Smoothing Spline Anova for Exponential Families, with Application to the Wisconsin Epidemiological Study of Diabetic Retinopathy

Grace Wahba, Yuedong Wang, Chong Gu, Ronald Klein, Barbara Klein

Annals of Statistics, Volume 23, Issue 6 (Dec., 1995), 1865-1895.

Stable URL:
http://links.jstor.org/sici?sici=0090-5364%28199512%2923%3A6%3C1865%3ASSAFEF%3E2.0.CO%3B2-0

The Annals of Statistics
1995, Vol. 23, No. 6, 1865–1895

THE 1994 NEYMAN MEMORIAL LECTURE

SMOOTHING SPLINE ANOVA FOR EXPONENTIAL FAMILIES, WITH APPLICATION TO THE WISCONSIN EPIDEMIOLOGICAL STUDY OF DIABETIC RETINOPATHY[1]

BY GRACE WAHBA,[2] YUEDONG WANG,[3] CHONG GU,[4] RONALD KLEIN[5] AND BARBARA KLEIN[6]

University of Wisconsin–Madison, University of Michigan, Purdue University, University of Wisconsin–Madison and University of Wisconsin–Madison

Let y_i, $i = 1, \ldots, n$, be independent observations with the density of y_i of the form $h(y_i, f_i) = \exp[y_i f_i - b(f_i) + c(y_i)]$, where b and c are given functions and b is twice continuously differentiable and bounded away from 0. Let $f_i = f(t(i))$, where $t = (t_1, \ldots, t_d) \in \mathscr{T}^{(1)} \otimes \cdots \otimes \mathscr{T}^{(d)} = \mathscr{T}$, the $\mathscr{T}^{(\alpha)}$ are measurable spaces of rather general form and f is an unknown function on $\mathscr{T}$ with some assumed "smoothness" properties. Given $\{y_i, t(i),\ i = 1, \ldots, n\}$, it is desired to estimate $f(t)$ for t in some region of interest contained in $\mathscr{T}$. We develop the fitting of smoothing spline ANOVA models to this data of the form $f(t) = C + \Sigma_\alpha f_\alpha(t_\alpha) + \Sigma_{\alpha<\beta} f_{\alpha\beta}(t_\alpha, t_\beta) + \cdots$. The components of the decomposition satisfy side conditions which generalize the usual side conditions for parametric ANOVA. The estimate of f is obtained as the minimizer, in an appropriate function space, of $\mathscr{L}(y, f) + \Sigma_\alpha \lambda_\alpha J_\alpha(f_\alpha) + \Sigma_{\alpha<\beta} \lambda_{\alpha\beta} J_{\alpha\beta}(f_{\alpha\beta}) + \cdots$, where $\mathscr{L}(y, f)$ is the negative log likelihood of $y = (y_1, \ldots, y_n)'$ given f, the $J_\alpha, J_{\alpha\beta}, \ldots$ are quadratic penalty functionals and the ANOVA decomposition is terminated in some manner. There are five major parts required to turn this program into a practical data analysis tool: (1) methods for deciding which terms in the ANOVA decomposition to include (model selection), (2) methods for choosing good values of the smoothing parameters $\lambda_\alpha, \lambda_{\alpha\beta}, \ldots$, (3) methods for making confidence statements concerning the estimate, (4) numerical algorithms for the calculations and, finally, (5) public software. In this paper we carry out this program, relying on earlier work and filling in important gaps. The overall scheme is applied

Received January 1995; revised May 1995.

[1]This work formed the basis for the Neyman Lecture at the 57th Annual Meeting of the Institute of Mathematical Statistics, Chapel Hill, North Carolina, June 23, 1994, presented by Grace Wahba.

[2]Research supported in part by NIH Grant EY09946 and NSF Grant DMS-91-21003.

[3]Research supported in part by NIH Grants EY09446, P60-DK20572 and P30-HD18258.

[4]Research supported by NSF Grant DMS-93-01511.

[5]Research supported by NIH Grant EY03083.

[6]Research supported by NIH Grant EY03083.

AMS 1991 *subject classifications.* Primary 62G07, 92C60, 68T05, 65D07, 65D10, 62A99, 62J07; secondary 41A63, 41A15, 62M30, 65D15, 92H25, 49M15.

Key words and phrases. Smoothing spline ANOVA, nonparametric regression, exponential families, risk factor estimation.

to Bernoulli data from the Wisconsin Epidemiologic Study of Diabetic Retinopathy to model the risk of progression of diabetic retinopathy as a function of glycosylated hemoglobin, duration of diabetes and body mass index. It is believed that the results have wide practical application to the analysis of data from large epidemiologic studies.

1. Introduction. We, along with many others, are interested in building flexible statistical models for prediction (a.k.a. multivariate function estimation). Desirable features of such models include the ability to simultaneously handle continuous variables on various domains, ordered categorical variables and unordered categorical variables. A crucial feature is the availability of a set of methods for adaptively controlling the complexity or degrees of freedom of the model (sometimes called the bias–variance tradeoff) and for comparing different candidate models in the same or related families of models. Other desirable features include the reduction to simple parametric models if the data suggest that such models are adequate, readily interpretable estimates even when several predictor variables are involved, reasonable accuracy statements after the model has been fitted and publicly available software.

Smoothing spline ANOVA (SS-ANOVA) models, which are the subject of this paper, are endowed with all of these features to a greater or lesser extent, although the development of both theory and practice is by no means complete. Briefly, these models represent a function $f(t)$, $t = (t_1, \ldots, t_d)$ of d variables as $f(t) = C + \Sigma_\alpha f_\alpha(t_\alpha) + \Sigma_{\alpha<\beta} f_{\alpha\beta}(t_\alpha, t_\beta) + \cdots$, where the components satisfy side conditions which generalize the usual side conditions for parametric ANOVA to function spaces, and the series is truncated in some manner. Independent observations y_i, $i = 1, \ldots, n$, are assumed to be distributed as $h(y_i, f(t(i)))$ with parameter of interest $f(t(i))$, and $f(\cdot)$ is assumed to be "smooth" in some sense; f is estimated as the minimizer, in an appropriate function space, of $\mathscr{L}(y, f) + \Sigma_\alpha \lambda_\alpha J_\alpha(f_\alpha) + \Sigma_{\alpha<\beta} \lambda_{\alpha\beta} J_{\alpha\beta}(f_{\alpha\beta}) + \cdots$, where $\mathscr{L}(y, f)$ is the negative log likelihood of $(y_1, \ldots, y_n)$ given f, the $J_\alpha, J_{\alpha\beta}, \ldots$ are quadratic penalty functionals and the $\lambda_\alpha, \lambda_{\alpha\beta}, \ldots$ are smoothing parameters to be chosen.

These models have been developed extensively for Gaussian data, and the $d = 1$ special case has been developed for exponential families. Our goal here is to extend this work to the $d > 1$ case for exponential families and to demonstrate its usefulness by analyzing data from the Wisconsin Epidemiologic Study of Diabetic Retinopathy (WESDR). We build an SS-ANOVA model to estimate the risk of progression of diabetic retinopathy, an important cause of blindness, at followup, given values of the predictor variables glycosylated hemoglobin, duration of diabetes and body mass index at baseline, and the response (progression of retinopathy or not) at followup. From the data set analyzed here we have been able to describe interesting relations that were not found using more traditional methods.

CART [Breiman, Friedman, Olshen and Stone (1984)], MARS [Friedman (1991)], projection pursuit [Friedman and Steutzle (1981)] and the π method

[Breiman (1991)] are some of the more popular methods that have been proposed in the statistical literature for multivariate function estimation. Certain supervised machine learning methods, in particular feedforward neural nets and radial basis functions, are also used for this purpose. See Geman, Bienenstock and Doursat (1992), Ripley (1994), Cheng and Titterington (1994), Wahba (1992, 1995) and references therein, for a discussion of relationships between neural nets and statistical nonparametric regression methods. The popular additive models of Hastie and Tibshirani (1990), when fitted with smoothing splines, are a special case of the smoothing spline ANOVA model. The varying-coefficient models [Hastie and Tibshirani (1993)] are a very interesting subfamily. Roosen and Hastie (1994) have taken a further interesting step by combining the additive spline models with projection pursuit. Two basic reference works for smoothing splines are Eubank (1988) and Green and Silverman (1994). Stone (1994) has recently studied theoretical properties of sums of tensor products of polynomial splines used in a regression approach (as opposed to the smoothing approach here) to estimate components of an ANOVA decomposition of a target function. In Stone's regression context numbers of basis functions play the role of smoothing parameters. They are chosen there theoretically, although it would be possible to choose them, as well as the knot locations, adaptively. It is tantalizing to conjecture the circumstances under which Stone's convergence rates could be obtained in the smoothing context employed in the present paper. See Chen (1991), Gu and Qiu (1994) and references cited there.

SS-ANOVA models for Gaussian data are described in some (but not complete) generality in Wahba [(1990), Chapter 10], where references to the previous literature are given. GCV (generalized cross-validation), UBR (unbiased risk) and GML (generalized maximum likelihood) are all discussed there for choosing the smoothing parameters in the Gaussian case. See Craven and Wahba (1979) and Li (1985, 1986) for properties of GCV and UBR estimates. Gu, Bates, Chen and Wahba (1989), Chen, Gu and Wahba (1989), Gu (1992b), Gu and Wahba (1991a, b, 1993a, b), Chen (1991, 1993) and others discuss further various aspects of these models. The code RKPACK [Gu (1989), available from `statlib@lib.stat.cmu.edu`] will fit specified SS-ANOVA models given Gaussian data.

O'Sullivan (1983) and O'Sullivan, Yandell and Raynor (1986), in the $d = 1$ case, proposed penalized log-likelihood estimates with spline penalties for data from general exponential families. Methods for choosing a single smoothing parameter in the $d = 1$ non-Gaussian case have been a matter of lively activity. O'Sullivan, Yandell and Raynor (1986), Green and Yandell (1985), Yandell (1986), Cox and Chang (1990), Wahba (1990), Moody (1991), Liu (1993), Gu (1990, 1992a, c) and Xiang and Wahba (1995) have addressed this issue, all considering methods related to ordinary leaving out one, GCV or UBR adapted to the non-Gaussian case. Wong (1992) has examined the existence of exactly unbiased estimates for the expected Kullback–Leibler information distance as well as predictive mean square error in several non-Gaussian cases. See also Hudson (1978). One can conclude from Wong's

work that there is no exact unbiased risk estimate of the Kullback–Leibler information distance in the Bernoulli case. It is clear, however, that for dense data sets and smooth unknown true functions, good approximations must exist. This may explain why no unique, completely definitive result is available in the Bernoulli case. We will use the approach in Wang (1994), which represents a multiple smoothing parameter extension of Gu's (1992a) extension of the UBR estimate originally obtained for Gaussian data with known variance [Mallows (1973), Craven and Wahba (1979)].

Bayesian "confidence intervals" were proposed for the cross-validated smoothing spline with Gaussian data by Wahba (1983) and their properties were studied by Nychka (1988, 1990). Generalization to the componentwise case in SS-ANOVA appears in Gu and Wahba (1993b). Gu (1992c) discussed their extension to the single smoothing parameter non-Gaussian case. In this work, we develop and employ the componentwise generalization of Gu (1992c) to the non-Gaussian componentwise SS-ANOVA case.

Model selection in the context of non-Gaussian SS-ANOVA has many open questions. The first model selection question might be: will the parametric model which is built into the SS-ANOVA as a special case do as well as a model which contains nonparametric terms? A method for answering this question in the Gaussian case from a hypothesis testing point of view was given by Cox, Koh, Wahba and Yandell (1988) and by Xiang and Wahba (1995) in the Bernoulli case. In the general case where one is comparing one nonparametric model with another, the problem is more complicated. Chen (1993) proposed an approximate hypothesis testing procedure in the general Gaussian case. Gu (1992b) proposed cosine diagnostics as an aid in model selection. The use of componentwise confidence intervals to eliminate terms was suggested in Gu and Wahba (1993b). Of course model selection from a hypothesis testing point of view (i.e., is a simple model correct?) is not the same as model selection from a prediction point of view (i.e., no model is correct, which model is likely to predict best?). In our analysis of the WESDR data we carry out informal model selection procedures including deletion of terms small enough to be of no practical significance, and examination of the componentwise Bayesian "confidence intervals." We will discuss a number of open questions related to model selection in this context from a prediction point of view.

It is clear that the existence of user-friendly software is essential for this and any other sophisticated nonparametric regression method to be useful. A computer code GRKPACK [Wang (1995)], which calls RKPACK as a subroutine, has been developed to carry out the SS-ANOVA analysis for Bernoulli and other non-Gaussian data. We use GRKPACK to carry out the WESDR data analysis.

In Section 2 we review penalized GLIM models with a single smoothing parameter, and then review the SS-ANOVA decomposition of a function and established methods for fitting SS-ANOVA models in the Gaussian case. Although this review is fairly detailed, the presentation of this detail eases greatly the exposition of the generalization of the fitting of these models in

the non-Gaussian case. In Section 3 we describe the extension of SS-ANOVA models to the non-Gaussian exponential family no nuisance parameter case, including a numerical algorithm and methods for choosing the smoothing parameters. In Section 4 Bayesian "confidence intervals" are extended to the componentwise exponential family case and a procedure for computing them is described. In Section 5 we discuss model selection and in Section 6 we carry out the WESDR data analysis. Section 7 discusses some computational considerations and Section 8 gives some conclusions.

