Statistical Science
1988, Vol. 3, No. 4, 425-461

Monotone Regression Splines in Action

J. O. RAMSAY

Abstract. Piecewise polynomials or splines extend the advantages of polynomials to include greater flexibility, local effects of parameter changes and the possibility of imposing useful constraints on estimated functions. Among these constraints is monotonicity, which can be an important property in many curve estimation problems. This paper shows the virtues of monotone splines through a number of statistical applications, including response variable transformation in nonlinear regression, transformation of variables in multiple regression, principal components and canonical correlation, and the use of monotone splines to model a dose-response function and to perform item analysis. Computational and inferential issues are discussed and illustrated.

Key words and phrases: Regression splines, monotone transformation, nonlinear regression, semiparametric principal components analysis, Eckart-Young approximation, dose-response, item regression function.

1. INTRODUCTION

The problem of estimating some transformation of a variable $x \rightarrow f(x)$ is so basic to statistics that it often goes unrecognized. For example, were linear regression discussed as a data transformation problem more frequently, there might be a stronger desire to move beyond this most elementary of transformations. The motives for nonlinear transformations include:

Estimation: An interest in the function f itself (as in nonlinear regression).

Reformatting: Transforming the data so that some prespecified model or relationship fits better (as in transformation to normality).

Redefining: Moving the data from one domain such as the entire real line to another such as [0, 1] (as in dose-response function estimation or bioassay).

Approximation: Replacing a complex and/or noisy relationship by something simple but reasonable (as in interpolation and smoothing).

Whatever the motive, the statistician needs a flexible set of functions appropriate to the job. Spline functions are fast replacing polynomials in this role, and this article tries to show why. By illustrating splines in action in a wide variety of settings, it aims to convince the reader that they work remarkably well.

Monotonicity can be essential in a data transformation. Often the inverse transformation is required in order to describe effects in terms of the original variables, and in some instances theoretical considerations require this feature. Attempting to impose monotonicity on polynomials quickly becomes unpleasant. A spline function, on the other hand, is easily constrained to be everywhere nondecreasing (or nonincreasing) while still retaining its essential flexibility. This is achieved by imposing nonnegativity on the parameters which define it. Because these simple linear inequality constraints can always be relaxed where monotonicity is inessential, the models used in this paper lose no generality by being restricted to monotone splines.

It is hoped that enough "how to" material can be found in this paper to enable the reader to move quickly to trying out splines in his favorite context. Emphasis here is on computation rather than mathematics, because there are now many excellent references on the mathematics of splines at all levels (e.g., Chambers, 1977; de Boor, 1978; Schumaker, 1981).

2. WHAT ARE SPLINES?

In this section the concept of a *polynomial regression spline* defined on an interval $[L, U]$ is introduced as a piecewise polynomial with specified continuity constraints. The continuity characteristics of the spline and the number of independent parameters which define it depend on a *knot sequence* t which partitions this interval into a number of subintervals. Associated with t is a suitable set of basis splines that can be combined linearly to yield any other spline associated with this knot sequence. This set of basis functions is

J. O. Ramsay is Professor and Chairman, Department of Psychology, McGill University, 1205 Dr. Penfield Avenue, Montreal, Quebec, Canada H3A 1B1.

then modified to provide a computationally convenient and meaningful basis for monotone splines, referred to as I-splines. A monotone spline function can be constructed from these I-splines by taking a nonnegative linear combination. The approach is displayed in Figure 1 showing such a monotone spline. Further discussion and interpretation of Figure 1 is given later in this section.

In the first part of this section the properties of the knots are described, and the second part defines the M-spline and I-spline bases. Readers unconcerned or already familiar with these details may wish to skip to the next section discussing computational issues or to Section 4 in which applications are taken up.

Only a very limited introduction to the vast and fascinating topic of splines is possible here. Fortunately, there are now many excellent introductions to both the mathematics of splines, and also to their applications to data analysis. The review of splines in statistics by Wegman and Wright (1983) is particularly recommended.

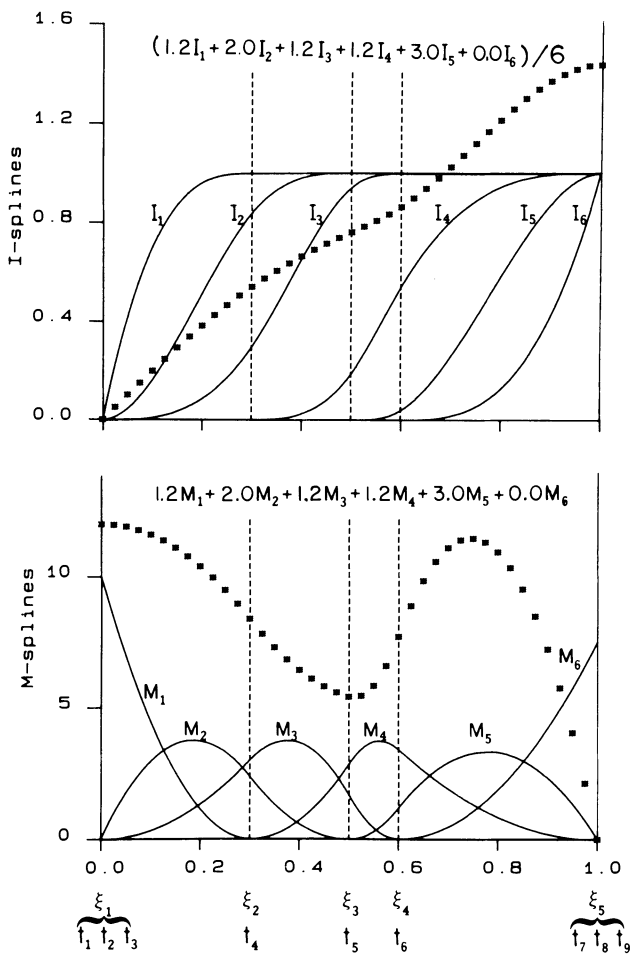

FIG. 1. *The six M-splines and associated I-splines of order 3 associated with the interior knots .3, .5 and .6. The dotted lines indicate the linear combination of splines resulting from the coefficients* 1.2, 2.0, 1.2, 1.2, 3.0 *and* 0.0 *(and divided by* 6 *for the I-splines). Note that the effect of the final zero coefficient is an M-spline which returns to zero and an I-spline which has zero slope at unity.*

Polynomials, $f(x) = \sum_{i=1}^{k} a_i x^{i-1}$, owe their central role in practical mathematics to two features: they are linear in the parameters a_i to be estimated, and the functions x^{i-1} that are linearly combined are easy to manipulate algebraically and numerically, especially with respect to differentiation and integration. A transformation family which does not have both of these features will not be likely to find wide applicability.

However, polynomials do have one serious limitation: a lack of flexibility in the sense that changing the behavior of f near one value x_1 has radical implications for its behavior for any other value x_2. Thus, a polynomial transformation which is quite satisfactory for values of x in, say, the central portion of their distribution, may exhibit very unpleasing features in the tails. This can be disastrous for certain problems such as density or distribution function estimation requiring some predetermined characteristics such as nonnegativity and small derivatives in these regions. This poses the problem of how to retain flexibility where it is needed, while leaving the function elsewhere either relatively unaffected or constrained as desired.

A *polynomial spline* achieves these objectives by constructing f from polynomials joined end-to-end. That is, an interval $[L, U]$ is subdivided by a mesh Δ consisting of points $L = \xi_1 < \cdots < \xi_q = U$. Within any subinterval $[\xi_j, \xi_{j+1})$ the function is a polynomial P_j of specified degree $k - 1$ or order k.

Crucial to the practical value of splines is the fact that adjacent polynomials are required to join with a specified degree of smoothness at the boundaries ξ_j. Smoothness is usually defined as the equality of the derivatives, $(D^{m-1}P_j)(\xi_{j+1}) = (D^{m-1}P_{j+1})(\xi_{j+1})$, $m = 1, \cdots, v_j$, where the notation $(D^{m-1}P_j)(\xi)$ means $d^{m-1}P/dx^{m-1}$ evaluated at argument ξ if $m > 1$ and $(D^0P)(\xi) = P(\xi)$. In the most common case, all orders of continuity v_j are specified as the degree $k - 1$ of the polynomial, so that adjacent polynomials have matching derivatives up to order $k - 2$. That is, if $k = 2$ the spline consists of straightline segments required to match at the boundaries ξ_j, and if $k = 3$, the spline is piecewise quadratic with matching first derivatives. However, provision must be made for other situations, and, for example, if $v_j = 0$ then the spline is permitted to be discontinuous at ξ_j. Thus the spline for $k = 1$ is a step function discontinuous at the boundaries. At the other extreme, if all values of v_j are k, the entire spline function is a single polynomial of degree $k - 1$ or order k.

In order to provide a useful system of notation and to develop a representation of a spline function that is convenient from the points of view of application

and computation, the mesh Δ and the continuity conditions v_j are incorporated into a *knot sequence* $t = \{t_1, \cdots, t_{n+k}\}$ where n is the number of free parameters that specify the spline function having the specified continuity characteristics. The knot sequence has the properties:

1. $t_1 \le \cdots \le t_{n+k}$.
2. For all i there is some j such that $t_i = \xi_j$.
3. The continuity characteristics are determined by:
   3.1. $t_1 = \cdots = t_k = L$ and $U = t_{n+1} = \cdots = t_{n+k}$;
   3.2. $t_i < t_{i+k}$ for all i;
   3.3. if $t_i = \xi_j$ and $t_{i-1} < \xi_j$, then

$$t_i = \cdots = t_{i+k-v_j-1}. \tag{1}$$

That is, the knot sequence t is derived from the mesh Δ by placing the number of knots at a boundary value ξ_j according to the order of continuity at that boundary. In the simplest situation, a single knot occurs at a boundary value, and property 3.3 then implies that the order of continuity is $k - 1$ and thus that $k - 2$ derivatives match. Multiple knots at a boundary value implies a lower order of continuity, and k knots at a boundary point implies discontinuity. In particular, property 3.1 indicates that the spline is permitted to be discontinuous at the end points of the interval. An example of the relationship between the knot sequence t and the mesh Δ generating it is displayed in Figure 1.

