The Canadian Journal of Statistics
Vol. 29, No. 3, 2001, Pages 343–378
La revue canadienne de statistique

Flexible regression modeling with adaptive logistic basis functions

Peter M. HOOPER

Key words and phrases: Approximation; data mining; least absolute deviation; neural networks; nonparametric multiple regression; radial basis functions; smoothing.

MSC 2000: Primary 62G07; secondary 62J02.

Abstract: The author proposes a new method for flexible regression modeling of multi-dimensional data, where the regression function is approximated by a linear combination of logistic basis functions. The method is adaptive, selecting simple or more complex models as appropriate. The number, location, and (to some extent) shape of the basis functions are automatically determined from the data. The method is also affine invariant, so accuracy of the fit is not affected by rotation or scaling of the covariates. Squared error and absolute error criteria are both available for estimation. The latter provides a robust estimator of the conditional median function. Computation is relatively fast, particularly for large data sets, so the method is well suited for data mining applications.

Un modèle de régression flexible défini à partir d'une base de fonctions logistiques adaptatives

Résumé : L'auteur propose une nouvelle méthode de régression flexible pour la modélisation de données multivariées dans laquelle la fonction de régression est approchée par une combinaison linéaire de fonctions logistiques. Cette méthode adaptative permet de choisir des modèles plus ou moins complexes selon les besoins. Le nombre, la localisation et (jusqu'à un certain point) la forme des fonctions logistiques de base sont automatiquement déterminés à partir des données. La méthode étant équivariante par transformations affines, la précision de l'ajustement n'est pas affectée par une rotation ou un changement d'échelle des variables exogènes. L'estimation peut s'appuyer sur le critère de l'erreur quadratique ou absolue. Dans le second cas, on obtient un estimateur robuste de la médiane conditionnelle. La méthode se prête bien au forage de données, car les calculs nécessaires se font rapidement, même pour de grands ensembles de données.

1. INTRODUCTION

Consider the problem of estimating a regression function $f(\mathbf{x}) = \mathrm{E}(y \mid \mathbf{x})$, where y is a response variable and $\mathbf{x}$ is a vector of d covariates. Estimators often approximate f by a linear combination of basis functions:

$$f(\mathbf{x}) \approx f_K(\mathbf{x}) = \sum_{k=1}^{K} \delta_k \phi_k(\mathbf{x}). \tag{1}$$

Examples include tensor-product splines (Gu, Bates, Chen & Wahba 1989; Friedman 1991), thin-plate splines (Wahba 1990), and ridge functions (Friedman & Stuetzle 1981). This article investigates a new family of estimators $\hat{f}$ defined by logistic basis functions:

$$\phi_k(\mathbf{x}) = \exp(\alpha_k + \boldsymbol{\beta}_k'\mathbf{x}) \Bigg/ \sum_{m=1}^{K} \exp(\alpha_m + \boldsymbol{\beta}_m'\mathbf{x}). \tag{2}$$

There is some redundancy in the parameterization. We have $\sum \phi_k(\mathbf{x}) = 1$ for all $\mathbf{x}$, so approximation (1) does not require a constant term. Dividing the numerator and denominator of (2) by $\exp(\alpha_K + \boldsymbol{\beta}_K'\mathbf{x})$ shows that, without loss of generality, one pair $(\alpha_K, \boldsymbol{\beta}_K)$ can be set to zero.

The effective number of parameters used in approximation (1) is thus

$$p = 1 + (K - 1)(d + 2). \tag{3}$$

I refer to the methodology developed in this article as adaptive logistic basis (ALB) regression. The method is "adaptive" in that both K and the parameters defining f_K are determined from the data. ALB estimators $\hat{f}$ are defined for a family of location measures, including the conditional mean and median. Suppose that $(\mathbf{x}, y)$ is a random vector, choose $q \geq 1$, and let f be a function minimizing $\mathrm{E}\{|y - f(\mathbf{x})|^q\}$. It is assumed that this expectation is finite. Conditional mean and median functions are obtained by taking $q = 2$ and $q = 1$, respectively. The conditional median need not be uniquely defined. Suppose that we have a sample $\{(\mathbf{x}_i, y_i), i = 1, \ldots, n\}$. For given K, an ALB L_q estimator $\hat{f}_K$ is calculated by minimizing $\sum |y_i - f_K(\mathbf{x}_i)|^q$. The parameter values defining $\hat{f}_K$ are determined separately for different numbers K, and a generalized cross-validation technique is used to select K.

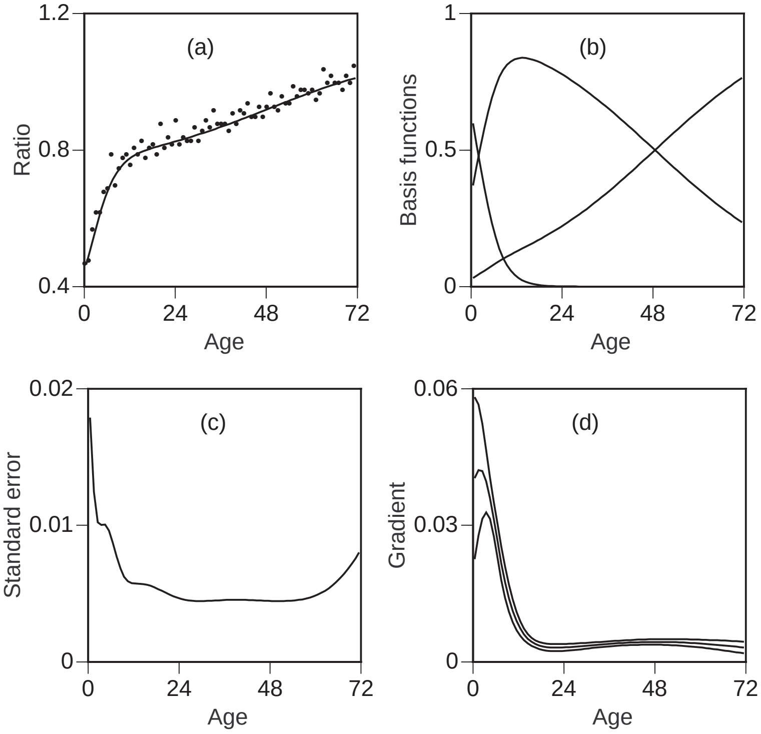

FIGURE 1: Preschool boys' weight/height ratio by age. (a) ALB L_2 estimate $\hat{f}$. (b) Basis functions. The corresponding δ_k estimates, from left to right, are 0.25, 0.75, and 1.09. (c) Standard error of $\hat{f}$. (d) Gradient $\hat{g} \pm 2\text{se}(\hat{g})$.

This article introduces the ALB regression methodology and investigates its potential usefulness through theory, examples, and simulations. First, in section 2, using five data sets, I will illustrate the method. Section 3 contains some theoretical results and comparisons of ALB with related statistical and neural network methods. Algorithms to obtain $\hat{f}_K$ for given K and to select K are described in Section 4. A simple formula for approximate standard errors is developed in Section 5. The results of simulation studies on predictive performance of the ALB L_2 estimator are reported in Section 6. Finally, in section 7, several extensions of the ALB methodology are discussed.

Logistic basis functions have proved to be useful in many applications. They have long been used in regression models for binary responses (Cox & Snell 1989). They have recently been applied in classification problems to construct flexible classification boundaries (Hooper 1999) and to model conditional probabilities of class membership (Hooper 2001). The estimation algorithm described in Section 4.1 is based on a stochastic approximation algorithm developed for the related classification problem. The classification methodology is an important component of a program to predict genetic structure in DNA sequences (Hooper, Zhang & Wishart 2000). ALB regression has been used to model a covariance function, as part of a model relating ultrasound estimates of fetal weight to gestational age (Hooper, Mayes & Demianczuk 2001).

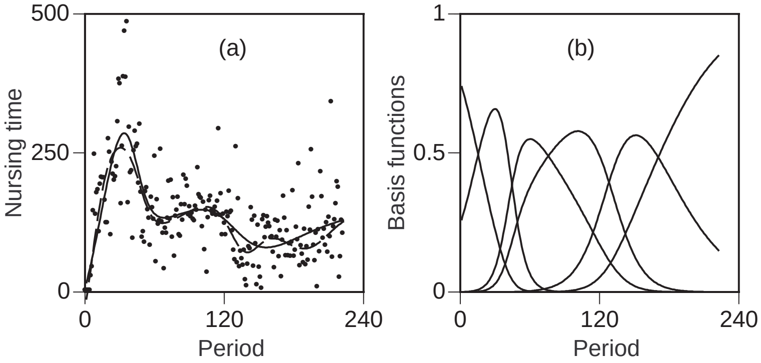

FIGURE 2: Nursing time of a beluga whale calf by time period.
(a) ALB L_2 (solid line) and L_1 (broken line) estimates. (b) Basis functions for the L_2 estimate.
The corresponding δ_k estimates, from left to right, are -137, 453, 46, 225, 16, and 149.

2. EXAMPLES

The examples in this section illustrate properties of $\hat{f}$, its gradient $\hat{\mathrm{g}} = \partial\hat{f}/\partial\mathbf{x}$, standard errors, and basis functions. Usually the basis functions are not interpretable and would not be examined when analysing a data set. They are displayed here to provide insight concerning the construction of $\hat{f}$.

2.1. Weight/height ratio.

Figure 1 presents data relating weight/height ratio (in lb/in.) to age (in months) for preschool boys [source: Gallant (1987), Eppright *et al.* (1972)]. The ALB L_2 estimate in (a) is a linear combination of the $\hat{K} = 3$ basis functions plotted in (b). The coefficients $\hat{\delta}_k$ are listed in the caption below Figure 1. This example provides a simple illustration of how the basis functions are used to construct the curved and linear portions of $\hat{f}$. The plot (c) of the standard error $\mathrm{se}\{\hat{f}(x)\}$ shows how the standard deviation of $\hat{f}(x)$ increases at the boundaries of the data. The plot (d) of the gradient estimate includes approximate 95% confidence intervals $\hat{g}(x) \pm 2\mathrm{se}\{\hat{g}(x)\}$. Note how $\hat{g}(x)$ shrinks slightly toward zero at the data boundaries. This is likely an artifact that is related to the shape of the basis functions, which causes $\hat{f}$ to flatten as x moves away from the data. Note also how $\mathrm{se}\{\hat{g}(x)\}$ is large when x is close to the boundary and $|\hat{g}(x)|$ is large. These two effects occur quite generally with single and multiple covariates.

2.2. Beluga whale.

Figure 2 presents data on nursing patterns for Hudson, a beluga whale calf born at the New York Aquarium [source: Chatterjee, Handcock & Simonoff (1995)]. Here x is the six-hour time period postpartem index and y is the nursing time in seconds. The ALB L_2 and L_1 estimates are superimposed on scatterplot (a). Plots of residuals (not shown) indicate right skewness and heteroscedasticity, so it is not surprising that the L_1 estimate is slightly less than the L_2 estimate over much of the interval. The L_2 estimate uses $\hat{K} = 6$ basis functions, plotted in (b), while the L_1 estimate uses $\hat{K} = 8$. Simonoff (1996, Fig. 5.18) estimated the regression function using a local quadratic kernel smoother with varying bandwidth. His bandwidths were selected informally to account for varying smoothness and scatter. The adaptive selection of the logistic basis functions has a similar motivation.

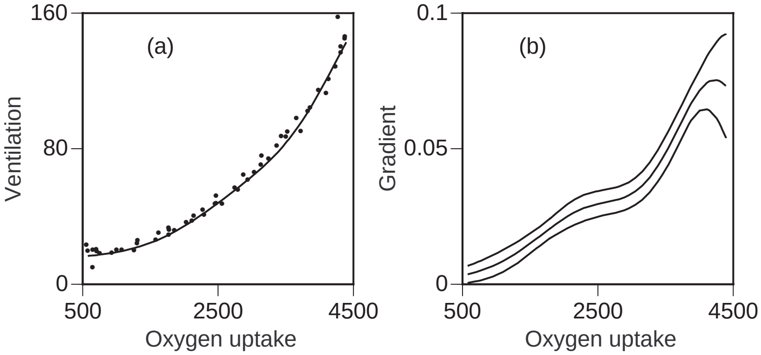

FIGURE 3: Anaerobic threshold data. (a) ALB L_2 estimate. (b) Gradient $\hat{g} \pm 2\text{se}(\hat{g})$.

2.3. Anaerobic threshold.

This example illustrates how gradient estimates can be applied to the problem of estimating anaerobic threshold levels. Routledge (1991) described this problem in applied physiology as follows. Physiologists believe that during a progressive exercise test, there comes a point when aerobic metabolic processes are supplemented by anaerobic processes, producing an additional source of CO_2. Some investigators have estimated this “anaerobic threshold” by locating an upward bend in a plot of expired ventilation against oxygen uptake. Several authors have noted that these plots often curve smoothly with no apparent bend. Figure 3(a) presents data for a single individual in a single exercise test [source: Bennett (1988)]. The location (or even the existence) of a threshold is not obvious from the plot of the ALB L_2 estimate ($\hat{K} = 3$). The plot of its gradient in (b), however, suggests an upward bend. It is possible that the threshold effect is introduced more gradually in some cases, producing an upward bend in the gradient but not in the original function.

2.4. Viking formation.

The Viking formation is a sandstone layer, the floor of an ancient ocean, lying beneath the surface of western Canada. ALB regression can be used to model the elevation of this layer (in feet above sea level) as a function of latitude and longitude. The data were obtained from 74229 drill holes in Alberta [source: Stefan Bachu, Alberta Geological Survey, personal communication]. The drill holes range from the Saskatchewan border to the foothills of the Rocky Mountains and from the 50th to the 57th parallel. Elevations vary from $+560$ in the northeast to -1900 near the foothills in the southwest. The ALB L_1 and L_2 estimators produce similar results, with respective $\hat{K}$ values of 12 and 13. The more robust L_1 estimator seems preferable in this application. The distribution of the residuals is symmetric and very long-tailed, with an interquartile range of 17 and a range of 1913.

Figure 4 presents a contour plot of the L_1 estimate $\hat{f}$, with a sample of drill hole locations superimposed. The plot shows a rapid decline in elevation as one approaches the mountains. The flattening of $\hat{f}$ in the southwest corner of the plot is an artifact associated with an absence of drill holes in this region. Figure 4 also displays a contour plot of $\max_k \phi_k$, illustrating the location, orientation, and relative height of the 12 basis functions. Two of the basis functions, located near $(-111, 52)$, are difficult to identify in the plot because their maximum height is overshadowed by neighbouring peaks. Reasonably good fits $\hat{f}_K$ can be obtained with substantially fewer basis functions. As K increases from 1 to 12, the standardized predictive absolute error risk decreases as follows: $1.000, .295, .078, .069, .046, .042, .040, \ldots, .034$. An examination of the residuals from $\hat{f}$ reveals many outliers (likely due to measurement or data entry errors) but also some localized effects that appear to represent structure missed by the ALB fit. These finer details might be investigated by fitting surfaces to the residuals over smaller subregions.

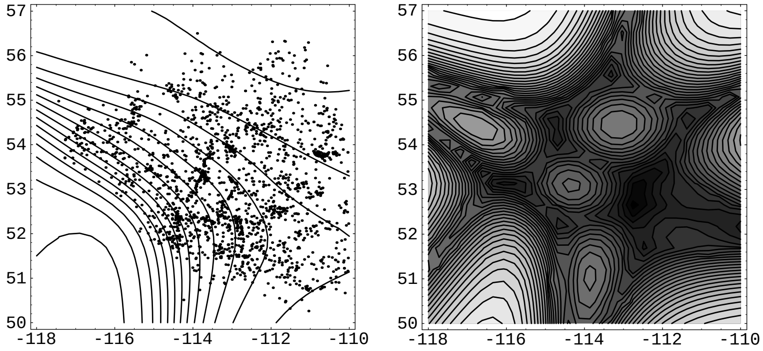

FIGURE 4: Viking formation. (a) Contours of the L_1 estimate $\hat{f}$, with 2000 randomly sampled locations superimposed. Latitude is plotted on the vertical axis. Negative longitude is plotted on the horizontal axis, producing the usual east-west orientation. (b) Contours of $\max(\phi_1, \ldots, \phi_{12})$, with lighter shaded regions closer to 1.

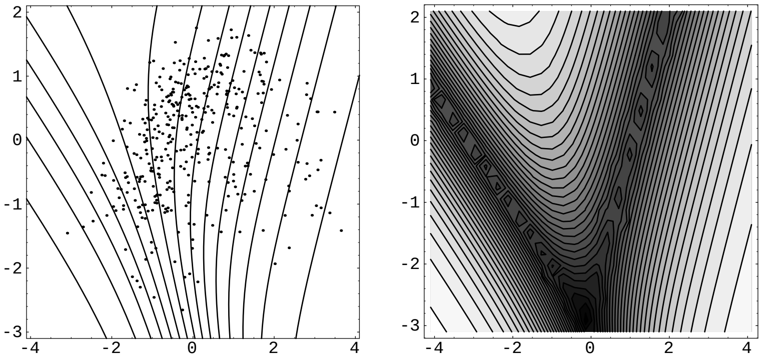

FIGURE 5: Boston housing data. (a) Contours of the L_2 estimate $\hat{f}$, with the 374 data points superimposed. The $\hat{f}$ values increase from 9.1 to 10.8 as the contours proceed from left to right. The horizontal and vertical axes are linear combinations $\mathbf{b}_1'\mathbf{x}$ and $\mathbf{b}_2'\mathbf{x}$ of the original 13 predictors. (b) Contours of $\max(\phi_1, \phi_2, \phi_3)$, with lighter shaded regions closer to 1. The δ_k estimates for the lower left, upper middle, and lower right basis function are 8.85, 9.80, and 10.92.

