FULLY NONPARAMETRIC EMPIRICAL BAYES ESTIMATION VIA PROJECTION PURSUIT*

M. Vernon Johns

Stanford University

The fully nonparametric formulation of the empirical Bayes estimation problem considers m populations characterized by conditional (sampling) distributions chosen indpendently by some unspecified random mechanism. No parametric constraints are imposed on the family of possible sampling distributions or on the prior mechanism which selects them. The quantity to be estimated subject to squared-error loss for each population is defined by a functional T(F) where F is the population sampling cdf. The empirical Bayes estimator is based on n iid observations from each population where n > 1. Asymptotically optimal procedures for this problem typically employ consistent nonparametric estimators of certain nonlinear conditional expectation functions. In this study a particular projection pursuit algorithm is used for this purpose. The proposed method is applied to the estimation of population means for several simulated sets and one familiar real world data set. Certain possible extensions are discussed.

1. Introduction.

The purpose of this paper is to show how an old idea may be effectively implemented using new technology. The old idea is the notion of fully nonparametric empirical Bayes estimation, which was introduced by the author in a paper (Johns, 1957) directly inspired by the fundamental paper of

The research was supported by the U.S. Office of Naval Research Grant Number N00014-83-K-0472.

AMS 1980 subject classifications. Primary 62G05; Secondary 62J02.

Key words and phrases. Empirical Bayes, nonparametric estimation, projection pursuit, nonlinear regression, multivariate estimation.

Robbins (1955). The new technique is computer based projection pursuit regression analysis.

The fully nonparametric appproach to empirical Bayes estimation differs from the original Robbins formulation in that it does not require the specification of a parametric family for the conditional (sampling) distributions of the independent component populations. Neither formulation makes parametric assumptions about the prior distribution of the quantity being estimated. This is in contrast to the case of "parametric" empirical Bayes estimation (see e.g., Efron-Morris, 1975) where parametric models are specified for both the conditional and prior distributions, and the "restricted" case where the estimators are constrained to have a particular simple form (see Robbins 1983). It should be noted that the fully nonparametric version of the problem requires that at least two observations be obtained from each component population.

When the empirical Bayes approach was first introduced, and for some time thereafter, it seemed that application of the methods to real world data would not often be feasible because of computational difficulties and the possibility that a very large number of component populations might be needed before approximately optimal results could be obtained. Indeed, one advantage of the parametric approach, or the restriction to linear forms of estimation, is the increased capacity to deal with real data sets of modest size at the cost of some potential loss of asymptotic efficiency. The original version of the fully nonparametric methodology (Johns, 1957) with which this paper is principally concerned, was of little practical use in a world where large scale digital computers had barely appeared on the scene. Fortunately, the present widespread availability of computational power and the development of sophisticated statistical software has opened up new possibilities.

One of the central requirements for dealing with the fully nonparametric empirical Bayes problem is the estimation of a conditional expectation function of unknown form involving several variables. In the original paper (Johns, 1957) a pointwise consistent estimator was proposed based on successive refinements of a partition of d-dimensional space. A convergence result (Lemma 5), which in a later incarnation has become known as the generalized Lebesgue dominated convergence theorem, was then used to show convergence to the Bayes optimal risk for the proposed empirical Bayes estimator. Some of these results could be regarded as primitive precursors of the more recent work of Stone (1977). In the last few years several other sophisticated methods for the nonparametric estimation of conditional expectation (regression) have been proposed. These include kernal smoothers, nearest neighbor estimates, recursive partitioning and notably, projection pursuit regression as proposed by Friedman and Stuetzle (1981). A comprehensive discussion of projection pursuit methods may be found in Huber (1985) where it is noted that, almost alone among multivariate procedures, they avoid many of the difficulties associated with high dimensionality and the presence of uninformative observations.

In the present study the regression aspect of the fully nonparametric empirical Bayes estimation procedure has been dealt with by substituting a projection pursuit regression scheme for the original conditional expectation estimator. The particular algorithm used is called The Smooth Multiple Additive Regression Technique (SMART) and is detailed in Friedman (1984). In section 2 the problem and the proposed solution are described more formally. In section 3 the proposed method is applied to several data sets generated by computer simulation and the results are discussed. The method is also applied to the famous Efron-Morris baseball data. Section 4 contains concluding remarks and acknowledgements.