Because a spline of degree $k - 1$ is a polynomial at any point ξ, it is determined in the absence of any continuity constraints by k free coefficients on the subinterval containing that point. However, the continuity conditions impose v_j linear equality constraints on the coefficients defining adjacent polynomials. It turns out that the total degrees of freedom or number of independent variables in a spline with specified degree and continuity conditions is given by the value n defined implicitly in the above definition of the knot sequence t. Thus, there are a total of $n + k$ knots to be allocated, and the number of mathematical parameters or the dimension of the function space associated with a spline given Δ and the v_j's is n. In the most common situation where maximal continuity $k - 1$ holds at each boundary point in the interior of $[L, U]$, the dimension n equals the number of knots interior to $[L, U]$ plus the order k.

Although the idea of a spline as a piecewise polynomial is intuitively simple, its widespread application required the development of a suitable set of *basis splines*, $M_i(\cdot \mid k, t)$, $i = 1, \cdots, n$, such that any piecewise polynomial or spline f of order k and associated with knot sequence t could be represented as the linear combination $f = \sum a_i M_i$. The earliest basis was provided by the truncated power functions

$$\phi_{jm}(x) = [(x - \xi_j)_+]^m, \quad m = v_j, \cdots, k - 1,$$

where $u_+ = \max\{u, 0\}$. Although the simplicity of this basis makes it attractive for statistical work, and Smith (1979) and others have used it effectively in applications, it has the rather serious disadvantage of generating considerable rounding error except for very low values of k. Moreover, truncated power functions do not seem to have a natural interpretation in most applications, unlike the splines comprising the bases specified below.

A set of basis splines particularly appealing to statisticians is the M-spline family in which M_i, $i = 1, \cdots, n$, is defined such that it is positive in (t_i, t_{i+k}), zero elsewhere, and has the normalization $\int M_i(x)\, dx = 1$ (Curry and Schoenberg, 1966). Although they can be defined in terms of divided differences of truncated power functions, from a computational perspective they are more appropriately specified for $t_i \le x < t_{i+1}$ by the recursion

$$M_i(x \mid 1, t) = \frac{1}{t_{i+1} - t_i}, \quad t_i \le x < t_{i+1}, \text{ and 0 otherwise,}$$

$$M_i(x \mid k, t) = \frac{k[(x - t_i)M_i(x \mid k-1, t) + (t_{i+k} - x)M_{i+1}(x \mid k-1, t)]}{(k-1)(t_{i+k} - t_i)}, \quad k > 1. \tag{2}$$

The fact that $M_i(x \mid k, t) > 0$ only when $t_i \le x < t_{i+k}$ and is zero otherwise is a very important property from the standpoint of controlling rounding error in the computations.

Each M_i has the properties of a probability density function over the interval $[t_i, t_{i+k}]$, and in particular $M_i(x \mid 1, t)$ is the uniform density on the interval $[t_i, t_{i+1}]$ and $M_i(x \mid 2, t)$ is the triangular density on $[t_i, t_{i+2}]$ with mode at t_{i+1}. In general the random variable X having distribution $M_i(x \mid k, t)$ has moments

$$E[X \mid M_i(x \mid k, t)] = (t_i + \cdots + t_{i+k})/(k + 1),$$

$$\operatorname{Var}[X \mid M_i(x \mid k, t)] = \sum_{j=l+1}^{i+k} \sum_{l=i}^{i+k} (t_j - t_l)^2/[(k + 1)^2(k + 2)]. \tag{3}$$

From a computational perspective, because each M_i is itself a spline, and hence a piecewise polynomial, all the desirable features of polynomials such as computability, linearity and differentiability obviously carry over to splines. Moreover, because the value of M_i is zero outside of the interval $[t_i, t_{i+k}]$ and positive within it, a change in coefficient a_i will only affect f within

this interval, thus achieving a very desirable local sensitivity to coefficient values. The nonnegativity of $f = \sum a_i M_i$ is assured by $a_i \geq 0$ and its unit area by $\sum a_i = 1$. Moreover, because only M_1 and M_n are nonzero at L and U, respectively, $f(L) = 0$ or $f(U) = 0$ is achieved by setting a_1 or a_n to zero, respectively.

A closely related basis is given by the B-splines, $B_i = (t_{i+k} - t_i)M_i/k$. These have the alternative normalization $\sum B_i(x) = 1$ for all x, which can be especially useful in modeling discrete data. Although consideration of nonpolynomial splines is beyond the scope of this article, it is worth pointing out that splines composed of piecewise linear combinations of exponential, trigonometric and other function families are also possible (de Boor, 1978; Schumaker, 1981).

The focus of this article is monotone splines, and one technique for defining these is to employ a basis consisting of monotone splines. Because M-splines are nonnegative, one obvious approach is to define the *integrated splines* I_i, or *I-splines* for the sake of brevity, as

$$I_i(x \mid k, t) = \int_L^x M_i(u \mid k, t)\, du, \tag{4}$$

and this provides a set of splines which, when combined with nonnegative values of the coefficients a_i, yields monotone splines. Because each M_i is a piecewise polynomial of degree $k - 1$, each I_i will be a piecewise polynomial of degree k.

We shall use the term *order* k to refer to an M-spline of degree $k - 1$ or the associated I-spline of degree k. This permits us to refer in either case to the number of free parameters n as being the number of interior knots plus the order for simple knot sequences having only one knot at each interior boundary. For such simple knot sequences, for which $t_j \leq x < t_{j+1}$ for all x, the I-spline I_i can be put in the more convenient form

$$I_i(x \mid k, t) = \begin{cases} 0, & i > j, \\ \sum_{m=i}^{j} (t_{m+k+1} - t_m) M_m(x \mid k+1, t)/(k+1), & j - k + 1 \leq i \leq j, \\ 1, & i < j - k + 1. \end{cases} \tag{5}$$

In the useful special case of $k = 2$, when I_i will be piecewise quadratic, zero for $x < t_i$ and unity for $x > t_{i+2}$, one has the direct expressions

$$I_{i-1}(x \mid 2, t) = 1 - \frac{(t_{i+1} - x)^2}{(t_{i+1} - t_i)(t_{i+1} - t_{i-1})},$$

$$I_i(x \mid 2, t) = \frac{(x - t_i)^2}{(t_{i+1} - t_i)(t_{i+2} - t_i)}. \tag{6}$$

Figure 1 displays the family of M-splines defined on [0, 1] of order 3 and associated with interior knots .3, .5 and .6. Each M-spline is piecewise quadratic and nonzero over at most three intervals. An M-spline of order k will have $k - 2$ continuous derivatives at the location of a knot within its interval of support. It loses one level of continuity for each additional knot positioned at that location. If the usual practice of positioning k knots at L and U is followed, then the lower and upper splines will be discontinuous at these boundary values, the adjacent M-splines will have discontinuous first derivatives, and so on.

Also displayed in Figure 1 is the result of combining these six M-splines with the coefficients indicated. Note that there is a close relationship between the value of the resulting function and the size of the coefficient for the nearest M-spline. The I-splines associated with these M-splines are also indicated in the top figure, along with the behavior of the resulting linear combination. The fact that the final coefficient is zero imposes a zero slope on this function at unity.

In the case of natural monotonic ordering of variable values considered in this paper, the isotonic regression techniques discussed by Barlow, Bartholomew, Bremner and Brunk (1972) and Kruskal (1965) can be viewed as a particular case of monotone splines. The piecewise linear monotone functions that these procedures generate correspond to monotone splines defined as above with $k = 1$ and a knot positioned at each data point. This achieves optimal flexibility, but one pays a rather heavy price in the sense that such functions may not be pleasing to the eye in terms of smoothness, and they imply a very large number of estimated parameters. Illustrations in this article will concentrate on very much smoother splines with the goal of keeping the number of estimated parameters as small as possible.

Wegman and Wright (1983) and Smith (1979) review some of the large literature on statistical applications of splines, including the important work on smoothing splines by Wahba (1978) and Craven and Wahba (1979). The splines used in this paper are called *regression splines* to distinguish them from *interpolating* and *smoothing* splines, where the objective is to fit to data a curve which minimizes the loss function

$$\min_f \left\{ \sum_r^R [Y_r - f(X_r)]^2 + \lambda \int [f''(x)]^2\, dx \right\},$$

where (X_r, Y_r), $r = 1, \cdots, R$, are bivariate observations. The first term in this smoothing criterion indicates the lack of fit of f to the data, and the second measures the degree of smoothness of f and is therefore a roughness penalty. Smoothing parameter $\lambda \geq 0$ controls the trade-off between smoothness and fit, with interpolation resulting from $\lambda \to 0$. No

assumptions are made about f aside from having an integrable second derivative. The solution is a polynomial spline of degree 3 with knots positioned at the values of x. More general splines are obtained by using an arbitrary linear differential operator L instead of the operator D^2 in the above integration. Silverman (1985) offers an extensive discussion of smoothing in a statistical context.

What primarily distinguishes interpolation and smoothing splines from regression splines is the matter of knot placement; knots are positioned at fixed values and are usually much fewer in number in the regression spline case. From this perspective regression splines imply in general far fewer coefficients to be estimated from the data, although necessarily at some expense in terms of flexibility. In statistical applications, where confirmation or prediction is important, the estimation and inferential advantages of working with a very restricted number of parameters tends to favor regression splines.

Regression splines also offer more control over the characteristics of the transformation. Although Wright and Wegman (1980) describe an approach to fitting monotone smoothing splines, and Passow (1974) and Passow and Roulier (1977) discuss monotone interpolation splines, in general monotonicity and other boundary conditions are much easier to impose for regression splines, where the basis splines determined by the knot sequence t and order k are regarded as a fixed basis for a finite dimensional function space. Computing with regression splines involves well-established techniques and is discussed in the next section. Moreover, as will be indicated in the next section, it is a relatively simple matter to constrain one or more derivatives to be zero as well as to impose normalization constraints.

The Box-Cox transformation family, $f(x; \lambda) = (x^\lambda - 1)/\lambda$ (Box and Cox, 1964), also requires mention in any discussion of data transformation. This is an attempt to provide a useful range of transformations using only a single parameter λ. Where the resulting curve fits the data sufficiently well, this is a very desirable approach. Most of the applications discussed here, however, will require more flexibility than this technique can provide.

3. COMPUTING WITH MONOTONE SPLINES

Once a suitable interval $[L, U]$ for transformation $f = \sum a_i I_i$ and a knot sequence t have been chosen, the estimation of an f assumes the familiar form of optimizing a suitable criterion with respect to the transformation parameters a_i subject to a variety of linear equality or inequality constraints. The conditions $a_i \geq 0$ are sufficient to impose monotonicity, and $\sum a_i = 1$ implies $f(U) = 1$. Moreover, if for some $m \leq k$ we have all $a_i = 0$, $i = 1, \cdots, m$, then $(D^{i-1}f)(L) = 0$, $i = 1, \cdots, m$, so that the initial value of f and its order of contact can be easily controlled. Similarly, if $a_{n-i} = 0$, then $(D^{i+1}f)(U) = 0$, $i = 0, \cdots, m$.

