DEPARTMENT OF STATISTICS
University of Wisconsin
1210 West Dayton St.
Madison, WI 53706

TECHNICAL REPORT NO. 940

December 24,1994

Smoothing Spline ANOVA for Exponential Families, with Application to the Wisconsin Epidemiological Study of Diabetic Retinopathy [1] [2]

Grace Wahba [3]
Department of Statistics, University of Wisconsin, Madison WI

Yuedong Wang [4]
Department of Biostatistics, University of Michigan Ann Arbor MI

Chong Gu [5]
Department of Statistics, Purdue University, West Lafayette IN

Ronald Klein, MD[6] Barbara Klein, MD[7]
Department of Ophthalmalogy, University of Wisconsin,Madison WI

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

[2]Corresponding author address: Prof. Grace Wahba, Department of Statistics, University of Wisconsin, 1210 W. Dayton St., Madison, WI 53706.

[3]Research supported in part by NIH Grant EY09946 and NSF Grant DMS9121003

[4]Research supported in part by NIH Grant EY09446, P60 DK20572 and P30 HD18258

[5]Research supported by DMS9301511

[6]Research supported by NIH Grant EY03083

[7]Research supported by NIH Grant EY03083

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

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

Abstract

Let $y_i, i = 1, \cdots n$ 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 $t = (t_1, \cdots, t_d) \in \mathcal{T}^{(1)} \otimes \cdots \otimes \mathcal{T}^{(d)} = \mathcal{T}$, the $\mathcal{T}^{(\alpha)}$ are measureable spaces of rather general form, and f is an unknown function on $\mathcal{T}$ with some assumed 'smoothness' properties. Given $\{y_i, t(i), i = 1, \cdots, n\}$, it is desired to estimate $f(t)$ for t in some region of interest contained in $\mathcal{T}$. We develop the fitting of smoothing spline ANOVA models to this data of the form $f(t) = C + \sum_\alpha f_\alpha(t_\alpha) + \sum_{\alpha<\beta} f_{\alpha\beta}(t_\alpha, t_\beta) + \cdots$. 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 $\mathcal{L}(y, f) + \sum_\alpha \lambda_\alpha J_\alpha(f_\alpha) + \sum_{\alpha<\beta} \lambda_{\alpha\beta} J_{\alpha\beta}(f_{\alpha\beta}) + \cdots$, where $\mathcal{L}(y, f)$ is the negative log likelihood of $y = (y_1, \cdots, y_n)'$ given f, the $J_\alpha, J_{\alpha\beta}, \cdots$ 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 $\lambda_\alpha, \lambda_{\alpha\beta} \cdots$, (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 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 continous variables on various domains, ordered categorical variables, and unordered categorical variables. A crucial feature is the availability of a set of methods for 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 $f(t), t = (t_1, \cdots, t_d)$ of d variables as $f(t) = C + \sum_\alpha f_\alpha(t_\alpha) + \sum_{\alpha<\beta} f_{\alpha\beta}(t_\alpha, t_\beta) + \cdots$, 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 $y_i, i = 1, \cdots, n$ 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 $\mathcal{L}(y, f) + \sum_\alpha \lambda_\alpha J_\alpha(f_\alpha) + \sum_{\alpha<\beta} \lambda_{\alpha\beta} J_{\alpha\beta}(f_{\alpha\beta}) + \cdots$, where $\mathcal{L}(y, f)$ is the negative log likelihood of $(y_1, \cdots, y_n)$ given f, the $J_\alpha, J_{\alpha\beta}, \cdots$ are quadratic penalty functionals, and the $\lambda_\alpha, \lambda_{\alpha\beta}, \cdots$ 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), the π method (Breiman 1991), tensor product regression splines (Stone 1994), 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 there, for a discussion of relationships between neural nets and statistical nonparametric regression methods. The popular additive models of Hastie and Tibshirani (1990), when fit 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(1989) and Green and Silverman(1994).

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), 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), 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,1992c), Xiang and Wahba (1994) 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-Liebler 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-Liebler 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 (1994b), 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 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 component-wise generalization of Gu (1992c) to the non-Gaussian component-wise SSANOVA 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 an hypothesis testing point of view, was given by Cox, Koh, Wahba and Yandell (1988) and by Xiang and Wahba (1994) 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 component-wise confidence intervals to eliminate terms was suggested in Gu and Wahba (1993b). Of course model selection from an 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 component-wise 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 component-wise exponential family case and a procedure for computing them described. In Section 5 we discuss model selection, and in Section 6 we carry out the WESDR data analysis. Section 7 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/under dispersion situations, see Wang (1994b) for details. We consider random variables y_i with density $h(y_i, f_i)$ of the form

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

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 $\{y_i, t(i), i = 1, \cdots, n\}$, 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

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

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 estimated f as the minimizer, in an appropriate function space (reproducing kernel Hilbert space, RKHS) of $\mathcal{L}(y, f) + \lambda J(f)$, where J is a roughness penalty. An SS-ANOVA model provides a decomposition of f of the form

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

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

$$y_i = f(t_1(i), \cdots, t_d(i)) + \epsilon_i, \quad i = 1, \cdots, n, \tag{2.4}$$

where $\epsilon = (\epsilon_1, \cdots, \epsilon_n)' \sim N(0, \sigma^2 I_{n\times n})$, $t_\alpha \in \mathcal{T}^{(\alpha)}$, where $\mathcal{T}^{(\alpha)}$ is a measurable space, $\alpha = 1, \cdots, d; (t_1, \cdots, t_d) = t \in \mathcal{T} = \mathcal{T}^{(1)} \otimes \cdots \otimes \mathcal{T}^{(d)}$, 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 $\mathcal{T}^{(\alpha)}$ and define the averaging operator $\mathcal{E}_\alpha$ on $\mathcal{T}$ by

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

Then the identity is decomposed as

$$I = \prod_\alpha (\mathcal{E}_\alpha + (I - \mathcal{E}_\alpha)) = \prod_\alpha \mathcal{E}_\alpha + \sum_\alpha (I - \mathcal{E}_\alpha) \prod_{\beta \neq \alpha} \mathcal{E}_\beta + \sum_{\alpha < \beta} (I - \mathcal{E}_\alpha)(I - \mathcal{E}_\beta) \prod_{\gamma \neq \alpha, \beta} \mathcal{E}_\gamma + \cdots + \prod_\alpha (I - \mathcal{E}_\alpha). \tag{2.6}$$

The components of this decomposition generate the ANOVA decomposition of f of the form (2.3) by $C = (\prod_\alpha \mathcal{E}_\alpha) f, f_\alpha = ((I - \mathcal{E}_\alpha) \prod_{\beta \neq \alpha} \mathcal{E}_\beta) f, f_{\alpha\beta} = ((I - \mathcal{E}_\alpha)(I - \mathcal{E}_\beta) \prod_{\gamma \neq \alpha, \beta} \mathcal{E}_\gamma) f$, and so forth. Efron and Stein (1981) discuss this kind of ANOVA decomposition in a different context.

