result indicates that some improvement in the set of classes is called for. To do so, two additional training fields were selected and added to the training set.

Final Classification. The data were again classified using ECHO and the new training set. The result is shown above. An examination of this result shows that it provides a reasonably accurate thematic presentation of the scene. If further refinement of the result would be desired, it might be desirable to increase the size of the training sets by choosing additional pixels representative of each of the classes. It might also be desirable to apply the Statistics Enhancement scheme. An additional possibility might be to use the Decision Boundary Feature Extraction (DBFE) algorithm to the original data using the augmented training sets. Recall that DBFE does not have the limitation of DAFE with regard to classes that have nearly equal mean vectors, but it does require larger numbers of training samples to deliver good performance.

Concluding Remarks

Advances made in sensor system technology in recent years have effectively removed one of the most significant barriers to improved performance in multispectral remote sensing systems, namely the limited spectral dimensionality of these systems previously. The much increased dimensionality makes possible substantially more accurate and more detailed discriminations. However, the price to be paid for this improvement is the need to be able to more precisely define quantitatively the classes to be discriminated, as the above material makes clear.

There are, of course, many valid ways to approach the analysis of multispectral data, and a great range of user needs, from the mapping of simple classes to a desire to discriminate between classes with only very subtle differences. In exploring the nature of hyperspectral data, we have focused upon methods for the most challenging cases, in order to probe the limits of the potential of such data. To provide a simple explanation of the potential of high dimensional data, consider the case of hyperspectral data of 100 bands, gathered with a signal-to-noise ratio high enough to justify a ten bit data system. In this case, there would be $2^{10} = 1024$ possible values in each band, and a total of $1024^{100} \approx 10^{300}$ possible different discrete locations in the feature space. This is a huge number, so large that even for a data set containing 10^6 pixels, the probability of any two pixels landing in the same feature space cell is vanishingly small. This means that, before even considering the cause/effect relationship between the physical characteristics of pixel areas on the ground and their spectral response or atmospheric or other such effects, one can conclude that, since there is no overlap of pixels in such a feature space, anything is theoretically separable from anything. The problem is that to approach this possibility, one must be able to locate a decision boundary between pixels of different desired classes correctly and precisely. As the dimensionality and therefore the volume available in such feature spaces grows, the estimation precision must grow as well and very rapidly so.

In this chapter, we studied the characteristics of high dimensional spaces quantitatively in order to understand, document, and make credible such potential. The results suggest a fundamental shift in the analysis paradigm from that common today. Much current literature on multispectral analysis suggests studying cause/effect relationships of spectral responses with the intent to using that knowledge directly to achieve classification. Implied in this is the thought that, documenting the spectral response of various materials will allow the analysis of later spectral data sets based upon such documented responses.

From a purely scientific standpoint, such work is certainly valuable, for it increases the understanding of the interaction of electromagnetic energy with scene materials. However, natural scenes are not only very complex but very dynamic as well. Thus the spectral response of a given material is not very stable over time and place. Increasing the dimensionality of such data, while enhancing the potential for discrimination, does not materially change this stability problem. A now common line of approach to solving this dilemma has been to "correct" the data, i.e. to attempt to adjust the data for the various factors that have changed from one observation time and place to another. Such factors include the atmospheric effects, the illumination and view angle effects due to the non-Lambertian characteristics of most scene materials, changes in hemispheric illumination, the adjacency effect, and many more. This has lead to very extended studies of some very daunting problems.

After many years of study of the effects of the atmosphere, for example, about the best that is claimed for atmospheric adjustment is an accuracy of 2 to 5%, a level that does not favorably compare with the 0.1% measurement level implied by 10 bit data common today. Thus it may not be helpful nor perhaps even wise to use such calculations directly in the classification of data. At the very least, it introduces a complex step into the analysis process, one that may have little positive effect on the accuracy achieved in the generation of a thematic map. Further, it is only one of a number of such adjustments that would be necessary to reconcile conditions between data collected at different times or places.

The field of wireless communication faced a similar problem in its development many years ago. In that case, rather than attempting to adjust for or subtract out the many sources of interference (e.g., noise generated in the atmosphere, by other competing signal sources, etc.) and distortion (e.g., multipath, fading, etc.) and the like, the approach was initially to model the corrupting influences, as well as the signals desired, and then to construct optimal detection procedures that discriminate between them. The success of this approach, obviously after much further development, is seen in the clarity of modern digital cellular phone messages through complex urban environments, for example, or in the communication possible from planetary and deep space probes using very low transmitter power.

Spectra collected at another time or place can certainly represent useful information, and there are usually many other pieces of information, some quantitative, some subjective, that are available about a scene when one begins an analysis process. Rather than attempting to use such spectra collected at another time, place, or with another instrument directly in the analysis process, a more viable approach would seem to be to use the knowledge gained from such measurements together with other ancillary information to label examples of classes of interest in the data set to be analyzed. In this way, a suitable collection of spectra, representing the range of conditions and circumstances existing in the current data set, could be constructed that would indeed define the classes of interest to adequate precision.

Using such information to label "training samples," rather than attempting to use it directly in the analysis process, means that the analysis can take place on the original, "uncorrected" data. This has several advantages. First of all, it greatly reduces the amount and complexity of the processing that must be done to the data, since it does not involve further calibration of the data, removing the atmospheric and other effects, conversion from radiance to reflectance, and other such adjustments. If any of these processes need to be carried out to label training samples, it would involve such processing on a much smaller quantity of data. However, perhaps more importantly, use of this approach means that any unsuspected corrupting influences of such processing, which necessarily will always be imperfect, would be avoided. Since the analyst has no way to tell if such processing, which seems on the face of it, so logical, actually helps or hinders, a cleaner, simpler analysis process results. The secret, then, is in the adequacy of the precision and detail by which the user quantitatively specifies the classes of interest, and doing so in the data set to be analyzed

by labeling an adequate number of training samples. It is basically the inverse of the computer user's mantra, "Garbage in, Garbage out," namely, precise and careful specification of what is desired put into the analysis process can lead to precise and accurate output.