REGRESSION WITH OPTIMAL SCALING OF THE DEPENDENT VARIABLE

Jan de Leeuw
Department of Data Theory FSW
University of Leiden

Abstract. There are various ways in which the usual linear model can be generalized to deal with nonlinearities. We discuss two convenient extensions of the linear model, with normally distributed disturbances, that allow for transformations of the dependent and independent variables. The methods are related to other developments in the recent statistical and data analytical literature, and they are applied to an example from physics.

Introduction

In this paper we treat two different nonlinear generalizations of the linear regression model. Both generalizations are comparatively simple, and stay relatively close to the usual linear model. We also compare these generalizations with some related nonlinear data analysis techniques, proposed by Kruskal (1965), De Leeuw, Young, and Takane (1976), and Breiman and Friedman (1985). Because the model-based techniques are somewhat less familiar, and somewhat more complicated, we present them first.

The first model (called model I) supposes that the true, or error-free, response variables satisfy the usual linear model. Thus we are dealing with n unobservable random variables $\underline{\eta}_i$ which satisfy the model

$$\underline{\eta}_i = \Sigma_j x_{ij}\beta_j + \underline{\zeta}_i, \tag{1a}$$

$$\underline{\zeta}_i \text{ are iid } N(0,\sigma^2). \tag{1b}$$

Throughout the paper we use the convention of underlining random variables. The x_i are n given fixed vectors with m elements, the $\underline{\zeta}_i$ are unobservable random variables called disturbances. The unobserved $\underline{\eta}_i$ are connected with the n observable responses $\underline{y}_i$ by the rule

$$\underline{y}_i = g(\underline{\eta}_i). \tag{1c}$$

We suppose that g is at most partially known, and that our prior information about g is in the form $g \in \mathbb{G}$, with $\mathbb{G}$ a known set of functions. If g is completely known, and equal to the identity, then model (1) is the ordinary nonlinear regression model. If g is known, and invertible, then we can compute $\underline{\eta}_i = g^{-1}(\underline{y}_i)$ and apply the ordinary regression model to the $\underline{\eta}_i$. If g belongs to a particular one-parameter family of (continuous and monotone) functions, then (1) defines a Box-Cox model. There have been a great many publications about Box-Cox models, of which we mention Box and Cox (1964, 1982), Hinkley (1975), Hernandez and Johnson (1980), Bickel and Doksum (1981), Carroll and Ruppert (1981), Bunke (1982), Carroll (1982), Doksum and Wong (1982)

In another familiar special case of this model g is an increasing step function. This is sometimes called the discrete normal linear regression model. It is reviewed by Maddala (1983, section 2.13) and by De Leeuw (1984). We can distinguish the case in which the location of the steps is known, and the case in which only the number of steps is known.

In this paper we do not analyze the case in which g is a step function, but we assume that g is differentiable and strictly increasing, and has a differentiable inverse h. In stead of formulating the constraints on g it is more convenient to require $h \in \mathbb{H}$. Clearly

$$\begin{aligned} &\text{prob}[\underline{y}_i < y] = \text{prob}[g(\underline{\eta}_i) < y] = \text{prob}[\underline{\eta}_i < h(y)] = \\ &= \Phi\{(h(y) - \textstyle\sum_j x_{ij}\beta_j))/\sigma\}. \end{aligned} \tag{2}$$

Here Φ is the cumulative standard normal. It follows that L, which is -2 times the log-likelihood, is given by

$$L = \sigma^{-2} \textstyle\sum_i (h(y_i) - \sum_j x_{ij}\beta_j)^2 + n \ln \sigma^2 - 2 \sum_i \ln h'(y_i). \tag{3}$$

The idea is to minimize this loss function over the unknown parameters β and σ^2, but also over all h in $\mathbb{H}$.

Model II is somewhat more simple. It assumes that

$$\underline{y}_i = f(\Sigma_j x_{ij}\beta_j) + \underline{\zeta}_i, \tag{4a}$$

$$\underline{\zeta}_i \text{ are iid } N(0,\sigma^2). \tag{4b}$$

Again f is at most partially known, and $f \in \mathbb{F}$. Now

$$L = \sigma^{-2} \Sigma_i \{y_i - f(\Sigma_j x_{ij}\beta_j)\}^2 + n \ln \sigma^2 . \tag{5}$$

This must be minimized over β, σ^2, and over all $f \in \mathbb{F}$. Model II is closely related to projection pursuit regression (Friedman and Stützle, 1981, Huber, 1985), of which it can be considered to be a special case.

