Rejoinder

Charles J. Stone, University of California at Berkeley
Mark Hansen, Bell Laboratories
Charles Kooperberg, University of Washington
Young K. Truong, University of North Carolina at Chapel Hill
Jianhua Huang, University of California at Berkeley

December 30, 1996

We wish to thank our discussants for their interesting and thoughtful comments. With these comments in mind, we begin our rejoinder by clarifying some features of spline-based estimation in the context of an extended linear model.

1 Extended linear modeling revisited

Since our paper was written, the theoretical investigation of extended linear modeling has continued to expand. In particular, in Huang (1996a,b) the rate of convergence results for polynomial splines in Section 2 of our paper have been extended to general approximating spaces including polynomials and trigonometric polynomials, as well as polynomial splines. In Huang and Stone (1996) this more general framework is used to extend the rate of convergence results for hazard regression in Kooperberg, Stone and Truong (1995b) to event history analysis involving repeated events of multiple kinds and time-dependent covariates. In light of this work by Huang and his further work in progress on extended linear modeling, the authors of this paper have invited him to be a coauthor of our rejoinder.

In Section 2 of the paper, we have presented a rather general asymptotic result for maximum likelihood estimates defined over spaces of (possibly smooth) piecewise polynomials. Suppose for simplicity that the separate intervals used to define the different (univariate) polynomial pieces have common length δ, corresponding to equally spaced knots. To achieve the optimal rate of convergence, δ must shrink to zero as the sample size tends to infinity, with the rate being chosen to balance the bias and variance of the resulting estimate. Stone and Koo (1986a,b) followed a similar nonadaptive recipe for knot placement in the context of generalized linear models and logspline density estimation. In practical applications, however, fixed knot splines are rarely adequate, so adaptive knot spline procedures have been developed that alternate between adding knots in regions where the unknown function being estimated exhibits significant features and deleting knots in regions where, subject to noise considerations, this function seems relatively unstructured. A long but still incomplete list of such developments includes Smith (1982); Breiman, Friedman, Olshen and Stone (1984); Friedman and Silverman (1989); Breiman (1991); Friedman (1991); Kooperberg and Stone (1991, 1992); Breiman (1993); Zhang (1994); Kooperberg, Stone and Truong (1995a,c); Hansen, Kooperberg and Sardy (1996); and Kooperberg, Bose and Stone (1997).

Knot placement is not variable selection. Being based on classical model building techniques, adaptive knot spline procedures are readily understood by practitioners. There is, however, a crucial distinction between knot placement and variable selection. Let G be the space of twice continuously differentiable cubic splines on a bounded interval $[a, b]$ having knots (jumps in the third derivative) at the points $x_1, \ldots, x_M$ inside this interval. Aside from its poor numerical properties, the truncated power basis 1, x, x^2, x^3, $(x-x_1)_+^3, \ldots, (x-x_M)_+^3$ is convenient for dealing with the addition and deletion of knots. Specifically, deleting a knot is equivalent to deleting one of the last M columns from the design matrix corresponding to the indicated basis and adding a knot is equivalent to adding the appropriate column to this matrix. It is important, however, to restrict these column operations on the design matrix so that at each step we have made sensible alterations to the underlying function space. Thus we have been careful to treat adaptation in terms of spline spaces, stressing the notion of allowable spaces.

To illustrate what can go wrong, consider knot deletion. By treating this process as a problem in variable selection, we could remove any column of the design matrix corresponding to the truncated power basis. Unfortunately, however, if we attempt to remove the intercept or linear term x in the truncated power basis before the other basis functions have been removed, the resulting space will have potentially poor approximation power in the interval $[a, x_1]$. This is essentially the observation of Härdle et al. in the arguments leading to their Proposition 1. While they are mainly concerned with the intercept being removed, the same effect will hold if the linear term is removed prematurely.

In the top row of Figure 1, we display the results of repeating the simulation described by Härdle et al. 100 times. For each run, we created a data set of 200 observations: $y_i = 0.6(i-0.5)/200 + 0.2 + \epsilon_i$, $i = 1, \ldots, 200$, where $\epsilon_i \sim N(0, 0.25)$. We then performed backward deletion on the columns of the design matrix based on $M = 9$ equally spaced knots, ignoring the implications on the underlying space G. For each data set, we selected the best BIC fit from the sequence of models encountered during deletion, and in the three plots in the top row of Figure 1 we present the top third, middle third, and bottom third of the fits ordered by mean squared error. The gray regions denote the theoretical 95% (pointwise) prediction intervals for the simulated data. For comparison, we repeated this process on the same data sets, this time respecting the structure of the linear spline space; that is, we considered constant, linear, quadratic, and cubic polynomial models together with cubic spline models corresponding to the original collection of knots and its nonempty, proper subsets. For each fit, we also imposed the "tail-linear constraints" $g''(0) = g''(1) = 0$ that are commonly employed in adaptive knot cubic spline procedures such as LOGSPLINE and Friedman's MARS. Because this constraint eliminated two degrees of freedom, we enlarged the space and considered $M = 11$ initial knots. The resulting fits are presented in the middle row of Figure 1, again ordered by mean squared error.