2. Penalized GLIM and Gaussian SS-ANOVA models. For simplicity of notation, we will be primarily concerned with data from a member of an exponential family with no nuisance parameter and semiparametric generalizations of the generalized linear models (GLIM's) introduced by Nelder and Wedderburn (1972); see also McCullagh and Nelder (1989). Our method can also deal with over/underdispersion situations; see Wang (1994) for details. We consider random variables y_i with density $h(y_i, f_i)$ of the form

$$h(y_i, f_i) = \exp\left[y_i f_i - b(f_i) + c(y_i)\right], \tag{2.1}$$

where b and c are given functions with b twice continuously differentiable and uniformly bounded away from 0. This includes binomial, Poisson and other random variables as well as normal random variables with variance 1. Letting t be a vector of predictor variables taking values in some fairly arbitrary index set $\mathcal{T}$, we observe pairs $\{y_i, t(i),\ i = 1, \ldots, n\}$, where the y_i are independent observations with distribution $h(y_i, f(t(i)))$. Our goal is to estimate $f(t)$ for t in some region in the space $\mathcal{T}$ of interest. GLIM models represent f as a linear combination of simple parametric functions of the components of t, typically as low degree polynomials. Usually the unknown coefficients are then estimated by minimizing the negative log likelihood, that is, by minimizing

$$\mathscr{L}(y, f) = -\sum_{i=1}^{n} \left[y_i f(t(i)) - b(f(t(i)))\right]. \tag{2.2}$$

O'Sullivan, Yandell and Raynor (1986) replaced the parametric assumption on f by the assumption that f is a member of some "smooth" class of functions of t, and they estimated f as the minimizer, in an appropriate function space [reproducing kernel Hilbert space (RKHS)] of $\mathscr{L}(y, f) + \lambda J(f)$, where J is a roughness penalty. An SS-ANOVA model provides a decomposition of f of the form

$$f(t_1, \ldots, t_d) = \mu + \sum_{\alpha} f_\alpha(t_\alpha) + \sum_{\alpha\beta} f_{\alpha\beta}(t_{\alpha\beta}) + \cdots \tag{2.3}$$

and the penalty $\lambda J(f)$ is replaced by $\sum_\alpha \lambda_\alpha J_\alpha(f_\alpha) + \sum_{\alpha\beta} \lambda_{\alpha\beta} J_{\alpha\beta}(f_{\alpha\beta}) + \cdots$.

The SS-ANOVA model with Gaussian data has the form

$$y_i = f(t_1(i), \ldots, t_d(i)) + \varepsilon_i, \qquad i = 1, \ldots, n, \tag{2.4}$$

where $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)' \sim N(0, \sigma^2 I_{n\times n})$, $t_\alpha \in \mathcal{T}^{(\alpha)}$, where $\mathcal{T}^{(\alpha)}$ is a measurable space, $\alpha = 1, \ldots, d$, $(t_1, \ldots, t_d) = t \in \mathcal{T} = \mathcal{T}^{(1)} \otimes \cdots \otimes \mathcal{T}^{(d)}$ and σ^2 may

be unknown. For f satisfying some measurability conditions a unique ANOVA decomposition of the form (2.3) can always be defined as follows. Let $d\mu_\alpha$ be a probability measure on $\mathscr{T}^{(\alpha)}$ and define the averaging operator $\mathscr{E}_\alpha$ on $\mathscr{T}$ by

$$(\mathscr{E}_\alpha f)(t) = \int_{\mathscr{T}^{(\alpha)}} f(t_1, \ldots, t_d)\, d\mu_\alpha(t_\alpha). \tag{2.5}$$

Then the identity is decomposed as

$$\begin{aligned} I &= \prod_\alpha (\mathscr{E}_\alpha + (I - \mathscr{E}_\alpha)) \\ &= \prod_\alpha \mathscr{E}_\alpha + \sum_\alpha (I - \mathscr{E}_\alpha) \prod_{\beta \neq \alpha} \mathscr{E}_\beta \\ &\quad + \sum_{\alpha < \beta} (I - \mathscr{E}_\alpha)(I - \mathscr{E}_\beta) \prod_{\gamma \neq \alpha, \beta} \mathscr{E}_\gamma + \cdots + \prod_\alpha (I - \mathscr{E}_\alpha). \end{aligned} \tag{2.6}$$

The components of this decomposition generate the ANOVA decomposition of f of the form (2.3) by $C = (\Pi_\alpha \mathscr{E}_\alpha)f$, $f_\alpha = ((I - \mathscr{E}_\alpha)\Pi_{\beta \neq \alpha} \mathscr{E}_\beta)f$, $f_{\alpha\beta} = ((I - \mathscr{E}_\alpha)(I - \mathscr{E}_\beta)\Pi_{\gamma \neq \alpha, \beta} \mathscr{E}_\gamma)f$ and so forth. Efron and Stein (1981) discuss this kind of ANOVA decomposition in a different context. Further details in the RKHS context may be found in Gu and Wahba (1993a, b).

The idea behind SS-ANOVA is to construct an RKHS $\mathscr{H}$ of functions on $\mathscr{T}$ so that the components of the SS-ANOVA decomposition represent an orthogonal decomposition of f in $\mathscr{H}$. Then RKHS methods can be used to explicitly impose smoothness penalties of the form $\Sigma_\alpha \lambda_\alpha J_\alpha(f_\alpha) + \Sigma_{\alpha\beta} \lambda_{\alpha\beta} J_{\alpha\beta}(f_{\alpha\beta}) + \cdots$, where, however, the series will be truncated at some point. This is done as follows. Let $\mathscr{H}^{(\alpha)}$ be an RKHS of functions on $\mathscr{T}^{(\alpha)}$ with $\int_{\mathscr{T}^{(\alpha)}} f_\alpha(t_\alpha)\, d\mu_\alpha = 0$ for $f_\alpha(t_\alpha) \in \mathscr{H}^{(\alpha)}$ and let $[1^{(\alpha)}]$ be the one-dimensional space of constant functions on $\mathscr{T}^{(\alpha)}$. Construct $\mathscr{H}$ as

$$\begin{aligned} \mathscr{H} &= \prod_{\alpha=1}^{d} (\{[1^{(\alpha)}]\} \oplus \{\mathscr{H}^{(\alpha)}\}) \\ &= [1] \oplus \sum_\alpha \mathscr{H}^{(\alpha)} \oplus \sum_{\alpha < \beta} [\mathscr{H}^{(\alpha)} \otimes \mathscr{H}^{(\beta)}] \oplus \cdots, \end{aligned} \tag{2.7}$$

where $[1]$ denotes the constant functions on $\mathscr{T}$. With some abuse of notation, factors of the form $[1^{(\alpha)}]$ are omitted whenever they multiply a term of a different form. Thus $\mathscr{H}^{(\alpha)}$ is shorthand for $[1^{(1)}] \otimes \cdots \otimes [1^{(\alpha-1)}] \otimes \mathscr{H}^{(\alpha)} \otimes [1^{(\alpha+1)}] \otimes \cdots \otimes [1^{(d)}]$ (which is a subspace of $\mathscr{H}$). The components of the ANOVA decomposition are now in mutually orthogonal subspaces of $\mathscr{H}$. Note that the components will depend on the measure $d\mu_\alpha$, and these should be chosen in a specific application so that the fitted mean, main effects, two factor interactions and so forth have reasonable interpretations.

Next, $\mathscr{H}^{(\alpha)}$ is decomposed into a parametric part and a smooth part, by letting $\mathscr{H}^{(\alpha)} = \mathscr{H}_\pi^{(\alpha)} \oplus \mathscr{H}_s^{(\alpha)}$, where $\mathscr{H}_\pi^{(\alpha)}$ is finite dimensional (the "parametric" part) and $\mathscr{H}_s^{(\alpha)}$ (the "smooth" part) is the orthocomplement of $\mathscr{H}_\pi^{(\alpha)}$ in $\mathscr{H}^{(\alpha)}$. Elements of $\mathscr{H}_\pi^{(\alpha)}$ are not penalized through the device of letting $J_\alpha(f_\alpha) = \|P_s^{(\alpha)} f_\alpha\|^2$, where $P_s^{(\alpha)}$ is the orthogonal projector onto $\mathscr{H}_s^{(\alpha)}$; $[\mathscr{H}^{(\alpha)} \otimes$

$\mathscr{H}^{(\beta)}]$ is now a direct sum of four orthogonal subspaces: $[\mathscr{H}^{(\alpha)} \otimes \mathscr{H}^{(\beta)}] = [\mathscr{H}_\pi^{(\alpha)} \otimes \mathscr{H}_\pi^{(\beta)}] \oplus [\mathscr{H}_\pi^{(\alpha)} \otimes \mathscr{H}_s^{(\beta)}] \oplus [\mathscr{H}_s^{(\alpha)} \otimes \mathscr{H}_\pi^{(\beta)}] \oplus [\mathscr{H}_s^{(\alpha)} \otimes \mathscr{H}_s^{(\beta)}]$. By convention the elements of the finite-dimensional space $[\mathscr{H}_\pi^{(\alpha)} \otimes \mathscr{H}_\pi^{(\beta)}]$ will not be penalized. Continuing this way results in an orthogonal decomposition of $\mathscr{H}$ into sums of products of unpenalized finite-dimensional subspaces, plus main effects "smooth" subspaces, plus two-factor interaction spaces of the form parametric $\otimes$ smooth $[\mathscr{H}_\pi^{(\alpha)} \otimes \mathscr{H}_s^{(\beta)}]$, smooth $\otimes$ parametric $[\mathscr{H}_s^{(\alpha)} \otimes \mathscr{H}_\pi^{(\beta)}]$ and smooth $\otimes$ smooth $[\mathscr{H}_s^{(\alpha)} \otimes \mathscr{H}_s^{(\beta)}]$ and similarly for the three and higher factor subspaces.

Now suppose that we have selected the model $\mathscr{M}$; that is, we have decided which subspaces will be included. Collect all of the included unpenalized subspaces into a subspace, call it $\mathscr{H}^0$, of dimension M, and relabel the other subspaces as $\mathscr{H}^\beta$, $\beta = 1, 2, \ldots, p$; $\mathscr{H}^\beta$ may stand for a subspace $\mathscr{H}_s^{(\alpha)}$, or for one of the three subspaces in the decomposition of $[\mathscr{H}^{(\alpha)} \otimes \mathscr{H}^{(\beta)}]$ which contains at least one "smooth" component, or for a higher order subspace with at least one "smooth" component. Collecting these subspaces as $\mathscr{M} = \mathscr{H}^0 \oplus \sum_\beta \mathscr{H}^\beta$, the estimation problem in the Gaussian case becomes: find f in $\mathscr{M} = \mathscr{H}^0 \oplus \sum_\beta \mathscr{H}^\beta$ to minimize

$$\frac{1}{n}\sum_{i=1}^{n}\left(y_i - f(t(i))\right)^2 + \lambda \sum_{\beta=1}^{p} \theta_\beta^{-1}\|P^\beta f\|^2, \tag{2.8}$$

where P^β is the orthogonal projector in $\mathscr{M}$ into $\mathscr{H}^\beta$. The overparameterization $\lambda\theta_\beta^{-1} = \lambda_\beta$ is convenient for both expository and computational purposes [see Gu (1989) and Gu and Wahba (1991b)] and is accounted for in RKPACK. The minimizer $f_{\lambda,\theta}$ of (2.8) is known to have a representation [Wahba (1990), Chapter 10] in terms of a basis $\{\phi_\nu\}$ for $\mathscr{H}^0$ and the reproducing kernels (RK's) $\{R_\beta(s,t)\}$ for the $\mathscr{H}^\beta$. Letting

$$Q_\theta(s,t) = \sum_{\beta=1}^{p} \theta_\beta R_\beta(s,t), \tag{2.9}$$

it is

$$f_{\lambda,\theta}(t) = \sum_{\nu=1}^{M} d_\nu \phi_\nu(t) + \sum_{i=1}^{n} c_i Q_\theta(t(i),t) = \phi(t)'d + \xi(t)'c, \tag{2.10}$$

where

$$\phi'(t) = (\phi_1(t), \ldots, \phi_M(t)),$$
$$\xi'(t) = (Q_\theta(t(1),t), \ldots, Q_\theta(t(n),t)).$$

$c_{n\times 1}$ and $d_{M\times 1}$ are vectors of coefficients which satisfy

$$\begin{aligned}(Q_\theta + n\lambda I)c + Sd &= y,\\ S'c &= 0,\end{aligned} \tag{2.11}$$

where here and below we are letting Q_θ be the $n \times n$ matrix with ijth entry $Q_\theta(t(i), t(j))$ and S be the $n \times M$ matrix with $i\nu$th entry $\phi_\nu(t(i))$. This system will have a unique solution for any $\lambda > 0$ provided S is of full column

rank, which we will always assume. This condition on S is equivalent to the uniqueness of least squares regression onto span$\{\phi_\nu\}$. Since the RK of a tensor product space is the product of the RK's of the component spaces, the computation of the R_β's is straightforward. For example, the RK corresponding to the subspace $\mathscr{H}_\pi^{(\alpha)} \otimes \mathscr{H}_s^{(\beta)}$ is (in an obvious notation) $R_{\mathscr{H}_\pi^{(\alpha)}}(s_\alpha, t_\alpha) R_{\mathscr{H}_s^{(\beta)}}(s_\beta, t_\beta)$. Of course any positive-definite function may in principle play the role of a reproducing kernel here. Special properties of RK's related to splines are noted in Wahba (1990). Conditionally positive-definite functions as occur in thin plate splines [Wahba and Wendelberger (1980)] can be accommodated; see Gu and Wahba (1993a) and references cited therein. Examples on the sphere can be found in Wahba (1981, 1982) and Weber and Talkner (1993); examples on a discrete index set, as might occur in large contingency tables, can be found in Gu and Wahba (1991a). It is not hard to modify reproducing kernels so that a given particular set of functions plays the role of a spanning set for $\mathscr{H}_\pi^{(\alpha)}$; see Wahba [(1978), Section 3]. Arbitrary functions including functions containing breaks and jumps and indicator functions may be added to $\mathscr{H}^0$; see Shiau, Wahba and Johnson (1986), Wahba (1990), Wahba, Gu, Wang and Chappell (1995).

