Additive Splines in Statistics

Charles J. Stone and Cha-Yong Koo,
University of California, Berkley

Reprinted from the 1985 Statistical Computing Section,
Proceedings of the American Statistical Association. Original pagination is p. 45–48.

Abstract

Additive models for regression functions and logistic regression functions are considered in which the component functions are fitted by cubic splines constrained to be linear in the tails. Rules and strategies for knot placement are discussed, and two illustrative applications of the resulting methodology are presented.
Key words: Regression; Logistic regression; Additivity; Spline.

1 Additive Spline Models

Let Y be a response variable whose distribution depends on the values of input variables $x_1, \ldots, x_J$ and let $\eta = f(x_1, \ldots, x_J)$ denote a parameter defined in terms of this distribution. The function f is referred to as the *response function*. If η is the mean of Y, f is the usual regression function. Suppose instead that Y takes on only the values 0 and 1 and that η is the logit of the probability that $Y = 1$, where $\text{logit}(\pi) = \log(\pi/(1-\pi))$. Then f is the logistic regression function.

Additive models for a response function have recently been considered by Hastie and Tibshirani (1986) and Stone (1985, 1986). In such models

$$f(x_1, \ldots, x_J) = f_0 + \sum_1^J f_j(x_j),$$

where f_0 is a constant and $f_1, \ldots, f_J$ are called *component functions*.

A natural approach in fitting an additive model based on data is to use a parametric form for the component functions. Obvious candidates are:

linear:	$f_j(x_j) = \beta_j x_j$
power:	$f_j(x_j) = \beta_j (x_j - \alpha_j)^{\gamma_j}$
log:	$f_j(x_j) = \beta_j \log(x_j - \alpha_j)$
quadratic:	$f_j(x_j) = \beta_j x_j + \gamma_j x_j^2$
cubic:	$f_j(x_j) = \beta_j x_j + \gamma_j x_j^2 + \delta_j x_j^3$

The commonly employed linear form is inflexible since it has only one unknown parameter (the constant term being absorbed into f_0). The power and log forms are somewhat more flexible but hard to fit, since they are nonlinear in the unknown parameters. Quadratic, cubic and higher degree polynomials are linear in the unknown parameters and increasingly flexible. But they are typically too flexible in the tails in relation to the amount of noisy data available there, especially when used for extrapolation.

Splines are an attractive alternative to polynomials. A spline of degree q is a piecewise qth degree polynomial, perhaps subject to some smoothness constraints at the knots (boundaries between consecutive pieces). Commonly employed are piecewise constants ($q = 0$), linear splines ($q = 1$), quadratic splines ($q = 2$) and, especially, cubic splines ($q = 3$).

Twice continuously differentiable cubic splines are particularly attractive since modest discontinuities in the third derivative cannot be detected visually. Consider a finite sequence ξ of strictly increasing knot locations $\xi_1, \ldots, \xi_N$; and let $\mathcal{S}_\xi$ denote the collection of twice continuously differentiable functions on $\mathbb{R}$ that reduce to cubic polynomials on each of the intervals $(-\infty, \xi_1], [\xi_1, \xi_2], \ldots, [\xi_{N-1}, \xi_N], [\xi_N, \infty)$. Then $\mathcal{S}_\xi$ is a $(4+N)$-dimensional vector space. This collection of cubic splines is better suited for curve fitting than the $(4+N)$-dimensional space of polynomials of degree $3+N$. But the functions in $\mathcal{S}_\xi$ are still too flexible in the tails. Let $\mathcal{S}_{L\xi}$ denote the N-dimensional space consisting of those functions in $\mathcal{S}_\xi$ whose restrictions to $(-\infty, \xi_1]$ and to $[\xi_N, \infty)$ are each linear. These functions satisfy the "natural" boundary condition that their second derivative vanish at ξ_1 and ξ_N. (Fuller, 1969, proposed such linear restrictions in the context of extrapolating a time series trend.) Let ${}_\xi k, 1 \le k \le N$, denote a basis of $\mathcal{S}_{L\xi}$ chosen so that $A_{\xi N} = 1$. Such a basis can be constructed starting from either a truncated power basis or a B-spline basis of $\mathcal{S}_\xi$ (see de Boor, 1978).