The welter of curves in the top row of Figure 1 exhibit the anticipated behavior when knot deletion is confused with variable selection. As can be readily verified from the subplots in this row, the intercept appears in only 39 of the 100 final fits. The only fit that actually included both the intercept and the linear term x also contained the other two polynomial terms x^2 and x^3 and two additional spline functions from the power basis. The essential problem is that in the

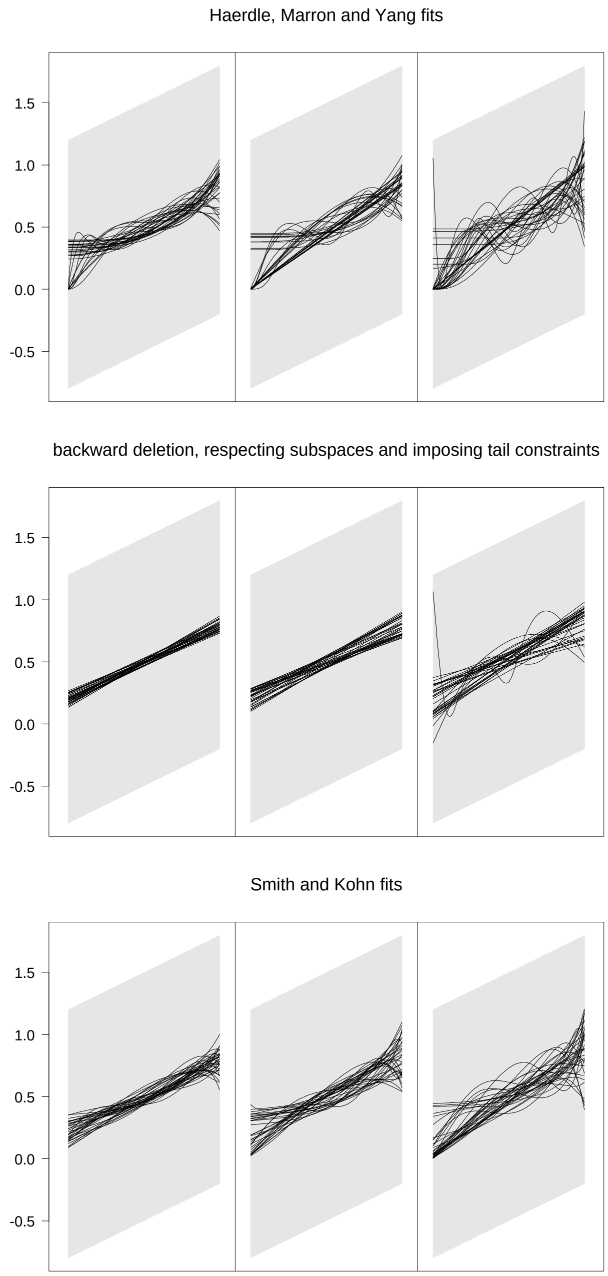

Figure 1: *The simulations of Härdle, Marron and Yang. Three separate approaches to model selection.*

first few deletion steps, the models being considered have very similar residual sums of squares, and a decision that compromises the approximation power of the underlying space can easily be made. This is the essential thrust of the arguments by Härdle et al. By contrast, consider the fits displayed in the second row of Figure 1, only four of which contain terms other than 1 and x. These fits underscore the need to perform stepwise deletion properly, respecting the underlying approximation space. When tail-linear constraints are not imposed, the second row changes only in the high MSE or third subplot, with the extra variation being seen near the boundaries of the interval.

The stepwise procedures suggested in our paper represent computationally efficient approaches to knot placement that have proved effective in practice, but, as the above simulation underscores, it is important to impose the correct heirarchical structure when adding or deleting columns from the design matrix. Not only did Härdle et al. fail to appreciate the distinction between knot placement and variable selection, but this point was also missed in the original paper, Smith and Kohn (1996a), on Bayesian nonparametric regression that they cite. However, this Bayesian approach can be corrected and it provides us with a nice framework to discuss more general schemes for allocating knots efficiently.

Other knot placement procedures. The idea behind the Smith–Kohn technique is to introduce a binary vector $\gamma = (\gamma_1, \ldots, \gamma_{M+4})$ that indexes the columns of the design matrix X corresponding to the truncated power basis: γ_i equals zero or one according as the coefficient β_i of the ith basis function does or does not equal zero. The components of γ are assumed to be a priori independent, with probability one-half of equaling zero. This corresponds to giving all possible subsets of the set of $M+4$ variables the same prior probability. After also specifying prior distributions for $\beta = (\beta_1, \ldots, \beta_{M+4})|(\gamma, \sigma^2)$ and $\sigma^2|\gamma$, Smith and Kohn derive the posterior distribution of γ given the vector of n observations $y = X\beta + \epsilon$, where $\epsilon \sim N(0, \sigma^2 I_n)$. Specifically, if we let β_γ and X_γ denote the coefficient vector and design matrix, respectively, corresponding to a model containing exactly those variables for which γ_i equals one, then by setting

$$p(\beta_\gamma|\gamma, \sigma^2) = N(0, c(X_\gamma^T X_\gamma)^{-1}) \quad \text{and} \quad p(\sigma^2|\gamma) \sim 1/\sigma^2$$

we find that the posterior probability function of γ is given by

$$p(\gamma|y) \sim (1+c)^{-q_\gamma/2}\Big(y^T y - \frac{c}{c+1} y^T X_\gamma (X_\gamma^T X_\gamma)^{-1} X_\gamma^T y\Big)^{-n/2},$$