2. The Problem and the Proposed Method.

We consider m populations from each of which n observations are obtained. Let these observations be given by

$$X_{ij} = \text{the i-th observation from the j-th population,}$$

$$i=1,2,\ldots,n; \quad j=1,2,\ldots,m.$$

We assume that for each j the X_{ij}'s are i.i.d. with common random cdf F_j, where $F_1,F_2,\ldots,F_m$ are assumed to be selected independently according to some unknown prior probability measure over all cdf's. Let T(F) = a real-valued functional defined on all cdf's which represents the "parameter" to be estimated for each population subject to squared-error loss, i.e., $\Theta_j = T(F_j)$, for any estimator $\hat{\Theta}_j$ the loss incurred is $(\hat{\Theta}_j - \Theta_j)^2$. If $\hat{\Theta} = (\hat{\Theta}_1,\hat{\Theta}_2,\ldots,\hat{\Theta}_m)$ then the average loss for the m component populations is

$$L(\hat{\Theta},\Theta) = (\hat{\Theta} - \Theta)(\hat{\Theta} - \Theta)'/m. \tag{1}$$

The corresponding average risk is then

$$R(\hat{\Theta}) = E\{L(\hat{\Theta},\Theta)\}, \tag{2}$$

where the expectation operator E reflects the randomness in the selection of the F_j's as well as the X_{ij}'s. Initially, we consider functionals of the form

$$T(F) = E_F\{h(X)\}, \tag{3}$$

where h(.) is a specified function and X has cdf F. For example, if the quantity we wish to estimate is the mean of F we would set

$$T(F) = \int_{-\infty}^{\infty} x \, dF(x).$$

In section 4 we indicate a method for dealing with more general functionals.

We observe that for each j, the Bayes optimal estimate of $\Theta_j = T(F_j)$ under squared-error loss is

$$\hat{\Theta}_j = E\{\Theta_j | X_{ij}, 1 \leq i \leq n\}.$$

If the observation X_{kj} is omitted from the data for the j-th population for some k, $1 \leq k \leq n$, then the corresponding Bayes estimator for Θ_j is

$$
\begin{aligned}
\hat{\Theta}_j(k) &= E\{\Theta_j | X_{ij}\ 1 \leqq i \leqq n,\ i \neq k\}, \\
&= E\{E\{h(X_{kj})|\ 1 \leqq i \leqq n,\ i \neq k\}, \\
(4) \qquad &= E\{h(X_{kj})|X_{ij},\ 1 \leq i \leq n,\ i \neq k\}, \\
&\overset{\text{def}}{=} \phi(X_{ij},\ 1 \leq i \leq n,\ i \neq k),
\end{aligned}
$$

where ϕ is a fixed symmetric function of $n - 1$ arguments independent of j and k. Since ϕ is a conditional expectation function, it may be estimated using any suitable nonparametric regression method applied to the data from all m populations. To make maximum use of the information available for the estimation of ϕ, we may organize the mn observations as follows:

$$
(5) \qquad
\begin{array}{cc}
\underline{\text{"Dependent"}} & \underline{\text{"Independent"}} \\
h(X_{11}) & X_{21}, X_{31}, \dots, X_{n1} \\
h(X_{21}) & X_{11}, X_{31}, \dots, X_{n1} \\
\cdot & \cdot \\
\cdot & \cdot \\
h(X_{n1}) & X_{11}, X_{21}, \dots, X_{n-1,1} \\
h(X_{12}) & X_{22}, X_{32}, \dots, X_{n2} \\
\cdot & \cdot \\
\cdot & \cdot \\
h(X_{nm}) & X_{1m}, X_{2m}, \dots, X_{n-1,m}
\end{array}
$$

Because of the symmetry of the function ϕ we should increase this list by including all permutations of the "independent" values, but this may be avoided by first ordering the observations from each population so that $X_{ij} \leq X_{2j} \leq \dots \leq X_{nj}$ for each j. This, of course, leads to a different (nonsymmetric) regression function, say ψ, which is defined only for ordered arguments but contains the same information as ϕ. Henceforth, we shall assume that the X_{ij}'s are ordered in this fashion. If $\hat{\psi}_m$ represents a suitable nonparametric regression estimate of ψ based on the available data, then the proposed empirical Bayes estimator of Θ_j is

$$(6) \qquad \hat{\Theta}_j = \frac{1}{n} \sum_{k=1}^{n} \hat{\psi}_m(X_{ij}, 1 \leqslant i \leqslant n, i \neq k),$$

for j=1,2,...,m. The averaging over n values of $\hat{\psi}$ indicated in (6) results in a slight improvement in the performance of the estimator (see (2.47), p.656 of Johns, 1957).