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

$$\mathcal{H} = \prod_{j=1}^{d} (\{[1^{(\alpha)}]\} \oplus \{\mathcal{H}^{(\alpha)}\}) = [1] \oplus \sum_j \mathcal{H}^{(\alpha)} \oplus \sum_{\alpha < \beta} [\mathcal{H}^{(\alpha)} \otimes \mathcal{H}^{(\beta)}] \oplus \cdots, \tag{2.7}$$

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

Next, $\mathcal{H}^{(\alpha)}$ is decomposed into a parametric part and a smooth part, by letting $\mathcal{H}^{(\alpha)} = \mathcal{H}_\pi^{(\alpha)} \oplus \mathcal{H}_s^{(\alpha)}$, where $\mathcal{H}_\pi^{(\alpha)}$ is finite dimensional (the "parametric" part) and $\mathcal{H}_s^{(\alpha)}$ (the "smooth" part) is the orthocomplement of $\mathcal{H}_\pi^{(\alpha)}$ in $\mathcal{H}^{(\alpha)}$. Elements of $\mathcal{H}_\pi^{(\alpha)}$ 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 $\mathcal{H}_s^{(\alpha)}$. $[\mathcal{H}^{(\alpha)} \otimes \mathcal{H}^{(\beta)}]$ is now a direct sum of four orthogonal subspaces: $[\mathcal{H}^{(\alpha)} \otimes \mathcal{H}^{(\beta)}] = [\mathcal{H}_\pi^{(\alpha)} \otimes \mathcal{H}_\pi^{(\beta)}] \oplus [\mathcal{H}_\pi^{(\alpha)} \otimes \mathcal{H}_s^{(\beta)}] \oplus [\mathcal{H}_s^{(\alpha)} \otimes \mathcal{H}_\pi^{(\beta)}] \oplus [\mathcal{H}_s^{(\alpha)} \otimes \mathcal{H}_s^{(\beta)}]$. By convention the elements of the finite dimensional space $[\mathcal{H}_\pi^{(\alpha)} \otimes \mathcal{H}_\pi^{(\beta)}]$ will not be penalized. Continuing this way results in an orthogonal decomposition of $\mathcal{H}$ 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 $[\mathcal{H}_\pi^{(\alpha)} \otimes \mathcal{H}_s^{(\beta)}]$, smooth $\otimes$ parametric $[\mathcal{H}_s^{(\alpha)} \otimes \mathcal{H}_\pi^{(\beta)}]$ and smooth $\otimes$ smooth $[\mathcal{H}_s^{(\alpha)} \otimes \mathcal{H}_s^{(\beta)}]$ and similarly for the three and higher factor subspaces.

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

f in $\mathcal{M} = \mathcal{H}^0 \oplus \sum_\beta \mathcal{H}^\beta$ to minimize

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

where P^β is the orthogonal projector in $\mathcal{M}$ onto $\mathcal{H}^\beta$. The overparametrization $\lambda\theta_\beta^{-1} = \lambda_\beta$ is convenient for both expository and computational purposes, see Gu (1989), 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 $\mathcal{H}^0$ and the reproducing kernels (RK's) $\{R_\beta(s,t)\}$ for the $\mathcal{H}^\beta$. 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

$$\begin{aligned} \phi'(t) &= (\phi_1(t), \cdots, \phi_M(t)), \\ \xi'(t) &= (Q_\theta(t(1),t), \cdots, Q_\theta(t(n),t)). \end{aligned}$$

$c_{n\times 1}$ 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 $\mathcal{H}_\pi^{(\alpha)} \otimes \mathcal{H}_s^{(\beta)}$ is (in an obvious notation), $R_{\mathcal{H}_\pi^{(\alpha)}}(s_\alpha, t_\alpha) R_{\mathcal{H}_s^{(\beta)}}(s_\beta, t_\beta)$. 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 accomodated, see Gu and Wahba(1993a) and references cited there. Examples on the sphere can be found in Wahba(1981, 1982), Weber and Talkner (1993), and on a discrete index set, as might occur in large contingency tables, 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 $\mathcal{H}_\pi^{(\alpha)}$, see Wahba (1978) Section 3. Arbitrary functions including functions containing breaks and jumps and indicator functions may be added to H^0, see Shiau, Wahba and Johnson(1986), Wahba(1990), Wahba, Gu, Wang and Chappell (1994).

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

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

and

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

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

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

The properties of GCV and UBR estimates in the Gaussian case are well known, see Wahba(1990) and the references cited there, 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 distingishable'. 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 exponential exponential families.

The generalization of an ANOVA estimate for Gaussian data to the general exponential family is obtained by replacing $\frac{1}{n}\sum(y_i - f(t(i)))^2$ by $\frac{1}{n}\mathcal{L}(y,f)$ where $\mathcal{L}(y,f)$ is given by (2.2), and then solving the variational problem: find f in $\mathcal{M}$ to minimize

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

The minimizer of (3.1) is also known to have a representation of the form (2.10) (O'Sullivan *et al*, 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) can not 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 $u_i = -dl_i/df_i$, $u' = (u_1,\cdots,u_n)$, $w_i = -d^2 l_i/df_i^2$, $W = \text{diag}(w_1,\cdots,w_n)$, and $S = (\phi(t(1)),\cdots,\phi(t(n)))'$. 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 $(f_1, \cdots, f_n)'$. 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

$$\begin{aligned} (W\_Q_\theta + n\lambda I)c + W\_Sd &= (W\_f\_ - u\_), \\ S'c &= 0. \end{aligned} \tag{3.4}$$

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

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

where $\tilde{y}_i = f_{i-} - u_{i-}/w_{i-}$ 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 (1994b) for further details. Wang (1994b) 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*

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

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

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

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

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

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, $V(\lambda)$ evaluated, a new λ be chosen, the new $V(\lambda)$ be valuated 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 $u_i^2/w_i = (y_i - Ey_i)^2/var(y_i)$. As a result, Gu (1992a) suggested that V be replaced by U with $\sigma^2 = 1$, giving the U criteria

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

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 $\mathcal{T}$. We define the symmetrized Kullback-Liebler distance $SKL_\nu(f, f_{\lambda,\theta})$ with respect to ν as $SKL_\nu(f, f_{\lambda,\theta}) = \frac{1}{2}[KL_\nu(f, f_{\lambda,\theta}) + KL_\nu(f_{\lambda,\theta}, f)]$ where the KL distance $KL_\nu(f, f_{\lambda,\theta})$ in the exponential family case of (2.1) is given by $KL_\nu(f, f_{\lambda,\theta}) = \int[\{\mu(t)f(t) - b(f(t))\} - \{\mu(t)f_{\lambda,\theta}(t) - b(f_{\lambda,\theta}(t))\}]d\nu(t)$, 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 (1994b) 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 reasonable estimate of the minimizer of SKL_ν with respect to λ, θ with ν the sample design measure.