2.5. Boston housing.

Following Li (1997), I examined a low crime rate subset of 374 census tracts from the Boston housing data [source: Harrison & Rubinfeld (1978), Breiman & Friedman (1985)]. Here y is the log median housing price per census tract and $\mathbf{x}$ consists of the remaining 13 variables. Ten-fold cross-validation indicates that the ALB L_2 estimator accounts for 89% of the variance of y, i.e., the predictive squared error risk estimate divided by the sample variance of y equals 0.11.

Before examining a plot of the fitted model, consider how one might try to visualize an ALB model in higher dimensions. It is straightforward to show that an estimate $\hat{f}_K$ can be expressed as a function of $r = \min(d, K-1)$ linear combinations $\mathbf{b}_1'\mathbf{x}, \ldots, \mathbf{b}_r'\mathbf{x}$, where $\mathbf{b}_1, \ldots, \mathbf{b}_r$ span the subspace that is spanned by the contrasts $\hat{\boldsymbol{\beta}}_1 - \hat{\boldsymbol{\beta}}_K, \ldots, \hat{\boldsymbol{\beta}}_{K-1} - \hat{\boldsymbol{\beta}}_K$. When $r = 2$, we can visualize $\hat{f}_K$ in a 3-dimensional plot. More generally, we can identify the directions in the covariate space that best represent variation in $\hat{f}_K$ by carrying out a principal components analysis of the gradient sum-of-products matrix $\mathbf{G} = \sum \hat{\mathrm{g}}(\mathbf{x}_i)\hat{\mathrm{g}}(\mathbf{x}_i)'$. The gradient vectors $\hat{\mathrm{g}}(\mathbf{x}_i)$ lie in the above-mentioned contrast subspace, so the rank of $\mathbf{G}$ is at most r. Let $\mathbf{b}_j$ be the eigenvector corresponding to the jth largest eigenvalue e_j of $\mathbf{G}$. The first eigenvector $\mathbf{b}_1$ maximizes the sum of squared gradients $\sum\{\mathbf{b}'\hat{\mathrm{g}}(\mathbf{x}_i)\}^2$. If $(e_1 + e_2)/(e_1 + \cdots + e_r) \approx 1$, then a plot of $\hat{f}_K$ versus $(\mathbf{b}_1'\mathbf{x}, \mathbf{b}_2'\mathbf{x})$ accounts for nearly all of the variation in $\hat{f}_K$. Otherwise, $\hat{f}_K$ cannot be visualized in a single plot.

For the Boston housing data, the ALB L_2 estimate $\hat{f}$ has $\hat{K} = 3$, so $\hat{f}$ can be fully represented in a 3-dimensional plot. Figure 5(a) shows a scatter plot of the data and a contour plot of $\hat{f}$, with horizontal and vertical axes defined by the principal gradient components $\mathbf{b}_1'\mathbf{x}$ and $\mathbf{b}_2'\mathbf{x}$. The leading eigenvalue of $\mathbf{G}$ is relatively large, so most of the variation in $\hat{f}$ occurs in the horizontal direction. The plot reveals a partial helix effect similar to that reported by Li (1997). Contours of the 3 basis functions are shown in Figure 5(b). The ALB estimate with $K = 4$ is similar to that with $K = 3$. A plot (not shown) of $\hat{f}_4$ against its first two principal gradient components accounts for most of the variation in $\hat{f}_4$ because the third eigenvalue of the $\mathbf{G}$ matrix for $\hat{f}_4$ is relatively small.

To interpret an ALB model, we must relate $\hat{f}$ to the individual covariates. I address this problem by examining regions where the axis of steepest ascent/descent remains fairly stable. This can be done by clustering gradient direction vectors $\tilde{\mathrm{g}}(\mathbf{x}_i) = \hat{\mathrm{g}}(\mathbf{x}_i)/||\hat{\mathrm{g}}(\mathbf{x}_i)||$ about direction "centroids" $\mathbf{c}_j$ using the "distance" $1 - |\mathbf{c}_j{}'\tilde{\mathrm{g}}(\mathbf{x}_i)|$. Choosing three clusters leads to regions on roughly the left side, middle, and right side of Figure 5(a), corresponding to census tracts with low, middle, and high median housing prices. Within each region, $\hat{f}$ is well approximated by a one-dimensional function. One can attempt to interpret local directions of steepest ascent (cluster centroids) by examining within-cluster correlations between directions and individual covariates. The interpretation is, of course, less clear when there are stong dependencies among the covariates. A gradient clustering approach suggests the following interpretation of the ALB model. In regions where median prices are high, it appears that median prices are well predicted by median size of the house alone. Where prices are low, additional variables are needed for the best effect. This analysis supports the conclusion of Li (1997) that a linear model is inadequate.

3. THEORY AND COMPARISONS

This section presents some simple results about logistic basis functions, together with discussion and comparisons. The results shed some light on how f is approximated by f_K and on what kind of functions f are well approximated with small K. The first proposition presents some useful derivatives. The second describes key features of the basis functions when $d = 1$.

PROPOSITION 1.

(i) $\partial\phi_k/\partial\mathbf{x} = \phi_k(\mathbf{x})\{\boldsymbol{\beta}_k - \bar{\boldsymbol{\beta}}(\mathbf{x})\}$, *where* $\bar{\boldsymbol{\beta}}(\mathbf{x}) = \sum_{k=1}^{K}\phi_k(\mathbf{x})\boldsymbol{\beta}_k$.

(ii) $\partial\bar{\boldsymbol{\beta}}'/\partial\mathbf{x} = \sum_{k=1}^{K}\phi_k(\mathbf{x})\{\boldsymbol{\beta}_k - \bar{\boldsymbol{\beta}}(\mathbf{x})\}\{\boldsymbol{\beta}_k - \bar{\boldsymbol{\beta}}(\mathbf{x})\}'$.

(iii) $\partial^2\log\phi_k/\partial\mathbf{x}\partial\mathbf{x}' = -\partial\bar{\boldsymbol{\beta}}'/\partial\mathbf{x}$.

(iv) *Each* $\log\phi_k$ *is concave.*

Proof. Direct calculation. □

PROPOSITION 2. *Let $d = 1$ and suppose $\beta_1 \leq \cdots \leq \beta_K$ with $\beta_1 < \beta_K$.*

(i) *Each $\log \phi_k$ is strictly concave.*

(ii) *ϕ_1 is strictly decreasing, approaching 0 at $+\infty$. If $\beta_1 < \beta_2$, then ϕ_1 approaches 1 at $-\infty$.*

(iii) *ϕ_K is strictly increasing, approaching 0 at $-\infty$. If $\beta_{K-1} < \beta_K$, then ϕ_K approaches 1 at $+\infty$.*

(iv) *If $\beta_1 < \beta_k < \beta_K$, then ϕ_k is strictly increasing for $x < x_k^*$ and strictly decreasing for $x > x_k^*$, with x_k^* determined by $\bar{\beta}(x_k^*) = \beta_k$. Furthermore, ϕ_k approaches 0 at $\pm\infty$. We also have $x_k^* < x_m^*$ if $\beta_1 < \beta_k < \beta_m < \beta_K$.*

Proof. The results follow from Proposition 1, which shows that $\bar{\beta}$ is a strictly increasing function mapping $\Re$ onto the open interval (β_1, β_K). □

When $d = 1$, ALB can be viewed as a scatterplot smoother. A large variety of nonparametric smoothers have been developed, e.g., see Simonoff (1996) or Eubank (1999). Various methods often provide similar accuracy given appropriate choices of smoothing parameters. The concavity of $\log \phi_k$ suggests connections with b-splines, in that basis functions contribute to $\hat{f}$ in local, overlapping regions. The adaptive estimation of the ϕ_k suggests comparison with free-knot splines, where the number and location of the knots are chosen adaptively (Lindstrom 1999). The ALB and spline methods differ with regard to smoothness of the fitted models. ALB models are infinitely differentiable. Splines possess a finite number of derivatives at knot locations (two for cubic splines), and the number can be reduced by moving knots together. This suggests that free-knot splines may be more efficient than ALB in fitting curves with sharp bends. Applications of ALB to examples from Lindstrom (1999) support this view, although differences in predictive performance appear to be minor.

The following proposition allows an interpretation of higher-dimensional basis functions through lower-dimensional projections. It also establishes the important property of affine invariance.

PROPOSITION 3.

(i) *Let $\mathbf{B}$ be a $d_0 \times d$ matrix, with $d_0 \leq d$, and let $\mathbf{a} \in \Re^d$. The restriction of $\{\phi_k\}$ to the d_0-dimensional linear manifold $\{\mathbf{a} + \mathbf{B}'\mathbf{z} : \mathbf{z} \in \Re^{d_0}\}$ produces a set of d_0-dimensional logistic basis functions, i.e.,*

$$\phi_k(\mathbf{a} + \mathbf{B}'\mathbf{z}) = \exp(\tilde{\alpha}_k + \tilde{\boldsymbol{\beta}}_k'\mathbf{z}) \Big/ \sum_{m=1}^{K} \exp(\tilde{\alpha}_m + \tilde{\boldsymbol{\beta}}_m'\mathbf{z}),$$

where $\tilde{\alpha}_k = \alpha_k + \boldsymbol{\beta}_k'\mathbf{a}$ and $\tilde{\boldsymbol{\beta}}_k = \mathbf{B}\boldsymbol{\beta}_k$.

(ii) *If $\{\boldsymbol{\beta}_1 - \boldsymbol{\beta}_K, \ldots, \boldsymbol{\beta}_{K-1} - \boldsymbol{\beta}_K\}$ spans $\Re^d$, then each $\log \phi_k$ is strictly concave.*

(iii) *Each ϕ_k is quasiconcave; i.e., upper level sets $\{\mathbf{x} : \phi_k(\mathbf{x}) \geq c\}$ are convex.*

(iv) *If $\boldsymbol{\beta}_k$ is in the interior of the convex hull of $\{\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K\}$, then the upper level sets of ϕ_k are compact for $c > 0$. Otherwise, the upper level sets are either unbounded or empty.*

(v) *If $\boldsymbol{\beta}_k = \boldsymbol{\beta}_m$ for some $k \neq m$, then $f_K = \sum \delta_k \phi_k$ can be re-expressed as a linear combination of $K - 1$ logistic basis functions.*

(vi) *The family of functions f_K is invariant under affine transformations of $\mathbf{x}$ and f_K.*

Proof. (i) Immediate. (ii) Concavity follows from Proposition 1. Strict concavity follows from (i) with $d_0 = 1$, and from Proposition 2. The spanning assumption implies that if $\mathbf{b} \neq 0$, then $\mathbf{b}'\boldsymbol{\beta}_k \neq \mathbf{b}'\boldsymbol{\beta}_m$ for some (k, m). (iii) Follows from concavity of $\log \phi_k$. (iv) Follows from (i) and Proposition 2. (v) Replace α_k and α_m by $\alpha = \log\{\exp(\alpha_k) + \exp(\alpha_m)\}$. Replace δ_k and δ_m by $\delta/2$, where $\delta = \delta_k \exp(\alpha_k - \alpha) + \delta_m \exp(\alpha_m - \alpha)$. (vi) Follows from (i) with $d_0 = d$ and $\mathbf{B}$ invertible, and from the identity $a + bf_K = \sum(a + b\delta_k)\phi_k$. □

Propositions 2 and 3 suggest limitations on the complexity of f_K. It appears likely that when $d = 1$, the function f_K can have at most $K - 2$ local extrema; i.e., local minima and maxima. Although I do not have a proof of this conjecture, I have been unable to construct a counterexample. Proposition 3(i) shows that this limitation would also apply to fluctuations along one-dimensional linear manifolds in higher dimensions. A large number of basis functions would thus be required to approximate functions with many bumps or ripples occurring in multiple directions.

Linear and quadratic functions are often used for local approximation. One may ask whether these functions are well-approximated by ALB. Linear functions are well-approximated with $K = 2$ basis functions. From expression (3), the effective number of ALB parameters $p = d + 3$ is just slightly larger than the number $d + 1$ defining a linear function. For a quadratic function $f(\mathbf{x}) = a + \mathbf{b}'\mathbf{x} + \mathbf{x}'\mathbf{C}\mathbf{x}$, simulations with $\mathbf{x}$ multivariate normal indicate that $K = 2d + 1$ generally yields an adequate approximation, e.g., two basis functions for each dimension plus a single basis function in the centre. The number of ALB parameters $p = 1 + 2d(d + 2)$ is substantially larger than the number $1 + d + d(d + 1)/2$ defining a quadratic function. Of course, f_K with $K = 2d + 1$ can approximate many nonquadratic functions as well. Furthermore, substantially fewer basis functions may be required, depending on the rank of $\mathbf{C}$ and the domain of interest within $\Re^d$.

This raises a key question: What kind of functions f are well-approximated by f_K with small K? For $K = 2$ we have a one-dimensional sigmoidal function, which can aproximate linear functions and monotonic curves with an asymptote. This latter shape often arises in applications of parametric nonlinear regression models (Bates & Watts 1988). For $K = 3$, there are a few basic patterns that arise in one and two dimensions; see Examples 2.1, 2.3, and 2.5, and the example from Gu, Bates, Chen & Wahba (1989) discussed in Section 6. Hooper, Mayes & Demianczuk (2001) obtained a useful approximation for a covariance function using $K = 3$. For $K \geq 4$ the possibilities are more varied and harder to characterize. Given the properties of the basis functions, I would expect ALB to work well when the covariate space can be covered with a small number of overlapping regions where f is well approximated by simple low-dimensional functions. ALB allows the local subspace on which $\hat{f}$ is implicitly defined to vary smoothly from one region of the covariate space to another. In the Boston housing example, different one-dimensional approximations are obtained for regions with low and high prices.

The affine invariance of ALB suggests comparison with projection pursuit regression (Friedman & Stuetzle 1981). Both methods employ a linear combination of simpler functions and neither is affected by rotation or scaling of the covariates. Projection pursuit approximates f by a sum of one-dimensional ridge functions $f(\mathbf{x}) \approx \sum h_k(\boldsymbol{\beta}_k'\mathbf{x})$. The ridge functions h_k are estimated using one-dimensional smoothers and can incorporate several bumps. The logistic basis functions employed by ALB are more complex than ridge functions in one respect, being multi-dimensional, but are simpler in other respects, with configuration and quasi-concave shape constrained by a parametric family.

Affine invariance is a mixed blessing. For some applications, it is a desirable property. In the Viking formation example, there is no reason to think that latitude and longitude are well-suited for modeling elevation. Alternative characterizations of spatial location should work just as well. For other applications, the function f may exhibit simple structure related to the covariates, such as $f(\mathbf{x}) = \sum h_k(x_k)$. Methods that exploit this structure have an advantage, e.g., generalized additive models (Hastie and Tibshirani 1990), multivariate adaptive regression splines (Friedman

1991), and the Π method (Breiman 1991). The performance of such methods will usually suffer if the covariates are rotated. An affine invariant method provides a minimax performance that is unaffected by linear transformation.

Logistic functions are usually defined using the linear parameterization in expression (2). An alternative parameterization is also useful. The ALB functions can be expressed in terms of Euclidean distance from reference points $\boldsymbol{\xi}_k$ in the covariate space

$$\phi_k(\mathbf{x}) = \exp(\gamma_k - \tau^{-2}||\mathbf{x} - \boldsymbol{\xi}_k||^2) / \sum_{m=1}^{K} \exp(\gamma_m - \tau^{-2}||\mathbf{x} - \boldsymbol{\xi}_m||^2) \; . \tag{4}$$

Note that the term $\exp(-\tau^{-2}||\mathbf{x}||^2)$ factors out in the numerator and denominator of (4). The two parameterizations can thus be related by

$$\alpha_k = \gamma_k - \tau^{-2}||\boldsymbol{\xi}_k||^2 \quad \text{and} \quad \boldsymbol{\beta}_k = 2\tau^{-2}\boldsymbol{\xi}_k .$$

The reference point parameterization is easier to interpret than the linear parameterization. Roughly speaking, the location of ϕ_k can be controlled by $\boldsymbol{\xi}_k$, the relative influence of ϕ_k can be controlled by γ_k, and the smoothness of ϕ_k can be controlled by τ. This interpretation is useful when initializing parameter values for estimation. The interpretation oversimplifies matters to some degree. The roles of the parameters are actually not so clearly separated due to redundancies among the parameters. For example, τ can be fixed without limiting the generality of (4), and smoothness can be controlled by adjusting the remaining parameters. This is in fact the approach adopted in Section 4 when estimating f_K. The details underlying the interpretation are spelled out as follows.