The original formulation of the fully nonparametric empirical Bayes estimation problem considered the component problems in sequence and concentrated on the risk for the m-th problem using the estimated conditional expectation based on the data from the previous m-1 problems. Strictly speaking, the original asymptotic optimality result applies to the present case only if we modify the procedure indicated above so that for each j the estimate of ψ involves only data from the other m-1 component problems. Then, for the modified procedure and the original partition estimate of ψ, if we let $\hat{\Theta}_j$'s given by (6) the following result holds:

THEOREM (Johns, 1957) If $E\{h^2(X)\} < \infty$, then

$$(7) \qquad R_n^* < \lim_{m\to\infty} R_n(\hat{\Theta}) < R_{n-1}^*$$

where R_n^* = the Bayes optimal risk for a component problem with sample size n, and $R_n(\hat{\Theta})$ is the average risk using the empirical Bayes estimator $\hat{\Theta}$ where the sample size is n for each component problem.

The modified procedure is too cumbersome for application to actual data since it entails repeated estimation of the function ψ. It seems plausible that (7) will hold for the unmodified procedure based on any well behaved estimator of the function ψ for which the pointwise convergence in probabiliity to ψ as m becomes large is asymptoticaly unaffected by the values of the X_{ij}'s for any fixed j.

In applications, if n is large and m is not very large, the estimate of $\hat{\psi}_m$ may be unstable and it may be desirable to substitute a summary statistic of lower dimension for the n-1 arguments of ψ. If this summary statistic is well chosen the resulting loss of asymptotic efficiency may be slight. One possibility would be to replace the conditioning X_{ij}'s by a two dimensional statistic consisting of robust estimators of location and scale. In some of the examples considered in the present paper, a less drastic reduction in dimension has been obtained by replacing the n-1 ordered X_{ij}'s by d averages of s successive ordered values where ds = n-1. It may be shown (see, e.g., Johns (1974)) that such averages of blocks of order statistics retain most of the sample information about the underlying distribution.

As was mentioned in the introduction, the method used to estimate the required conditional expectation in the present study is the SMART algorithm of Friedman (1984). Given a number of i.i.d. observations of a dependent variable Y and the correspondig values of "independent" variables $X_1,X_2,\dots,X_p$, the algorithm estimates $E\{Y|X_1,X_2,\dots,X_p\}$ nonparametrically by an expression of the form

$$(8) \qquad \overline{Y} + \sum_{r=1}^{R} \beta_r f_r(aX'),$$

where $X = (X_1,X_2,\dots,X_p$ (and $a_r = (a_{1r},a_{2r},\dots,a_{pr})$. The a_{ir}'s and the functions $f_r()$ are suitably normalized to avoid identifiability difficulties. For each r and each direction vector a_r the functions f_r are obtained by a smoother which fits a least squares line to the Y's associated with a symmetric set of nearest neighbors for each value of $a_r\mathbf{X}'$. The resulting smoothed f_r's are in turn chosen to minimize the sum of squared deviations of (8) from the observed Y's. The process proceeds stagewise starting with a maximum number R of terms (= three in the present application) and reducing R until the sum of squared residuals increases sharply. The convergence properties of projection pursuit algorithms are discussed in e.g., Donoho, et al. (1985).