where $q_\gamma = \sum_i \gamma_i$ is the number of terms in the model and c is a user-specified constant. Smith and Kohn apply the Gibbs sampler to simulate from the posterior distribution of γ and either report the posterior mode of γ or the posterior mean of β. The model has been specified so that the sampling procedure steps through many models with high posterior probability in a computationally efficient manner. After K Gibbs iterations, two alternative approaches are applied to the samples $\gamma^{[k]}$, $k = 1, \ldots K$, to estimate β: (i) $\hat\beta$ is obtained by an OLS fit to those variables included in the model specified by the vector $\gamma^{[k]}$ that maximizes $p(\gamma^{[k]}|y)$; or (ii) the posterior mean $E(\beta|y)$ is estimated by the average value of $E(\beta|y, \gamma^{[k]})$, where the indicated conditional expectation is computed exactly using the fact that $\beta|(y, \gamma)$ has a multivariate t distribution.

Transforming the expression for the posterior probability function of γ into something more familiar to the smoothing community, Foster and George (1996) have found that under certain conditions the value of γ that maximizes this function also minimizes the quantity $\mathrm{RSS}(\gamma) + (1 + c^{-1})\log(c+1)q_\gamma\hat{\sigma}^2$, where $\mathrm{RSS}(\gamma)$ is the residual sum of squares for the model specified by γ, and $\hat{\sigma}^2$ is estimated from the full model with all $M+4$ variables. By selecting c properly, we can perform model selection with respect to Mallow's C_p, AIC, or BIC. Making this connection, we see that the Gibbs sampler of Smith and Kohn is in fact an alternative to our approach of minimizing BIC in a stepwise fashion. Unfortunately, because of the way the vector γ treats all variables as candidates for inclusion or exclusion, we are left with the deficiencies described in the previous section.

As an illustration of the last remark, we applied the Smith–Kohn technique to the same data and basis as in our previous simulations. To this end we downloaded the S-PLUS function `br()` from Statlib, which returns only the posterior mean estimate of β, and applied it with the option `density = 20`, thus specifying twenty data points per knot (or $M = 9$). The resulting fits, shown in the bottom row of Figure 1, exhibit behavior that is similar to (but somewhat less variable than) those in the top row. The problem is that the posterior model probability is roughly the same for the simple model involving 1 and x as it is for a large number of models that include a single polynomial term and one or more of the spline functions from the power basis. In short, because Smith and Kohn search over models that would not be considered when the underlying approximation space G is respected, their final fits average together models with the same types of degeneracies as are found in the top row of Figure 1. Mike Smith (personal communication) has indicated several ways in which their technique could provide for the proper structure of spline spaces by imposing corresponding structure on the prior distribution of γ. Further results on the general model considered by Smith and Kohn and empirical Bayes procedures for estimating c and specifying different prior distributions for γ can be found in Foster and George (1996). Ultimately, we agree with Härdle et al. that, properly implemented, the Smith–Kohn technique is a very attractive method for efficiently identifying good knot locations in the context of least-squares regression.

The performance of this technique, however, is heavily dependent on the number of knots used to define the truncated power basis. An alternative approach followed by Denison, Mallick and Smith (1996) involves defining prior distributions for the number and location of knots as well as the coefficients in a spline expansion. The resulting "automatic Bayesian curve fitting" procedure makes use of reversible jump Markov chain Monte Carlo methods [Green (1995)] to compute the posterior distribution, this time over collections of models having different numbers and positions of knots. At each step in their sampling procedure, one of several possible transitions is chosen at random. These transitions include adding a new knot and either moving or deleting an existing knot, and they can in principle be made efficient through the use of Rao and Wald statistics as discussed in Section 3 of our paper.

Recently, Hansen and Kooperberg (1997) have applied Markov chain Monte Carlo to the problem of identifying promising triangulations for fitting the Triogram models discussed in Section 9. In this investigation, the collection of addition and deletion steps described in Figures 15 and 16 in Section 9.2 has been augmented with two new moves, illustrated here in Figure 2: (i) swapping the diagonal of a convex quadrilateral; (ii) moving a vertex within the union of triangles that contain it. With these additional moves, a much wider variety of triangulations can efficiently be examined than was possible by means of simple stepwise adaptation. For a similar technique applied to piecewise constant modeling, see Nicholls (1996).