Assuming the model (2.4), the smoothing parameters λ, θ may be chosen by generalized cross-validation (GCV) (σ^2 unknown) or unbiased risk (UBR) (σ^2 known). The GCV and UBR estimates are the minimizers of V and U, respectively, given by

$$V(\lambda, \theta) = \frac{1/n \|(I - A(\lambda, \theta))y\|^2}{[(1/n)\mathrm{tr}(I - A(\lambda, \theta))]^2} \tag{2.12}$$

and

$$U(\lambda, \theta) = \frac{1}{n}\|(I - A(\lambda, \theta))y\|^2 + \frac{2}{n}\sigma^2 \operatorname{tr} A(\lambda, \theta), \tag{2.13}$$

where $A(\lambda, \theta)$ satisfies

$$(f_{\lambda,\theta}(t(1)), \ldots, f_{\lambda,\theta}(t(n)))' = A(\lambda, \theta)y. \tag{2.14}$$

The properties of GCV and UBR estimates in the Gaussian case are well known; see Wahba (1990) and the references cited therein, especially Li (1985). Loosely speaking, under appropriate assumptions they provide good estimates of the λ, θ which minimize $\sum_{i=1}^n (f_{\lambda,\theta}(t(i)) - f(t(i)))^2$. The code RKPACK [Gu (1989)] may be used to compute the GCV and UBR estimates of the $\lambda\theta_\beta^{-1}$, along with $f_{\lambda,\theta}$ and the components of $f_{\lambda,\theta}$ in the ANOVA decomposition. Of course to estimate $\lambda\theta_\beta^{-1}$ the component matrices with i, jth entry $R_\beta(t(i), t(j))$ must be "sufficiently distinguishable." One way to quantify this would be to examine the Fisher information matrix for the θ_β based on the associated Bayes model [Wahba (1978)] and (4.1) below.

3. SS-ANOVA for general exponential families. The generalization of an ANOVA estimate for Gaussian data to the general exponential family is obtained by replacing $(1/n)\Sigma(y_i - f(t(i)))^2$ by $(1/n)\mathscr{L}(y, f)$, where $\mathscr{L}(y, f)$

is given by (2.2), and then solving the variational problem: find f in $\mathscr{M}$ to minimize

$$\mathscr{L}(y, f) + \frac{n}{2}\lambda \sum_{\beta=1}^{p} \theta_\beta^{-1} \|P^\beta f\|^2. \tag{3.1}$$

The minimizer of (3.1) is also known to have a representation of the form (2.10) [O'Sullivan, Yandell and Raynor (1986), Wahba (1990)] and it is well known that now c and d are the minimizers of

$$I(c, d) = -\sum_{i=1}^{n} l_i(\phi'(t(i))d + \xi'(t(i))c) + \frac{n}{2}\lambda c' Q_\theta c, \tag{3.2}$$

where $l_i(f_i) = y_i f_i - b(f_i)$ with $f_i = \phi'(t(i))d + \xi'(t(i))c$ and where Q_θ is as in (2.11). Since the l_i's are not quadratic, (3.2) cannot be minimized directly. If all $l_i(f_i)$'s are strictly concave, we can use a Newton–Raphson procedure to compute c and d for fixed λ and θ. Let $u_i = -dl_i/df_i$, $u' = (u_1, \ldots, u_n)$, $w_i = -d^2 l_i/df_i^2$, $W = \operatorname{diag}(w_1, \ldots, w_n)$ and $S = (\phi(t(1)), \ldots, \phi(t(n)))'$. Note that from the properties of the exponential family, the vector u and the diagonal entries of the matrix W contain the means and variances for the distributions with parameter f_i. We have $\partial I/\partial c = Q_\theta u + n\lambda Q_\theta c$, $\partial I/\partial d = S'u$, $\partial^2 I/\partial c\, \partial c' = Q_\theta W Q_\theta + n\lambda Q_\theta$, $\partial^2 I/\partial c\, \partial d' = Q_\theta W S$ and $\partial^2 I/\partial d\, \partial d' = S'WS$. The Newton–Raphson iteration satisfies the linear system

$$\begin{pmatrix} Q_\theta W_- Q_\theta + n\lambda Q_\theta & Q_\theta W_- S \\ S'W_- Q_\theta & S'W_- S \end{pmatrix} \begin{pmatrix} c - c_- \\ d - d_- \end{pmatrix} = \begin{pmatrix} -Q_\theta u_- - n\lambda Q_\theta c_- \\ -S'u_- \end{pmatrix}, \tag{3.3}$$

where the subscript minus indicates quantities evaluated at the previous Newton–Raphson iteration; see Gu (1990). With some abuse of notation when the meaning is clear, we will here let f stand for the vector $(f_1, \ldots, f_n)'$. Then, as in Gu (1990), $f = Sd + Q_\theta c$ is always unique as long as S is of full column rank. So only a solution of (3.3) is needed. If Q_θ is nonsingular, (3.3) is equivalent to the system

$$\begin{aligned} (W_- Q_\theta + n\lambda I)c + W_- S d &= (W_- f_- - u_-), \\ S'c &= 0. \end{aligned} \tag{3.4}$$

If Q_θ is singular, any solution to (3.4) is also a solution to (3.3). Let $Q_{W_-,\theta} = W_-^{1/2} Q_\theta W_-^{1/2}$, $c_{W_-} = W_-^{-1/2} c$, $S_{W_-} = W_-^{1/2} S$ and $\tilde{\tilde{y}} = W_-^{-1/2}(W_- f_- - u_-)$. Then (3.4) becomes

$$\begin{aligned} (Q_{W_-,\theta} + n\lambda I)c_{W_-} + S_{W_-} d &= \tilde{\tilde{y}}, \\ S'_{W_-} c &= 0; \end{aligned} \tag{3.5}$$

compare (2.11).

So far, the smoothing parameters $\lambda_\beta = \lambda\theta_\beta^{-1}$ are fixed. We now consider their automatic choice. It is easy to see that the solution of (3.4) gives the

minimizer of

$$\sum_{i=1}^{n}\left(\tilde{\tilde{y}}_i - w_i^{1/2} f_i\right)^2 + \frac{n}{2}\lambda \sum_{\beta=1}^{p} \theta_\beta^{-1}\|P^\beta f\|^2$$

$$= \sum_{i=1}^{n} w_{i-}\left(\tilde{y}_i - f_i\right)^2 + \frac{n}{2}\lambda \sum_{\beta=1}^{p} \theta_\beta^{-1}\|P^\beta f\|^2, \tag{3.6}$$

where $\hat{y}_i = f_{i-} - u_{i-}/w_{i-}$ and the $\tilde{\tilde{y}}_i = w_i^{1/2}\tilde{y}_i$ are the components of $\tilde{\tilde{y}}$ defined before (3.5). The $\tilde{y}_i$'s are called the pseudo-data. The Newton–Raphson procedure iteratively reformulates the problem to estimate the f_i's from the pseudo-data by weighted penalized least squares. See Wang (1994) for further details. Wang (1994) proved the following lemma, which shows that the pseudo-data approximately have the usual data structure if f is the canonical parameter and f_- is not far from f.

LEMMA 1. *Suppose that b of* (2.1) *has two continuous derivatives and b'' is uniformly bounded away from* 0. *If* $|f_{i-} - f_i| = o(1)$ *uniformly in* i, *then*

$$\tilde{y}_i = f_i + \varepsilon_i + o_p(1),$$

where ε_i *has mean* 0 *and variance* w_i^{-1}.

See also Gu (1990).

Wahba [(1990), Section 9.2] suggested (in the single smoothing parameter case) that λ be chosen by minimizing the generalized cross-validation (GCV) score

$$V(\lambda, \theta) = \frac{1/n\|(I - A(\lambda, \theta))\tilde{\tilde{y}}\|^2}{\left[(1/n)\mathrm{tr}(I - A(\lambda, \theta))\right]^2}, \tag{3.7}$$

where $A(\lambda, \theta)$ satisfies

$$\left(w_{1-}^{1/2} f_{\lambda,\theta}(t(1)), \ldots, w_{n-}^{1/2} f_{\lambda,\theta}(t(n))\right)' = A(\lambda, \theta)\tilde{\tilde{y}} \tag{3.8}$$

and $f_{\lambda,\theta}(t(i))$'s are computed from the solution of (3.5). She suggested that λ be fixed, the Newton–Raphson iteration (3.3) be run to convergence, $V(\lambda)$ be evaluated, a new λ be chosen, the new $V(\lambda)$ be evaluated at convergence and then λ be chosen to minimize the $V(\lambda)$ so obtained. Gu (1992a) provided an argument why a better estimate would result from carrying out one step of the Newton–Raphson iteration, minimizing the GCV score, carrying out a second iteration with the new value of λ and iterating to convergence. See also Yandell (1986).

In the case $l_i(f_i) = y_i f_i - b(f_i)$, the dispersion parameter is 1 and $u_i^2/w_i = (y_i - Ey_i)^2/\mathrm{var}(y_i)$. As a result, Gu (1992a) suggested that V be replaced by

U with $\sigma^2 = 1$, giving the U criteria

$$U(\lambda, \theta) = \frac{1}{n}\|(I - A(\lambda, \theta))\tilde{y}\|^2 + \frac{2}{n}\operatorname{tr} A(\lambda, \theta), \tag{3.9}$$

again arguing that U should be minimized at each step of the iteration.

Various criteria can be adopted to measure the goodness of fit of $f_{\lambda,\theta}$ to f. Let ν be a given probability distribution on $\mathscr{T}$. We define the symmetrized Kullback–Leibler distance $\mathrm{SKL}_\nu(f, f_{\lambda,\theta})$ with respect to ν as $\mathrm{SKL}_\nu(f, f_{\lambda,\theta}) = \frac{1}{2}[\mathrm{KL}_\nu(f, f_{\lambda,\theta}) + \mathrm{KL}_\nu(f_{\lambda,\theta}, f)]$, where the KL distance $\mathrm{KL}_\nu(f, f_{\lambda,\theta})$ in the exponential family case of (2.1) is given by $\mathrm{KL}_\nu(f, f_{\lambda,\theta}) = \int[\{\mu(t)f(t) - b(f(t))\} - \{\mu(t)f_{\lambda,\theta}(t) - b(f_{\lambda,\theta}(t))\}]\, d\nu(t)$; here $\mu(t)$ is the expected value of $y|t$ under the distribution $h(y, f(t))$ of (2.1). Following the same argument as in Gu (1992a), it is shown in Wang (1994) that $U(\lambda, \theta)$ is a proxy for SKL_ν with ν the sample design measure for the $t(i)$ and $f_{\lambda,\theta}$ calculated from the solution of (3.5). That is, the minimizer of U with respect to λ, θ can be expected to be a reasonable estimate of the minimizer of SKL_ν with respect to λ, θ with ν the sample design measure.

By comparing (2.11) and (3.5), it can be seen that RKPACK can be called at each step of a Newton–Raphson iteration to solve (3.5) and can then be used to minimize the V or U score at each step.

A simulation study to compare the iterated GCV criteria of (3.7) and the iterated UBR criteria of (3.9) for Bernoulli data was carried out in Wang (1994) and further reported in Wang, Wahba, Chappell and Gu (1995). In that study, the iterated UBR outperformed the iterated GCV criteria in terms of minimizing $\mathrm{SKL}(f, f_{\hat{\lambda},\hat{\theta}})$, and we will be using the former criteria in the analysis of Bernoulli data from WESDR.

4. Approximate Bayesian confidence intervals for exponential families. Bayesian "confidence intervals" for the cross-validated univariate smoothing spline with Gaussian data were introduced by Wahba (1983) and their "across-the-function" properties were suggested there, for functions in an appropriate function space, and λ was chosen according to a predictive mean square criterion. The across-the-function property means that, if for example, $n = 100$, then the 95% Bayesian "confidence intervals" will cover about 95 of the 100 true values of the function being estimated, evaluated at the data points. Nychka (1988, 1990), Wang and Wahba (1995) and others studied the properties of these intervals. Gu and Wahba (1993b) extended these confidence intervals componentwise to the Gaussian SS-ANOVA case, and simulation results there suggested that the across-the-function property was excellent for $f_{\lambda,\theta}$ with λ, θ estimated by GCV and that the componentwise intervals generally behaved reasonably well in the examples studied. Gu (1992c) discussed the extension of these confidence intervals in the univariate case for data from non-Gaussian distributions with convex log likelihood. In this section we review these previous results and describe their extension to the non-Gaussian convex log-likelihood smoothing spline ANOVA case.