3. Examples.

The proposed nonparametric empirical Bayes estimation procedure incorporating the SMART algorithm as implemented on a VAX11/750 computer was applied to six sets of simulated data and one set of real data. For each example, the quantities being estimated (i.e., the Θ_j's) are the means of the component populations. The simulated data sets consist in each case of either 50 or 100 component populations. These numbers are perhaps larger than would be expected in some applications to real world data but were chosen to yield reasonably stable and interpretable results. The numerical results obtained from the six simulations are summarized in Table 1. The conditional distributions were either normal (cases (a)-(d)) or logistic (cases (e), (f)). The prior distribution for the conditional means were either normal (cases (a)-(c), and (e)) or the longtailed distribution having density

$$g(\Theta) = \frac{\sqrt{2}}{\pi(1+\Theta^4)} \tag{9}$$

which has mean = 0 and standard deviation = 1. The normal priors all had standard deviation = 2 and mean = 25, except for case (e) which had mean.= 0. For cases (a) and (f) the scale parameter for the conditional distribution was selected independently with equal prior probabilities from three values; 2.0, 4.0, and 6.0 for case (a) and 4.0, 5.0, and 6.0 for case (f). For cases (b), (c), (d) and (e) the fixed values of the scale parameters were 4.0, 6.0, 2.0 and 3.0 respectively. The summary statistic on which the predicted values of Θ were based was either the set of all n-1 available observations or, for cases (c) and (f) for which n = 11, the set of five averages of two adjacent order statistics.

Table 1
Summary of the Simulation Results

Case Label	Conditional Distr.	Prior Distr. (for Θ)	No. of Pops.& Sample Size (m,n)	Bayes Opt. Risk (Approx.)	Asymptotic BLUE M.S.E.	Observed BLUE M.S.E.	Observed EB M.S.E.
(a)	Normal	Normal*	(100,5)	1.67	3.73	4.29	1.98
(b)	Normal	Normal	(100,5)	1.78	3.20	3.32	1.57
(c)	Normal	Normal	(50,11)	1.80	3.27	3.38	2.58
(d)	Normal	Longtail	(100,5)	0.44	0.80	0.69	0.45
(e)	Logistic	Normal	(100,6)	2.12	4.50	4.50	2.69
(f)	Logistic	Longtail*	(50,11)	0.86	7.00	6.98	3.40

* The values of sigma are selected randomly from among three values.

The last column of Table 1 shows the actual mean squared error (M.S.E.) produced by the fully nonparametric empirical Bayes procedure. For comparison purposes both the average observed variances and the true (asymptotic) variances for the best linear unbiased estimators (BLUE's) are shown. For the normal cases, of course, the BLUE is simply the sample mean. Approximate values for the Bayes optimal risk are also given. These are based on linear Bayes estimators and asymptotic variances so they are only exact for cases (b) and (c) where both the conditional and the prior distributions are normal. It is encouraging to note that the empirical Bayes M.S.E. is substantially smaller than the BLUE variance for each of the examples. Furthermore, the empirical Bayes M.S.E. is in the vicinity f the Bayes optimal risk for all cases but one (example (f)).

The actual regression functions produced by the SMART algorithm are plotted in Figures 1 and 2. In all cases only a single function f_1 was required in expression (8) for an adequate description of the data. When interpreting the plots it should be borne in mind that a different direction vector a is associated with each function. The vector $\mathbf{X}$ represents the appropriate set of "independent" variables.