Suppose now that for each $j, 1 \le j \le J$, we are given a sequence of N_j knots. Write the corresponding restricted cubic spline space and its basis as $\mathcal{S}_{Lj}$ and $A_{jk}, 1 \le k \le N$, respectively; and set $K_j = N_j - 1$. Let $\mathcal{A}$ denote the collection of all additive functions a of the form

$$a(x_1, \ldots, x_j) = a_0 + \sum_1^J a_j(x_j),$$

where a_0 is a constant and $a_j \in \mathcal{S}_{Lj}$ for $1 \le j \le J$. Then $\mathcal{A}$ is a vector space of dimension $1 + \sum_1^J K_j$ and every $a \in \mathcal{A}$ can be written in the form

$$a(x_1, \ldots, x_j) = \theta_0 + \sum_1^J \sum_1^{K_j} \theta_{jk} A_{jk}(x_j).$$

Although the shape of the function $\sum_1^{K_j} \theta_{jk} A_{jk}(\cdot)$ is meaningful, its individual values are not identifiable. To remedy this defect, let F_j be a probability distribution on $\mathbb{R}$ and set $\bar{A}_{jk} = \int A_{jk}(\cdot)\, dF_j$. Then the function $a_j(\cdot)$, defined by

$$a_j(x_j) = \sum_1^{K_j} \theta_{jk}(A_{jk}(x_j) - \bar{A}_{jk})$$

is such that $\int a_j(\cdot)\, dF_j = 0$. Thus a positive value of $a_j(x_j)$ is above "average" and a negative value is below average. To assess the statistical significance of any such departure from zero, however, a rough standard error (SE) formula is required.

A training sample $(x_{i1}, \ldots, x_{iJ}, y_i), 1 \le i \le n$, is necessary in order to obtain estimates $\hat{\theta}_{jk}$,

$$\hat{a}_j(x_j) = \sum_1^{K_j} \hat{\theta}_{jk}(A_{jk}(x_j) - \bar{A}_{jk}),$$

and

$$\hat{a}(x_1, \ldots, x_J) = \hat{\theta}_0 + \sum_1^J \hat{a}_j(x_j).$$

Suppose that the $\hat{\theta}_{j_k}$s have approximately a multivariate normal distribution with covariance matrix $\Gamma_j = (\Gamma_{jkl})$. Then the positive square root of

$$\sum_1^{K_j} \sum_1^{K_j} \Gamma_{jkl}(A_{jk}(x_j) - \bar{A}_{jk})(A_{jl}(x_j) - \bar{A}_{jl})$$

yields a rough SE formula for $\hat{a}_j(x_j)$. From now on, let F_j be the empirical distribution of the values $x_{1j}, \ldots, x_{nj}$ of the jth input variable in the training sample.

In order to implement these techniques, rules or strategies must be used to select the knot locations corresponding to each input variable. They may be automatic or subjective and may involve F_j; but if the SE formulas are roughly to be valid, they should not significantly involve values of the response variable in the training sample. When F_j is regular and the size of the training sample is at least in the hundreds and at least twenty times the number of input variable, we tentatively recommend putting knots at the 5th smallest and 5th largest values (counting multiplicities) of the jth input variable in the training sample and putting three additional knots in between so that the resulting five knots are equally spaced. There will then be four degrees of freedom for each input variable, one degree of freedom for the constant term, and $4J + 1$ degrees of freedom in total. Five parameters, including the constant term, should be enough to model the overall shape of the regular (smooth and either monotonic or unimodal) component functions that are likely to arise in practice. Putting in more than three additional knots allows for increased flexibility, especially better ability to determine the fine details of the component function. But it also increases the variance of the estimator, and the number of knots should not be increased to the point where the SEs become unacceptably large. (A similar point is made on page 132 of Backus and Gilbert, 1970, in the context of determining the interior structure of the earth.)

When F_j is highly irregular it may be best to place knots corresponding to the jth input variable subjectively, in an iterative manner, by examining histogram or quantile plots of F_j and SE plots of the estimated component function. It seems natural first to choose the knot locations based on a separate analysis for each input variable and then to modify the choices, if necessary, to take into account stochastic dependencies among the input variables. There are some other advantages in choosing knot locations subjectively. In particular, substantive knowledge and foresight can sometimes be used; in this regard, Poirier (1973) emphasized situations in which the knots correspond to known points of structural change. Also, one can use subjective judgements to make reasonable bias–variance tradeoffs.

