The Annals of Statistics
2000, Vol. 28, No. 6, 1570–1600

SMOOTHING SPLINE ANOVA MODELS FOR LARGE DATA SETS WITH BERNOULLI OBSERVATIONS AND THE RANDOMIZED GACV

By Xiwu Lin,[1] Grace Wahba,[1] Dong Xiang,[2] Fangyu Gao,[1]
Ronald Klein, MD,[3] Barbara Klein, MD[3]

*Cendant Corporation, University of Wisconsin, SAS Institute,
University of Wisconsin, University of Wisconsin and
University of Wisconsin*

We propose the randomized Generalized Approximate Cross Validation ($ranGACV$) method for choosing multiple smoothing parameters in penalized likelihood estimates for Bernoulli data. The method is intended for application with penalized likelihood smoothing spline ANOVA models. In addition we propose a class of approximate numerical methods for solving the penalized likelihood variational problem which, in conjunction with the $ranGACV$ method allows the application of smoothing spline ANOVA models with Bernoulli data to much larger data sets than previously possible. These methods are based on choosing an approximating subset of the natural (representer) basis functions for the variational problem. Simulation studies with synthetic data, including synthetic data mimicking demographic risk factor data sets is used to examine the properties of the method and to compare the approach with the GRKPACK code of Wang (1997c). Bayesian "confidence intervals" are obtained for the fits and are shown in the simulation studies to have the "across the function" property usually claimed for these confidence intervals. Finally the method is applied to an observational data set from the Beaver Dam Eye study, with scientifically interesting results.

1. Introduction.

1.1. *Overview.* Smoothing spline ANOVA (SS-ANOVA) models have been shown to provide a large, flexible class of methods for non and semiparametric statistical model building and supervised machine learning from databases. These models have a number of advantages, including the facts that they reduce to commonly used parametric models when the data so warrant; the

Received September 1998; revised August 2000.

[1] Supported in part by NIH Grant EY09946 and NSF Grant DMS-91-21003.

[2] Supported by NSF Grant DMS-93-01511. Supported in part by NIH Grant EY09946 and NSF Grant DMS-91-21003.

[3] Supported by NIH Grant EY06594.

AMS 2000 *subject classifications*. Primary 62G07, 92C60, 68T05, 65D07, 65D10, 62A99, 62J07; secondary 41A63, 41A15, 62G07, 62M30, 65D15, 92H25, 49M15.

Key words and phrases. Smoothing spline ANOVA, nonparametric regression, exponential families, risk factor estimation, degrees of freedom, representers, reproducing kernel Hilbert spaces, penalized likelihood.

results are generally interpretable, reasonable accuracy statements concerning the models are available; and mixtures of continuous, ordered discrete and unordered discrete variables can be accomodated. The main drawback of these models as developed to date is that they are highly computationally intensive, especially with non-Gaussian data, where the fits are no longer linear in the observations. In that case they become infeasible for sample sizes of the order of a few thousand or so. The increasing availability, and desire to exploit large data bases to model and understand the relationships between predictor variables and response variables via flexible methods, makes it desirable to develop methods which will allow these models to be applied to much larger data sets. Fitting these models with large sets of non-Gaussian data provides new theoretical and computational challenges.

In this paper we contribute to these techniques by providing an overall method which allows the building of SS-ANOVA models on data from general exponential families with no nuisance parameter, on much larger data sets than previously possible. The primary tricks are two: Firstly we develop a technique for efficiently choosing a (reduced) collection of approximating basis functions, which is much smaller than the full collection which is usually used to solve the SS-ANOVA variational (penalized likelihood) problem exactly, yet the technique provides an answer which is still quite accurate when compared with solutions based on the full set. Secondly we develop and use a (new) randomized version of the Generalized Approximate Cross Validation method ($ranGACV$) for choosing multiple smoothing parameters in penalized likelihood equations. This $ranGACV$ is the major contribution of this paper, along with the basis function technique. They are what allow the SS-ANOVA models to be applied to very large data sets, while retaining, or even improving upon the favorable results available in previous work with these models. The $ranGACV$, and the method as a whole, is developed for general exponential families with no nuisance parameter; however, the simulation studies and data analysis here are carried out specifically for Bernoulli data.

SS-ANOVA models represent a function $f(t)$, $t = (x_1, \dots, x_d)$ of d variables as

$$f(t) = C + \sum_{\alpha} f_\alpha(x_\alpha) + \sum_{\alpha<\beta} f_{\alpha\beta}(x_\alpha, x_\beta) + \cdots, \tag{1}$$

where the main effects $\{f_\alpha\}$, two factor interactions $\{f_{\alpha\beta}\}$ etc. satisfy side conditions which generalize the usual side conditions for parametric ANOVA to function spaces and the series is truncated in some manner. Indicator functions and other parametric functions may be added to the model of (1). Independent observations y_i, $i = 1, \dots, n$ are assumed to be distributed with the density $g(y_i, f(t(i)))$, where $g(y, f) = \exp[yf - b(f) + c(y)]$, with parameter of interest $f(\cdot)$ and $f(\cdot)$ is assumed to be in an appropriate function space $\mathscr{H}$, a reproducing kernel Hilbert space. For Bernoulli data ($y = 0$ or 1), $c(y) = 0$, $b(f) = \log(1+\exp f)$, f is the log odds ratio, a.k.a. logit, and $f = log[\mu/(1-\mu)]$, where $\mu = Ey$. f is estimated as f_λ, the minimizer in $\mathscr{H}$ of the penalized log likelihood functional $I_\lambda(f, Y)$ given by

$$I_\lambda(f, Y) = -\frac{1}{n}\sum_{i=1}^{n} l(y_i, f_i) + \frac{1}{2}J_\lambda(f), \tag{2}$$

where $f_i = f(t(i))$, $Y = (y_1, \dots, y_n)'$, $l(y_i, f_i) = y_i f_i - b(f_i)$ is the log likelihood of $(y_i|f_i)$, and $J_\lambda(f)$ is of the form

$$J_\lambda(f) = \sum_\alpha \lambda_\alpha J_\alpha(f_\alpha) + \sum_{\alpha<\beta} \lambda_{\alpha\beta} J_{\alpha\beta}(f_{\alpha\beta}) + \cdots. \tag{3}$$

The $J_\alpha, J_{\alpha\beta}, \dots$ are quadratic penalty functionals, the series is truncated somewhere and the $\lambda_\alpha, \lambda_{\alpha\beta}, \dots$ are smoothing parameters to be chosen. In important examples the components of the fit f_λ are splines of various kinds, but may be much more general.

The exact minimizer of (2) is known to be in an n dimensional subspace of $\mathscr{H}$ [Kimeldorf and Wahba (1971)] and an essentially n dimensional matrix decomposition problem is solved to obtain the solution. In the typical model selection and model fitting problems that arise in demographic medical risk factor studies that we are concerned with here, the degrees of freedom for signal is very much less than n. It can be shown theoretically under certain circumstances [Gu (personal communication), Xiang (1996), Gao (1999)] that one can use fewer, appropriately chosen basis functions and obtain an estimate which is asymptotically indistinguishable from the "exact" estimate f_λ obtained using all the basis functions. We will demonstrate that excellent approximations to an "exact" solution in the "correct" n dimensional subspace can be obtained in a much smaller subspace spanned by $k \ll n$ basis functions, if the basis functions are chosen appropriately for that approximating task. The k basis functions we use here are a subset of the n representers of evaluation functionals at the observation points, in a subspace of $\mathscr{H}$; we describe a clustering technique for choosing the subset that works very well for this purpose.

To obtain the smoothing parameter estimates proposed here we begin with the generalized approximate cross validation ($GACV$) estimate proposed in Xiang and Wahba (1996), which was derived as a transformation of an approximate leaving-out-one cross validation estimate (ACV) for data from an exponential family with no nuisance parameter. The $GACV$ estimate was shown in Xiang and Wahba (1996) via simulation to have favorable properties for Bernoulli data, as measured by the Comparative Kullback-Leibler distance (CKL). However, the estimate there was computed directly with small sample sizes and the computation described there is not stable for large data sets. In the present work we develop $ranGACV$, which gets around the numerical instabilities of the original $GACV$ and which in fact is very cheap and stable to compute, given an algorithm for solving the penalized likelihood problem. As judged by the CKL in simulations, we find that $ranGACV$ behaves at least as well, and sometimes better than the original $GACV$ as well as previous related estimates, noted below. But at the same time it can be readily adapted for use in very large data sets with multiple smoothing parameters, and, possibly, complex model selection procedures, using the basis selection procedure noted above. The Bayesian "confidence intervals" of earlier work are also developed here in the context of the present algorithm. Selected simulation studies are presented which illustrate the properties of the method. Finally, the method is applied to the study of risk factors associated with pigmentary abnormalities, based on data from the Beaver Dam Eye study. Some new insights into this data set appear as a result of the analysis.

1.2. *Related work.* The work here may be considered a sequel to Wahba, Wang, Gu, Klein and Klein (1995) and Wang, Wahba, Gu, Klein and Klein (1997). In those papers, the iterative unbiased risk (*UBR*) estimate given in Gu (1992), Wang (1995, 1997) was extended for the purpose of choosing multiple smoothing parameters in the smoothing spline (penalized likelihood) ANOVA models. The *UBR* method is targeted at minimizing the *CKL* distance of the estimate from the unknown "truth." The penalized likelihood estimate is obtained using the fact that the solution to the variational problem is known to lie in an n-dimensional space based on the n representers of evaluation in $\mathscr{H}$ and the span of the null space of the penalty functional [see, e.g., Wahba (1990)]. A Newton-Raphson iteration is used to find the coefficients, for any fixed set of smoothing parameters $\lambda = \{\lambda_\alpha, \lambda_{\alpha\beta} \dots\}$. At each step of the iteration, a local quadratic approximation to the penalized likelihood functional is obtained. When the penalized likelihood problem is exactly quadratic (the Gaussian case), and the variance is known, there is an exact unbiased risk estimate. The iterative *UBR* estimate alternates back and forth between solving the penalized likelihood problem, and choosing the smoothing parameter(s) to minimize the unbiased risk estimate for the Gaussian problem associated with the local quadratic approximation. This method is available as an option in *GRKPACK*. It is specifically designed for SS-ANOVA models and has been successfully used by a number of authors. We will be comparing the present algorithm to the algorithm in *GRKPACK*, as well as an earlier randomization technique proposed in Xiang (1996) and Xiang and Wahba (1997).

As many readers will know, the search for data based smoothing or bandwidth parameters is a very active field of research, pursued in the context of various kinds of nonparametric regression estimates, for example, kernel, orthogonal series, regression splines, wavelets, sparse representations, and so forth, and various kinds of data, including Bernoulli data. We will not attempt to discuss the extensive literature here, with the exception of mention of recent work related to smoothing spline/penalized likelihood ANOVA models, and smoothing parameter estimates that are relatively closely related to the *ranGACV*. More extensive references to the literature can be found in Wahba, Wang, Gu, Klein and Klein (1995) and more recently in Wood and Kohn (1998).Wood and Kohn (1998) propose a Bayesian method for choosing smoothing parameters in the multicomponent spline context based on endowing the smoothing parameter(s) with a flat prior, and a Gibbs sampler for the computations. Interestingly, they provide simulation examples which show that the method has favorable properties compared to *GRKPACK*, which is targeted more directly to the *CKL* distance. We do not provide any comparisons here, but it would be interesting to further understand theoretically how methods directly targeted at the *CKL* might compare with the Woods-Kohn approach. We note in passing other recent references making use of smoothing spline ANOVA models in various contexts: Wang (1998),Gu (1998), Verbyla, Cullis, Kenward and Welham (1997), Brumback and Rice (1998), Luo (1997). Lin (1998c), Lin (1998b) has recently obtained some general convergence results for these models. We note that the popular additive spline models in Hastie and Tibshirani (1990) are the special case of (1) restricted to main effects.

