Penalized Discriminant Analysis of *In Situ* Hyperspectral Data for Conifer Species Recognition

Bin Yu, *Senior Member, IEEE,* I. Michael Ostland, Peng Gong, and Ruiliang Pu

***Abstract*—Using *in situ* hyperspectral measurements collected in the Sierra Nevada Mountains in California, we discriminate six species of conifer trees using a recent, nonparametric statistics technique known as penalized discriminant analysis (PDA). A classification accuracy of 76% is obtained. Our emphasis is on providing an intuitive, geometric description of PDA that makes the advantages of penalization clear. PDA is a penalized version of Fisher's linear discriminant analysis (LDA) and can greatly improve upon LDA when there are a large number of highly correlated variables.**

I. Introduction

CLASSIFICATION of forest species is important in natural resource management, environmental protection, biodiversity, and wildlife studies. Conventionally reliable methods for tree species recognition depend mainly on costly, time-consuming, and labor-intensive inventory in the field or on interpretation of large-scale aerial photographs. The use of these methods is frequently limited by cost and time and is not applicable to large areas. Another option is the use of hyperspectral data such as those obtained from field and imaging spectrometers. An important step toward large-scale application of this approach is the successful classification of individual trees using ground-based hyperspectral measurements.

Such data were used to estimate biochemistry constituents [1]–[4] and were shown to detect subtle spectral changes of various targets [5]. But many studies used spectra measured either from tree leaves only [6], [7] or from selected components of forest stands such as branches of needles, shoot stacks, barks, and litter and soil [8], [9]. Although valuable for understanding the underlying biophysics and biochemistry, this "decomposed" approach requires difficult, nonlinear models in order to characterize properties of remotely sensed forest canopies in terms of these constituent parts [10], [11].

Different from the "decomposed" approach, our goal is to use hyperspectral data measured directly from above forest canopies in the field. We believe that the high spectral resolution data in hundreds of bands provide a wealth of information for forest species discrimination. Our belief is supported by initial results, in which an artificial neural network algorithm using the spectral derivative technique successfully discriminated six conifer species [12].

However, neural networks can be tricky to tune, and the parameters are difficult to interpret. We desire a classifier that is easy to implement and that provides high accuracy and a ready physical interpretation. Penalized discriminant analysis (PDA) as developed by Hastie *et al.* [13] is reviewed here as a promising technique for realizing these goals. Like linear discriminant analysis (LDA), PDA produces linear combinations that show how the components of the predictor vector contribute to the discrimination rule. Unlike LDA, which is known to fail when faced with the high dimension and high correlation of adjacent spectral bands, PDA often performs well with hyperspectral data. Furthermore, the penalization of PDA has a nice geometric interpretation that makes clear how PDA escapes the pitfalls of LDA in situations such as ours. On our data, PDA roughly doubled the accuracy of LDA and narrowly outperformed a well-tuned neural network similar to those in our previous work [12].

Our particular example is not definitive about the general applicability of PDA for forest species recognition. As Section II will describe, our data are exclusively young conifers measured vertically tens of centimeters above tree canopies, and spectral reflectance properties from such data may not scale up to the conifers' adult counterparts. However, the success in this specific problem is clearly promising.

Instead of an exhaustive survey, we would like to mention briefly a few discrimination methods applicable to hyperspectral data. A binary coding algorithm (BCA) [14] encodes each hyperspectral band to zero or one according to the sign of the first order derivative and the difference between the spectral value and the mean of all the bands. Any of a number of multivariate binary data classification procedures [15] could then be used. This encoding characterizes the general pattern of a spectral curve and can be useful in discriminating populations with very different general patterns. However, this is not suitable for conifer species discrimination, because their spectral curves have similar shapes with only small differences in magnitudes. Bensmail and Celeux [16] use cross validation to choose among classes of models that impose restrictions and/or varying degrees of commonality on the components of the eigenvalue decompositions of the within group covariance matrices. Unfortunately, the estimated eigenvectors (whether common or not) will suffer from the same pitfalls that we describe in Section III.

Manuscript received October 30, 1997; revised July 28, 1998. The work of B. Yu was supported in part by ARO DAAH04-94-G-32 and NSF DMS 9305601. The work of I. M. Ostland was supported in part by ARO DAAH04-94-G-32. The work of P. Gong was supported in part by IHRMP of California.

B. Yu and M. Ostland are with the Department of Statistics, University of California, Berkeley, CA 94720 USA.

P. Gong and R. Pu are with the Center for Assessment and Monitoring of Forest and Environmental Resources, Department of Environmental Science, Policy, and Management, University of California, Berkeley, CA 94720 USA.

Publisher Item Identifier S 0196-2892(99)06269-5.

This paper is organized as follows. Section II describes the study area and data collection process. Section III carefully introduces PDA, emphasizing a geometric point of view that is helpful in understanding how PDA can improve over LDA. A simulation in that section demonstrates the geometric intuition in high dimensions. In Section IV, we give an example of PDA analysis using our data. Section V contains some concluding remarks.

II. Study Area and Data Description

A. Study Area

Hyperspectral measurements were taken at the Blodgett Forest Research Station of the University of California, Berkeley, located in the American River watershed on the western slope of the central Sierra Nevada, El Dorado County, CA. The vegetation consists of the normal associates of the Sierra mixed-conifer forest type, the major tree species include five conifers: sugar pine (SP, *pinus lambertiana*), Ponderosa pine (PP, *pinus ponderosa*), white fir (WF, *abies concolor*), Douglas fir (DF, *pseudotsuga menziesii*), incense cedar (IC, *calocedrus decurrens*), and one hardwood, California black oak (*quercus kelloggii*). All but the black oak are present in our data. In addition, we also measured the giant sequoia (GS, *sequoiadendron giganteum*), a species native to the Sierra Nevada but not found in the Blodgett Forest and which has been planted in selected sites since the 1900's. Major shrub species include manzanita, deerbrush, white thorn, and bear clover.