By comparing (2.11, 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 (1994b), and further reported in Wang, Wahba, Chappell and Gu, in preparation. In that study, the iterated UBR outperformed the iterated GCV criteria in terms of minimizing $SKL(f, f_{\hat{\lambda},\hat{\theta}})$, 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 suggested there, for functions in an appropriate functon space, and λ chosen according to a predictive mean square criteria. 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(1994) and others studied the properties of these intervals. Gu and Wahba (1993b) extended these confidence intervals component-wise to the 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 component wise 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 $\mathcal{M} = \mathcal{H}^0 \oplus \sum_{\beta=1}^q \mathcal{H}^\beta$

be the model space as before, with $\mathcal{H}^0 = \text{span}\{\phi_1, \cdots, \phi_M\}$, $R_\beta(s,t)$ the RK for $\mathcal{H}^\beta$ and $Q_\theta(s,t) = \sum_{\beta=1}^{q} \theta_\beta R_\beta(s,t)$. Define the stochastic process $X_\xi(t), t \in \mathcal{T}$ by

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

where $\tau = (\tau_1, \cdots, \tau_M)' \sim N(0, \xi I)$, 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

$$y_i = X_\xi(t(i)) + \epsilon_i, \ i = 1, \cdots, n, \ \ (\epsilon_1, \cdots, \epsilon_n)' \sim N(0, \sigma^2 W^{-1}) \tag{4.2}$$

with W positive definite and known. Let $n\lambda = \sigma^2/b$. Following Gu(1992b), Gu and Wahba (1993b) and using (1.5.11) and (1.5.12) of Wahba(1990) we have that (for each $t \in \mathcal{T}$), $f_{\lambda,\theta}(t) = \lim_{\xi \to \infty} E(X_\xi(t)|y)$, where $f_{\lambda,\theta}(\cdot)$ is the minimizer in $\mathcal{M}$ 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 component-wise confidence intervals will be used to aid in model selection in the data analysis below.

Gu (1992c) considers the univariate case where $\mathcal{L}(y, f)$ 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 $\lim_{\xi \to \infty} X_\xi(t), t \in \mathcal{T}$. He shows, that, upon letting $n\lambda = \sigma^2/b$, the conditional density of $f(t)|y$ is approximately Gaussian with conditional mean $f_\lambda(t)$, where f_λ is the minimizer of (the one- dimensional version of) (3.1) and the conditional covariance is computed as though the pseudo-data at convergence had the (prior distribution) Gaussian with mean $(f(t(1)), f(t(2))), \cdots, f(t(n)))'$ and covariance the converged value $\sigma^2 W^{-1}$ 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 a reasonable 'across-the-function properties. Gu's argument extends word for word to the multicomponent smoothing spline ANOVA case, see Wang(1994b), resulting in conditional covariances for the components of the model. The result is stated in Appendix A.

To compute the approximate component-wise 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 component-wise 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 dicussed here for choosing λ and θ. The code may also be used to compute the confidence intervals component-wise 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).

Results of a simulation study of the overall and component-wise Bayesian confidence intervals with Bernoulli data may be found in Wang (1994b), 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 component-wise 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 below. 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 $\mathcal{M}$.

Recently Ragahavan (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 $\mathcal{M}$, 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 the minimum could be compared for different models. However, for nested models, this is not quite 'fair', since $min_{\lambda_1,\cdots,\lambda_p} U(\lambda_1, \cdots, \lambda_p) \leq min_{\lambda_1,\cdots,\lambda_{p-1}} U(\lambda_1, \cdots, \lambda_{p-1}, \infty)$: 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 $KL_\nu(f, f_{\hat{\lambda},\hat{\theta}})$ 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 $\hat{KL}_\nu = -\frac{1}{\#\ in\ test\ set} \sum_{j\in testset}[y_j f_{\hat{\lambda},\hat{\theta}}(t(j)) - b(f_{\hat{\lambda},\hat{\theta}}(t(j)))]$ 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 (1994). 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 k th 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 re-estimation. 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 1994b), and some reasonable results with small data sets ($n \sim 200$) and a main effects model were obtained in Wang (1994b). 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 wether 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 1980-82, then again in 1984-86 and 1990-92. Detailed descriptions of the data have been given by Klein *et al* (1988) and Klein *et al* (1989c) and references there. 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 follow-up 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 reported in Wang(1994b), 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 reported in the various WESDR manuscripts. See Klein *et al*, 1984b,c,d, 1989a,b, 1994 and *in press*. 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 diabeties at baseline (`agd` + `dur` = `agb`)
4. `gly`: glycosylated hemoglobin, a measure of hyperglycemia, %
5. `bmi`: body mass index-weight in kg /(height in m)^2

At the baseline and follow-up 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 *et al* 1989a,b. At baseline and the 4-year follow-up, 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 *et al*, 1984a). 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 non-proliferative retinopathy (scores of 51/51 or better) at the start, and no missing data for the variables we studied. This group has been called the 'Younger Onset Progression' group, and was analyzed in Klein *et al* (1988). [1] The remainder of the 891 subjects either had proliferative retinopathy at the start, or, had missing data.

Klein *et al* (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, Inc. 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 wether there was an effect of age over and above that explained by `dur` it was decided to fit a model using `agd, dur, gly, 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 $f(\mathtt{agd},\mathtt{dur},\mathtt{gly},\mathtt{bmi}) = f_1(\mathtt{agd}) + \mathtt{f}_2(\mathtt{dur}) + \mathtt{a}_3 \cdot \mathtt{gly} + \mathtt{f}_4(\mathtt{bmi}) + \mathtt{f}_{14}(\mathtt{agd},\mathtt{bmi})$ was fitted and the Bayesian confidence intervals computed for f. The components for `agd` were not obviously negligible at the fitting stage. However, plots of cross sections of f *vs* `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. Analysis of all three models are in Wang (1994b).

[1] The sample size differs slightly from Klein *et al* (1988) due to different missing data patterns.

