Some Fundamentals and Methods for Hyperspectral Image Data Analysis

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

This paper was published in Systems and Technologies for Clinical Diagnostics and Drug Discovery II, G.E. Cohn,J.C. Owicki, Editors, Proc. of SPIE Vol. 3603, and is made available as an electronic reprint with permission of SPIE. Single print or electronic copies for personal use only are allowed. Systematic or multiple reproduction, or distribution to multiple locations through an electronic listserver or other electronic means, or duplication of any material in this paper for a fee or for commercial purposes is prohibited. By choosing to view or print this document, you agree to all the of the copyright law protecting it.

Abstract

Multispectral image data has been a key data type for land observational remote sensing from aircraft and spacecraft since the 1960's[1]. Sensor technology was a primary limiting factor for many years for this method, as sensors such as Landsat could only collect data in four to seven spectral bands at once. In the last few years, advances in sensor technology have made possible the collection of such image data in as many as several hundred spectral bands at once. In this paper, some results obtained in the study of data analysis methods for such high dimensional data will be overviewed. They show that such data have substantially increased potential for deriving more detailed and more accurate information, but to achieve it, the primary limiting factor has become the precision with which a user can specify the analysis classes of interest. Some methods and procedures for mitigating this limitation in practical circumstances will be described.

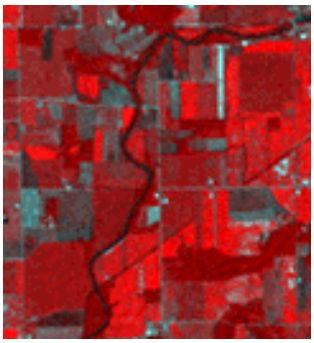
Though multispectral technology has been an important topic to Earth observational remote sensing for some years, it no doubt has substantial applicability in many other disciplines.

Introduction and History

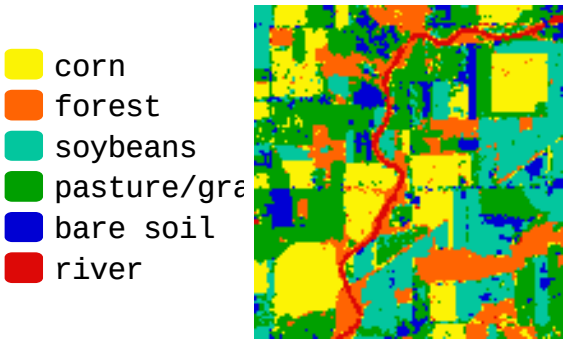
Multispectral methods for deriving information about the Earth's resources using spaceborne sensors began to be studied in the mid-1960's, not long after the launch of the first Earth-looking satellites. The question of how to use aerospace technology for gathering Earth resources data, such as that for the fields of agriculture and food production, geology and the location of oil and mineral resources, geography and urban and non-urban land use was the focus. The motivation for this was to take advantage of both the synoptic view space provides and the economies of scale, since data over large areas could be gathered very quickly and economically from such platforms.

Via such platforms, information is available from the electromagnetic fields that emanate from the Earth's surface, and in particular, from the spatial, spectral, and temporal variations of those electromagnetic fields. The first question that had to be addressed was how to take advantage of these spectral, spatial and temporal variations to derive useful information. Since utilizing temporal variations would require multiple looks at a given area, the study of this one was postponed, because of the added complexity multiple looks and their merging would entail. Use of spatial variations was also postponed, since the spatial resolution that would be required to identify classes of ground cover of interest would be very high. For example, to use airborne or spaceborne data to discrimination between two agricultural crops such as soybeans and corn would require a spatial resolution of the order of a few centimeters, something that would result in extremely large quantities of data if a large area were to be surveyed. This meant that conventional image processing methods would not be appropriate.

These factors placed the focus upon the use of spectral variations, i.e., using the spectral distribution of energy emanating from a pixel to label the contents of that pixel. Then, with the use of pattern recognition methods, a thematic map of a region could be made displaying the amount and distribution of a set of classes of ground cover, both subjectively and quantitatively. Figure 1 illustrates the process. On the left is presented a simulated color infrared image form of the data. This is a form of imagery common in aerial photography since World War II in which bands from the green, red, and infrared are used as the blue, green, and red colors, respectively, of the image. On the right is a thematic map resulting from the analysis of this data.