We first review the Bayes model associated with smoothing spline ANOVA for Gaussian data and generalize the results to the case where the sampling errors are not iid. Let $\mathscr{M} = \mathscr{H}^0 \oplus \sum_{\beta=1}^{q} \mathscr{H}^\beta$ be the model space as before, with $\mathscr{H}^0 = \text{span}\{\phi_1, \ldots, \phi_M\}$, $R_\beta(s,t)$ the RK for $\mathscr{H}^\beta$ and $Q_\theta(s,t) = \sum_{\beta=1}^{q} \theta_\beta R_\beta(s,t)$. Define the stochastic process $X_\xi(t)$, $t \in \mathscr{T}$, by

$$X_\xi(t) = \sum_{\nu=1}^{M} \tau_\nu \phi_\nu(t) + b^{1/2} \sum_{\beta=1}^{q} \sqrt{\theta_\beta} Z_\beta(t), \tag{4.1}$$

where $\tau = (\tau_1, \ldots, \tau_M)' \sim N(0, \xi I)$ and Z_β are independent, zero mean Gaussian stochastic processes, independent of τ, with $EZ_\beta(s)Z_\beta(t) = R_\beta(s,t)$. Let $Z(t) = \sum_{\beta=1}^{q} \sqrt{\theta_\beta} Z_\beta(t)$. Then $EZ(s)Z(t) = Q_\theta(s,t)$. Suppose observations have the form

$$y_i = X_\xi(t(i)) + \varepsilon_i, \qquad i = 1, \ldots, n, (\varepsilon_1, \ldots, \varepsilon_n)' \sim N(0, \sigma^2 W^{-1}) \tag{4.2}$$

with W positive-definite and known. Let $n\lambda = \sigma^2/b$. Following Gu (1992c) and Gu and Wahba (1993b) and using (1.5.11) and (1.5.12) of Wahba (1990) we have that (for each $t \in \mathscr{T}$) $f_{\lambda,\theta}(t) = \lim_{\xi \to \infty} E(X_\xi(t)|y)$, where $f_{\lambda,\theta}(\cdot)$ is the minimizer in $\mathscr{M}$ of

$$\min (y - f)'W(y - f) + n\lambda \sum_{\beta=1}^{p} \theta_\beta^{-1} \|P_\beta f\|^2. \tag{4.3}$$

The derivation of posterior means and covariances for the components of the model (4.2) is a straightforward generalization of Gu and Wahba (1993b), who provide the result with $W = I$. For reference below, the result is stated in Appendix A. The componentwise confidence intervals will be used to aid in model selection in the data analysis below.

Gu (1992c) considers the univariate case where $\mathscr{L}(y, f)$ is no longer Gaussian, but convex and completely known except possibly for (division by) an unknown dispersion parameter σ^2, and f is assumed to have the same prior distribution as $\lim_{\xi \to \infty} X_\xi(t)$, $t \in \mathscr{T}$. Considering the single smoothing parameter case, he shows, upon setting $n\lambda = \sigma^2/b$, that the posterior distribution of $f(t)|y$ is approximately Gaussian with mean $f_\lambda(t)$ and covariance (the converged value of) $\sigma^2 W^{-1}$. Gu makes some remarks concerning the precision of the estimate, remarking that it is likely to be better for larger λ and noting that it is primarily useful for obtaining the Bayesian confidence intervals. He carried out a Monte Carlo experiment with a single predictor variable and Bernoulli data and the results were highly suggestive that these intervals do have reasonable across-the-function properties. Gu's argument extends word for word to the multicomponent smoothing spline ANOVA case [see Wang (1994)], resulting in posterior covariances for the components of the model. The result is stated in Appendix A.

To compute the approximate componentwise Bayesian confidence intervals, we need to calculate the posterior variances given in Appendix A, based on converged values. Gu and Wahba (1993b) discussed calculation of these

quantities for the Gaussian case when $W = I$, and a demonstration program with examples was added to the original RKPACK [Gu (1989)] in 1992. An outline of how the computational algorithm in RKPACK is exploited to compute the componentwise confidence intervals for $W \neq I$ is given in Appendix A. GRKPACK [Wang (1995)] can be used to obtain SS-ANOVA estimate of f with Bernoulli data as well as general binomial data and Poisson and gamma data. GRKPACK minimizes (3.2) via the Newton–Raphson iteration of (3.5), using RKPACK as a subroutine, and provides for V, U and a third option not discussed here for choosing λ and θ. The code may also be used to compute the confidence intervals componentwise and for the entire function $f_{\hat{\lambda},\hat{\theta}}$, using the estimated λ and θ, and converged values of u_- and W_-. Computational details and program documentation may be found in Wang (1995). The code is available by ftp to `netlib.att.com` in the file `netlib/gcv/grkpack.shar.Z`.

Results of a simulation study of the overall and componentwise Bayesian confidence intervals with Bernoulli data may be found in Wang (1994), using the U option for the smoothing parameters. The means of the nominal coverages were quite good even with sample sizes of only 200 and 400. As in the Gaussian case the componentwise intervals were somewhat less reliable than the intervals for the whole function. Although further study of the properties of these intervals is warranted, they turned out to be quite useful in our applications; see Section 6. We note that these across-the-function studies are typically being carried out as though the unknown f is in fact an element of the model space $\mathscr{M}$.

Recently Raghavan (1993) carried out an exhaustive study of the properties of the posterior distribution of the logit f in the case of Bernoulli data, where f is considered to be a realization of the associated stochastic process. She raises some interesting questions concerning the tail behavior of the posterior. At this time we do not know what implications of these results might be for the use of the Bayesian confidence intervals under the assumption that f is an element of $\mathscr{M}$, since this is a different assumption than f a realization of the stochastic process associated with the reproducing kernel of the model space.

5. Selecting the model. In this discussion we assume that our goal is prediction and that no model under consideration may be correct. We want to select, from among the models being entertained, one or several which are likely to have the best predictive capability, in some sense to be defined. The value of U at the minimum could be compared for different models. However, for nested models, this is not quite "fair," since $\min_{\lambda_1,\ldots,\lambda_p} U(\lambda_1,\ldots,\lambda_p) \leq \min_{\lambda_1,\ldots,\lambda_{p-1}} U(\lambda_1,\ldots,\lambda_{p-1},\infty)$: setting $\lambda_p = \infty$ is equivalent to deleting the component in the pth penalized subspace from the model. It appears that one should apply a "charge" to this minimization procedure for allowing the minimization over λ_p. What this charge should be in the SS-ANOVA context is an interesting question to which we do not have the answer at the present time.

For sufficiently large data sets, we have the trivial but highly defensible answer to the model selection problem, which is much favored in the supervised machine learning community: divide the data into a "training" set and a "testing" set, fit each candidate model on the training set (including choosing the smoothing parameters) and select one or more of the fitted models, on the basis of their predictive ability on the "testing" set. For example, letting $\mathrm{KL}_\nu(f, f_{\hat\lambda,\hat\theta})$ be the selection criteria, we need only be concerned with the so-called comparative KL distance $-\int[\,\mu(t)f_{\hat\lambda,\hat\theta}(t) - b(f_{\hat\lambda,\hat\theta}(t))]\,d\nu(t)$ since this quantity differs from the KL distance by a quantity which does not depend on $f_{\hat\lambda,\hat\theta}$. This may be estimated on the testing set by

$$\mathrm{KL}_\nu = -\frac{1}{\#\text{ in test set}}\sum_{j\in\text{test set}}\left[y_j f_{\hat\lambda,\hat\theta}(t(j)) - b(f_{\hat\lambda,\hat\theta}(t(j)))\right],$$

where $f_{\hat\lambda,\hat\theta}$ has been fitted on the training set. That was done in the conference proceedings [Wahba, Gu, Wang and Chappell (1995)]. In practice several models may appear to be "close" by this procedure. In that case one might like to retain all of the models which are not (in *some* sense) significantly worse than the best model. How to define and quantify "significantly" here is again an interesting question for which we do not have an answer.

If n is not large enough to set aside a test set, a second-level k-fold (or n-fold) cross-validation may be used to estimate the comparative KL distance by dividing the data into k subsets, fitting the candidate model on the data with the kth subset left out (presumably including refitting the smoothing parameters), estimating the comparative KL distance on the omitted (test) subset and averaging over k estimates. This procedure can be extremely computer intensive due to the smoothing parameter reestimation. It would be interesting to develop a defensible variant of this procedure which does not involve repeated reestimation of the smoothing parameters. A bias-corrected bootstrap (BCB) can also be defined for estimating the comparative KL distance [Efron and Tibshirani (1993), Wang (1994)], and some reasonable results with small data sets ($n \sim 200$) and a main effects model were obtained in Wang (1994). The BCB estimates were nearly the same as those obtained by a second-level n-fold cross-validation. However, the BCB was not satisfactory on the WESDR data set of $n = 669$ (below) because the smoothing parameter estimates on the bootstrap samples tended to seriously overfit the data in a substantial fraction of the bootstrap samples. We attributed this to the fact that the UBR estimate for the smoothing parameters being used here is assuming a dispersion parameter of 1, while what might be considered the effective dispersion parameter of the bootstrap samples must be less than 1, since in large data sets some observations are likely to be resampled many times. It remains an open question whether some variant of the BCB can be successfully developed in this context.

6. Wisconsin Epidemiological Study of Diabetic Retinopathy. The WESDR is an ongoing epidemiological study of a cohort of patients receiving their medical care in an 11-county area in southern Wisconsin, who were first

examined in 1980–1982, then again in 1984–1986 and 1990–1992. Detailed descriptions of the data have been given by Klein, Klein, Moss, Davis and DeMets (1988, 1989c) and references therein. All younger onset diabetic persons (defined as less than 30 years of age at diagnosis and taking insulin) and a probability sample of older onset persons receiving primary medical care in an 11-county area of southwestern Wisconsin in 1979–1980 were invited to participate; 1210 younger onset patients were identified, of which 996 agreed to participate in the baseline examination, and of those, 891 participated in the first followup examination. The older onset persons fell into two groups: older onset taking insulin and older onset not taking insulin. Data from these groups were also analyzed and the results were reported in Wang (1994), but not here.

A large number of medical, demographic, ocular and other covariates were recorded at the baseline and later examinations along with a retinopathy score for each eye (to be described). Relations between various of the covariates and the retinopathy scores have been extensively analyzed by standard statistical methods including categorical data analysis and parametric GLIM models, and the results have been reported in the various WESDR manuscripts. See Klein, Klein, Moss, Davis and DeMets (1984a, b, 1989a, b), Klein, Klein, Moss, DeMets, Kauffman and Voss (1984) and Klein, Klein, Moss and Cruickshanks (1994a, b). Thus, the present study has benefited from the previous analyses. It was our goal to see whether or not further information might be extracted via SS-ANOVA, and if so, to demonstrate its use. We limited this first study to developing a predictive model for progression (to be defined) of diabetic retinopathy at the first followup, as a function of some of the covariates available at baseline. We only list the covariates pertinent to our analysis:

1. `agb`: age at the baseline examination (years);
2. `agd`: age at diagnosis (years);
3. `dur`: duration of diabetes at baseline `(agd + dur = agb)`;
4. `gly`: glycosylated hemoglobin, a measure of hyperglycemia (%);
5. `bmi`: body mass index (weight in kilograms/height in meters, squared).

At the baseline and followup examinations, stereoscopic color fundus photographs of each eye were graded in a masked fashion using the modified Airlie House classification system. Grading protocols have been described in detail elsewhere; see Klein, Klein, Moss, Davis and DeMets (1989a, b). At baseline and the four-year followup, each eye was given one of six retinopathy severity score grades: 10 (no retinopathy), 21, 31, 41 or 51 (nonproliferative retinopathy) or 60 + (proliferative retinopathy). In the WESDR, a retinopathy severity score was also assigned to each participant by giving the eye with the higher score greater weight. [See Klein, Davis, Segal, Long, Harris, Haug, Magli and Syrjala (1984)]. For example, the level for a participant with level 31 retinopathy in each eye is specified by the notation "level 31/31," whereas that for a participant with level 31 in one eye and less severe retinopathy in the other eye is noted as "level 31/< 31." This scheme

provided an 11-step scale: 10/10, 21/ < 21, 21/21, 31/ < 31, 31/31, 41/ < 41, 41/41, 51/ < 51, 51/51, 60 + / < 60 + and 60 + /60 + . In the WESDR study, progression (y_i) for a participant (with nonproliferative or no retinopathy at baseline) is defined to be 1 if the participant had his/her baseline level increased two steps or more (10/10 to 21/21 or greater, or 21/ < 21 to 31/ < 31 or greater, for instance), and 0 otherwise. Our aims are to find risk factors and to build models for prediction of progression of diabetic retinopathy.

We report an analysis of a subgroup of the younger onset population, consisting of 669 subjects with no or nonproliferative retinopathy (scores of 51/51 or better) at the start and no missing data for the variables we studied. Since we will be analyzing Bernoulli($\{0, 1\}$) data, $b(f) = \log(1 + e^f)$. The averaging operators, penalty functionals and reproducing kernels used are given in Appendix B. This group has been called the younger onset progression group and was analyzed in Klein, Klein, Moss, Davis and DeMets (1988). The sample size differs slightly from Klein, Klein, Moss, Davis and DeMets (1988) due to different missing data patterns. The remainder of the 891 subjects either had proliferative retinopathy at the start or had missing data.

Klein, Klein, Moss, Davis and DeMets (1988) reported that gly is a strong predictor of progression of diabetic retinopathy in the younger onset group. Figure 1 there suggests that dur has a nonlinear effect on the probability of progression. Four individual univariate spline fits for risk of progression as functions, respectively, of gly, agb, dur and bmi suggested that the effect of gly was very strong and fairly linear in the logit and that agb, dur and bmi were strong and nonlinear. Some exploratory GLIM modeling using the SAS procedure LOGISTIC [SAS Institute (1989)] suggested (agb, bmi) and/or (dur, bmi) interactions might be present. We entertained the model

$$f(\mathtt{dur}, \mathtt{gly}, \mathtt{bmi}) = \mu + f_1(\mathtt{dur}) + a_2 \cdot \mathtt{gly} + f_3(\mathtt{bmi}) + f_{13}(\mathtt{dur}, \mathtt{bmi}) \tag{6.1}$$

and also the model (6.1) with agb replacing dur.