Ye and Wong (1997a), Ye and Wong (1997b) define the generalized degrees of freedom (*GDF*) in the general exponential family case, and by an interesting theorem show that it is the key to model fitting and model selection when the goal is to minimize the *CKL*. The *GDF* generalizes the degrees of freedom for signal for the Gaussian penalized likelihood case, given in Wahba (1983), where it is defined as the trace of the influence matrix. Ye and Wong's theorem holds for *any* model fitting procedure, not just penalized likelihood estimates with prespecified terms, and they argue that it justifies the use of the *GDF* in very general model selection procedures. Interesting examples of the use of the estimated *GDF* in model selection in the Gaussian case are given in Ye (1998), where randomization techniques are used in the estimation process. In the general non-Gaussian exponential family case, the *GDF* depends on the true but unknown parameter f that one is attempting to estimate. Ye and Wong (1997b) outline an approach based on sensitivity analyses to estimate the *GDF* in the Bernoulli case, which is similar in spirit, but not exactly the same as the estimate proposed here.

1.3. *Outline of paper.* In Section 2 we review unbiased risk estimates in the penalized likelihood, general exponential family case with no nuisance parameter when the *CKL* is the target, and we describe the role of the *GDF* in obtaining these estimates, or approximations to them. In Section 3 we present the the *GACV* estimate from Xiang and Wahba (1996) in order to set the stage for the derivation of the randomized version, $ranGACV$, which takes place in Section 4. It is noted here that the technique may be applicable in much more general contexts. Section 5 discusses approximate solutions to the penalized likelihood optimization problem, Section 5.1 describes a clustering technique for extracting an approximating basis set, and Section 5.2 brings the clustering technique and the $ranGACV$ together. Section 5.3 describes the Bayes model behind this approximate estimate, as well as Bayesian "confidence intervals" which generalize the confidence intervals given in Wahba, Wang, Gu, Klein and Klein (1995), Wahba (1983) to this approximate estimate. Section 6 presents a suite of simulation studies, to examine by illustration the properties of the method, and to compare the estimates with the iterated *UBR* estimates of Wahba, Wang, Gu, Klein and Klein (1995). The first set of simulations is based on simple "truth" functions with regular data. The second set of simulations is based on "truth" functions which were previously obtained smoothing spline ANOVA model fits to two epidemiological data sets. For the simulated data points we used the observed epidemiological design points, which are quite irregular, as is typical in many observational studies. These two epidemiological data sets are from the Pima Indian Diabetes Data Set and the Wisconsin Epidemiologic Study of Diabetic Retinopathy. In Section 7 we use the method to analyze the association of pigmentary abnormalities with various risk factors in the women in the Beaver Dam Eye Study ($n = 2585$, 5 smoothing parameters). We found via the use of this method an association of lower cholesterol with the presence of pigmentary abnormalities and an association of hormone replacement therapy with their absence. Finally, Section 8 gives a summary and conclusions.

2. Unbiased risk estimates and the generalized degrees of freedom. Let $y_i, i = 1, \dots, n$ be independent random variables from an exponential family with no nuisance parameter, with density of the form

$$g(y_i, f_i) = \exp\{y_i f_i - b(f_i) + c(y_i)\}, \tag{4}$$

with b a strictly convex function of f on any bounded set. We have that $Ey_i = b'(f_i) = \mu_i$, and $var\ y_i = b''(f_i) = \frac{\partial \mu_i}{\partial f_i} \equiv \sigma_i^2$, say. Our examples in this paper are all for Bernoulli data. Thus $y_i = 1$ with probability μ_i and 0 with probability $(1-\mu_i)$. In this case $b(f_i) = log(1+e^{f_i})$, $b'(f_i) = e^{f_i}/(1+e^{f_i}) \equiv \mu_i$, $b''(f_i) = e^{f_i}/(1+e^{f_i})^2 \equiv \mu_i(1-\mu_i) \equiv \sigma_i^2$, and $c(y_i)$ is 0. Furthermore, $f_i = \log[\mu_i/(1-\mu_i)]$ is the log odds ratio, also known as the logit. We assume that $f_i \equiv f(t(i))$, where $t(i)$ is a vector of covariates, $t(i) \in \mathscr{T}$, where $\mathscr{T}$ is some possibly multivariate index set. $f(\cdot)$ is assumed to be in some reproducing kernel Hilbert space $\mathscr{H}$ of real valued functions of $t \in \mathscr{T}$. If a component of t is continuous, then typically $f(\cdot)$ will be a smooth function of that component. It is desired to estimate $f(\cdot)$. f is estimated as f_λ, the minimizer of $I_\lambda(f, Y) = -\frac{1}{n}\sum_{i=1}^n l(y_i, f_i) + \frac{1}{2}J_\lambda(f)$ of (2) where $l(y_i, f_i) = y_i f_i - b(f_i)$ is the log likelihood minus $c(y_i)$. In the examples we will study, as the components of λ become large, f_λ is shrunk into a low dimensional (parametric) subspace, and as λ becomes small, $\mu_\lambda \equiv \mu(f_\lambda)$, where $\mu(f_\lambda) = b'(f_\lambda)$, becomes closer to the observations. We have a family of estimates indexed by λ, and the goal is to choose λ from the observations to minimize the comparative Kullback-Leibler distance $CKL(\lambda)$ of f_λ from f. Letting $f_{\lambda i} = f_\lambda(t(i)), \mu_{\lambda i} = \mu(f_{\lambda i})$, $CKL(\lambda)$ is given by

$$CKL(\lambda) = KL(f, f_\lambda) - \frac{1}{n}\sum_{i=1}^n [E_\mu y_i f_i - b(f_i)] \tag{5}$$

$$\equiv \frac{1}{n}\sum_{i=1}^n [-\mu_i f_{\lambda i} + b(f_{\lambda i})] \tag{6}$$

where

$$KL(f, f_\lambda) = \frac{1}{n}\sum_{i=1}^n E_\mu \left(\log \frac{g(y_i, f_i)}{g(y_i, f_{\lambda i})}\right).$$

Here E_μ indicates expectation with respect to the true $\mu = b'(f)$, equivalently the true f. The *CKL* differs from the Kullback-Leibler distance *KL* by quantities not depending on λ. The *CKL* is, up to a factor of 2, the same as in Hastie and Tibshirani [(1990), equation (6.29)], where it is called the *PE* (prediction error). For Gaussian observations, minimizing the *CKL* is equivalent to minimizing the predictive mean square error, hence the *CKL* can be viewed as a generalization of the predictive mean square error. It is tempting to estimate the *CKL* by

$$OBS(\lambda) = \frac{1}{n}\sum_{i=1}^{n}[-y_i f_{\lambda i} + b(f_{\lambda i})] \tag{7}$$

(where "OBS" stands for "observed" and corresponds in the Bernoulli case to one half the deviance), but it is well known that $OBS(\lambda)$ is an underestimate of the $CKL(\lambda)$; see, for example, Efron (1986). This is related to the fact that y_i and $f_{\lambda i}$ are correlated.

Let $D(\lambda)$ be defined by

$$CKL(\lambda) = OBS(\lambda) + D(\lambda). \tag{8}$$

Then $D(\lambda) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \mu_i)f_{\lambda i}$ and

$$E_\mu D(\lambda) = \frac{1}{n}\sum_{i=1}^{n} E_\mu (y_i - \mu_i) f_{\lambda i}. \tag{9}$$

Ye and Wong (1997b) have provided the following interesting theorem concerning (9):

Let $\hat{f}$ be any *estimate of f. Then*

$$\frac{1}{n}\sum_{i=1}^{n} E_{\mu_i}(y_i - \mu_i)\hat{f}_i = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial f_i}E_{\mu_i}(\hat{f}_i) = \frac{1}{n}\sum_{i=1}^{n}\sigma_i^2\frac{\partial}{\partial \mu_i}E_{\mu_i}(\hat{f}_i). \tag{10}$$

Here $E_{\mu_i}(\hat{f}_i)$ is the expectation with respect to y_i conditional on the y_j, $j \neq i$ being fixed. The proof is short:

$$\begin{aligned}\frac{\partial}{\partial f}E_\mu(\hat{f}) &= \frac{\partial}{\partial f}\int \hat{f}(z)\exp\{[zf - b(f) + c(z)]\}dz\\ &= \int z\hat{f}(z)\exp\{[zf - b(f) + c(z)]\}dz\\ &\quad -b'(f)\int \hat{f}(z)\exp\{[zf - b(f) + c(z)]\}dz\\ &= E_\mu(y\hat{f} - \mu\hat{f}).\end{aligned}$$

The second equality follows from the fact that $\sigma_i^2 = \frac{\partial\mu_i}{\partial f_i}$. Ye and Wong call n times the right hand side of (10) the generalized degrees of freedom for $\hat{f}$ ($GDF(\hat{f})$), generalizing the degrees of freedom for signal proposed in Wahba

(1983) and elsewhere. In the Gaussian case with the variance σ_i^2 known (w.l.o.g. set $\sigma_i^2 = 1$) we have $b(f) = f^2/2, \mu(f) = f$. If f_λ is the minimizer of (2) there is a symmetric, nonnegative definite (smoother) matrix $A(\lambda)$ such that

$$\begin{pmatrix} f_{\lambda 1} \\ \vdots \\ f_{\lambda n} \end{pmatrix} = A(\lambda)Y.$$

See Wahba (1990). In this case $\sigma_i^2 \frac{\partial}{\partial \mu_i} E_\mu f_{\lambda i} = a_{ii}(\lambda)$, where $a_{ii}(\lambda)$ is the iith entry of $A(\lambda)$ and $nE_\mu D(\lambda) \equiv GDF(f_\lambda) = trA(\lambda)$. This result is the well known unbiased risk estimate of λ as the minimizer of

$$\frac{1}{2n}\left[\sum_{i=1}^{n}(y_i - f_{\lambda i})^2 + 2trA(\lambda)\right] \tag{11}$$

[Mallows (1973), Craven and Wahba (1979)]. We see from (10) that the problem of choosing λ to minimize the *CKL* in a penalized likelihood estimate in the non-Gaussian, exponential family case, can be reduced to the problem of estimating $GDF(f_\lambda)$. Wong (1992) and Ye and Wong (1997b) give an exact unbiased risk estimate in the Poisson case. However they also show that in the Bernoulli case, no unbiased estimate of $GDF(\hat{f})$ exists, so that only approximations are possible. The minimizer of the *GACV*, which will be discussed in the next section, has been shown to provide a good estimate of the minimizer of the *CKL*. The *GACV* is the sum of *OBS* and an additional term which may be thought of as an estimator of the *GDF*.

3. The GACV estimate of λ.

In the general penalized likelihood problem where J_λ is a seminorm in $\mathscr{H}$, the minimizer $f_\lambda(\cdot)$ of (2) has a representation

$$f_\lambda(t) = \sum_{\nu=1}^{M} d_\nu \phi_\nu(t) + \sum_{i=1}^{n} c_i Q_\lambda(t(i), t) \tag{12}$$

where the ϕ_ν span the null space of J_λ, $Q_\lambda(s,t)$ is a reproducing kernel (positive definite function) for the penalized part of $\mathscr{H}$, and $c = (c_1, \dots, c_n)'$ satisfies M linear conditions, so that there are (at most) n free parameters in f_λ. Typically the unpenalized functions ϕ_ν are low degree polynomials. If $f_\lambda(\cdot)$ is of the form (12) then $J_\lambda(f_\lambda) = \sum_{i,j=1}^{n} c_i c_j Q_\lambda(t(i), t(j))$. Substituting this and (12) into (2) results in I_λ a convex functional in c and $d = (d_1, \dots, d_M)'$, and c and d are obtained numerically via a Newton Raphson iteration. For large n, the second sum on the right of (12) may in some applications be replaced by an approximation of the form $\sum_{\ell=1}^{k} c_{i_\ell} Q_\lambda(t(i_\ell), t)$ for some $k \ll n$. The rationale for this approximation, and the choice of the $t(i_\ell)$ will be discussed later.