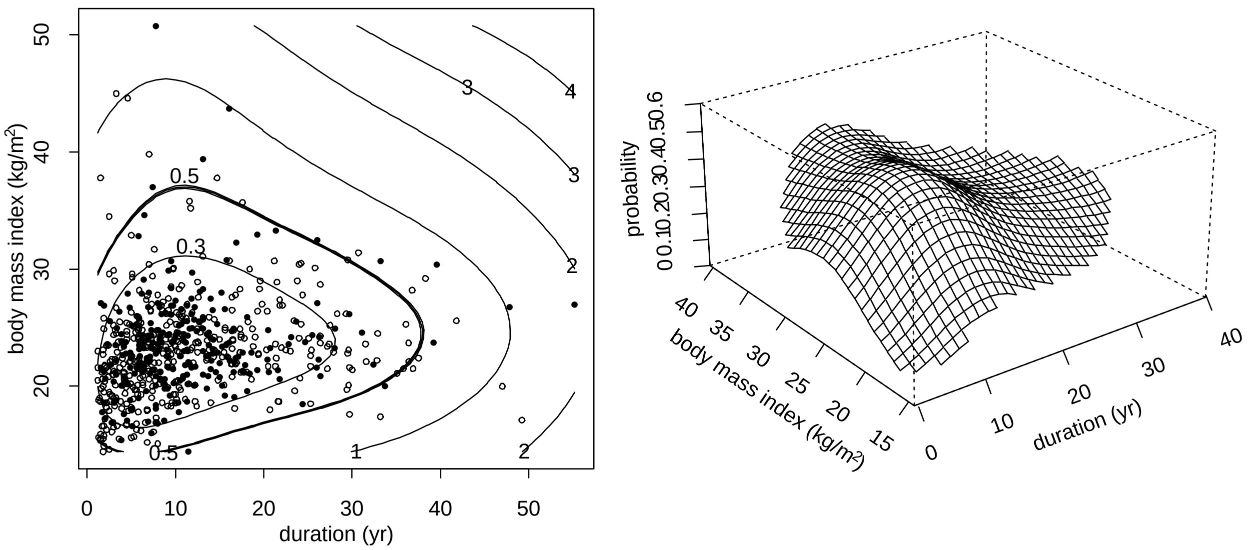

Figure 6.1: Left: data and contours of constant posterior standard deviation. Right: estimated probability of progression as a function of duration and body mass index for glycosylated hemoglobin fixed at its median.

The left panel in Figure 6.1 gives a scatterplot of `dur` *vs* `bmi`, with the solid circles representing patients with 4-year progression of retinopathy 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 .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 $p(\texttt{dur}, \texttt{bmi}, \texttt{gly})$ 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 .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 6.2 and 6.3 give slices of $p(\texttt{dur}, \texttt{bmi}, \texttt{gly})$, with the cross sections of Figure 6.2 plotted as a function of `dur` for four levels of `bmi` and `gly`, and the cross sections of Figure 6.3 plotted as a function of `bmi` for four levels of `dur` and `gly`. These levels are at the 12.5 th, 37.5 th 62.5 th and 87.5 th percentiles of the `bmi` and `gly` values. Figure 6.4 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 6.5 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. As previously reported(Klein *et al* 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 4-year progression of retinopathy increases with increasing body mass index at baseline until a value of about 25 kg/m^2, after which

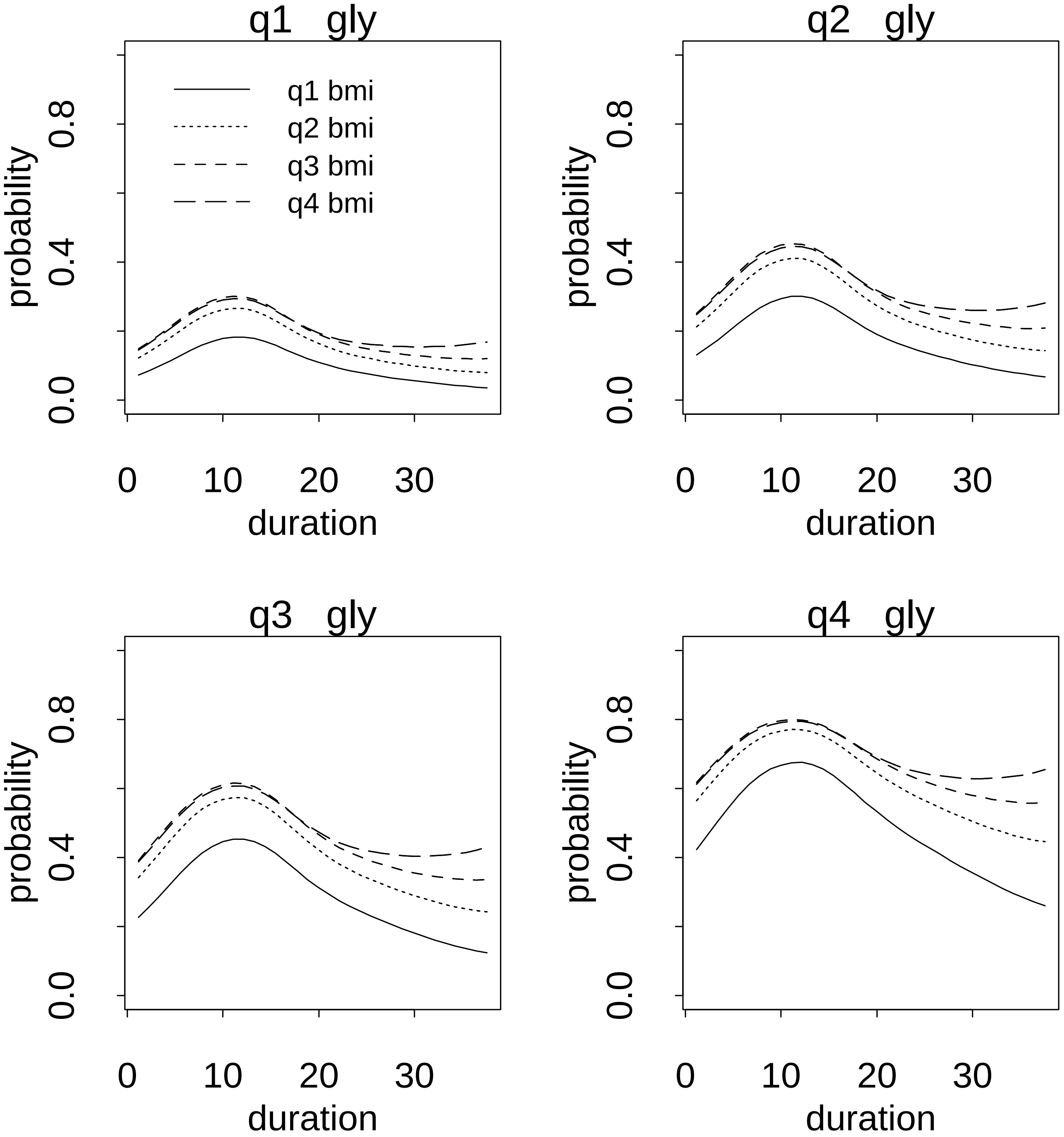

Figure 6.2: Estimated probability of progression as a function of `dur` for four levels of `bmi` by four levels of `gly`.