These two models gave qualitatively very similar results, suggesting the possibility, previously considered by the WESDR study, that agb may be considered as a proxy for dur, the relevant predictor being the length of time the subject is exposed to diabetes. To see whether there was an effect of age over and above that explained by dur, it was decided to fit a model using agd, dur, gly and bmi. (Recall that agd + dur = agb.) General main effects for agd, dur, gly and bmi were included, along with general interaction terms for agd, dur and agd, bmi. The gly smooth part and the agd, bmi interaction term turned out to be of negligible size in a practical sense compared to the other terms. After deleting these terms, the model

$$f(\mathtt{agd}, \mathtt{dur}, \mathtt{gly}, \mathtt{bmi}) = f_1(\mathtt{agd}) + f_2(\mathtt{dur}) + a_3 \cdot \mathtt{gly} + f_4(\mathtt{bmi}) + f_{14}(\mathtt{agd}, \mathtt{bmi})$$

was fitted and the Bayesian confidence intervals were computed for f. The components for agd were not obviously negligible at the fitting stage. However, plots of cross sections of f versus agd at the median gly and several levels of dur plotted along with the confidence intervals for f showed that a constant function of agd was in the interior of all of the confidence bands, suggesting that the (somewhat difficult to interpret) marginal dependence on agd, taking into account dur, was probably not meaningful. Therefore, we adopted the model of (6.1) as our SS-ANOVA model. Without the f_{13} term, the model (6.1) would correspond to an additive model with cubic smoothing splines for the smooth, also described in Hastie and Tibshirani (1990) and Chambers and Hastie (1992). An examination of the size of the fitted f_{13} term, along with cross sections of its confidence intervals revealed that it was not negligible in a practical sense, (i.e., negligibly small compared to the other terms in the model) and, although the confidence intervals were wide, there was a reasonably sized region where they did not cover 0. For this reason f_{13} was retained in the model. The fitting procedure, which is based on a proxy for the SKL distance of the fit from the truth, has, by its choice of smoothing parameter, suggested that this term does belong in the model. Other (in sample) procedures for reexamining whether to retain f_{13} are possible (as noted by a referee); for example, we could have examined the minimum $U(\lambda, \theta)$ with and without the f_{13} term, and the cosine diagnostics in Gu (1992a) could also have been examined. Absent objective methods for interpreting the results of these examinations, we have used informal methods and confidence intervals here. Analysis of all three models are in Wang (1994).

The left panel in Figure 1 gives a scatterplot of dur versus bmi, with the solid circles representing patients with four-year progression of retinopathy

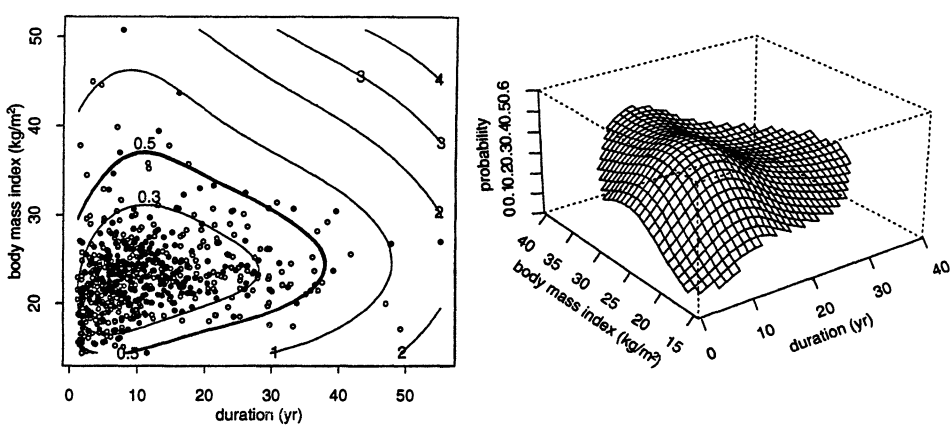

FIG. 1. *Left: Data and contours of constant posterior standard deviation. Right: Estimated probability of progression as a function of duration and body mass index for glycosylated hemoglobin fixed at its median.*

and open circles representing those without progression of retinopathy. The contour lines in this panel are level curves of constant posterior standard deviation of the overall fit $f_{\hat{\lambda},\hat{\theta}}$, evaluated at the median value of gly. The heavy curve is the 0.5 contour. Note that this heavy curve provides a reasonable boundary defining a data-dense region. The right panel gives a three-dimensional plot of the estimated probability of progression $p(\mathtt{dur},\mathtt{bmi},\mathtt{gly})$ as a function of dur and bmi for gly fixed at its median value, from the SS-ANOVA fitted model (6.1). This three-dimensional plot covers only the region enclosed by the 0.5 level curve of the left panel. Outside this region, the fit is not considered reliable. Plots of multivariate SS-ANOVA fits carried into regions of very sparse or no data can be assumed to be meaningless and can appear visually ugly and misleading. Therefore, it is useful to have this readily computable method for determining a reasonable region over which the fits are to be taken at face value. Figures 2 and 3 give slices of $p(\mathtt{dur},\mathtt{bmi},\mathtt{gly})$, with the cross sections of Figure 2 plotted as a function of dur for four levels of bmi and gly, and the cross sections of Figure 2 plotted as a function of bmi for four levels of dur and gly. These levels are at the 12.5th, 37.5th, 62.5th and 87.5th percentiles of the bmi and gly values.

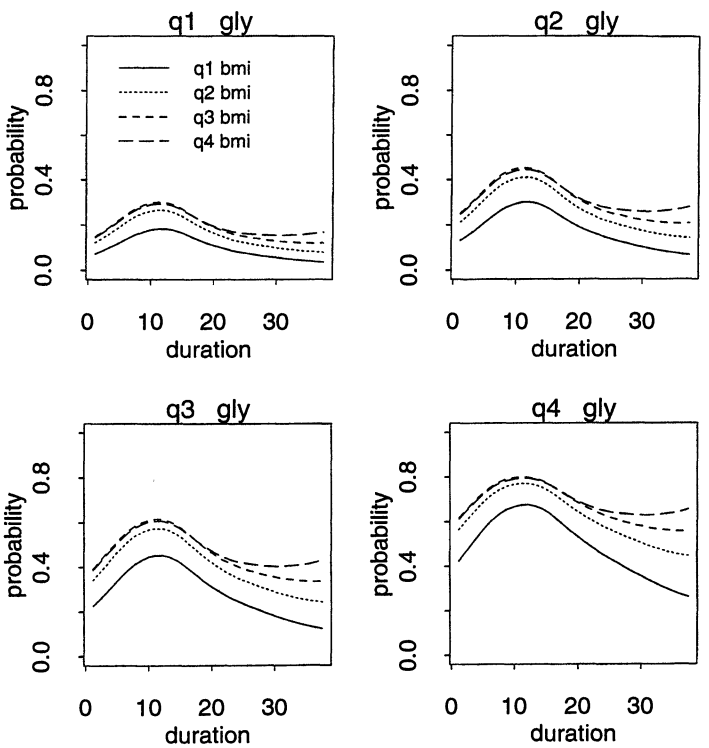

FIG. 2. *Estimated probability of progression as a function of* dur *for four levels of* bmi *by four levels of* gly.

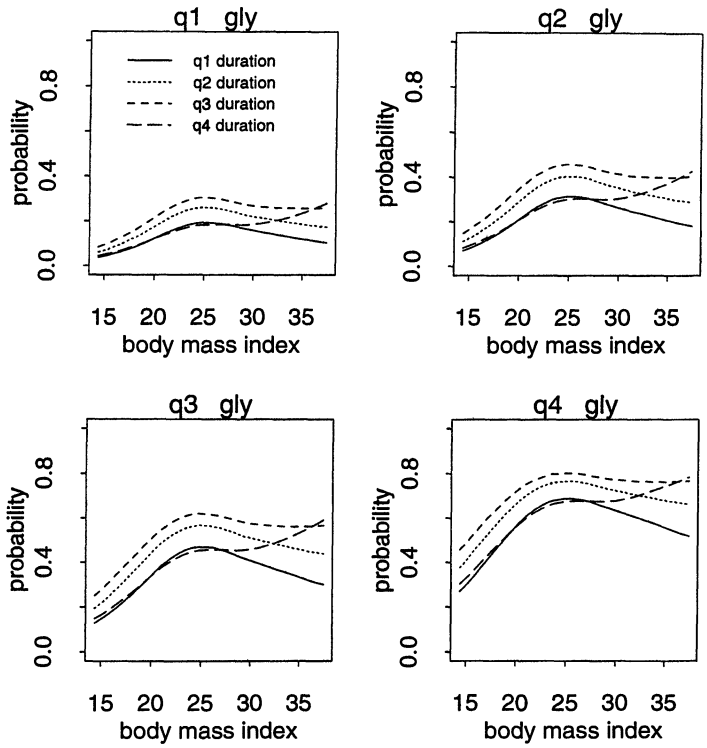

FIG. 3. *Estimated probability of progression as a function of* bmi *for four levels of* dur *by four levels of* gly.

Figure 4 gives slices of p (solid lines) and their Bayesian confidence intervals (dotted lines) plotted as a function of dur for three levels of bmi, gly, and Figure 5 gives slices and their Bayesian confidence intervals as a function of bmi for three levels of dur, gly. The three levels are at the 25th, 50th and 75th percentiles. For the convenience of a reader who might like to find his or her probability of four-year progression in Figures 2–5, Table 1 gives the correspondence between the percentiles and the physical units.

As previously reported [Klein, Klein, Moss, Davis and DeMets (1988)], increases in glycosylated hemoglobin at baseline are associated with increases in the risk of progression of diabetic retinopathy over the first four years of the study. At most durations of diabetes or glycosylated hemoglobin levels at baseline, the risk of four year progression of retinopathy increases with increasing body mass index at baseline until a value of about 25 kg/m^2, after which there was flattening, except at the longer durations, where risk of progression continues to increase with body mass index. However, the confidence intervals are fairly wide at this part of the surface. These relations of body mass index to progression of retinopathy were not found in earlier analysis and the reasons for these findings are not known.

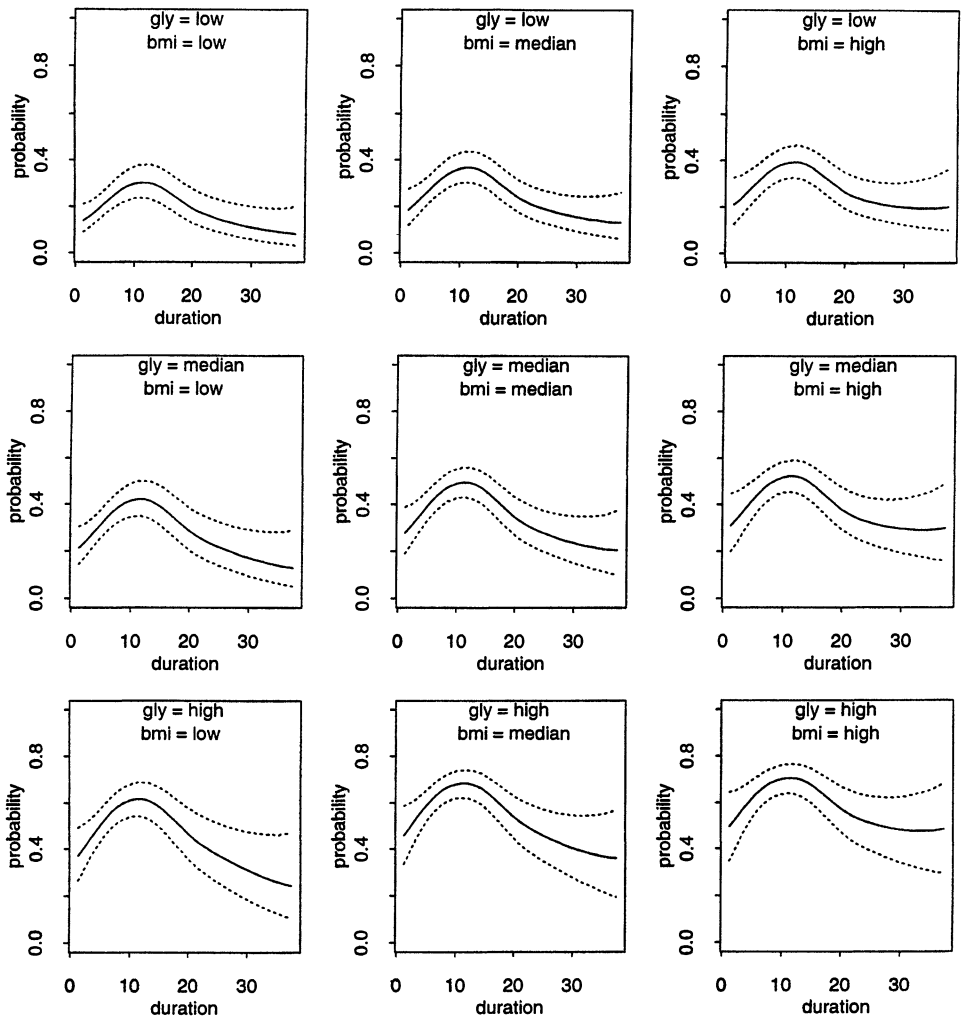

FIG. 4. *Estimated probability of progression and Bayesian confidence intervals as a function of* `dur` *for three levels of* `bmi` *by three levels of* `gly`.