PROPOSITION 4. *Set $\zeta_k = \tau^2\gamma_k$ and define*

$$A_k = \{\mathbf{x} : ||\mathbf{x} - \boldsymbol{\xi}_k||^2 - \zeta_k < ||\mathbf{x} - \boldsymbol{\xi}_m||^2 - \zeta_m \;\; \text{for all } m \neq k\}. \tag{5}$$

(i) *We have $A_k = \{\mathbf{x} : \phi_k(\mathbf{x}) > \phi_m(\mathbf{x})$ for all $m \neq k\}$. Each A_k is a convex set, possibly empty. The boundary between two neighbouring sets A_k and A_m is a subset of a hyperplane orthogonal to $\boldsymbol{\xi}_k - \boldsymbol{\xi}_m$.*

(ii) *If the ζ_k are all equal, then $\{A_k\}$ forms a Dirichlet tessellation of $\Re^d$; i.e., A_k consists of all $\mathbf{x}$ nearest to $\boldsymbol{\xi}_k$. If the ζ_k differ substantially, then the spatial interpretation of the $\boldsymbol{\xi}_k$ is less clear; e.g., it is possible that $\boldsymbol{\xi}_k \notin A_k$.*

(iii) *We have $\partial\phi_m/\partial\gamma_k = \phi_k(1 - \phi_k)$ for $m = k$, and $-\phi_k\phi_m$ for $m \neq k$, so increasing γ_k increases the influence of ϕ_k and diminishes that of other ϕ_m.*

(iv) *Fix $\zeta_1, \boldsymbol{\xi}_1, \ldots, \zeta_K, \boldsymbol{\xi}_K$. As τ approaches 0, $\phi_k(\mathbf{x})$ converges to the indicator function of A_k, for all $\mathbf{x}$ not on the boundary of A_k. As τ approaches ∞, $\phi_k(\mathbf{x})$ converges to $1/K$.*

Proof. Direct calculation. □

The functions f_K can be viewed as neural networks. If the reference point parameterization is employed and $\gamma_k = \gamma$ is fixed, then f_K is a radial basis functions network of a type introduced by Moody & Darken (1989). It is common practice in applications of such networks to replace τ with varying parameters τ_k. This permits variation in the receptive field size, i.e., the volume of upper level sets of the basis functions. Proposition 4(iii) shows that fixing τ and varying γ_k has a similar effect. The (τ, γ_k) family of functions has a potential advantage over the (τ_k, γ) family. Logistic basis functions are affine invariant while radial basis functions are not.

When the linear parameterization (2) is employed, f_K can be represented as a network with d inputs, K nodes in a single hidden layer, and one output. This is not a feed-forward network, however, because the logistic transformation (called softmax in neural network literature) involves all nodes in the hidden layer. In feed-forward networks, such as

$$\sum_{k=1}^{K} \delta_k \exp(\alpha_k + \boldsymbol{\beta}_k'\mathbf{x})/\{1+\exp(\alpha_k + \boldsymbol{\beta}_k'\mathbf{x})\}, \tag{6}$$

each hidden node is transformed separately. The basis functions in (6) are ridge functions with sigmoidal shape. The basis functions in (2) appear to have an advantage over those in (6) with regard to estimation. While similar training algorithms can be applied to both models, the effectiveness of these algorithms depends on the initial values chosen for the parameters. A spatial interpretation of the reference points in (4) permits an effective initialization using a clustering algorithm, which is described in the next section. This approach is not feasible for the $\boldsymbol{\beta}_k$ in (6). Initial weights in feed-forward neural networks are usually generated randomly (Ripley 1996).

The family $\{f_K, K \geq 1\}$ possesses the universal approximation property (Hornik, Stinchcombe & White 1989), i.e., if f is a continuous function defined on a compact set A, then there exists a sequence (f_K) converging uniformly to f on A. This result is easily demonstrated using Proposition 4. Given $\varepsilon > 0$, choose K and subsets A_k of the form (5) such that f varies by at most $\varepsilon/2$ within each A_k. This can be arranged, for example, by setting all γ_k to zero and choosing the $\boldsymbol{\xi}_k$ so that the maximum of $\{\|\mathbf{x} - \boldsymbol{\xi}_k\| : \mathbf{x} \in A_k, k \leq K\}$ is sufficiently small. Let τ approach 0, so that f is essentially approximated by a piecewise constant function. Details of the proof are omitted because the result follows from the universal approximation property of radial basis functions (Xu, Kryzak & Yuille 1994).

4. ESTIMATION

4.1. Estimation of f_K by stochastic approximation.

Stochastic approximation was introduced by Robbins & Monro (1951). The following brief review follows Benveniste, Métivier & Priouret (1990). Consider minimizing a function $Q(\boldsymbol{\theta})$ using an iterative algorithm driven by a sequence of independent and identically distributed random vectors $\mathbf{z}_m$,

$$\boldsymbol{\theta}_m = \boldsymbol{\theta}_{m-1} + a_m \mathbf{H}(\boldsymbol{\theta}_{m-1}, \mathbf{z}_m). \tag{7}$$

Let the gain function a_m satisfy $a_m > 0$, $\sum a_m = \infty$, and $\sum a_m^\alpha < \infty$ for some $\alpha > 1$. Write $\boldsymbol{\theta}_m = \boldsymbol{\theta}(t_m)$, where $t_m = \sum_{i=1}^m a_i$. After an initial transient phase, the behaviour of process (7) is represented to a first approximation by that of the differential equation $d\boldsymbol{\theta}(t)/dt = \mathrm{E}[\mathbf{H}\{\boldsymbol{\theta}(t), \mathbf{z}\}]$. In stochastic gradient algorithms, the updating function $\mathbf{H}$ is defined so that $-\mathrm{E}\{\mathbf{H}(\boldsymbol{\theta}, \mathbf{z})\}$ is proportional to the gradient of $Q(\boldsymbol{\theta})$.

This section describes a stochastic gradient algorithm to estimate f by $f_K = \sum_{k=1}^K \delta_k \phi_k$ for given K. While the underlying idea is simple, its implementation involves several engineering details, e.g., choosing the number of iterations and the form of the gain function. The choices described below were made largely on empirical grounds, guided by experience with similar classification training algorithms (Hooper 1999, 2001). The implementation of the algorithm treats these choices as default values, to be ignored in most applications, but subject to adjustment by the user.

Our underlying aim is to minimize the predictive risk

$$R(f_K) = \mathrm{E}_P\{|y - f_K(\mathbf{x})|^q\},$$

where E_P denotes the expected value when sampling $(\mathbf{x}, y)$ from a population of interest. Let $\hat{P}$ denote the empirical distribution, assigning probability $1/n$ to each observation $(\mathbf{x}_i, y_i)$. The L_q

estimator $\hat{f}_K$ minimizes the training risk

$$\mathrm{E}_{\hat{P}}\{|y - f_K(\mathbf{x})|^q\} = \frac{1}{n}\sum_{i=1}^{n} |y_i - f_K(\mathbf{x}_i)|^q. \tag{8}$$

The response and covariates are centered and scaled to have zero mean and unit standard deviation. This standardization makes it easier to initialize parameter values and specify updating formulae. After estimation, $\hat{f}_K$ is transformed back to the original scale. The reference point parameterization (4) is employed, with τ fixed at a convenient value described below. Set

$$\boldsymbol{\theta}' = (\delta_1, \gamma_1, \boldsymbol{\xi}_1', \ldots, \delta_K, \gamma_K, \boldsymbol{\xi}_K')$$

and let $\hat{\boldsymbol{\theta}}$ denote a parameter vector defining $\hat{f}_K$. Since the parameters are not uniquely determined, $\hat{\boldsymbol{\theta}}$ is not regarded as an estimator but as one of many equivalent parameterizations of $\hat{f}_K$.

Initial parameter values are motivated by Proposition 4. The initial γ_k are set to zero. The initial $\boldsymbol{\xi}_k$ are obtained as a spatially representative set of points in the covariate space (see below). The initial δ_k are then defined as the average of the y_i values for $\mathbf{x}_i$ in the region nearest to $\boldsymbol{\xi}_k$. The parameter τ is set to the average distance between nearest neighbours among the K initial points $\boldsymbol{\xi}_k$. This choice for τ yields a reasonable amount of overlap among neighbouring basis functions.

A representative set of K points $\boldsymbol{\xi}_k$ can be obtained by minimizing

$$\sum_{i=1}^{n} \min\{\|\mathbf{x}_i - \boldsymbol{\xi}_k\|^2 : k = 1, \ldots, K\}. \tag{9}$$

The resulting points have been called K-means cluster centroids (MacQueen 1967) and principal points (Flury 1990). The latter term is more appropriate here, as we are not searching for clusters. The initial $\boldsymbol{\xi}_k$ can be calculated using a K-means clustering algorithm (Hartigan & Wong 1979). My preference, however, is to initialize both $\boldsymbol{\xi}_k$ and δ_k simultaneously, using a vector quantization algorithm (Kohonen 1995). Begin by generating $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_K$ randomly from $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ and set all δ_k to zero. Then repeat the following steps for $3000\sqrt{K}$ iterations. At the mth iteration, sample $(\mathbf{x}, y)$ from $\hat{P}$, determine the point $\boldsymbol{\xi}_k$ nearest $\mathbf{x}$, replace $\boldsymbol{\xi}_k$ by $(1 - a_m)\boldsymbol{\xi}_k + a_m\mathbf{x}_k$, and replace δ_k by $(1 - a_m)\delta_k + a_m y$. The gain is defined as $a_m = 100\sqrt{K}/(m + 100\sqrt{K})$. This algorithm produces approximate principal points and y-averages, which serve as useful starting values.

After selecting initial values, the training risk (8) is minimized by stochastic approximation. In successive iterations, an observation $(\mathbf{x}, y)$ is randomly sampled (with replacement) from $\hat{P}$ and the parameter vector $\boldsymbol{\theta}$ is updated as in expression (7). Set

$$h_k(\mathbf{x}, y, \boldsymbol{\theta}) = |y - f_K(\mathbf{x})|^{q-1}\mathrm{sign}\{y - f_K(\mathbf{x})\}\phi_k(\mathbf{x}).$$

Differentiation of $-|y - f_K(\mathbf{x})|^q$ with respect to the parameters yields the following updating formulae at the mth iteration:

$$\begin{array}{lcl} \delta_k & \leftarrow & \delta_k + a_m^{\delta} h_k(\mathbf{x}, y, \boldsymbol{\theta}) \ , \\ \gamma_k & \leftarrow & \gamma_k + a_m^{\gamma} h_k(\mathbf{x}, y, \boldsymbol{\theta})\{\delta_k - f_K(\mathbf{x})\} \ , \\ \boldsymbol{\xi}_k & \leftarrow & \boldsymbol{\xi}_k + a_m^{\xi} h_k(\mathbf{x}, y, \boldsymbol{\theta})\{\delta_k - f_K(\mathbf{x})\}(\mathbf{x} - \boldsymbol{\xi}_k). \end{array} \tag{10}$$

The parameter τ remains fixed. The updates make sense, based on the interpretation of the parameters in Proposition 4. If $y - f_K(\mathbf{x})$ is positive (negative), then δ_k is increased (decreased). If the product $\{y - f_K(\mathbf{x})\}\{\delta_k - f_K(\mathbf{x})\}$ is positive (negative), then γ_k is increased (decreased) and $\boldsymbol{\xi}_k$ is shifted toward (away from) $\mathbf{x}$. The magnitude of each change depends on $\phi_k(\mathbf{x})$ and, if $q > 1$, on $|y - f_K(\mathbf{x})|$.

The number of iterations is set at $M = 50000\sqrt{K}$. The gains are defined as follows:

$$a_m^\xi = \begin{cases} a_0^\xi \frac{c_g M}{m + c_g M}, & 1 \le m \le M/2, \\ a_{M/2}^\xi \frac{2(M-m)}{M}, & M/2 < m \le M, \end{cases}$$

$a_m^\delta = a_m^\xi$, $a_m^\gamma = a_m^\xi/2$, $a_0^\xi = 0.25$, and $c_g = 0.01$. The updating functions and gain functions were scaled in an attempt to make the three perturbations in (10) have effects of similar magnitude on $f_K(\mathbf{x})$.

The theory of stochastic approximation indicates that after an initial transient phase, the training process typically converges toward a local optimum (Benveniste, Métivier & Priouret 1990). There is no guarantee that a global optimum will be found, and replication of the process could produce varying results, but the algorithm typically yields reasonable results. The quality of the estimator is improved and variation under replication is reduced by restarting the process, i.e., replicate the first 10% of the process ten times, calculating the training risk each time, then continue the process with the most promising vector of parameter values.

Four comments. First, note that the updates (10) for the L_1 estimator depend on the deviation $y - f_K(\mathbf{x})$ only through its sign. This shows that the L_1 estimator is robust against outliers and heteroscedasticity in the response variable. Second, the number M of iterations increases slowly with the complexity of the fitted model but does not depend on n. When n is small to moderate, each observation is sampled many times, but when n is very large, some observations may not be sampled at all. In the Viking formation example, with $K = 12$ and $n = 74229$, each observation is sampled on average 2.3 times. One might want to increase M in such situations. Third, although I have not done so, one could exploit the conditional linearity of the model when estimating the δ_k coefficients for the L_2 estimator, e.g., obtain preliminary estimates by stochastic approximation, then fix the basis functions and obtain exact least squares estimates of the δ_k. Care would be needed to deal with potential problems of multicollinearity. Fourth, for the L_2 estimator, stochastic approximation can be replaced by a nonlinear least squares algorithm, such as Gauss–Newton or Newton–Raphson (Bates & Watts 1988). These "batch" algorithms require fewer iterations than the "on-line" algorithm described above, but they typically employ the entire data set at each iteration. I have found stochastic approximation to be highly effective in problems with large data sets and large numbers of parameters. Randomness may help in the search for a good local optimum, given a poor initial function estimate.

4.2. Selection of K by generalized cross-validation.

Our aim is to select a number $\hat{K}$ so that $R(\hat{f}_{\hat{K}}) \approx \min_K R(\hat{f}_K)$. To this end, $\hat{K}$ is obtained by minimizing an adjusted training risk

$$R_{\mathrm{GCV}}(\hat{f}_K) = \left(\frac{n}{n-p}\right)^q \frac{1}{n} \sum_{i=1}^n |y_i - \hat{f}_K(\mathbf{x}_i)|^q, \tag{11}$$

where $p = 1 + (K-1)(d+2)$. This adjustment, called generalized cross-validation, was originally introduced for L_2 loss and linear smoothers (Craven & Wahba 1979). Its application here is justified primarily on empirical grounds. In simulation studies, with R approximated using a large test set, $R(\hat{f}_{\hat{K}})$ was typically close to $\min_K R(\hat{f}_K)$. A straightforward search is employed to minimize (11). The GCV risk is evaluated for successive values of K, starting with $K = 1$. The search halts when the minimum GCV risk remains unchanged for m consecutive values of K. The selection of $\hat{K}$ therefore involves the calculation of $\hat{K} + m$ estimates $\hat{f}_K$. This simple strategy works well because computation time increases rapidly with K (see below) and typically $\hat{K} \leq 10$. The stopping value $m = 3$ is adequate in most situations. A larger value may be useful if $R_{\mathrm{GCV}}(\hat{f}_K)$ is unusually flat as a function of K.

The adjusted risk $R_{\mathrm{GCV}}(\hat{f}_{\hat{K}})$ is not always a good estimator of $R(\hat{f}_{\hat{K}})$. Simulation studies reported in Section 6 show that when the sample size is small, the ratio $R_{\mathrm{GCV}}(\hat{f}_{\hat{K}})/R(\hat{f}_{\hat{K}})$ could be highly variable and the average ratio could be significantly less than one. These findings suggest that if an estimate of $R(\hat{f}_{\hat{K}})$ is required, then $R_{\mathrm{GCV}}(\hat{f}_{\hat{K}})$ should be supplemented with a more reliable estimate, such as a 10-fold cross-validated risk estimate or a bootstrap estimate. The findings do not invalidate the use of GCV in selecting K, because risk estimates for consecutive values of K are highly correlated. In simulations, plots of $R_{\mathrm{GCV}}(\hat{f}_K)$ and $R(\hat{f}_K)$ against K often differ substantially while still attaining their minima at the same value K.

