Last revised: June 16, 2003

A Note on Linear Classifiers

James H. Stock
Department of Economics
Harvard University

1. Introduction and Notation

This note provides background information for the authorship attribution study of Stock and Trebbi (2003), but it might be of broader interest as well.

The classification problem is to assign an object of unknown type, or class, to one of several possible classes. This note considers the two-class classification problem (class 1 and 2), where the object to be classified is characterized by a *k*×1 vector *X* of observed attributes. The class of the attribute is summarized by the variable *Y*:

$$Y = \begin{cases} +1 \text{ if the object is in class 1} \\ -1 \text{ if the object is in class 2} \end{cases}$$

The distributions of the attributes *X* differ between the two classes, and it is because of this difference that it is possible to use *X* to assist in classification. Denote the distribution of *X* in the two classes by:

$$X \sim \begin{cases} \text{If } X \text{ is in class 1:} \\ \qquad f_1 \text{ with mean } \mu_1 \text{ and variance } E[(X-\mu_1)(X-\mu_1)' \,|\, Y=+1) = \Sigma_1 \\ \text{If } X \text{ is in class 2:} \\ \qquad f_2 \text{ with mean } \mu_2 \text{ and variance } E[(X-\mu_2)(X-\mu_2)' \,|\, Y=-1) = \Sigma_2. \end{cases} \tag{1}$$

A classifier is said to be linear if it is a linear function of *X*.

It is assumed that there is an estimation or "training" data set of n_c observations known to belong to class c. Denote the estimation data set by (X_i, Y_i), $i = 1,\ldots, n$, and let n_1 and n_2 denote the number of observations from each class. Let $\boldsymbol{X}$ be the $n\times k$ matrix $[X_1 \ldots X_n]'$ and let $\boldsymbol{Y}$ be the $n\times 1$ vector $[Y_1 \ldots Y_n]'$. For example, in Stock and Trebbi (2003), the object to be classified is a text of unknown authorship, the two classes are the two potential authors, the attributes are up to $k = 87$ stylometric indicators, and the estimation data set consists of $n_1 = 25$ observations on texts known to be written by author 1 and $n_2 = 20$ observations on texts known to be written by author 2.

This note provides a brief exposition of four classifiers: the Gaussian Bayes classifier, Fisher's linear discriminant, simple OLS regression, and principal components regression. Decision theory tells us that the optimal classifier assigns the object to the class that has the greatest posterior probability. As is shown in the next section, when f_1 and f_2 are normal distributions, the optimal Gaussian Bayes classifier has a simple form. However, the Gaussian Bayes classifier is not feasible when the number of attributes (k) exceeds the number observations (n) in the estimation or "training" set. In contrast, two of the classifiers considered in this note, Fisher's linear discriminant and principal components regression, are feasible even if $k \geq n$. The remaining classifier, unrestricted linear regression, is not feasible when $k \geq n$, but provides a familiar benchmark based on standard econometric tools.

Although these four classifiers have different motivations and their algorithms appear to be quite different, in fact they are closely related. In fact, under certain conditions these classifiers are equivalent, at least in large samples. More precisely, the main relationships among the various methods are:

1. If X is normally distributed, then the optimal Bayes ("Gaussian Bayes") classifier assigns the unknown object to the closest class (based on the Mahalanobis distance), using a cutoff that involves Σ_1, Σ_2, and the prior probabilities of being in class 1 or 2.

2. If $\Sigma_1 = \Sigma_2$, then:
   - the Gaussian Bayes classifier is linear;

- Fisher's linear discriminant (in its general form) is asymptotically equivalent to the Gaussian Bayes classifier when the prior probability of being in one class or the other is the same;
- if in addition $n_1 = n_2$, the Fisher linear discriminant weights are proportional to the OLS regression weights in the linear regression $Y_i = \beta_0 + \beta' X_i + u_i$, from which it follows that the linear regression classifier is asymptotically equivalent to the Gaussian Bayes classifier.

3. If $\Sigma_1 = \Sigma_2 = \sigma^2 I_k$, then:
   - The largest eigenvector of $\boldsymbol{X}'\boldsymbol{X}$ is approximately proportional to the linear regression weights;
   - If in addition $n_1 = n_2$, and n and k are large, classification based on the first principal component approaches classification based on the Gaussian Bayes classifier.

The bottom line, then, is that these linear classifiers are related to the optimal Gaussian Bayes classifier, and in fact equal (or equal asymptotically) the Gaussian Bayes classifier if some additional conditions, such as $\Sigma_1 = \Sigma_2 = \sigma^2 I_k$, hold. These conditions presumably do not hold in a given application, so in this sense the different classifiers are only approximations to the optimal Gaussian Bayes classifier. The advantage of Fischer's linear discriminant and the principal components classifier over the Bayes classifier is that they are feasible when k is of the same order as, or greater than, n, whereas the Bayes classifier is not. The disadvantage of these linear classifiers is that, if the data are strongly nonGaussian, they can perform quite poorly relative to nonlinear classifiers; see Devroye, Györfi, and Lugosi (1991, Section 4.3).

