Discussion: Polynomial Splines and their Tensor Products in Extended Linear Modeling

Chong Gu

The Annals of Statistics, Vol. 25, No. 4. (Aug., 1997), pp. 1432-1443.

Stable URL:
http://links.jstor.org/sici?sici=0090-5364%28199708%2925%3A4%3C1432%3ADPSATT%3E2.0.CO%3B2-C

The Annals of Statistics is currently published by Institute of Mathematical Statistics.

fitting. Their sampling properties such as bias and variance and asymptotic distributions can be derived. Estimators of their biases and variances can easily be formulated.

The global modeling and the local modeling approach both have strengths in their own domain of applications. Together they provide invaluable tools for nonlinear data analyses and foundational insights.

REFERENCES

EUBANK, R. and SPECKMAN, P. (1993). Confidence bands in nonparametric regression. *J. Amer. Statist. Assoc.* **88** 1287–1301.

FAN, J. (1996). Test of significance based on wavelet thresholding and Neyman's truncation. *J. Amer. Statist. Assoc.* **91** 674–688.

FAN, J. and GIJBELS, I. (1996). *Local Polynomial Modeling and Its Applications*. Chapman and Hall, London.

FAN, J., FARMEN, M. and GIJBELS, I. (1997). Local maximum likelihood estimation and inference. *J. Roy. Statist. Soc. Ser. B*. To appear.

FAN, J., HÄRDLE, W. and MAMMEN, E. (1995). Direct estimation of additive and linear components for high dimensional data. Mimeo 2339, Inst. Statistics, Univ. North Carolina, Chapel Hill.

GREEN, P. J. and SILVERMAN, B. W. (1994). *Nonparametric Regression and Generalized Linear Models: a Roughness Penalty Approach.* Chapman and Hall, London.

HÄRDLE, W. and STOKER, T. M. (1989). Investigating smooth multiple regression by the method of average derivatives. *J. Amer. Statist. Assoc.* **84** 986–995.

HASTIE, T. and TIBSHIRANI, R. (1990). *Generalized Additive Models*. Chapman and Hall, London.

LI, K.-C. (1991). Sliced inverse regression for dimension reduction (with discussion). *J. Amer. Statist. Assoc.* **86** 316–342.

STONE, C. J. (1977). Consistent nonparametric regression. *Ann. Statist.* **5** 595–645.

WONG, W. H. and SHEN, X. (1996). Dimension reduction in regression. Unpublished manuscript.

DEPARTMENT OF STATISTICS
UNIVERSITY OF NORTH CAROLINA
CHAPEL HILL, NORTH CAROLINA 27599-3260
E-MAIL: jfan@stat.unc.edu

DISCUSSION

CHONG GU

Purdue University

Stone, Hansen, Kooperberg and Truong are to be congratulated for their fine article summarizing the adaptive regression spline approach to nonparametric function estimation. With the unified asymptotic theory, the successful applications to a broad spectrum of problems and the availability of user-friendly software, the developments present very impressive achievements that leave many people envious.

Comprehensive and coherent as the authors' treatment is, there still exists an alternative approach that can achieve about as much. This other approach is the penalized likelihood method, pioneered by Good and Gaskins (1971)

and extensively developed over the years by the Wisconsin spline school led by Wahba. My mandate here is to present in a nutshell what has been going on with this alternative line of research, and to provide some comparative comments where fit.

In Sections 2, 3 and 4, I will briefly describe what one can do with the penalized likelihood method. Before that, a bit more discussion of the analysis of variance (ANOVA) decomposition is presented in Section 1, which plays a pivotal role in many of the subsequent developments. Brief comparative comments appear here and there as we move along. Further thoughts on model selection are collected in Section 5.

1. ANOVA decomposition. Let us first look at a generic construction of ANOVA decomposition of functions on arbitrary product domains, one that does not involve the notion of inner product. Despite its extensive application in recent developments of the penalized likelihood method, that to some may seem to tie it with the specific method, the construction does have its independent conceptual identity.

Consider a function $\phi(x_1, \ldots, x_M)$ on a product domain $\prod_{m=1}^M \mathscr{X}_m$. Let A_m be averaging operators acting on arguments x_m that satisfy $A_m^2 = A_m$. An ANOVA decomposition of the function can be defined as

$$\begin{aligned} \phi &= \left\{ \prod_{m=1}^{M} (I - A_m + A_m) \right\} \phi \\ &= \sum_{\mathscr{S} \subseteq \{1,\ldots,M\}} \left\{ \prod_{m \in \mathscr{S}} (I - A_m) \prod_{m \in \mathscr{S}^c} A_m \right\} \phi \\ &= \sum_{\mathscr{S} \subseteq \{1,\ldots,M\}} \phi_{\mathscr{S}}, \end{aligned} \tag{1.1}$$

where $\mathscr{S}$ is the index set of active arguments in a component. $\phi_{\varnothing} = [\prod_{m=1}^M A_m]\phi$ is a constant, $\phi_m = \phi_{\{m\}} = \{(I - A_m)\prod_{l \neq m} A_l\}\phi$ are the x_m main effects, $\phi_{m,l} = \phi_{\{m,l\}} = \{(I - A_m)(I - A_l)\prod_{k \neq m,l} A_k\}\phi$ are the x_m-x_l interactions, and so on. The identifiability of such a decomposition is assured by the side conditions $A_m \phi_{\mathscr{S}} = 0$, $\forall \mathscr{S} \ni m$. The decomposition can also be obtained through recursive hierarchical construction.