Color infrared image display of Multispectral data.



Thematic map derived from the data by associating a cover class to each pixel using pattern recognition methods.

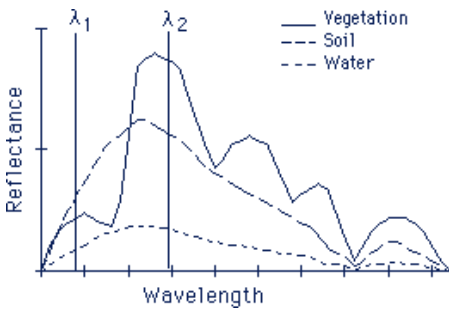
Figure 1. Initial Assumptions About Multispectral Data Analysis (Original in Color)

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

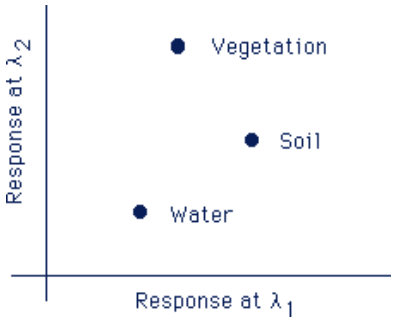
- In image form, i. e., pixels displayed in geometric relationship to one another,
- As spectra, i. e., variations within pixels as a function of wavelength,
- In feature space, i. e., pixels displayed as points in an N-dimensional space.

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.

Image Space



Spectral Space



Feature Space

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, as suggested above, 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 supports area mensuration activities usually associated with remote sensing techniques. For this reason, 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 that 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. Though an attractive and straightforward idea, spectral matching and signature banks have not proven to be very powerful in terms of their ability to extract information in a robust and practical sense.

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, however, it, too, has its limitations in practical circumstances.

Feature Space. The third basis for data representation also begins with a spectral focus, i.e., that energy radiance 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 is to sample the continuous function of emitted and reflected energy vs. wavelength and convert 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. Figure 3 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.

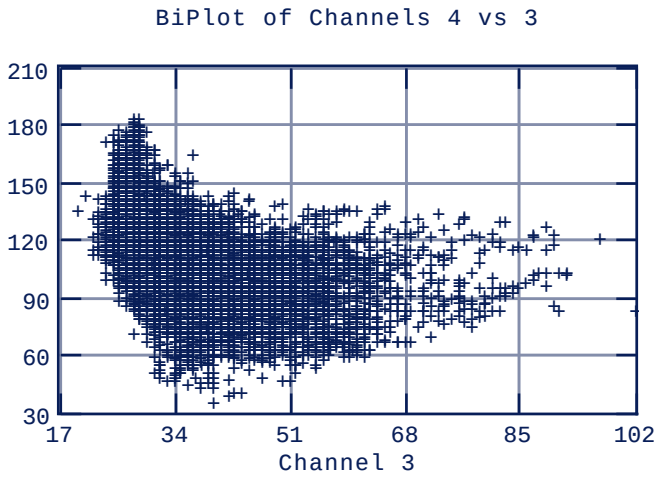


Figure 3. Scatter plot of Landsat Thematic Mapper Channel 4 (0.76-0.90 μm) vs. Channel 3 (0.63-0.69 μm) for an agricultural area containing a small number of crop types.

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 useless or even harmful variability. This scatter is indeed information-bearing, if appropriate means are used to model it.

Another key characteristic that 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 that 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[2,3]. 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 that 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[4]. Shown in the 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 spread at each wavelength is not noise, but is due to the natural variability of reflectance present from the leaves of such vegetation.