The *GACV* was obtained in Xiang and Wahba (1996). They began with the ordinary leaving-out-one cross validation function $CV(\lambda)$ as an estimate for the for the *CKL* of (6):

$$
\begin{aligned}
&(13)\quad CV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left[-y_i f_{\lambda i}^{[-i]} + b(f_{\lambda i})\right] \\
&(14)\qquad\qquad = OBS(\lambda) + \frac{1}{n}\sum_{i=1}^{n}\left[y_i\left(y_i - \mu_{\lambda i}^{[-i]}\right)\right]\left[\frac{f_{\lambda i} - f_{\lambda i}^{[-i]}}{y_i - \mu_{\lambda i}^{[-i]}}\right] \\
&(15)\qquad\qquad = OBS(\lambda) + \frac{1}{n}\sum_{i=1}^{n} y_i(y_i - \mu_{\lambda i})\left[\frac{f_{\lambda i} - f_{\lambda i}^{[-i]}}{y_i - \mu_{\lambda i}^{[-i]}}\right]\Bigg/\left[1 - \frac{\mu_{\lambda i} - \mu_{\lambda i}^{[-i]}}{y_i - \mu_{\lambda i}^{[-i]}}\right] \\
&\qquad\qquad\quad \approx OBS(\lambda) \\
&(16)\qquad\qquad\quad + \frac{1}{n}\sum_{i=1}^{n} y_i(y_i - \mu_{\lambda i})\left[\frac{f_{\lambda i} - f_{\lambda i}^{[-i]}}{y_i - \mu_{\lambda i}^{[-i]}}\right]\Bigg/\left[1 - \sigma_{\lambda i}^2\left(\frac{f_{\lambda i} - f_{\lambda i}^{[-i]}}{y_i - \mu_{\lambda i}^{[-i]}}\right)\right]
\end{aligned}
$$

where $f_{\lambda}^{[-i]}$ is the solution to the variational problem of (2) with the ith data point left out and $f_{\lambda i}^{[-i]}$ is the value of $f_{\lambda}^{[-i]}$ at $t(i)$. (Observe that y_i and $f_{\lambda i}^{[-i]}$ are uncorrelated.) Here $\sigma_{\lambda i}^2 = \sigma^2(f_{\lambda i})$ and the last approximation follows upon recalling that $\frac{\partial\mu}{\partial f} = \sigma^2$. Xiang and Wahba (1996) after a series of steps approximating the terms in brackets in (16) obtain the *GACV* as

$$
\begin{aligned}
&\qquad GACV(\lambda) = OBS(\lambda) \\
&(17)\qquad\qquad + \frac{1}{n}\sum_{i=1}^{n} y_i(y_i - \mu_{\lambda i})\left[\frac{1}{n}trH\right]\Bigg/\left[\frac{1}{n}tr(I - (W^{1/2}HW^{1/2})\right],
\end{aligned}
$$

where $W = W(f)$ is the $n \times n$ diagonal matrix with $\sigma_{\lambda i}^2$ in the iith position and $H = [W + n\Sigma_\lambda]^{-1}$, where, to define Σ_λ we need some notation as follows: Where there is no confusion between functions $f(\cdot)$ and vectors $(f_1, \ldots, f_n)'$ of values of f at $t_1, \ldots, t_n$, let $f = (f_1, \ldots, f_n)'$. For any $f(\cdot)$ of the form (12), $J_\lambda(f)$ also has a representation as a non-negative definite quadratic form in $(f_1, \ldots, f_n)'$, Σ_λ is the matrix of this quadratic form. We can then rewrite (2) as

$$
(18)\qquad I_\lambda(f, Y) = \frac{1}{n}\sum_{i=1}^{n}[-y_i f_i + b(f_i)] + \frac{1}{2}f'\Sigma_\lambda f.
$$

Using the fact that σ_i^2 is the second derivative of $b(f_i)$, we note that $H = [W+n\Sigma_\lambda]^{-1}$ is the inverse Hessian of this variational problem (18). The inverse Hessian of the variational problem plays a key role in perturbation methods here, as can be seen clearly in (20) below. [Note that in the Gaussian case $A(\lambda)$ is the inverse Hessian.]

Numerical results based on an exact calculation of (17), which provide evidence that the minimizer of $GACV(\lambda)$ is a good estimate of the minimizer of $CKL(\lambda)$, appear in Xiang and Wahba (1996). This exact calculation is limited to small n however, since the direct calculation of Σ_λ will generally be unstable for large n.

The reader may compare (17) with a similar, but not the same, estimate given in equation (6.30) of Hastie and Tibshirani (1990) called AIC, and reproduced here, namely $AIC = D(y;\hat{u})/n + 2df\phi/n$. Their $D(y;\hat{\mu})/n$ is what we call $OBS(\lambda)$. The second term in (17), estimating the $GDF(\lambda)$, plays the same role as their $df\phi/n$, but differs from their suggestions for this quantity.

4. The randomized GACV estimate. Given any "black box" which, given λ, and a training set $\{y_i, t(i)\}$ produces $f_\lambda(\cdot)$ as the minimizer of (2), and hence $f_\lambda = (f_{\lambda 1}, \dots, f_{\lambda n})'$, we can produce randomized estimates of trH and $tr[I - W^{1/2}HW^{1/2}]$ without having any explicit calculations of these matrices. This is done by running the "black box" on perturbed data $Y + \varepsilon$, where the components of ε come from a random number generator. When the y_i are from a Gaussian distribution, randomized trace estimates of the inverse Hessian of the variational problem (the "influence matrix") have been studied extensively and shown to be essentially as good as exact calculations for large n; see, for example, Girard (1998). Randomized trace estimates are based on the fact that if A is any square matrix and ε is a zero mean random n-vector with independent components with variance σ_ε^2, then $\frac{1}{\sigma_\varepsilon^2}E\varepsilon' A\varepsilon = trA$. See Gong, Wahba, Johnson and Tribbia (1998) and references cited there for experimental results with multiple regularization and other parameters in the Gaussian case. In practice σ_ε^2 is replaced by $\frac{1}{n}\sum_{i=1}^n \varepsilon_i^2$. Xiang and Wahba (1997), following the argument in Xiang and Wahba (1996), obtained the approximation $f_\lambda^{Y+\varepsilon} - f_\lambda^Y \approx [W(f_\lambda^Y) + n\Sigma_\lambda]^{-1}\varepsilon$, which suggests that $\frac{1}{\sigma_\varepsilon^2}\varepsilon'(f_\lambda^{Y+\varepsilon} - f_\lambda^Y)$ with $\varepsilon \sim \mathscr{N}(0, \sigma_\varepsilon^2 I)$ provides an estimate of $tr[W(f_\lambda^Y) + n\Sigma_\lambda]^{-1}$.

In this work we make the key observation that if we take the solution f_λ^Y to the nonlinear system for the original data Y as the initial value for a Newton-Raphson calculation of $f_\lambda^{Y+\varepsilon}$ things become even simpler. Letting $f_\lambda^{Y+\varepsilon,1}$ be the result of the first step in a Newton-Raphson iteration gives

$$(19)\qquad f_\lambda^{Y+\varepsilon,1} = f_\lambda^Y - \left[\frac{\partial^2 I_\lambda}{\partial f'\partial f}(f_\lambda^Y, Y+\varepsilon)\right]^{-1}\frac{\partial I_\lambda}{\partial f}(f_\lambda^Y, Y+\varepsilon).$$

Since $\frac{\partial I_\lambda}{\partial f}(f_\lambda^Y, Y+\varepsilon) = -\varepsilon + \frac{\partial I_\lambda}{\partial f}(f_\lambda^Y, Y) = -\varepsilon$, and $[\frac{\partial^2 I_\lambda}{\partial f'\partial f}(f_\lambda^Y, Y+\varepsilon)]^{-1} = [\frac{\partial^2 I_\lambda}{\partial f'\partial f}(f_\lambda^Y, Y)]^{-1}$, we have $f_\lambda^{Y+\varepsilon,1} = f_\lambda^Y + [\frac{\partial^2 I_\lambda}{\partial f'\partial f}(f_\lambda^Y, Y)]^{-1}\varepsilon$ so that

$$(20)\qquad f_\lambda^{Y+\varepsilon,1} - f_\lambda^Y = [W(f_\lambda^Y) + n\Sigma_\lambda]^{-1}\varepsilon.$$

The result is the following *ranGACV* function:

$$\begin{aligned} &ranGACV(\lambda) \\ &\quad = OBS(\lambda) + \frac{1}{n}\sum_{i=1}^{n} y_i(y_i - \mu_{\lambda i}) \\ &\qquad \times \left[\varepsilon'(f_\lambda^{Y+\varepsilon,1} - f_\lambda^Y)\right] / \left[\varepsilon'\varepsilon - \varepsilon' W(f_\lambda^Y)(f_\lambda^{Y+\varepsilon,1} - f_\lambda^Y)\right]. \end{aligned} \tag{21}$$

To reduce the variance in the term after the "+" in (21), we may draw R independent replicate vectors $\varepsilon_1, \dots, \varepsilon_R$, and replace the term after the "+" in (21)by

$$\begin{aligned} &\frac{1}{n}\sum_{i=1}^{n} y_i(y_i - \mu_{\lambda i}) \\ &\quad \times \frac{1}{R}\sum_{r=1}^{R} \left[\varepsilon_r'(f_\lambda^{Y+\varepsilon_r,1} - f_\lambda^Y)\right] \Big/ \left[\varepsilon_r'\varepsilon_r - \varepsilon_r' W(f_\lambda^Y)(f_\lambda^{Y+\varepsilon_r,1} - f_\lambda^Y)\right] \end{aligned} \tag{22}$$

to obtain an R-replicated $ranGACV(\lambda)$ function.

We remark that *ranGACV* of (21) was derived assuming a particular penalized likelihood estimate, and various continuity properties were assumed. However, following Ye (1998), Ye and Wong (1997a), it can be defined for a variety of other procedures which produce an estimate $\hat{f}^Y$ given a data vector Y, by replacing $f_\lambda^{Y+\varepsilon,1} - f_\lambda^Y$ in (21) by $\hat{f}^{Y+\varepsilon} - \hat{f}^Y$, if necessary. If the second term in (21) is still a reasonable estimate of the *GDF*, then this estimate could be used in data mining and model selection in the same way as the estimated *GDF* was used in Ye (1998). At present, this observation is, however, only conjectural. Randomized trace estimates for the inverse Hessian in a Gaussian problem where, however the optimization problem was non quadratic in a fairly complicated way, have been quite successful; see Gong, Wahba, Johnson and Tribbia (1998).

5. Approximate solutions to the penalized likelihood problem.

5.1. *Clustering the design points.* The "exact" minimizer of (2) is, as noted, in an n dimensional subspace consisting of *span* ϕ_ν and an $n-M$ dimensional subspace of $\mathscr{H}$ spanned by the $Q_{\lambda i}(\cdot)$; where $Q_{\lambda i}(t) = Q_\lambda(t(i), t)$, and the coefficient vector c satisfies $T'c = 0$ with T the $n \times M$ matrix with i, νth entry $\phi_\nu(t(i))$. Various authors have suggested (in this and other contexts) that, especially for large data sets, f be found in a smaller subspace, say that spanned by the ϕ_ν and $Q_{\lambda i_\ell}$, $\ell = 1, \dots, k$, say. See, for example, Hutchinson (1984), Silverman (1985). Given that f_λ will be constrained to be of the form

$$f_\lambda = \sum_{\nu=1}^{M} d_\nu \phi_\nu + \sum_{\ell=1}^{k} c_{i_\ell} Q_{\lambda i_\ell} \tag{23}$$

and observing that then $J_\lambda(f_\lambda) = \sum_{\ell,\ell'=1}^k c_{i_\ell} c_{i_{\ell'}} Q_\lambda(t(i_\ell), t(i_{\ell'}))$, the coefficients d and $c_k = (c_{i_1}, \ldots, c_{i_k})'$ minimizing (2) can be found by a similar Newton-Raphson iteration as that described for the case $k = n$, in Wahba, Wang, Gu, Klein and Klein (1995). It is

$$\begin{pmatrix} Q_\lambda^{kn} W_- Q_\lambda^{nk} + nQ_\lambda^{kk} & Q_\lambda^{kn} W_- T \\ T' W_- Q_\lambda^{nk} & T' W_- T \end{pmatrix} \begin{pmatrix} c_k - c_{k-} \\ d - d_- \end{pmatrix} = \begin{pmatrix} -Q_\lambda^{kn} \mu_{\lambda -} - nQ_\lambda^{kk} c_{k-} \\ -T' \mu_{\lambda -} \end{pmatrix}, \tag{24}$$

where Q_λ^{nk} is the $n \times k$ matrix with i, ℓ entry $Q_\lambda(t(i), t(i_\ell))$, $i = 1, \cdots, n$, $\ell = 1, \ldots, k$, Q_λ^{kn} is its transpose, Q_λ^{kk} is the $k \times k$ matrix $Q(t(i_\ell), t(i_{\ell'}))$, $\ell, \ell' = 1, \ldots, k$, and $\mu_\lambda = (\mu_{\lambda 1}, \ldots, \mu_{\lambda n})'$. The subscript "_" indicates the value from the previous iteration. The $k = n$ case corresponds to iteratively reweighted least squares, see Wahba, Wang, Gu, Klein and Klein (1995), Hastie and Tibshirani (1990).