The Appendix describes the gradient projection constrained optimization algorithm used in the applications described in the next section, and further computational remarks will be made in discussing them.

The choice of the interior knots is obviously an important consideration in any application, because these determine the shapes of the basis splines M_i or I_i, and thus in turn the shape of the final function. This choice should be guided by two considerations:

1. The more knots in a region, the greater the flexibility of the function in that region.
2. The more data points there are between a pair of knots, the better defined the curve in that interval.

The second consideration tends to dominate, because there is little point to positioning many knots in a region where a lack of data points would imply poor definition of the function. Nevertheless, if other considerations imply more nonlinearity in one region than in another, both having a reasonable amount of data points, it makes sense to use more knots in the former than in the latter. In practice, however, there is often enough flexibility in a curve defined by a single interior knot, and it is not usually necessary in a statistical environment to use large numbers of knots. Moreover, unless great flexibility is required (accompanied inevitably by a large number of data points), very low values of k such as 2 will suffice. This will be illustrated below.

A useful preliminary knot placement is to position a single interior knot at the median, or two interior knots at the terciles, and so on. There are techniques for automatic refinement of knot placement (de Boor, 1978), but it remains to be seen whether these will prove useful in statistical applications. A number of papers have studied knot placement as a design problem (Agarwal and Studden, 1978; Park, 1978, for example). In general the shape of a spline function is not very sensitive to knot placement.

4. APPLICATIONS

In this section a number of applications of monotone splines to real data analytic problems are presented with a view to displaying some of their potential. Not included, however, is the most obvious application: splines can and have often been used in nonlinear regression to transform an independent variable x so as to minimize $\sum [y_i - f(x_i)]^2$ (Wegman and Wright, 1983; Wold, 1974). Instead, this problem is subsumed in a couple of other applications.

4.1 Linear Regression with Response Transformation: The Yarn Data

The first application is to data frequently used to illustrate dependent variable transformation techniques. Box and Cox (1964) and Kruskal (1965), among others, transformed 27 observations of the breaking strength of yarn gathered in a three-way factorial design. The optimal Box-Cox power transform was $(Y^{-0.06} - 1)/(-0.06)$, a curve that very closely approximates the natural logarithm.

Let Y be a random response variable, and let f be a transformation of Y which is determined by transformation parameter vector a. Assume that the transformed response $f(Y; a)$ is distributed as $N[g(x; \theta), \sigma^2]$, where g is a function of a known covariate vector x and structural parameter vector θ.

The additive model for observation Y_{ijk} at levels i, j and k of the three factors for the yarn data is $E[f(Y_{ijk}; a)] = \mu + \alpha_i + \beta_j + \gamma_k$. Thus, the structural parameter vector θ contains the constant term μ and three regression coefficient vectors α, β and γ for the dummy variables coding the effects subject to the constraints $\sum \alpha_i = \sum \beta_j = \sum \gamma_k = 0$.

Using more general notation, the problem is to estimate the dependent variable transformation f which is defined by parameter vector a of dimension n, and the structural parameter vector θ, which contains a constant term θ_0 and p regression coefficients θ_j. The data consist of a set of dependent and independent variable observations, y_r and x_{rj}, $r = 1, \cdots, R, j = 1, \cdots, p$, respectively. In the yarn data, the y_r's are the 27 observations in the 3^3 design and the x_{rj}'s are elements of the design matrix for the additive model.

The fitting criterion could be least squares, but this is not desirable when the dependent variable is being transformed, and maximum likelihood or Bayesian approaches which take into account the Jacobian $J(y; a) = \partial f/\partial y$ of the transformation are preferable (Box and Cox, 1964). The maximum log likelihood is

$$
\begin{aligned}
L_{\max} &= \max_{a,\theta,\sigma} \left\{ -R \ln \sigma \right. \\
&\qquad \left. - \frac{1}{2} S(y, x; a, \theta)/\sigma^2 + \ln J(y; a) \right\} \\
&= \max_{a,\theta} \left\{ -\frac{R}{2} [\ln S(y, x; a, \theta) - \ln R + 1] \right. \\
&\qquad \left. + \ln J(y; a) \right\},
\end{aligned} \tag{7}
$$

where the error sum of squares $S(y, x; a, \theta)$ and the Jacobian $J(y, a)$ are

$$
\begin{aligned}
S(y, x; a, \theta) &= \sum_r^R \left[f(y_r; a) - \sum_j^p x_{rj}\theta_j - \theta_0) \right]^2, \\
\ln J(y; a) &= \sum_r^R \ln \frac{\partial f}{\partial y} (y_r; a).
\end{aligned} \tag{8}
$$

It may be practical to determine $L_{\max}$ by alternating between maximizing with respect to a and maximizing with respect to θ.

The monotone spline transformation $f(y; a) = \sum a_i I_i(y)$ was specified to be over the interval [0, 4000]. Figure 2 shows the result of using order $k = 2$ and placing a single interior knot at the median value of 566. This implies a total of three transformation parameters, and their estimated values are .533, .442 and .024. The resulting transformation closely resembles a logarithm, and in fact the resulting $L_{\max} = -137.57$ did not exceed the value of -136.44 resulting from the optimal power transformation. Moreover, increasing the flexibility of the spline transformation by placing interior knots at the terciles and quartiles yielded $L_{\max}$ values of -138.69 and -137.49, respectively, so that it seems clear that monotone spline transformations cannot improve on a power or logarithmic transformation for these data.

Interval estimates for the transformation and structural parameters may be obtained by asymptotic techniques, jackknifing or bootstrapping. Parameter estimates which are zero require some special consideration, but a useful procedure is simply to treat them as fixed and to derive interval estimates for the remainder. Interval estimates for the values of the

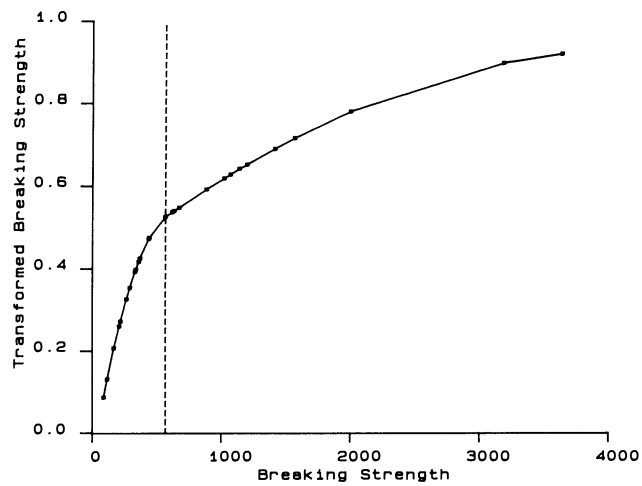

FIG. 2. *Estimated monotone spline transformation of the breaking strength of yarn fit by a linear model with three qualitative independent variables. The interval of transformation was* [0, 4000] *and the vertical dashed line indicates the location of the single interior knot, which was positioned at the median. Points correspond to data values.*

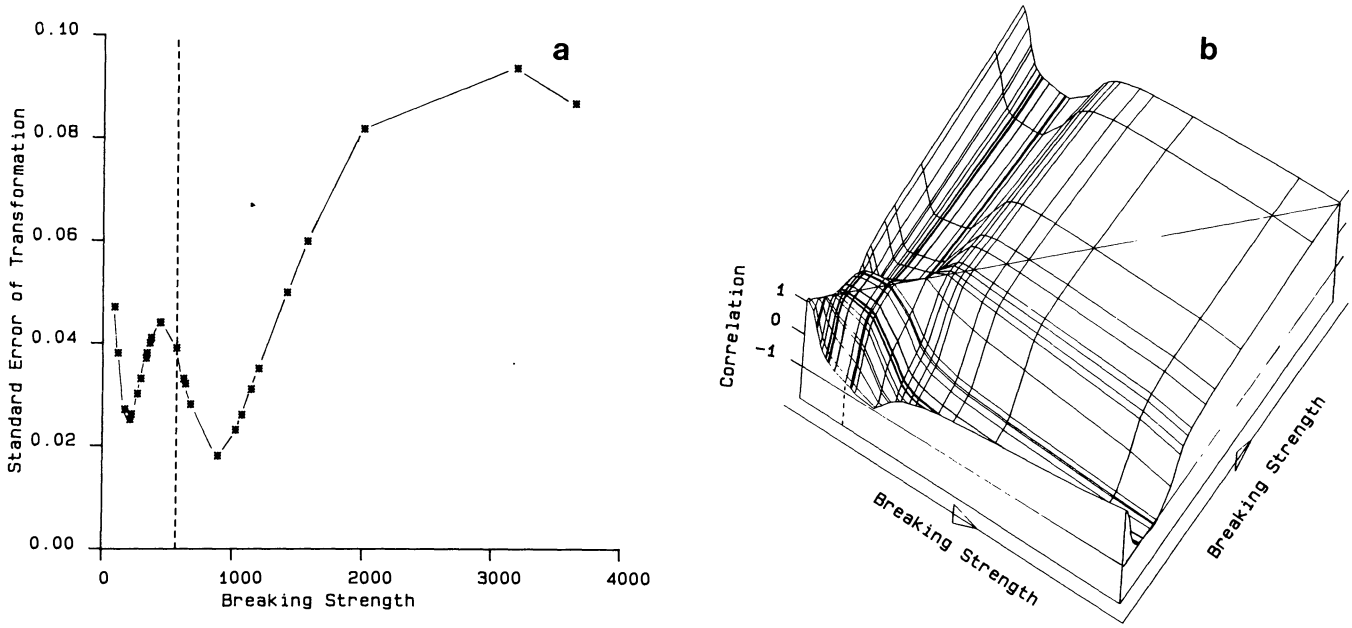

FIG. 3. a, *standard error of transformation value estimated by jackknifing.* b, *correlation among transformation values estimated by jackknifing. The height of the surface at any pair of breaking strength values indicates the amount of correlation between corresponding transformation values. Cross-hatching is at data values, and the edges are at the least and greatest values. The location of the single interior knot is indicated by the vertical dashed line.*

transformation at various dependent variable values can then be obtained either by the same techniques or from the estimated sampling variance of a.

Because in this application only three iterations were typically required to achieve convergence, jackknifing by deleting one observation at a time was not expensive. Figure 3a displays the estimated standard error of the estimated curve location. Standard error tends to be high near the end points of the transformation where there are fewer data points, and near the single interior knot where the curve has the greatest flexibility.