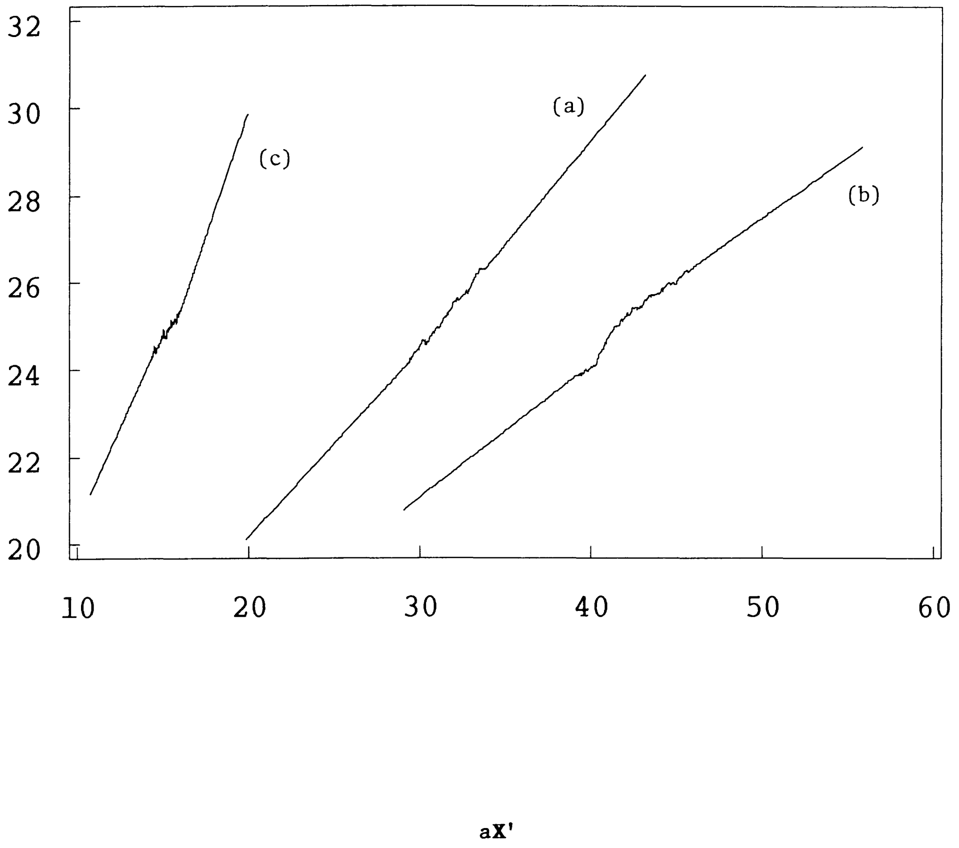

Figure 1. SMART Regression Functions

We observe that the plots are quite linear for all cases with normal conditional distributions as one might expect, but distinctly nonlinear for the logistic cases. It was thought that example (a) might yield a nonlinear regression because of the random prior on σ. A numerical calculation of the actual conditional expectation of the mean given the sample mean and the sample variance verified that the regression surface was in fact fairly linear.