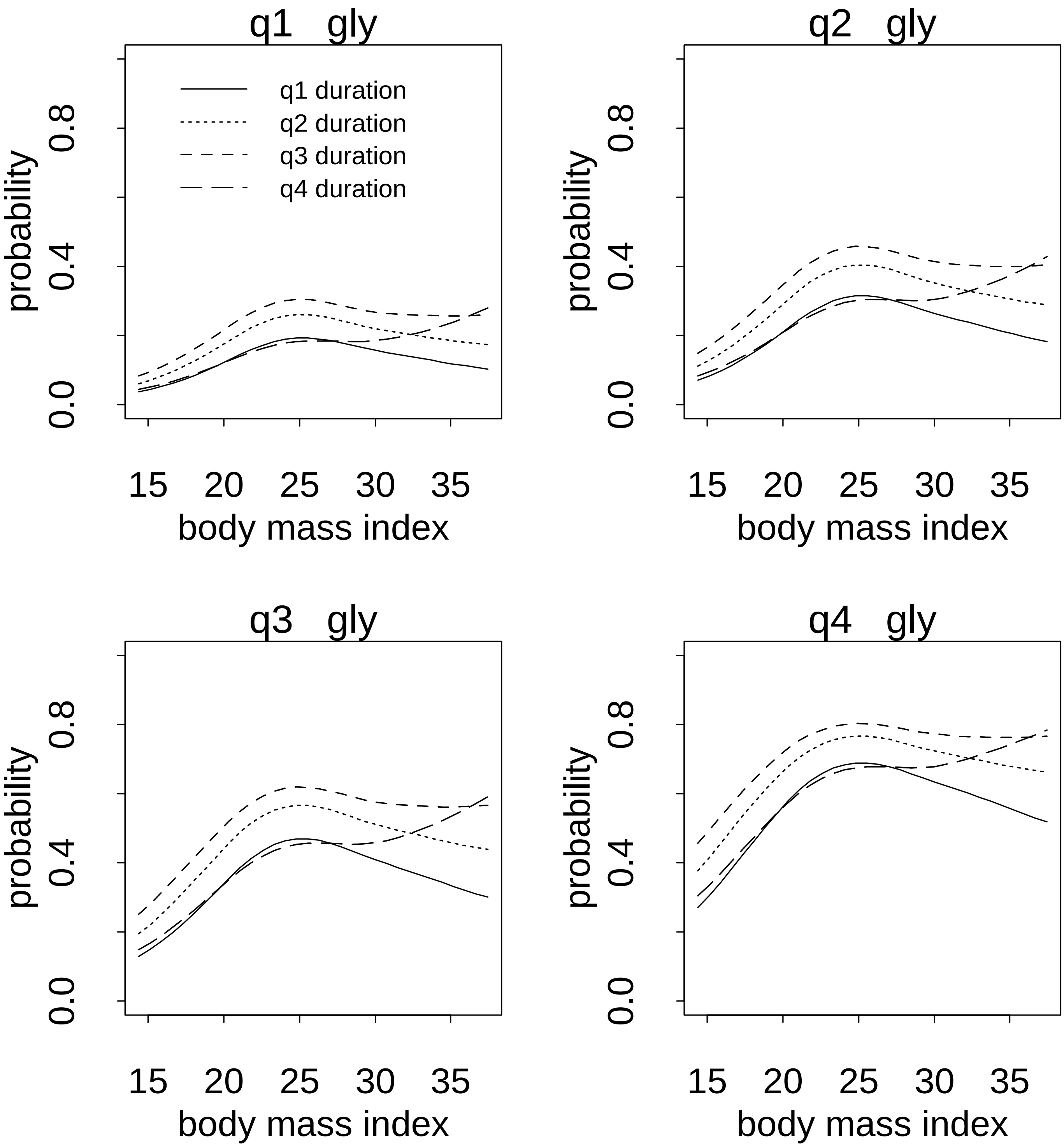

Figure 6.3: Estimated probability of progression as a function of `bmi` for four levels of `dur` by four levels of `gly`.