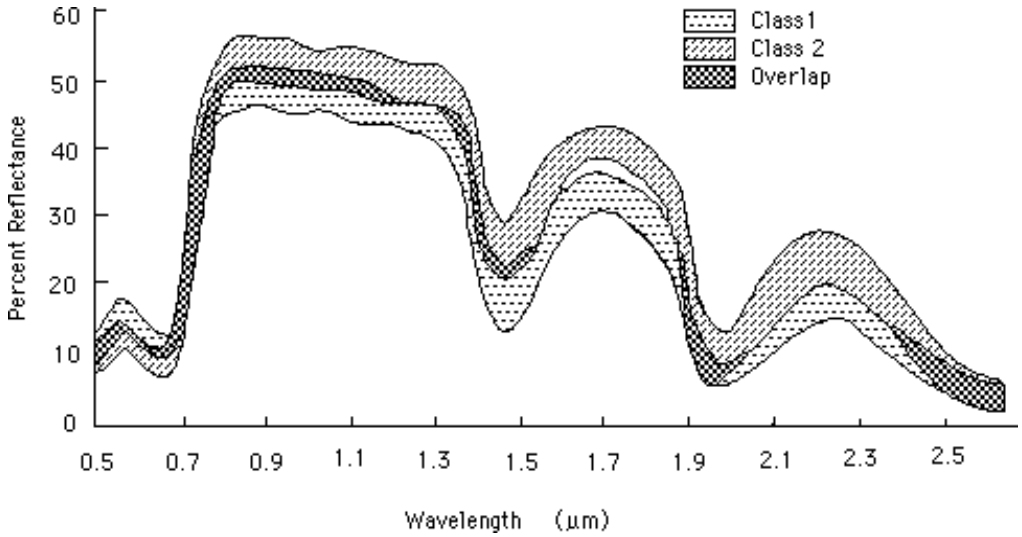


Figure 4. Spectral responses for a typical pair of classes, showing the interval at each wavelength into which they fall.

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.

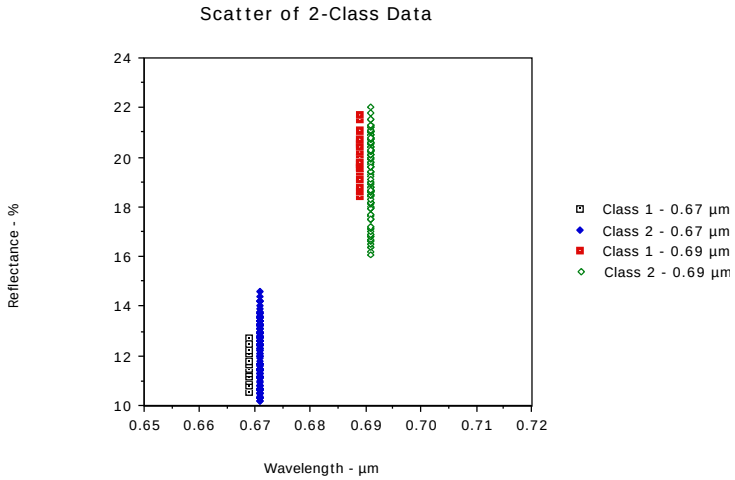


Figure 5. Data points for two vegetation classes in two spectral bands.

However, by plotting the same data in a two dimensional 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. If this were not the case, the two classes would be distributed over a more circular area about their means with a diameter near to the length of the major axis of that shown. 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.