B. Spectral Reflectance Collection

Field measurements were taken with the PSD1000 [17], a high spectral resolution spectrometer designed for use with a portable computer and capable of precise measurements from 210 to 1050 nm. The PSD1000 covers over 1500 bands with an average band width of about 0.5 nm and spectral resolution of approximately 2.6 nm. The field of view of the spectrometer is approximately 22°. Three types of spectral measurements can be made: dark current (the response of the system with no light being exposed to detectors), white reference (spectra from a standard white panel with close to perfect diffusion), and sample (spectra obtained from the target of interest). To avoid saturation or shortage, an integration time for collecting photons is selected based on the illumination condition and by adjusting the sampling frequency. A reflectance spectrum can be generated by dividing the sample radiance by the radiance from the standard white reference under the same light condition.

At Blodgett Forest, six sites (see Table I) were chosen for hyperspectral measurements at different times in multiple years for long-term monitoring of selected tree species. Canopy sizes at sites 2, 3, and 6 are smaller than those at sites 1, 4, and 5. Site 1 has the largest canopies with a dry soil background free of litter and understory vegetation. Sites 4 and 5 have more litter and understory vegetation surrounding the tree canopies than all the other sites. Our measurements were collected between June 2 and 3, 1996 between 11:00 and 13:00 local time under a clear sky with air temperatures ranging from 20 to 30 °C. We measured young conifer trees (four to seven years old) only. Measurements were made at heights less than 1.5 m from vertical directions, 15–20 cm above canopies. We do not believe that a result based on a sample of only young trees can be directly generalized to the entire population. Rather, we believe that our work is suggestive of possible benefits, requiring further investigation on the adult population.

TABLE I
Site Names and Sample Distributions

Z	Site Description	Trees per Species	Reps per Tree	No. of Obs
1	Sequoia	2	6	72
2	Hand-cut	6	2	70*
3	Grazed	10	1	60
4	Flat-area	5	2	60
5	Valley	2	2	24
6	Hand-weeded	6	1	36

*Two observations from site 2 were mistakenly lost.

Dark current and white reference were measured every 5–10 min as necessary to reduce the effects of possible illumination differences. A total of 322 reflectance spectra were measured from the six conifer species in equal proportions at each of the six locations. For some locations, each tree was measured multiple times. We denote our data by (X_i, Y_i, Z_i) where $X_i = (x_{i1}, \cdots, x_{iK})$ is the vector of spectral reflectances corresponding to our K bands, and (Y_i, Z_i) are categorical variables indicating the species and location, respectively, of the associated tree. We will make use of standard statistical terminology by referring to (X_i, Z_i) as predictors.

C. Preprocessing and Aggregating Data

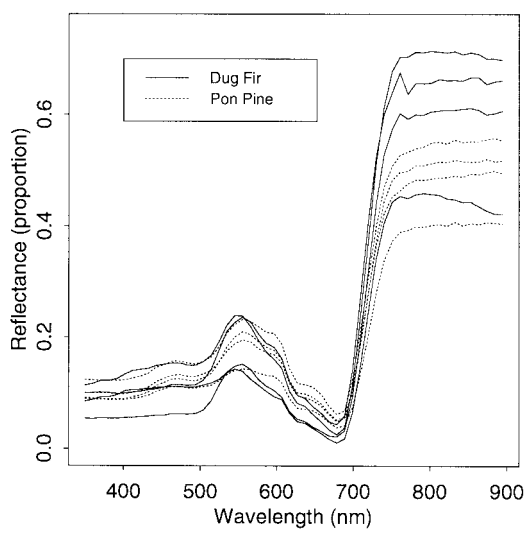
For all analyses that follow, we perform the following preprocessing of our data. First, spectral curves are truncated below 350 nm and above 900 nm, since the measurements are extremely noisy outside of this range. This leaves us with 1073 bands, each with a width of about 0.51 nm. Next, we take simple averages over blocks of six neighboring bands, leaving us with $K = 179$ bands. We then normalize the spectral curves for constant area by dividing by the mean reflectance for that curve. That is, we replace x_{ij} with

$$x_{ij} \Big/ \left(\frac{1}{K} \sum_j x_{ij} \right). \tag{1}$$

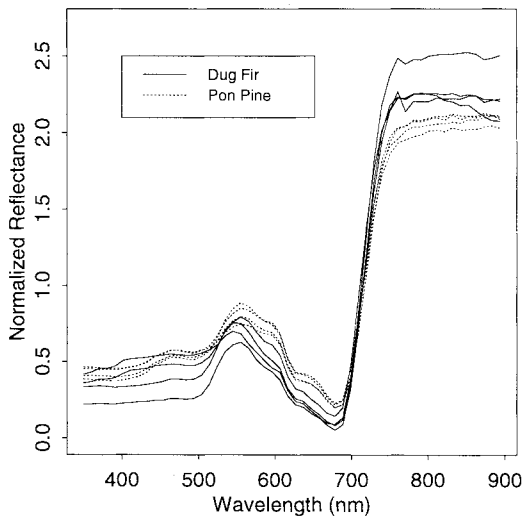
The benefit of such a normalization is the suppression of illumination differences. Fig. 1(a) shows a plot of unnormalized x_{ij} versus band wavelength for eight observations (four DF and four PP). Fig. 1(b) shows the same curves after normalization. Notice the clearer separation between the species over a wide range of frequencies in Fig. 1(b).

III. Methods

High dimension, strong correlation within the vector X_i, and similarity of classes make our discrimination problem difficult. Standard techniques, such as LDA, are known to



(a)



(b)

Fig. 1. (a) Unnormalized spectral reflectance curves for two species and (b) same curves now normalized by (1).

perform poorly in such contexts. Hastie *et al.* [13] introduced a penalized variation on LDA, the aforementioned PDA, that could be a considerable improvement. In this section, we carefully describe LDA and PDA in order to give interested researchers the tools to perform PDA and the intuition to understand the source of the improvements. In Section IV, we give an example. Readers interested in theoretical details of PDA are referred to [13].

A. Fisher's Linear Discriminant Analysis (LDA)

LDA [18] is a classical technique that assumes only that the data are drawn from G groups with K-dimensional group mean vectors $M_j, j = 1, \cdots, G$ common within group covariance matrix Σ_W and proportions $\pi_1, \cdots, \pi_G$ of the groups in the population. LDA searches for successive linear combinations of the data such that the group means of the linear combinations are spread out as much as possible relative to the within group variation. Specifically, we find $\beta \in R^K$ with $\beta^T \Sigma_W \beta = 1$ such that $f = \sum_{j=1}^{G} \pi_j (\beta^T m_j - \beta^T \overline{M})^2$ is maximized. Here, $\overline{M} = \sum_j \pi_j M_j$ is the overall population mean vector. Some trivial algebra yields $f = \beta^T \Sigma_B \beta$, where by definition $\Sigma_B = \sum_{j=1}^{G} \pi_j (M_j - \overline{M})(M_j - \overline{M})^T$.