When p/n is small, the GCV criterion (11) is closely related to AIC (Akaike 1973). Suppose the conditional density of y given $\mathbf{x}$ has the form

$$\frac{c}{\sigma} \exp\left\{-\frac{|y - f_K(\mathbf{x})|^q}{q\sigma^q}\right\},$$

where c is a normalizing constant and σ is a scale parameter. For $q = 2$, we have the normal density, and for $q = 1$, the double exponential. Let $L(\boldsymbol{\theta}, \sigma)$ denote the likelihood function based on the conditional density of $(y_1, \ldots, y_n)$ given $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$. A maximum likelihood estimator for $\boldsymbol{\theta}$ yields the ALB L_q estimator $\hat{f}_K$, and the mle for σ is

$$\hat{\sigma}_K = \left\{\frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{f}_K(\mathbf{x}_i)|^q\right\}^{1/q}.$$

Using AIC, we would select K to minimize $-\log L(\hat{\boldsymbol{\theta}}_K, \hat{\sigma}_K) + p$. This is equivalent to minimizing

$$\log(\hat{\sigma}_K^q) + qp/n \approx \log(\hat{\sigma}_K^q) - q\log(1 - p/n). \qquad (12)$$

The right-hand side of (12) is the logarithm of the GCV criterion (11). As p/n increases, GCV assigns an increasingly heavier penalty relative to AIC. In particular, GCV imposes the restriction $p < n$, while AIC does not.

A potential misconception about model fitting warrants the following comment. Many regression methods select from a large set of potential basis functions using forward selection and/or backward elimination strategies. ALB regression adopts a different approach. While K is selected by sequentially calculating $\hat{f}_K$, parameters are optimized separately for each K. The parameters and basis functions determining $\hat{f}_K$ play no role in the calculation of $\hat{f}_{K+1}$. The estimator $\hat{f}$ would not be improved by pruning basis functions because parameters are optimized jointly for all K basis functions.

The computation time required to estimate f, including selection of $\hat{K}$, is typically between 5 and 30 seconds, fast enough for interactive use. The time increases with $\hat{K}$ and d (but does not depend on q) and increases very slowly with n. Table 1 lists ALB estimation times for a 360 MHz SUN UltraSparcII workstation. Each value includes the total time needed to obtain $\hat{f}_K$ for $K = 1, \ldots, \hat{K} + 3$. The sample size was $n = 500$. The time is roughly linear in d, with intercept and slope depending on $\hat{K}$, and roughly linear in $\hat{K}^2$ with slope depending on d. The sample size n has relatively little effect on time because of the sampling technique used in the training algorithm. The number M of iterations is proportional to $\sqrt{K}$, each iteration requires the evaluation of K distances, and each distance calculation time is proportional to d. Times increase slowly with n because of increased overhead (data input and transformation, calculation of the GCV risk) and a tendency to select larger $\hat{K}$. In the Viking formation example with $n = 74299$ and $\hat{K} = 12$, the computing time was 150 seconds.

TABLE 1: Time (in seconds) to calculate $\hat{f}$, including selection of $\hat{K}$.

	d			
$\hat{K}$	1	5	10	20
1	2.6	3.1	3.6	5.1
2	3.5	4.8	5.6	8.1
5	9.2	12.6	15.2	22.4
10	27.0	37.5	44.6	66.1
15	56.6	79.3	94.6	140.7

5. STANDARD ERRORS

This section presents approximate standard errors for the ALB L_2 estimator and the components of its gradient. The reader is advised that the derivation makes unwarranted assumptions and the approximations are unreliable in some situations. The derivation assumes that $K = \hat{K}$ is fixed, $f(\mathbf{x}) = f_K(\mathbf{x})$ for some $\boldsymbol{\theta} \in \Re^{(2+d)K}$, and the $y_i - f(\mathbf{x}_i)$ are normally distributed with zero mean and constant variance σ_ε^2. Standard errors are obtained using a standard asymptotic technique in nonlinear regression analysis, e.g., see Bates & Watts (1988, Section 2.3). The regression function $f_K(\mathbf{x})$ is approximated locally near $\hat{\boldsymbol{\theta}}$ by a linear function of $\boldsymbol{\theta}$. Linear regression formulae are then applied. The redundant parameterization of $\boldsymbol{\theta}$ presents a potential problem here. Further difficulties could arise from multicollinearity among the estimated basis functions. Both problems are resolved in a simple manner by a ridge regression technique.

Set

$$\mathbf{v}(\mathbf{x}, \boldsymbol{\theta}) = \partial f_K(\mathbf{x})/\partial \boldsymbol{\theta} \quad \text{and} \quad \mathbf{A} = \sum_{i=1}^{n} \mathbf{v}(\mathbf{x}_i, \hat{\boldsymbol{\theta}})\{\mathbf{v}(\mathbf{x}_i, \hat{\boldsymbol{\theta}})\}'.$$

The matrix $\mathbf{A}$ is singular and nonnegative definite symmetric with positive diagonal elements. An invertible matrix $\mathbf{B}$ is obtained by slightly increasing (multiply by 1.01) each diagonal element of $\mathbf{A}$ and leaving off-diagonal elements unchanged. The error variance is estimated by

$$\hat{\sigma}_\varepsilon^2 = \frac{1}{n-p} \sum_{i=1}^{n} \{y_i - \hat{f}(\mathbf{x}_i)\}^2,$$

where $p = 1 + (K-1)(d+2)$. Given $\mathbf{x}$, the standard deviation of $\hat{f}(\mathbf{x})$ is estimated by

$$\mathrm{se}\{\hat{f}(\mathbf{x})\} = \hat{\sigma}_\varepsilon \left[\{\mathbf{v}(\mathbf{x}, \hat{\boldsymbol{\theta}})\}'\mathbf{B}^{-1}\mathbf{v}(\mathbf{x}, \hat{\boldsymbol{\theta}})\right]^{1/2}. \tag{13}$$

The interval $\hat{f}(\mathbf{x}) \pm 2\mathrm{se}\{\hat{f}(\mathbf{x})\}$ provides an approximate confidence interval for $f(\mathbf{x})$ with nominal 95% coverage probability. One may note four potential problems with this simple confidence interval. First, the ridge modification to $\mathbf{A}$ produces a slight downward bias in the standard error. Second, the quality of the linear approximation of f_K may be poor. Bates & Watts (1988) suggested methods for identifying curvature effects and modifying confidence regions, but their methods may not be tractable in the applications considered here. Third, standard errors account for variance, but not bias. If f is poorly approximated by f_K, then $\hat{f}_K(\mathbf{x})$ may have substantial bias relative to its standard deviation. Fourth, and perhaps most important, the derivation assumes that K is fixed. The effect of adaptive selection is unknown. It seems likely that variation in $\hat{K}$ will increase variation in $\hat{\sigma}_\varepsilon$, and hence in the standard error.

The simulation studies in Section 6 suggest that the confidence interval can be moderately liberal with coverage probabilities (averaged over $\mathbf{x}_1, \ldots, \mathbf{x}_n$) between 85% and 97%. These results reveal only part of the story, since coverage could also vary across the covariate space. For a simple example, suppose $d = 1$, x is uniformly distributed over $(0, 2\pi)$, $f(x) = \sin(x)$, and $\sigma_\varepsilon = 0.5$. The sine function is well approximated by f_K with $K = 4$, but smaller values of $\hat{K}$ are often selected when n is not large, e.g., $P(\hat{K} \leq 3) \approx 0.20$ when $n = 100$. When $\hat{K} = 3$, $\hat{f}$ typically fits f well over much of the interval, but fits an asymptote at one end. In the region with the asymptote, $|\hat{f} - f|$ tends to be large while $\mathrm{se}(\hat{f})$ tends to be small. Coverage probabilities are thus lower for values of x near the ends of the interval. This problem of K being underestimated, and bias thereby being increased, appears to diminish as n increases and/or the signal-to-noise ratio increases.

Approximate standard errors for gradient components can be defined in a manner similar to (13). Set

$$g_j(\mathbf{x}) = \frac{\partial}{\partial x_j} f(\mathbf{x}) \ , \quad \hat{g}_j(\mathbf{x}) = \frac{\partial}{\partial x_j} \hat{f}_K(\mathbf{x}) \ , \quad \mathbf{w}_j(\mathbf{x}, \boldsymbol{\theta}) = \frac{\partial^2}{\partial \boldsymbol{\theta} \partial x_j} f_K(\mathbf{x}) \ ,$$

where $\mathbf{x}' = (x_1, \ldots, x_d)$. Given $\mathbf{x}$, the standard deviation of $\hat{g}_j(\mathbf{x})$ is estimated by

$$\mathrm{se}\{\hat{g}_j(\mathbf{x})\} = \hat{\sigma}_\varepsilon \left[\{\mathbf{w}_j(\mathbf{x}, \hat{\boldsymbol{\theta}})\}' \mathbf{B}^{-1} \mathbf{w}_j(\mathbf{x}, \hat{\boldsymbol{\theta}}) \right]^{1/2} .$$

The interval $\hat{g}_j(\mathbf{x}) \pm 2\mathrm{se}\{\hat{g}_j(\mathbf{x})\}$ provides an approximate confidence interval for $g_j(\mathbf{x})$ with nominal 95% coverage probability. The actual coverage probability is less stable than that for the $f(\mathbf{x})$ interval, perhaps due to increased bias in the gradient estimator. The estimator $\hat{f}_K$ tends to flatten near the edge of the data, shrinking $\hat{g}_j$ toward zero. The simulation studies in Section 6 (results omitted from Table 3) revealed coverage probabilities (averaged over $\mathbf{x}_1, \ldots, \mathbf{x}_n$) between 75% and 100%. The higher coverage probabilities occur when g_j is identically zero, i.e., when the covariate x_j is a nuisance variable. Variation in coverage across the covariate space appears to be greater for derivatives $g_j(\mathbf{x})$ than for $f(\mathbf{x})$.

In view of the problems noted above, one may wish to restrict application of the standard errors to exploratory analysis and adopt alternative methods, such as the bootstrap, for more formal inference. I have found the standard errors useful in two regards. First, plots of $\mathrm{se}\{\hat{f}(\mathbf{x})\}$ can be used to detect outliers in the covariate space. When using $\hat{f}$ to predict a response at a new point $\mathbf{x}$, it is not always clear whether $\mathbf{x}$ lies within the available data. A large standard error suggests that the prediction involves extrapolation and is thus likely to be affected by increased bias and variance. Second, boxplots of the standardized gradients $\hat{g}_j(\mathbf{x})/\mathrm{se}\{\hat{g}_j(\mathbf{x})\}$ may suggest possible nuisance variables. If f does not involve the covariate x_j, then $g_j(\mathbf{x}) = 0$ for all $\mathbf{x}$. Elimination of such variables can substantially improve the fit. It should noted that standardized gradient plots, like t statistics for linear regression coefficients, could be misleading given dependencies among the covariates.

6. SIMULATION STUDIES

The accuracy of the ALB L_2 estimator was investigated in simulation experiments. Various examples were chosen to investigate how comparative performance depends on the target function f, the dimensionality d, and the sample size n. In each example, 100 samples of n independent observations of $(\mathbf{x}, y)$ were generated from a model: $y = f(\mathbf{x}) + \varepsilon$, where $\mathbf{x}$ and ε are independent, $\mathbf{x}$ is distributed uniformly on a hypercube $(a, b)^d$, and ε is a $N(0, \sigma_\varepsilon^2)$ random variable. The ALB L_2 estimate $\hat{f}$ was calculated and several performance measures were evaluated for each sample. Averages and standard deviations of the performance measures are reported in Table 3. There are ten basic examples. Each has several values of n to demonstrate how accuracy improves with increased sample size. Some have several values of d to demonstrate the adverse effects of nuisance variables. Table 2 lists characteristics of the basic examples: the function f, the side (a, b), σ_f = the standard deviation of $f(\mathbf{x})$, and σ_ε. The signal-to-noise ratio is $\sigma_f / \sigma_\varepsilon$.

TABLE 2: Examples used in simulation studies.

No.	Function f	Side	σ_f	σ_ε	$\sigma_\varepsilon^2/\sigma_y^2$
1	constant	$(0,1)$	0.00	1.00	1.000
2	$\exp\{x_1 \sin(\pi x_2)\}$	$(-1,1)$	0.45	0.50	0.552
3	$3\sin(x_1 x_2)$	$(-2,2)$	1.90	1.00	0.217
4	$x_1 x_2 x_3$	$(-2,2)$	1.50	1.00	0.308
5	additive (16)	$(0,1)$	2.52	1.00	0.136
6	partly additive (17)	$(0,1)$	4.90	1.00	0.040
7	impedance (18)	$(0,1)$	354.	118.	0.100
8	phase shift (19)	$(0,1)$	0.33	0.11	0.100
9	GBCW (20)	$(0,1)$	3.10	1.00	0.094
10	ALB $(K=5, d=4)$	$(0,1)$	0.53	0.18	0.100

Two measures of predictive accuracy are reported for each $\hat{f}$. Each measure is scaled by dividing by $\sigma_y^2 = \sigma_f^2 + \sigma_\varepsilon^2$. The scaled mean predictive squared error evaluates accuracy at points within the observed sample:

$$\mathrm{MPSE} = \left[\frac{1}{n}\sum\{f(\mathbf{x}_i) - \hat{f}(\mathbf{x}_i)\}^2 + \sigma_\varepsilon^2\right] \Big/ \sigma_y^2 \,.$$

The scaled integrated predictive squared error evaluates accuracy over the hypercube:

$$\mathrm{IPSE} = \left[\int\{f(\mathbf{x}) - \hat{f}(\mathbf{x})\}^2 P_u(d\mathbf{x}) + \sigma_\varepsilon^2\right] \Big/ \sigma_y^2 \,, \tag{14}$$

where P_u is the uniform distribution on $(a,b)^d$. The numerator of (14) was approximated by averaging the predictive squared error of $\hat{f}$ over an independent sample of $99n$ observations. Usually IPSE is greater than MPSE, and the difference could be large when n is small or d is large. Both predictive measures are bounded below by $\sigma_\varepsilon^2/\sigma_y^2$ listed in Table 2.

Averages for three additional measures are reported as well: $\hat{K}$, CPf, and GCV/IPSE. CPf is the observed coverage probability of a nominal 95% confidence interval for $f(\mathbf{x})$, averaged over the sample:

$$\mathrm{CP}f = \frac{1}{n}\sum_{i=1}^{n} I\left[|f(\mathbf{x}_i) - \hat{f}(\mathbf{x}_i)| \le 2\mathrm{se}\{\hat{f}(\mathbf{x}_i)\}\right].$$

GCV/IPSE denotes the ratio of scaled GCV risk to IPSE, where the unscaled GCV risk is defined in (11). The tabulated values suggest that the bias in the GCV risk is small when MPSE $\approx$ IPSE but that the GCV risk tends to underestimate the integrated predictive squared error when MPSE is substantially less than IPSE. The average ratio of scaled GCV risk to MPSE (which is not tabulated) is typically greater than one. As the sample size n increases, the difference IPSE $-$ MPSE and the bias in the GCV risk both decrease.

Adaptive estimators should detect real structure where it exists and ignore spurious structure caused by random variation. Example 1 focuses on this second goal by examining performance when f is constant. We would hope that, in most samples, ALB selects $\hat{K} = 1$ so that $\hat{f} = \bar{y}$ and IPSE $\approx 1 + 1/n$. Table 3 shows how accuracy deteriorates as d increases and improves as n increases. In the 100 replicates for each of the four (d, n) pairs, ALB selected $\hat{K} = 1$ with frequencies 96, 74, 64, and 82.

Examples 2, 3, and 4 were used by Breiman (1991) to illustrate the Π method for multivariate function estimation. The Π method employs approximations of the form

$$f(\mathbf{x}) \approx \sum_{m=1}^{M} \prod_{j=1}^{d} h_{mj}(x_j), \tag{15}$$

where the h_{mj} are smooth functions. The target functions f in the three examples are all well approximated by (15) with $M \leq 2$, so the Π method performs very well here. Breiman's root mean squared error results (for $n = 100$) are equivalent to MPSE values of 0.64, 0.25, and 0.33. The less efficient performance of ALB, observed in Table 3, stems from the the target functions being less easily approximated by logistic basis functions. This can be seen by comparing average numbers of degrees of freedom used in the two approaches. Breiman (1991) reported $\text{ave}\,df$ = 6.5, 13.1, and 5.0. Corresponding values for ALB (with $n = 100$) are $1 + (\text{ave}\,\hat{K} - 1)(d + 2) =$ 14.9, 20.8, and 34.5. The difference in comparative performance is greatest in Example 4, where the target function is extremely simple for the Π method but relatively complex for ALB. In support of ALB, however, note that the Π method is sensitive to the coordinate system used to describe the covariates. If the coordinate axes were randomly rotated, then a larger number of products would likely be needed for a good approximation in (15), and the performance of the Π method would suffer. The performance of ALB would be unaffected by such a rotation. The target functions are relatively complex, with several ripples or bumps. ALB adaptively selects larger values of $\hat{K}$ as n increases.