There are several different criteria for choosing the i_ℓ, and the resulting methods may, roughly, be divided into two categories, namely those which involve both y_i and $t(i)$, and those which involve only the $t(i)$. In the former category are included methods designed to capture different amounts of structure in the estimate in different parts of $\mathscr{T}$. Luo and Wahba (1997) is in this category. A greedy algorithm is used to choose the $t(i_\ell)$ and the result is an estimate which allows more flexibility in the solution where the data are more dense and/or the responses more variable. MARS [Friedman (1991)] and related methods for "knot selection" which restrict the knots to a subset of the data points, and choose them via a greedy algorithm, are in this spirit. The second category of methods is based on the assumption that the minimizer of (2) is the "gold standard" and it is desired to choose k and the i_ℓ as a compromise to obtain a good approximation to the minimizer of (2) while reducing the computational cost of performing a Newton-Raphson iteration in $n + M$ unknowns to one in $k + M$ unknowns. Gu (personal communication), Xiang (1996) and Gao (1999) have shown in some special cases, that if k increases at an appropriate (quite slow) rate, the same convergence rates are obtainable as with the exact solution. In the problems that we will consider in the rest of the paper, which concern Bernoulli observations from demographic data sets, the desired f_λ will generally not be expected to have a lot of fine structure, and, furthermore, there is no *a priori* reason to believe that the desired estimate is more "wiggly" in one part of $\mathscr{T}$ than another. Thus, we consider only the second category of methods here. Furthermore, we can expect that k may be substantially less than n in many cases and still provide an excellent approximation to the minimizer of (2) when the desired solution has relatively few degrees of freedom.

If $t(i)$ is close to $t(j)$ in some sense, then $Q_{\lambda i}$ will be "close" to $Q_{\lambda j}$, so for fixed k, for the purpose of approximating the minimizer of (2) it is desirable that the $t(i_\ell)$ have maximal separation while being "representative" of the full set of $t(i)$. A random or stratified sampling scheme on the $t(i)$ (after suitable scaling) is possible. Xiang and Wahba (1995), Xiang (1996), Xiang and Wahba (1997) utilize a clustering scheme to "thin out" basis functions.

5.2. *The algorithm.* In this work we utilize a similar clustering scheme. The FASTCLUS procedure in SAS [SAS Institute (1989)],which is designed for the disjoint clustering of very large data sets in minimal time, is used to obtain K clusters from the n data points. Here, within each cluster, one data point $t(i_\ell)$ is chosen at random to be representative of each of the K clusters. Given the basis functions $Q_{\lambda i_\ell}, \ell = 1, \dots, K$, (2) is minimized in the span of the ϕ_ν, $Q_{\lambda i_\ell}$ and $ranGACV(\lambda)$ is computed. The minimizer $\hat{\lambda}(K)$, say, of $ranGACV(\lambda)$ is found, along with the fit, $f_{\hat{\lambda}(K)}$. (The minimization scheme for λ is described in Section 5.4 below.) Then K is increased, say, by a factor of 2, and the process repeated to obtain $f_{\hat{\lambda}(2K)}$, say. Then K is increased again, until the difference between two consecutive fits is smaller than a given tolerance, as judged by

$$\frac{\|f_{\hat{\lambda}(2K)} - f_{\hat{\lambda}(K))}\|}{\|f_{\hat{\lambda}(K)}\|} \le 10^{-4}.$$

It is possible that the coefficient matrix of the linear system (24) would be computationally singular even if it is nonsingular in theory. In order to get a stable solution, the QR factorization with pivoting is used. Also, when solving the linear system using the QR decomposition, a cutoff parameter τ is selected (such as the machine precision times the largest absolute diagonal element of the R matrix). Whenever $|r_{ii}| \le \tau$ (where r_{ii} denotes the diagonal element of the R matrix in the QR decomposition), the corresponding component of the solution is set to be zero.

5.3. *The Bayes model and Bayesian "confidence intervals" for the approximate solution.* Bayesian "confidence intervals" based on the Bayes model corresponding to the variational problem (2) are discussed in Wahba, Wang, Gu, Klein and Klein (1995). They are based on approximating the posterior distribution of $\{f_\lambda^Y|Y, \lambda\}$ in the exponential family case, by the posterior distribution of $\{f_\lambda^Y|Y, \lambda\}$ in the Gaussian case which corresponds to the quadratic optimization problem appearing in the last step of the Gauss-Newton iteration for minimizing (2). See Wahba, Wang, Gu, Klein and Klein (1995), Gu (1990), Gu (1992). Experiments described there [see also Wang and Wahba (1995)] tend to provide empirical evidence that these estimates have the "across the function" property when an optimum value of λ is used. The "across the function" property says that about 95% of the true values at the observation points will be covered by the 95% Bayesian "confidence intervals." See also Wahba (1983), Nychka (1988).

The Gaussian model is

$$y_i = f(t(i)) + \varepsilon_i, i = 1, \dots, n, \tag{25}$$

where $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)' \sim \mathscr{N}(0, W^{-1})$ and

$$f(t) \sim \sum_{\nu=1}^{M} d_\nu \phi_\nu(t) + Z_\lambda(t), t \in \mathscr{T}, \tag{26}$$

where $d \sim \mathscr{N}(0, \xi I_M)$ with $\xi \to \infty$ and $Z_\lambda(t)$ zero mean Gaussian with

$$EZ_\lambda(s) Z_\lambda(t) = Q_\lambda(s, t). \tag{27}$$

The Bayes model behind f_λ of (23) is obtained by replacing $Z_\lambda(t)$ of (27) by the projection $Z_\lambda^k(\cdot)$ of $Z_\lambda(\cdot)$ onto $span\{Z_\lambda(t_{i_\ell}), \ell = 1, \dots, k\}$, that is

$$Z_\lambda^k(t) = E[Z_\lambda(t) | Z_\lambda(t_{i_1}), \dots, Z_\lambda(t_{i_k})]. \tag{28}$$

We have that

$$EZ_\lambda^k(s) Z_\lambda^k(t) = (Q_{\lambda i_1}(s), \dots, Q_{\lambda i_k}(s))(Q_\lambda^{kk})^{-1} \begin{pmatrix} Q_{\lambda i_1}(t) \\ \cdots \\ Q_{\lambda i_k}(t) \end{pmatrix}. \tag{29}$$

Following Silverman (1985), another representation for the zero mean Gaussian stochastic process with the covariance (29) is

$$Z_\lambda^k(t) = \sum_{\ell=1}^{k} c_{i_\ell} Q_{\lambda i_\ell}(t) \tag{30}$$

where we endow the vector $c_k = (c_{i_1}, \dots, c_{i_k})'$ with the distribution $c_k \sim \mathscr{N}(0, (Q_\lambda^{kk})^{-1})$. Replacing $Z_\lambda(t)$ in (26) by $Z_\lambda^k(t)$ of (30) into the joint distribution of Y and (c_k, d) we have that the posterior log likelihood of (c_k, d) given (Y, λ) is proportional to

$$-\tfrac{1}{2}\Big[(Y - Q_\lambda^{nk} c_k - Td)' W (Y - Q_\lambda^{nk} c_k - Td) + c_k' Q_\lambda^{kk} c_k\Big]. \tag{31}$$

This gives that the posterior distribution of $(c_k' : d')'$ given Y, λ has mean

$$E[(c_k' : d')' | Y, \lambda] = M^{-1} \begin{pmatrix} Q_\lambda^{kn} \\ \cdots \\ T' \end{pmatrix} Y, \tag{32}$$

and covariance M^{-1}, where

$$M = \begin{pmatrix} Q_\lambda^{kn} W Q_\lambda^{nk} + n Q_\lambda^{kk} & Q_\lambda^{kn} WT \\ T' W Q_\lambda^{nk} & T'WT \end{pmatrix}. \tag{33}$$

Letting $v(t) = (Q_{\lambda i_1}(t), \dots, Q_{\lambda i_k}(t) : \phi_1(t), \dots, \phi_M(t))'$ then gives the posterior distribution of $f_\lambda(\cdot)$ as

$$\begin{gathered} E[f_\lambda(t) | Y, \lambda] = v(t)' M^{-1} \begin{pmatrix} Q_\lambda^{kn} \\ \cdots \\ T' \end{pmatrix} Y, \\ \operatorname{cov}[f_\lambda(s), f_\lambda(t) | Y, \lambda] = v(s)' M^{-1} v(t). \end{gathered} \tag{34}$$

Since the Newton-Raphson iteration for minimizing (2) solves $Mx = y$ [refer to (24)], the (approximate, Bayesian) posterior variance of $f_\lambda(t)$ is at hand.

5.4. *Minimizing $ranGACV(\lambda)$ in the case $\lambda = (\lambda_1, \ldots, \lambda_p)$.* In the cases we will be interested in, which include the well known additive models [Hastie and Tibshirani (1990)] smoothing spline ANOVA models [Wahba, Wang, Gu, Klein and Klein (1995)] and others, $Q_\lambda(s, t)$ has a representation of the form

$$Q_\lambda(s,t) = \sum_{\beta=1}^{p} \theta_\beta R_\beta(s,t), \tag{35}$$

with ${\theta_\beta}^{-1} = \lambda_\beta$; see Wahba (1990) and the examples below. Since first or second derivatives of $ranGACV(\lambda_1, \ldots, \lambda_p)$ are not at hand in this formulation of the problem, standard optimization methods such as the Newton method or conjugate gradient algorithm for minimizing $ranGACV(\lambda_1, \ldots, \lambda_p)$ are not available. In the experiments here, we have found that for p as many as 6 or more, the downhill simplex method [Press, Teukolvsky, Vetterling and Flannery (1992)], possibly in conjunction with computer experimental design techniques [Bowman, Sacks and Chang (1993)] works well. Details will be given with the experiments below.

6. Simulation studies.

In this section we present several simulation studies, designed to illustrate various aspects of the overall approach, and to compare the $ranGACV$ estimate with the iterative UBR estimate. Comparison with other methods remains for the future. The models here are the same smoothing spline ANOVA models described in detail in Wahba, Wang, Gu, Klein and Klein (1995). These models have main effects which are cubic splines. Other models may be found in Gu and Wahba (1993) and elsewhere. In the models here, the $t(i)$ are all rescaled to the unit interval or the unit cube. For $t \in E^d, t = (x_1, \ldots, x_d)$, the R_β of (35) are built up from the basic linear function and cubic spline reproducing kernels (rk's) $r_o(u, v) = (u - 1/2)(v - 1/2)$ and $r_1(u, v) = k_2(u)k_2(v) - k_4([u - v])$, where $u, v \in [0, 1]$, $\ell! k_\ell(u)$ is the ℓth Bernoulli polynomial and $[x]$ is the fractional part of x. The main effects rk's involve only r_1 with values of x_α and the two factor interaction terms involve tensor products of r_o and r_1 and of r_1 with itself, with values of x_α and x_β as arguments. The ϕ_ν are of the form $u - 1/2$ with u taking values of x_α, and tensor products of these terms when interactions are present. See Wahba, Wang, Gu, Klein and Klein (1995). In each of the examples below, only a subset of the n representers $Q_{\lambda i}$ were used, chosen by the clustering method previously described. In every case K was 25, and $2K = k = 50$ was sufficient to meet the tolerance requirement. The randomized trace estimates were based on iid $\mathscr{N}(0, \sigma_\varepsilon^2)$ random variables. In theory the randomized trace estimates based on one step of the Newton-Raphson iteration are independent of σ_ε^2, since they are linear in ε, although in the (multi-step) method which iterates $f_\lambda^{Y+\varepsilon}$ to convergence they are not. In practice the (one step) randomized trace estimates were found to be insensitive to σ_ε^2 over seven orders of magnitude. In the multi-step method, some experimentation in Xiang and Wahba (1997) found that $\sigma_\varepsilon = .001$ worked well. $\sigma_\varepsilon = .001$ is used in most of the simulations below.