The risk of progression of retinopathy as a function of duration at baseline increases up to a duration of about 10 years, when it then decreases. Several explanations for this decrease are possible. The frequency of other factors associated with higher risk of progression of retinopathy, which were not included in these analyses, may decrease in people with longer duration of diabetes. These findings may also be due to censoring due to death in people with longer duration of diabetes (if people with longer duration of diabetes whose retinopathy progressed in the interim are more likely to not get

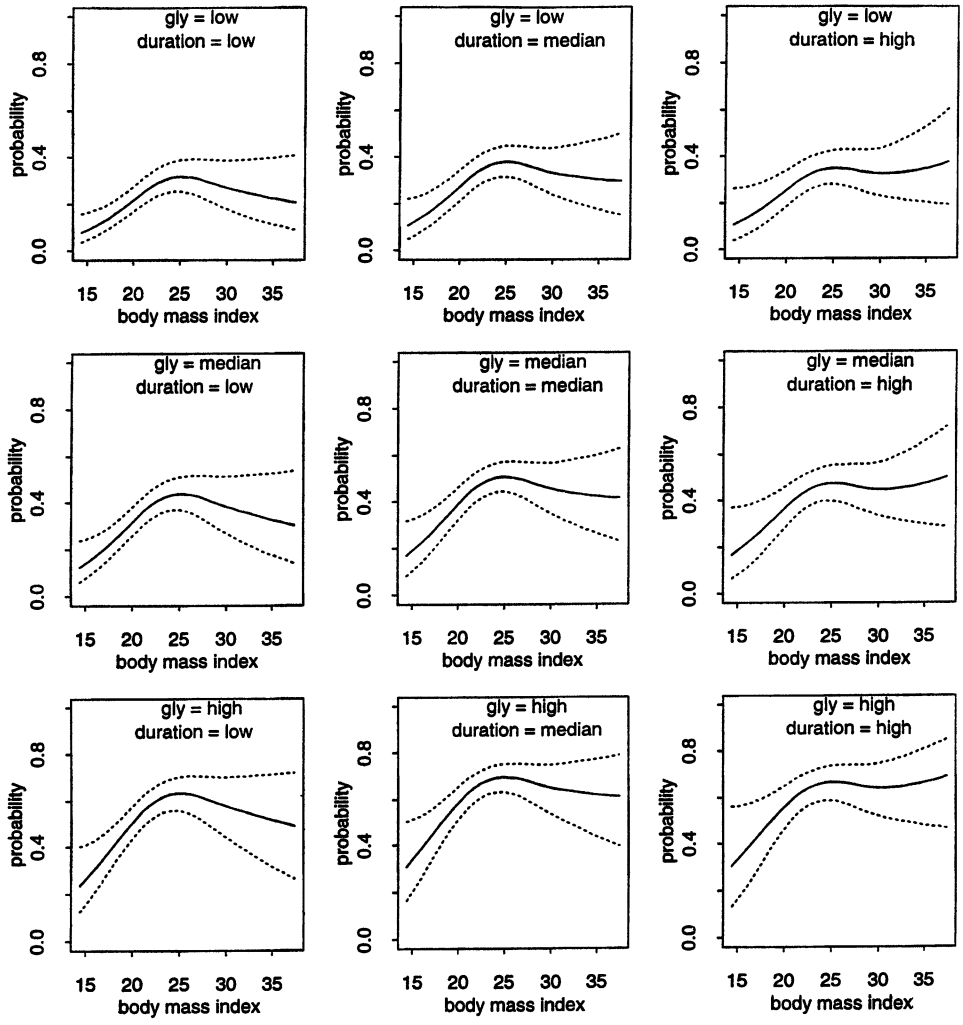

FIG. 5. *Estimated probability of progression and Bayesian confidence intervals as a function of* bmi *for three levels of* dur *by three levels of* gly.

TABLE 1

Percentile	**12.5**	**25**	**37.5**	**50**	**62.5**	**75**	**87.5**
dur (years)	3.3	5.4	7.1	9.2	11.5	15.3	21.6
gly (%)	9.6	10.7	11.5	12.2	13.2	14.1	15.4
bmi (kg/m^2)	18.7	20.6	21.7	22.9	23.9	25.2	27.0

examined at the four-year followup due to death than people with long duration of diabetes whose retinopathy did not progress).

7. Computational considerations. A detailed discussion of the computations may be found in Wang (1995). RKPACK is based on matrix decompositions, generally $O(n^3)$. Since the GRKPACK algorithm calls RKPACK iteratively, it also uses $O(n^3)$ operations. For the models we tried, it usually took from one to up to as much as eight hours on our Decstation 3000/400 alpha workstation, depending on the number of smoothing parameters being estimated. The calculations as presently performed require the storage of two $n \times n$ matrices for each smoothing parameter to be fitted, so that storage requirements are also relatively high. The biggest data set we have run with five to seven smoothing parameters is about 800. The cost is still negligible, compared with the cost of collecting data like the WESDR data, and, we trust, not large considering the potential for extracting interpretable information from very expensive data sets. Numerical methods for efficiently applying SS-ANOVA with multiple smoothing parameter estimation to much larger data sets is an area of active research. One tool to consider is to approximate the span of the $(n + M)$ basis functions in (2.10) by a carefully chosen subset. The variational problem of (3.2) is then solved in this lower dimensional subspace. This approach was proposed in Wahba (1980) for the special case of thin plate splines, and has been developed and implemented by Hutchinson (1984) in ANUSPLIN, Hutchinson and Gessler (1994), O'Sullivan (1990), Luo and Wahba (1995) and others. See also the discussion in Section 7 of Nychka, Wahba, Goldfarb and Pugh (1984) and references cited therein. Numerical methods appropriate for solving the variational problem on a set of basis functions with a single smoothing parameter are implemented in GCVPACK [Bates, Lindstrom, Wahba and Yandell (1987)] and can be extended to the multiple smoothing parameter case. Extension of these so-called hybrid methods to the exponential family SS-ANOVA case are an area of active research (Z. Luo and D. Xiang, personal communications). Criteria with various goals in mind may be adopted to choose the basis functions. The second tool is to exploit and extend the randomized trace technique in Girard (1987, 1989, 1991) and Hutchinson (1989), possibly in conjunction with iterative methods for solving the variational problem that avoids matrix decompositions. This too is an area of active research [Golub and Von Matt (1995), J. Gong, personal communication]. This has been done in a single smoothing parameter Gaussian case in Wahba, Johnson, Gao and Gong (1995), where a conjugate gradient algorithm for the variational problem is employed and the number of conjugate gradient iterations is also considered as a parameter.

A referee has asked us to discuss the relationship between the estimation procedure here and the backfitting algorithm [Hastie and Tibshirani (1990) (HT), Chambers and Hastie (1992)]. We do that here. HT (Section 5.2.3) discussed backfitting in the context of the general SS-ANOVA problem of (2.8) as earlier set up by Chen, Gu and Wahba (1989). Referring to (2.8)–(2.11),

let $f_0(t) = \sum_{\nu=1}^{M} d_\nu \phi_\nu(t)$ and let $f_\beta(t) = \sum_{i=1}^{n} c_i \theta_\beta R_\beta(t(i), t)$. Then $f_{\lambda,\theta}(\cdot) = f_0(\cdot) + \sum_{\beta=1}^{n} f_\beta(\cdot)$, with c and d satisfying (2.11), is the minimizer over f in $\mathscr{M}$ of

$$\sum_{i=1}^{n} \left(y_i - f(t(i))\right)^2 + \sum_{\beta=1}^{p} \lambda_\beta \|P^\beta f\|^2, \tag{7.1}$$

where we have set $\lambda_\beta = n\lambda\theta_\beta^{-1}$. Now define $\tilde{f}_0(\cdot) = f_0(\cdot)$ and $\tilde{f}_\beta(\cdot) = \sum_{i=1}^{n} c_{i\beta} R_\beta(t(i), \cdot)$, for arbitrary $c_{i\beta}$. In what follows it will be useful to recall that $\|\tilde{f}_\beta\|^2 \equiv \|P^\beta \tilde{f}_\beta\|^2 = c_\beta' R_\beta c_\beta$, where $c_\beta = (c_{1\beta}, \ldots, c_{n\beta})'$ and R_β is the $n \times n$ matrix with ijth entry $R_\beta(t(i), t(j))$; see Wahba (1990). HT observe that the minimizer of (7.1) in the span of all functions of the form $\{\tilde{f}_0(\cdot), \tilde{f}_1(\cdot), \ldots, \tilde{f}_p(\cdot)\}$ is obtained by finding d and c_β, $\beta = 1, \ldots, p$, to minimize $\|y - Sd - \sum_{\beta=1}^{p} R_\beta c_\beta\|^2 + \sum_{\beta=1}^{p} \lambda_\beta c_\beta' R_\beta c_\beta$. They remark that there are $pn + M$ unknowns, since the c_i have been replaced by $c_{i\beta}$. They note that the (vector) smooths corresponding to the minimizers $\tilde{\mathbf{f}}_0 \equiv S\tilde{d}$ and $\tilde{\mathbf{f}}_\beta \equiv R_\beta c_\beta$ satisfy the backfitting equations

$$\tilde{\mathbf{f}}_\gamma = S_\gamma\left(y - \sum_{\alpha \neq \gamma} \tilde{\mathbf{f}}_\alpha\right), \qquad \gamma = 0, 1, \ldots, p, \tag{7.2}$$

with the smoother matrix S_0 given by $S_0 = S(S'S)^{-1}S'$ and the other smoother matrices S_β given by $S_\beta = R_\beta(R_\beta + \lambda_\beta I)^{-1}$, $\beta = 1, \ldots, p$. The backfitting algorithm solves for $\tilde{\mathbf{f}}_0$ and $\tilde{\mathbf{f}}_\beta$, $\beta = 1, \ldots, p$, by cycling through $\tilde{\mathbf{f}}_\gamma = \mathbf{S}_\gamma(\mathbf{y} - \sum_{\alpha \neq \gamma} \tilde{\mathbf{f}}_\alpha)$, $\gamma = 0, 1, \ldots, p$. The backfitting algorithm is known to converge if the Frobenius norm of each product $S_\alpha S_\beta$ is less than 1.

The last p backfitting equations are equivalent to

$$R_\beta\left(\lambda_\beta c_\beta + \sum_{\alpha=1}^{p} R_\alpha c_\alpha\right) = R_\beta(y - Sd), \qquad \beta = 1, \ldots, p. \tag{7.3}$$

Now suppose $\lambda_\alpha c_\alpha = n\lambda c$ for some c, $\alpha = 1, \ldots, p$. Recalling that $\lambda_\alpha = n\lambda\theta_\alpha^{-1}$, this would give

$$R_\beta\left(n\lambda I + \sum_{\alpha=1}^{p} \theta_\alpha R_\alpha\right)c = R_\beta(y - Sd), \qquad \beta = 1, \ldots, p. \tag{7.4}$$

Thus if c satisfies $(n\lambda I + \sum_{\beta=1}^{p} \theta_\beta R_\beta)c = y - Sd$, then $c_\beta = \theta_\beta c$, $\beta = 1, \ldots, p$, satisfies the backfitting equations. Thus, despite the apparently larger number $np + M$ of unknowns compared to the $n + M$ unknowns in the present formulation, the backfitting solutions $\tilde{\mathbf{f}}_\gamma$, $\gamma = 0, 1, \ldots, p$, are, at convergence, equivalent to solving

$$(Q_\theta + n\lambda I)c + Sd = y, \tag{7.5}$$

$$S'c = 0 \tag{7.6}$$

for c and d and setting $\tilde{\mathbf{f}}_0 = Sd$, $\tilde{\mathbf{f}}_\beta = \theta_\beta R_\beta c$. [Equation (7.6) follows by observing that the first backfitting equation becomes $Sd = S_0(y - Q_\theta c)$. Substituting this into $(n\lambda I + Q_\theta)c = (y - Sd)$ results in $n\lambda c = (I - S_0) \times (y - Q_\theta c)$, which entails (7.6).] This result, while not immediately evident

from HT [see (ii) on page 113] is not surprising since we are solving a variational problem in $\mathscr{M}$, and the setup in HT is equivalent to solving the same variational problem in a certain $(np + M)$-dimensional subspace of $\mathscr{M}$ which contains the n-dimensional subspace in which the solution lies. Note that c is not necessarily unique but the $\tilde{\mathbf{f}}_\gamma$ are, provided that S is of full column rank. In the SS-ANOVA context with given smoothing parameters, the backfitting algorithm, then, is essentially an alternative method for solving (7.5) and (7.6). As noted in HT, in many important applications there is special structure to the smoother matrices S_β, and when this can be taken advantage of, implementations of the backfitting algorithm can be much faster than the methods of the present paper. In the case of completely general S_β, however, direct solution of (7.5) and (7.6) can be expected to be faster.

The algorithms embodied in RKPACK and GRKPACK are driven by two considerations: (1) the requirement that the smoothing parameters be chosen by GCV or UBR and (2) no special structure is assumed in the R_β. The algorithmic procedures were specifically designed to handle the unstructured case. When special structure is available, cheaper algorithms are available.

We now turn to choosing multiple smoothing parameters via GCV and then UBR in the context of the backfitting algorithm.