Examples 5 through 8 were used by Friedman (1991) to illustrate MARS. The target in Example 5 is an additive function of the first five covariates:

$$f(\mathbf{x}) = 0.1\exp(4x_1) + 4/[1 + \exp\{-20(x_2 - 0.5)\}] + 3x_3 + 2x_4 + x_5\ . \tag{16}$$

Friedman reported IPSE values for MARS (with $d = 10$ and $n = 50, 100, 200$) of 0.26, 0.18, and 0.16. The target in Example 6 is a partly additive function:

$$f(\mathbf{x}) = 10\sin(\pi x_1 x_2) + 20(x_3 - 0.5)^2 + 10x_4 + 5x_5\ . \tag{17}$$

Friedman's IPSE values (again with $d = 10$ and $n = 50, 100, 200$) were 0.24, 0.074, and 0.056. ALB results are reported in Table 3. In both examples the results for $d = 10$ are substantially worse than those of MARS, but the results for $d = 5$ are only slightly worse. The higher accuracy of MARS is obtained to some extent by exploiting the additive and partly additive structure of f and by effectively eliminating the nuisance variables $x_6, \ldots, x_{10}$. A rotation of the coordinate axes would create higher-order interactions, adversely affecting MARS but not ALB.

Plots of the ALB gradient functions $\hat{g}_j(\mathbf{x})$ can be used to detect additive and partially additive structure. If the effect of x_1 is additive, as in Example 5, then the gradient $g_1(\mathbf{x})$ is a function of x_1. If the joint effect of (x_1, x_2) is additive, as in Example 6, then the gradients $g_1(\mathbf{x})$ and $g_2(\mathbf{x})$ are functions of (x_1, x_2). In these examples, plots of the ALB gradient estimates reveal little scatter about the gradient curve. These ALB diagnostics would suggest the use of alternative methods, such as MARS, that exploit additivity. As noted in Section 5, plots of standardized gradient functions can be used to detect nuisance variables.

TABLE 3: Performance measures: averages (and standard deviations) from 100 replicated samples of size n. MPSE and IPSE are bounded below by $\sigma_\varepsilon^2/\sigma_y^2$ listed in Table 2.

No.	d	n	$\hat{K}$	MPSE	IPSE	GCV/IPSE	CPf
1	1	100	1.04(.20)	1.01(.02)	1.01(.03)	1.00(.16)	0.95(.22)
	5	100	1.38(.76)	1.06(.11)	1.07(.14)	0.94(.17)	0.90(.23)
	10	100	1.60(.97)	1.14(.21)	1.23(.38)	0.86(.22)	0.88(.23)
	10	400	1.26(.63)	1.02(.04)	1.02(.05)	0.98(.08)	0.90(.25)
2	2	100	4.47(.83)	0.66(.03)	0.70(.04)	1.02(.14)	0.92(.06)
	2	200	5.49(1.2)	0.63(.02)	0.64(.03)	1.00(.12)	0.90(.08)
	2	400	6.64(.94)	0.59(.01)	0.60(.02)	1.02(.08)	0.93(.05)
3	2	100	5.96(1.7)	0.29(.03)	0.34(.04)	1.03(.23)	0.92(.08)
	2	200	8.34(1.5)	0.25(.02)	0.27(.03)	1.05(.15)	0.96(.07)
	2	400	9.51(.78)	0.23(.006)	0.24(.01)	1.05(.08)	0.98(.03)
4	3	100	7.69(.80)	0.41(.03)	0.55(.09)	0.96(.21)	0.94(.05)
	3	200	8.76(.43)	0.35(.02)	0.38(.04)	1.01(.14)	0.97(.04)
	3	400	9.04(.20)	0.32(.007)	0.33(.01)	1.02(.09)	0.97(.03)
5	5	50	3.05(.74)	0.22(.02)	0.31(.08)	0.81(.23)	0.88(.08)
	5	100	3.87(.92)	0.20(.01)	0.23(.03)	0.89(.15)	0.84(.09)
	5	200	5.09(.95)	0.18(.01)	0.19(.01)	0.96(.10)	0.85(.10)
	10	50	2.54(.56)	0.25(.02)	0.42(.13)	0.80(.30)	0.92(.07)
	10	100	3.36(.88)	0.23(.02)	0.33(.07)	0.79(.20)	0.88(.08)
	10	200	4.15(.89)	0.20(.01)	0.24(.02)	0.88(.13)	0.83(.06)
6	5	50	4.58(.73)	0.080(.022)	0.19(.08)	0.87(.40)	0.97(.05)
	5	100	5.42(.64)	0.063(.004)	0.089(.025)	0.91(.22)	0.95(.04)
	5	200	6.10(.93)	0.056(.003)	0.063(.005)	0.95(.12)	0.89(.05)
	10	50	2.85(.72)	0.15(.06)	0.46(.14)	0.71(.29)	0.92(.09)
	10	100	4.93(.57)	0.077(.016)	0.17(.06)	0.86(.27)	0.98(.03)
	10	200	5.29(.50)	0.063(.003)	0.081(.010)	0.95(.15)	0.92(.03)
7	4	25	2.35(.50)	0.16(.02)	0.25(.10)	0.86(.43)	0.93(.09)
	4	50	2.84(.53)	0.14(.01)	0.16(.03)	0.91(.23)	0.92(.09)
	4	100	3.16(.44)	0.12(.007)	0.12(.01)	0.97(.16)	0.93(.06)
	4	200	3.33(.55)	0.11(.003)	0.11(.005)	0.99(.12)	0.91(.06)
8	4	25	2.38(.53)	0.17(.03)	0.52(.23)	0.47(.23)	0.91(.12)
	4	50	3.13(.73)	0.15(.02)	0.33(.12)	0.58(.21)	0.91(.11)
	4	100	4.13(.77)	0.14(.01)	0.20(.05)	0.77(.18)	0.93(.05)
	4	200	4.60(.68)	0.13(.009)	0.16(.02)	0.89(.14)	0.89(.06)
9	2	25	3.07(.26)	0.12(.02)	0.15(.05)	1.12(.47)	0.98(.04)
	2	50	3.16(.39)	0.11(.008)	0.12(.02)	1.01(.23)	0.97(.05)
	2	100	3.25(.56)	0.10(.004)	0.10(.007)	1.00(.17)	0.94(.06)
	10	50	2.99(.61)	0.19(.07)	0.50(.29)	0.80(.28)	0.91(.16)
	10	100	3.14(.38)	0.12(.01)	0.15(.05)	0.95(.21)	0.96(.03)
10	4	50	4.16(.47)	0.14(.01)	0.20(.05)	0.92(.32)	0.97(.04)
	4	100	4.59(.62)	0.13(.007)	0.14(.02)	0.95(.19)	0.95(.04)
	4	200	5.06(.45)	0.12(.004)	0.12(.006)	0.97(.10)	0.94(.04)
	4	400	5.06(.28)	0.11(.002)	0.11(.003)	1.00(.08)	0.94(.04)

The functions f in Examples 7 and 8 relate impedance and phase shift to four other variables in an alternating current series circuit (Friedman 1991):

$$\begin{aligned}\text{impedance} &= \left(Q^2+R^2\right)^{1/2} \quad \text{and} && (18)\\ \text{phase shift} &= \arctan(Q/R), && (19)\end{aligned}$$

where $R = 100x_1$ is the resistance, $\omega = 2\pi(20 + 260x_2)$ is the angular frequency, $L = x_3$ is the inductance, $C = 1 + 10x_4$ is the capacitance, and $Q = \omega L - 1/(\omega C)$. These target functions include interactions of all orders, although capacitance has only a slight effect over the specified domain. Friedman (1991) reported the following IPSE results (with $d = 4$ and $n = 100, 200, 400$): 0.35, 0.21, 0.16 for impedance and 0.32, 0.25, 0.21 for phase shift. O'Sullivan (1991) reported improvements over MARS in these examples using a smoothed version of CART. For $n = 100$, he obtained IPSE ≈ 0.22 for impedance and IPSE ≈ 0.25 for phase shift. ALB is substantially more accurate in these examples. Results with $n = 25, 50, 100, 200$ are reported in Table 3.

Example 9 was used by Gu, Bates, Chen & Wahba (1989) to illustrate interaction spline smoothing:

$$f(\mathbf{x}) = \frac{40h(\mathbf{x}, 0.5, 0.5)}{h(\mathbf{x}, 0.2, 0.7) + h(\mathbf{x}, 0.7, 0.2)}\,, \tag{20}$$

where

$$h(\mathbf{x}, a_1, a_2) = \exp\left[8\left\{(x_1-a_1)^2 + (x_2-a_2)^2\right\}\right].$$

This example was also used by Breiman (1991) and Friedman (1991) to illustrate the Π method and MARS, respectively. Plots of f and several estimates $\hat{f}$ can be found in these references. The target f is well approximated by f_K with $K = 3$, and ALB provides accurate estimates with small sample sizes. Other estimators are less efficient. Gu, Bates, Chen & Wahba (1989) and Friedman (1991) reported MPSE = 0.11 for $n = 300$ and $d = 2$. Breiman (1991) reported MPSE = 0.125 for $n = 100$ and $d = 2$.

TABLE 4: IPSE results for MARS in Example 10, based on 100 replicated samples of size n. The tuning constant "degree" is the maximum number of covariates permitted in MARS interaction terms.

	degree			
n	1	2	3	4
50	0.74(.10)	0.61(.20)	0.70(.34)	0.72(.38)
100	0.66(.05)	0.40(.46)	0.47(.48)	0.49(.48)
200	0.63(.03)	0.29(.09)	0.32(.09)	0.33(.33)
400	0.61(.02)	0.28(.09)	0.29(.07)	0.30(.06)

In Example 10, the target function f is an ALB regression function defined on a 3-dimensional projection of $\Re^4$, i.e., $f(\mathbf{x}) = f_K(\mathbf{z})$ where $K = 5$, $\mathbf{z} = (z_1, z_2, z_3)'$,

$$\begin{aligned} z_1 &= \sqrt{3}(x_1 + x_2 + x_3 + x_4 - 2),\\ z_2 &= \sqrt{3}(x_1 + x_2 - x_3 - x_4),\\ z_3 &= \sqrt{3}(x_1 - x_2 + x_3 - x_4).\end{aligned}$$

The z_j have mean 0 and standard deviation 1. The reference point parameterization is used to specify f_K: $\boldsymbol{\xi}_1 = (1,0,0)'$, $\boldsymbol{\xi}_2 = (-1,0,0)'$, $\boldsymbol{\xi}_3 = (0,1,0)'$, $\boldsymbol{\xi}_4 = (0,0,1)'$, $\boldsymbol{\xi}_5 = (0,0,0)'$,

$\gamma_1 = \cdots = \gamma_5 = 0$, $\delta_1 = \delta_2 = 1$, $\delta_3 = \delta_4 = -1$, $\delta_5 = 0$, and $\tau = 1$. The target f can be expressed directly as an ALB function of $\mathbf{x}$, with suitable parameters, and f has interactions of all orders among the four covariates. The interpretation of f is simpler in terms of $\mathbf{z}$. For fixed (z_2, z_3), f is a bowl-shaped function of z_1. For fixed (z_1, z_3), f is a decreasing sigmoidal function of z_2. For fixed (z_1, z_2), f is a decreasing sigmoidal function of z_3. Table 3 lists ALB results for n = 50, 100, 200, and 400. In previous examples, the average value of $\hat{K}$ increases with n. Here, the results suggest that $\hat{K}$ converges in probability to the true value $K = 5$. Whether this convergence holds as $n \to \infty$ is unknown.

In this example, the performance of ALB is substantially better than that of MARS (see Table 4). I carried out simulations using the version of MARS in the mda Package (Hastie & Tibshirani 2001, http://lib.stat.cmu.edu), varying the "degree" argument and setting all other arguments to their default values. The degree specifies the maximum number of covariates permitted in MARS interaction terms. Setting the degree to 1 yields an additive model. The best MARS results were obtained with the degree set to 2. It appears that MARS has difficulty modeling the higher-order interactions in f, even when n is large. In further simulations, the average IPSE values for MARS remain roughly constant as n increases from 400 to 4000.

7. DISCUSSION

I have attempted to show that ALB provides a useful addition to regression methodology. Its strengths, such as affine invariance, complement those of other flexible regression techniques. ALB appears well-suited for exploration of large multidimensional data sets where the target function f contains higher-order interactions. Some of ALB's limitations may be addressed by extending its methodology or by combining it with other techniques. The following are some ideas under investigation.

The simulation studies show that when nuisance variables are added to the predictors, ALB tends to compensate by reducing the number of basis functions. This results in smoother estimates $\hat{f}$ and reduced predictive performance. The behaviour is related to the effective number of parameters $1 + (K-1)(d+2)$ employed by ALB. In many applications, f is well-approximated by a function defined on a lower-dimensional projection, i.e., $f(\mathbf{x}) \approx f_0(\mathbf{Bx})$, where $\mathbf{B}$ is a $d_0 \times d$ matrix with $d_0 < d$. There is a substantial body of literature describing stable methods for dimension reduction, see, e.g., Li (1991, 1992), Cook (1998a,b) and Ferré (1998). Such methods can be used to estimate d_0 and the column space of $\mathbf{B}$, before applying ALB to the lower-dimensional predictor space. The affine invariance of ALB implies that the subspace basis chosen to define $\mathbf{B}$ will not affect the resulting estimator.

When a residual analysis indicates heteroscedasticity, one may wish to employ a weighted least squares estimator; i.e., minimize

$$\sum_{i=1}^{n} \left[\{y_i - \hat{f}(\mathbf{x}_i)\}/s(\mathbf{x}_i) \right]^2 ,$$

where $s(\mathbf{x})$ is an estimate of some measure of scale for the conditional distribution of y. The extension of ALB to weighted least squares is straightforward. A robust scale estimate can be obtained in two steps: first calculate the ALB L_1 regression estimate $\hat{f}$, then calculate $s(\mathbf{x})$ as the ALB L_1 regression of $|y_i - \hat{f}(\mathbf{x}_i)|$ against either $\hat{f}(\mathbf{x}_i)$ or $\mathbf{x}_i$. This conditional scale estimator is motivated by the MAD (median of absolute deviations from the median) estimator used in robust methods.

Regression quantiles can be estimated by modifying the ALB training risk. Set $0 < \alpha < 1$ and define the check function, $\rho_\alpha(z) = (1-\alpha)(-z)^+ + \alpha z^+$, where $z^+ = \max(z, 0)$. Minimizing $\sum \rho_\alpha\{y_i - \hat{f}(\mathbf{x}_i)\}$ yields an ALB α-quantile estimator suitable for multi-dimensional applications. He (1997) described methods for constructing several quantile curves that avoid crossing. This constraint can be implemented by using a common set of basis functions for all quantiles and order restrictions on the δ coefficients.

ALB methods can be applied in the context of generalized linear models. The main adjustment is to replace the training risk with an appropriate log likelihood. If the response variable is nonnegative (e.g., count data) then $f(\mathbf{x})$ can be approximated by $\exp\{\sum \delta_k \phi_k(\mathbf{x})\}$. Modifications to the updating functions (10) are straightforward. Techniques for an unordered polytomous response variable were developed in Hooper (2001).

ACKNOWLEDGEMENTS

Part of this work was carried out during a sabbatical year at the Department of Statistics, University of British Columbia. The author wishes to thank Nancy Heckman for helpful advice. Financial support from the Natural Sciences and Engineering Research Council of Canada is gratefully acknowledged.

REFERENCES

H. Akaike (1973). Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory* (B. N. Petrov & F. Cáski, eds.), Akademiai Kaidó, Budapest, pp. 267–281. Reprinted in *Breakthroughs in Statistics*, Volume I (S. Kotz & N. L. Johnson, eds.), Springer, New York, 1992, pp. 599–624.

D. M. Bates & D. G. Watts (1988). *Nonlinear Regression Analysis and its Applications*. Wiley, New York.

G. W. Bennett (1988). Determination of anaerobic threshold. *The Canadian Journal of Statistics*, 16, 307–316.

A. Benveniste, M. Métivier & P. Priouret (1990). *Adaptive Algorithms and Stochastic Approximations*. Springer-Verlag, New York.

L. Breiman (1991). The Π method for estimating multivariate functions from noisy data (with discussion). *Technometrics*, 33, 125–160.

L. Breiman & J. Friedman (1985). Estimating optimal transformations for multiple regression and correlation. *Journal of the American Statistical Association*, 80, 580–597.

S. Chatterjee, M. S. Handcock & J. S. Simonoff (1995). *A Casebook for a First Course in Statistics and Data Analysis*. Wiley, New York.

R. D. Cook (1998a). Principal Hessian directions revisited (with discussion). *Journal of the American Statistical Association*, 93, 84–100.

R. D. Cook (1998b). *Regression Graphics: Ideas for Studying Regressions through Graphics*. Wiley, New York.

D. R. Cox & E. J. Snell (1989). *Analysis of Binary Data*. Chapman & Hall, London.

P. Craven & G. Wahba (1979). Smoothing noisy data with spline functions. Estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerical Mathematics*, 31, 317–403.

E. S. Eppright, H. M. Fox, B. A. Fryer, G. H. Lamkin, V. M. Vivian & E. S. Fuller (1972). Nutrition of infants and preschool children in the north central region of the United States of America. *World Review of Nutrition and Dietetics*, 14, 269–332.

R. L. Eubank (1999). *Nonparametric Regression and Spline Smoothing*. Marcel Dekker, New York.

L. Ferré (1998). Determining the dimension in sliced inverse regression and related methods. *Journal of the American Statistical Association*, 93, 132–140.