For $\mathscr{X}_m = [a, b]$ a real interval, one may choose $A_m\phi = (b-a)^{-1}\int_a^b \phi\, dx_m$ or $A_m\phi = \phi(a)$, anything that satisfies $A_m^2 = A_m$.

For $\mathscr{X}_m = \{1, \ldots, K\}$ a discrete domain, one may choose $A_m\phi = K^{-1}\sum_{x_m=1}^K \phi(x_m)$ or $A_m\phi = \phi(1)$ and so forth.

For $\mathscr{X}_m$ logically univariate but mathematically multivariate such as the geography, one does not need to decompose things further into say the longitude effect and the latitude effect that do not always make practical sense. A possible choice for the averaging operator is $A_m\phi = N^{-1}\sum_{j=1}^N \phi(x_{j,m})$, where $x_{j,m} \in \mathscr{X}_m$ provide a "normalizing mesh" on the domain.

Technically, the decomposition of (1.1) can be constructed explicitly using the tensor product spline technique based on the construction of tensor product reproducing kernel Hilbert spaces, with possibly a mixture of continuous,

discrete, univariate or multivariate marginal domains. Technical details can be found in Aronszajn (1950), Wahba (1990) and other references to follow throughout this discussion. For the cursory exposition in this discussion, the reader only needs to know that the components can be independently attached or detached in the construction. For example, on $\mathscr{X}_1 \times \mathscr{X}_2$, one may well choose to consider only functions of the form $\phi = \phi_1 + \phi_{1,2}$, with the constant and the x_2 main effect eliminated. The decomposition obviously is dependent on the choices of A_m, which are usually based on the ease of interpretation or implementation.

Aside from the asymptotic theory, the ANOVA decomposition does not seem to play much of a role in the authors' treatment. For one thing, the authors do not seem to get an explicit ANOVA decomposition from their fit, which can be useful in the interpretation of the fit. Also, a mechanism to enforce selective exclusion of certain interaction terms would be very useful, if one is not already at work.

2. Penalized likelihood function estimation. The penalized likelihood estimate of a function ϕ can be defined by the minimizer of

$$L(\phi|\text{data}) + (\lambda/2)J(\phi), \tag{2.1}$$

where $L(\phi|\text{data})$ is usually the minus log-likelihood that measures the goodness-of-fit of ϕ to the data, $J(\phi)$ often is a quadratic functional that measures the roughness of ϕ and λ is a tunable smoothing parameter that balances the two conflicting goals of goodness-of-fit and smoothness. The minimizer of (2.1) is sought in a function space $\mathscr{H}$ in which $J(\phi) < \infty$. For ϕ on a product domain, the ANOVA decomposition of (1.1) can be built into the procedure via modular constructions of $\mathscr{H}$ and J using the tensor product spline technique. A penalized likelihood estimate is also called a smoothing spline.

Regression. Consider response data from exponential family distributions $Y|x \sim \exp\{(y\phi(x) - b(\phi(x)))/\sigma^2 + c(y, \sigma^2)\}$, where the dependence of the canonical parameter ϕ on the covariate x is to be estimated and the possibly unknown nuisance dispersion parameter σ^2 is assumed common to all observations. Based on observed pairs (x_i, Y_i), ϕ is estimated by minimizing

$$-\frac{1}{n}\sum_{i=1}^{n}\{Y_i\phi(x_i) - b(\phi(x_i))\} + \frac{\lambda}{2}J(\phi), \tag{2.2}$$

where σ^2 is absorbed into λ.

For ϕ the normal mean, (2.2) reduces to the classical penalized least squares procedure. Other common examples include ϕ the logit for binary data and ϕ the log intensity for Poisson data. The general formulation of (2.2) appeared in the literature no later than O'Sullivan, Yandell and Raynor (1986).

The covariate x resides on a generic domain $\mathscr{X}$, which, in particular, can be a product domain with a mixture of marginals. Unified numerical and theoretical treatments have been developed over the years; further discussion can be found in the next two sections.

Some of the recent developments in regression include diagnostics for aliasing or negligible terms in an ANOVA decomposition [Gu (1992a)], the incorporation of multivariate marginals in an ANOVA decomposition [Gu and Wahba (1993a)], interval estimates for the individual terms of an ANOVA decomposition [Gu and Wahba (1993b); Wahba, et al. (1995); Wang and Wahba (1995)] and the treatment of dependent observations and longitudinal data [Wang (1996a, b)].