If values of the response variable for the cases in the training sample are used in knot placement, the SE formulas are no longer to be trusted. Smith and Klein (1982) and Smith (1982) described and successfully applied attractive automatic knot placement procedures based on stepwise regression starting with a reasonably large initial pool of equally spaced knot locations. For a review of these and other proposals for automatic knot placement, see Eubank (1984). Alternatively, one could choose the knots subjectively by looking at plots of the estimated component functions and at the corresponding residual plots. This may lead to more knots in one part of the region than another if interesting

things seem to be happening on a smaller scale in the former region. If there appears to be a sharp change of slope at a point, one could model this by a triple knot at that point. Inessential knots (perhaps revealed by a modest sized t-statistic) could be moved or removed.

Subjective, interactive determination of knot locations allows for the inclusion with hindsight of knowledge of the substantive field from which the data arose. Suppose, for example, that examination of the plot of the estimates of a component function and the corresponding residual plot suggests the presence of an additional peak at a certain location, which was not picked up because of lack of flexibility of the corresponding cubic spline in a neighborhood of the peak. If substantive knowledge suggests that such a peak is plausible and worth detecting, a knot could be moved to the center of the conjectured peak or an additional knot could be placed there. Alternatively, substantive knowledge might suggest that such a peak is spurious or irrelevant and that it is proper to leave it smoothed out. Even when it does appear worthwhile to assess such a peak, examination of the new estimate of the component function along with the corresponding SE plot, residual plot, and t-statistic for the knot may suggest that there is no such peak or there is not enough data to determine it and hence that the effort to mirror the peak in the estimate should be abandoned. (This is also in line with the viewpoint of Backus and Gilbert referred to above.) Cleveland and McGill (1984) emphasized that "*Making a graph of a set of data is an iterative procedure.*" Suppose now that η is the mean of Y, so that f is the regression function. Then the estimates $\hat{\theta}_{j_k}$ can be obtained by the method of least squares. Under the assumption that the variance σ^2 of Y is constant, the covariance matrix Γ_j is the appropriate $K_j \times K_j$ submatrix of $s^2(X^TX)^{-1}$, where X is the $n \times (1 + \sum_1^J K_j)$ design matrix and s^2 is the usual unbiased estimate of σ^2.

In order to illustrate these techniques, we apply them to the data on atmospheric ozone concentration treated previously by Breiman and Friedman (1985) and Hastie and Tibshirani (1986). The data consist of 330 observations on 10 variables. The response variable is

UPO3: Upland Ozone Concentration (ppm).

Since the data is observational, the input variables are referred to as predictors. Following Hastie and Tibshirani, we use five of the predictors, the others being judged to be statistically insignificant. These five are

SBTP: Sandburg Air Force Base Temperature (°C)

IBHT: Inversion Base Height (feet)

DGPG: Daggot Pressure Gradient (mmHg)

VSTY: Visibility (miles)

DOYR: Day of Year

Figure 1 shows the graphs of the estimated component functions which are smoother than, but otherwise very similar to those of Hastie and Tibshirani, and the corresponding standardized residual plots.

2 Additive Logistic Regression

Suppose instead that Y takes on only the values 0 and 1 and that η is the logit of the probability that $Y = 1$. The f is the logistic regression function. The estimates $\hat{\theta}_{j_k}$ can now be obtained by the maximum likelihood method, and Γ_j can be obtained from the inverse information matrix.

A study conducted between 1958 and 1970 at the University of Chicago's Billings Hospital concerned the survival of patients who had undergone surgery for breast cancer (Haberman, 1976; Hastie and Tibshirani, 1986). There were 306 observations on four variables:

$$y = \begin{cases} 1 & \text{patient survived 5 or more years} \\ 0 & \text{otherwise} \end{cases}$$

$x_1 =$ age of patient at time of operation

$x_2 =$ year of operation (minus 1900

$x_3 =$ number of positive auxilliary nodes detected in patient

Figure 2 shows the empirical frequency functions, estimated component functions (again similar to those of Hastie and Tibshirani), and appropriate standardized residuals corresponding to each of the three predictors.

3 Concluding Remarks

Additive splines are a promising tool in confirmatory and exploratory statistics, especially in a workstation environment with fast computation, interactive graphics and hardcopy plotting capability. But further research and convenient software are needed.

4 References

Backus, G. and Gilbert, F. (1970), "Uniqueness in the inversion of inaccurate gross earth data," *Philosophical Transactions of the Royal Society of London, Series A*, 266, 123–192.