To maximize f, note that the ratio $g = \beta^T \Sigma_B \beta / \beta^T \Sigma_W \beta$ is identical to f under the constraint $\beta^T \Sigma_W \beta = 1$. But since g does not depend on $\|\beta\|$, we can do (unconstrained) maximization of g and then rescale any solution by $\sqrt{\beta^T \Sigma_W \beta}$ to satisfy the constraint. Differentiation leads to the eigensystem $\Sigma_W^{-1} \Sigma_B \beta = g\beta$. In this way, the β's are seen to be the properly scaled eigenvectors corresponding to the $d \leq G - 1$ nonzero eigenvalues of $\Sigma_W^{-1} \Sigma_B$. We have $d < G - I$ if the G means lie in a hyperplane of dimension less than $G - 1$, since this implies Σ_B has reduced rank.

Let $X = (x_1, \cdots, x_K)$ denote a test sample to be classified. LDA forms the $K \times d$ matrix $\mathbf{B}_{\text{LDA}}$ whose jth column is β_j. It next forms the d-vector $\mathbf{B}_{\text{LDA}}^T X$, whose components are sometimes called *discriminant variables*. LDA then classifies according to Euclidean distances between discriminant variables and the transformed class means. Namely, X is classified into group

$$\arg\min_j \|\mathbf{B}_{\text{LDA}}^T X - \mathbf{B}_{\text{LDA}}^T M_j\|^2 - 2 \log \pi_j. \tag{2}$$

There are several appealing aspects of this methodology. First, it is intuitive, since groups are easier to tell apart if their means are well spread out relative to the within group variability. Second, the reduction from K to d dimensions allows easy graphical inspection of the training and test data. Third, an equivalent classification results from replacing the first term in (2) with the Mahalanobis distance $(X - M_j)^T \Sigma_W^{-1} (X - M_j)$ (cf. [13]). This in turn is the well-known Bayes Rule when the populations are Gaussian.

We note that many authors define LDA in terms of the training sample estimates of M_j, Σ_W, and π_j. This is certainly reasonable, since the population parameters are rarely (if ever) available. However, we draw the distinction between the population and sample quantities since, as we will show in Section III-D, the high variance of the plug-in estimate of Σ_W is the primary obstacle to a successful LDA in our context. It is precisely this obstacle that penalization addresses.

B. Geometric Perspective of LDA

In this section, we give a geometric interpretation of LDA for $K = 2$ dimensions and $G = 3$ classes. Although it is impossible to graph in higher dimensions, the ideas here are the key to understanding the benefits of penalization. Some helpful geometric discussion of LDA also can be found in [19].

Fig. 2 graphically demonstrates LDA. In Fig. 2(a), the numbered ellipses are 50% probability contours of bivariate Gaussian densities with common covariance and means marked 1, 2, and 3. The constraint $\beta^T \Sigma_W \beta = 1$ corresponds to the ellipse about the origin. Our discussion centers on the following trivial reexpression of f:

$$f = \|\beta\|^2 u^T \Sigma_B u \tag{3}$$

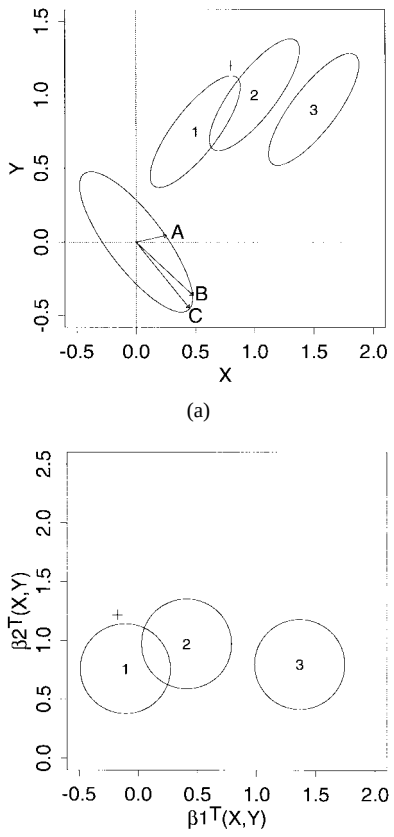


Fig. 2. (a) Equal probability contours of three densities and constraint $\beta^T \Sigma_W \beta = 1$. A: direction most separating the means; B: first LDA direction; C: direction keeping within class variation smallest and (b) same curves after transformed into discriminant variable space. Plus marks a new observation to be classified.

where $u = \beta/||\beta||$. Thus, f is a product of two quantities, one depending only on the norm of β and the other only on the direction. Point A in Fig. 2(a) is one of two points on the ellipse (the other being A reflected about the origin) that maximizes $u^T \Sigma_B u$. Similarly, point C is one of the two points that maximizes $||\beta||$. Neither A nor C is a satisfactory choice for separating the groups. In direction A, the means are well separated, but groups two and three overlap substantially from the high within group variation. In direction C, the within group variation is small, but the means from group one and two are nearly on top of each other. As we move along the constraint ellipse from A to C, we gain in (3) from increasing $||\beta||$ on the one hand, but on the other hand we lose, since the right term decreases as u moves away from the direction passing through A. The point B is exactly where the losses start to overtake the gains. Thus, f is locally maximized at B, which is therefore the first LDA direction β_1. Another local maxima β_2 can be found between A and the reflection of C about the origin (not shown in Fig 2).

Fig. 2(b) shows the density contours transformed by the two LDA linear combinations β_1 and β_2. Specifically, let $z = (x, y)^T$ denote the coordinates with respect to the original basis in Fig. 2(a), then the x axis in Fig. 2(b) is $\beta_1^T z$ and the y-axis $\beta_2^T z$. The small cross symbol represents a new observation to be classified. Although in original coordinates this new observation is closer to the first class mean in a Euclidean sense, once transformed by the LDA linear combinations, it is closer to the second mean (which is clearly where it should be classified given the within class covariance).