Figure 3b shows the correlations among curve locations as a surface whose height at any point is the correlation between function values at corresponding abscissa values. Note that when both values are in regions of high standard error due to the lack of data, the correlation between curve locations on opposite sides of the single interior knot falls off to nearly -1.0, but is near unity when both points are on the same side. On the other hand, values at opposite sides of the knot but each in a region of minimal standard error have nearly zero correlation due to the flexibility of the curve in these regions. This lack of coupling of curve values among distinct regions where the curve is most flexible is one of the great virtues of splines.

4.2 Nonlinear Regression: Dissimilarity Judgments

That this first example fails to do better than previous analyses is not at all disappointing; one is reassured when more powerful tools reproduce results obtainable more simply. The next example is drawn from multidimensional scaling where the response variable is a judgment on the degree of dissimilarity of two stimuli on a rating scale. In the typical experiment, each of the $\binom{V}{2}$ unordered pairs among V objects is presented to a respondent, who is requested to rate the degree of similarity or dissimilarity of the two objects. The objective of the analysis is a representation of each object v by a point in a Euclidean space of the appropriate dimensionality J, with its coordinates relative to some coordinate system being ξ_{vj}. This representation is arrived at by treating the dissimilarity rating d_{uvr} for object pair (u, v) by respondent r as an index of distance.

People vary remarkably in the way in which they use rating scales in general, with some showing tendencies to avoid extreme ratings, others using specific categories disproportionately often and still others piling their judgments up against one extreme of the scale. Thus, it is more realistic to seek a monotone transformation $f_r(d_{uvr})$ of a dissimilarity rating which can be treated as a distance measure rather than using raw judgment. These and other idiosyncracies of human judgmental data require very flexible transformation tools.

In the multidimensional scaling model used in this analysis (Ramsay, 1982), the judgment d_{uvr} of the degree of dissimilarity of stimulus objects u and v, $u, v = 1, \cdots, V$, by subject r, $r = 1, \cdots, R$, is

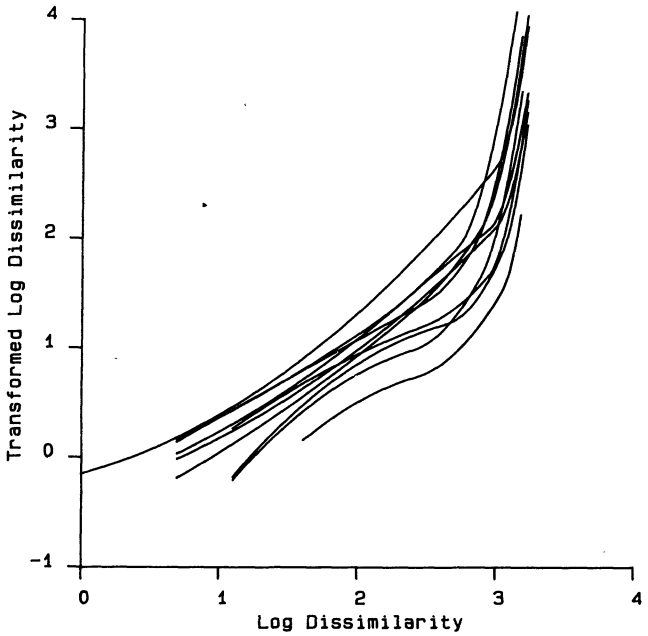

FIG. 4. *Ten monotone spline transformations of dissimilarity judgments in a multidimensional scaling analysis. Order two splines were used, and knots were placed at the lowest and highest responses, and at the ordered observations dividing the data in equal thirds.*

approximated by a distance $\delta_{uvr} = [\sum_j^J (\xi_{uj} - \xi_{vj})^2]^{1/2}$. The parameters to be estimated are the VJ coordinates ξ_{uj}. It is assumed that transformed independent judgments for a specific stimulus pair have a log-normal distribution about the model value, or $f_r(\log D_{uv}) \sim N(\log \delta_{uv}, \sigma_r^2)$, so that the R variance parameters σ_r^2 along with the coordinates comprise the structural parameter θ. Because each subject's data are transformed separately, there are R transformation parameter vectors a_r, each containing the coefficients a_{ir} determining the spline transformations. In this application each transformation also includes a constant term a_{0r}.

The stimulus objects in this example were fifteen types of recreational activities, and each of ten subjects rated all unordered 105 pairs on a rating scale with 25 categories. The first category was labeled "extremely similar" and the last "extremely different." Order 2 spline transformations were applied to the log responses, and knots were located at the extreme responses and at the terciles, which divide the responses for a specific subject into equal thirds. Figure 4 shows the ten estimated spline transformations, and indicates that all ten subjects required an expansion of the upper end of the response scale to be consistent with the model. A comparison of these results, for which the log likelihood was −1760, with those using a linear transformation for each subject's log dissimilarities, which yielded a log likelihood of −2267, indicates that the use of 20 extra parameters was well worthwhile.

4.3 Monotone Spline Regression: City Gasoline Consumption

The next few illustrations involve the analysis of the data in Table 1, comprising various measures on 44 automobiles reported in the April 1986 issue of *Consumer Reports*. Among these are the two measures of gasoline consumption that are of interest as dependent variables, and two other measures, engine displacement and weight, that are obvious explanatory variables. Weight is relevant because much of the energy expended in city driving goes into acceleration, but one may wonder whether a car with a large engine will use more gas than is explainable by its extra weight.

A conventional regression analysis in which city gas consumption is predicted by displacement and weight yields a squared multiple correlation of 0.799 and shows a contribution of displacement which is significant at only the 0.03 level. However, these variables may not be linearly related, and we wish to know whether a nonlinear but monotone transformation of each variable will yield different conclusions about an additive model as well as provide a better fit.

Thus, we seek a transformation of the dependent variable $f_0(Y)$ and independent variable transformations $f_j(X_j)$, $j = 1, \cdots, p$, such that the following additive model holds:

$$(9) \quad f_0(y_r) = \sum_{j=1}^{p} f_j(x_{rj}) + a_0 + e_r, \quad r = 1, \cdots, R,$$

where e_r is the residual for the rth observation and a_0 is a constant term. Each of the transformations has the form

$$f_j(\bullet) = \sum_{i=1}^{n_j} a_{ij} I_{ij}(\bullet \mid k_j, t_j), \quad j = 0, \cdots, p,$$

where U is the upper limit on the interval of transformation and k_j and t_j are the order and knot sequence for the jth transformation, respectively. In order to fix the scale of the model, we can impose the constraint $f_0(U) = 1$, which implies that $\sum a_{i0} = 1$.

Figure 5 displays the order 2 monotone spline transformations of these three variables using a single interior knot positioned at the median for each variable. The domain of the transformation in each case was the range of the variables. Also shown in Figure 5 are estimated 95% asymptotic confidence limits on curve location, where each coefficient which was given a zero estimate was regarded as fixed (Winsberg and Ramsay, 1980). The FORTRAN 77 computer program used in these analyses can be obtained from the author by sending a diskette.

No explicit regression coefficients are required in (9) because the scale of each f_j is unrestricted. The derivative $f_j'(X_j) = \sum a_{ij} M_{im}(X_j)$ is therefore an

TABLE 1
Automobile data from 1986 *Consumer Reports*

Automobile	Plot symbol in Fig. 9	Price	Displacement	City gas	Expwy gas	Weight
Chevrolet Chevette	A	56	1.6	13.3	6.8	10.0
Chevrolet Nova	B	74	1.6	10.5	5.5	10.2
Chevrolet Spectrum	C	67	1.5	10.1	5.3	8.7
Dodge Colt	D	56	1.5	11.0	5.6	9.9
Dodge Omni	E	68	1.6	11.5	5.6	9.5
Ford Escort	F	61	1.9	12.0	6.2	10.9
Honda Civic	G	56	1.5	11.5	6.5	9.2
Mitsubishi Tredia	H	75	2.0	12.0	6.3	10.9
Nissan Sentra	I	56	1.6	10.5	5.6	9.5
Renault Alliance	J	60	1.4	12.0	5.6	9.1
Subaru	K	80	1.8	12.0	5.6	10.4
Toyota Corolla	L	73	1.6	11.0	5.3	10.3
Toyota Tercel	M	56	1.5	11.0	5.5	9.7
Volkswagen Golf	N	74	1.8	12.0	6.3	10.1
Volkswagen Jetta	O	83	1.8	12.6	6.3	10.5
Chrysler Laser	P	94	2.2	15.8	7.4	12.7
Honda Civic CRX	Q	73	1.5	9.4	5.6	9.0
Honda Prelude	R	110	1.8	10.5	6.5	10.6
Isuzu Impulse	S	109	1.9	14.9	6.8	12.4
Mitsubishi Cordia	T	89	2.0	12.0	6.3	11.2
Nissan 200SX	U	95	1.8	12.6	6.5	12.2
Nissan Pulsar NX	V	87	1.6	9.7	5.1	9.2
Pontiac Fiero	W	89	2.5	12.6	6.7	11.5
Audi 4000S	X	142	2.2	14.1	8.2	10.7
Chevrolet Cavalier	Y	67	2.0	15.8	7.2	11.6
Ford Tempo	Z	74	2.3	13.3	6.2	11.8
Honda Accord	1	88	2.0	13.3	6.3	11.8
Mazda 626	2	92	2.0	12.6	6.7	11.8
Mitsubishi Galant	3	132	2.4	14.9	6.7	12.9
Nissan Stanza	4	101	2.0	11.0	5.6	11.1
Oldsmobile Calais	5	93	2.5	14.1	6.7	12.0
Saab 900	6	126	2.0	14.1	7.9	12.9
Toyota Camry	7	97	2.0	11.0	5.5	12.2
Volvo	8	144	2.3	14.1	7.9	13.3
Buick Century	9	101	2.5	16.9	6.5	12.6
Chrysler Fifth Avenue	0	149	5.2	23.0	9.7	16.2
Chrysler Le Baron	@	100	2.2	13.3	7.9	11.7
Dodge Aries	#	72	2.2	13.3	7.9	11.5
Dodge Lancer	$	94	2.2	15.8	8.7	12.7
Mercury Cougar	%	114	3.8	15.8	8.7	14.1
Oldsmobile Cutlass	/	107	3.8	19.5	9.0	15.2
Pontiac 6000	&	95	2.8	16.9	8.7	12.6
Buick Electra	<	154	3.8	18.1	7.7	15.0
Chevrolet Caprice	>	106	5.0	18.1	8.2	16.5

Notes: Price is measured in 100 U. S. dollars, displacement in liters, gasoline consumption in liters/100 km, and weight in 100 kg.

indicator of the dependency of $E(f_0)$ on the independent variable X_j, and thus plays the role of the regression coefficient in a conventional linear model. Thus, the relative roles of weight and displacement in determining gas consumption may now be assessed for various regions of the latter by examining the relative slopes f_j' of the three transformations. From these one may conclude that (a) weight is much more important than displacement for small cars (under 1200 kg), (b) both weight and displacement are important for mid-sized cars and cars with mid-sized engines, (c) neither weight nor displacement has as much explanatory power for very large cars with very large engines as for small or medium sized cars and (d) city gas consumption measured in liters/100 km might be improved as a measure by taking its square root.