B. D. Flury (1990). Principal points. *Biometrika*, 77, 33–41.

J. H. Friedman (1991). Multivariate adaptive regression splines (with discussion). *The Annals of Statistics*, 19, 1–141.

J. H. Friedman & W. Stuetzle (1981). Projection pursuit regression. *Journal of the American Statistical Association*, 76, 817–823.

A. R. Gallant (1987). *Nonlinear Statistical Models*. Wiley, New York.

C. Gu, D. M. Bates, Z. Chen, & G. Wahba (1989). The computation of GCV functions through householder tridiagonalization with application to the fitting of interaction spline models. *SIAM Journal of Matrix Analysis*, 10, 457–480.

D. Harrison & D. L. Rubinfeld (1978). Hedonic housing prices and the demand for clean air. *Journal of Environmental Economics and Management*, 5, 81–102.

J. A. Hartigan & M. A. Wong (1979). Algorithm AS136. A K-means clustering algorithm. *Applied Statististics*, 28, 100–108.

T. J. Hastie & R. J. Tibshirani (1990). *Generalized Additive Models*. Chapman & Hall, London.

X. He (1997). Quantile curves without crossing. *The American Statistician*, 51, 186–192.

P. M. Hooper (1999). Reference point logistic classification. *Journal of Classification*, 16, 91–116.

P. M. Hooper (2001). Reference point logistic regression and the identification of DNA functional sites. *Journal of Classification*, 18, 81–107.

P. M. Hooper, D. C. Mayes & N. N. Demianczuk (2001). A model for foetal growth and diagnosis of intrauterine growth restriction. *Statistics in Medicine*, to appear.

P. M. Hooper, H. Zhang & D. S. Wishart (2000). Prediction of genetic structure in genomic DNA, using reference point logistic regression models and sequence alignment. *Bioinformatics*, 16, 425–438.

K. Hornik, S. Stinchcombe & H. White (1989). Multilayer feed-forward networks are universal approximators. *Neural Networks*, 2, 359–366.

T. Kohonen (1995). *Self-Organizing Maps*. Springer, New York.

K. C. Li (1991). Sliced inverse regression for dimension reduction (with discussion). *Journal of the American Statistical Association*, 86, 316–342.

K. C. Li (1992). On principal Hessian directions for data visualization and dimension reduction: another application of Stein's Lemma. *Journal of the American Statistical Association*, 87, 1025–1039.

K. C. Li (1997). Nonlinear confounding in high-dimensional regression. *The Annals of Statistics*, 25, 577–612.

M. J. Lindstrom (1999). Penalized estimation of free-knot splines. *Journal of Computational and Graphical Statistics*, 8, 333–352.

J. MacQueen (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1, 281–297.

J. E. Moody & C. Darken (1989). Fast learning in networks of locally-tuned processing units. *Neural Computation*, 1, 281-294.

F. O'Sullivan (1991). Comment on 'Multivariate adaptive regression splines'. *The Annals of Statistics*, 19, 99–101.

B. D. Ripley (1996). *Pattern Recognition and Neural Networks*. Cambridge University Press, Cambridge.

H. Robbins & S. Monro (1951). A stochastic approximation method. *The Annals of Mathematical Statistics*, 22, 400–407.

R. D. Routledge (1991). Using time lags in estimating anaerobic threshold. *The Canadian Journal of Statistics*, 19, 233–236.

J. S. Simonoff (1996). *Smoothing Methods in Statistics*. Springer, New York.

G. Wahba (1990). *Spline Models for Observational Data*. SIAM, Philadelphia, PA.

L. Xu, A. Kryzak & A. Yuille (1994). On radial basis function nets and kernel regression: statistical consistency, convergence rates, and receptive field sizes. *Neural Networks*, 7, 609–628.

Received 22 July 2000
Accepted 12 January 2001

Peter M. HOOPER: hooper@stat.ualberta.ca
Dept. of Mathematical and Statistical Sciences
The University of Alberta, Edmonton, Alberta, Canada T6G 2G1

Discussion[1]

Comment 1: Mary J. LINDSTROM

The author is to be commended for a very thorough and insightful presentation of a promising new class of multidimensional nonparametric estimators. From this discussant's point of view, the most intriguing aspect of these estimators is that they are based on multivariate basis functions whose locations are estimated from the data.

1. PARAMETERIZATIONS

Our first challenge is to understand the structure of the ALB estimator. If we start with the reference point parameterization (equation 4) and (without loss of generality) we set τ equal to 1, the parameters which define each basis function are the K d-dimensional "centres" $\boldsymbol{\xi}_k$ and the K scalars γ_k. The form of the denominator, viz.

$$\sum_{m=1}^{K} \exp(\gamma_m) \exp(-||\mathbf{x} - \boldsymbol{\xi}_m||^2),$$

is what makes the ALB functions interesting (and complex). If the denominator did not involve $\mathbf{x}$, the exp(γ_k) term in the numerator and the entire denominator could be absorbed into δ_k (in equation 1) and we would have a set of very simple, radially-symmetric, Gaussian-density type basis functions, i.e.,

$$\phi_k'(x) = \exp(-||\mathbf{x} - \boldsymbol{\xi}_k||^2).$$

However, the denominator does involve all the distances $||\mathbf{x} - \boldsymbol{\xi}_m||$ with relative influence controlled by γ_m. Note that the relative influence of $||\mathbf{x} - \boldsymbol{\xi}_m||$ is the same for all basis functions (other than the mth) since γ_m does not vary by basis function. Thus the interpretation of $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_K)'$ is more complex than might be expected at first. One interpretation which seems helpful is that $\boldsymbol{\gamma}$ allows for estimation of the appropriate orientation of the basis functions, i.e., it allows for affine invariance. It would be interesting to compare the ALB estimator to a (nonaffine invariant) version with fixed $\boldsymbol{\gamma}$ and potentially more basis functions.

It is interesting to note that while fixing τ does not reduce the generality of $\phi_k(\mathbf{x})$, both $\boldsymbol{\gamma}$ and the "centres" $\boldsymbol{\xi}_k$ must be adjusted. That is, if we substitute $\tau^\star = c\tau$ for τ, then to obtain the same basis functions we must also substitute $\boldsymbol{\xi}_k^\star = c^2\boldsymbol{\xi}_k$ for $\boldsymbol{\xi}_k$ and $\gamma_k^\star = \gamma_k - (1-c^2)||\boldsymbol{\xi}_k||^2/\tau^2$ for γ_k. Note that nothing limits the new "centres" $\boldsymbol{\xi}_k^\star$ (or, for that matter, the original ones $\boldsymbol{\xi}_k$) to be in the span of the predictors. This underlines the interdependence of the parameters $\boldsymbol{\gamma}$, $\boldsymbol{\xi}_k$, and τ and the risks involved in interpreting them separately. The approach taken by the author seems very sensible. That is, reduce the over-parameterization of the basis by fixing τ to produce reasonably centered "centres." Unfortunately this must be done separately for each data set. Setting $\tau = 1$ is theoretically acceptable, but starting values for $\boldsymbol{\gamma}$ and $\boldsymbol{\xi}_k$ may be difficult to guess.

2. COMPARISON TO B-SPLINES

The author points out the similarities between one-dimensional ALB functions and the b-spline basis commonly used in fixed and free-knot regression splines. The term "free-knot" refers to location parameters (knots) which are estimated from the data. Like an ALB function, the value of a b-spline at a point x depends on the distance from x to the active "centres" (or knots) in the b-spline. In the cubic regression spline that is commonly used, there are five active knots for each b-spline. B-splines have truly local support in that if x is outside the range of the five active knots, then the b-spline value is zero. This is not a major advantage over ALB functions though

[1] These comments reflect the discussion of the paper which occurred after its public reading at the 29th Annual Meeting of the Statistical Society of Canada in Burnaby (British Columbia) on 14 June 2001.

since, depending on the parameter values, ALB functions can have effective (computationally) local support.

There are at least two major differences between ALB functions and b-splines. First, the knot locations completely determine the form of the b-splines. There are no parameters analogous to τ and γ modulating the shape. This makes them less flexible but also eliminates the problem of over-parameterization. Thus free-knot splines may be easier to estimate than ALB estimators. The second major difference is that replicate centres in a regression spline (two or more centres or knots with the same value) correspond to the loss of one or more derivatives at that location. As the author points out, this could be an advantage when modeling non-smooth functions but typically, it is more a nuisance when estimating knot positions. The enforced smoothness of the ALB estimators may make them easier to estimate than free-knot splines but only when using optimization methods which can handle over-parameterized models. These conflicting conclusions indicate that a detailed comparison would be worthwhile.

The author mentions the over-parameterization of the ALB functions but does not specifically discuss the problem of exchangeable parameters (which is shared by free-knot splines), i.e., exchanging the values of $\boldsymbol{\xi}_i$ and $\boldsymbol{\gamma}_i$ with $\boldsymbol{\xi}_j$ and $\boldsymbol{\gamma}_j$ for any i and j will not change the fitted values, but does change the parameter vector. Exchangeable parameters contribute to the numerous local optima which make least squares estimates of the knots in free-knot splines typically very difficult to find. Not only are there multiple global optima with relabeled parameters, but the exchangeability introduces extra local optima because of the symmetry induced in the objective function along lines and surfaces where two exchangeable parameters are equal (see Lindstrom 1999 for details). In one dimension, we can eliminate exchangeability by transforming the centres to a log-ratio parameterization which enforces ordering. Unfortunately, there is no obvious analogy in multiple dimensions due to the lack of a strict ordering.

Even when using the log-ratio parameterization, there are typically many local optima in a free-knot spline objective function which may or may not correspond to fits that are similar to the global optimum. In other words, the global optimum can be difficult to find and it may be important to find it. It seems that the estimation methods described in Section 4 would not, in general, identify this condition. Also, an objective surface with multiple optima corresponding to similar fits creates difficulties when estimating standard errors. The variability of the estimator may be much greater than can be inferred from the local characteristics of the objective function (a fifth caveat to the approximate standard errors).

B-splines do not generalize directly to multiple dimensions but there are many multivariate basis functions suggested in the literature. Typically, however, the kth basis function depends on only the kth centre through a term of the form $||\mathbf{x} - \boldsymbol{\xi}_k||$, i.e., the locations of the other centres do not influence the kth basis function at all. It is intriguing that the ALB functions generalize the dependence of b-splines on multiple centres to multiple dimensions.

I am not aware of any previous proposal for true estimation of the locations of multidimensional basis functions. As the author points out, there are many proposed stepwise deletion and insertion algorithms but these do not allow for the inclusion of the variability of the estimated centre locations in estimates of the variability of the fit. It may be that most researchers, perhaps given the known difficulties in estimating the knots in a unidimensional spline, felt that the computational difficulties of estimating multivariate locations would be overwhelming. The author is to be commended for finding a computational approach which is fast and seems to find useful solutions.

Mary J. LINDSTROM: lindstro@biostat.wisc.edu
Dept. of Biostatistics and Medical Informatics, University of Wisconsin
600 Highland Avenue, Room K6/446, Madison, WI 53792, USA

Comment 2: James O. RAMSAY

This paper showcases a wide spectrum of interesting ideas and issues in the smoothing of data and the estimation of response functions and surfaces. The adaptive logistic basis system seems well worth considering in some applications, and the optimization technology used to fit the data has some distinct merits. The five examples are interesting, and this article would make great a discussion piece for a graduate course in data analysis.

The implementation of localized basis functions in the 1960's, especially in the literature on spline functions, was an enormous leap forward with respect to either orthogonal polynomials or Fourier series as basis systems for basis function expansions. It meant that complex curve features could be accurately captured while still retaining the sparse coefficient matrix in the linear equation system defining least squares coefficient estimates. This made $O(n)$ calculations possible in practice, a crucial advantage when curve-fitting technology was transported to image analysis where the number n of data points could routinely be in the thousands. Moreover, a local basis system implies a local response to either changes in data or changes in coefficients, whereas in polynomial and Fourier series bases, a change in a single coefficient changes the fit everywhere, and often catastrophically at extreme sampling points.

Local basis systems led naturally to two strategies: use a lot of basis functions, and trim or down-weight those not needed in the fit; or keep the number K of basis functions small, and move them to where they were needed. Both strategies have advantages and disadvantages. Certainly one plus for the second adaptive approach is that conventional statistical theory can be appealed to in constructing confidence regions, since the number of parameters can be kept to a reasonably small fraction of the number of data values.

However, adaptive systems could also be unstable, and the author alludes to the problem of multicollinearity when two basis functions get too close together. The adaptive logistic basis for curve estimation was used by Bock & Thissen (1980) to model human growth data. They worked with three basis functions, implying nine parameters, but found that two parameters had to be collapsed to assure stable estimation. This parameter-collapse issue is also well known in the free-knot spline literature.

The real challenge is now in image analysis, whether over two or three spatial dimensions; we probably have more reliable curve-fitting technology than we need at this point. Both of the example response surfaces are rather benign in the sense of being fairly flat, with sampling points distributed over most of a rectangular region. I wonder what advantages the adaptive logistic fit would have relative to those from other approaches such as kernel or local polynomial smoothing or tensor product splines.

There is a third basis selection strategy in the image situation that seems promising, illustrated in Ramsay (2000). This is to position a great number of local basis functions exactly where they are needed, and then to control the smoothness through the use of a roughness penalty. The finite element method for solving partial differential equation systems can be adapted easily to the smoothing problem, and it also permits adaption to complex boundaries, both around the exterior of the data and also around "holes" in the interior of their distribution. T. Ramsay (2001) has taken the finite element approach substantially further.

I would like to focus some remarks on the optimization strategy, stochastic approximation, used in the paper. I was delighted to see this applied so successfully, and I am sure that we will see many more applications in the near future as we confront more and more data sets with n's of huge size.

Stochastic approximation may seem shocking at first sight, since it is the method that will not produce the same answer every time, and almost guarantees that the answer settled on is not as good as something out there. In this sense, it is in the same spirit as Markov chain Monte Carlo techniques. But in my opinion, statisticians have been too preoccupied by optimality, a view well argued by Tukey (1962). What matters is that we can find a good answer to a question, especially when the best answer is going to be indistinguishably better in a sense that really matters, such as predictive efficiency, risk, and other criteria. Those of us who work a lot with multicollinear predictors in regression settings are already used to seeing a dozen models with values of R^2 within 0.005 of the least squares estimate, even with K fixed.

Nevertheless, stochastic approximation could be slow, and when other methods are available that yield answers quickly enough to enable bootstrapping and other resampling approaches to interval estimation, they are likely to be preferred. In this regard, the first order differential equation corresponding to the stochastic gradient method used in the paper can be solved directly for the numbers of parameters involved in the illustration. The solution can be computed using existing numerical methods, such as those available in the base version of the Matlab system, for example. See Ramsay (1970) for a discussion of this approach.

James O. RAMSAY: ramsay@psych.mcgill.ca
Dept. of Psychology, 1205 avenue Docteur-Penfield
McGill University, Montréal (Québec), Canada H3A 2B1

Comment 3: Nancy E. HECKMAN

1. INTRODUCTORY COMMENTS

Parametric regression methods can be used to estimate an arbitrary smooth regression function provided one uses a flexible set of basis functions. Common bases include trigonometric functions yielding a Fourier series expansion and B-splines (see, e.g., Eubank 1988). The Fourier method does not accurately estimate functions that are relatively constant in some regions but rapidly changing in others. B-spline methods are able to adapt to this type of local variation provided one chooses the B-spline basis appropriately. Choosing a B-spline basis is equivalent to selecting a finite set of points, called knots, in the independent variable space. This is typically done by a somewhat cumbersome combination of forward selection and backward elimination. Extending these B-spline methods to high dimensions is straightforward in principle, but is computationally prohibitive due to the knot selection. See Stone, Hansen, Kooperberg & Truong (1997) and Zhou & Shen (2001).

The author's Adaptive Logistic Basis (ALB) regression method can be used to estimate functions that are relatively constant in some regions but rapidly changing in others. The method works well in high dimensions, and the speed is impressive (see the author's Table 1). The method is virtually automatic, following easily understandable criteria in a non-ad hoc manner. This is quite an accomplishment, and so the method shows great promise.

I will comment on some useful extensions of the methodology and also on one of the high-dimensional exploratory techniques introduced by the author.

2. ADDITIVE MODELS AND MODEL TESTING

The author notes, in Section 3 of his paper, that "the function f may exhibit simple structure related to the covariates... Methods that exploit this structure have an advantage." Additive models and semiparametric models have just such a simple structure.

With an additive model, one avoids some of the problems of high-dimensional regression. Computations are faster and one eliminates the "curse of dimensionality," the large mean squared errors inherent in high dimensional estimation. Moreover, additive models are often easily interpretable. As an example of an additive model, suppose the covariate vector x can be split into two components $x = (x_1, x_2)$. If there are no interactions between x_1 and x_2, we model the expected response additively as $f(x_1, x_2) = f_1(x_1) + f_2(x_2)$. In principle, the ALB method can be adapted to fit this additive model by modeling $f_j(x_j)$ as $\sum_1^{K_j} \delta_k^j \phi_k^j(x_j)(j = 1, 2)$, with the ϕ_k^j as in (2) of the paper. Furthermore, the ALB method might be used to construct a test to determine that there are indeed no interactions between x_1 and x_2. We would compare the fit of the additive model with the fit gotten from the full model, bootstrapping to calculate a p-value.