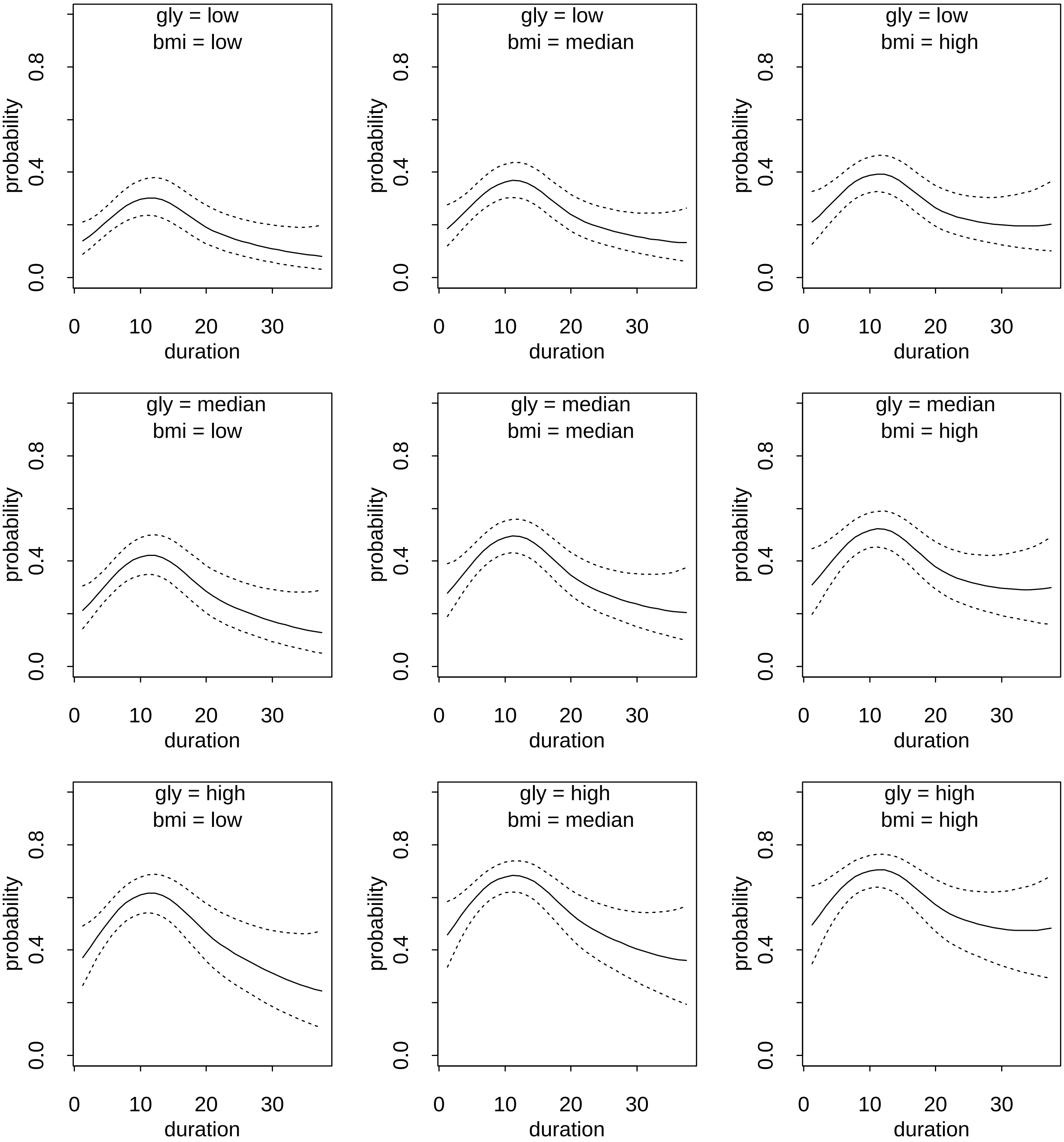

Figure 6.4: Estimated probability of progression and Bayesian confidence intervals, as a function of `dur` for three levels of `bmi` by three levels of `gly`.

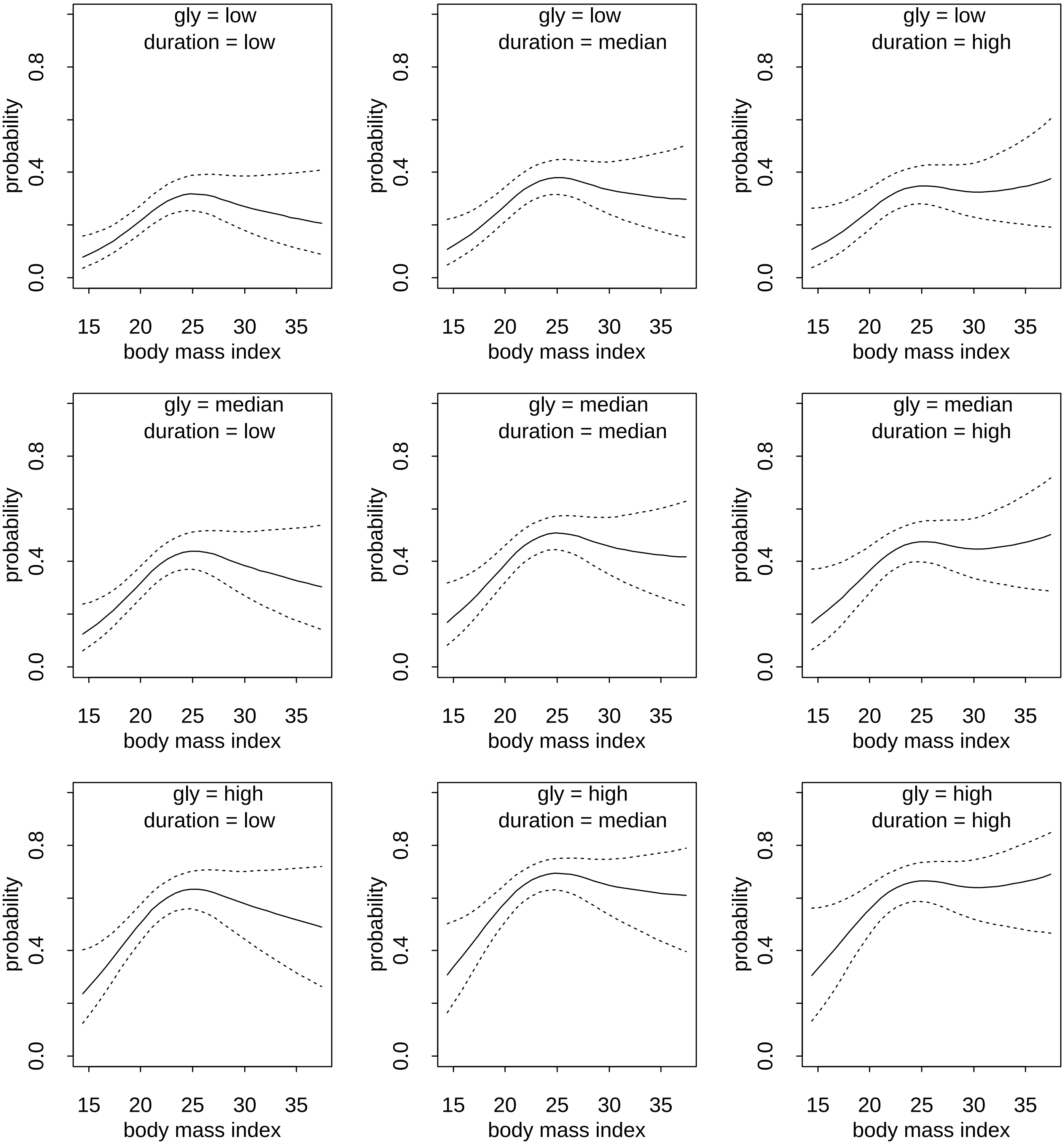

Figure 6.5: Estimated probability of progression and Bayesian confidence intervals, as a function of `bmi` for three levels of `dur` by three levels of `gly`.

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.

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 examined at the 4-year follow-up due to death than people with long duration of diabetes whose retinopathy did not progress).

7 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 non-trivial to implement.