The exposition here is not rigorous, rather, the purpose is to explain these procedures and their relation to the optimal Bayes classifier. Although the discussion focuses on the two-class problem, the conceptual framework, the optimal Bayes classifier and the principal components classifier generalize in a straightforward way to the multiclass classification problem. The multiclass generalization of Fischer’s linear discriminant is canonical discriminant analysis.

This primer draws on Duda and Hart (1973). For a more advanced treatment focusing on nonlinear classifiers when $k << n$, see Devroye, Györfi, and Lugosi (1991).

2. Optimal Bayes Classifiers

Let π_c be the prior probability of being in class c. Given the distributions of X, the ratio of the posterior odds of being in class 2, relative to class 1, is

$$\frac{P(Y=-1\,|\,X)}{P(Y=+1\,|\,X)}=\frac{f_2(X)}{f_1(X)}\times\frac{\pi_2}{\pi_1}\,. \tag{2}$$

where the distributions f_1 and f_2 are treated as known.

If f_1 and f_2 are k-dimensional normal distributions, then the log Bayes ratio is,

$$\begin{aligned} r(X) &= \ln[f_2(X)/f_1(X)] \\ &= ½[(X-\mu_1)'\Sigma_1^{-1}(X-\mu_1)-(X-\mu_2)'\Sigma_2^{-1}(X-\mu_2)+\ln(|\Sigma_2|/|\Sigma_1|)], \end{aligned} \tag{3}$$

where $|\Sigma_1|$ is the determinant of Σ_1. The optimal Bayes classifier places the object in class 2 if $r(X) > \ln(\pi_1/\pi_2)$; with equal prior odds, it places the object in class 2 if $r > 0$.

One useful way to rewrite r is,

$$r(X) = ½[d_1(X,\mu_1) - d_2(X,\mu_2) + \ln(|\Sigma_2|/|\Sigma_1|)] \tag{4}$$

where $d_c(X,\mu_c)$ is the Mahalnobis distance from X to μ_c,

$$d_c(X,\mu_c) = (X-\mu_c)'\Sigma_c^{-1}(X-\mu_c). \tag{5}$$

A second useful way to rewrite r is,

$$r(X) = w_0 + w'X + X'\mathrm{A}X, \qquad (6)$$

where

$$w_0 = ½[\mu_1'\Sigma_1^{-1}\mu_1 - \mu_2'\Sigma_2^{-1}\mu_2 + \ln(|\Sigma_2|/|\Sigma_1|)]$$

$$w = \Sigma_2^{-1}\mu_2 - \Sigma_1^{-1}\mu_1, \text{ and}$$

$$A = ½(\Sigma_1^{-1} - \Sigma_2^{-1}).$$

As is clear from (6), in general the Gaussian Bayes classifier is quadratic in X. It is useful, however, to consider three cases in which the Gaussian Bayes classifier is linear.

Case 1: $\Sigma_1 = \Sigma_2 = \Sigma$.

In this case, A = 0 so the Gaussian Bayes classifier is linear, specifically,

$$r(X) = w_0 + w'X, \qquad (7)$$

where

$$w = \Sigma^{-1}(\mu_2 - \mu_1). \qquad (8)$$

Case 2: $\Sigma_1 = \Sigma_2 = \mathrm{diag}(\sigma_1^2,\ldots,\sigma_k^2)$.

In this case, the Gaussian Bayes classifier is given by (7) with

$$w_j = \frac{\mu_{j,2} - \mu_{j,1}}{\sigma_j^2}, j = 1,\ldots, k, \qquad (9)$$

where $\mu_{j,c}$ is the mean of attribute j in class c.

Case 3: $\boldsymbol{\Sigma_1 = \Sigma_2 = \sigma^2 I_k}$.

In this case, the Gaussian Bayes classifier is given by (7) with

$$w = (\mu_2 - \mu_1)/\sigma^2, \tag{10}$$

so the optimal linear classifier is proportional to $(\mu_2 - \mu_1)'X$.