6.1. *Replicates of $ranGACV$ follow CKL.* Figure 1 is an example illustrating the ability of $ranGACV(\lambda)$ to follow $CKL(\lambda)$ in a simple univariate case. $n = 500$ observations were generated based on $t \in [0, 1]$, $f(t) = 2\sin 10t$, with $t(i) = (i - 1/2)/500$, $i = 1, \dots, n$.

The solid curves in both panels are a plot of the $CKL(\lambda)$ for this data set with the minimum marked with a filled in square. Each of the 10 dashed lines in panel (a) represents a plot of $ranGACV(\lambda)$ using $R = 1$ replicate of ε to compute $ranGACV$, and the minimum of each is marked with a circle. It can be seen that any of these ten versions of the $ranGACV$ provides a rather good estimate of the λ that minimizes the CKL. Panel (b) presents results from the same experiment except that this time the number R of replicates in (22) was taken as 5. It can be seen that all 10 minimizers of the 10 $ranGACV$ curves are even more reliable estimates of the minimizer of CKL. The direct calculation of Σ_λ in order to calculate $traceH$ directly is not feasible with n this large because the calculation is very ill-posed. [An explicit formula for Σ_λ appears in Xiang and Wahba (1997), equation (3.4).]

6.2. *$ranGACV$ behaves similarly to the multi-step randomized $GACV$.* Figure 2 provides a comparison, based on the CKL, of the $ranGACV$ based on one step of the Newton-Raphson iteration versus the multi-step randomized $GACV$ as described in Xiang and Wahba (1996).

The multi-step randomized $GACV$ involves carrying the Newton-Raphson iteration to convergence to obtain $f_\lambda^{Y+\varepsilon}$. Plotted is the CKL. Each point represents one observational data set. The observations were generated from $f(t) = 2sin10t$ as before with $n = 500$ equally spaced points. It can be seen that based on the CKL the two randomized versions are about equally good. Since the (one step) $ranGACV$ is faster to compute, we have adopted it in the studies here. This figure and Figure 1 were typical of a number of simulation studies, with different “truth functions.”

6.3. *Minimizing $ranGACV(\lambda_1, \lambda_2)$.* In order to implement the method in large data sets with multiple smoothing parameters a workable method for finding the minimizer of $ranGACV$ which does not use derivatives is necessary. After a fair amount of experimentation [see also Gong, Wahba, Johnson and Tribbia (1998), Lin (1998a)], we have found that the downhill simplex method works well for $ranGACV$ functions encountered in the demographic data sets with Bernoulli data that we have analyzed. Starting guesses for the downhill simplex method may be obtained by trial and error, via a default (e.g., $\log \lambda_\beta = -5$) which works well at the scale of the present experiments, or via computer experimental design methods, see Bowman, Sacks and Chang (1993). In computer experimental design methods a multivariate design is selected, for example, a Latin hypercube design, and the function to be minimized is evaluated at the design points. It is then interpolated via some appropriate multivariate interpolation scheme using functions that can be minimized easily, and the minimum of the interpolant found. Figure 3 describes a typical two smoothing parameter case.

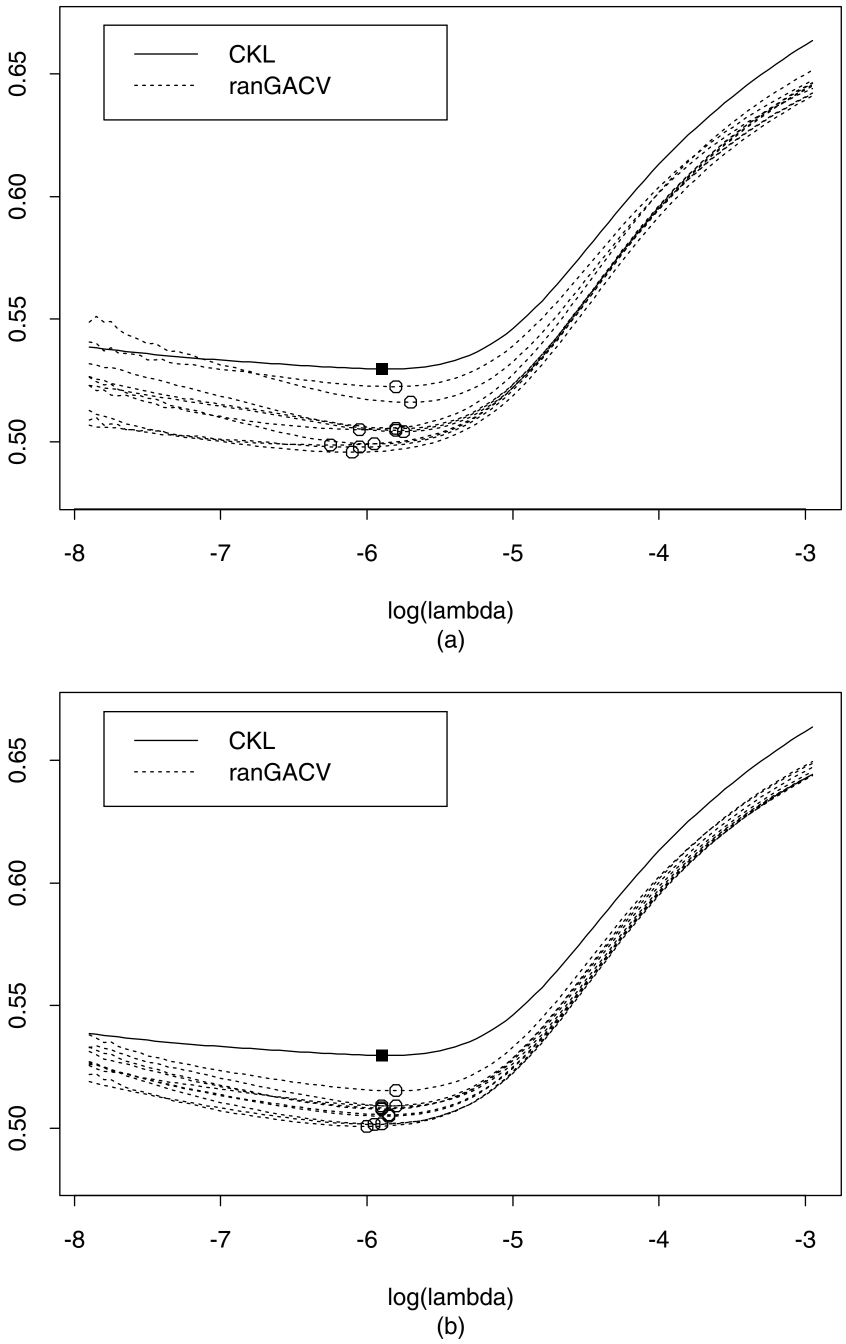


FIG. 1. 10 *replicates of* $ranGACV(\lambda)$ *compared with* $CKL(\lambda)$: (a) *One replicate for each curve.* (b) *Five replicates for each curve.*

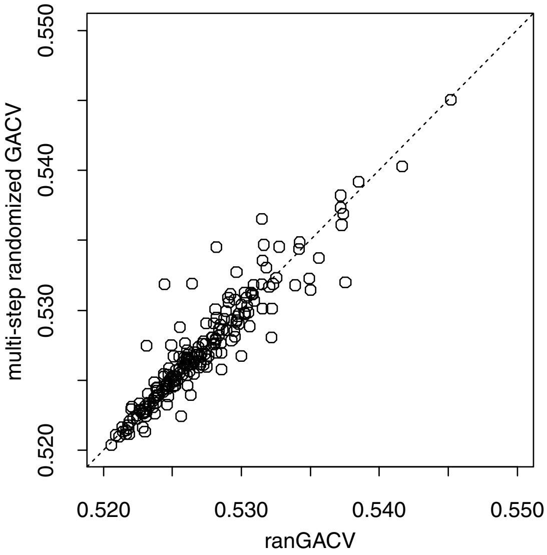


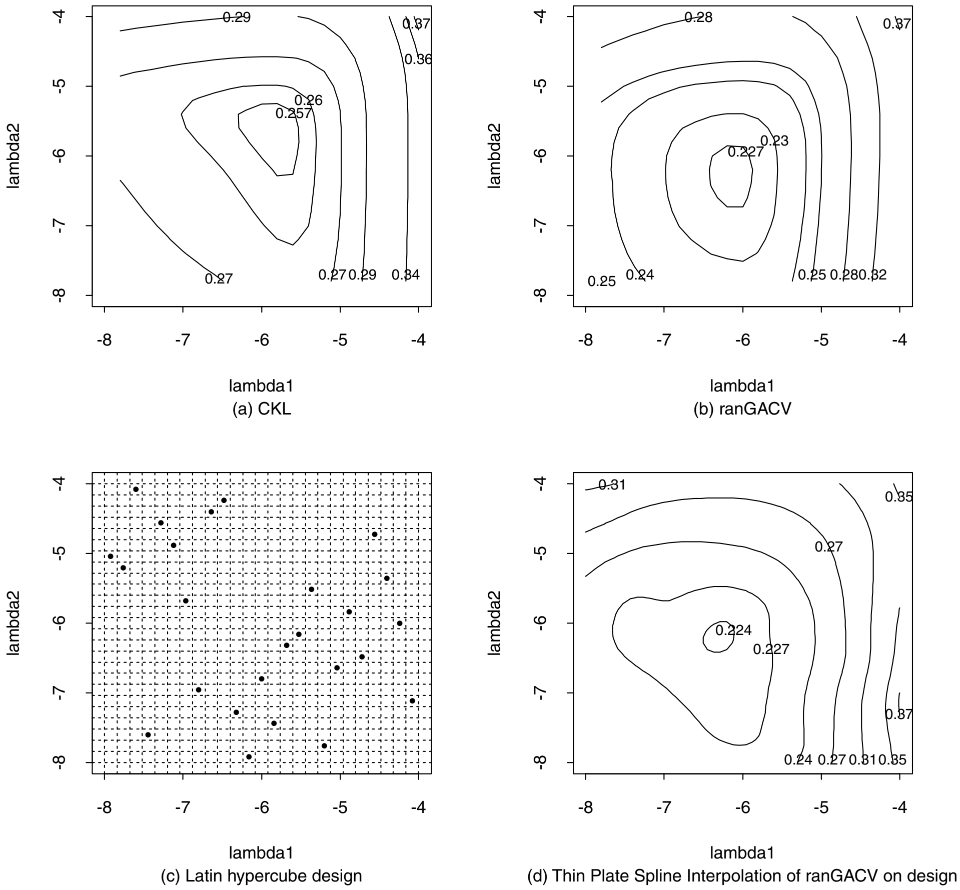
FIG. 2. *CKL comparison of the (one step) ranRGACV with the multi-step randomized GACV.*

Data were generated according to the additive model $f(x_1, x_2) = \sin x_1 - \sin x_2$ with $n = 500$ and the values of $t = (x_1, x_2)$ chosen according to a uniform distribution on the unit square. An additive cubic spline model was fit, with one smoothing parameter for each variable. Panel (a) gives the $CKL(\lambda_1, \lambda_2)$ and panel (b) gives the corresponding $ranGACV(\lambda_1, \lambda_2)$. It can be seen that the inefficiency of the $GACV$ estimate, as judged by

$$\frac{CKL(\hat{\lambda_1}, \hat{\lambda_2})}{min_{\lambda_1, \lambda_2} CKL(\lambda_1, \lambda_2)}$$

is very close to 1. Panel (c) gives an example of a Latin hypercube design, and panel (d) gives a thin plate spline interpolant to the $ranGACV$ function at the design points, which could be used to provide starting guesses for a downhill simplex search. For examples tried with dimensions $p = 3, \ldots, 6$, we have found a least squares quadratic polynomial interpolant adequate. Other methods are discussed in Bowman, Sacks and Chang (1993).