Attempts to improve the fit using more interior knots or higher order splines were fruitless. A comparison of these results with the monotone spline regression of gas consumption on weight alone yielded an asymptotic χ^2 test value of 3.3 with 3 d.f., so that it appears that transformed displacement has little to

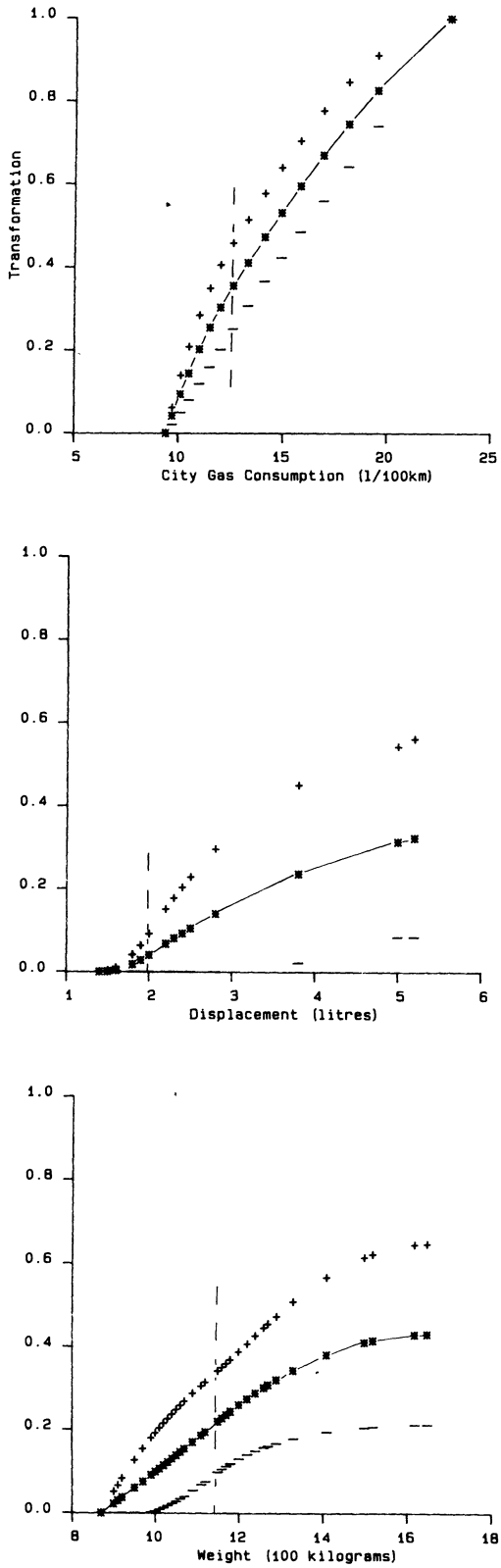

FIG. 5. *Estimated monotone spline transformations yielding the best regression of transformed city gas consumption on transformed displacement and weight. Vertical dashed lines indicate the location of the single interior knot, positioned at the median. "Plus" and "minus" signs indicated estimated upper and lower 95% confidence bands, respectively, for the transformation values. The dependent variable transformation is fixed at the boundaries of the transformation interval, and thus has zero standard error at these points.*

contribute to transformed weight as an explanatory variable, and its significance in the untransformed case was due to the high leverage of the large cars resulting from its skewed distribution.

Breiman and Friedman (1985) describe an algorithm for estimating the optimal transformation of variables in multiple regression which they refer to as ACE, an acronym for alternating conditional expectation. This procedure uses a smoothing process for each variable in turn conditioned on the transformations of the other variables estimated in previous iterations. If monotonicity is required, this can be incorporated into the particular smoothing algorithm employed. Although splines are not used in the algorithm that these authors use in their illustrations, they can be used as a basis for smoothing. Their procedure as well as the somewhat similar approach of Hastie and Tibshirani (1986, 1987) may be called semiparametric in that the smoothing process is not based on explicit parameters.

In these procedures optimization is with respect to each variable in turn rather than with respect to all transformations simultaneously. Unfortunately, as Figure 6 indicates, the ACE algorithm turned out to be sensitive to the ordering of the independent variables. The solution in which weight was the first independent variable resembles the monotone spline solution, but reversing the order yields a much stronger dependency on displacement. This is no doubt due to the fact that weight and displacement are strongly correlated ($r = 0.90$), and the sequential procedure employed by ACE seems to be the problem here.

There is also a theoretical explanation for this phenomenon. The authors explain that the algorithm is considered to approximate the solution to a continuous eigenvalue problem in function space, for which there are in principle a countable number of discrete solutions. However, there is no reason to assume that the eigenfunction of the continuous problem is associated with the largest eigenvalue, and hence with the maximum possible canonical correlation.

4.4 Monotone Spline Canonical Correlation: Gasoline Consumption

A natural extension of monotone spline regression involves a comparison between two sets of transformed variables via canonical correlation. Thus, for variables X_u, $u = 1, \cdots, p$, and Y_v, $v = 1, \cdots, q$, we seek two sets of monotone spline transformations $f_u = \sum a_{iu} I_{iu}$ and $g_v = \sum b_{iv} I_{iv}$. The goal is to choose these transformations plus coefficients c_u and d_v in such a way that the two linear combinations or canonical variables

$$\sum_u^p c_u f_u(x_{ru}) \quad \text{and} \quad \sum_v^q d_v g_v(y_{rv}), \quad r = 1, \cdots, R,$$

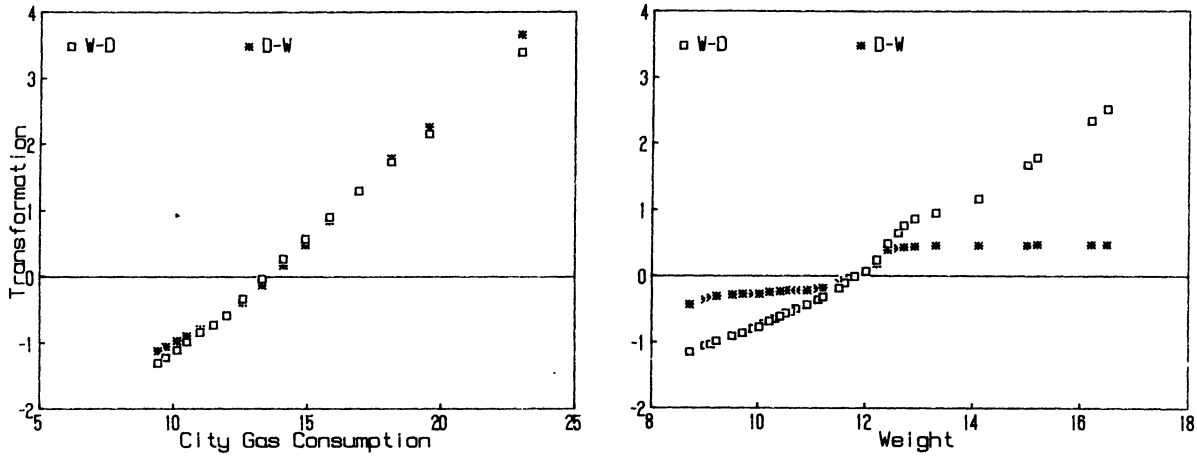

FIG. 6. *Two sets of estimated transformations using the ACE algorithm of Breiman and Friedman* (1985) *illustrating the order sensitivity of that algorithm. Results labeled "D-W" are obtained when displacement is used as the first independent variable, and those labeled "W-D" correspond to those using weight first. In the latter case the transformation of displacement was identically zero.*

have the largest possible correlation. Putting this problem in matrix notation, let matrices F and G, of dimensions $R \times p$ and $R \times q$, respectively, contain the values of the transformed observations, $f_u(x_{ru})$ and $g_v(y_{rv})$. We seek simultaneously canonical coefficient vectors c and d and monotone transformations such that the canonical correlation $\rho = c^t F^t G d$ is maximized subject to the normalizations $c^t F^t F c = d^t G^t G d = 1$.

It is convenient to split the computational problem into the familiar one of computing canonical coefficients c and d given transformations F and G on the one hand, and optimizing the canonical correlation ρ with respect to the transformation coefficients a_{iu} and b_{iv} that determine F and G on the other. Most applications will require a preliminary step in which the I-spline values $I_{iu}(X_{ru})$, $i = 1, \cdots, n_u$, are replaced by

$$I_{iu}(x_{ru}) - R^{-1} \sum_{r=1}^{R} I_{iu}(x_{ru}),$$

and similarly for I-spline values associated with the y_{rv}'s. This implies that the matrices F and G will have zero column sums, and thus define a value of ρ which is a correlation in the usual sense.

In this application one seeks to explore the relationship between the two gasoline consumption variables and the two size variables, weight and displacement, with a view to finding two pairs of transformations with the best correlation. These transformations are shown in Figure 7. Those for city gas consumption, displacement and weight resemble those obtained in the monotone regression analysis, except for the very small cars. Expressway gas consumption is transformed so that the role of very small and very large cars is enhanced by the transformation. The canonical correlation is $\rho^2 = 0.806$ and the associated canonical weights for city and expressway gas consumption are .85 and .52, respectively, and for weight and displacement .75 and .66, respectively. Thus, the inclusion of expressway gas consumption provides a role for engine displacement, especially for cars of extreme sizes.

4.5 Monotone Spline Principal Components

The previous examples have shown that gasoline consumption can be nicely accounted for in terms of weight with some small contribution from displacement. One suspects that price, although not unrelated to weight, represents a significant additional dimension of variation, and that it, too, will benefit from some nonlinear transformation. One way to address this question is to perform an analysis of the R observations in which the transformations optimize a bilinear approximation of fixed dimensionality K to the transformed data. This is often referred to in the psychometric literature as an Eckart-Young (1936) approximation of the transformed data (Winsberg and Ramsay, 1983).

