Comment

Matthew Goldstein

Mark Schervish is to be applauded for providing readers an opportunity to reflect on how far multivariate analysis has come while at the same time giving us his thoughtful views on the direction theoreticians and practitioners of statistics appear to be headed. To take a body of knowledge as vast as multivariate analysis and attempt to distill its most salient features in the limited space of his article, takes, at the very least, much courage. By restricting his universe of material to that contained within the pages of the two books forming the basis of his discussion, he successfully presents a balanced overview of many of the important developments in multivariate analysis. Although I am pleased by the selection and in particular of his generally kind words of the text by Dillon and Goldstein there are critically important omissions. An article at least the size of that presented by Schervish would be required. Because this is not practical I will limit my remarks to just a few areas.

1. DISCRIMINANT ANALYSIS

Schervish comments on the sparse treatment that error rate analysis is given (by both books) within the context of the classification problem. Aside from rather ad hoc methods like cross-validation and a rich body of material using asymptotic methods to approximate the actual and apparent error estimates of the true error (assuming a multivariate normal structure), most of the interesting work until quite recently assumed discrete multivariate data. Readers of Dillon and Goldstein were referred to an earlier book (1978) by the same authors where a full chapter was devoted to error rate analysis for the discrete classification problem. Recent work has added new insights and results.

Suppose we observe a set of data given as

$$x_1 = (\mathbf{t}_1, y_1),\ x_2 = (\mathbf{t}_2, y_2)\ \ldots,\ x_n = (\mathbf{t}_n, y_n),$$

where the $\mathbf{t}_i$ are observed p-dimensional covariate vectors and the y_i are independent binary variables such that $y_i \sim B(\pi_i)$. Let us further assume that the binomial parameters π_i are given by the logistic formula

$$\pi_i = 1/(1 + \exp(\mathbf{t}_i'\alpha)) \qquad i = 1, 2, \ldots, n, \tag{1.1}$$

where α is an unknown p-dimensional vector of parameters. Estimates π_i can be found by plugging in the maximum likelihood estimates for α. In the classical discrimination problem a new observation $\mathbf{t}_{n+1}$ is observed and a choice needs to be made regarding the value of y_{n+1}, that is, if $y_{n+1} = 1$, we say that the covariates come from group 1, otherwise from group 2.

Consider the sample-based prediction rule $\hat{\eta}$ given by

$$\hat{\eta}_i = \begin{cases} 1 & \text{if } \hat{\pi}_i > C, \\ 0 & \text{if } \hat{\pi}_i \le C, \end{cases} \tag{1.2}$$

for some given cutoff point C. The proportion of times this rule is in error is

$$\text{Err} = \#\{y_i \neq \hat{\eta}_i\}/n. \tag{1.3}$$

Err is commonly referred to as the apparent error rate. It is well known that the apparent error is optimistically biased because the data that are used to construct the rule are the same as those used to evaluate how well it performs.

Gong (1986) considered estimates of the excess error, the difference between the true and apparent errors. By using simulated and real data, she compared three estimates of the excess error—the jackknife, cross-validation and the bootstrap. Although the jackknife and cross-validation showed little improvement, the bootstrap substantially reduced the size of excess error.

Efron (1986), in the most far reaching and unified treatment to date, derived among other estimates, an estimate for the expected excess error $w(\pi)$ assuming the logistic formulation. His estimate is

$$w(\hat{\pi}) = 2/n \sum_{i=1}^{n} \hat{\pi}_i(1 - \hat{\pi}_i)\phi\left(\frac{\hat{C}_i}{\sqrt{\hat{d}_i}}\right)\sqrt{\hat{d}_i}, \tag{1.4}$$

where $\phi(\cdot)$ is the standard normal,

$$\hat{C}_i = \log(C/(1 - C)) - \mathbf{t}_i'\hat{\alpha},$$

$$\hat{d}_i = \mathbf{t}_i'\,\hat{\Sigma}\,\mathbf{t}_i, \quad \hat{\Sigma} = \sum_{i=1}^{n} \hat{\pi}_i(1 - \hat{\pi}_i)\mathbf{t}_i\mathbf{t}_i'.$$

$w(\hat{\pi})$ has the usual asymptotic optimality properties of maximum likelihood estimates. A small sampling experiment showed that $w(\hat{\pi})$ is nearly unbiased for $w(\pi)$ with quite small standard deviation. Practitioners using (1.2) can better assess the magnitude of the optimistic bias of Err by using $w(\hat{\pi})$.

Matthew Goldstein is President of the Research Foundation and Professor of Statistics at The City University of New York, 79 Fifth Avenue, New York, New York 10003.

2. PROJECTION PURSUIT METHODS

I was somewhat surprised to see that projection pursuit (PP) methods were not mentioned in Dr. Schervish's paper. In 1985, Huber pulled together many of the ideas of PP in a clear and cogent manner and established the beginning of a theoretical framework to study such methods.