It has been suggested (see the discussion of the BRUTO algorithm in HT Section 9.4.3) that $\operatorname{tr} A(\lambda, \theta) \equiv \operatorname{tr} A(\lambda_1, \ldots, \lambda_p)$ be approximated by $\sum_{\gamma=0}^{p} \operatorname{tr} S_\gamma$ in the definition of the GCV function V of (2.12), to give

$$V^B(\lambda_1, \ldots, \lambda_p) = \frac{\|y - \sum_{\gamma=1}^{p} \tilde{\mathbf{f}}_\gamma\|^2}{\left(1 - (1/n)\left[M + \sum_{\beta=1}^{p} \operatorname{tr} S_\beta\right]\right)^2}. \tag{7.7}$$

The minimization of V^B with respect to λ_β for the other smoothing parameters fixed could be done at the βth step in each cycle of the backfitting algorithm. This could be done in general with a matrix decomposition of each R_β. In contrast, RKPACK uses a variant of Newton's method (including derivative information) to minimize the GCV or UBR function in all of the smoothing parameters simultaneously, where each iteration of the Newton descent costs one matrix decomposition. In the case of the main effects only model, where each main effect $f_\alpha(\cdot)$ is a univariate polynomial spline, the smooth of $z_\beta \equiv y - \sum_{\gamma \neq \beta} \tilde{\mathbf{f}}_\gamma$, as well as $\operatorname{tr} S_\gamma$ can be obtained at a cost $O(n)$ from, for example, `spline.smooth` or one of the univariate spline codes in `netlib/gcv`. These codes make use of the special structure that obtains for certain matrices associated with the computation of the polynomial smoothing splines in one variable. A good place to read about this special structure is Green and Silverman [(1994), Section 2.6].

If all the S_γ are pairwise orthogonal, $S_\alpha S_\beta = 0$ for $\alpha \neq \beta$, then $A(\lambda_1, \ldots, \lambda_p) = \sum_{\gamma=0}^{p} S_\gamma$ and the BRUTO approximation V^B to V is exact. This approximation can be expected to be better or worse according as S_β is "close" to orthogonal to S_α. A cheap diagnostic for this might be $\operatorname{tr} R_\alpha R_\beta / \sqrt{\operatorname{tr} R_\alpha^2} \sqrt{\operatorname{tr} R_\beta^2}$. In the case of the main effects model with each

main effect a polynomial smoothing spline, the implied S_α and S_β can be expected to be close to orthogonal for reasonably distributed data, and an $O(n)$ rather than $O(n^3)$ algorithm employing backfitting is available via the use of the backfitting algorithm, including the use of GCV to choose the smoothing parameters; see HT. Due to this near orthogonality, it is to be expected that the results would be similar to those obtained (more expensively) by RKPACK. As the S_β become less mutually orthogonal, this approximation may eventually break down as a good procedure for implementing GCV or UBR. In any case, properties of this approximation in the general nonorthogonal case remain to be obtained. Similar remarks in the Bernoulli case using UBR could also be made. In the non-Gaussian, nonorthogonal case, it can be expected that the estimation of the smoothing parameters by UBR will be sensitive to the particular way in which the backfitting iteration, the iterative calculation of the c_β and the search for the λ_β are arranged; see Gu (1992a). Properties and relative timing in the unstructured, nonorthogonal case remain to be investigated.

Since cheap diagnostics for orthogonality are available, it is intriguing to speculate whether GRKPACK could be combined with backfitting to take advantage of the strengths of both. For example, the components (subspaces) in data space could be grouped so that groups are orthogonal while within group subspaces are not. One might then call GRKPACK as a subroutine for each group within a backfitting algorithm.

8. Conclusions. We have developed a flexible family of models for risk factor estimation (and other statistical problems) which provide an interpretable alternative to the rigid parametric GLIM models, for use when the GLIM models may not be adequate. The models can be used as tools to check whether GLIM models are adequate. We were motivated by the possibility of describing interesting relationships in data from large epidemiologic studies that might not be found by more traditional methods, and we have demonstrated this possibility through an analysis of WESDR data. Further work remains to be done in formalizing model selection procedures and in developing computational techniques which will allow analysis of much larger data sets than we have analyzed here. The extension of the approach to survival data, to longitudinal data and to a variety of other data structures and types of responses arising in epidemiologic studies is certainly feasible, although the details may be nontrivial to implement.

APPENDIX

A. Details of Bayesian confidence intervals. The calculation of posterior means and covariances for the components of the model (4.1) and (4.2) as $\xi \to \infty$ and $n\lambda = \sigma^2/nb$ is a straightforward generalization of Gu and Wahba [(1993b), Theorem 1] with $M = Q_\theta + n\lambda I$ there replaced by $M = Q_\theta + n\lambda W^{-1}$. We reproduce the result here to use in the description below of

how they can be calculated in the non-Gaussian case with the aid of RKPACK. Letting $M = Q_\theta + n\lambda W^{-1}$, $g_{0,\nu}(t) = \tau_\nu \phi_\nu(t)$ and $g_\beta(t) = b^{1/2}\sqrt{\theta_\beta} Z_\beta(t)$, $\nu = 1, \dots, M$, $\beta = 1, \dots, q$, then

$$E(g_{0,\nu}(t)|y) = d_\nu \phi_\nu(t),$$

$$E(g_\beta(t)|y) = \sum_{i=1}^{n} c_i \theta_\beta R_\beta(t, t(i)),$$

$$\text{(A.1)} \quad \begin{aligned} \frac{1}{b}\mathrm{Cov}(g_{0,\nu}(t), g_{0,\mu}(t)|y) &= \phi_\nu(t)\phi_\mu(t) e_\nu'(S'M^{-1}S)^{-1} e_\mu, \\ \frac{1}{b}\mathrm{Cov}(g_\beta(s), g_{0,\nu}(t)|y) &= -d_{\nu,\beta}(s)\phi_\nu(t), \\ \frac{1}{b}\mathrm{Cov}(g_\beta(s), g_\beta(t)|y) &= \theta_\beta R_\beta(s,t) - \sum_{i=1}^{n} c_{i,\beta}(s)\theta_\beta R_\beta(t, t(i)), \\ \frac{1}{b}\mathrm{Cov}(g_\gamma(s), g_\beta(t)|y) &= -\sum_{i=1}^{n} c_{i,\gamma}(s)\theta_\beta R_\beta(t, t(i)), \end{aligned}$$

where e_ν is the νth unit vector and $(d_{1,\beta}(t), \dots, d_{M,\beta}(t)) = d_\beta(t)'$ and $(c_{1,\beta}(t), \dots, c_{n,\beta}(t)) = c_\beta(t)'$ are given by

$$\text{(A.2)} \quad d_\beta(t) = (S'M^{-1}S)^{-1} S'M^{-1} \begin{pmatrix} \theta_\beta R_\beta(t, t(1)) \\ \vdots \\ \theta_\beta R_\beta(t, t(n)) \end{pmatrix},$$

$$\text{(A.3)} \quad c_\beta(t) = \left[M^{-1} - M^{-1}S(S'M^{-1}S)^{-1}S'M^{-1}\right] \begin{pmatrix} \theta_\beta R_\beta(t, t(n)) \\ \vdots \\ \theta_\beta R_\beta(t, t(n)) \end{pmatrix}.$$

Gu's (1992c) Theorem 3.1 extends directly to the SS-ANOVA model considered here; see Wang (1994). We state the result:

THEOREM. *Let ζ, η be any one of $\tau_\nu\phi_\nu(t)$, $\tau_\mu\phi_\mu(t)$, $\sqrt{\theta_\beta} Z_\beta(t)$ and $\sqrt{\theta_\alpha} Z_\alpha(t)$ for arbitrary points s and t. The posterior density $\hat{\pi}(\zeta, \eta|y)$ is approximately Gaussian with mean and covariance given in* (A.1).

To compute the componentwise Bayesian confidence intervals in the non-Gaussian case we need to calculate the posterior covariances as in (A.1); W is taken as the converged value of W_- and λ and θ are taken as the converged estimates; $\sigma^2 = 1$ if there is no nuisance variance parameter and $b = \sigma^2/n\lambda$. The computational algorithm in RKPACK accommodates this calculation since

$$\text{(A.4)} \quad \begin{aligned} d &= (S'M^{-1}S)^{-1} S'M^{-1}y, \\ c &= M^{-1} - M^{-1}S(S'M^{-1}S)^{-1}S'M^{-1}y \end{aligned}$$

is the solution to the system $Mc + Sd = y$, $S'c = 0$, which is solved by RKPACK. Thus, by making the right substitutions in (A.4), that is, by

replacing y by

$$\begin{pmatrix} \theta_\beta R_\beta(t, t(1)) \\ \vdots \\ \theta_\beta R_\beta(t, t(n)) \end{pmatrix},$$

the numerical methods in RKPACK can be exploited to obtain $d_\beta(t)$ and $c_\beta(t)$ in the $W = I$ case. Let $Q_{W,\theta} = W^{1/2} Q_\theta W^{1/2}$, $S_W = W^{1/2} S$, $R_{W,\beta}(t, t(i)) = \sqrt{w_i} R_\beta(t, t(i))$ and $M_W = Q_{W,\theta} + n\lambda I$. We can then calculate $(S_W' M_W^{-1} S_W)^{-1}$, $d_\beta(t)$ and $c_{W,\beta}(t) = W^{-1/2} c_\beta(t)$ exactly the same way as in Gu and Wahba (1993b) by replacing R_β there by $R_{W,\beta}$. We then have $(S' M^{-1} S)^{-1} = (S_W' M_W^{-1} S_W)^{-1}$, $d_\beta(t)$ and $c_\beta(t) = W^{1/2} c_{W,\beta}(t)$.

B. RK's used in the WESDR example. In the analysis of the WESDR data, all of the predictor variables were considered as continuous variables on the real line, and each variable was rescaled to $[0, 1]$ by mapping the smallest and largest values to 0 and 1, respectively. Thus, $\mathscr{T}^{(\alpha)} = [0, 1]$, all α. The measures μ_α were all taken as Lebesgue measure on $[0, 1]$, $\mathscr{H}^{(\alpha)}$ was taken as the reproducing kernel space $\{g\colon g, g' \text{ abs. cont.}, \int_0^1 g(u)\,du = 0, \int_0^1 [g''(u)]^2\,du < \infty\}$, $\mathscr{H}_\pi^{(\alpha)}$ was taken as the one-dimensional space of multiples of $u - 1/2$ (i.e., linear functions averaging to 0) and $\mathscr{H}_s^{(\alpha)}$ was the subspace of $\mathscr{H}^{(\alpha)}$ of functions satisfying $g(0) - g(1) = 0$; $\int_0^1 [g''(u)]^2\,du$ is then a square norm on $\mathscr{H}_s^{(\alpha)}$. With this norm, the reproducing kernels for $\mathscr{H}_\pi^{(\alpha)}$ and $\mathscr{H}_s^{(\alpha)}$, respectively, are given by

$$\begin{aligned} R_{\mathscr{H}_\pi^{(\alpha)}}(u, v) &= (u - 1/2)(v - 1/2), \\ R_{\mathscr{H}_s^{(\alpha)}}(u, v) &= k_2(u) k_2(v) - k_4([u - v]), \end{aligned} \tag{B.1}$$

where $l! k_l(u)$ is the lth Bernoulli polynomial and $[x]$ is the fractional part of x; $\mathscr{H}_\pi^{(\alpha)}$ and $\mathscr{H}_s^{(\alpha)}$ will be orthogonal subspaces of $\mathscr{H}^{(\alpha)}$ if $\mathscr{H}^{(\alpha)}$ is endowed with the square norm $\|g\|^2 = [g(1) - g(0)]^2 + \int_0^1 [g''(u)]^2\,du$. Further details may be found in Wahba [(1990), Chapter 10]. In a preliminary conference proceedings study of non-Gaussian SS-ANOVA [Wahba, Gu, Wang and Chappell (1995)], a categorical variable was included in $\mathscr{H}^0$.

Acknowledgment. We thank Scot Moss for his assistance in obtaining the data and for helpful conversations.

REFERENCES

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

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

BREIMAN, L., FRIEDMAN, J., OLSHEN, R. and STONE, C. (1984). *Classification and Regression Trees*. Wadsworth, Belmont, CA.

CHAMBERS, J. and HASTIE, T. (1992). *Statistical Models in S*. Wadsworth and Brooks/Cole, Belmont, CA.

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

CHEN, Z. (1993). Fitting multivariate regression functions by interaction spline models. *J. Roy. Statist. Soc. Ser. B* **55** 473–491.

CHEN, Z., GU, C. and WAHBA, G. (1989). Comment on "Linear smoothers and additive models" by A. Buja, T. Hastie and R. Tibshirani. *Ann. Statist.* **17** 515–521.

CHENG, B. and TITTERINGTON, D. (1994). Neural networks: a review from a statistical perspective (with discussion). *Statist. Sci.* **9** 2–54.

COX, D. and CHANG, Y. (1990). Iterated state space algorithms and cross validation for generalized smoothing splines. Technical Report 49, Dept. Statistics, Univ. Illinois, Champaign.

COX, D., KOH, E., WAHBA, G. and YANDELL, B. (1988). Testing the (parametric) null model hypothesis in (semiparametric) partial and generalized spline models. *Ann. Statist.* **16** 113–119.

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.

EFRON, B. and STEIN, C. (1981). The jackknife estimate of variance. *Ann. Statist.* **9** 586–596.

EFRON, B. and TIBSHIRANI, R. (1993). *An Introduction to the Bootstrap*. Chapman and Hall, London.

EUBANK, R. (1988). *Spline Smoothing and Nonparametric Regression*. Dekker, New York.

FRIEDMAN, J. (1991). Multivariate adaptive regression splines. *Ann. Statist.* **19** 1–141.