Let $R \times p$ matrix $F(A)$ contain the transformed values of R observations of each of p variables, where $f_j(x_{rj}) = \sum a_{ij} I_{ij}(x_{rj})$, $r = 1, \cdots, R; j = 1, \cdots, p$, and A contains the coefficients a_{ij}, $i = 1, \cdots, n_j$, defining monotone spline transformations. The objective is to approximate F as follows:

$$\min_{A,B,C} \{\mathrm{tr}[F(A) - BC^t]^t[F(A) - BC^t]\}, \tag{10}$$

where B is the $R \times K$ matrix of scores on the principal components and C is the $p \times K$ matrix containing the regression coefficients on these components. This problem requires some normalization of the transformations f_j, and we shall require that they run from zero at the lowest observation to unity at the largest.

Winsberg and Ramsay (1983) discuss this problem as well as the closely related issue of transformation to multinormality with specified covariance structure. They also provide a computer program in the PROC MATRIX language of SAS (SAS Institute Inc., 1985).

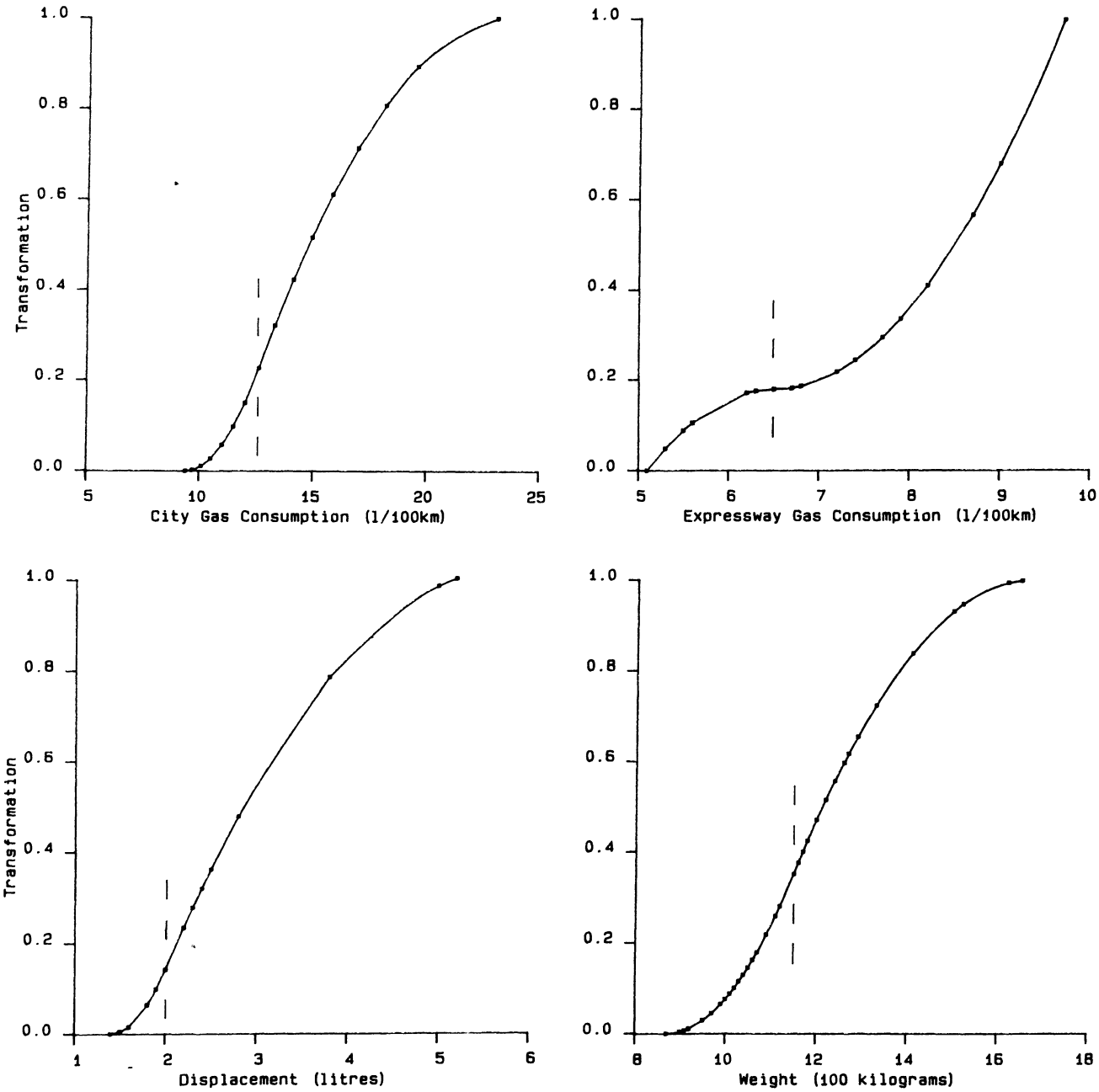

FIG. 7. *Estimated monotone spline transformations resulting from a canonical correlation between the transformed gas consumption variables in one set and transformed displacement and weight in the other. Dashed lines indicate locations of single interior knots at medians.*

Figure 8 displays the five transformations resulting from the monotone spline principal components analysis of the automobile data in two dimensions. Note that the strongly skewed variables, price and displacement, are transformed so as to have more symmetric distributions. Figure 9 shows by means of a biplot the nature of the two principal components. The first clearly measures the overall size of the car, while the second, forced to be orthogonal to the first, might be characterized as the extent to which the car costs less and consumes more gas than is consistent with its size.

4.6 Binomial Response: Test Item Analysis

The probability of one of two dichotomous responses often depends monotonically on the level of some covariate or linear combination of covariates. That is, given a sequence of Bernoulli variables, Y_j, and associated covariate values x_j, $j = 1, \cdots, p$, we have the model

$$\text{Prob}\{Y_j = 1 \mid x_j\} = P(x_j). \tag{11}$$

Because I-splines map a portion of the real line into $[0, 1]$, they are especially well suited to the analysis of

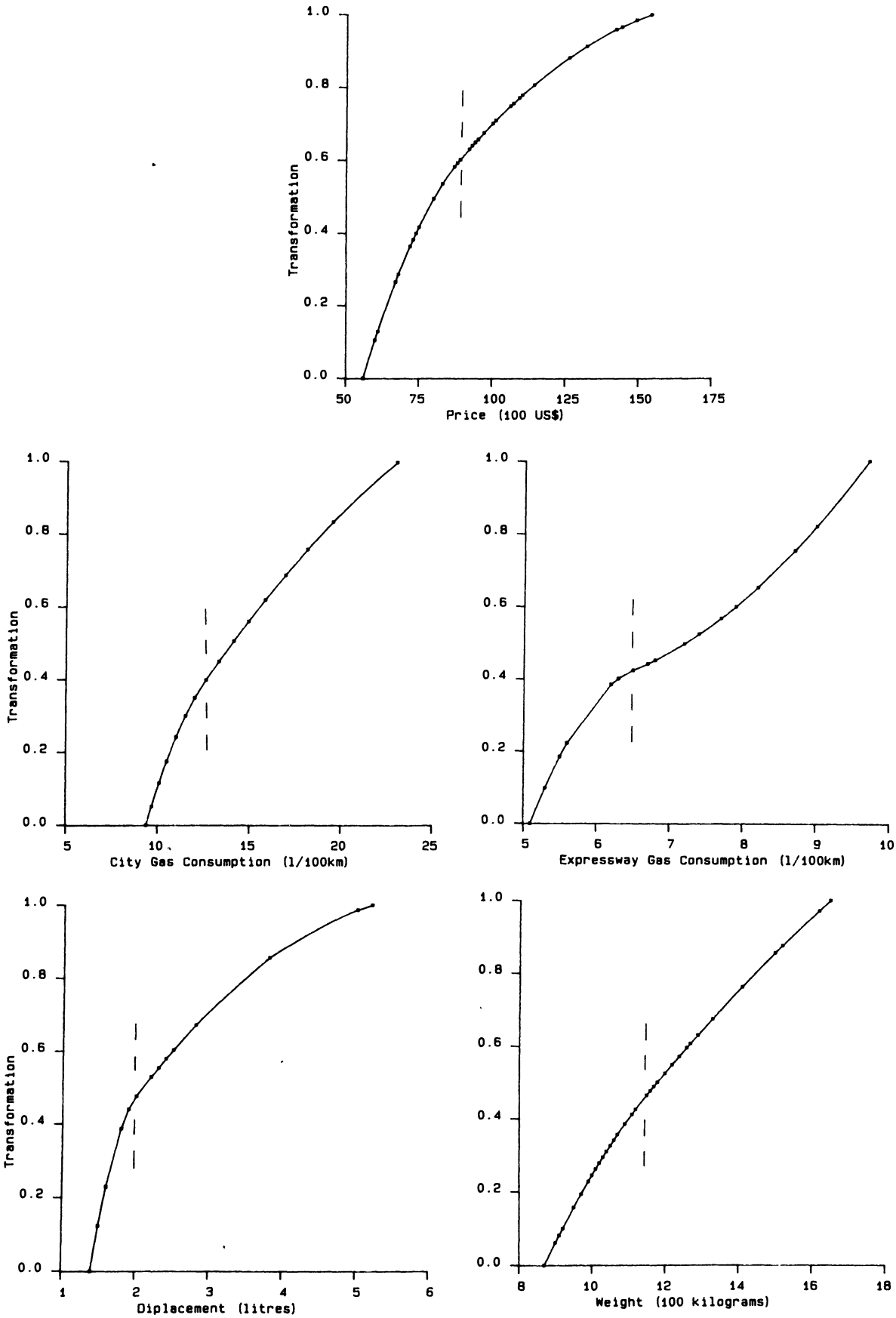

FIG. 8. *Estimated monotone spline transformations resulting from an approximation of five transformed variables by the first two principal components.*