3. Fisher's Linear Discriminant

Fisher's linear discriminant is the linear combination $\omega'X$ that maximizes the ratio of its "between" sum of squares to its "within" sum of squares. That is, ω solves,

$$\max_\omega J(\omega), \text{ where } J(\omega) = \frac{(\omega'\mu_2 - \omega'\mu_1)^2}{\omega'\Sigma_1\omega + \omega'\Sigma_2\omega} = \frac{\omega'\Sigma_B\omega}{\omega'\Sigma_W\omega}, \tag{11}$$

where Σ_W and Σ_B are the "within" and "between" variance matrices,

$$\Sigma_B = (\mu_2 - \mu_1)(\mu_2 - \mu_1)' \text{ and } \Sigma_W = \Sigma_1 + \Sigma_2. \tag{12}$$

The maximization problem (11) is a generalized eigenvalue problem. Because Σ_B has rank 1, it can be verified that the solution to this problem is proportional to

$$\omega = \Sigma_W^{-1}(\mu_2 - \mu_1), \tag{13}$$

This expression simplifies further in the following case:

Case 4: $\boldsymbol{\Sigma_1 = \text{diag}(\sigma_{1,1}^2, \ldots, \sigma_{k,1}^2)}$ and $\boldsymbol{\Sigma_2 = \text{diag}(\sigma_{1,2}^2, \ldots, \sigma_{k,2}^2)}$.

In this case, ω in (13) becomes,

$$\omega_j = \frac{\mu_{j,2} - \mu_{j,1}}{\sigma^2_{j,1} + \sigma^2_{j,2}}, \tag{14}$$

which is the expression for the linear discriminant weight used by Mosteller and Wallace (1963). If in addition $\Sigma_1 = \Sigma_2$, then this case becomes Case 3 above, and comparison of (14) and (9) shows that the linear discriminant weights are proportional to the Gaussian Bayes weights. Thus, in Case 3 Fisher's linear discriminant is equivalent to the Gaussian Bayes classifier.

If $k < n_1$, n_2, then it is feasible to estimate to estimate the full "within" matrix, which leads to the empirical linear discriminant weights in which the population means and variances in (13) are replaced by their sample counterparts:

$$\hat{\omega} = \hat{\Sigma}_W^{-1}(\hat{\mu}_2 - \hat{\mu}_1), \tag{15}$$

where $\hat{\Sigma}_W = \hat{\Sigma}_1 + \hat{\Sigma}_2$, where

$$\hat{\mu}_c = \frac{1}{n_c}\sum_{i \in \text{class } c} X_i \text{ and } \hat{\Sigma}_c = \frac{1}{n_c}\sum_{i \in \text{class } c} (X_i - \hat{\mu}_c)(X_i - \hat{\mu}_c)', \; c = 1, 2. \tag{16}$$

If $k \geq n_1$ or n_2, then (15) is infeasible ($\hat{\Sigma}_c$ is singular) and instead a practical solution is to ignore the off-diagonal terms, that is, to implement the formula in Case 4 as,

$$\hat{\omega}_j = \frac{\hat{\mu}_{j,2} - \hat{\mu}_{j,1}}{\hat{\sigma}^2_{j,1} + \hat{\sigma}^2_{j,2}}, \tag{17}$$

where the population means and variances in (14) are replaced by their sample counterparts. This is the version implemented in Stock and Trebbi (2003) and (using medians and interquartile ranges rather than means and variances) by Mosteller and Wallace (1963).

4. Unrestricted OLS Regression

The unrestricted linear regression classification model is,

$$Y_i = \beta_0 + \beta' X_i + u_i. \tag{18}$$

Without loss of generality, let $X_1,\ldots, X_n$ be standardized using all n observations, so that $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = 0$ and the diagonal elements of $\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})'$ all equal 1.

The OLS estimator satisfies the first order conditions, $(\boldsymbol{X'X}/n)\hat{\beta} = \boldsymbol{X'Y}/n$. It is readily verified that $\boldsymbol{X'Y}/n = p_2\hat{\mu}_2 - p_1\hat{\mu}_1$ and

$$\boldsymbol{X'X}/n = p_1\hat{\Sigma}_1 + p_2\hat{\Sigma}_2 + b(p_2\hat{\mu}_2 - p_1\hat{\mu}_1)(p_2\hat{\mu}_2 - p_1\hat{\mu}_1)' \tag{19}$$

where $p_1 = n_1/n$, $p_2 = n_2/n$, and $b = (2p_1p_2)^{-1}$. Thus the OLS estimator satisfies,

$$[p_1\hat{\Sigma}_1 + p_2\hat{\Sigma}_2 + b(p_2\hat{\mu}_2 - p_1\hat{\mu}_1)(p_2\hat{\mu}_2 - p_1\hat{\mu}_1)']\hat{\beta} = p_2\hat{\mu}_2 - p_1\hat{\mu}_1. \tag{20}$$

Consider the special case of $n_1 = n_2$ (so $p_1 = p_2 = ½$). Then the empirical Fisher linear discriminant weights $\hat{\omega}$, defined in (15), are proportional to the OLS regression weights in the unrestricted regression. (This is shown by substituting $\hat{\beta} = d\hat{\omega}$ into (20), where $\hat{\omega}$ is defined in (15) and $d = [\,1 + (\hat{\mu}_2 - \hat{\mu}_1)'\hat{\Sigma}_B^{-1}(\hat{\mu}_2 - \hat{\mu}_1)]^{-1}$.) Thus, under Case 1 with $n_1 = n_2$, the linear predictor constructed using unrestricted OLS is equivalent to the general form of the linear discriminant. It follows that the linear regression classifier is asymptotically equivalent to the Gaussian Bayes classifier as $n \to \infty$, $n_1/n_2 \to 1$, and k is fixed.