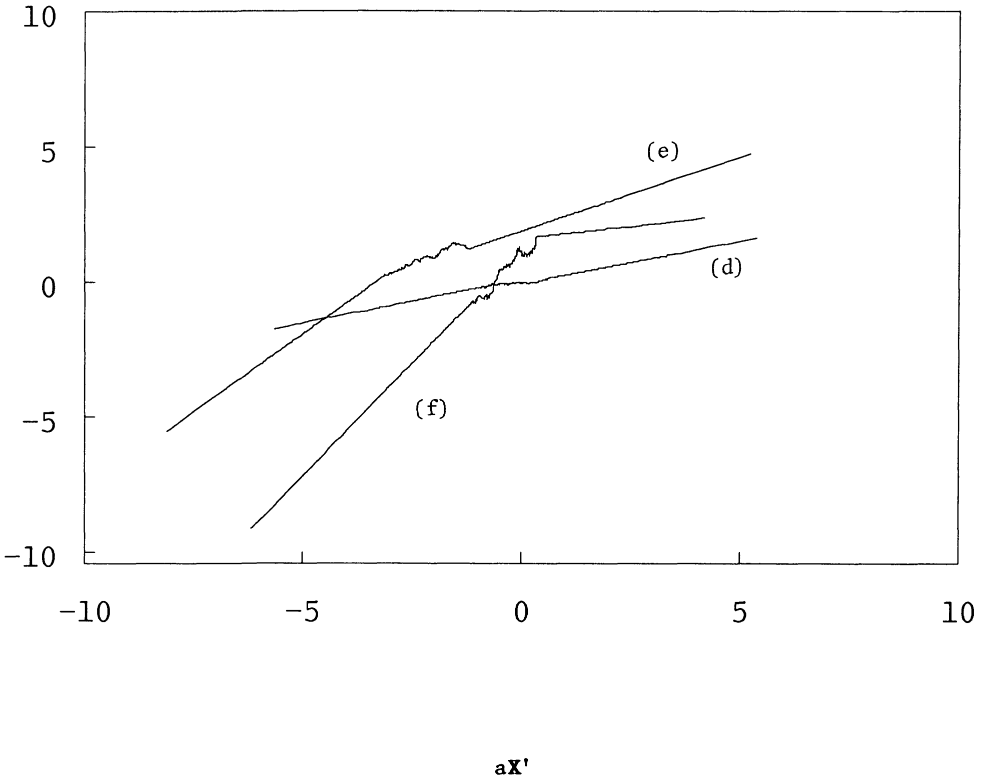

Figure 2. SMART Regression Functions

An actual real world data set was also analyzed using the fully nonparametric empirical Bayes scheme. The data was obtained from Efron-Morris (1975) and consists of the batting averages for 18 major league baseball players for their first 45 times at bat and their averages for the remainder of the season which represent the 'true' values one wishes to predict. Efron-Morris first transform the data to approximate normality using the arcsine transformation. They then compute the Stein estimator (Stein, 1955) and their own proposed estimator based on a linear empirical Bayes formula modified to limit the maximum component risk. The results are then converted back to proportions. For the present study the data was considered in this original form as a set of Bernoulli observations (hits or non-hits) and the fully nonparametric empirical Bayes method was applied. The results are shown in Table 2. The third column gives the maximum likelihood estimate (MLE) which is just the observed proportion of hits in the first 45 at bats. The nonparametric empirical Bayes estimate is given in the fourth column and Stein's estimate in the fifth. The Efron-Morris limited risk estimate with index .8 is given in the last column. The corresponding mean squared errors of prediction are shown in the last row.

Table 2
Batting Averages and their Estimates

i	"TRUE"	MLE	NP-EB	STEIN	EMEST(.8)
1	.346	.400	.306	.290	.351
2	.298	.378	.293	.286	.329
3	.276	.356	.281	.281	.308
4	.222	.333	.269	.277	.287
5	.273	.311	.256	.273	.273
6	.270	.311	.256	.273	.273
7	.263	.289	.247	.268	.268
8	.210	.267	.247	.264	.264
9	.269	.244	.254	.259	.259
10	.230	.244	.254	.259	.259
11	.264	.222	.258	.254	.254
12	.256	.222	.258	.254	.254
13	.303	.222	.258	.254	.254
14	.264	.222	.258	.254	.254
15	.226	.222	.258	.254	.254
16	.285	.200	.266	.249	.242
17	.316	.178	.274	.244	.218
18	.200	.156	.283	.239	.194
M.S.E.		.00419	.00105	.00120	.00139

We observe that the procedure proposed in this study has the smallest overall mean squared error of prediction and also does better than the Efron-Morris estimator in three out of the five cases (i=1,2,3,17,18) where their procedure limits the risk. The highly nonlinear regression function which SMART produces for this case is plotted in Figure 3. The abscissa of this figure is a linear function of the number of hits in 44 at bats.