Authors commenting after Huber's paper are quite valuable and some of their ideas are included here. PP techniques are tools for multivariate data analysis useful for finding interesting and useful low dimensional projections of higher dimensional data by maximizing (or minimizing) a certain projection index. Advocates of these tools proclaim their virtue in part because of the claim that it is one of the few multivariate methods able to bypass the "curse of dimensionality" caused by the fact that high dimensional space is mainly empty, manifesting itself in less robustness, increasing mean square error or slowing down convergence to limiting distributions. A further claim is that some PP methods are able to ignore information-poor variables giving them an advantage over methods based upon interpoint distances like multidimensional scaling and most clustering algorithms. Others refute these claims stating that the sample size may need to be very large relative to dimensionality to avoid the vexing problems associated with extreme sparseness. Secondly, although PP methods might be able to deal more effectively with "noisy" variables than interpoint distance methods, it does not appear to be the solution for handling large numbers of variables simultaneously.

A linear projection from R^d to R^k is any $k \times d$ matrix $\mathbf{A}$ of rank k:

$$\mathbf{Z} = \mathbf{AX}, \qquad \mathbf{X} \in R^d, \quad Z \in R^k.$$

The projection is orthogonal if the row vectors of $\mathbf{A}$ are orthogonal to each other and have length 1. If $\mathbf{X}$ is a d-dimensional random vector then $\mathbf{Z}$ is a k-dimensional vector. Denote its induced distribution function by F_A. When $k = 1$, $\mathbf{A}$ reduces to a row vector $\mathbf{a}'$ and we represent the distribution function by F_a. Projection pursuit searches for a linear projection $\mathbf{A}$ maximizing a certain projection index $Q(\mathbf{AX})$.

An interesting observation is that a number of classical multivariate techniques are special cases of PP including principal components, discriminant analysis and multiple linear regression. In the case of principal components suppose that $\mathbf{X} \sim F$ is a p-dimensional random vector with covariance matrix $\mathbf{V}$. Let $\mathbf{a}'$ be a p-dimensional random vector and let the linear projection $\mathbf{a}'\mathbf{X}$ have distribution function F_a. Denote the eigenvalues of $\mathbf{V}$ by $v_1, v_2, \ldots, v_p$ and let $\sigma(F_a)$ be the standard deviation of $\mathbf{a}'\mathbf{X}$. Recall that the first principal component is that linear projection $\mathbf{a}'\mathbf{X}$ which satisfies

$$\sigma(F_{a_1}) = \max_{|\mathbf{a}|=1} \sigma(F_a) = \max_{|\mathbf{a}|=1} (\mathbf{a}'\mathbf{va})^{1/2}$$

and that $\sigma^2(F_{a_1}) = \mathbf{a}_1'\mathbf{va}_1$ is the largest eigenvalue v_1 associated with $\mathbf{a}_1$. Further, the second principal component is determined by

$$\sigma(F_{a_2}) = \max_{|\mathbf{a}|=1, \mathbf{a} \perp \mathbf{a}_1} \sigma(F_a)$$

or

$$\sigma^2(F_{a_2}) = v_2.$$

Thus, the search for principal components is the search for low dimensional linear projections that maximize the projection index.

In projection pursuit regression (PPR) the vector directions $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_M$ are given and we wish to find projection functions, $f_1, f_2, \ldots, f_M$ that minimize the expected residual sum of squares

$$E(r_M^2) = E\left(Y - \sum_{j=1}^{M} f_j(\mathbf{a}_j'\mathbf{x})\right)^2. \tag{2.1}$$

It is straightforward to see that (2.1) is minimized if for each $k = 1, 2, \ldots, M$

$$E(Y \mid \mathbf{a}_k'\mathbf{x} = z) = E\left(\sum_{j=1}^{M} f_j(\mathbf{a}_j\mathbf{x}) \;\middle|\; \mathbf{a}_k'\mathbf{x} = z\right)$$

or

$$E\left(Y - \sum_{j \neq k} f_j(\mathbf{a}_j'\mathbf{x}) \;\middle|\; \mathbf{a}_k'\mathbf{x} = z\right) = f_k(z).$$

This is a system of linear equations for the f_j that can be solved by using Gauss-Seidel.

Friedman's comments after Huber's paper relating to implementation of projection pursuit procedures are particularly valuable. He emphasizes quite thoughtfully that the performance judgment of a new method will be based upon how well the complete implementation performs, and as "data algorithms become more complex, this problem becomes more acute. The best way to guard against this is to become as literate as possible in algorithms, numerical methods and other aspects of software implementation." In some sense the jury is still out on how profound the impact PP methods will be in analyzing multivariate data. Statisticians have to broaden their experience with these tools, examining and reflecting upon their advantages and disadvantages with a variety of real data and simulated problems.