Interval estimates seem to be lacking in the authors' treatment of regression even for the function ϕ itself.

Density estimation. Based on independent samples X_i from a probability density $f(x)$ on a domain $\mathscr{X}$, one may write $f = e^{\phi}/\int_{\mathscr{X}} e^{\phi}$, known as a logistic density transform [Leonard (1978)], and estimate ϕ by minimizing

$$-\frac{1}{n}\sum_{i=1}^{n}\left\{\phi(X_i) - \log\int_{\mathscr{X}} e^{\phi}\right\} + \frac{\lambda}{2}J(\phi). \tag{2.3}$$

To make the logistic density transform one-to-one, one may enforce a side condition $A\phi = 0$ with some averaging operator A on $\mathscr{X}$ [Gu and Qiu (1993)], as the authors also do. This can be done by the elimination of the constant term in an ANOVA decomposition, possibly one-way.

When $\mathscr{X}$ is a product domain, selective inclusion/exclusion of the ANOVA terms may be employed to incorporate (conditional) independence structures of the marginals, providing a means to the nonparametric fitting of certain graphical models [cf. Whittaker (1990)]. When $\mathscr{X}$ consists of only a portion of a product domain due to sampling truncation, such as in the protein data example in Section 9.4 of the paper under discussion, the ANOVA structure can be used to enforce pretruncation independence of the marginals, if desired.

Further details can be found in Gu and Qiu (1993) and Gu (1993, 1997). Earlier work on univariate density estimation can be found in Good and Gaskins (1971), Leonard (1978), Silverman (1982), O'Sullivan (1988a) and Cox and O'Sullivan (1990).

Conditional density estimation and polychotomous regression. Now consider a product domain $\mathscr{X} \times \mathscr{Y}$, with both marginals generic. Observing pairs (x_i, Y_i), the objective is to estimate the conditional probability density $f(y|x)$. Write the joint density as

$$f(x, y) = \frac{\exp(\phi_x + \phi_y + \phi_{x,y})}{\int_{\mathscr{X}\times\mathscr{Y}} \exp(\phi_x + \phi_y + \phi_{x,y})}, \tag{2.4}$$

where an ANOVA decomposition is explicitly spelled out and the constant term is trimmed for a one-to-one logistic density transform. The conditional density is easily seen to be $f(y|x) = \exp(\phi_y + \phi_{x,y})/\int_{\mathscr{Y}} \exp(\phi_y + \phi_{x,y})$. This can be written as $f(y|x) = e^{\phi}/\int_{\mathscr{Y}} e^{\phi}$, with side conditions $A_y\phi = 0$, $\forall x$, where A_y is the averaging operator on domain $\mathscr{Y}$ that helps to define the ANOVA decomposition. The side conditions ensure a one-to-one logistic conditional density

transform. The penalized likelihood estimation of $f(y|x)$ is then through the minimization of

$$-\frac{1}{n}\sum_{i=1}^{n}\left\{\phi(x_i, Y_i) - \log\int_{\mathscr{Y}}\exp(\phi(x_i, y))\right\} + \frac{\lambda}{2}J(\phi) \tag{2.5}$$

in a function space with $A_y\phi = 0$.

While the conditional density can be derived from the joint density estimated from random pairs (X_i, Y_i) via (2.3) [with $\mathscr{X}$ in (2.3) replaced by $\mathscr{X} \times \mathscr{Y}$], the use of (2.5) is necessary when observations on the $\mathscr{X}$ domain are considered "fixed," as in a typical regression setting.

Unlike the regression procedure (2.2), which assumes a parametric model on the $\mathscr{Y}$ axis and estimates a parameter ϕ "univariate" in x, the present procedure estimates a "bivariate" function nonparametrically on both axes. The words "univariate" and "bivariate" are put in quotes for x (and y) can itself be multivariate in a hierarchical structure.

For $\mathscr{Y}$ a real interval, the procedure gets conditional mean and conditional quantiles all at once, without ever running into the quantile crossover problem that may trouble methods which target individual quantiles separately.

For $\mathscr{Y}$ discrete, (2.5) naturally reduces to a procedure for nonparametric polychotomous regression. When the class number is 2, the method reduces to exactly what one would get by applying (2.2) to Bernoulli data.

Further details can be found in Gu (1995a).

Density estimation under sampling bias. Distribution data may not always come from the generating density directly, and they may actually come from a variety of sources. The penalized likelihood method provides a convenient way to combine information in the estimation process.

Observing X_i on $\mathscr{X}$ from a density proportional to $f(x)w_i(x)$ with $w_i(x)$ known, the estimation of $f = e^{\phi}/\int_{\mathscr{X}} e^{\phi}$ is simply through the minimization of

$$-\frac{1}{n}\sum_{i=1}^{n}\left\{\phi(X_i) - \log\int_{\mathscr{X}} w_i e^{\phi}\right\} + \frac{\lambda}{2}J(\phi). \tag{2.6}$$

Ordinary samples, length-biased samples, randomly truncated samples or a mixture of these are among those covered by (2.6). Further details can be found in Gu (1992b).