Winsberg and Ramsay (1983) have compared the two models I and II (in a somewhat different context). They call model I the 'model-error-transformation' formulation, and model II the 'model-transformation-error' formulation. According to them model I is preferable when any nonlinear transformation in $\mathbb{G}$ of the dependent variable 'is about as meaningful as any other', while model II seems more appropriate if the response variable 'is uniquely significant in its original form.' (l.c., pag 580-581)

It is of some interest that both projection pursuit regression and the additive regression models of Stone (1985a, 1985b) assume that both x and y are (realizations of) random variables, and estimate the conditional expectation of y given x. Thus they are not linear model techniques in the classical sense.

The models and techniques mentioned above readily admit further nonlinear generalizations of various forms. The data analysis techniques of Kruskal (1965), De Leeuw,

Young, and Takane (1976), and Young, De Leeuw, and Takane (1976) minimize a loss function of the form

$$S = \Sigma_i \{h(y_i) - \Sigma_j t_j(x_{ij})\}^2, \tag{6a}$$

over both $h \in \mathbb{H}$ and $t_j \in \mathbb{T}_j$, subject to some normalization condition such as

$$\Sigma_i h(y_i)^2 = 1. \tag{6b}$$

Thus all predictors or independent variables are transformed optimally as well.

Minimization of (6a) under condition (6b) has aspects in common with both model I and model II, but it cannot be formulated very well in a likelihood form. It is not based on any probabilistic model, and can best be interpreted as a form of data reduction and approximation. In stead of maximizing the likelihood, we actually maximize the multiple correlation coefficient by our transformations. Or, to use a distinction emphasized by Jöreskog and Wold (1982), we maximize predictability and not structural simplification. It is well known that the alternating least squares methods based on (6) do not give consistent estimates of the structural parameters if model I or model II is true, but this is hardly a valid objection to them, because model I and model II are not assumed to be true. We might as well criticize the likelihood methods for model I and model II discussed below, because they do not maximize the multiple correlation. The ACE-method of Breiman and Friedman (1985) does something very similar to minimizing (6), but it works in the theoretical framework in which both x and y are random variables. In the finite sample case ACE does not explicitly optimize a loss function.

We do not pay much attention to these further nonlinear generalizations, because they do not really introduce anything new. In the context of the linear model the nonlinearity in the response variable is the interesting part, the nonlinearity in the predictors is usually much easier to handle.

Algorithms

In this section we discuss methods to minimize the loss functions (3) and (5). As in Winsberg and Ramsay (1980) and Stone (1985a, 1985b) we use splines to define $\mathbb{F}$ and $\mathbb{H}$. More precisely we shall use increasing B-splines on a given knot-sequence (with all knots different). Thus

$$\mathbb{F} = \{f \mid f = \Sigma_r \alpha_r \phi_r\}, \tag{7a}$$

$$\mathbb{H} = \{h \mid h = \Sigma_r \gamma_r \psi_r\}, \tag{7b}$$

where ϕ_r and ψ_r are the B-splines bases for the particular knot-sequences chosen, and where α_r and γ_r are sequences of coefficients chosen in such a way that the splines are increasing. If we work with parabolic splines, we can use the convenient result that the spline is increasing if and only if the sequence of coefficients is increasing (De Boor, 1978, pag. 163). Moreover the parabolic splines are differentiable everywhere, the derivative is a broken-line function (a piecewise linear B-spline) which is consequently continuous. Because of this reason we have used parabolic splines in our examples, but higher orders are also possible, of course.

We can now formulate our minimization problems more precisely. The first one is minimization of

$$L_1 = \sigma^{-2} \Sigma_i (\Sigma_r \gamma_r \psi_r(y_i) - \Sigma_j x_{ij}\beta_j)^2 + n \ln \sigma^2 - 2 \Sigma_i \ln \Sigma_r \gamma_r \psi_r'(y_i), \tag{8}$$

and the second one minimization of

$$L_2 = \sigma^{-2} \Sigma_i \{y_i - \Sigma_r \alpha_r \phi_r(\Sigma_j x_{ij}\beta_j)\}^2 + n \ln \sigma^2 . \tag{9}$$

The principle of algorithm construction that we use throughout is that of alternating the minimization over sets of parameters, while keeping the other parameters fixed at their current values. This principle has worked very well in many related data analysis contexts (Young, 1981, Wold, 1982).