A problem with LDA is that with many highly correlated predictors there is too much flexibility in the choice of the β's for the method to be robust to a poor estimate of Σ_W. High correlation yields a constraint ellipsoid with the major axis much longer than the others. In three dimensions, this is much like a long, thin cigar with tips very far from the origin. In higher dimensions, we still can think of the two regions of the ellipsoid's surface near the intersections with the major axes as "tips." A poor estimate of Σ_W causes the ellipsoid to be poorly oriented compared to the ellipsoid based on the true Σ_W. This is clearly undesirable since, just as in the two-dimensional example above, the left term of the product (3) encourages movement toward the tips of the ellipsoid. But with the cigar potentially misoriented, the negative impact on the right term in (3) may not always compensate for the positive effect of increasing $||\beta||$. The net result is a β too far out in the tip of $\beta^T \Sigma_W \beta = 1$. This is less of a problem in low dimensions for two reasons. First, with few parameters, Σ_W is easy to estimate, and second, there are not as many dimensions in which to find a route toward the tip of the cigar for which $||\beta||$ dominates the right term of (3). For instance, in Fig. 2(a) the ellipsoid has a one-dimensional surface. However, with thirty to hundreds of dimensions, as in problems such as ours, the constraint surface is very high-dimensional, and LDA is almost sure to find a route along the misoriented ellipsoid to get near the tip.

This tendency of LDA to favor β's too far in the tips of the constraint ellipsoid results in grossly inflated $||\beta||$'s and overly rough or wiggly (when plotted against index) directions. Fig. 5 gives an example. The extreme roughness is a property of the tips of the constraint ellipsoid (true or misoriented), as determined by the extreme correlation structure in Σ_W. It is the misorientation from the poor estimate of Σ_W that allows LDA to drift too far out into these tips. This drifting is what statisticians refer to as over-fitting, and manifests itself (as we will see) in perfect classification on the training set, but very poor performance on the test set. In later sections, we show how penalization can be used to limit this drifting.

C. Penalized Discriminant Analysis (PDA)

To improve the performance of LDA, Hastie *et al.* [13] add a penalty term to the within species covariance matrix Σ_W. Specifically, they replace Σ_W with $\Sigma'_W = \Sigma_W + \Omega$, where Ω is a K by K matrix such that $\beta^T \Omega \beta$ is large for "undesirable" β. In our context, undesirable could mean spatially rough or having large $||\beta||$. We discuss details of selecting such a penalty matrix below. One then proceeds exactly as before. Let $\mathbf{B}_{\mathrm{PDA}}$ denote the $K \times d$ matrix, whose columns are eigenvectors of $\Sigma'_W \Sigma_B$ [classify using the analog of (2)].

Penalizing is a fairly common practice in the statistical literature [20]. In fact, Friedman's regularized discriminant analysis [19] is basically PDA with an additional parameter that controls how much the individual within group covariance

matrices are shrunk toward a common value. Trading a small amount of bias for a reduction in variance is a standard parameter estimation point of view of penalization (cf. [20]). Since estimating a large number of parameters with limited data results in high variance and Σ_W has $K(K+1)/2$ parameters to estimate $(K \approx 100)$, it is easy to see why such a tradeoff could be beneficial in our setting.

Here, we return to geometry to interpret the effect of penalization. The new constraint ellipsoid defined by $\beta^T \Sigma_W' \beta = 1$ will differ from the unpenalized ellipsoid in that the penalty term will cause $||\beta||$ to be smaller in the undesirable directions. This of course impacts (3), driving down the objective function in these directions and forcing more desirable directions into preference.

In the sequel, we consider two types of penalty matrices. For penalizing high local variation, we consider a second derivative-type penalty matrix Ω_D. Namely, let D_k denote $k-1$ by k-dimensional first difference operator matrix. For example, D_4 is

$$\begin{pmatrix} 1 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & -1 \end{pmatrix}. \tag{4}$$

Then define

$$\Omega_D = \lambda D_K^T D_{K-1}^T D_{K-1} D_K. \tag{5}$$

The nonnegative parameter λ is called the *smoothing parameter*, since it controls how much of a price is paid for local variation. A second shrinkage-type penalty is of the form

$$\Omega_s = \lambda I_K \tag{6}$$

where I_K is the $K \times K$ identity matrix. This is a similar idea to ridge regression analysis. The name shrinkage comes from the fact that up to a λ, the penalty term $\beta^T \Omega \beta$ reduces to the usual Euclidean norm $||\beta||^2$. In other words, the penalty favors β's that are close to the origin.

We now apply our geometric interpretation to Ω_S. By adding a constant to the diagonal elements of Σ_W, the shrinkage penalty simply rounds and shrinks the constraint ellipsoid (cf. Fig. 3). Compared with the unpenalized ellipsoid, the new, more circular constraint allows less of a reward in (3) in the form of a larger $||\beta||$ for moving into the tip. So the PDA direction moves only to the point b instead of all the way down to the point B as with LDA.

The penalty Ω_D is not so simple to interpret. We note that one can change basis in such a way that Ω_D is diagonal (but with differing elements along the diagonal) with respect to the new basis [21]. Thus, the penalty amounts to shrinkage with respect to the new basis, where the different coordinates are shrunk unequally.

There are other possibilities for the choice of Ω, and one should use the science underlying the application to guide this decision. For instance, if there is a reason to insist on more local smoothness at some wavelengths than others, this could be accommodated easily by modifying Ω_D.

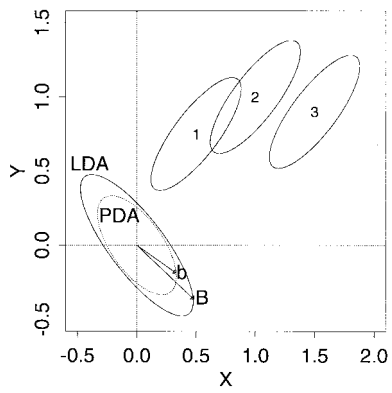


Fig. 3. Constraint $\beta^T(\Sigma_W + \Omega_S)\beta = 1$ and $\beta^T \Sigma_W \beta = 1$. B: first LDA direction; b: first PDA direction.