5. Principal Components Regression

In practice, k can be too large to make unrestricted regression feasible. If so, one method for reducing the dimension of the regressors is principal components. As above, we assume that the regressors are standardized (this assumption is *not* without loss of generality, but it is conventional when computing principal components). The principal components are the linear combinations formed by the eigenvectors of $\boldsymbol{X}'\boldsymbol{X}$. The first principal component corresponds to the largest eigenvector.

The key insight of principal components regression is that the eigenvector of $\boldsymbol{X}'\boldsymbol{X}$ corresponding to its largest eigenvector will tend towards the direction $p_2\hat{\mu}_2 - p_1\hat{\mu}_1$, at least when k is large. This follows from the identity (19). Let $\alpha = p_2\hat{\mu}_2 - p_1\hat{\mu}_1$; then from (19),

$$\alpha'(\boldsymbol{X}'\boldsymbol{X}/n)\,\alpha = \alpha'[p_1\hat{\Sigma}_1 + p_2\hat{\Sigma}_2]\alpha + b(\alpha'\alpha)^2. \qquad (21)$$

The first term on the right hand side of (21) satisfies $\alpha'[p_1\hat{\Sigma}_1 + p_2\hat{\Sigma}_2]\alpha \leq$ $(\alpha'\alpha)[p_1\text{mineval}(\hat{\Sigma}_1) + p_2\text{mineval}(\hat{\Sigma}_2)] = O(k)$, because $\alpha'\alpha = O(k)$ and by assumption the eigenvalues of Σ_1 and Σ_2 are bounded. Similarly, the second term on the right hand side of (21) is $O(k^2)$, so $\alpha'(\boldsymbol{X}'\boldsymbol{X}/n)\alpha = O(k^2)$. However, any other linear combination (not in the direction α) will be $O(k)$. This heuristic reasoning suggests that, for k large, the largest eigenvector of $\boldsymbol{X}'\boldsymbol{X}$ will tend towards the direction α.

Accordingly, if in addition $n_1 = n_2$, the first principal component is approximately

$$\textit{first principal component} \approx (\hat{\mu}_2 - \hat{\mu}_1)'X. \qquad (22)$$

The results (22) and (10) suggest that the first principal component is, in large samples, essentially proportional to the optimal Gaussian Bayes classifier in Case 3 if $n_1/n_2 \to 1$, $n \to \infty$, and $k \to \infty$.

Formalizing the argument in this section is rather involved technically because it entails limits in both n and k. I am not aware of a reference that works out the details. One route towards proving this is to build on the factor model/principal component consistency results in Ding and Hwang (1999) and Stock and Watson (2002).

In practice, the first principal component will not be exactly proportional $\mu_2 - \mu_1$ for sampling reasons. Also, the strong assumptions of Case 3 might not hold, so the Gaussian Bayes classifier weights might not be proportional to $\mu_2 - \mu_1$ in any event. For these reasons, when $k \geq n$, it makes sense to compute the first few (m) principal components and use them as regressors, or to compute the Gaussian Bayes classifier for this reduced m-dimensional vector of attributes constructed from the full set of k attributes.

References

Devroye, L., L. Györfi, and G. Lugosi. 1991. *A Probabilistic Theory of Pattern Recognition*. New York: Spring.

Ding, A.A. and J.T. Hwang. 1999. "Prediction Intervals, Factor Analysis Models, and High-Dimensional Empirical Linear Prediction," *Journal of the American Statistical Association*, 94, 446 – 455.

Duda, Richard O. and Peter E. Hart. 1973. *Pattern Classification and Scene Analysis*. New York: Wiley.

Mosteller, Frederick and David L. Wallace. 1963. "Inference in an Authorship Problem." *Journal of the American Statistical Association*, 58, pp. 275–309.

Stock, James H. and Mark. W. Watson. 2002. “Forecasting Using Principal Components from a Large Number of Predictors,” *Journal of the American Statistical Association*, 97, pp. 1167–1179.

Stock, James H. and Francesco Trebbi. 2003. “Who Invented Instrumental Variable Regression?” *Journal of Economic Perspectives*, September 2003.