On a product domain $\mathscr{X} \times \mathscr{Y}$, one sometimes collects data from the "wrong" conditional density $f(x|y)$, but is interested in aspects of the other conditional density $f(y|x)$. This is the case with (unmatched) case-control studies in biostatistics and choice-based sampling in econometrics, collectively known as response-based sampling. When information comes only from $f(x|y)$, (2.5) can be used, with x and y interchanged, to estimate ϕ_x and $\phi_{x,y}$, where ϕ_x and $\phi_{x,y}$ are as in (2.4). The odds ratio that interests most is characterized by $\phi_{x,y}$. When supplemental information concerning the joint density is also available, such as in an enriched choice-based sample [cf. Cosslett (1981)], a simple modification of (2.5) combines all relevant information and all three terms ϕ_x, ϕ_y and $\phi_{x,y}$ are estimable. Further details can be found in Gu (1996a).

Hazard estimation. Let T be the lifetime of an item with a survival function $S(t, u) = P(T > t|u)$ and hazard function $e^{\phi(t,u)} = -\partial \log S(t, u)/\partial t$, where u is a covariate. Let Z be the left truncation time and let C be the right censoring time, independent of T and of each other. Observing $(Z_i, X_i, \delta_i, U_i)$, where $X = \min(T, C)$, $\delta = I_{[T \le C]}$ and $Z < X$, one may estimate ϕ by minimizing

$$-\frac{1}{n}\sum_{i=1}^{n}\left\{\delta_i\phi(X_i, U_i) - \int_{Z_i}^{X_i}\exp(\phi(t, U_i))\,dt\right\} + \frac{\lambda}{2}J(\phi). \tag{2.7}$$

With an ANOVA decomposition $\phi = \phi_\varnothing + \phi_t + \phi_u + \phi_{t,u}$, the elimination of $\phi_{t,u}$ characterizes a proportional hazard model, and the inclusion of $\phi_{t,u}$ takes one beyond the proportional hazard model. The covariate domain $\mathscr{U}$ can be a product domain itself, on which hierarchical ANOVA structures can be recursively constructed. The procedure (2.7) estimates all components of ϕ simultaneously via penalized *full* likelihood.

When the covariate domain $\mathscr{U}$ degenerates to a singleton, (2.7) reduces to the log-hazard estimation procedure originally proposed by O'Sullivan (1988a).

Treating $\phi_\varnothing + \phi_t$ as nuisance parameters, penalized *partial* likelihood was used by O'Sullivan (1988b) to estimate ϕ_u in a proportional hazard model and by Zucker and Karr (1990) to estimate $\phi_u + \phi_{t,u}$ of the form $u\beta(t)$, a parametric model with time-varying parameter.

Further details concerning (2.7) can be found in Gu (1994, 1996b, 1997).

Spectral density estimation. Spectral density estimation was a major motivation for the early development of nonparametric function estimation, and the smoothing of a periodogram or log periodogram has been the main tool since day one. Cogburn and Davis (1974) appear to have been the first to use smoothing splines in spectral density estimation.

Based on the first two moments of the log periodogram, Wahba (1980) proposed a certain penalized least squares estimate for the log spectral density, and developed an optimal strategy for the selection of the smoothing parameter. As a refinement of Wahba's (1980) work, Pawitan and O'Sullivan (1994) replaced the least squares by the so-called Whittle log-likelihood of the log periodogram, and developed their version of an optimal smoothing parameter selector. The Whittle log-likelihood is virtually the same log-likelihood the authors use in their LSPEC procedure.

Wahba (1980) and Pawitan and O'Sullivan (1994) both reported extensive empirical studies to justify the optimality of their methods.

3. Asymptotics. Along with the recent methodological developments outlined in Section 2, a unified theme for the calculation of asymptotic convergence rates for penalized likelihood estimates has also emerged. Actually, the asymptotics has played an important role in bringing the method to practice.

Our asymptotic analysis is different from that of the authors. Instead of using the L_2 loss as the universal criterion, we use specific stochastic loss functions customized to specific problem settings. What is common is the an-

alytical approach, together with the routine we follow to customize the loss functions and the regularity conditions.