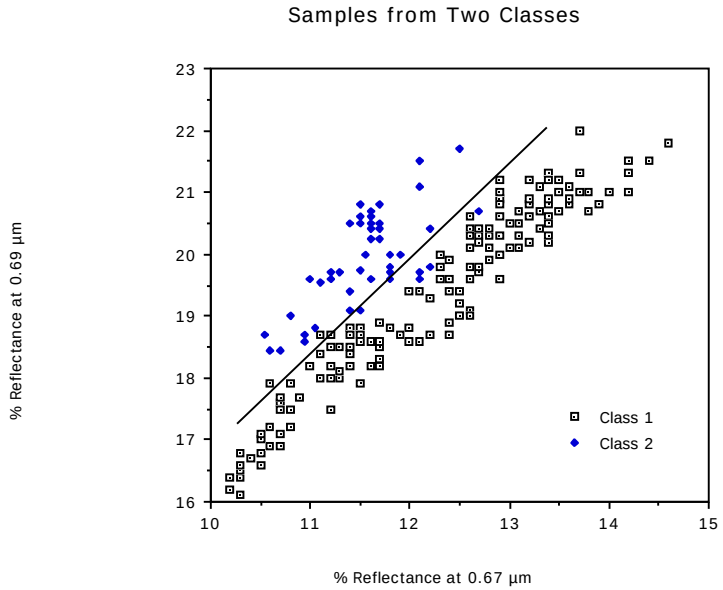


Figure 6. Data points for two vegetation classes in two dimensional feature space.

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[5]. In this paper, the result shown in Figure 7 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 result shown in the figure is 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.

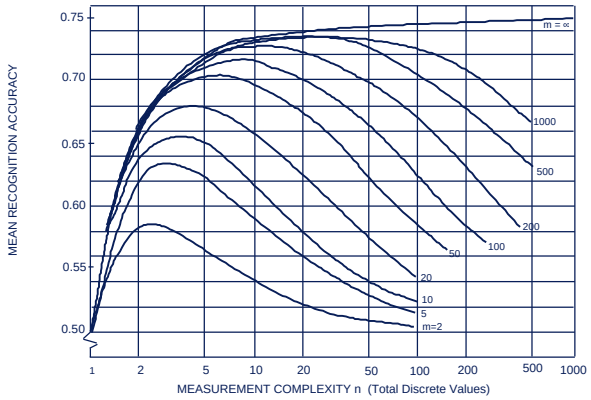


Figure 7. Mean Recognition Accuracy vs. Measurement Complexity for the finite training case.

Some Properties of High Dimensional Spaces

The previous sections of this paper are primarily in the context of conventional, low dimensional multispectral data. However, as the dimensionality is increased, the characteristics of such feature spaces change in a quite counter-intuitive fashion. The volume in the space grows very rapidly with dimensionality[6], leading to such conclusions as a) The volume of a hypercube concentrates in the corners, b) The volume of a hypersphere concentrates in an outside shell, and c) The diagonals are nearly orthogonal to all coordinate axes. However, a conclusion especially relevant as a practical matter is that the required number of labeled samples for supervised classification increases as a function of dimensionality. Fukunaga[7], 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[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[9,10].

It is reasonable to expect that high dimensional data contain more information in the sense of a capability to detect more classes with more accuracy. As an illustrative point of view, assume that one has data from 100 spectral bands, and that the data are of 10 bit precision. Then in each of the 100 bands there are 1024 discrete possible values, and in the whole space

there are $1024^{100} \approx 10^{300}$ possible discrete locations. This number is so large that even with a data set of a million pixels, the probability of any two pixels falling in the same discrete location in the space is vanishingly small. Thus, theoretically, any thing is separable from anything, since there is no overlap. However, to achieve such discrimination, one must be able to locate a decision boundary very precisely. This implies a very precise estimation of the class statistics is needed. Thus as the dimensionality of data available is increased, the finite size of a training set must grow to realize the potential of the data, and a more complex and sophisticated classification algorithm may make matters worse rather than better.

It has also been shown that 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 normal (Gaussian) distribution with probability approaching one. 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.

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[11] 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[12,13,14]. 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 that 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 based upon a method known as projection pursuit[15,16]. This algorithm does the necessary calculations in the projected space, rather than the original, high dimensional space. Figure 8 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.