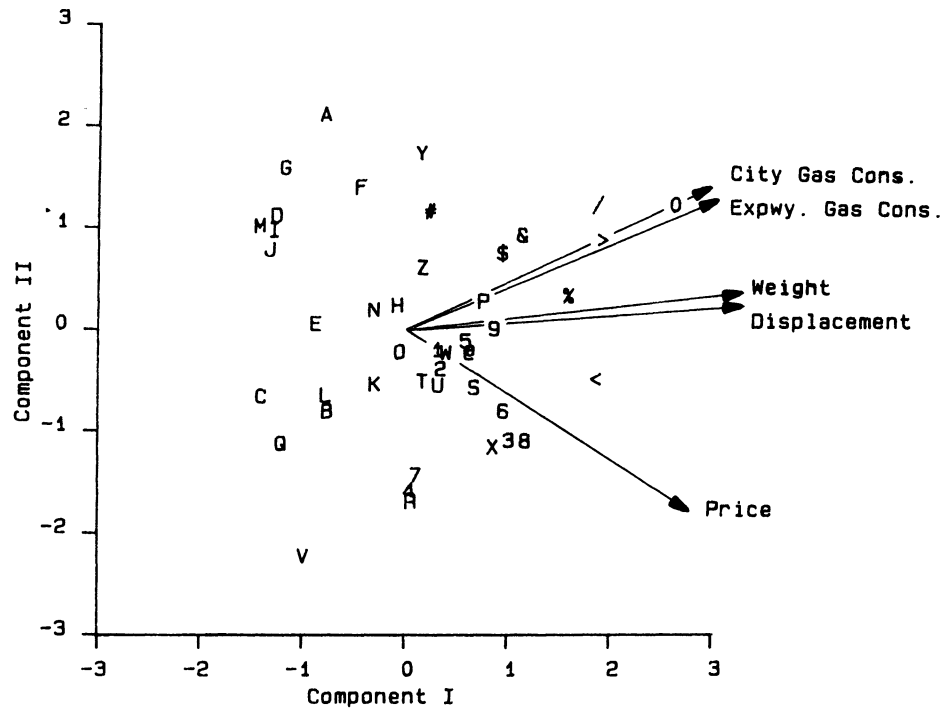

FIG. 9. *Biplot of the scores on the first two principal components of the transformed automobile variables. Vectors indicate the directions corresponding to the transformed variables. The correspondence between symbols and automobiles is in Table* 1.

dose-response, bioassay, event incidence, test item response and other binary data requiring models of this kind. Suppose that for each value x_j, $j = 1, \cdots, p$, N_j responses are observed, of which R_j are "successes," and that the distribution of R_j is binomial with probability $P(x_j)$. It is proposed to model the dependence of probability of success on the covariate values by the monotone spline

$$P(x_j) = \sum_{1}^{n} a_i I_i(x_j). \tag{12}$$

Results can then be compared with more traditional procedures such as the logistic model (Cox, 1970)

$$\ln\{P(x_j)/[1 - P(x_j)]\} = \beta_1 x_j + \beta_0. \tag{13}$$

As a first illustration, consider the problem of summarizing the probability of correctly answering a single test item as a function of some covariate such as the score on the entire test. One would expect that the larger the total test score, the more likely the examinee would be to get any particular item correct. In the extreme cases, examinees with a total score of zero will have zero probability of success on any particular item, and those with perfect scores will have success with probability one on any item. In this problem N_j for a specific test item is the number of examinees whose total test score is x_j, and observation r_j the number at that score level who passed the item.

The resulting item-test *regression curve* for a particular item can reveal a number of useful things about the role of this item in the test (Lord and Novick, 1968). In particular, a test item is termed highly discriminating over a specific score interval if the probability of a correct response rises steeply in that interval. A difficult item shows up as a curve which remains near the zero or chance response level for all total test score values except the very highest.

These features are illustrated in Figure 10, which shows the item-test regression curves for the last five items of a multiple choice test taken by 300 examinees and having 100 items. Each item had four options. The covariance x_j was the total test score, which in principle could range from 0 to 100, but in fact was not less than 24 or greater than 91 among the 300 examinees who took the test. The points displayed in Figure 10 show the frequencies with which the various total score values were obtained.

The five monotone spline curves shown in Figure 10 were estimated by maximum likelihood using procedures described in detail elsewhere (Ramsay, 1988). The interval of transformation in this case was [0, 100], with the restrictions $P(0) = 0$ implying that $a_0 = 0$ and $P(100) = 1$ implying that $\sum a_i = 1$. Three interior knots were positioned at 56, 64 and 73, which are the quartiles of the test score distribution and these are indicated by the vertical dashed lines in the figure. Order 3 splines were

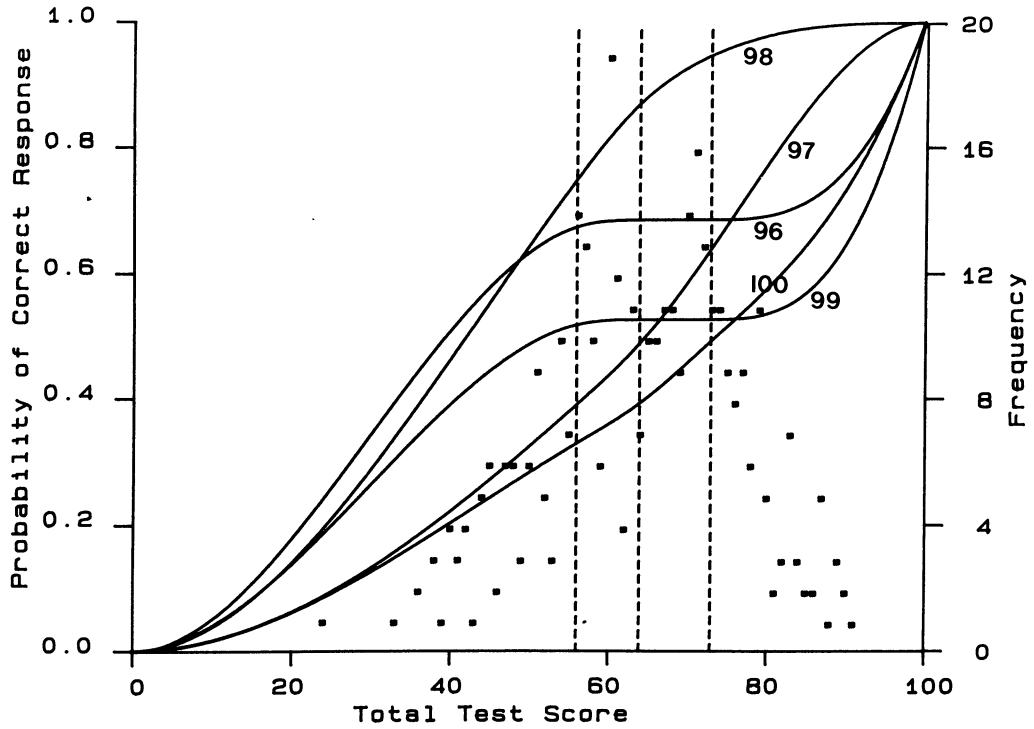

FIG. 10. *Estimated order 3 monotone spline functions showing the relationship between probability of a correct response on five multiple choice test items and total test score. The points indicate the frequencies of total test score values, and the vertical dashed lines indicate the knot value.*

TABLE 2
Incidence of Down's Syndrome in Australia from 1942 to 1957

Age of mother	Total number of births	No. of cases
19 or less	35,555	15
20–24	207,931	128
25–29	253,450	208
30–34	170,970	194
35–39	86,046	297
40–44	24,498	240
45 or over	1,707	37

employed, implying six coefficients determining each curve. However, the two restrictions imposed on each curve leave four mathematically independent parameters to be estimated per curve.

The results display the power of monotone splines to take on shapes difficult to model using standard functions such as the logistic. Item 98 is very easy, with only those in the lowest quartile having any danger of failure. On the other hand, items 97 and 100 are comparatively difficult. The plateaus in the curves for items 96 and 99 show that these items are very nondiscriminating for the majority of students, suggesting that there are two plausible answers and that students choose between them using criteria unrelated to ability as measured by total test score. For example, item 99 reads, "A technique called brainstorming, in group problem-solving: (a) helps generate solutions that prove to be superior, (b) actually causes poorer and fewer solutions to be generated, (c) generates a greater number of solutions than those given by individuals working alone, (d) generates fewer but better solutions than those given by individuals working alone." Alternative (c) was scored as correct, but alternative (a) is also highly plausible. Thus, the shapes of item response curves can provide clues to useful modifications of items.

4.7 Binomial Response: Down's Syndrome Data

The data in Table 2 were collected by Collmann and Stoller (1962) and display the incidence of mongolism as a function of mother's age in Victoria, Australia. Two features of the data invite modeling by monotone spline response functions: the covariate does not entirely have metric properties due to the discretization and lumping of all ages below 20 and of those above 44, and there is a sharp rise in incidence at about age 35 that suggests a compound explanatory process.

Figure 11 displays the results using both monotone splines and logistic regression. In these analyses the categories were given the integer values 1, ···, 7, and the transformation was over [1, 7]. The order 2 spline fit is based on a single knot placed at 4, and thus uses

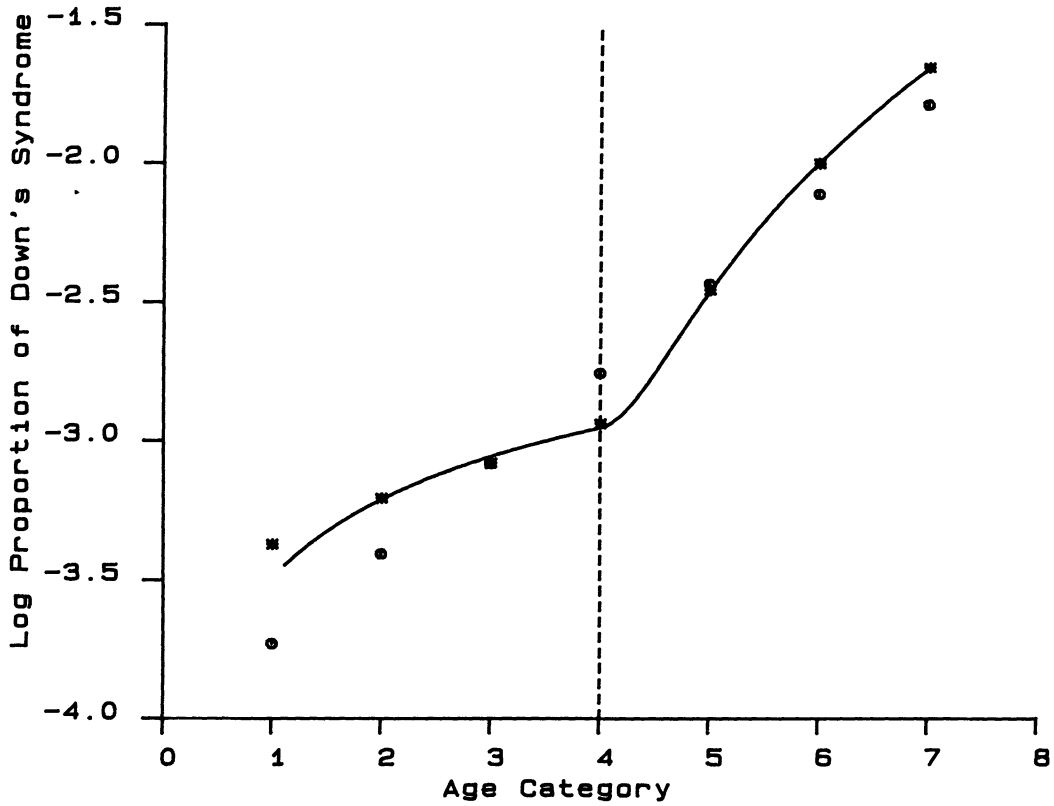

FIG. 11. *Relation between log proportion of births having Down's syndrome and age of mother. Observed points are fit by logistic regression (dashed line) and monotone spline function (solid line). The single interior knot was located at age category 4.*

three degrees of freedom, whereas the log linear model uses two. The spline fit is obviously much superior and thus worth the extra parameter. Its χ^2 value is 2.2 (4 d.f.) as compared to 92.2 (5 d.f.) for the log linear model. Going to the log quadratic model did not improve the fit substantially.