For the minimization of (8) it is clear that in the first substep we can easily minimize over σ^2 and the β_j, while keeping the γ_r fixed at their current values. In the second substep we must minimize over the γ_r, with σ^2 and the β_j fixed. The first substep is merely a linear regression problem with a currently optimally transformed variable. The second substep is also quite simple. We have to remember that $\psi_r(y_i)$ and $\psi_r'(y_i)$ are known constants. Thus, for given β_j and σ^2, the function L_1 is a convex quadratic minus a concave 'log-linear' function of the γ_r, and consequently it is convex. Because of the monotonicity constraints finding the optimal γ_r is a fairly simple convex programming problem, that can be solved by using the rapid manifold suboptization methods outlined by Zangwill (1969, section 8.3). The two substeps are alternated until there is convergence.

Clearly such a procedure is very much like the alternating least squares methods discussed by Young (1981). It alternates a data transformation step with a parameter fitting step. In the context of the discrete normal linear regression model a very similar algorithm was proposed by De Leeuw (1984), who also related it to the familiar EM-algorithm of Dempster, Laird, and Rubin (1977).

One may wonder why a normalization condition such as (6b) is not needed here. This is simply because the minimum of L_1 over σ^2 and the β_j, for fixed γ, can be written, except for some irrelevant constants, as

$$L_{1*} = n \ln \gamma' U' Q U \gamma - 2 \Sigma_i \ln v_i'\gamma. \qquad (10)$$

Here U contains the $\psi(y_i)$ and V, with rows v_i, contains the $\psi'(y_i)$. Moreover Q is the projector $I - X(X'X)^{-1}X'$. Loss function L_{1*} does not change its value if we multiply γ by a positive constant, i.e. it is scale-free. Thus a normalization is not required. Or, to make a comparison with the alternating least squares methods more precise, in the ALS-formulation or the ACE formulation we need a condition such as (6b) to normalize the transformation. In model I the second part of (10) takes care of the normalization and serves a function similar to (6b).

Minimizing (9) seems simpler than minimizing (8), because (9) does not have the term derived from the Jacobian. To minimize (9) it suffices to minimize the simple sum of squares

$$L_{2*} = \Sigma_i \{y_i - \Sigma_r \alpha_r \phi_r(\Sigma_j x_{ij}\beta_j)\}^2. \qquad (11)$$

No normalization is required here either. Minimization over increasing α for fixed β is a monotone regression problem, minimization over β for fixed α is a rather complicated nonlinear regression problem. In our program we use a version of the Levenberg-Marquand method. Compare Schwetlick (1979, ch. 10) or Gill, Murray, and Wright (1981, section 4.7). In many cases our results can be expected to be similar to those of Friedman and Stützle (1981), who do not exactly minimize (9) but perform closely related operations. Their projection pursuit regression algorithm does not require f to be a monotone spline, but merely requires f to be 'smooth'. Smoothness is not imposed as a constraint, however, but in each step of the algorithm the optimal f without constraints is smoothed by an appropriate subroutine.

Finally we indicate how nonlinear generalizations such as in (6) can easily be incorporated. If we have $\Sigma_j t_j(x_{ij})$, with $t_j \in \mathbb{T}_j$, replacing $\Sigma_j x_{ij}\beta_j$ then expanding each t_j in its B-spline form $t_j = \Sigma_r \lambda_{jr}\delta_{jr}$, and thus $\Sigma_j t_j(x_{ij}) = \Sigma_j\Sigma_r \lambda_{jr}\delta_{jr}(x_{ij})$, which is simply linear again. This illustrates that transforming the predictors nonlinearly does not add anything new or spectacular, and need not be modelled separately.

Statistics

We briefly mention the properties of our estimators. Model II is perhaps most convenient in this sense. It is a straightforward nonlinear regression model, and the theory developed by Jennrich (1969) and Wu (1981) applies. By checking the regularity conditions they impose we can conclude that our estimators of the parameters are consistent and asymptotically normal.

Model I has been studied in considerable detail for the case of one-parameter families of transformations. Compare the references given for the Box-Cox model above, especially Hernandez and Johnson (1980), and Bunke (1982). The idea is to use the results of Huber (1967) to prove consistency and asymptotic normality.

It is our feeling that the asymptotics for nonlinear regression models are certainly mathematically interesting, but necessarily somewhat artificial. Perhaps the safest policy is to use the regression versions of Bootstrap and Jackknife to assess stability.