6.4. *Comparison of ranGACV with UBR of GCVPACK, two smoothing parameters.* Figure 4 gives the results of a comparison of the $ranGACV$ and the iterative UBR estimate of GRKPACK. Here 200 data sets of $n = 500$ observations each, from the model of the previous example were generated. For each data set, the CKL was obtained both for the UBR estimate and the $ranGACV$ estimate. The $ranGACV$ estimate was also based on a subset of $k = 50$ basis functions and $R = 5$ replications of the randomized trace estimate. Each point represents one comparison. It can be seen that the CKL is roughly the same for about 90% of the cases, while in the preponderance of the remaining cases



(a) CKL

(b) ranGACV

(c) Latin hypercube design

(d) Thin Plate Spline Interpolation of ranGACV on design

FIG. 3. *$CKL(\lambda_1, \lambda_2)$, $ranGACV(\lambda_1, \lambda_2)$ and thin plate spline interpolation of ranGACV on a Latin hypercube design.*

the *ranGACV* estimate is better. This kind of result is typical of a number of similar experiments.

6.5. *Simulation experiments with realistic designs.*

6.5.1. *Pima Indian Diabetes data set.* Typical designs in observational studies tend to be very irregular. This example and the example in the next section are used to compare the *ranGACV* estimate with the *UBR* estimate in GRKPACK on more realistic simulated data sets. This example is based on the Pima Indian Diabetes data set from the UCI Repository of Machine Learning Databases (`http://www.ics.uci.edu/~mlearn/MLRepository.html`) that was analyzed in Wahba, Gu, Wang and Chappell (1995). Here we use the same randomly chosen subset of $n = 500$ of the subjects that was used as the training set in Wahba, Gu, Wang and Chappell (1995). Only two variates, x_1 = body mass index (`bmi`) and x_2 = plasma glucose concentration (`pgc`)

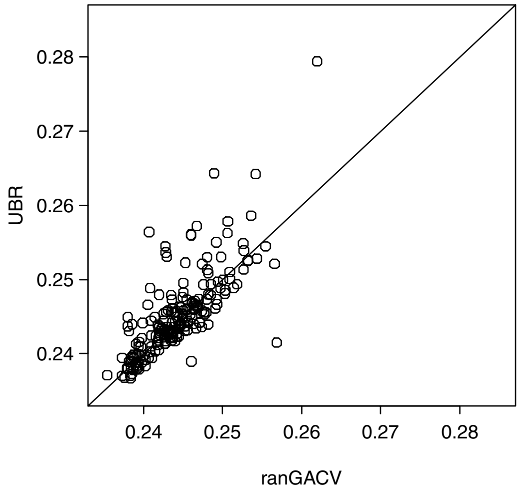


FIG. 4. *ranGACV vs. UBR comparison of the CKL*, 200 *runs*.

are used here. The response is a positive or negative test for diabetes. The smoothing spline ANOVA model

$$f(x_1, x_2) = C + f_1(x_1) + f_2(x_2) + f_{12}(x_1, x_2) \tag{36}$$

was fitted using GRKPACK and then used as the "truth" function. Five smoothing parameters were used in this model, one each for the main effects and three for the interaction term. [The interaction term has three components, smooth⊗parametric, parametric⊗smooth and smooth⊗smooth, see Wahba, Wang, Gu, Klein and Klein (1995), pages 1870–1871.] A downhill simplex search with starting guesses $\log \lambda_p = -5$ and changing each $\log \lambda_p$ successively to -4 to get the simplex was used. Figure 5 gives a scatter plot of the two covariates, and the fitted probability surface $\mu(\hat{f})$.

We generated 200 data sets from $\mu(\hat{f})$ and the design of Figure 5 and fitted each set using GRKPACK along with *ranGACV*. The *ranGACV* estimate used $k = 50$ basis functions and $R = 5$ replicates in the randomized trace calculations. The results are plotted in Figure 6.

For 8 of the 200 replications the *UBR* algorithm failed to converge, so that the plot has 192 points. It can be seen that about 90–95% of the points form a roughly circular cloud, which indicates that the *UBR* and *ranGACV* estimates do about equally well, with the remaining 5–10% of points indicating a larger *CKL* for the *UBR* estimate.

6.5.2. *The Wisconsin Epidemiologic Study of Diabetic Retinopathy* (*WESDR*). This example is based on a model obtained from data from the WESDR study. This study is described in more detail in Wahba, Wang, Gu, Klein and Klein (1995) and Klein, Klein, Moss, Davis and Demets (1984). The predictor variables are duration of diabetes (`dur`), glycosylated hemoglobin (`gly`) and body mass index (`bmi`) at baseline. The response is four year pro-

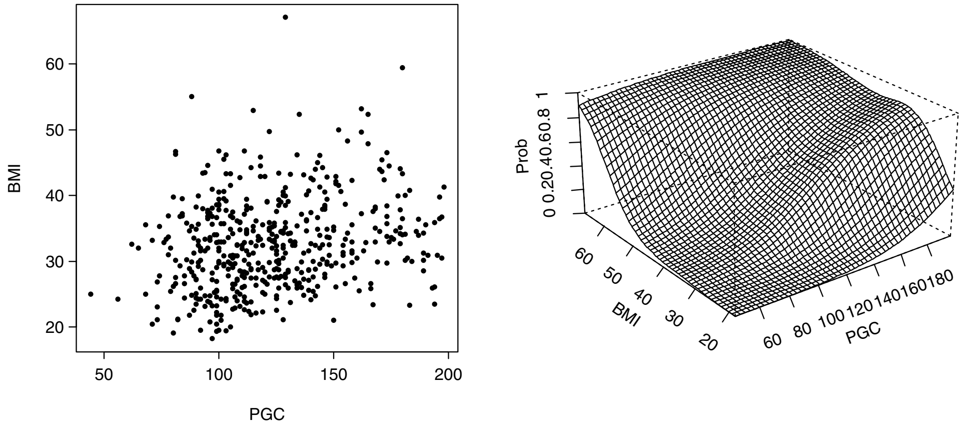


FIG. 5. *Left*: *Scatter plot of the covariates*. *Right*: *"Truth" probability surface* $\mu(\hat{f})$, *Pima Indian example*.

gression of diabetic retinopathy. $n = 669$ subjects were in the data set that was analyzed in Wahba, Wang, Gu, Klein and Klein (1995), and that data set is available as part of the documentation for GRKPACK. The following smoothing spline ANOVA model was fitted to this data using GRKPACK with *UBR*:

$$f(\mathtt{dur},\ \mathtt{gly},\mathtt{bmi}) = C + f_1(\mathtt{dur}) + f_2(\mathtt{gly}) + f_3(\mathtt{bmi}) + f_{13}(\mathtt{dur},\mathtt{bmi}). \tag{37}$$

This model is similar to the model in Wahba, Wang, Gu, Klein and Klein (1995) the only difference being that there $f_2(\mathtt{gly})$ was replaced by $const \cdot \mathtt{gly}$, since it was found there that the fitted $f_2(\mathtt{gly})$ was visually indistinguishable from a straight line. Thus (37) represents a conservative approach for our example below. The right panel of Figure 7 gives this fitted probability surface, as a function of $(\mathtt{bmi},\mathtt{dur})$ with $\mathtt{gly}$ fixed at its median. The surface is plotted only over the region for which the posterior standard deviation (in f) is .5 or less, although the entire surface is retained for the purposes of this experiment. For this experiment data were simulated from this fitted model at the design points of the original data set. The left panel gives a scatterplot of the $\mathtt{bmi}$, $\mathtt{dur}$ design. The first simulated data set has been used to mark simulated 1 responses by filled in circles and the simulated 0 responses by open circles.

100 data sets were simulated. Then the ANOVA model is fitted for each data set, first, using GRKPACK and then *ranGACV* with $k = 50$ and $R = 5$, and the *CKL* for each fit is determined. There are six smoothing parameters. For the *ranGACV* estimate the same downhill simplex search as in the previous example was used. Figure 8 shows the comparison results.

The results are similar to those in the previous examples. In about 90%–95% of the replicates GRKPACK and the *ranGACV* with approximating basis functions give about the same *CKL* values, while in the remaining cases *UBR* estimates are worse.

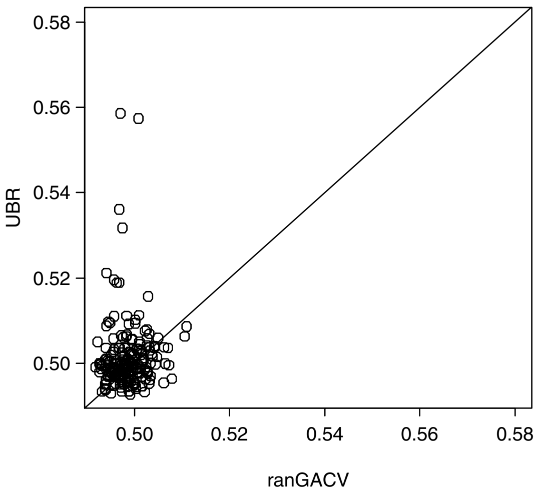


FIG. 6. *Pima Indian example*: *UBR vs ranRGACV compared on the basis of the CKL*, 192 *simulated data sets*.

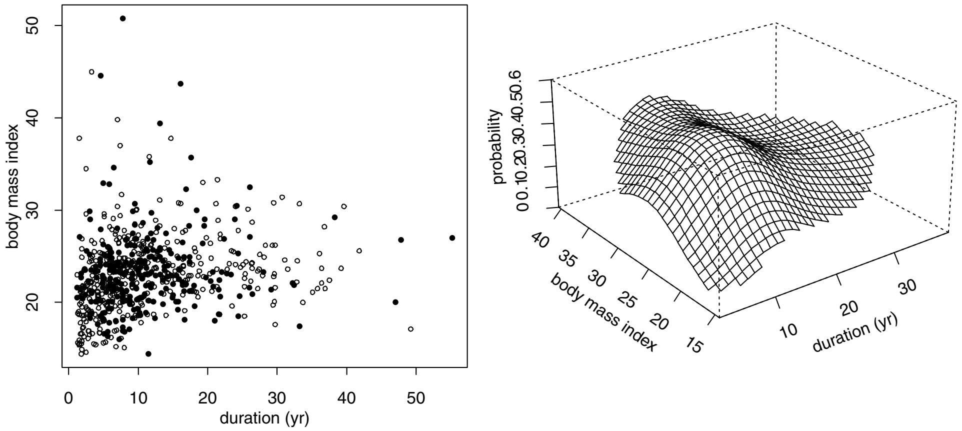


FIG. 7. *Left*: *scatter plot of responses from one simulated data set*, *as a function of* dur *and* bmi. *Right*: *"Truth" probability surface*, *as a function of* dur *and* bmi, *at the median value of* gly, *plotted over a restricted region*.

Several other experiments were performed with this experimental data set. Two other search methods were tried with the $ranGACV$, first a computer experimental design method followed by a downhill simplex search, and secondly a two stage computer experimental design method. The computer experimental design method used 28 design points and a least squares fit to a quadratic polynomial. The two stage method generated a new set of points around the minimum of the quadratic polynomial from the first stage. While the minimizer varied slightly by search method, the three pairwise comparisons based on the *CKL* showed that, based on this criterion they were all close, and no one method was superior to the others. Since the downhill simplex method is conceptually and practically simpler, we are continuing to use it.