Take density estimation of (2.3) for example. The loss we target is the symmetrized Kullback–Leibler,

$$\mathrm{SKL}(\phi, \phi_0) = \mu_\phi(\phi - \phi_0) - \mu_{\phi_0}(\phi - \phi_0), \tag{3.1}$$

where $\mu_g(h) = \int_{\mathscr{X}} he^g / \int_{\mathscr{X}} e^g$ and ϕ_0 is the "true" function. A related normed distance is

$$V(\phi - \phi_0) = \mu_{\phi_0}((\phi - \phi_0)^2) - \mu^2_{\phi_0}(\phi - \phi_0). \tag{3.2}$$

Under appropriate conditions, the minimizer $\hat{\phi}$ of (2.3) converges to ϕ_0 at a rate

$$\mathrm{SKL}(\hat{\phi}, \phi_0) \sim V(\hat{\phi} - \phi_0) = O_p(n^{-1}\lambda^{-1/r} + \lambda), \tag{3.3}$$

where r is the decay rate of the eigenvalues of V with respect to J, which characterizes the smoothness of functions in space $\mathscr{H} \subseteq \{f\colon J(f) < \infty\}$ in which $\hat{\phi}$ is sought. In general, the space $\mathscr{H}$ is infinite dimensional and $\hat{\phi}$ is not computable. To bring the method to practice, an adaptive finite-dimensional subspace of $\mathscr{H}$, denoted by $\mathscr{H}_n$, is identified, and the minimizer $\hat{\phi}_n$ of (2.3) in $\mathscr{H}_n$ is shown to have the same convergence rate as given in (3.3). Technical details can be found in Gu and Qiu (1993). Customizations for conditional density estimation and for density estimation under sampling bias can be found in respective references cited in Section 2.

For regression, the loss functions are customized to be

$$\begin{aligned} \mathrm{SKL}(\phi, \phi_0) &= \int_{\mathscr{X}} (\phi - \phi_0)(\mu - \mu_0) f, \\ V(\phi - \phi_0) &= \int_{\mathscr{X}} (\phi - \phi_0)^2 v_0 f, \end{aligned} \tag{3.4}$$

where $\mu(x) = \dot{b}(\phi) = E(Y|x)$, $v(x) = \ddot{b}(\phi) \propto \mathrm{Var}(Y|x)$ and $f(x)$ is the limiting density of x_i. The rate given in (3.3) is established for the minimizer $\hat{\phi}$ of (2.2). Technical details are given in Gu and Qiu (1994). In regression problems, $\hat{\phi}$ is known to be in a finite-dimensional space, so no $\hat{\phi}_n$ is necessary.

For hazard estimation, the loss functions are customized to be

$$\begin{aligned} \mathrm{SKL}(\phi, \phi_0) &= \int_{\mathscr{U}} \int_{\mathscr{T}} (\exp(\phi) - \exp(\phi_0))(\phi - \phi_0) \tilde{S} m, \\ V(\phi - \phi_0) &= \int_{\mathscr{U}} \int_{\mathscr{T}} (\phi - \phi_0)^2 \dot{\exp}(\phi_0) \tilde{S} m, \end{aligned} \tag{3.5}$$

where $\tilde{S}(t, u) = \mathrm{Prob}(Z < t \leq X | u)$ is the at-risk probability and $m(u)$ is the limiting density of U_i. The counting process and martingale structure of survival data are employed to obtain the convergence rate given in (3.3) for the corresponding $\hat{\phi}$ and $\hat{\phi}_n$. Gu (1996b) gives details.

When ϕ_0 resides outside of $\mathscr{H}$, say an additive model is fitted while ϕ_0 does contain interaction, as the authors discuss in Section 2 of the paper, minimal

modification of the analysis yields the same rate for $\hat{\phi}$ and $\hat{\phi}_n$ converging toward the Kullback–Leibler projection of ϕ_0 in $\mathscr{H}$. For density estimation, the projection is the minimizer of the relative Kullback–Leibler, $\mathrm{RKL}(\phi|\phi_0) = \log \int_{\mathscr{X}} e^{\phi} - \mu_{\phi_0}(\phi)$, in $\mathscr{H}$. Further details and customizations in other settings are to be found in Gu (1995b).

References of influence include Silverman (1982), Cox and O'Sullivan (1990), Zucker and Karr (1990) and O'Sullivan (1993).

4. Model selection and computation. We shall now discuss smoothing parameter selection strategies, the single most important factor that determines the practical performance of the estimates. To facilitate the use of the methods in data analysis, software that implements the methods is made available to the public.

The most popular smoothing parameter selection method for penalized least squares regression is Craven and Wahba's (1979) generalized cross-validation (GCV), which was shown by Li (1986) to asymptotically minimize the mean square error, $n^{-1}\sum_{i=1}^{n}(\phi(x_i) - \phi_0(x_i))^2$. Generic algorithms have been developed by Gu, Bates, Chen and Wahba (1989) and Gu and Wahba (1991) for the calculation of automatic fits using GCV selected smoothing parameters. The algorithms are implemented in RKPACK [Gu (1989)], a collection of self-documented Fortran compatible routines (available at `http://www.stat.purdue.edu/~chong/software.html`). Information needed for the construction of interval estimates is also available from RKPACK routines.

Among earlier numerical work are GCVPACK by Bates, Lindstrom, Wahba and Yandell (1987) for models without an ANOVA decomposition (`ftp://ftp.stat.wisc.edu/pub/wahba/software`) and BART by O'Sullivan (1985) for the fast calculation of smoothing splines in one dimension (`http://www.netlib.org/gcv`).