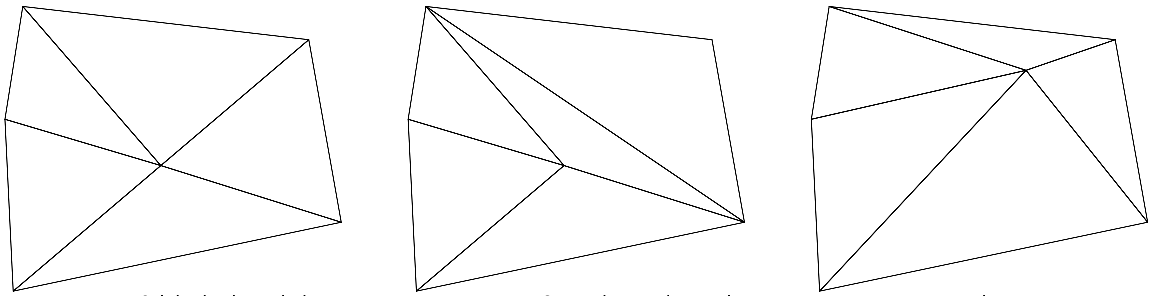

Figure 2: *Additional moves to improve the search for promising triangulations.*

2 Specific responses to our discussants

Härdle, Marron and Yang. Härdle et al. question the use of stepwise knot deletion, indicating that this may be "an inefficient method of adaptation." Their concerns stem primarily from their proof that if we are cavalier in our definition of a minimal space, then the resulting estimates can be inconsistent. As noted in Section 1 of this rejoinder , their proposal to remove the constant function from the minimal allowable space is not natural when this space is thought of as a linear space of functions rather than as the span of columns of the design matrix. Härdle et al. also suggest that these inconsistencies could develop for other "important" basis functions. Like other locally adaptive procedures, the stepwise addition and deletion process can produce spurious effects. The arguments of Härdle et al., however, paint a needlessly pessimistic picture of the problems with our approach. We know, for example, that without adaptive knot selection, spline estimates yield an L_∞ rate of $O[\,n^{1/3}(n^{-1}h^{-1}\log n)^{1/2}+h\,]$; here $h=1/J$ is the distance between uniformly spaced knots. As noted by Härdle et al., this expression is optimized when $h^{-1}=J=O[\,n^{1/9}(\log n)^{-1/3}\,]$, not when $J\sim n^{1/3}$ as they used in their theoretical work. Therefore, a fair comparison between the fixed bandwidth Nadaraya–Watson estimate and the nonadaptive spline estimate with equally spaced knots indicates that the two procedures achieve the same asymptotic performance when measured in the L_∞ norm.

As a practical matter, we depend on a mixture of simulated and real data to guide the development of each new application. For example, when testing the "deletion-only" version of the LOGSPLINE procedure, we experimented with various rules for selecting the number and location of the initial knots. If we apply stepwise deletion starting with too many knots, we encounter an unacceptable number of spurious fits to data drawn from various known densities. If the initial number of knots is too small, on the other hand, the procedure is unable to adapt to the essential features of any but the most regular data sets. Ultimately, we decided that a sensible default for the LOGSPLINE software is to place $O(n^{1/5})$ knots at suitable order statistics. Interestingly enough, the theoretical inconsistency result derived by Härdle et al. fails to hold unless the initial number of knots is much larger than $n^{1/5}$.

To summarize, this discussion reinforces the notion that we must be mindful that our adaptive procedure is defined on an underlying spline space and not on the columns of a particular parameterization of the space. This immediately defines a set of allowable operations on the resulting design matrix that we must not ignore. Their theoretical treatment also adds support to the prevailing notion that these procedures are sensitive to the size of the maximum allowable space and hence that this size must be tuned through extensive simulation.

We agree with these discussants that kernel and local polynomial methods will continue to have an enduring attraction to many statisticians. In particular, it is simple to understand what "vanilla" kernel smoothing does to data, which is one reason it was used as a *rough diagnostic* in Kooperberg and Stone (1991). On the other hand, once we include variable bandwidths [Silverman (1986)], bandwidth selection [Sheather and Jones (1991)], and transformations [Wand, Marron and Ruppert (1991)] in kernel density estimation, it loses its advantages in simplicity, intuitive understanding, and mathematical elegance over logspline density estimation.

These three discussants also question the terminology "polynomial spline" and would prefer that we use the "older" names of "B-spline" or "regression spline." Our use, however, is compatible with the bulk of the literature on splines, which is in the field of numerical approximation. For example, Schumaker (1981) starts out Chapter 9 (which along with the first eight chapters is about univariate splines) with the following:

> In the first eight chapters of this book we have dealt exclusively with polynomial splines. Here and in the following two chapters we develop the theory of similar spaces of generalized splines. We begin this development by studying Tchebycheffian splines, where, as we shall see, almost all of the results for polynomial splines can be carried over.

Tchebycheffian splines include, for example, smooth piecewise exponential functions. It should also be noted that "B-spline" and "polynomial spline" are not synonymous terms. While B-splines are a basis for certain spaces of polynomial splines in a single variable, in certain methodologies dealing with polynomial splines, a different choice of basis is often more convenient (see Section 5.2). Finally, it seems perverse to us to use "regression spline" to refer to a polynomial spline model for a log-density function.

Fan. The approximation power of adaptive knot splines makes them attractive tools for modeling functions with unknown smoothness properties [Devore and Lorentz (1991)], but the usefulness of these tools in a statistical context depends on how well knot placement algorithms can be tuned in the presence of noise. As Fan points out, the general problem of knot placement is very intricate. The use of Rao and Wald statistics in a stepwise fashion is a practical attempt to deal with this intricacy while preserving the flexibility of adaptive knot splines. In order to control the variability of estimation based on this approach, however, we need to limit the maximum number of knots and the minimum spacing between knots.

As Fan suggests, in the context of fitting constrained models, additional ideas may be required to apply our approach and other techniques are sometimes preferable. In particular, Li's approach to dimensionality reduction is an attractive alternative to the use of low-order ANOVA models. Our concern about using the local polynomial method to estimate the effective directions in Li's model is that the resulting methodology lacks the simplicity of Li's original SIR method. The same issue would arise if we were to substitute our approach for the local polynomial method in the procedure described by Fan. However, our approach could be used after the directions have been selected by SIR.