A Details of Bayesian confidence intervals

The calculation of posterior means and covariances for the components of the model (4.1,4.2) 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 $M = Q_\theta + n\lambda W^{-1}$, $g_{0,\nu}(t) = \tau_\nu \phi_\nu(t)$, $g_\beta(t) = b^{\frac{1}{2}}\sqrt{\theta_\beta} Z_\beta(t)$, $\nu = 1, \cdots, M$, $\beta = 1, \cdots, q$, then

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

$$\frac{1}{b}\mathrm{Cov}(g_\gamma(s), g_\beta(t)|y) = -\sum_{i=1}^{n} c_{i,\gamma}(s)\theta_\beta R_\beta(t, t(i)), \tag{A.1}$$

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

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

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

Gu's (1993c) Theorem 3.1 extends directly to the SS-ANOVA model considered here, see Wang(1994b). 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 component-wise 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 accomodates this calculation since

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

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), *i. e.* 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 rescaled to $[0, 1]$ by mapping the smallest and largest values to 0 and 1 respectively. Thus, $\mathcal{T}^{(\alpha)} = [0, 1]$, all α. The measures μ_α were all taken as Lebesgue measure on $[0, 1]$, $\mathcal{H}^{(\alpha)}$ was taken as the reproducing kernel space $\{g : g, g'\, abs.cont., \int_0^1 g(u)du = 0, \int_0^1 [g''(u)]^2 du < \infty\}$. $\mathcal{H}_\pi^{(\alpha)}$ was taken as the one dimensional space of multiples of $u - 1/2$, (that is, linear functions averaging to 0), and $\mathcal{H}_s^{(\alpha)}$ was the subspace of $\mathcal{H}^{(\alpha)}$ of functions satisfying $g(0) - g(1) = 0$. $\int_0^1 [g''(u)]^2 du$ is then a square norm on $\mathcal{H}_s^{(\alpha)}$. With this norm, the reproducing kernels for $\mathcal{H}_\pi^{(\alpha)}$ and $\mathcal{H}_s^{(\alpha)}$ respectively are given by

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

where $\ell!k_\ell(u)$ is the ℓth Bernoulli polynomial, and $[x]$ is the fractional part of x. $\mathcal{H}_\pi^{(\alpha)}$ and $\mathcal{H}_s^{(\alpha)}$ will be orthogonal subspaces of $\mathcal{H}^{(\alpha)}$ if $\mathcal{H}^{(\alpha)}$ 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, 1994), a categorical variable was included in H^0.

References

Breiman, L. (1991), 'The π method for estimating multivariate functions from noisy data', *Technometrics* **3**, 125–143.

Breiman, L., Friedman, J., Olshen, R. & Stone, C. (1984), *Classification and Regression Trees*, Wadsworth.

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. Stat. Soc. B* **55**, 473–491.

Chen, Z., Gu, C. & Wahba, G. (1989), 'Comments to 'Linear Smoothers and Additive Models', by Buja, Hastie and Tibshirani', *Ann. Statist.* **17**, 515–521.

Cheng, B. & Titterington, D. (1994), 'Neural networks: A review from a statistical perspective', *Statistical Science* **9**, 2–54.

Cox, D. & Chang, Y. (1990), Iterated state space algorithms and cross validation for generalized smoothing splines, Technical Report 49, University of Illinois, Dept. of Statistics, Champaign, IL.

Cox, D., Koh, E., Wahba, G. & Yandell, B. (1988), 'Testing the (parametric) null model hypothesis in (semiparametric) partial and generalized spline models', *Ann. Statist.* **16**, 113–119.

Craven, P. & 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. & Stein, C. (1981), 'The jackknife estimate of variance', *Ann. Statist.* **9**, 586–596.

Efron, B. & Tibshirani, R. (1993), *An Introduction to the Bootstrap*, Chapman and Hall.

Eubank, R. (1989), *Spline Smoothing and Nonparametric Regression*, Marcel Dekker.

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

Friedman, J. H. & Stuetzle, W. (1981), 'Projection pursuit regression', *J. Amer. Statist. Assoc.* **76**, 817–823.

Geman, S., Bienenstock, E. & Doursat, R. (1992), 'Neural networks and the bias/variance dilemma', *Neural Computation* **4**, 1–58.

Green, P. & Silverman, B. (1994), *Nonparametric Regression and Generalized Linear Models*, Chapman and Hall.

Green, P. & Yandell, B. (1985), Semi-parametric generalized linear models, *in* R. Gilchrist, ed., 'Lecture Notes in Statistics, Vol. 32', Springer, pp. 44–55.

Gu, C. (1989), RKPACK and its applications: fitting smoothing spline models, *in* 'Proceedings of the Statistical Computing Section', American Statistical Association, pp. 42–51. Code available thru `netlib`.

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

Gu, C. (1992*a*), 'Cross-validating non-Gaussian data', *J. Comput. Graph. Stats.* **1**, 169–179.

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

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

Gu, C. & Wahba, G. (1991*a*), 'Comments to 'Multivariate Adaptive Regression Splines', by J. Friedman', *Ann. Statist.* **19**, 115–123.

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

Gu, C. & Wahba, G. (1993*a*), 'Semiparametric analysis of variance with tensor product thin plate splines', *J. Royal Statistical Soc. Ser. B* **55**, 353–368.

Gu, C. & Wahba, G. (1993*b*), 'Smoothing spline ANOVA with component-wise Bayesian "confidence intervals"', *J. Computational and Graphical Statistics* **2**, 97–117.

Gu, C., Bates, D., Chen, Z. & Wahba, G. (1989), 'The computation of GCV functions through householder tridiagonalization with application to the fitting of interaction spline models', *SIAM J. Matrix Anal.* **10**, 457–480.

Hastie, T. & Tibshirani, R. (1990), *Generalized additive models*, Chapman and Hall.

Hastie, T. & Tibshirani, R. (1993), 'Varying-coefficient models', *J. Roy. Statist. Soc. B* **55**, 757–796.

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

Klein, B. E. K., Davis, M. D., Segal, P., Long, J. A., Harris, W. A., Haug, G. A., Magli, Y. & Syrjala, S. (1984*a*), 'Diabetic retinopathy: Assessment of severity and progression', *Ophthmlmology* **91**, 10–17.

Klein, R., Klein, B. E. K. & Moss, S. E. Cruickshanks, K. J. (in press), 'The relationship of hyperglycemia to long-term incidence and progression of diabetic retinopathy', *Arch. Intern. Med.* **xxx**, xxx–xxx.

Klein, R., Klein, B. E. K., Moss, S. E. & Cruickshanks, K. J. (1994), 'The Wisconsin Epidemiologic Study of Diabetic Retinopathy. XIV. Ten year incidence and progression of diabetic retinopathy.', *Arch. Ophthalmol.* **112**, 1217–1228.

Klein, R., Klein, B. E. K., Moss, S. E., Davis, M. D. & DeMets, D. L. (1984*b*), 'The Wisconsin Epidemiologic Study of Diabetic Retinopathy. II. Prevalence and risk of diabetic retinopathy when age at diagnosis is less than 30 years', *Arch. Ophthalmol.* **102**, 520–526.