For non-Gaussian regression, an iterative algorithm with a certain adaptation of GCV has been developed and justified in Gu (1990, 1992c). Through semitheoretical analysis and simulation, the GCV adaptation was shown to asymptotically minimize the symmetrized Kullback–Leibler of (3.4). The computation is conveniently conducted by direct calls to existing RKPACK routines in each step of the iteration. Portable code was put together by Wang (1995) in GRKPACK (available at `http://www.stat.purdue.edu/~chong/software.html`). An alternative GCV adaptation was developed by Xiang and Wahba (1996).

The computation of density estimates, including conditional densities and possibly with sampling bias, and that of hazard estimates have much in common numerically. A smoothing parameter selection strategy, designed to minimize the loss functions of (3.1), (3.2), (3.5) or others in their respective settings, was built into certain performance-oriented iteration algorithms by Gu (1993, 1994, 1997) and was shown to demonstrate favorable performance in simulation studies. Fortran compatible routines implementing the algorithms have been put together by the discussant in RKPACK-II (currently available in beta version at `http://www.stat.purdue.edu/~chong/software.html`).

What I like the most in the authors' treatment is their provision of user-friendly S functions, which we also hope to do, but probably not in the immediate future. By working with a selected few basis functions, the authors were able to confine the numerical task to a manageable magnitude even for large data sets. With execution speed $O(n^3)$ and memory requirement $O(n^2)$, however, the algorithms implemented in RKPACK and RKPACK-II are likely to hang S with large data sets. Progress is being made to improve the situation [Wahba and Luo (1995)], and with the chip capacity magnifying and the chip price declining at the current rate, larger and larger problems will soon come within reach.

5. Further thoughts on model selection. Being the single most important issue in function estimation, model selection is probably also the softest spot, because "ad hocness" is often the name of the game. The authors' strategy guided by the Wald statistic, the Rao statistic and AIC or BIC is certainly very appealing, and the examples presented indeed demonstrate adequate performance, yet more can be desired, especially in view of how first intuitions can be grossly misleading in this area [Gu (1992c, 1995c)].

What appears missing in the authors' treatment is a systematic assessment of the performance of the method. Rigorous theoretical justification such as Li's (1986) results on Craven and Wahba's (1979) GCV is probably too much to ask, but *systematic* empirical evidence ought to be supplied to present a real convincing case. With a systematic model indexing, such as that by λ for penalized likelihood estimates or that by bandwidth for kernel estimates, one can (and should) always assess the performance of a model selection strategy, at least in relatively simple settings, by gauging its choices against the best possible fits in simulation studies. Such an assessment is understandably less feasible with a recursive growing/pruning approach that the authors have adopted, for the best possible fits are almost impossible to identify. Until some assessment as convincing yet feasible is developed, however, one may not be fully confident that the method is likely to return a nearly optimal fit. Note that AIC, BIC or GCV are not loss functions themselves and do not define the notion of optimality.

In a promising recent development, Shen and Hu (1994) directly tracked some consistent estimate of the relative Kullback–Leibler during the addition/deletion of knots in an adaptive regression spline approach known as the universal sieve method. Backed by rigorous theoretical justifications, the method is somewhat more excusable of a systematic empirical performance assessment.

The lack of performance assessment may also translate into user's confusion in practice. Take the Buffalo snowfall data for example. Facing a rich selection of four possible recipes with neither a track record for each nor a house recommendation, and with markedly different results at least between three of them, I don't know whether a consumer will choose to roll a die or simply leave the house. With a sample size of $n = 63$, as the authors point out, it is virtually impossible to accurately estimate the number of modes, so the simulation presented in Table 2 of the paper offers little help.

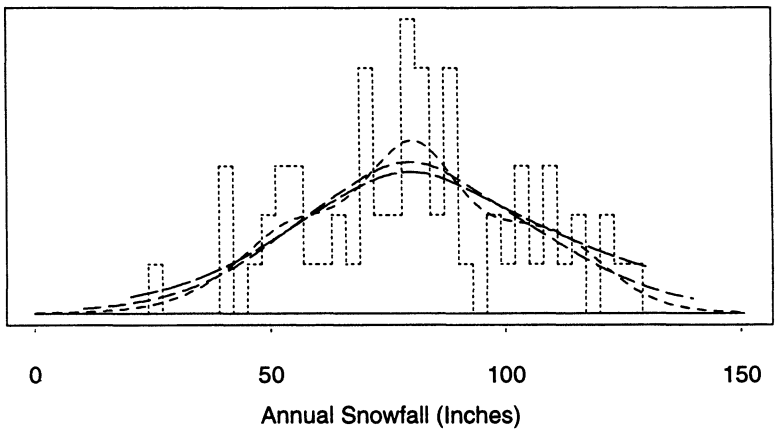

FIG. 1. *The distribution of Buffalo annual snowfalls. The three dashed lines are the automatic estimates under the three domain assumptions indicated by their running lengths. The dotted lines plot a finely binned histogram of the data.*