D. Effect of Covariance Estimation on LDA and PDA

In this section, we use a simulation to demonstrate the ideas of the previous sections in a high-dimensional setting. We simulate 20 vectors of training data of dimension $K = 50$ from each of three Gaussian distributions, the mean vectors of which are parabolic when considered as a function of index. Namely, $M_i(i) = 0.01 * (25-i)^2$, $M_2(i) = 0.6 + 0.008 * (25-i)^2$, and $M_3(i) = -0.6 + 0.012 * (25-i)^2$ for $i = 1, \cdots, 50$. The common within group covariance matrix Σ_W is defined such that the (m, n)th element is $1 - 0.01 * |m-n|$. These precise numbers are not so important. The means and covariance structure are selected to resemble our hyperspectral data in terms of very high correlation and smooth, similarly shaped underlying means. We then generate 50 vectors of test samples from each of the three distributions. Next, we compare eight different classifiers of the test data. The first four are LDA with:

1) Σ_W and Σ_B known;
2) only Σ_B known;
3) only Σ_W known;
4) neither known.

The next four are the same but use a PDA with a shrinkage penalty and a value of λ chosen because it performed well on preliminary simulations. When Σ_W or Σ_B is unknown, it is estimated from the training sample in the standard way. For all eight classifiers, we record the test set classification accuracy and β_1, the first linear combination.

Table II reports the twenty-fifth, fiftieth, and seventy-fifth percentiles of three quantities of interest from 25 independent simulations as described above. The first quantity is the test-set classification accuracy (or "rate"). Let β^* denote the first LDA direction based on the true Σ_W and Σ_B (i.e., the Bayes Rule), and then the remaining quantities of interest are the ratio of norms $r = ||\beta_1||/||\beta^*||$ and the angle between directions $\theta = \cos^{-1}(\beta_1^T \beta^* / ||\beta_1|| ||\beta^*||)$ in degrees. The four broad columns of the table correspond to the states of knowledge as defined in the previous paragraph.

When both parameter matrices are known, neither LDA nor PDA depends on the training data. Hence, the only numbers that vary in the left column are the classification rates, which still depend on the random test samples. In terms of accuracy,

TABLE II
RESULTS FROM 25 SIMULATIONS. RATE IS TEST SET ACCURACY. r IS RATIO OF THE NORM OF THE ESTIMATED LINEAR COMBINATION TO THE NORM OF THE OPTIMAL (BAYES RULE) LINEAR COMBINATION. θ IS THE ANGLE IN DEGREES BETWEEN THE ESTIMATED AND OPTIMAL DIRECTIONS. (NOTE: FOR A LIST OF 25 NUMBERS WHOSE VALUES SORTED IN ASCENDING ORDER ARE $x_{(1)}, \cdots, x_{(25)}$, THE TWENTY-FIFTH, FIFTIETH, AND SEVENTY-FIFTH PERCENTILES ARE $x_{(7)}$, $x_{(13)}$, AND $x_{(19)}$, RESPECTIVELY)

known		Σ_W & Σ_B			Σ_B			Σ_W					
percentile		25	50	75	25	50	75	25	50	75	25	50	75
rate	LDA	.87	.90	.93	.52	.57	.62	.80	.83	.88	.48	.55	.58
	PDA	.87	.90	.93	.82	.82	.90	.85	.87	.88	.77	.83	.88
r	LDA	1.0	1.0	1.0	23.8	29.5	40.0	1.6	1.7	1.8	22.3	25.9	37.1
	PDA	0.8	0.8	0.8	1.9	2.0	2.0	0.9	1.0	1.0	2.0	2.0	2.1
θ	LDA	0.0	0.0	0.0	80.2	84.9	87.7	49.7	54.3	59.6	83.2	85.9	87.7
	PDA	15.5	15.5	15.5	58.6	62.9	68.4	37.7	40.6	47.5	59.8	64.7	69.9

PDA matches LDA very closely in the left column. Hence, one loses very little by invoking a modest penalty in the ideal situation when all is known. More importantly, when Σ_W is unknown, we see that PDA does appreciably better than LDA. Where the accuracy of the median LDA result falls from 90% to under 60%, PDA falls only 8%. To explain this, we see that the LDA β_1 is far too large ($r \approx 30$) and is nearly orthogonal to β^* ($\theta \approx 85°$), while PDA keeps $||\beta_1||$ under control ($r \approx 2$) and closer to the optimal direction ($\theta \approx 65°$). It is interesting that the θ's for PDA are still quite large, but the minor, consistent improvement is enough to yield the benefit in terms of test set accuracy. The differences between cases 2 (Σ_B known) and 4 (nothing known) are minor.

Neither method suffers much from needing to estimate Σ_B in case 3 (Σ_w known). This agrees with our geometric discussion of a misoriented constraint ellipsoid as the main source of our problems. Knowing Σ_W gives us a perfectly oriented ellipsoid. Consequently, r is only inflated by about 70% or so for LDA, as opposed to a factor of about 30 when Σ_W must be estimated. Similarly, θ is much better behaved. However, it is noteworthy that even here penalization improves slightly on θ, r and test set accuracy.

IV. EXAMPLE WITH HYPERSPECTRAL DATA

We now demonstrate PDA using the hyperspectral data detailed in Section II. From the 322 observations, we form a test sample consisting of the 60 observations from site 3 plus the 60 observations from site 4. The 202 observations from the other four sites form the training sample. We compare classification accuracy using PDA with two types of penalties and a range of values for the smoothing parameter λ. The case $\lambda \approx 0$ corresponds to LDA and is included for comparison. We present plots of the discriminant variables and directions in order to highlight the effect of penalization on discriminant analysis. The PDA analysis itself is done in Splus using the $mda()$ collection of functions written by Hastie and Tibshirani. These functions are documented and publicly available from the S archive of StatLib at http://lib.stat.cmu.edu.

Fig. 4 shows the classification accuracy for PDA with second-derivative Ω_D and shrinkage Ω_S penalties for various choices of the smoothing parameter λ. For the smallest levels of λ, neither penalty has any effect, and the classification is just that of LDA. The overall classification accuracy (number correctly classified divided by number of test samples times 100%) is about 38.3% for LDA. As the level of smoothing increases, the accuracies of both forms of PDA improve until peaking and then declining again. PDA with Ω_D has a slightly higher peak than with Ω_S (76.7 versus 74.2%), and the former has a less abrupt dropoff in performance for nonideal choices of λ. Since the best λ will have to be estimated, this second property is quite important. Cross validation is one tool for such estimation. A discussion of the strengths and weaknesses of cross validation can be found in Efron [22]. Neural networks similar to [12] achieved accuracies between 60–75% depending on a variety of tuning parameters such as the number of hidden nodes.