Fan claims that the smoothing spline approach provides a conceptually simple solution to the monotone regression problem, but this is not obvious to us. Ramsay (1988) and Kelly and Rice (1990) provide alternative solutions involving B-splines. When the monotonicity assumption is valid, our approach should yield estimates that are monotone or nearly so; otherwise, the estimates should provide a warning that this assumption may be invalid.

Some constrained models are naturally handled by our approach. For example, in partly linear models without monotonicity constraints, the unknown function has the form $f_1(X_1)+\cdots+f_p(X_p)+Z^T\beta$. Here, to apply our approach we simply need to enforce the additive and linear constraints in choosing the allowable spaces.

The surface averaging technique developed by Fan, Härdle and Mammen (1995) provides a theoretically interesting alternative to our approach in estimating the components in ANOVA models, at least in the regression context. It is surprising that lack of knowledge of f_2 does not cost us anything asymptotically in estimating f_1. However, this result depends on smoothness assumptions on both f_2 and the joint density function of the covariates. Moreover, in estimating an additive component of the regression function at a single point, we need to obtain a local linear estimate of the multidimensional regression function at every "design" point. This necessity makes their procedure computationally intensive. Also, in light of the curse of dimensionality and their smoothness requirement on the joint density function, we are concerned that a very large sample size may be needed for their procedure to be competitive in statistical efficiency with ours. Fan mentions that a special case of their procedure will estimate the average of the regression surface defined with respect to a weight function of product form even if their additive model is misspecified, but we wonder about the motivation for such a weight function when the covariates may be dependent.

Gu. The smoothing spline approach to extended linear modeling developed by Wahba and her school, nicely summarized by Gu in his discussion, is an attractive alternative to our approach. While both approaches emphasize the use of low-order ANOVA models to overcome the curse of dimensionality, the actual implementations are quite different. In the smoothing spline approach, the penalization technique is used after the various main effect and interaction components in the ANOVA model are specified, with other techniques being employed to select these components [Wahba et al. (1995)]. This approach is competitive with ours in applications involving a few covariates, but it is computationally infeasible in applications such as SOLVD and phoneme recognition that involve many covariates.

In his discussion, Gu concentrates on a single smoothing parameter λ as a systematic model index. In the context of fitting low-order ANOVA models, however, it is more natural to use one such parameter per component. When there are many smoothing parameters, their automatic selection becomes computationally challenging. In the context of additive regression and additive generalized regression with polynomial splines, Burman (1990) treated the adaptive selection of a single smoothing parameter (common number of knots). Presumably, similar arguments would apply to other extended linear models.

An interesting modification to the adaptive knot spline approach is to perform a penalized fit using a basis built by stepwise addition. This alternative to stepwise deletion was suggested by Hastie [see Buja, Duffy, Hastie and Tibshirani (1991)] in his discussion of MARS, and it has been implemented by Hastie and Pregibon (1990) in the context of shrinking tree models. Luo and Wahba (1996), cited by Gu, narrow the gap between the smoothing spline and adaptive knot spline approaches both computationally and philosophically by considering the stepwise addition of cubic reproducing kernel functions. A penalized regression is then performed on the smaller basis, reducing the amount of computation considerably and also improving the local adaptability of the smoothing spline approach. In the simple univariate regression problems considered in their paper, for example, they consider adding up to a maximum of 150 such basis functions in problems with samples of size 2048. Interestingly, the authors state that they realized only small improvements in mean squared error when a penalized rather than ordinary least squares fit was performed on the basis built by stepwise addition.

We are pleased to see the many recent and interesting papers on the theoretical properties of the smoothing spline approach that were cited in Gu's discussion. In these papers, however, the underlying assumptions involve conditions on the eigenvalues and eigenfunctions of certain bilinear forms. In general, these conditions are collectively very difficult to verify from more primitive and statistically more natural conditions such as those in Stone (1994), Hansen (1994), and Kooperberg, Stone and Truong (1995b). This difficulty is heightened in the context of applying the smoothing spline approach to the fitting of additive and other unsaturated ANOVA models so as to ameliorate the curse of dimensionality (see Section 2 of our paper). Chen (1991) does deal successfully with such heightened difficulties, but only for regression. Moreover, for mathematical tractability, he was forced to replace the random points $\mathbf{X}_1, \ldots, \mathbf{X}_n$ by deterministic points that form a suitably regular balanced complete factorial design.

Although it is not clear to us exactly what has been rigorously established in the paper by Shen and Hu that Gu cites, we agree that this paper is a promising recent development at least in the sense of suggesting a worthwhile direction for future research.

Hastie and Tibshirani. Hastie and Tibshirani's example illustrates that the use of multiple respose linear regression is at best a cheap substitute for POLYCLASS and that it is much better either to use POLYCLASS directly or to fit a POLYCLASS model with basis functions selected by POLYMARS. Obviously, the situation in their toy example is not as bad when POLYMARS is used to select basis functions. In particular, when we applied POLYMARS to a similar data set, it immediately positioned two knots in each of the two gaps between classes, so that the fitted probabilities almost exactly matched the true probabilities. In this example a POLYMARS model with three basis functions can achieve perfect classification. While POLYCLASS and POLYMARS can achieve both "perfect classification" and "perfect probabilities" in this example, POLYCLASS can do so with fewer basis functions.