3. DENSITY ESTIMATION

Nonparametric estimation of density functions is an important problem finding useful and interesting applications in nonparametric regression, classifica-

tion and hypothesis testing. Perhaps the best known and most widely studied are a class of estimates called kernel estimates defined for $\mathbf{x} \in R^d$ by

$$f_n(\mathbf{x}) = \frac{1}{nk_n^d} \sum_{i=1}^{n} K\left(\frac{\mathbf{x} - \mathbf{x}_i}{h_n}\right) \tag{3.1}$$

where the kernel $K(\cdot)$ is an arbitrary density and $\{h_n\}$ is a sequence of positive numbers often referred to as smoothing parameters or bandwidths having the properties

$$h_n \to 0 \quad \text{and} \quad nh_n^d \to 0 \quad \text{as} \quad n \to \infty.$$

For a good overview see Devroye and Györfi (1985) and Silverman (1986). Most of the research on kernel estimates revolve around conditions under which the sequence $f_n(\cdot)$ satisfies different consistency properties, "optimal" choices for the kernel $K(\cdot)$ and smoothing parameters h_n. Different regularity conditions ensure that $f_n(\cdot)$ is a consistent sequence of estimates in the sense that the L_1 error tends to zero in different modes. Perhaps the most critical problem is choosing the smoothing parameters, changes to which have profound impact on the estimate. Kernel estimates require large sample sizes especially for high dimensional problems, otherwise they may be severely biased.

Huber claims that some of the problems using kernel estimates can be avoided by using projection pursuit density estimation. The literature is quite limited in this regard and much more experience will be needed before more definitive comparisons can be made. For a good discussion see Friedman, Stuetzle and Schroeder (1984). Their estimates take the form

$$P_M(\mathbf{x}) = P_0(\mathbf{x}) \prod_{m=1}^{M} f_m(\mathbf{a}_m'\mathbf{x}) \tag{3.2}$$

where P_M is the current density estimate after M iterations; P_0 is a given multivariate density function to be used as the initial model; $\mathbf{a}_m'$ is a vector of directions in R^d and f_m is a univariate function. In their examples P_0 was chosen as a multivariate normal with sample mean vector and sample covariance matrix. From (3.1) they use the recursive relation

$$P_M(\mathbf{x}) = P_{M-1}(\mathbf{x}) f_M(\mathbf{a}_M'\mathbf{x}). \tag{3.3}$$

Thus, given $P_{M-1}(\mathbf{X})$ we seek a new model $P_M(\mathbf{X})$ to serve as a better approximation to the data density $P(\mathbf{X})$. Relative goodness of fit is measured by the cross-entropy term of the Kullback-Leibler distance.

CONCLUSION

A basic criticism whether overt or below the surface of Anderson and Dillon and Goldstein is what Schervish calls the "α level mindset." I couldn't agree more. Once tests are available for hypotheses of interest there appears to be a propensity for the procedure to assume a life of its own with almost a preordained script to follow depending on the calculated p-value associated with a given test statistic. I would like to believe that most statisticians do not follow such myopic tendencies. However, with the wide availability of "packaged" programs to do most of the common multivariate techniques, we are at the very least setting ourselves up for benign abusive behavior. Notwithstanding some obvious lapses in tight writing that Dr. Schervish picked up (already corrected for a 2nd edition), I believe many of us who teach this subject matter should be mindful of how easy it is to go astray. I am however hopeful that, with students better trained in computation techniques, there will be a greater likelihood of individual experimentation and not sole reliance on neatly packaged techniques that for the uninitiated convey the impression that a set of values spewed out of a printer is the whole story.

ADDITIONAL REFERENCES

DEVROYE, L. and GYÖRFI, L. (1985). *Nonparametric Density Estimation: The L_1 View.* Wiley, New York.

EFRON, B. (1986). How biased is the apparent error rate of a prediction rule? *J. Amer. Statist. Assoc.* **81** 461–470.

FRIEDMAN, J. H., STUETZLE, W. and SCHROEDER, A. (1984). Projection pursuit density estimation. *J. Amer. Statist. Assoc.* **79** 599–608.

GOLDSTEIN, M. and DILLON, W. R. (1978). *Discrete Discriminant Analysis.* Wiley, New York.

GONG, G. (1986). Cross-validation, the jackknife, and the bootstrap: Excess error estimation in forward logistic regression. *J. Amer. Statist. Assoc.* **81** 108–113.

HUBER, P. J. (1985). Projection pursuit. *Ann. Statist.* **13** 435–475.

SILVERMAN, B. W. (1986). *Density Estimation for Statistics and Data Analysis.* Chapman and Hall, London.