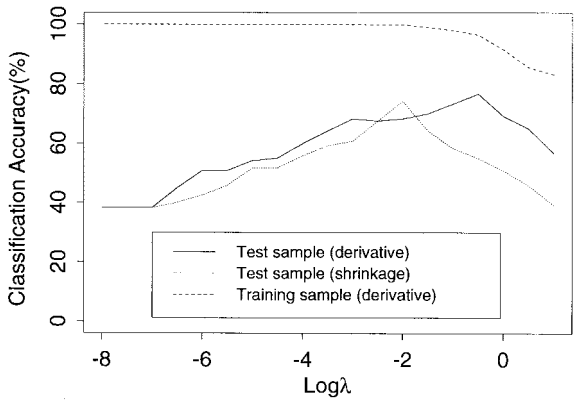


Fig. 4. Classification accuracy as a function of smoothing parameter λ.

Although not shown, we repeated the analysis on other partitions of test and training sample, and the results were very similar. LDA had classification accuracies between 18–50%, while the well-tuned PDA increased accuracies to between 60–90%. The peak accuracies were comparable for the two types of penalties, but as in Fig. 4, Ω_D yielded high accuracy over several orders of magnitude in the smoothing parameter, whereas Ω_S was more peaked. Overall, the best test set performance for Ω_D occurred with $\log_{10}\lambda \approx -2$, which is a bit lower than the best smoothing parameter value in Fig. 4.

A final interesting observation from Fig. 4 is that the range of λ's for which PDAΩ_D attains its highest test set accuracy is just about at the point where the method starts to misclassify on the training set. This phenomenon also occurred consistently in our examples. This observation could be used (instead of cross validation) as a rough guide to select λ.

Fig. 5 plots the first two of the five β's (rescaled to have norm 1) versus index for LDA (dashed) and PDA-Ω_D with $\log_{10}\lambda = -0.5$. Similar to our simulations in Section III-D,

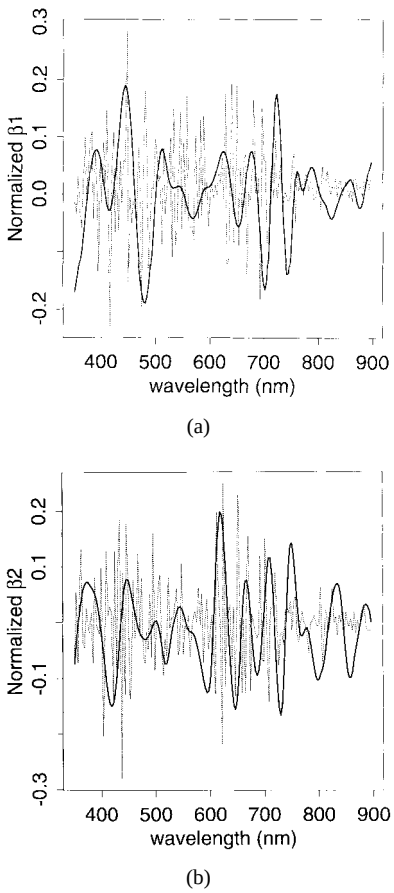


Fig. 5. (a) First direction, $\beta_1/\|\beta_1\|$, versus index for LDA (dashed) and PDA-Ω_D with $\log_{10}\lambda = -0.5$ (solid) and (b) second direction.

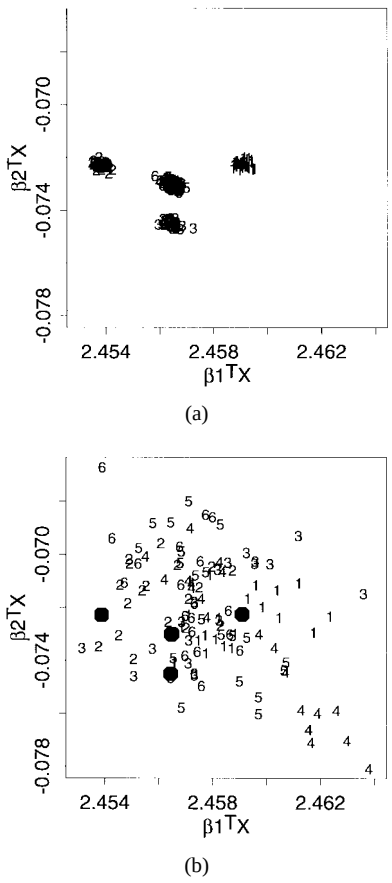


Fig. 6. First two LDA discriminant variables: (a) training sample: groups 1, 2, and 3 are well separated from each other and the other three groups (bunched together in the center) and (b) test sample: locations of the training group centroids are given by large dots.

$||\beta_1||$ was about 30 times as big for LDA as for PDA. This, along with the extreme difference in smoothness between the two methods, suggests that PDA has effectively prevented the β's from wandering off into the tips of the constraint ellipsoid.

The four plots in Figs. 6 and 7 further illustrate the benefit of penalization. Fig. 6(a) is a scatter plot of $\beta_2^T X$ versus $\beta_1^T X$ for the training sample and using LDA. The numbers denote the species labels Y_i, where for clarity, the numbers 1–6 are used to label species according to alphabetical order (i.e., $DF = 1, \cdots, WF = 6$). Clearly, β_1 and β_2 only separate the first three species (as one would expect given this information; β_3, β_4, and β_5 address the remaining three species). But the separation is extreme. Fig. 6(b) is the test sample analog of Fig. 6(a). The scales of the axes do not match, but shaded circles mark the location of the training sample centroids. Clearly, the classes are considerably mixed up on the test set. While this 2-D projection does not tell the entire story of the classification rule (which incorporates all five directions jointly), it is suggestive of the extent to which LDA has overfit the training data.

Fig. 7 is the PDA analog of Fig. 6. Again, we take PDA with Ω_D and $\log_{10}\lambda = -0.5$. When comparing the training sample discriminant variables between LDA and PDA, we see that the general orientation is the same, but the PDA classes are less separated, and the within class variation is larger. Consequently, by resisting the overfit, PDA performs better on the test sample, which can be seen by the closer agreement between the 1's, 2's, and 3's and their respective centroids in Fig. 7(b).