The extended linear modeling approach that we have discussed in the current paper always considers nonlinear functions of the predictors. This is precisely to prevent artifacts such as the one presented by Hastie and Tibshirani.

Ignoring computational efficiency, we agree with Hastie and Tibshirani in strongly prefering POLYCLASS to POLYMARS in the selection of basis functions for use in POLYCLASS because the latter approach would tend to add basis functions that are ultimately not needed at the POLYCLASS fitting stage and thus require more basis functions to ensure optimal performance. On the other hand, we doubt that this tendency is so great that the former approach should invariably be used whenever it is computationally feasible despite its heavy computational premium. This issue deserves investigation.

In any case, we ended up using the combination procedure for the phoneme recognition example because POLYCLASS itself is not computationally feasible when applied to such huge data sets with so many classes. Currently, we are studying the application of the stochastic gradient method to the fitting of POLYCLASS models. With this approach, the selection of 400 basis functions using POLYMARS and the fitting of a POLYCLASS model together take less than 20 hours of CPU time (on the SGI workstation at our disposal), which would allow us to experiment with different sets of features and use more basis functions. (In fact, we have experimented with different sets of features and larger models, and we now get test-set error rates of approximately 30%, which is competitive with neural networks.)

Clearly, the density estimation approach proposed by Hastie and Tibshirani will be feasible for larger data sets than POLYCLASS by itself. For example, to apply it to the phoneme recognition data, one would have to estimate $45 \times 63 = 2835$ densities based on sample sizes of $112115/45 \approx 2500$ on the average when no interactions are considered. But this is actually very feasible using LOGSPLINE, since this procedure takes only about 10 seconds of CPU time when applied to the income data ($n = 7125$). If we do not increase the maximum number of basis functions beyond the 25 used for the income data (and this may be very reasonable since presumably not too much detail in the density estimates is needed), the CPU time for LOGSPLINE would be linear in the sample size. Thus the density estimation approach to polychotomous regression would take approximately $2835 \times (2500/7125) \times 10$ seconds ≈ 3 hours of CPU time. Moreover, the computations could easily be divided over a number of workstations.

A related approach would be to fit separate logistic regressions for each of the $K+1$ classes; that is, we would obtain separate estimates of the functions

$$\phi(k|\mathbf{x}) = \log \frac{P(Y = k|\mathbf{X} = \mathbf{x})}{P(Y \neq k|\mathbf{X} = \mathbf{x})}, \qquad 1 \leq k \leq K+1.$$

We can then set

$$\theta(k|\mathbf{x}) = \phi(k|\mathbf{x}) - \phi(K+1|\mathbf{x}) = \log \frac{P(Y = k|\mathbf{X} = \mathbf{x})}{P(Y = K+1|\mathbf{X} = \mathbf{x})}, \qquad 1 \leq k \leq K.$$

As in the density estimation approach, this is a POLYCLASS model. The logistic regression approach could be considered as *partial discriminative* training.

For the logistic regression approach, we can use the POLYCLASS algorithm, which is quite feasible for very large data sets when there are only two classes. Kooperberg, Bose and Stone (1997) establish that the POLYCLASS algorithm requires approximately $O(K^2 P_{\max}^3 n)$ flops independently of the number of predictors. Thus for the logistic regression approach, we would have to carry out $K+1$ times an algorithm that takes $O(P_{\max}^3 n)$ flops, so that we would gain a factor of K. This may not make the approach feasible for the phoneme recognition data, but it will definitely be applicable to much larger data sets than will a direct application of POLYCLASS.

3 Some Publicly Available Software

Not surprisingly, the discussants introduce a number of viable alternative approaches to function estimation in various extended linear model settings. To practitioners, this diversity can represent new views of a given dataset and hence the potential for new insights. Unfortunately, current non-parametric procedures are sufficiently difficult to implement that without usable, publicly available software it is unlikely that a new technique would ever be applied. During our own work to refine the various spline-based estimates presented in our paper, we have found that a language like S/S-PLUS allows us easily to share both code and results, speeding up the entire learning process.

With this in mind, we were naturally led to inquire about the availability of software for comparing our approach with those suggested by the discussants. Both Fan and Härdle et al. mention local polynomial techniques, Härdle et al. suggest that kernel procedures are dependable and understandable, and Gu summarizes how the smoothing spline technology has been applied in many extended linear model settings. We decided to test the practicality of these various approaches by using each one to estimate the density associated with the income data discussed in Section 4 (see Figure 1) of our paper. Locating software that implements each of these approaches was fairly straightforward: A variety of local polynomial models can be fitted using LocFit, a collection of routines with a very polite S/S-PLUS interface being written by Clive Loader [Loader (1996)] and available from `http://cm.bell-labs.com/stat/project/locfit`; kernel density estimation comes with standard S/S-PLUS, and Simon Sheather kindly provided us with a Fortran subroutine for calculating the Sheather–Jones plug-in bandwidth estimate SJPI [Sheather and Jones (1991)]; Chong Gu's RKPK2 is available on the Web and is packaged with a number of density estimation examples [Gu (1993)].