Example

The example we analyze is taken from a very interesting paper by Wilson (1926). He discusses the question if statistical methods in general, and correlational methods such as regression in particular, can help us to discover natural laws. Wilson's point of view is that correlational procedures are not of much help if they are not combined with 'antecedent rationalism', i.e. which prior knowledge about the subject matter. He illustrates his point by a physical example, taken from the work of Willard Gibbs on the equilibrium of heterogeneous substances. Gibbs derived, from theoretical considerations, a formula

connecting density y, pressure x_1, and absolute temperature x_2 of a mixture of gases with convertible components. The formula is of the form

$$h(y) = \ln x_1 + \beta_2/x_2 + \beta_3, \tag{12}$$

with h a known monotonic function of y, and with β_2 and β_3 two constants which must be determined empirically. Gibbs determined the constants from experiments by Cahours and Bineau, and then used the formula to predict the outcomes of 65 new experiments by Neumann. He discusses the deviations he finds in terms of the rational formula (12).

Wilson uses ordinary linear least squares to predict y from x_1 and x_2, and he concludes that the results of this blind approach are quite useless from the point of view of physical theory. It seems to us that his conclusion is a bit pessimistic. In turns out that with the nonlinear techniques we can recover the rational transformations quite nicely. Let us illustrate this with some results.

Ordinary least squares on the raw data gives R^2 equal to .9166. The correlations of temperature, pressure and density with their rational transforms are $r_T = .9960$, $r_P = .9655$, and $r_D = .9751$. If we first use the rational transformations, and then apply linear regression we find $R^2 = .9826$, while of course $r_T = r_P = r_D = 1$. The alternating least squares method with parabolic spline transformations for pressure and density was tried out next. We used cardinal splines with about 10 equally spaced knots covering the range of the variables. For temperature, which assumes only nine discrete values, we used a step-function. This resulted in $R^2 = .9927$, $r_T = .9977$, $r_P = .9975$, and $r_D = .9979$. Other alternating least squares analyses, with step functions of varying degree and varying number of knots, are reported by Gifi (1981, page 292-300). They lead to the conclusion that alternating least squares with B-splines having about ten knots virtually gives us the exact shape of the unknown rational transformations. This is true both for linear and for parabolic splines.

We have also compared the spline analysis with the Breiman-Friedman ACE-method, using two very simple moving average smoothers. If the bandwidth of the smoother is 3, then we find $R^2 = .9973$, $r_T = .9966$, $r_P = .9899$, and $r_D = .9972$. With bandwidth 9 this becomes $R^2 = .9884$, $r_T = .9969$, $r_P = .9954$, and $r_D = .9922$. Obviously the performance of ACE in this example is also very satisfactory, even though it does not optimize an explicit criterion. Although alternating least squares with parabolic monotone B-splines seems to recover the 'true' functions just a little bit better, the differences are hardly noticable.

Finally we used the (much more complicated) maximum likelihood procedure based on model I. We found $R^2 = .9697$, $r_T = .9865$, $r_P = .9901$, and $r_D = .9967$. Although the method is (by definition) better in terms of likelihood, it is inferior in terms of multiple correlation and recovery of the rational functions. We have not computed the model II solutions for this example, because we were specifically interested in the optimal transformations of density.

The figure on the next page specifically compares the solutions for ALS with B-splines, ACE with small and large bandwidth, and model I. The asterisks indicate the transformations found by the technique, the line in the plot is the rational transformation of density. In the ACE-analysis with small bandwidth we see an endpoint effect at the higher end of the scale. This is probably caused by the peculiarities of the running average smoother. The B-spline, large bandwidth ACE, and maximum likelihood methods overestimate the true function somewhat close to the larger values. The B-spline ALS-method and the small bandwidth ACE seem to give the smoothest transformations. Both ML and large bandwidth ACE are more jagged.

It is quite clear that the example we have chosen is very well behaved, and that examples in the life sciences and the social sciences will generally have much more error. It is of considerable interest to apply the various methods we have discussed on a wider range of examples, and to consider their statistical stability as well.