Since the PDA β's are less influenced by training sample noise, we can hope that they (or at least the first few that explain the majority of the variability) have meaningful physical interpretations. Rewriting the left term in the argument of (2), the distance in discriminate space of a test sample $(x_1, \cdots, x_K)$ from the jth class mean is just

$$\sum_{i=k}^{d}\sum_{k=1}^{K}\beta_i(k)^2(x_k - M_j(k)). \tag{7}$$

Therefore, coordinates of the β_i vectors that are large in magnitude correspond to bands that are influential in distancing an observation from a potential class.

Focusing on the solid lines in Fig. 5(a), we see that there are no large bands above about 750 nm, which agrees with our understanding of the spectral information just beyond the red edge. β_2 has some moderate weights in this range but no very large weights. β_3 through β_5 (not shown) are similar. Elsewhere, we see the largest individual weights at approximately 350, 430, 480, 700, and 730 nm for β_1 and at 415, 610, and 725 nm for β_2. Most of those spectral bands are in the blue spectral range (400–500 nm) and the

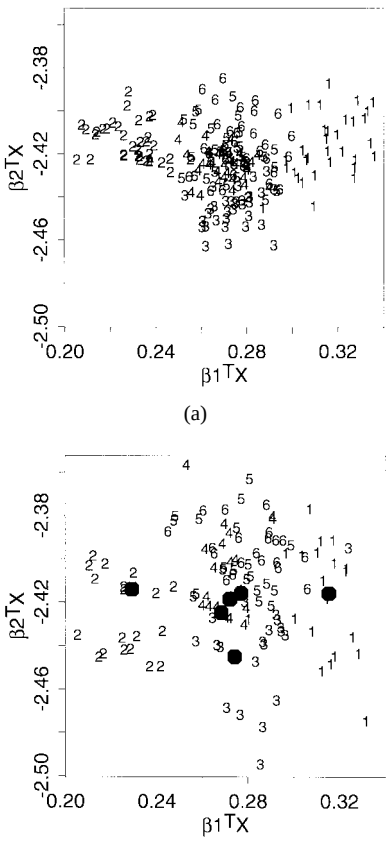


Fig. 7. First two discriminant variables for PDA with Ω_D and $\log_{10}\lambda = -0.5$: (a) training sample and (b) test sample: locations of the training group.

red (600–700 nm) and red edge (670–740 nm) regions. This information could be useful for selecting a much smaller subset of bands for classification in the event that collecting hyperspectral data is not feasible. Of course, in that situation one might also be interested in *regions* of fairly large $|\beta_i(k)|$ (such as around 615 nm for β_1) rather than large individual weights, since we expect a physically meaningful location to be more of a regional than pointwise phenomenon. Distilling such potentially useful information from the erratic LDA β's seems unlikely. The black box of neural networks is similarly not helpful when such physical interpretation is desired.

V. Conclusions

Emphasizing a geometric point of view, we describe a novel, nonparametric statistical classification technique known as PDA [13]. The geometry sheds light on how, in the proper context, penalization is able to improve substantially upon LDA. A simulation further demonstrates the dire consequences of poorly estimating Σ_W and how these are mitigated by penalization. Finally, an analysis of our *in situ* hyperspectral data on six conifer species demonstrates the possible benefits of PDA. The question of how well PDA can identify tree species in general is not answered by a single example on young conifers. However, our findings suggest that PDA is worthy of attention from researchers looking for an accurate, easy-to-use method of classification.

We employed PDA with two types of penalties and varying smoothing parameter. Our results suggest that derivative style penalties (5) are preferable to shrinkage style penalties (6) in that the former are more resistant to poor choices of smoothing parameter λ. With a well-chosen smoothing parameter, PDA classifies about twice as well as LDA: 76.7% to 38.3%. Moreover, when we split our data so that trees from the test site are included in the training set, our classification accuracy is around 90%, which is similar to previous results using neural networks under similar conditions [12]. Thus, PDA's accuracy seems comparable to neural networks. However, unlike the complicated nonlinear classifier, PDA is useful for data reduction, and the "principle component" directions are physically interpretable as directions where the important spectral bands for classification are emphasized. This interpretability may be useful in the selection of subsets of bands for classification, which we hope to address in the future.

Acknowledgment

The authors are grateful to M. Hansen of Bell Labs, Lucent Technologies, for many thought-provoking discussions. They also wish to thank the two anonymous reviewers for their highly valuable comments.

References

[1] P. Curran, "Remote sensing of foliar biochemistry," *Remote Sens. Environ.*, vol. 30, pp. 271–278, Jan. 1989.
[2] D. H. Card, D. L. Peterson, and P. A. Matson, "Prediction of leaf chemistry by use of visible and near infrared reflectance spectroscopy," *Remote Sens. Environ.*, vol. 26, no. 2, pp. 123–147, 1988.
[3] L. F. Johnson, C. A. Hlavka, and D. L. Peterson, "Multivariate analysis of aviris data for canopy biochemical estimation along the Oregon transect," *Remote Sens. Environ.*, vol. 47, no. 2, pp. 216–230, 1994.
[4] P. A. Matson, L. F. Johnson, J. R. Miller, C. R. Billow, and R. Pu, "Seasonal patterns and remote spectral estimation of canopy chemistry across the Oregon transect," *Ecol. Applicat.*, vol. 4, no. 2, pp. 280–298, 1994.
[5] A. F. H. Goetz, G. Vane, J. E. Solomonson, and B. N. Rock, "Imaging spectrometry for earth remote sensing," *Science*, vol. 228, pp. 1147–1153, 1985.
[6] J. R. Miller, J. Wu, M. G. Boyer, M. Belanger, and E. W. Hare, "Season patterns in leaf reflectance red edge characteristics," *Int. J. Remote Sensing*, vol. 12, no. 7, pp. 1509–1523, 1991.
[7] C. Daughtry, L. Biehl, and K. Ranson, "A new technique to measure the spectral properties of conifer needles," *Remote Sens. Environ.*, vol. 33, no. 1, pp. 55–64, 1989.
[8] S. N. Gonward, K. Huemmrich, and R. Waring, "Visible-near infrared spectral reflectance of landscape components in western Oregon," *Remote Sens. Environ.*, vol. 47, no. 2, pp. 190–203, 1994.
[9] D. Williams, "A comparison of spectral reflectance properties at the needle branch and canopy level for selected conifer species," *Remote Sens. Environ.*, vol. 35, no. 2/3, pp. 79–91, 1991.
[10] S. Liang and A. Strahler, "An analytic brdf model of canopy radiative transfer and its inversion," *IEEE Trans. Geosci. Remote Sensing*, vol. 31, pp. 1081–1092, Sept. 1993.
[11] X. Li and A. Strahler, " Geometric-optical bidirectional reflectance modeling of a coniferous forest canopy," *IEEE Trans. Geosci. Remote Sensing*, vol. 24, pp. 906–919, Nov. 1986.
[12] P. Gong, R. Pu, and B. Yu, "Conifer species recognition: An exploratory analysis of *in situ* hyperspectral data," *Remote Sens. Environ.*, vol. 62, pp. 189–200, Nov. 1997.
[13] T. Hastie, A. Buja, and R. Tibshirani, "Penalized discriminant analysis," *Ann. Statist.*, vol. 23, no. 1, pp. 73–102, 1995.
[14] A. Mazer, M. Martin, M. Lee, and J. Solomon, "Image processing software for imaging spectrometry data analysis," *Remote Sens. Environ.*, vol. 24, no. 1, pp. 201–210, 1988.
[15] J. Ott and R. Kronmal, "Some classification procedures for multivariate binary data using orthogonal functions," *J. Amer. Statist. Assoc.*, vol. 71, pp. 391–399, 1976.