Using an IRIS Challenger with twenty 150 MHZ IP19 processors, we tried each procedure on the mean-scaled data used to generate Figure 1 in Section 4 of our paper; based on the experience of Wand, Marron and Ruppert (1991), we attempted to fit the logarithm of the mean-scaled data as well. The CPU time for each procedure is recorded in Table 1. The nonadaptive LocFit time includes the time to select the global bandwidth via BIC. The available implementations of the adaptive LocFit and smoothing spline procedures required binning the data before the models could be fit. Unfortunately, variable bandwidth selection is not yet implemented for density estimation in LocFit, but by applying a Poisson approximation we can make use of the adaptive features implemented for generalized linear models. The smoothing spline procedure is known to be computationally intensive, requiring $O(n^3)$ operations, so this procedure was not able to deal directly with the income data having $n = 7125$. After mapping the income data to the unit interval, we partitioned the data into 400 equally sized bins and passed these counts to the variable bandwidth LocFit and the smoothing spline programs. The CPU times in seconds listed in Table 1 do not include this preprocessing.

TABLE 1
CPU times in seconds

	original data ($n = 7125$)	binned (400 bins)
LOGSPLINE	9.0	
LocFit	14.5	
Variable Bandwidth LocFit		19.7
Smoothing Splines		194.2
SJPI	341.0	

The results are compared in Figure 3:

(a) The LOGSPLINE fit from Section 4 of our paper, compared with a new version of the routine that uses simulated annealing to optimize knot locations [Hansen and Kooperberg (1997)]. SALSA (Simulated Annealing LogSpline Approximation) requires a few minutes CPU time, so it is still competitive with the slower density estimation routines that are compared in the other examples.

(b) LocFit applied with both a global, fixed bandwidth (chosen by BIC) and a variable bandwidth using a Poisson approximation.

(c) The smoothing spline estimate using both log-transformed and untransformed mean-scaled data.

(d) The SJPI estimate using both log-transformed and untransformed mean-scaled data. [We also tried using the implementation `width.SJ` of SPJI mentioned by Venables and Ripley (1994, page 139), which is available from Statlib, and found that it required similar CPU time. The CPU time for SPJI could undoubtedly be reduced to some extent by binning.]

The smoothing spline estimate, the SJPI estimate, and the fixed bandwith LocFit estimate clearly have problems capturing all the elements of the data, since each of these procedures selects only one (global) smoothing parameter. To truly capture the height of the spike, each estimate would have to be very rough near the larger mode and in the tail. In fact, the heights of the spikes of the smoothing spline and SJPI estimates based on the untransformed data are reduced by more than 20% from that suggested by the adaptive routines, but these estimates still produce riples in other parts of the density. A logarithmic transformation considerably improves both nonadaptive estimates. We believe, however, that the efficacy here of such a simple transformation with one global smoothing parameter is the consequence of a fortuitous “interaction” between the spike and the tail. [Wand, Marron and Ruppert (1991) consider a two parameter family of power transformations for kernel density estimation that yields a very reasonable estimate for the income data.]

Currently, ignoring computing time, we prefer the SALSA estimate. It seems to have the correct height for the spike; based on calculations not reported here, we also feel that the dip near 0.6 has approximately the right depth. Among the faster procedures, LOGSPLINE and the variable bandwidth LocFit (which required binning and using a Poisson approximation) produce very nice estimates.

4 Additional References

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

BUJA, A., DUFFY, D., HASTIE, T. and TIBSHIRANI, R. (1991). Discussion of "Multivariate adaptive regression splines" by J. H. Friedman. *Ann. Statist.* **19** 93–99.

BURMAN, P. (1990). Estimation of generalized additive models. *J. Multivariate Anal.* **32** 230–255.

CHEN, Z. (1991). Interaction spline models and their convergence rates. *Ann. Statist.* **19** 1855–1868.

DENISON, D. G. T, MALLICK, B. K. and SMITH, A. F. M. (1996). Automatic Bayesian curve fitting. Technical Report, Statist. Section of the Math. Dept., Imperial College of Science, Technology and Medicine.

DEVORE, R. and LORENTZ, G. (1991). *Constructive Approximation.* Springer-Verlag, New York.

FOSTER, D. P. and GEORGE, E. I. (1996). Empirical Bayes variable selection. Technical Report, Dept. MSIS, Univ. Texas at Austin.

GREEN, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82** 711–732.

GU, C. (1993). Smoothing spline density estimation: a dimensionless automatic algorithm. *J. Amer. Statist. Assoc.* **88** 495–504.

HANSEN, M. and KOOPERBERG, C. (1997). Strategies for spline adaptation. In preparation.

HASTIE, T. and PREGIBON, D. (1990). Shrinking trees. Technical Report, AT&T Bell Laboratories.

HUANG, J. (1996a). Projection estimation in multiple regression with application to functional ANOVA models. Technical Report 451, Dept. Statist., Univ. California, Berkeley.

HUANG, J. (1996b). Functional ANOVA models for generalized regression. Technical Report 458, Dept. Statist., Univ. California, Berkeley.

HUANG, J. and STONE, C. J. (1996). The L_2 rate of convergence for event history regression with time-dependent covariates. Technical Report 473, Dept. Statist., Univ. California, Berkeley.