5. SUMMARY

The flexibility of splines and the ease with which they can be constrained to satisfy desirable side conditions such as monotonicity are features of great value in a very wide range of statistical problems, only a few of which are illustrated here. The price paid is small; the usual numerical and computational problems associated with conventional estimation techniques are complicated only by the imposition of simple linear equality and inequality constraints. The resulting increase in program complexity and computation time is not great. Only a very few parameters need be devoted to defining the transformations, so that the increase in standard error of structural parameter estimates is often more than offset by the improvement in fit. In any case, standard techniques can be used to determine whether the additional power of splines is worthwhile. There seems to be a serious case for including spline transformation as a routine tool on the applied statistician's workbench.

APPENDIX

We consider the problem

$$\min_c \{f(c)\}, \quad c \in R^n,$$

$$\text{subject to} \quad c_m \geq 0 \quad \text{and} \quad \sum c_m = 1.$$

A variety of techniques are available for this optimization problem with linear equality and inequality constraints. The approach defined here is an example of the gradient projection method (Fletcher, 1981). This is based on two principles. First, at any iteration v the set $\mathcal{J}_v$ is defined to contain the indices of positive coefficients, termed the active coefficient set. Optimization at this iteration is only with respect to these coefficients. The second principle is the linear transformation of the active coefficients $c_{\mathcal{J}} = \{c_i, i \in \mathcal{J}\}$ by $\tilde{c}_{\mathcal{J}} = A_{\mathcal{J}} c_{\mathcal{J}}$ in order to transform the linearly constrained problem into an unconstrained one. In particular, for the constraint $\sum c_i = 1$, if there are L indices in $\mathcal{J}$, then $A_{\mathcal{J}}$ is any $L - 1$ by L matrix satisfying

$$A_{\mathcal{J}} A'_{\mathcal{J}} = I \quad \text{and} \quad A_{\mathcal{J}} 1 = 0,$$

where 1 is a vector of L 1's. This implies that $P_{\mathcal{J}} = A'_{\mathcal{J}}A_{\mathcal{J}}$ is a projection matrix which projects a vector onto the subspace orthogonal to 1. One way of computing such a matrix $A_{\mathcal{J}}$ is to use the orthonormal polynomial values of degree 1, $\cdots$, $L-1$ associated with the argument values 1, $\cdots$, L.

Let c_v and g_v denote the coefficient and gradient vectors at iteration v, respectively, and ε a convergence criterion. Then the algorithm proceeds as follows:

0. Set $\mathcal{J}$ for c_0 and $\tilde{g} = g$, $\tilde{g}_{\mathcal{J}} = A_{\mathcal{J}}\tilde{g}_{\mathcal{J}}$.
1. For $v = 1$ to $v_{\max}$:
   - 1.1. If $\| \tilde{g} \| < \varepsilon$ then
     - 1.1.1. $\tilde{g} = g$, $\tilde{g}_{\mathcal{J}} = A'_{\mathcal{J}}A_{\mathcal{J}}\tilde{g}_{\mathcal{J}}$;
     - 1.1.2. if $\min_{i \in \mathcal{J}} \{\tilde{g}_i\} = \tilde{g}_{\min} < 0$ then remove index of $\tilde{g}_{\min}$ from $\mathcal{J}$, revise $A_{\mathcal{J}}$ else exit;
   - 1.2. $\tilde{g} = g$, $\tilde{g}_{\mathcal{J}} = A_{\mathcal{J}}\tilde{g}_{\mathcal{J}}$;
   - 1.3. $H = E[D^2 f]$, $\tilde{H} = H$, $\tilde{H}_{\mathcal{J}} = A_{\mathcal{J}}\tilde{H}_{\mathcal{J}}A'_{\mathcal{J}}$;
   - 1.4. $\Delta = 0$, $\tilde{\Delta} = \tilde{H}^{-1}\tilde{g}$, $\Delta_{\mathcal{J}} = A'_{\mathcal{J}}\tilde{\Delta}_{\mathcal{J}}$;
   - 1.5. $\min_{\alpha} \{f(c_v) \mid c_v \geq 0\}$, $c_v = c_{v-1} + \alpha\Delta$;
   - 1.6. revise $\mathcal{J}$ and $A_{\mathcal{J}}$.

ACKNOWLEDGMENTS

This research was supported by Natural Sciences and Engineering Research Council of Canada Grant APA 320. The assistance of M. Abrahamowicz, C. Brewer, Y. Takane and the referees in the preparation of the manuscript was greatly appreciated. I would especially like to thank Ingram Olkin for his encouragement of the writing of this paper.

REFERENCES

AGARWAL, G. G. and STUDDEN, W. J. (1978). Asymptotic design and estimation using linear splines. *Comm. Statist. B—Simulation Comput.* **7** 309–320.

BARLOW, R. E., BARTHOLOMEW, D. J., BREMNER, J. M. and BRUNK, H. D. (1972). *Statistical Inference under Order Restrictions: The Theory and Application of Isotonic Regression.* Wiley, New York.

BOX, G. E. P. and COX, D. R. (1964). An analysis of transformations (with discussion). *J. Roy. Statist. Soc. Ser. B* **26** 211–252.

BREIMAN, L. and FRIEDMAN, J. H. (1985). Estimating optimal transformations for multiple regression and correlation (with discussion). *J. Amer. Statist. Assoc.* **80** 580–618.

CHAMBERS, J. M. (1977). *Computational Methods for Data Analysis.* Wiley, New York.

COLLMAN, R. D. and STOLLER, A. (1962). A survey of mongoloid births in Victoria, Australia, 1942–1957. *Amer. J. Public Health* **57** 813–829.

CONSUMERS UNION INC. (1986). 1986 automobiles. *Consumer Reports.* Consumers Union Inc., Mt. Vernon, N. Y.

COX, D. R. (1970). *The Analysis of Binary Data.* Chapman and Hall, London.

CRAVEN, P. and WAHBA, G. (1979). Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of generalized cross-validation. *Numer. Math.* **31** 377–403.

CURRY, H. B. and SCHOENBERG, I. J. (1966). On Polya frequency functions. IV. The fundamental spline functions and their limits. *J. Analyse Math.* **17** 71–107.

DE BOOR, C. (1978). *A Practical Guide to Splines.* Springer, New York.

ECKART, C. and YOUNG, G. (1936). The approximation of one matrix by another of lower rank. *Psychometrika* **1** 211–218.

FLETCHER, R. (1981). *Practical Methods of Optimization* **2**. Wiley, New York.

HASTIE, T. and TIBSHIRANI, R. (1986). Generalized additive models (with discussion). *Statist. Sci.* **1** 297–318.

HASTIE, T. and TIBSHIRANI, R. (1987). Generalized additive models: Some applications. *J. Amer. Statist. Assoc.* **82** 371–386.

KRUSKAL, J. B. (1965). Analysis of factorial experiments by estimating monotone transformations of the data. *J. Roy. Statist. Soc. Ser. B* **27** 251–263.

LORD, F. M. and NOVICK, M. R. (1968). *Statistical Theory of Mental Test Scores.* Addison-Wesley, Reading, Mass.

PARK, S. H. (1978). Experimental designs for fitting segmented polynomial regression models. *Technometrics* **20** 151–154.

PASSOW, E. (1974). Piecewise monotone spline interpolation. *J. Approx. Theory* **12** 240–241.

PASSOW, E. and ROULIER, J. (1977). Monotone and convex spline interpolation. *SIAM J. Numer. Anal.* **14** 904–909.

RAMSAY, J. O. (1982). Some statistical approaches to multidimensional scaling (with discussion). *J. Roy. Statist. Soc. Ser. A* **145** 285–312.

RAMSAY, J. O. (1988). Binomial regression with monotone splines: A psychometric application. Unpublished paper.

SAS INSTITUTE INC. (1985). *SAS/IML User's Guide, Version 5 Edition.* SAS Institute Inc., Cary, N. C.

SCHUMAKER, L. (1981). *Spline Functions: Basic Theory.* Wiley, New York.

SILVERMAN, B. W. (1985). Some aspects of the spline smoothing approach to non-parametric regression curve fitting (with discussion). *J. Roy. Statist. Soc. Ser. B* **47** 1–52.

SMITH, P. (1979). Splines as a useful and convenient statistical tool. *Amer. Statist.* **33** 57–62.

WAHBA, G. (1978). Improper priors, spline smoothing and the problem of guarding against model errors in regression. *J. Roy. Statist. Soc. Ser. B* **40** 364–372.

WEGMAN, E. J. and WRIGHT, I. W. (1983). Splines in statistics. *J. Amer. Statist. Assoc.* **78** 351–365.

WINSBERG, S. and RAMSAY, J. O. (1980). Monotonic transformations to additivity using splines. *Biometrika* **67** 669–674.

WINSBERG, S. and RAMSAY, J. O. (1981). Analysis of pairwise preference data using integrated B-splines. *Psychometrika* **46** 171–186.

WINSBERG, S. and RAMSAY, J. O. (1983). Monotonic spline transformations for dimension reduction. *Psychometrika* **48** 403–423.

WOLD, S. (1974). Spline functions in data analysis. *Technometrics* **16** 1–11.

WRIGHT, I. and WEGMAN, E. (1980). Isotonic, convex and related splines. *Ann. Statist.* **8** 1023–1035.