Returning to Figure 8, several data sets where the *CKL* from the two fitting methods was the same were selected and cross sections of the actual fits [corresponding to the cross sections presented in Wahba, Wang, Gu, Klein and Klein (1995)] were compared visually. All the cross sectional curves were found to be essentially visually indistinguishable. From this we conclude that the approximate solution using $k = 50$ basis functions is almost identical to the exact smoothing spline ANOVA solution using all the basis functions. From this one can make the reasonable assumption that in the cases where the *CKL* from the two methods is noticeably different, the reason is due to the method for choosing the smoothing parameters and not the approximation using a smaller number of basis functions.

The Bayesian confidence intervals were computed for the first replication where the two methods had the same *CKL*, and the fractions of the $n = 669$ data points for which the confidence intervals covered the true $f_{\hat{\lambda}i}$ (equiva-

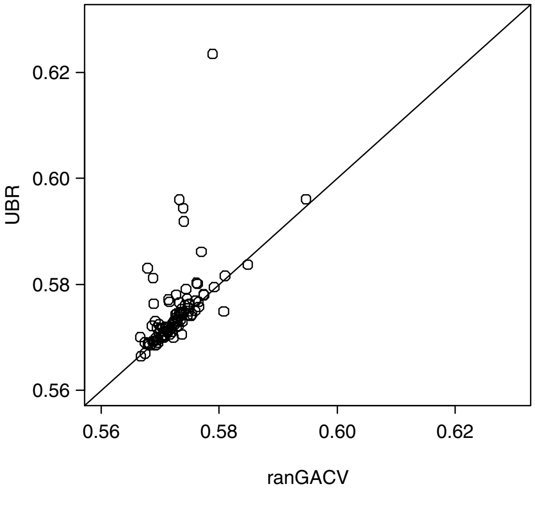


FIG. 8. *WESDR example: UBR and $ranGACV$ compared on the basis of the CKL,* 100 *simulated data sets.*

TABLE 1
Bayesian "confidence interval" coverage

Nominal Coverage	**GRKPACK**	**ranGACV, $R = 5$**
90%	88%	87%
95%	96%	94%

lently $\mu_{\hat{\lambda}i}$) were determined. Rather modest differences were found. The results are seen in Table 1.

The slightly wider coverage from GRKPACK may reflect the fact that the GRKPACK estimate is (slightly) more flexible.

7. Pigmentary abnormalities in the Beaver Dam study.

The Beaver Dam Eye Study is an ongoing population-based study of age-related ocular disorders. Subjects were a group of 4926 people aged 43–86 years at the start of the study who lived in Beaver Dam, WI and were examined at baseline, between 1988 and 1990. A description of the population and details of the study at baseline may be found in Klein, Klein, Linton and Demets (1991). Five year followup data is presently being analyzed; see, for example, Klein, Klein, Jensen and Meuer (1997). Here we consider only the $n = 2585$ women members of this cohort, and the baseline observations. We examine the association of pigmentary abnormalities with six other attributes $t = (x_1, x_2, \dots, x_6)$ at baseline. $\mu(t)$ is the probability that a subject with attribute vector t at baseline will be found to have a pigmentary abnormality in at least one eye, at baseline. Pigmentary abnormalities are an early sign of age-related macular degeneration and are defined by the presence of retinal depigmentation and increased retinal pigmentation. See Klein, Klein, Jensen and Ritter (1994), Klein, Klein and Linton (1992). 11.88% of the $n = 2585$ cohort studied here showed evidence of a pigmentary abnormality. The six "predictor" variables are shown in Table 2.

The model fitted here is

$$\begin{aligned} f(t) &= C + f_1(\mathtt{sys}) + f_2(\mathtt{chol}) + f_{12}(\mathtt{sys},\ \mathtt{chol}) \\ &\quad + d_{\mathtt{age}} \cdot \mathtt{age} + d_{\mathtt{bmi}} \cdot \mathtt{bmi} + d_{\mathtt{horm}} \cdot I_1(\mathtt{horm}) + d_{\mathtt{drin}} \cdot I_2(\mathtt{drin}). \end{aligned} \tag{38}$$

TABLE 2
Predictor variables for the Beaver Dam Pigmentary Abnormalities Model

Variable	**units**	**code**
current usage of hormone replacement therapy	yes/no	`horm`
history of heavy drinking	yes/no	`drin`
body mass index	kg/m^2	`bmi`
age	*years*	`age`
systolic blood pressure	$mmHg$	`sys`
serum cholesterol	mg/dL	`chol`

TABLE 3
Quantiles of the predictor variables

Percentile	Min	12.5	25	37.5	50	62.5	75	87.5	Max
sys	71	108	116	123	129	137	145	157	221
chol(mg/dL)	102	191	210	225	237	251	266	290	503
bmi (kg/m^2)	15	22.5	24.3	25.9	27.5	29.5	31.6	35.2	68.4
age $(years)$	43	48	52	58	62	67	71	77	86

I_1 and I_2 are indicator variables. Thus, there are 5 smoothing parameters [3 for the f_{12} interaction term which consists of $linear_{\texttt{sys}} \otimes smooth_{\texttt{chol}}$, $smooth_{\texttt{sys}} \otimes linear_{\texttt{chol}}$ and $smooth_{\texttt{sys}} \otimes smooth_{\texttt{chol}}$ terms, see Wahba, Wang, Gu Klein and Klein (1995)]. Originally, age and bmi were fitted as smooth main effects, but visual inspection of the smooth terms $f_3(\texttt{age})$ and $f_4(\texttt{bmi})$ indicated that they indistinguishable from linear terms, so that they were set to be linear in the final model.

Figures 9 and 10 give the estimated probability of finding pigmentary abnormalities as a function of chol, for various values of bmi, age and sys. In Figure 9, (horm, drin) = (no,no) and in Figure 10 (horm, drin) = (yes,no). Figure 11 gives cross sectional plots of the estimated probabilities along with the Bayesian confidence intervals intervals as a function of chol for four values of age, and both values of horm.

For reference, Table 3 gives the quantiles of the continuous predictor variables. A protective effect of hormones is evident, and a suggestion of a (nonlinear) protective effect of cholesterol, particularly at older values of age in the horm = no group may be seen as a result of fitting this model.

8. Summary and conclusions.

We have proposed a new method, the randomized Generalized Approximate Cross Validation ($ranGACV$) for choosing multiple smoothing parameters in the Bernoulli case. A clustering procedure for selecting an approximating set of basis functions is also proposed. These two techniques, taken together, make possible the fitting of Smoothing Spline ANOVA models for very large data sets. We have shown by example that the randomization technique coupled with the use of the approximating set of basis functions can substantially reduce computing requirements, without any appreciable loss of accuracy in solving the SS-ANOVA penalized likelihood equations. Furthermore the minimizer $\hat{\lambda}$ of $ranGACV$ is shown via a series of simulation experiments to be a good estimate of the smoothing parameters minimizing the *CKL* distance of the distribution defined by the estimate $f_{\hat{\lambda}}$ to the distribution defined by the true f. Care was taken that some of the simulations involved data scattered in a manner realistic of other, similar demographic studies. Our examples have demonstrated that the downhill simplex method coupled with computer experimental design techniques allow the efficient searching for optimal smoothing parameters in the multiple smoothing parameter case. The Bayesian "confidence intervals" are derived

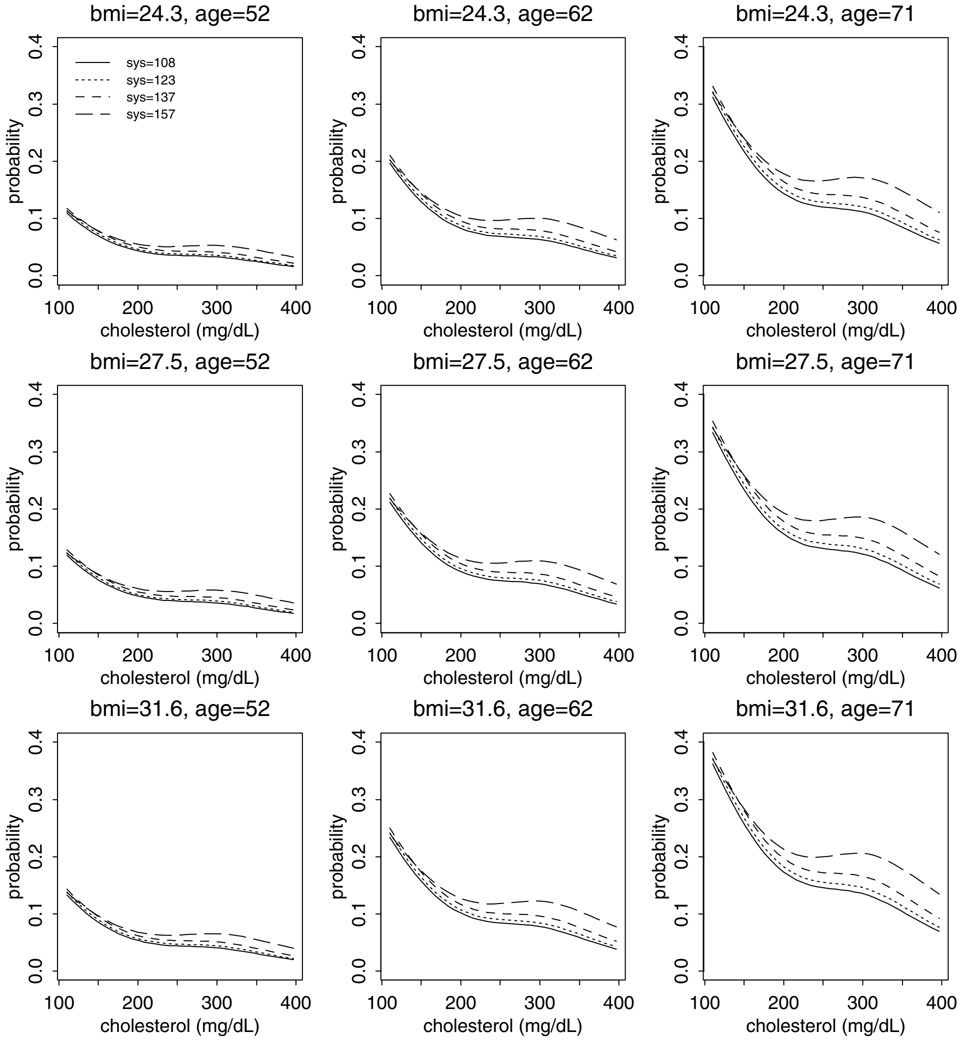


FIG. 9. *Estimated probability of pigmentary abnormality as a function of cholesterol by three levels of* bmi *and* age *and four levels of* sys, horm=*no*, drin=*no*.