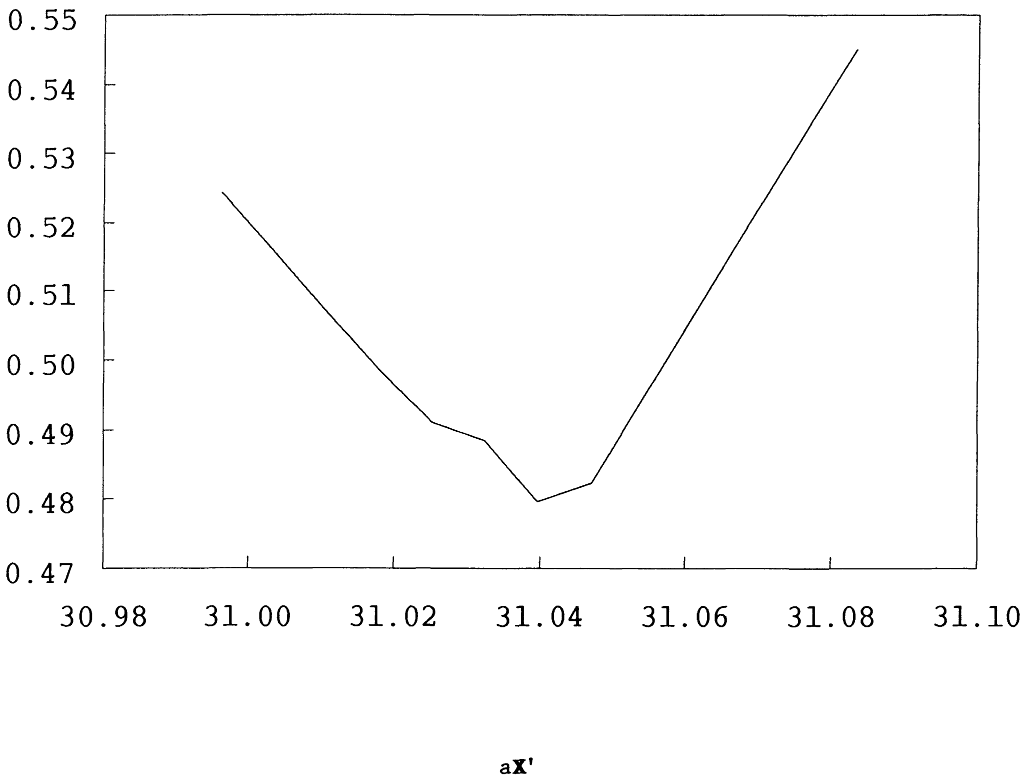

Figure 3. SMART Regression Function

4. Concluding remarks.

The estimation procedures discussed here may be modified and generalized in various ways. We may expect that ever more sophisticated nonparametric regression methods will be developed. Such procedures may then be substituted for the projection pursuit part of the scheme. The empirical Bayes problem described here assumes equal sample sizes for all component populations. The case of unequal sample sizes may be dealt with by various ad hoc methods some of which are discussed in the original paper (Johns, 1957). The question of the best way to proceed in such cases is still open.

In the preceding sections the quantities to be estimated were required to be represented as functionals of the form (3). However, within this framework we may estimate the conditional cdf F(t) for any fixed t by letting h(x) = the indicator function of the interval $(-\infty,t]$. Since F(t) can be recaptured, it should be possible to modify the procedure to permit the estimation other functionals T(F) such as, e.g., the median of F.

As is true of most empirical Bayes problems, the present one may be reinterpreted as a compound decision problem by dropping the assumption of the existence of a prior probability distribution, and replacing it with a suitable empirical distribution of unknown quantities. In the present case these quantities are the component cdf's $F_1,F_2,\ldots,F_m$. Presumably results paralleling the empirical Bayes results would be forthcoming here as in previously considered problems. (See Robbins (1951) for the original formulation of the key ideas and Gilliland (1968) and Johns (1967) for some further developments.)

Finally the author wishes to express his thanks to David J. Pasta who rendered invaluable assistance in the application of th SMART algorithm to the data of this study.

REFERENCES

Donoho, D., Johnstone, I., Rousseeuw, P. and Stahel, W. (1985). Discussion of P.J. Huber (1985), Projection Pursuit. Ann. Stat. **13** 435-475.

Efron, B. and Morris, C. (1975). Data analysis using Stein's Estimator and its generalization. JASA **70** 311-319.

Friedman, J.H. (1984). SMART User's Guide. Dept. of Statist., Stanford University, Report LCM001.

Friedman, J.H. and Stuetzle, W. (1981). Projection Pursuit Regression. J. Amer. Statist. Assoc. **76** 817-823.

Gilliland, D.C. (1968). Sequential compound estimation. Ann. Math. Stat. **39** 1890-1904.

Huber, P.J. (1985). Projection pursuit. Ann. Stat. **13** 435-475.

Johns, M.V. (1957). Non-parametric empirical Bayes procedures. Ann. Math. Statist. **28** 649-669.

Johns, M.V. (1967). Two-action compound decision problems. Proc. Fifth Berkeley Symp. Math. Statist. Prob. **1** 463-478.

Johns, M.V. (1974). Nonparametric estimation of location. JASA **69** 453-460.

Robbins, H. (1951). Asymptotically subminimax solutions of compound statistical decision problems. Proc. Second Berkeley Symp. Math. Statist. Prob. 131-148.

Robbins, H. (1956). An empirical Bayes approach to statistics. Proc. Third Berkeley Symp. Math. Statist. Prob. 157-163.

Robbins, H. (1983). Some thoughts on empirical Bayes estimation. Ann. Statist. **11** 713-723.

Stein, C. (1955). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. Proc. Third Berkeley Symp. Math. Statist. Prob. **197-206**.

Stone, C.J. (1977). Nonparametric regression and its applications. Ann. Statist. **5** 595-645.