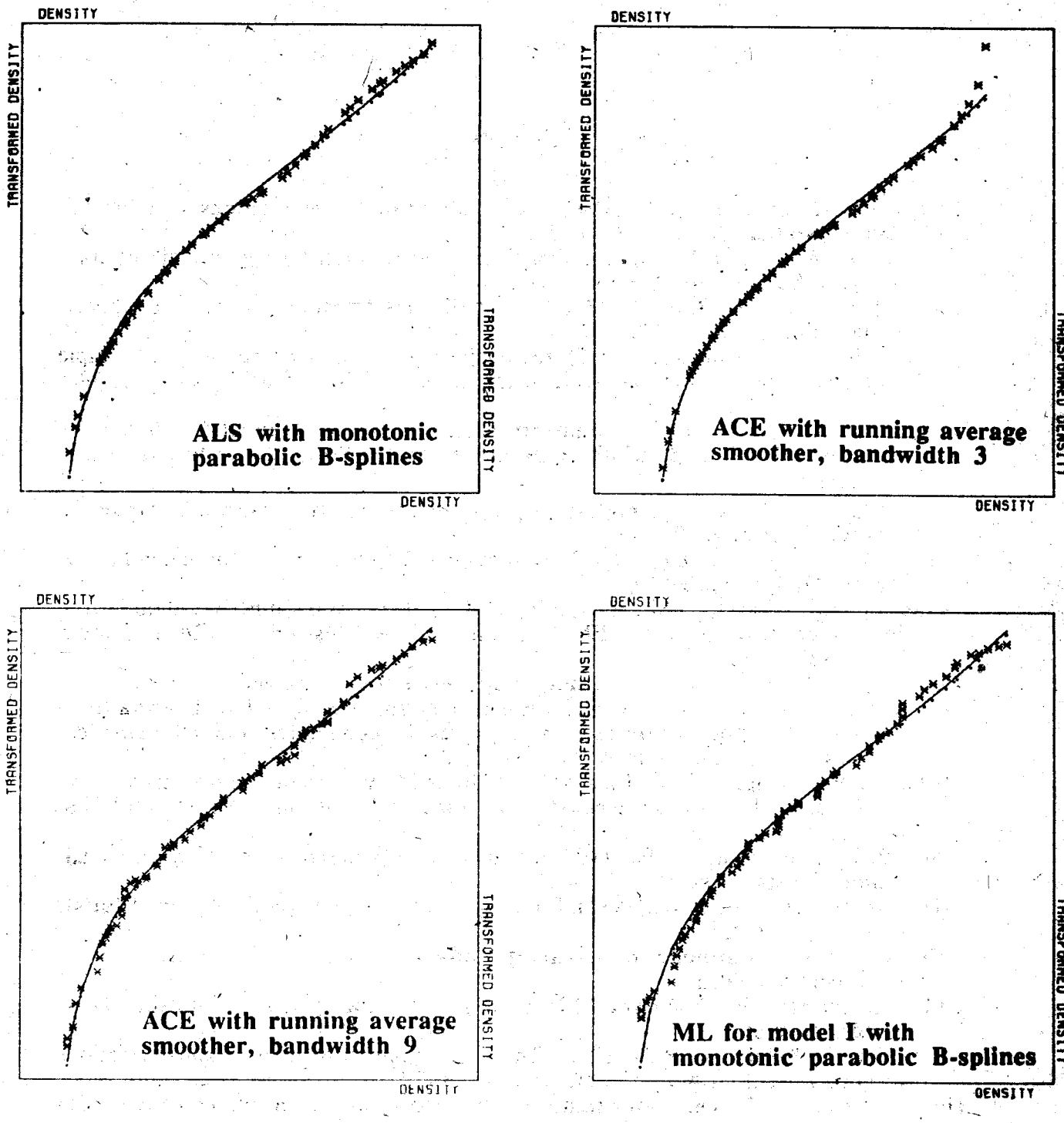

Figure 1:
Comparison of four optimal scalings of density.

References

Bickel, P.J., & Doksum, K.A. (1981). An analysis of transformations revisited. **J. Amer. Statist. Ass.**, 76, 296-311.

Box, G.E.P., & Cox, D.R. (1964). An analysis of transformations (with discussion). **J.Royal Statist. Soc..**, B26, 211-252.

Box, G.E.P., & Cox, D.R. (1982). An analysis of transformation revisited, rebutted.**J. Amer. Statist. Ass.**, 77, 209-210.

Breiman, L., & Friedman, J.H. (1985). Estimating optimal transformations for multiple regression and correlation (with discussion). **J. Amer. Statist. Ass.**, 80, 580-619.