Reproduced in Figure 1 are three automatic fits to the Buffalo snowfall data using the penalized likelihood method, taken from Gu (1993). The user again has to make some choices, but the choice here is not for different model selection strategies, but for the domain $\mathscr{X}$ on which the log density is assumed to be smooth. The data range from 25.0 to 126.4, and the three fits are supported on [20, 130], [10, 140] and [0, 150], respectively. As the support expands, the model selection strategy tries harder to take away the mass assigned to the empty space at the ends by smoothness, yielding rougher estimates. All three fits are unimodal, however, with the roughest barely showing two shoulders. Note that the relatively smoother fits are not due to a lack of flexibility in the estimation, as the space $\mathscr{H}_n$ (cf. Section 3) has a dimension of 64, but simply by the choice of the model selection procedure. To check out how well the model selection procedure tracks the optimal fits in terms of symmetrized Kullback–Leibler, the reader is referred to the simulation studies documented in Gu (1993).

REFERENCES

ARONSZAJN, N. (1950). Theory of reproducing kernels. *Trans. Amer. Math. Soc.* **68** 337–404.

BATES, D. M., LINDSTROM, M., WAHBA, G. and YANDELL, B. (1987). Gcvpack—routines for generalized cross validation. *Comm. Statist. Simulation Comput.* **16** 263–297.

COGBURN, R. and DAVIS, H. T. (1974). Periodic splines and spectral estimation. *Ann. Statist.* **2** 1108–1126.

COSSLETT, S. R. (1981). Maximum likelihood estimator for choice-based samples. *Econometrica* **49** 1289–1316.

COX, D. D. and O'SULLIVAN, F. (1990). Asymptotic analysis of penalized likelihood and related estimators. *Ann. Statist.* **18** 1676–1695.

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.

GOOD, I. J. and GASKINS, R. A. (1971). Nonparametric roughness penalties for probability densities. *Biometrika* **58** 255–277.

GU, C. (1989). Rkpack and its applications: fitting smoothing spline models. *Proceedings of the Statistical Computing Section* 42–51. Amer. Statist. Assoc., Alexandria, VA.

GU, C. (1990). Adaptive spline smoothing in non Gaussian regression models. *J. Amer. Statist. Assoc.* **85** 801–807.

GU, C. (1992a). Diagnostics for nonparametric regression models with additive terms. *J. Amer. Statist. Assoc.* **87** 1051–1058.

GU, C. (1992b). Smoothing spline density estimation: biased sampling and random truncation. Technical Report 92-03, Dept. Statistics, Purdue Univ.

GU, C. (1992c). Cross validating non Gaussian data. *J. Comput. Graph. Statist.* **1** 169–179.

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

GU, C. (1994). Penalized likelihood hazard estimation: algorithm and examples. In *Statistical Decision Theory and Related Topics 5* (S. S. Gupta and J. O. Berger, eds.) 61–72. Springer, New York.

GU, C. (1995a). Smoothing spline density estimation: conditional distribution. *Statist. Sinica.* **5** 709–726.

GU, C. (1995b). The destination and rates of convergence of penalized likelihood estimate when the model is wrong. Technical Report 255, Dept. Statistics, Univ. Michigan. (Available on-line at `http://www.stat.purdue.edu/~chong/manu.html`.)

GU, C. (1995c). Model indexing and smoothing parameter selection in nonparametric function estimation. Technical Report 93-55 (rev.), Dept. Statistics, Purdue Univ. (Available on-line at `http://www.stat.purdue.edu/~chong/manu.html`.)

GU, C. (1996a). Smoothing spline density estimation: response-based sampling. Technical Report 267, Dept. Statistics, Univ. Michigan. (Available on-line at `http://www.stat.purdue.edu/~chong/manu.html`.)

GU, C. (1996b). Penalized likelihood hazard estimation: a general procedure. *Statist. Sinica.* **6** 861–876.

GU, C. (1997). Structural multivariate function estimation: some automatic density and hazard estimates. *Statist. Sinica.* To appear. (Available on-line at `http://www.stat.purdue.edu/~chong/manu.html`.)

GU, C. and QIU, C. (1993). Smoothing spline density estimation: theory. *Ann. Statist.* **21** 217–234.

GU, C. and QIU, C. (1994). Penalized likelihood regression: a simple asymptotic analysis. *Statist. Sinica* **4** 297–304.

GU, C. and WAHBA, G. (1991). Minimizing GCV/GML scores with multiple smoothing parameters via the Newton method. *SIAM J. Sci. Comput.* **12** 383–398.

GU, C. and WAHBA, G. (1993a). Semiparametric ANOVA with tensor product thin plate splines. *J. Roy. Statist. Soc. Ser. B* **55** 353–368.