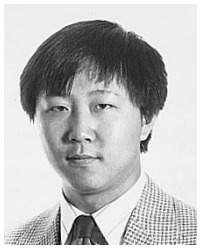
[16] H. Bensmail and C. Gilles, "Regularized Gaussian discriminatory analysis through eigenvalue decomposition," *J. Amer. Statist. Assoc.*, vol. 91, pp. 1743–1748, 1996.
[17] I. ANCAL, "C-SPEC data acquisition and manipulation program," *Users Guide and Operating Instructions, Version 1.5*. Las Vegas, NV: ANCAL, 1995.
[18] D. Mardia, J. Kent, and J. Bibby, *Multivariate Analysis*. London, U.K.: Academic, 1979.
[19] J. H. Friedman, "Regularized discriminant analysis," *J. Amer. Statist. Assoc.*, vol. 84, pp. 165–175, 1989.
[20] J. Ramsey and B. Silverman, *Functional Data Analysis*. New York: Springer-Verlag, 1997.
[21] G. Strang, *Introduction to Applied Mathematics*. Wellesley, MA: Cambridge Press, 1986.
[22] B. Efron, "Estimating the error rate of a prediction rule: Improvement on cross-validation," *J. Amer. Statist. Assoc.*, vol. 78, pp. 316–330, 1983.

Bin Yu (SM'97) received the B.S. degree in mathematics in 1984 from Peking University, China and the M.S. and Ph.D. degrees in statistics from the University of California, Berkeley, in 1987 and 1990, respectively.

From 1990 to 1992, she was an Assistant Professor of statistics at the University of Wisconsin, Madison. In the Fall of 1991, she was a postdoctoral Fellow at the Mathematical Science Research Institute at Berkeley, and in the Spring of 1993, she was a Visiting Assistant Professor of statistics at Yale University, New Haven, CT. From July 1993 to June 1997, she was an Assistant Professor of statistics at the University of California, Berkeley. Currently, she is an Associate Professor of statistics at the University of California, Berkeley, as well as a member of Technical Staff at Bell Labs, Lucent Technologies, Murray Hill, NJ. She has broad research interests, which currently include minimum description length (MDL) principle and model selection, information theory, signal (image and voice) denoising and compression, Markov chain Monte Carlo methods, and discriminant analysis based on curve data and its applications in remote sensing.

Dr. Yu is an Associate Editor for *The Annals of Statistics and Statistical Sinica*. She was on the program committee for the IEEE International Symposium on Information Theory at the Massachusetts Institute of Technology, Cambridge, in August, 1998, and for the 8th International Workshop on Algorithmic Learning Theory (ALT'97), Sendai, Japan, October 1997.

I. Michael Ostland is currently pursuing the Ph.D. degree in the Department of Statistics, University of California, Berkeley, where his primary research focuses on statistical problems in highway transportation. His other interests include classification, variable selection, and Markov chain Monte Carlo methods.



Peng Gong received the B.Sc. and M.Sc. degrees from Nanjing University, China, in 1984 and 1986, respectively, and the Ph.D. degree from the University of Waterloo, Waterloo, Ont., Canada, in 1990.

Currently, he is an Associate Professor of remote sensing and GIS and Director of the Center for Assessment and Monitoring of Forest and Environmental Resources at the University of California, Berkeley. From 1991 to 1994, he taught in the Department of Geomatics Engineering at the University of Calgary, Alta., Canada. His research interests are in photoecometrics, change detection, use of GIS, and remote sensing in epidemiology.

Dr. Gong is the author/coauthor of more than 120 technical papers and four books, a winner of three Best Paper Awards from the American Society for Photogrammetry and Remote Sensing, and the winner of an overseas Outstanding Young Scientist Award from the NSF of China. He is the Chief Editor for *Geographic Information Sciences*, an editor for *International Journal of Remote Sensing*, and is on the editorial board of *Journal of Remote Sensing* in China.

Ruiliang Pu received the B.Sc. degree and the M.Sc. degree from Nanjing Forestry University, China, in 1982 and 1985, respectively.

Until 1990, he was an Assistant Professor at Nanjing Forestry University. He was a visiting Scholar between 1990–1991 at the Earth-Observations Laboratory, Institute for Space and Terrestrial Science (ISTS), North York, Ont., Canada. Since 1993, he has been an Associate Professor at Nanjing Forestry University. He was a visiting Scientist in the Department of Geomatics Engineering, University of Calgary, Alta., Canada, in 1994, and since 1995, he has been a Research Associate in the Department of Environmental Science, Policy, and Management, University of California, Berkeley. His research has focused on applications of remote sensing and GIS technologies in forest resource management, ecosystem modeling and classification, and extracting biophysical and biochemical parameters with hyperspectral remotely sensed data.