Bunke, O. (1982). The roles of transformations and response functions chosen by maximum likelihood. **Math. Operationsforsch. Statist., Ser.Statistics**,13, 395-404.

Carroll, R.J. (1982). Tests for regression parameters in power transformation models. **Scand. J. Statist.**, 9, 217-222.

Carroll, R.J., & Ruppert, D. (1981). On prediction and the power transformation family. **Biometrika**, 68, 609-615.

Dempster, A.P., Laird, N.M., & Rubin, D.B. (1977). Maximum likelihood from incomplete data using the EM-algorithm (with discussion). **J. Royal Statist. Soc.**, B39, 1-38.

De Boor, C. (1978). **A practical guide to splines.** Berlin, Springer.

De Leeuw, J. (1984). Discrete normal linear regression models. In T.K. Dijkstra (ed.), **Misspecification Analysis**. Lecture notes in economics and mathematical systems, 237. Berlin, Springer.

De Leeuw, J., Young, F.W., & Takane, Y. (1976).Additive structure in qualitative data: an alternating least squares method with optimal scaling features. **Psychometrika**, 41, 471-503.

Doksum, K.A., & Wong, C.-W., (1983). Statistical tests based on transformed data.**J. Amer. Statist. Ass.**, 78, 411-417.

Friedman, J.H., & Stützle, W. (1981). Projection pursuit regression.**J. Amer. Statist. Ass.**, 76, 817-823.

Gifi, A. (1981). **Nonlinear multivariate analysis.** Department of Datatheory, University of Leiden.

Gill, P.E., Murray, W., & Wright, M.H. (1981). **Practical Optimization.**New York, Academic Press.

Hernandez, F., & Johnson, R.A. (1980). The large-sample behaviour of transformations to normality. **J. Amer. Statist. Ass.**, 75, 855-861.

Hinkley, D.V. (1975). On power transformations to symmetry.**Biometrika**, 62. 101-110.

Huber, P.J. (1967). The behaviour of maximum likelihood estimates under nonstandard conditions. **Proc. Fifth Berkeley Symp. Math. Statist. Prob.**, Berkeley, University of California Press.

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

Jennrich, R.I. (1969). Asymptotic properties of non-linear least squares estimators. **Ann. Math. Statist.**, 40, 633-643.

Jöreskog, K.G., & Wold, H. (1982). The ML and PLS techniques for modeling with latent variables: historical and comparative aspects. In K.G. Jöreskog & H. Wold (eds.), **Systems under indirect observation. Causality, structure, prediction.**Amsterdam, North Holland Publishing Co.

Kruskal, J.B. (1965). Analysis of factorial experiments by estimating monotone transformations of the data. **J. Royal Statist. Soc.**, B27,.251-263.

Maddala, G.S. (1983). **Limited-dependent and qualitative variables in econometrics.** Cambridge, Cambridge University Press.

Schwetlick, H. (1979). **Numerische Lösung nichtlinearer Gleichungen.**, Berlin, VEB Deutscher Verlag der Wissenschaften.

Stone, C.J. (1985a). Additive regression and other nonparametric models. **Ann. Statist.**,13, 689-705.

Stone, C.J. (1985b). **The dimensionality reduction principle for generalized additive models.**Technical report 41, Dept. of Statistics, University of California, Berkeley.

Wilson, E.B. (1926). Empiricism and rationalism. **Science**, 64, 47-57.

Winsberg, S., & Ramsay, J.O. (1980). Monotone transformations to additivity using splines. **Biometrika**, 67, 669-674.

Winsberg, S., & Ramsay, J.O. (1983). Monotone spline transformations for dimension reduction. **Psychometrika**, 48, 575-596.

Wold, H. (1982).Soft modeling: the basic design and some extensions. In K.G. Jöreskog & H.Wold (eds.),**Systems under indirect observation. Causality, structure, prediction.** Amsterdam, North Holland Publishing Co.

Wu, C.-F. (1981). Asymptotic theory of nonlinear least squares estimation. **Ann. Statist.**, 9, 501-513.

Young, F.W. (1981). Quantitative analysis of qualitative data. **Psychometrika**, 46, 357-388.

Young, F.W., De Leeuw, J., & Takane, Y. (1976). Regression with qualitative and quantitative variables: an alternating least squares method with optimal scaling features. **Psychometrika**, 40, 505-529.

Zangwill, W.I. (1969). **Nonlinear programming. A unified approach.**Englewood Cliffs, Prentice Hall.