GU, C. and WAHBA, G. (1993b). Smoothing spline ANOVA with component-wise Bayesian confidence intervals. *J. Comput. Graph. Statist.* **2** 97–117.

GU, C., BATES, D. M., CHEN, Z. and WAHBA, G. (1989). The computation of GCV functions through Householder tridiagonalization with application to the fitting of interaction spline models. *SIAM J. Matrix Anal. Appl.* **10** 457–480.

LEONARD, T. (1978). Density estimation, stochastic processes and prior information (with discussion). *J. Roy. Statist. Soc. Ser. B* **40** 113–146.

LI, K.-C. (1986). Asymptotic optimality of C_L and generalized cross-validation in the ridge regression with application to spline smoothing. *Ann. Statist.* **14** 1101–1112.

O'SULLIVAN, F. (1985). Discussion of "Some aspects of the spline smoothing approach to nonparametric regression curve fitting" by B. W. Silverman. *J. Roy. Statist. Soc. Ser. B* **47** 39–40.

O'SULLIVAN, F. (1988a). Fast computation of fully automated log-density and log-hazard estimators. *SIAM J. Sci. Comput.* **9** 363–379.

O'SULLIVAN, F. (1988b). Nonparametric estimation of relative risk using splines and cross-validation. *SIAM J. Sci. Comput.* **9** 531–542.

O'SULLIVAN, F. (1993). Nonparametric estimation in the Cox model. *Ann. Statist.* **21** 124–145.

O'SULLIVAN, F., YANDELL, B. and RAYNOR, W. (1986). Automatic smoothing of regression functions in generalized linear models. *J. Amer. Statist. Assoc.* **81** 96–103.

PAWITAN, Y. and O'SULLIVAN, F. (1994). Nonparametric spectral density estimation using penalized Whittle likelihood. *J. Amer. Statist. Assoc.* **89** 600–610.

SHEN, X. and HU, D. (1994). Universal sieve scheme and spline adaptation. Technical report, Dept. Statistics, Ohio State Univ.

SILVERMAN, B. W. (1982). On the estimation of a probability density function by the maximum penalized likelihood method. *Ann. Statist.* **10** 795–810.

WAHBA, G. (1980). Automatic smoothing of the log periodogram. *J. Amer. Statist. Assoc.* **75** 122–132.

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

WAHBA, G. and LUO, Z. (1995). Smoothing spline ANOVA fits for very large, nearly regular data sets, with application to historical global climate data. Technical Report 953, Dept. Statistics, Univ. Wisconsin.

WAHBA, G., WANG, Y., GU, C., KLEIN, R. and KLEIN, B. (1995). Smoothing spline ANOVA for exponential families, with application to the Wisconsin epidemiological study of diabetic retinopathy. *Ann. Statist.* **23** 1865–1895.

WANG, Y. (1995). Grkpack: fitting smoothing spline ANOVA models for exponential families. Technical Report 942, Dept. Statistics, Univ. Wisconsin. (Available on-line at `http://www.sph.umich.edu/~yuedong`.)

WANG, Y. (1996a). Smoothing spline models with correlated random errors. Technical report, Dept. Biostatistics, Univ. Michigan. (Available on-line at `http://www.sph.umich.edu/~yuedong`.)

WANG, Y. (1996b). Mixed-effects smoothing spline ANOVA. *J. Roy. Statist. Soc. Ser. B*. To appear. (Available on-line at `http://www.sph.umich.edu/~yuedong`.)

WANG, Y. and WAHBA, G. (1995). Bootstrap confidence intervals for smoothing splines and their comparison to Bayesian confidence intervals. *J. Statist. Comput. Simulation* **51** 263–279.

WHITTAKER, J. (1990). *Graphical Models in Applied Multivariate Statistics*. Wiley, Chichester.

XIANG, D. and WAHBA, G. (1996). A generalized approximate cross validation for smoothing splines with non-Gaussian date. *Statist. Sinica* **6** 675–692.

ZUCKER, D. M. and KARR, A. F. (1990). Nonparametric survival analysis with time-dependent covariate effects: a penalized partial likelihood approach. *Ann. Statist.* **18** 329–353.

DEPARTMENT OF STATISTICS
1399 MATH-SCI BUILDING
PURDUE UNIVERSITY
WEST LAFAYETTE, INDIANA 47907-1399
E-MAIL: chong@stat.purdue.edu

DISCUSSION

W. HÄRDLE,[1] J. S. MARRON[2] AND L. YANG[1]

Humboldt Universität zu Berlin, University of North Carolina, Chapel Hill and Humboldt Universität zu Berlin

Stone, Hansen, Kooperberg and Truong have written an excellent review of the fine work they have done in making one type of spline modeling useful

[1]Supported by Sonderforschungsbereich 373 "Quantifikation und Simulation Ökonomischer Prozesse" Deutsche Forschungsgemeinschaft.

[2]Supported by NSF Grant DMS-9504414.