FRIEDMAN, J. H. and STUETZLE, W. (1981). Projection pursuit regression. *J. Amer. Statist. Assoc.* **76** 817–823.

GEMAN, S., BIENENSTOCK, E. and DOURSAT, R. (1992). Neural networks and the bias/variance dilemma. *Neural Computation* **4** 1–58.

GIRARD, D. (1987). A fast "Monte Carlo cross-validation" procedure for large least squares problems with noisy data. Technical Report RR 687-M, IMAG, Grenoble, France.

GIRARD, D. (1989). A fast "Monte-Carlo cross-validation" procedure for large least squares problems with noisy data. *Numer. Math.* **56** 1–23.

GIRARD, D. (1991). Asymptotic optimality of the fast randomized versions of GCV and C_L in ridge regression and regularization. *Ann. Statist.* **19** 1950–1963.

GOLUB, G. and VON MATT, U. (1995). Generalized cross-validation in large scale problems. Technical Report, Scientific Computing/Computational Mathematics Program, Stanford Univ. To appear.

GREEN, P. and SILVERMAN, B. (1994). *Nonparametric Regression and Generalized Linear Models*. Chapman and Hall, London.

GREEN, P. and YANDELL, B. (1985). *Semi-Parametric Generalized Linear Models. Lecture Notes in Statist.* **32** 44–55. Springer, Berlin.

GU, C. (1989). RKPACK and its applications: fitting smoothing spline models. In *Proceedings of the Statistical Computing Section* 42–51. Amer. Statist. Assoc., Alexandria, VA. (Code available through `netlib`.)

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

GU, C. (1992a). Cross-validating non-Gaussian data. *Journal of Computational and Graphical Statistics* **1** 169–179.

GU, C. (1992b). Diagnostics for nonparametric regression models with additive terms. *J. Amer. Statist. Assoc.* **87** 1051–1057.

GU, C. (1992c). Penalized likelihood regression: a Bayesian analysis. *Statist. Sinica* **2** 255–264.

GU, C., BATES, D., 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.

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

GU, C. and WAHBA, G. (1991a). Comments on "Multivariate adaptive regression splines" by J. Friedman. *Ann. Statist.* **19** 115–123.

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

GU, C. and WAHBA, G. (1993a). Semiparametric analysis of variance 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." *Journal of Computational and Graphical Statistics* **2** 97–117.

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

HASTIE, T. and TIBSHIRANI, R. (1993). Varying-coefficient models. *J. Roy. Statist. Soc. Ser. B* **55** 757–796.

HUDSON, M. (1978). A natural identity for exponential families with applications in multiparameter estimation. *Ann. Statist.* **6** 473–484.

HUTCHINSON, M. (1984). A summary of some surface fitting and contouring programs for noisy data. Technical Report ACT 84/6, CSIRO Division of Mathematics and Statistics, Canberra.

HUTCHINSON, M. (1989). A stochastic estimator for the trace of the influence matrix for Laplacian smoothing splines. *Comm. Statist. Simulation Comput.* **18** 1059–1076.

HUTCHINSON, M. and GESSLER, P. (1994). Splines—more than just a smooth interpolator. *Geoderma* **62** 45–67.

KLEIN, B. E. K., DAVIS, M. D., SEGAL, P., LONG, J. A., HARRIS, W. A., HAUG, G. A., MAGLI, Y. and SYRJALA, S. (1984). Diabetic retinopathy: assessment of severity and progression. *Ophthalmology* **91** 10–17.

KLEIN, R., KLEIN, B. E. K., MOSS, S. E. and CRUICKSHANKS, K. J. (1994a). The relationship of hyperglycemia to long-term incidence and progression of diabetic retinopathy. *Archives of Internal Medicine* **154** 2169–2178.

KLEIN, R., KLEIN, B. E. K., MOSS, S. E. and CRUICKSHANKS, K. J. (1994b). The Wisconsin Epidemiologic Study of Diabetic Retinopathy. XIV. Ten year incidence and progression of diabetic retinopathy. *Archives of Ophthalmology* **112** 1217–1228.

KLEIN, R., KLEIN, B. E. K., MOSS, S. E., DAVIS, M. D. and DEMETS, D. L. (1984a). The Wisconsin Epidemiologic Study of Diabetic Retinopathy. II. Prevalence and risk of diabetic retinopathy when age at diagnosis is less than 30 years. *Archives of Ophthalmology* **102** 520–526.

KLEIN, R., KLEIN, B. E. K., MOSS, S. E., DAVIS, M. D. and DEMETS, D. L. (1984b). The Wisconsin Epidemiologic Study of Diabetic Retinopathy. III. Prevalence and risk of diabetic retinopathy when age at diagnosis is 30 or more years. *Archives of Ophthalmology* **102** 527–532.

KLEIN, R., KLEIN, B. E. K., MOSS, S. E., DAVIS, M. D. and DEMETS, D. L. (1988). Glycosylated hemoglobin predicts the incidence and progression of diabetic retinopathy. *Journal of the American Medical Association* **260** 2864–2871.

KLEIN, R., KLEIN, B. E. K., MOSS, S. E., DAVIS, M. D. and DEMETS, D. L. (1989a). Is blood pressure a predictor of the incidence or progression of diabetic retinopathy? *Archives of Internal Medicine* **149** 2427–2432.

KLEIN, R., KLEIN, B. E. K., MOSS, S. E., DAVIS, M. D. and DEMETS, D. L. (1989b). The Wisconsin Epidemiologic Study of Diabetic Retinopathy. IX. Four year incidence and progression of diabetic retinopathy when age at diagnosis is less than 30 years. *Archives of Ophthalmology* **107** 237–243.

KLEIN, R., KLEIN, B. E. K., MOSS, S. E., DAVIS, M. D. and DEMETS, D. L. (1989c). The Wisconsin Epidemiologic Study of Diabetic Retinopathy. X. Four year incidence and progression of diabetic retinopathy when age at dianosis is 30 or more years. *Archives of Ophthalmology* **107** 244–249.

KLEIN, R., KLEIN, B. E. K., MOSS, S. E., DEMETS, D. L., KAUFFMAN, I. and VOSS, P. S. (1984). Prevalence of diabetes mellitus in southern Wisconsin. *American Journal of Epidemiology* **119** 54–61.

LI, K. C. (1985). From Stein's unbiased risk estimates to the method of generalized cross-validation. *Ann. Statist.* **13** 1352–1377.

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

LIU, Y. (1993). Unbiased estimate of generalization error and model selection in neural network. Unpublished manuscript, Institute of Brain and Neural Systems, Dept. Physics, Brown Univ.

LUO, Z. and WAHBA G. (1995). Hybrid adaptive splines. Technical Report 947, Dept. Statistics, Univ. Wisconsin, Madison.

MALLOWS, C. (1973). Some comments on C_p. *Technometrics* **15** 661–675.

MCCULLAGH, P. and NELDER, J. (1989). *Generalized Linear Models*, 2nd ed. Chapman and Hall, London.

MOODY, J. (1991). The effective number of parameters: an analysis of generalization and regularization in nonlinear learning systems. In *Advances in Neural Information Processing Systems 4* (J. Moody, S. Hanson and R. Lippman, eds.) 847–854. Kaufmann, San Mateo, CA.

NELDER, J. and WEDDERBURN, R. (1972). Generalized linear models. *J. Roy. Statist. Soc. Ser. A* **35** 370–384.

NYCHKA, D. (1988). Bayesian confidence intervals for smoothing splines. *J. Amer. Statist. Assoc.* **83** 1134–1143.

NYCHKA, D. (1990). The average posterior variance of a smoothing spline and a consistent estimate of the average squared error. *Ann. Statist.* **18** 415–428.

NYCHKA, D., WAHBA, G., GOLDFARB, S. and PUGH, T. (1984). Cross-validated spline methods for the estimation of three dimensional tumor size distributions from observations on two dimensional cross sections. *J. Amer. Statist. Assoc.* **79** 832–846.

O'SULLIVAN, F. (1983). The analysis of some penalized likelihood estimation schemes. Ph.D. dissertation, Technical Report 726, Dept. Statistics, Univ. Wisconsin–Madison.

O'SULLIVAN, F. (1990). An iterative approach to two-dimensional Laplacian smoothing with application to image restoration. *J. Amer. Statist. Assoc.* **85** 213–219.

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.

RAGHAVAN, N. (1993). Bayesian inference in nonparametric logistic regression. Ph.D. dissertation Univ. Illinois, Urbana–Champaign.

RIPLEY, B. (1994). Neural networks and related methods for classification. *J. Roy. Statist. Soc. Ser. B* **56** 409–456.

ROOSEN, C. and HASTIE, T. (1994). Automatic smoothing spline projection pursuit. *Journal of Computational and Graphical Statistics* **3** 235–248.

SAS INSTITUTE (1989). *SAS/STAT User's Guide*, Version 6, 4th ed. SAS Institute, Inc., Cary, North Carolina.

SHIAU, J. J., WAHBA, G. and JOHNSON, D. (1986). Partial spline models for the inclusion of tropopause and frontal boundary information. *Journal of Atmospheric and Oceanic Technology* **3** 714–725.

STONE, C. (1994). The use of polynomial splines and their tensor products in multivariate function estimation (with discussion). *Ann. Statist.* **22** 118–184.

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

WAHBA, G. (1980). Spline bases, regularization, and generalized cross validation for solving approximation problems with large quantities of noisy data. In *Approximation Theory III* (W. Cheney, ed.) 905–912. Academic Press, New York.

WAHBA, G. (1981). Spline interpolation and smoothing on the sphere. *SIAM J. Sci. Statist. Comput.* **2** 5–16.

WAHBA, G. (1982). Erratum: spline interpolation and smoothing on the sphere. *SIAM J. Sci. Statist. Comput.* **3** 385–386.

WAHBA, G. (1983). Bayesian "confidence intervals" for the cross-validated smoothing spline. *J. Roy. Statist. Soc. Ser. B* **45** 133–150.

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

WAHBA, G. (1992). Multivariate function and operator estimation, based on smoothing splines and reproducing kernels. In *Nonlinear Modeling and Forecasting. Santa Fe Institute Studies in the Sciences of Complexity, Proceedings* (M. Casdagli and S. Eubank, eds.) **12** 95–112. Addison-Wesley, Reading, MA.

WAHBA, G. (1995). Generalization and regularization in nonlinear learning systems. In *Handbook of Brain Theory and Neural Networks* (M. Arbib, ed.) 426–430. MIT Press.

WAHBA, G., GU, C., WANG, Y. and CHAPPELL, R. (1995). Soft classification, a.k.a. risk estimation, via penalized log likelihood and smoothing spline analysis of variance. In *The Mathematics of Generalization. Santa Fe Institute Studies in the Sciences of Complexity, Proceedings* (D. Wolpert, ed.) **20** 329–360. Addison-Wesley, Reading, MA.

WAHBA, G., JOHNSON, D., GAO, F. and GONG, J. (1994). Adaptive tuning of numerical weather prediction models: randomized GCV in three and four dimensional data assimilation. *Monthly Weather Review* **123** 3358–3369.

WAHBA, G. and WENDELBERGER, J. (1980). Some new mathematical methods for variational objective analysis using splines and cross-validation. *Monthly Weather Review* **108** 1122–1145.

WANG, Y. (1994). Smoothing spline analysis of variance of data from exponential families. Ph.D. dissertation, Technical Report 928, Univ. Wisconsin–Madison.

WANG, Y. (1995). GRKPACK: fitting smoothing spline analysis of variance models to data from exponential families. Technical Report 942, Dept. Statistics, Univ. Wisconsin–Madison.

WANG, Y. and WAHBA, G. (1995). Bootstrap confidence intervals for smoothing splines and their comparison to Bayesian "confidence intervals." *J. Statist. Comput. Simulation.* **51** 263–280.

WANG, Y., WAHBA, G., CHAPPELL, R. and GU, C. (1995). Simulation studies of smoothing parameter estimates and Bayesian confidence intervals in Bernoulli SS-ANOVA models. *Comm. Statist. Simulation Comput.* To appear.

WEBER, R. and TALKNER, P. (1993). Some remarks on spatial correlation function models. *Monthly Weather Review* **121** 2611–2617.

WONG, W. (1992). Estimation of the loss of an estimate. Technical Report 356, Dept. Statistics, Univ. Chicago.

XIANG, D. and WAHBA, G. (1995). Testing the generalized linear model Null Hypothesis versus "smooth" alternatives. Technical Report 953, Dept. Statistics Univ. Wisconsin–Madison.

YANDELL, B. (1986). Algorithms for nonlinear generalized cross-validation. In *Computer Science and Statistics: 18th Symposium on the Interface* (T. Boardman, ed.). Amer. Statist. Assoc., Washington, DC.

GRACE WAHBA
DEPARTMENT OF STATISTICS
UNIVERSITY OF WISCONSIN
1210 WEST DAYTON STREET
MADISON, WISCONSIN 53706

CHONG GU
DEPARTMENT OF STATISTICS
PURDUE UNIVERSITY
MATH SCIENCES BUILDING
WEST LAFAYETTE, INDIANA 47907

YUEDONG WANG
DEPARTMENT OF BIOSTATISTICS
SCHOOL OF PUBLIC HEALTH
UNIVERSITY OF MICHIGAN
1420 WASHINGTON HEIGHTS
ANN ARBOR, MICHIGAN 48109

RONALD KLEIN, MD
BARBARA KLEIN, MD
DEPARTMENT OF OPHTHALMOLOGY
UNIVERSITY OF WISCONSIN
610 NORTH WALNUT STREET
MADISON, WISCONSIN 53705