Klein, R., Klein, B. E. K., Moss, S. E., Davis, M. D. & DeMets, D. L. (1984*c*), 'The Wisconsin Epidemiologic Study of Diabetic Retinopathy. III. Prevalence and risk of diabetic retinopathy when age at diagnosis is 30 or more years', *Arch. Ophthalmol.* **102**, 527–532.

Klein, R., Klein, B. E. K., Moss, S. E., Davis, M. D. & 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. & DeMets, D. L. (1989*a*), 'Is blood pressure a predictor of the incidence or progression of diabetic retinopathy', *Arch. Intern. Med.* **149**, 2427–2432.

Klein, R., Klein, B. E. K., Moss, S. E., Davis, M. D. & DeMets, D. L. (1989*b*), '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', *Arch. Ophthalmol.* **107**, 237–243.

Klein, R., Klein, B. E. K., Moss, S. E., Davis, M. D. & DeMets, D. L. (1989*c*), 'The Wisconsin Epidemiologic Study of Diabetic Retinopathy. X. Four incidence and progression of diabetic retinopathy when age at diagnosis is 30 or more years', *Arch. Ophthalmol.* **107**, 244–249.

Klein, R., Klein, B. E. K., Moss, S. E., DeMets, D. L., Kauffman, I. & Voss, P. S. (1984*d*), 'Prevalence of diabetes mellitus in southern Wisconsin', *Am. J. Epidemiol* **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, manuscript, Department of Physics, Institute of Brain and Neural Systems, Brown University.

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

McCullagh, P. & Nelder, J. (1989), *Generalized Linear Models, Second Edition*, Chapman and Hall.

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

Nelder, J. & Wedderburn, R. (1972), 'Generalized linear models', *J. Roy. Stat. Soc.* **135**, 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.

O'Sullivan, F. (1983), The analysis of some penalized likelihood estimation schemes, PhD thesis, Dept. of Statistics, University of Wisconsin, Madison, WI. Technical Report 726.

O'Sullivan, F., Yandell, B. & 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, PhD thesis, University of Illinois, Urbana-Champaign.

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

Roosen, C. & Hastie, T. (1994), 'Automatic smoothing spline projection pursuit', *J. Comp. Graph. Statist* **3**, 235–248.

SAS Institute, I. (1989), 'SAS/STAT user's guide', SAS Institute, Inc. Version 6, Fourth Edition.

Shiau, J. J., Wahba, G. & Johnson, D. (1986), 'Partial spline models for the inclusion of tropopause and frontal boundary information', *J. Atmos. Ocean Tech.* **3**, 714–725.

Stone, C. (1994), 'The use of polynomial splines anad their tensor products in multivariate function estimation, with discussion', *Annals of Statistics* **22**, 118–184.

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

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

Wahba, G. (1982), 'Erratum: Spline interpolation and smoothing on the sphere', *SIAM J. Sci. Stat. Comput.* **3**, 385–386.

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

Wahba, G. (1990), *Spline Models for Observational Data*, SIAM. CBMS-NSF Regional Conference Series in Applied Mathematics, v. 59.

Wahba, G. (1992), Multivariate function and operator estimation, based on smoothing splines and reproducing kernels, *in* M. Casdagli & S. Eubank, eds, 'Nonlinear Modeling and Forecasting, SFI Studies in the Sciences of Complexity, Proc. Vol XII', Addison-Wesley, pp. 95–112.

Wahba, G. (1995), Generalization and regularization in nonlinear learning systems, *in* M. Arbib, ed., 'Handbook of Brain Theory and Neural Networks', MIT Press, pp. xxx–xxx.

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

Wahba, G., Gu, C., Wang, Y. & Chappell, R. (1994), Soft classification, a. k. a. risk estimation, via penalized log likelihood and smoothing spline analysis of variance, *in* D. Wolpert, ed., 'The Mathematics of Generalization, Santa Fe Institute Studies in the Sciences of Complexity, Proc. Vol. XX', Addison-Wesley, Reading, MA, pp. 329–360.

Wang, Y. (1994), Smoothing spline analysis of Variance of Data from Exponential Families, PhD thesis, TR 928, University of Wisconsin-Madison, Madison, WI.

Wang, Y. (1995), GRKPACK: Fitting smoothing spline analysis of variance models to data from exponential families, Technical Report in preparation, Dept. of Statistics, University of Wisconsin, Madison, WI.

Wang, Y. & Wahba, G. (1994), Bootstrap confidence intervals for smoothing splines and their comparison to Bayesian 'confidence intervals', Technical Report 913, Dept. of Statistics, University of Wisconsin, Madison, WI, to appear, *J. Stat. Comp. Sim.*

Wang, Y., Wahba, G., Chappell, R. & Gu, C. (1995), Simulation studies of smoothing parameter estimates and Bayesian confidence intervals in Bernoulli SS-ANOVA models., Technical Report in preparation, Dept. of Statistics, University of Wisconsin, Madison, WI.

Weber, R. & Talkner, P. (1993), 'Some remarks on spatial correlation function models', *Mon. Wea. Rev.* **121**, 2611–2617.

Wong, W. (1992), Estimation of the loss of an estimate, Technical Report 356, Dept. of Statistics, University of Chicago, Chicago, Il.

Xiang, D. & Wahba, G. (1994), A generalized approximate cross validation for smoothing splines with non-gaussian data, Technical Report 930, Dept. of Statistics, University of Wisconsin, Madison, WI 53706.

Yandell, B. (1986), Algorithms for nonlinear generalized cross-validation, *in* T. Boardman, ed., 'Computer Science and Statistics: 18th Symposium on the Interface', American Statistical Association, Washington, DC.

Grace Wahba
Department of Statistics
University of Wisconsin
1210 W. Dayton St.
Madison WI 53706
`wahba@stat.wisc.edu`

Yuedong Wang
Department of Biostatistics
School of Public Health
University of Michigan
1420 Washington Heights
Ann Arbor, MI 48109
`yuedong@atlas.biostat.med.umich.edu`

Chong Gu
Department of Statistics
Purdue University
Math Sciences Building
West Lafayette, IN 47907
`chong@pop.stat.purdue.edu`

Ronald Klein, MD
Department of Ophthalmology
University of Wisconsin
610 N. Walnut St.
Madison, WI 53705
`kleinr%ophthc.decnet@biost.biostat.wisc.edu`

Barbara Klein, MD
Department of Ophthalmology
University of Wisconsin
610 N. Walnut St.
Madison, WI 53705
`kleinb%ophthc.decnet@biost.biostat.wisc.edu`