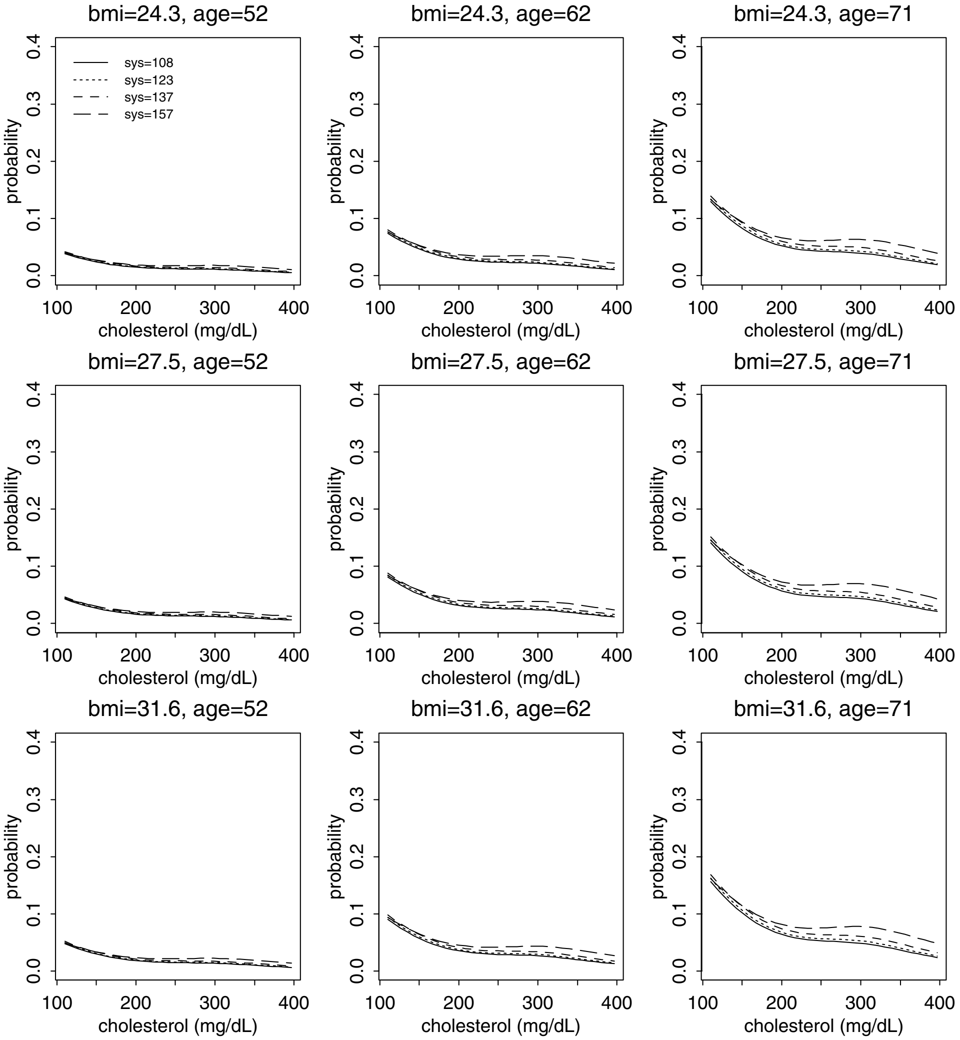


FIG. 10. *Estimated probability of pigmentary abnormality as a function of cholesterol by three levels of* `bmi` *and* `age` *and four levels of* `sys`, `horm`=*yes*, `drin`=*no*.

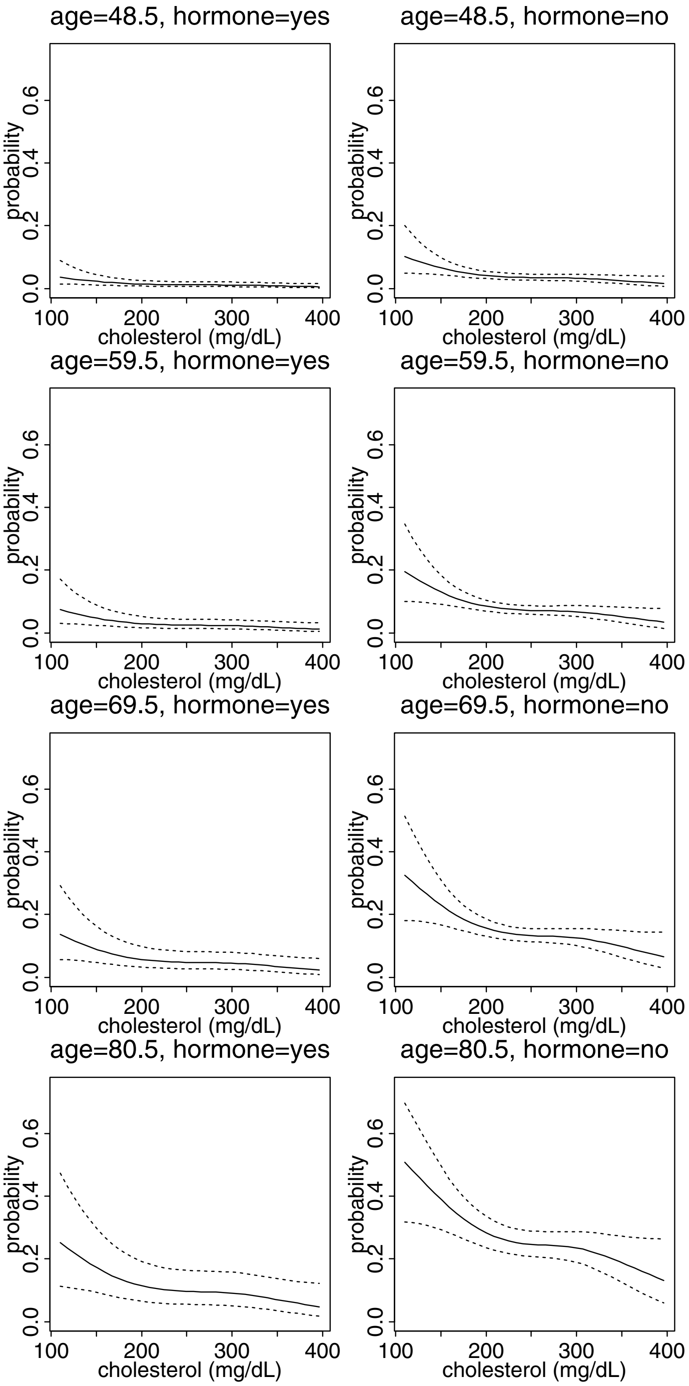


FIG. 11. *Bayesian confidence intervals as a function of* `chol`, `age`, `horm`. `bmi`, `sys` *fixed at their medians and* `drin` = *no*.

for the SS-ANOVA case based on the approximating set of basis functions following an argument of Silverman. It is shown that they are easily computed along with the estimate itself, but more importantly, the numerical experiments show that they appear to possess the "across the function property" common to their counterparts based on the entire set of basis functions. Finally, we have demonstrated that the overall algorithm can be applied to an important data set, with results that provide interesting insights into the data set. It is concluded that the techniques developed here provide a useful and effective method for statistical model fitting with favorable statistical properties as judged by the observed comparative Kullbak-Liebler distance and coverage properties of the Bayesian "confidence intervals." We also argue that the suite of techniques presented has potential for model fitting as well as model selection in more general settings than those studied here.

Acknowledgments. The authors wish to thank a referee, Associate Editor and the Editor for their careful reading of the manuscript and many helpful suggestions for improving the exposition.

REFERENCES

BOWMAN, K., SACKS, J. and CHANG, Y. (1993). Design and analysis of numerical experiments. *J. Atmos. Sci.* **50** 1267–1278.

BRUMBACK, B. and RICE, J. (1998). Smoothing spline models for the analysis of nested and crossed samples of curves. *J. Amer. Statist. Assoc.* **93** 961–991.

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. (1986). How biased is the apparent error rate of a prediction rule? *J. Amer. Statist. Assoc* **81** 461–470.

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

GAO, F. (1999). Penalized multivariate logistic regression with a large data set. Ph. D. dissertation, Dept. Statistics, Univ. Wisconsin–Madison.

GIRARD, D. (1998). Asymptotic comparison of (partial) cross-validation, GCV and randomized GCV in nonparametric regression. *Ann. Statist.* **26** 315–334.

GONG, J., WAHBA, G., JOHNSON, D. and TRIBBIA, J. (1998). Adaptive tuning of numerical weather prediction models: simultaneous estimation of weighting, smoothing and physical parameters. *Monthly Weather Rev.* **125** 210–231.

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

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

GU, C. (1998). Structural multivariate function estimation: Some automatic density and hazard estimates. *Statist. Sinica* **8** 317–336.

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

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

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, Australia.

KIMELDORF, G. and WAHBA, G. (1971). Some results on Tchebycheffian spline functions. *J. Math. Anal. Appl.* **33** 82–95.

KLEIN, B. E. K., KLEIN, R., JENSEN, S. and RITTER, L. (1994). Are sex hormones associated with age-related maculopathy in women? The Beaver Dam Eye Study. *Trans. Amer. Ophth. Soc.* **92** 289–297.

KLEIN, R., KLEIN, B. E. K. and LINTON, K. (1992). Prevalence of age-related maculopathy: The Beaver Dam Eye Study. *Ophthalmalogy* **99** 933–942.

KLEIN, R., KLEIN, B. E. K., JENSEN, S. and MEUER, S. (1997). The five-year incidence and progression of age-related maculopathy: The Beaver Dam eye study. *Ophthalmology* **104** 7–21.

KLEIN, R., KLEIN, B. E. K., LINTON, K. and DEMETS, D. (1991). The Beaver Dam eye study: Visual acuity. *Ophthalmology* **98** 1310–1315.

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

LIN, X. (1998a). Smoothing spline analysis of variance for polychotomous response data. Ph.D. dissertation, Dept. Statistics, Univ. Wisconsin–Madison.

LIN, Y. (1998b). Tensor product space ANOVA models. *Ann. Statist.* **28** 734–755.

LIN, Y. (1998c). Tensor product space ANOVA models in multivariate function estimation. Ph.D. dissertation, Univ. Pennsylvania, Philadelphia PA.

LUO, Z. (1998). Backfitting in smoothing spline ANOVA. *Ann. Statist.* **26** 1733–1759.

LUO, Z. and WAHBA, G. (1997). Hybrid adaptive splines. *J. Amer. Statist. Assoc.* **92** 107–114.

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

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

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

PRESS, W., TEUKOLSKY, S., VETTERLING, W. and FLANNERY, B. (1992). *Numerical Recipes in Fortran 77: The Art of Scientific Computing*. Cambridge Univ. Press.

SILVERMAN, B. (1985). Some aspects of the spline smoothing approach to non-parametric regression curve fitting. *J. Roy. Statist. Soc. Ser. B* **47** 1–52.

VERBYLA, A., CULLIS, B., KENWARD, M. and WELHAM, S. (1997). The analysis of designed experiments and longitudinal data using smoothing splines. Technical Report 97/4, Dept. Statistics, Univ. Adelaide.

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., 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* (D. Wolpert, ed.) 329–360. Addison-Wesley, Reading, MA.

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

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

WANG, Y. (1997). GRKPACK: Fitting smoothing spline analysis of variance models to data from exponential families. *Comm. Statist. Simul. Comput.* **26** 765–782.

WANG, Y. (1998). Smoothing spline models with correlated random errors. *J. Amer. Statist. Assoc.* **93** 341–348.

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

WANG, Y., WAHBA, G., GU, C., KLEIN, R. and KLEIN, B. (1997). Using smoothing spline ANOVA to examine the relation of risk factors to the incidence and progression of diabetic retinopathy. *Statistics in Medicine* **16** 1357–1376.

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

WOOD, S. and KOHN, R. (1998). A Bayesian aproach to robust binary nonparametric regression. *J. Amer. Statist. Assoc.* **93** 203–213.

XIANG, D. (1996), Model fitting and testing for non-Gaussian data with a large data set. Ph.D. dissertation, Univ. Wisconsin–Madison.

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

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

XIANG, D. and WAHBA, G. (1997). Approximate smoothing spline methods for large data sets in the binary case. In *Proceedings of the 1997 ASA Joint Statistical Meetings, Biometrics Section* 94–98. Amer. Statist. Assoc., Alexandria, VA.

YE, J. (1998). On measuring and correcting the effects of data mining and model selection. *J. Amer. Statist. Assoc.* **93** 120–131.

YE, J. and WONG, W. (1997a). Evaluation of highly complex modeling procedures with Binomial and Poisson data. Unpublished manuscript.

YE, J. and WONG, W. (1997b) Model uncertainty and correcting for selection bias. Unpublished manuscript.

X. LIN
CENDANT CORPORATION
707 SUMMER STREET
STAMFORD CONNECTICUT 06901
E-MAIL: xlin@cms.cendant.com

G. WAHBA
F. GAO
DEPARTMENT OF STATISTICS
UNIVERSITY OF WISCONSIN
1210 W. DAYTON STREET
MADISON, WISCONSIN 53706
E-MAIL: wahba@stat.wisc.edu
fgao@stat.wisc.edu

D. XIANG
SAS INSTITUTE INC.
SAS CAMPUS DRIVE, BLDG R
CARY, NORTH CAROLINA 27513
E-MAIL: sasdxx@wnt.sas.com

R. KLEIN, MD
B. KLEIN, MD
DEPARTMENT OF OPHTHALMOLOGY
UNIVERSITY OF WISCONSIN
610 N. WALNUT STREET
MADISON, WISCONSIN 53705
E-MAIL: kleinr@epi.ophth.wisc.edu
kleinb@epi.ophth.wisc.edu