Baker, R.J. and Nelder, J.A. (1978), *The GLIM system*, Release 3, *Generalized Linear Interactive Modelling*, Numerical Analysis Group, Oxford.

de Boor, C. (1978), *A Practical Guide to Splines*, Springer-Verlag, New York.

Breiman, L. and Friedman, J.H. (1985), "Estimating optimal transformations for multiple regression and

correlation," *Journal of the American Statistical Association*, 82, 580–619.

Cleveland, W.S. and McGill, R. (1984), "The many faces of a scatterplot," *Journal of the American Statistical Association*, 79, 807–822.

Eubank, R. L. (1984), "Approximate regression models and splines," *Communications in Statistics. Theory and Methods*, 13, 433–484.

Fuller, W.A. (1969), "Grafted polynomials as approximating functions," *Australian Journal of Agricultural Economics*, 13, 35–46.

Haberman, S.J. (1976), "Generalized residuals for log-linear models," *Proceedings of the 9th International Biometrics Conference*, Boston, 104–122.

Hastie, T.J. and Tibshirani, R.J. (1986), "Generalized additive models," *Statistical Science*, 1, 297–310.

Poirier, D.J. (1973), "Piecewise regression using cubic splines," *Journal of the American Statistical Association*, 68, 515–524.

Smith, P.L. (1982), "Curve fitting and modelling with splines using statistical variable selection techniques," Technical Report, Old Dominion University.

Smith, P.L. and Klein, V. (1982), "The selection of knots in polynomial splines using stepwise regression," Technical Report.

Stone, C.J. (1985), "Additive regression and other nonparametric models," *Annals of Statistics*, 13, 689–705.

Stone, C.J. (1986), "The dimensionality reduction principle for generalized additive models," *Annals of Statistics*, 14, 590–606.