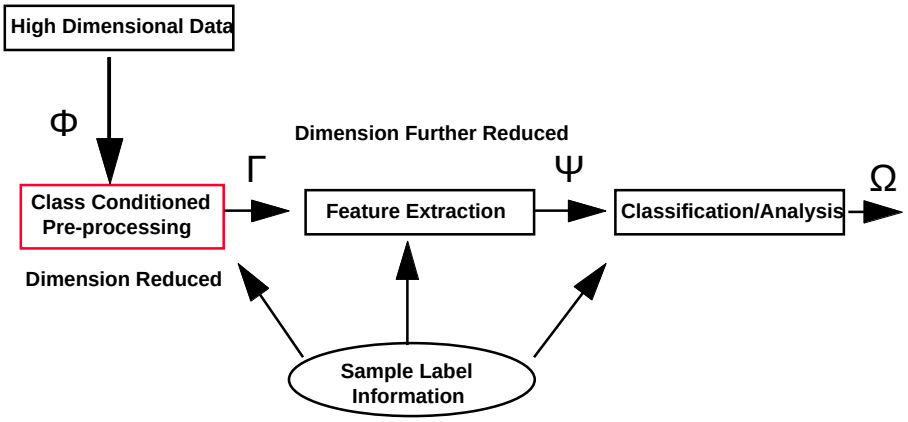


Figure 8. Classification of high dimensional data including reprocessing of high dimensional data.

Methods and Procedures.

The above concepts and theory then direct one to the following general procedure for analyzing a hyperspectral data set in order to divide the data set into a set of classes of interest.

1. **Identify and label training samples.** The analysis process begins by the user specifying what classes are of interest by labeling examples of each of the classes in the data set to be analyzed. It is from these samples that the quantitative statistical description of each class is generated, i.e., that the subjective list of classes that the user has in mind are converted to specific quantitative descriptions. Doing so also tends to normalize for various extraneous measurement variables, such as the effects of the atmosphere, goniometric variations due to the non-lambertian nature of most surface materials and the like, thus obviating the need for extensive preprocessing which might otherwise be used to minimize the impact of such extraneous measurement variables.

2. **Feature determination.** In low dimensional cases, this can be simply a matter of feature selection, i.e., choosing the best 4 features to use out of the 7 available, for example. Even in this simpler, low dimensional case, such reduction in dimensionality can aid in reducing the Hughes effect described above as a result of the training set being of limited size. In higher dimensional cases, the use of feature extraction methods such as Discriminate Analysis or Decision Boundary Feature Extraction would be used instead. In the case of a very difficult discrimination problem involving very subtle classes, the Projection Pursuit method may need to precede DAFE or DBFE, as illustrated above.

3. **Classification.** Having defined the classes and the features to be used, the next step is to apply a suitable classification algorithm to assign each pixel to one of the classes of interest.

Note that, rather than attempting to set up a fully automatic procedure in which there is no human participation, this procedure might properly be labeled a human assisted machine implemented one. The intent is to augment human intelligence in the analysis process rather than to replace it. Fully automated procedures are possible in very simple cases, but it has been found that for very complex and dynamic scenes such as those of the Earth's surface, automatic procedures are practical for only the simplest of discrimination problems.

The analysis of a high dimensional image data set is clearly not a simple task, and in this paper we have provided only a broad overview of some of the salient features. It is also by no means a "cookbook" process, and significant skill derived from practice is necessary to achieve the full capability that such data can deliver.

Further, the algorithms involved are not simple to implement. For this reason, a complete analysis system, containing the algorithms described above, has been constructed and is available in the form of a software system for personal computers. This system, called MultiSpec, is available for downloading from the web at

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

Versions are available for both Macintosh and Windows operating systems, however the more advanced algorithms are available only in the Macintosh version at this moment. This location also contains an extensive amount of documentation of MultiSpec and a number of examples, in some cases including the data used in them. MultiSpec and its documentation are available to anyone without charge.

References

[1] Landgrebe, David, "The Evolution of Landsat Data Analysis," Photogrammetric Engineering and Remote Sensing, Vol. LXIII, No. 7, July 1997, pp. 859-867 (Special issue commemorating the 25th anniversary of the launch of the first Landsat satellite.) This paper and several others referenced in this paper are available for downloading from http://dynamo.ecn.purdue.edu/~landgreb/publications.

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

[3] Papoulis, A., Probability, Random Variables, and Stochastic Processes, Second Edition, McGraw-Hill 1984.

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

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

[6] A more detailed indication of the characteristics of high dimensional feature spaces is given in 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 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.

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

[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.

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

[10] 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.

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

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.