In a semiparametric model, the expected response is linear in some of the covariates and smooth in the other covariates. Semiparametric models are popular due to their easy interpretability and the parametric rate of convergence of the linear fit. The simplest example of a semiparametric model arises when y is the response to some treatment, one of the covariates δ say, is the treatment indicator, and the other covariates, x are nuisance parameters such as age and weight. We model the expected response as $\beta\delta + f(x)$, f smooth. Is this model reasonable? If so, what is β, the treatment effect? A modified ALB would be able to answer these questions.

3. TESTING FOR SHAPE OF REGRESSION FUNCTIONS

In Example 2.3, the main interest is the existence and location of an anaerobic threshold. The illustration in Figure 3(b) indicates that this threshold may appear at oxygen intake of approximately 3000 units, when the gradient function becomes convex, that is, when f''' becomes negative. However, the $\pm 2\mathrm{se}(\hat{g})$ confidence bands are pointwise and subject to the problems discussed in Section 5, so any inference drawn is suspect.

Several nonparametric tests have been proposed to test the null hypothesis that $f'(x) \leq 0$ for all x, against the alternative that f' is positive for some region of x values. See Bowman, Jones & Gijbels (1998), Gijbels, Jones, Hall & Koch (2000), and Hall & Heckman (2000). Harezlak & Heckman (2001) have extended the technique in Bowman, Jones & Gijbels (1998) to test $f^{(k)}(x) \leq 0$ for all x. Can ALB be used to test the null hypothesis that $f'''(x) \leq 0$ for all x? One way to do this would be to find an ALB fit of f restricted so that $\hat{f}''' \leq 0$, then compare this fit to the unrestricted ALB fit. Can ALB be modified for shape-restricted estimation?

4. EXPLORATORY TECHNIQUES

I was intrigued by the techniques used to study the regression fit for the Boston housing data. The techniques don't seem to be specific to the ALB method, and so might be generally useful in regression analysis of high dimensional data. Two ideas are proposed: (i) defining x-directions of high variability in $\hat{f}$ as those eigenvectors of $G \equiv \sum \hat{g}(x_i)\hat{g}(x_i)'$ corresponding to large eigenvalues, where $\hat{g} = \bigtriangledown \hat{f}$ and (ii) clustering of gradient vectors.

I will only comment on the first method. The merits of (i) are clear for $\hat{f}$ of the form $\hat{f} = b'x$, where any reasonable method should say that the direction of variability of $\hat{f}$ is b. Here $\hat{g}(x) = b$ and $G = nbb'$, which has one non-zero eigenvalue, corresponding to eigenvector b. Below I'll argue that, for more complicated functions, the eigenvectors of G depend on both the gradient of $\hat{f}$ and the distribution of the x_i vectors.

Note that the directions chosen by this method are unchanged by a rotation of the axes or by a shift in origin. If we work in the coordinate system $x^* = P'x + c$ with P orthonormal, and if λ is an eigenvalue of G with corresponding eigenvector v in the original coordinates, then λ is an eigenvalue of G^*, as defined for the rotated coordinates, with corresponding eigenvector $P'v+c$. Therefore, assume that the x_i have sample mean 0 and that the sample covariance matrix is diagonal.

Consider a simple example, $\hat{f}(x) = x'Dx/2$ with D diagonal. So $\hat{g}(x) = Dx$ and $G = D\sum x_i x_i' D' \equiv DXD$ with X diagonal with X_{jj}/n equal to the sample variance of the jth covariate. The largest eigenvalue of G is equal to the maximum of $D_{jj}^2 X_{jj}$, with corresponding eigenvector completely in the direction of the corresponding covariate. The variation in this direction stems from both the variation in the function and the variation in the covariate. This does in some sense meet the author's claim of identifying "the directions in the covariate space that best represent variation in $\hat{f}_K$." Is this interaction between the variation in the function and the variation the covariates harder to interpret for more complicated $\hat{f}$'s?

Nancy E. HECKMAN: nancy@stat.ubc.ca
Dept. of Statistics, The University of British Columbia
Vancouver, British Columbia, Canada V6T 1Z2

Comment 4: Hugh A. CHIPMAN and Hong GU

1. INTRODUCTION

We congratulate the author on an interesting, broad, and practical approach to flexible regression. The paper includes many features one would expect from a methodology that has been around much longer. The ability to apply the method to large datasets is appealing, the standard errors a useful addition, the ability to do quantile and/or robust regression quite convenient, and there are many extra options, such as the ability to reduce dimensionality of the predictor space. We were struck by how many avenues for further development were either already developed or suggested in the paper. The paper also raises many interesting questions and should provide fertile ground for further research.

In this discussion, we consider two modifications of the algorithm. In Section 2 we look at how to deal with local optima of the parameters, and in Section 3 we modify the ALB algorithm to fit radial basis functions. Section 4 concludes with an assortment of other comments.

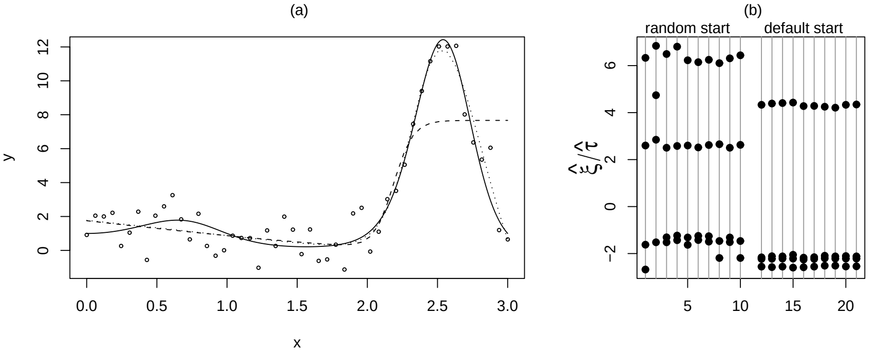

FIGURE C-4.1: A simulated example. In (a) is a simulated realization of 50 training cases, the true function $f(x)$ (—) and two local optima $\hat{f}(x)$ ($\cdots$, $---$) identified by different versions of ALB with $K = 4$. In (b) estimated reference points $\hat{\xi}_1, \ldots, \hat{\xi}_4$ (standardized by $\hat{\tau}$) from 20 runs of the algorithm are plotted, using either random starts or the default algorithm.

2. IMPROVING THE SEARCH

Local optima could be a problem for the stochastic approximation algorithm, especially if some optima fit poorly. The example in this section suggests that increased randomization of start points and stepwise deletion of bases could be useful in finding good local optima.

Figure C4-1(a) gives the example used to explore these strategies and shows two local optima of the ALB function with $K = 4$. We take $f(x) = \exp\{x \sin(\pi x)\} + \varepsilon$, with $\varepsilon \sim N(0, 1)$ and the training set having 50 equally spaced x values in the interval $(0, 3)$. The test set is the same 50 x values, with responses $y_i = f(x_i)$ instead of $f(x_i) + \varepsilon$.

In using the default parameters of the algorithm, $K = 5$ was usually chosen, which provided quite an accurate fit (R^2 for test set ≈ 0.98). Closer inspection revealed that for $K = 3, 4$, the estimated function fit poorly ($R^2 \leq 0.70$). With $K = 3$, it is possible to represent a single bump, such as the large one near $x = 2.5$. However, the algorithm tended to get stuck in poor local optima [$---$ in Figure C4-1(a)], perhaps due to minimal variation in the ten sets of starting values chosen by the vector quantization (VQ).

We considered random starting points for the parameters, with the hope that increasing variability would allow the algorithm to avoid poor local optima. We set $\gamma_m = 0$ and drew random pairs (x_i, y_i), $i = 1, \ldots, 4$ from the training set. We set $\delta_i = y_i$ and $\xi_i = 2x_i$. The x_i values were doubled because in the default runs of the algorithm, the ξ values were often outside the range of x. Using one simulated dataset, the optimization algorithm was run 100 times with different random number seeds and $K = 4$. All 100 $\hat{f}$ curves correctly identified the bump at $x = 2.5$ [$\cdots$ in Figure C4-1(a)]. Without randomization, the original algorithm missed the bump 96 times (— — —) and found it the other four ($\cdots$).

Although our implementation is primitive, an increase in the randomization of the initial parameter values seems to help the algorithm find better local optima. This better optimum is also found by about half the runs of the original algorithm if the initial step size in the vector quantization algorithm is doubled. A larger step size can be thought of as increasing the randomness of the algorithm, since the VQ algorithm samples the training cases one at a time in random order. Other randomization strategies, such as running VQ on small samples from the training data, might also prove successful. Simulated annealing might also be useful, although this would mean a substantial modification to the code.

In this example, we also found helpful stepwise deletion of basis functions. In Section 4.2, the author comments that because the parameters of the model must be simultaneously optimized, the stepwise addition or deletion of basis functions is not used. We think that deletion may be helpful in some cases, such as when several ξ are very close. Consider the ξ values in Figure C4-1(b), generated by 10 runs of the default algorithm and 10 runs with random starts. We standardize values by $\hat{\tau}$ which differs for each run. The model identified by the default algorithm fits poorly and has a group of three ξ around -2. In the reference point formulation (4) of $\phi_k(\mathbf{x})$, if two reference points are equal (say $\xi_1 = \xi_2$), then one basis function is redundant, since $\phi_2(\mathbf{x}) = \phi_1(\mathbf{x}) \exp(\gamma_2 - \gamma_1) = c\phi_1(\mathbf{x})$.

The near-duplication of reference points suggests a stepwise deletion strategy: if a model with reasonable fit has reference points that are quite close, delete one of the "near-duplicate" bases and use the remaining parameters as starting points for the algorithm. For the current example, one run of the default algorithm with $K = 5$ bases produced reference points $\xi = -1.183, -.473, -.472, 1.524, 2.646$. By deleting $\xi_2 = -.473$ and setting $K = 4$, the default algorithm identified a solution similar to the $\cdots$ curve in Figure C4-1(a). This solution offered comparable fit to the $K = 5$ case.

This illustration of two strategies for finding better optima should not be taken as an indication that the default algorithm fails — after all, $K = 5$ basis functions with good fit are identified. It does indicate, however, that the search for good parameter values could still be refined in some situations, perhaps leading to more parsimonious models.

3. RADIAL BASIS FUNCTIONS

The flexibility of the ALB family of models leads naturally to comparisons with other flexible models, such as radial basis functions or neural networks. In this section, we modify the stochastic approximation algorithm to estimate a radial basis function (RBF) model (Moody & Darken 1989). We consider the following parameterization of radial basis functions, as mentioned in the paper:

$$\phi_k(x) = \exp(-\tau_k^{-2}||x - \xi_k||^2) \Bigg/ \sum_{m=1}^{K} \exp(-\tau_m^{-2}||x - \xi_m||^2).$$

The parameter γ_k from ALB is dropped, and τ is allowed to vary across basis functions. As in ALB, a normalizing denominator is used. By allowing the radius τ_j of the jth basis to vary, the curvature of the function can be adjusted. Now we have

$$\frac{\partial \phi_m}{\partial \tau_k} = \begin{cases} 2\tau_k^{-3}||x - \xi_k||^2 \phi_k(1 - \phi_k) & \text{if } m = k; \\ -2\tau_k^{-3}||x - \xi_k||^2 \phi_k \phi_m & \text{if } m \neq k. \end{cases}$$

As with γ_k in ALB, increasing τ_k increases the influence of ϕ_k relative to other basis functions. Using the same gain a_m as defined by Hooper, the updating formulae in (10) become

$$\begin{aligned}
\delta_k &: \quad \delta_k + a_m^{\delta} h_k(x, y, \theta),\\
\tau_k &: \quad \tau_k + a_m^{\gamma} h_k(x, y, \theta)\delta_k - f_K(x)||x - \xi_k||^2(2\tau_k^{-3}),\\
\xi_k &: \quad \xi_k + a_m^{\xi} h_k(x, y, \theta)\delta_k - f_K(x)(x - \xi_k)(2\tau_k^{-2}).
\end{aligned}$$

Note that a_m^{γ} is used as the gain for τ. We used these updating formulae to modify the fortran code provided by the author to estimate radial basis functions.

As mentioned in Section 2, if $\xi_k = \xi_m$ for some $k \neq m$, one basis becomes redundant. This redundancy does not occur with radial basis functions, since if $\xi_k = \xi_m$ but $\tau_k \neq \tau_m$, a mixture of two bases with different radii results.

The accuracy of RBF and ALB was compared in simulation experiments for the following functions:

(i) $2\exp\{-(x_1^2 + x_2^2)/2\} + 3\exp\{-(x_1^2 + x_2^2)/5\}$;

(ii) $2\exp\{-(x_1^2 + x_2^2)/2\} + 3\exp[-\{(x_1 - 1)^2 + (x_2 - 1)^2\}/5]$;

(iii) ALB: $(2f_1 + 3f_2)/(f_1 + f_2)$,
where $f_1 = \exp\{1 - (x_1^2 + x_2^2)/4\}$ and $f_2 = \exp[2 - \{(x_1 - 1)^2 + (x_2 - 1)^2\}/4]$;

(iv) RBF: $(2f_1 + 3f_2)/(f_1 + f_2)$,
where $f_1 = \exp\{-(x_1^2 + x_2^2)\}$ and $f_2 = \exp[-\{(x_1 - 1)^2 + (x_2 - 1)^2\}/4]$;

(v)–(viii) Examples 2, 3, 6, 7 from the paper.

For each example, ten realizations of the dataset are simulated, and ALB and RBF models fit to each dataset. Table C4-1 gives average K and IPSE values, and also the results of paired t tests to compare IPSE values of the two models. A negative t statistic indicates that RBF has better accuracy (lower IPSE).

For Examples 1, 2 and 4, RBF significantly outperforms ALB, which one would expect when the true function is of the RBF form. For the ALB function in Example 3, ALB did slightly better than RBF. In Example 8, there is no significant difference. In other examples, ALB outperformed RBF. Does this mean ALB should be chosen over RBF? Not necessarily. In modifying the ALB algorithm to estimate a RBF model, we changed only the updating formulae and the basis functions. Other components of the ALB algorithm, which have been carefully optimized for the ALB function (e.g., the gains functions a_m), were left unchanged. The performance attained by RBF with a relatively straightforward modification of the algorithm is promising and indicates the effectiveness of the stochastic approximation algorithm.

4. OTHER COMMENTS

The ALB model is affine invariant, in the sense that if any affine transformation is applied to the predictors, there exists an ALB model using the transformed variables that provides exactly the same predictions as an ALB model using the original variables. This doesn't necessarily mean that the estimation algorithm can find this equivalent model, especially since there could be many local optima. A related issue is the fact that the algorithm is based on the reference point parameterization of the basis functions. By using Euclidean distance from reference points ξ_k in the covariate space, the accuracy of the fit is sensitive to multicollinear covariates. Under the affine transformation $z = B'x$, where B is invertible, the fits of the ALB regression based on the Euclidean distances in the x-space and in the z-space generally will not have the same accuracy. Elements of the algorithm, such as the update steps, may be affected, potentially yielding different models, even though the two forms of basis functions are one-to-one correspondent.

TABLE C4-1: Comparisons between RBF and ALB:
average and standard deviations from ten replicated samples of size n.

				ALB		RBF			
No.	d	n	$\sigma_f/\sigma_\varepsilon$	$\hat{K}$	IPSE	$\hat{K}$	IPSE	t	p-value
1	2	100	1	4.2	0.59 (.03)	2.1	0.54 (.03)	-4.08	0.003
	2	100	2	4.9	0.24 (.02)	2	0.21 (.008)	-4.16	0.002
	2	100	3	5	0.12 (.005)	2.2	0.11 (.003)	-10.15	0.000
2	2	100	1	3.7	0.59 (.02)	2	0.54 (.02)	-5.3	0.000
	2	100	2	4.9	0.24 (.01)	2.5	0.23 (.008)	-2.06	0.069
	2	100	3	5	0.12 (.006)	3.1	0.11 (0.006)	-3.01	0.015
3	2	100	1	2	0.52 (.02)	2.1	0.53 (.02)	1.84	0.098
	2	100	2	2	0.21 (.007)	2.2	0.21 (.009)	2.30	0.047
	2	100	3	2	0.10 (.003)	2.2	0.11(.005)	2.27	0.050
4	2	100	1	4.1	0.59 (.03)	2.2	0.53(.02)	-6.76	0.000
	2	100	2	5	0.24 (.009)	2.2	0.21(.008)	-9.01	0.000
	2	100	3	5.2	0.12 (.005)	2.1	0.11 (.005)	-8.00	0.000
5	2	100	0.9	4.6	0.68 (.02)	4.1	0.73 (.04)	3.44	0.007
	2	200	0.9	5.8	0.65 (.03)	4.5	0.69 (.04)	4.53	0.001
	2	400	0.9	6.5	0.60 (.02)	6.2	0.64 (.02)	6.04	0.000
6	2	100	1.9	5.5	0.33 (.03)	5.1	0.37 (.01)	3.38	0.008
	2	200	1.9	8	0.28 (.03)	7.6	0.32 (.02)	6.89	0.000
	2	400	1.9	9.2	0.24 (.008)	12.1	0.27 (.01)	5.73	0.000
7	5	50	4.9	4.7	0.16 (.04)	4.5	0.23 (.05)	3.03	0.014
	5	100	4.9	5.2	0.08 (.01)	5.7	0.10 (.015)	6.14	0.000
	5	200	4.9	5.9	0.06 (.005)	6	0.08 (.007)	6.68	0.000
	10	50	4.9	2.7	0.36 (.08)	3	0.38 (.09)	1.05	0.32
	10	100	4.9	5.1	0.18 (.07)	4.9	0.23 (.07)	2.56	0.03
	10	200	4.9	5.7	0.09 (.01)	6.3	0.11 (.03)	3.45	0.007
8	4	25	3	2.3	0.29 (.10)	2.5	0.31 (.14)	0.34	0.74
	4	50	3	2.9	0.17 (.04)	2.5	0.16 (.02)	-0.65	0.52
	4	100	3	3.3	0.12 (.009)	3	0.13 (.012)	1.17	0.27
	4	200	3	3.5	0.11 (.004)	3.8	0.12 (.007)	2.72	0.023

Any sensitivity that ALB has to affine transformations should be smaller than for methods that assume additivity, such as MARS.