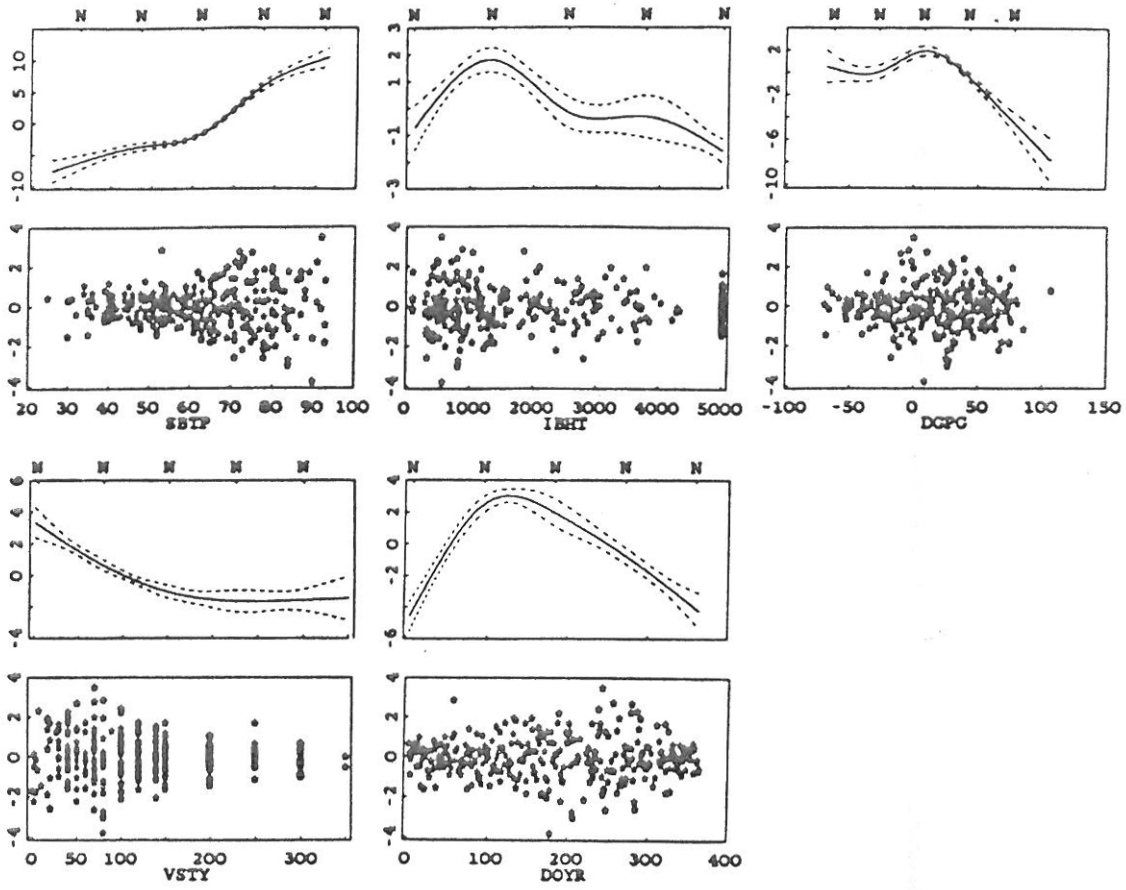

Figure 1: Additive regression. The response variable is UP03. Associated with each predictor is a pair of plots. The top plot shows the graphs of $\hat{a}_j(x_j)$ and $\hat{a}_j(x_j) \pm \mathrm{SE}(\hat{a}_j(x_j))$ and the corresponding knot locations, chosen in accordance with the recommendations in Section 1. The bottom plot shows the standardized residuals $r_i/s_i, 1 \leq i \leq n$, where r_i is the residual $Y_i - \hat{a}(x_{i1}, \ldots, x_{iJ})$ and s_i is the standard error of r_i.

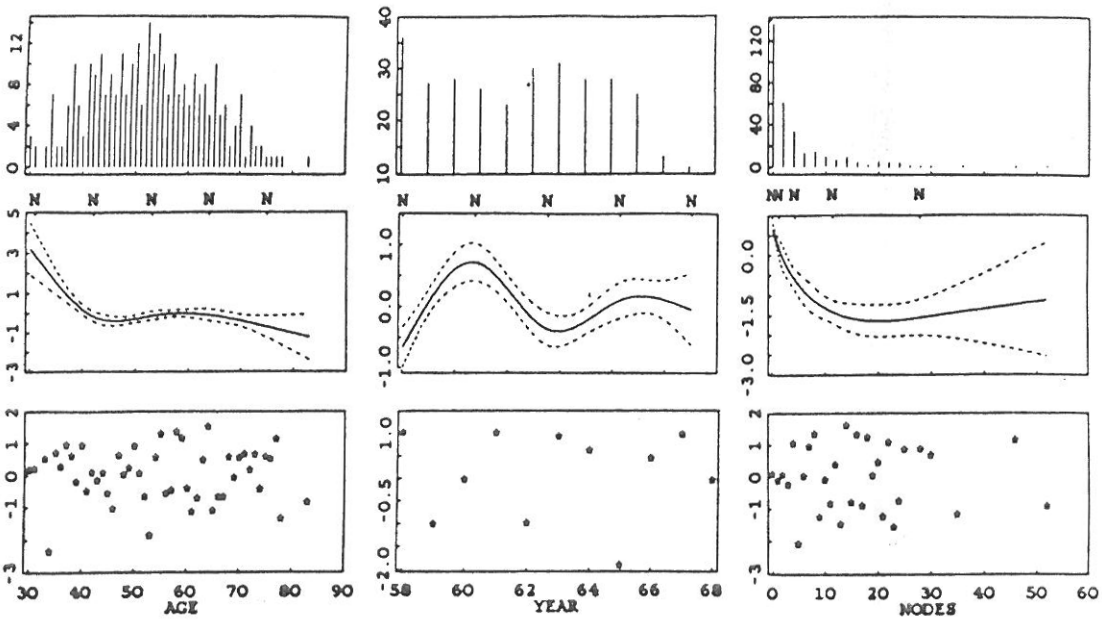

Figure 2: Additive logistic regression. Associated with each predictor is a triple of plots. The top plot shows the empirical frequency function of the corresponding predictor. The middle plot shows the graphs of $\hat{a}_j(x_j)$ and $\hat{a}_j(x_j) \pm \mathrm{SE}(\hat{a}_j(x_j))$ and the corresponding knot locations. (The empirical distribution of the third predictor is highly skewed. The three middle knots ξ_2, ξ_3, ξ_4 corresponding to this predictor were chosen so that $\log(1+\xi_l), 1 < l < 5$, are equally spaced.) The bottom plot shows the values, as a function of t, of the residual sum $\sum[(Y_i - \hat{a}(x_{i1}, \ldots, x_{iJ}); x_{ij} = t]$ divided by the standard error of this sum.