KELLEY, C. and RICE, J. (1990). Monotone smoothing with application to dose-response curves and the assessment of synergism. *Biometrics* **46** 1071–1085.

LOADER, C. (1996). Local likelihood density estimation. *Ann. Statist.* To appear.

LUO, Z. and WAHBA, G. (1996). Hybrid adaptive splines. *J. Amer. Statist. Assoc.* To appear.

NICHOLLS, G. (1996). Bayesian image analysis with Markov Chain Monte Carlo and colored continuum triangulation models. *J. Roy. Statist. Soc. Ser. B.* To appear.

RAMSAY, J. O. (1988). Monotone regression splines in action (with discussion). *Statist. Sci.* **3** 425–461.

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

SHEATHER, S.J. and JONES, M.C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *J. Roy. Statist. Soc. Ser. B* **53** 683–690.

SILVERMAN, B. W. (1986). *Density Estimation for Statistics and Data Analysis.* Chapman and Hall, London.

VENABLES, W. N. and RIPLEY, B. D. (1994). *Modern Applied Statistics with S-Plus.* Springer, New York.

WAND, M. P., MARRON, J. S. and RUPPERT, D. (1991). Transformations in density estimation (with discussion). *J. Amer. Statist. Assoc.* **86** 343–361.

ZHANG, H. (1994). Maximal correlation and adaptive splines. *Technometrics* **36** 196–201.

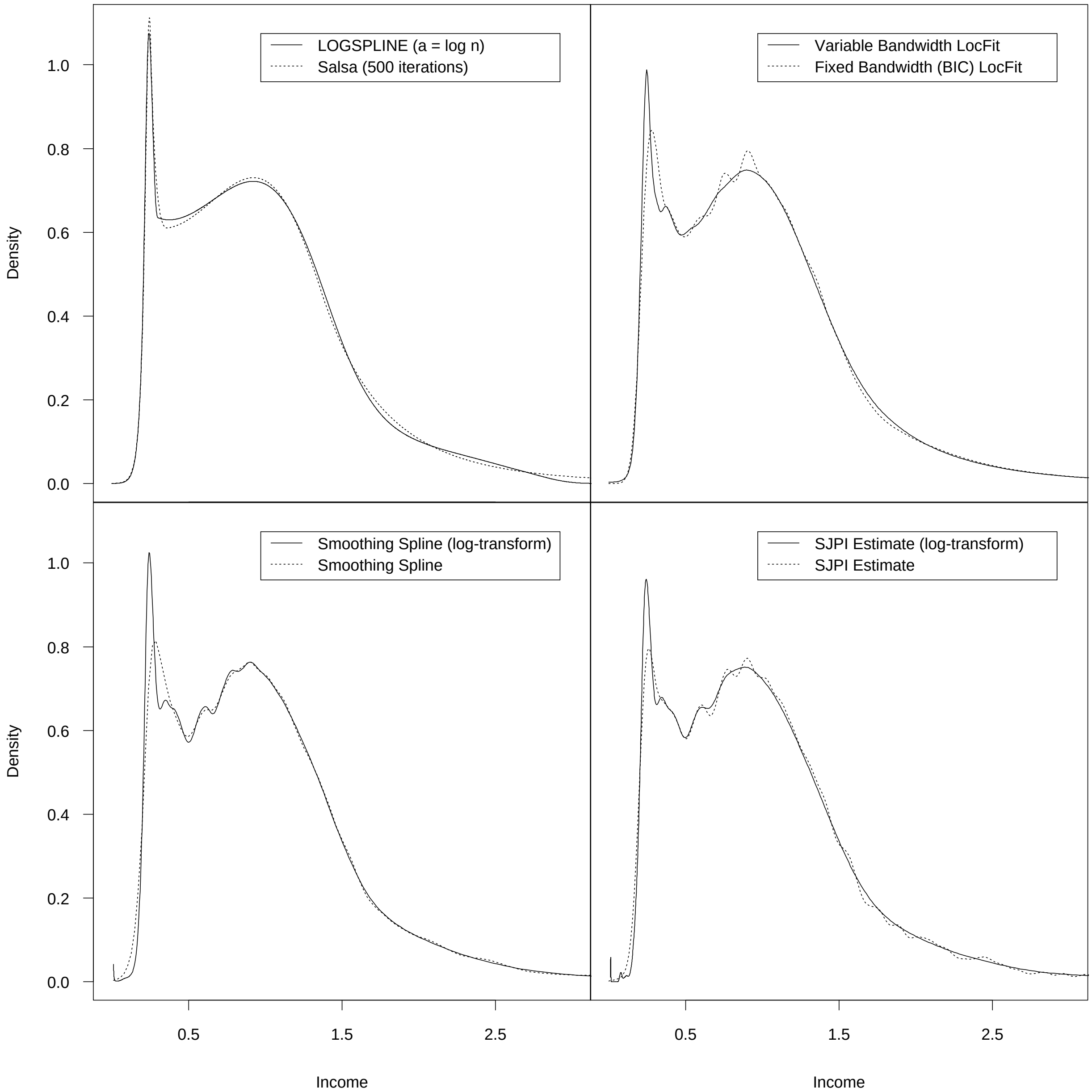

Figure 3: *Comparison of various density estimation methods on the income data.*