The inclusion of standard errors in Section 5 is a nice addition to the paper, allowing inference about the shape of the surface. The standard errors are obtained conditional on the number of bases (K), when in fact K is estimated from the data. Accounting for uncertainty in K might be accomplished via the bootstrap or a more complex Bayesian approach (such as Smith & Kohn 1996 or Chipman, George & McCulloch 1998). Bayesian (e.g., Draper 1995) or bootstrap

(Breiman 1996) model averaging might also improve predictions by combining multiple models. It is difficult to say whether model averaging will offer much of a gain with this form of model. Improvements are usually largest for families of models that are sensitive to small changes in the data, such as trees.

The stochastic approximation algorithm has been constructed so that the number of steps of the algorithm does not depend on the sample size. With sample sizes of more than a few hundred thousand, many points will never be used. This has a similar flavour to training the model on a sample of the data, a common technique for large data sets.

In Section 2.5, the paper uses principal components of the gradient sum-of-products matrix $\mathbf{G}$, suggesting that if the first two eigenvalues are large, a two-dimensional plot will represent most of the variation in the response model. We wonder whether this strategy could be taken further, using the directions defined by the eigenvectors to reduce the dimensionality of the original problem, perhaps yielding better models. This might also be an effective means to accomplish variable selection, eliminating variables with loadings near zero in all large principal components.

Hugh A. CHIPMAN: hachipman@uwaterloo.ca
Department of Statistics and Actuarial Science
University of Waterloo, Waterloo, Ontario, Canada N2L 3G1

Hong GU: hgu@mathstat.dal.ca
Department of Mathematics and Statistics
Dalhousie University, Halifax, Nova Scotia, Canada B3H 3J5

Rejoinder

Peter M. HOOPER

I thank the discussants for their thoughtful comments. My response is organized under three topics. A FORTRAN implementation of ALB is available at ftp.stat.ualberta.ca/pub/research/hooper/

1. OPTIMIZATION

Lindstrom provides an interesting comparison of ALB with free-knot splines. She suggests that over-parameterization of the ALB model makes optimization more difficult. I am not sure that this is necessarily so. Redundancy may improve optimization by creating more pathways toward good local optima. In complex applications with many parameters, it may be unrealistic to hope that a global optimum will be attained. The randomness inherent in stochastic approximation could also have a beneficial effect, similar to simulated annealing, assisting escape from poor local optima. I agree with Lindstrom that multiple local optima imply greater variance in $\hat{f}$ than is indicated by the standard errors.

Ramsay notes that alternative non-stochastic optimization methods are available. Such methods should greatly increase computational speed in smaller problems, allowing the use of bootstrap standard errors. I would expect stochastic approximation to remain competitive in applications where n, d, or K is large.

Chipman and Gu suggest that optimization may be improved by increasing variability in the initial reference points $\boldsymbol{\xi}_k$. This modification can be implemented by reducing the number of iterations in the vector quantization algorithm described at Expression (10). In effect, Chipman and Gu replace the number $3000\sqrt{K}$ by 0. I suspect their modification may introduce too much variation in some higher-dimensional applications, but an intermediate reduction may improve performance in such cases. I have recently investigated an alternative vector quantization algorithm allowing the initial reference points to depend on the joint distribution of $\mathbf{x}$ and y. This alternative seems to improve optimization in some examples and is included as an option in the current implementation of ALB.

Chipman and Gu note that optimization can produce near-duplication of reference points. They suggest a deletion strategy to construct a more parsimonious model. I have also observed this phenomenon in several examples. There may be little advantage in deleting duplicates since these have essentially no effect on $\hat{f}$ and do not contribute to overfitting. Duplication may be viewed as fortuitous. We may miss the optimal $K = 4$ model, due to poor initial values, but then effectively obtain the optimal $K = 4$ model by taking $K = 5$.

2. ALB AND RBF AS LATENT VARIABLE MODELS

Chipman and Gu show how a radial basis function (RBF) model can be estimated by modifying the ALB updating formulae. Some care is needed in applying the new formulae to ensure that the τ_k are bounded above zero. The comparison of ALB with RBF illustrates how accuracy depends on the nature of the estimand f. In my comparisons of ALB with MARS and with Pi, the method requiring fewer parameters typically yields more accurate estimates. Chipman and Gu obtain a similar relationship in their Table C4-1. When comparing average $\hat{K}$ values for ALB and RBF, it should be noted that the RBF parameterization is not redundant. The effective number of RBF parameters is $K(d + 2)$, as compared with $1 + (K - 1)(d + 2)$ for ALB.

The two models differ with respect to affine invariance. The RBF model is invariant under location shifts and orthogonal transformations, but not under scale transformations. The choice of covariate scales can thus affect the number of basis functions required for adequate approximation of f. Orthogonal invariance of RBF implies that, like ALB, it is unable to capitalize on additive properties of f. Chipman and Gu correctly note that, while the ALB model is affine invariant, the ALB estimator is not. The choice of covariate scales affects the initial parameter values, and the initial values in turn can affect the outcome of the stochastic approximation. Potential problems are reduced by routinely scaling all covariates to have unit variance, but problems could still arise from dependencies among the covariates. The initial reference points are more spread out in directions of greater variation in the covariate space. This behaviour is advantageous when the gradient of f is small in directions with little variation in $\mathbf{x}$, but it is problematic when the gradient is large in these directions. Transformation of the covariates to principal components could sometimes help, but it could also aggravate problems associated with high dimensionality; see the discussion at Expression (14) in Hooper (1999). To some extent, ALB shares the advantages and disadvantages of principal components regression.

ALB and RBF models can be viewed as latent variable models. This perspective sheds light on the ALB parameterizations, supplementing comments by Lindstrom, and reveals a somewhat unusual property of the RBF model. In the following expressions, $p(\cdot)$ denotes various probability and density functions. Suppose k is a discrete random variable, distributed jointly with $(\mathbf{x}, y)$, and suppose the conditional mean of y given $(\mathbf{x}, k)$ depends only on k, i.e., $\mathrm{E}(y \mid \mathbf{x}, k) = \delta_k$. We then have $\mathrm{E}(y \mid \mathbf{x}) = \sum \delta_k p(k \mid \mathbf{x})$. The ALB and RBF models for the conditional mean adopt different parametric models for $p(k \mid \mathbf{x})$. Both models represent the regression relationship as a consequence of a latent discrete variable k, with (some degree of) conditional independence between y and $\mathbf{x}$ given k. The latent variable is an abstraction and would typically not represent a “real” or interpretable category. The models place no restrictions on the marginal distribution of $\mathbf{x}$. It is interesting, however, to consider implicit restrictions on $p(k)$ and $p(\mathbf{x} \mid k)$ arising from Bayes’s formula:

$$p(k \mid \mathbf{x}) \propto p(k)p(\mathbf{x} \mid k).$$

ALB selects a multinomial logistic model:

$$p(k \mid \mathbf{x}) \propto \exp(\alpha_k + \boldsymbol{\beta}_k' \mathbf{x}).$$

The implications of Bayes’s formula are well known, reflecting the relationship between linear and logistic discriminant analysis. The conditional density of $\mathbf{x}$ is in the exponential family:

$$p(\mathbf{x}\,|\,k) \propto \exp\{\boldsymbol{\beta}_k'\mathbf{x} + h(\mathbf{x})\}.$$

We may, for example, take $p(\mathbf{x}\,|\,k)$ to be the $N_d(\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$ density, so that $\boldsymbol{\beta}_k = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k$. The distribution $p(k)$ is unrestricted and is not related to $p(\mathbf{x}\,|\,k)$. The factor $\exp(\alpha_k)$ accounts for both $p(k)$ and the normalizing factor for $p(\mathbf{x}\,|\,k)$.

RBF sets

$$p(k\,|\,\mathbf{x}) \propto \exp\{-\tau_k^{-2}||\mathbf{x} - \xi_k||^2\}.$$

The conditional density has the form

$$p(\mathbf{x}\,|\,k) \propto \exp\{-\tau_k^{-2}||\mathbf{x} - \xi_k||^2 + h(x)\}.$$

The distribution $p(k)$ is then determined by the normalizing factor:

$$p(k) \propto \int \exp\{-\tau_k^{-2}||\mathbf{x} - \xi_k||^2 + h(\mathbf{x})\}d\mathbf{x}.$$

We may, for example, take $p(\mathbf{x}\,|\,k)$ to be the $N_d(\boldsymbol{\xi}_k, \sigma_k^2\mathbf{I}_d)$ density. We would then have $p(k) \propto \sigma_k^d$. In the context of discriminant analysis, it would be unusual to adopt a model where the prior probabilities $p(k)$ are determined by the densities $p(\mathbf{x}\,|\,k)$. The restriction on $p(k)$ can be avoided by setting

$$p(k\,|\,\mathbf{x}) \propto \exp\{\gamma_k - \tau_k^{-2}||\mathbf{x} - \boldsymbol{\xi}_k||^2\}.$$

This model extends both ALB and RBF, but is not affine invariant.

3. EXTENSIONS

ALB adjusts the potential complexity of the model by varying the number of basis functions. This strategy produces a flexible but relatively simple family of models. Heckman suggests extending ALB to semiparametric and additive models. I briefly investigated semiparametric models comprised of an ALB component and a linear combination of fixed basis functions. The attempt was only partially successful due to difficulties in extending the optimization technique. My approach was fairly simplistic, however, and further study is warranted. I have not explored additive models with several ALB components. Here the task of optimization appears to be more difficult and may require combining stochastic approximation with a backfitting algorithm. Model selection issues also arise. I suspect that the ALB methodology will be less successful in implementing additivity constraints than methods, such as MARS, that employ basis functions designed for this purpose.

Heckman asks whether ALB can be modified for shape-restricted estimation. This is an intriguing question. Let $d = 1$. By using Proposition 1 and the latent variable perspective described above, the derivatives of f_K can be expressed in terms of conditional covariances given x. First note that $f_K(x) = \mathrm{E}(\delta_k\,|\,x)$ and $\bar{\beta}(x) = \mathrm{E}(\beta_k\,|\,x)$. The first two derivatives of $\bar{\beta}$ are $\bar{\beta}^{(1)}(x) = \mathrm{Var}(\beta_k\,|\,x)$ and $\bar{\beta}^{(2)}(x) = \mathrm{E}[\{\beta_k - \bar{\beta}(x)\}^3\,|\,x]$. The first three derivatives of f_K are

$$\begin{aligned}
f_K^{(1)}(x) &= \mathrm{Cov}(\delta_k, \beta_k\,|\,x),\\
f_K^{(2)}(x) &= \mathrm{Cov}[\delta_k, \{\beta_k - \bar{\beta}(x)\}^2\,|\,x],\\
f_K^{(3)}(x) &= \mathrm{Cov}[\delta_k, \{\beta_k - \bar{\beta}(x)\}^3\,|\,x] - 3\mathrm{Cov}(\delta_k, \beta_k\,|\,x)\mathrm{Var}(\beta_k\,|\,x).
\end{aligned}$$

With these, it should be possible to estimate f_K subject to constraints on its derivatives.

Heckman gives an interesting analysis of the global visualization technique based on gradient principal components. As she observes, the directions chosen depend on variation in both the function and the covariates. I think this interaction can be interpreted as a consequence of the gradient depending on the scale of the response and the covariates, so the idea in Heckman's example applies in general. Chipman and Gu ask whether gradient principal components can be used to improve the estimator by applying ALB to a smaller set of linear combinations. I have made some progress in this regard, and an optional dimension reduction strategy is included in the ALB implementation. It may also be useful to automate methods for identification and deletion of nuisance variables.

Ramsay asks what advantages ALB has relative to other methods in the context of the Viking formation and Boston housing examples. In the former example, affine invariant methods seem more suitable than tensor-product splines. Also, robust methods are desirable given the long-tailed distribution of $y - f(\mathbf{x})$. Regarding robustness, I neglected to mention work by Forsythe (1972) on L_q estimators in linear regression. He suggested the choice of $q = 1.5$ as a good compromise, with efficiency near that of the L_2 estimator when the errors are Gaussian, and with substantially higher relative efficiency when the errors are heavily contaminated by outliers. In the Boston housing example, kernel methods may perform poorly due to the high dimensionality. ALB describes relationships between housing price and the 13 covariates in a parsimonious two-dimensional model.

ACKNOWLEDGEMENTS

I would like to express my thanks to the past Editor of the Journal, Christian Genest, for arranging this forum for the presentation and discussion of this work at the 29th Annual Meeting of the Statistical Society of Canada in Burnaby (British Columbia) on 14 June 2001.

Peter M. HOOPER: hooper@stat.ualberta.ca
Department of Mathematical and Statistical Sciences
The University of Alberta, Edmonton, Alberta, Canada T6G 2G1

COMBINED REFERENCES IN THE DISCUSSION AND REJOINDER

R. D. Bock & D. Thissen (1990). Statistical problems of fitting individual growth curves. In *Human Physical Growth and Maturation: Methodologies and Factors* (F. E. Johnston, A. F. Roche & C. Susanne, eds.), Plenum, pp. 265–290.

A. W. Bowman, M. C. Jones & I. Gijbels (1998). Testing monotonicity of regression. *Journal of Computational and Graphical Statistics*, 7, 489–500.

L. Breiman (1996). Bagging predictors. *Machine Learning*, 26, 123–140.

H. A. Chipman, E. I. George & R. E. McCulloch (1998). Bayesian CART model search (with discussion). *Journal of the American Statistical Association*, 93, 935–960.

D. Draper (1995). Assessment and propagation of model uncertainty. *Journal of the Royal Statistical Society Series B*, 57, 45–97.

R. Eubank (1988). *Spline Smoothing and Nonparametric Regressions*. Dekker, New York.

A. B. Forsythe (1972). Robust estimation of straight line regression coefficients by minimizing pth power deviations. *Technometrics*, 14, 159–166.

I. Gijbels, P. Hall, M. C. Jones & I. Koch (2000). Tests for monotonicity of a regression mean with guaranteed level. *Biometrika*, 87, 663–673.

P. Hall & N. E. Heckman (2000). Testing for monotonicity of a regression mean by calibrating for linear functions. *The Annals of Statistics*, 28, 20–39.

J. Harezlak & N. E. Heckman (2001). $\mathcal{C}$ri$\mathcal{SP}$ — a tool for bump hunting. *Journal of Computational and Graphical Statistics*, to appear.

P. M. Hooper (1999). Reference point logistic classification. *Journal of Classification*, 16, 91–116.

M. J. Lindstrom (1999). Penalized estimation of free-knot splines. *Journal of Computational and Graphical Statistics*, 8, 333–352.

J. E. Moody & C. Darken (1989). Fast learning in networks of locally-tuned processing units. *Neural Computation*, 1, 281-294.

J. O. Ramsay (1970). A family of gradient methods for optimization. *The Computer Journal*, 13, 413–417.

J. O. Ramsay (2000). Differential equation models for statistical functions. *The Canadian Journal of Statistics*, 29, 225–240.

T. O. Ramsay (2001). Spline smoothing over difficult regions. Unpublished manuscript.

M. Smith & R. Kohn (1996). Nonparametric regression using Bayesian variable selection. *Journal of Econometrics*, 75, 317–344.

C. J. Stone, M. Hansen, C. J. Kooperberg & Y. K. Truong (1997). Polynomial splines and their tensor products in extended linear modeling. *The Annals of Statistics*, 25, 1371–1470.

J. W. Tukey (1962). The future of data analysis. *The Annals of Mathematical Statistics*, 33, 1–66.

S. Zhou & X. Shen (2001), Spatially adaptive regression splines and accurate knot selection schemes. *Journal of the American Statistical Association*, 96, 247–259.