GRKPACK: FITTING SMOOTHING SPLINE ANOVA MODELS FOR EXPONENTIAL FAMILIES

Yuedong Wang

Department of Biostatistics, University of Michigan,
Ann Arbor, Michigan 48109, U.S.A.

Key Words and Phrases: generalized cross validation; Newton-Raphson iteration; RKPACK; smoothing parameter; smoothing spline ANOVA; unbiased risk estimate.

ABSTRACT

Wahba, Wang, Gu, Klein and Klein (1995) introduced Smoothing Spline ANalysis of VAriance (SS ANOVA) method for data from exponential families. Based on RKPACK, which fits SS ANOVA models to Gaussian data, we introduce GRKPACK: a collection of Fortran subroutines for binary, binomial, Poisson and Gamma data. We also show how to calculate Bayesian confidence intervals for SS ANOVA estimates.

1 Introduction

Generalized linear models (GLM's) for analysis of data from exponential families have been extensively studied and widely used since 1970's (Nelder and Wedderburn, 1972; McCullagh and Nelder, 1989). As the popularity of these methods has increased, so has the need for more sophisticated model building and diagnostic checking techniques. In the context of nonparametric estimate of the GLM regression surface, O'Sullivan, Yandell and Raynor (1986) and

Gu (1990) used penalized likelihood method with smoothing splines and thin plate splines. Hastie and Tibshirani (1990) used additive models. Wahba et al. (1995) introduced the SS ANOVA models using the penalized likelihood and Smoothing Spline ANalysis of Variance methods. See also Wahba, Gu, Wang and Chappell (1994a), Wahba, Wang, Gu, Klein and Klein (1994b), Wang (1994), Wang, Wahba, Chappell and Gu (1995) and Wang, Wahba, Gu, Klein and Klein (1996) for details of SS ANOVA models. In this paper, we describe a package for estimations of the SS ANOVA models with binary, binomial, Poisson and Gamma data. We call this package as GRKPACK, which stands for generalized RKPACK.

First, we describe the computational part of the SS ANOVA model. Suppose data have the form $(y_i; \boldsymbol{t}_i)$, $i = 1, 2, \cdots, n$, where y_i are independent observations and $\boldsymbol{t}_i = (t_{1i}, \cdots, t_{di})$. The distribution function of y_i is from an exponential family with density function

$$g(y_i; f_i, \phi) = \exp((y_i h(f_i) - b(f_i))/a(\phi) + c(y_i, \phi)),$$

where $f_i = f(\boldsymbol{t}_i)$ is the parameter of interest and $h(f_i)$ is a monotone transformation of f_i known as the *canonical parameter*. ϕ is an unknown scale parameter. Let $\boldsymbol{t} = (t_1, ..., t_d)$, and let $t_j \in \mathcal{T}^{(j)}$, where $\mathcal{T}^{(j)}$ is a measurable space. Let $\mathcal{T} = \mathcal{T}^{(1)} \otimes \cdots \otimes \mathcal{T}^{(d)}$, then $\boldsymbol{t} \in \mathcal{T}$. Denote the log likelihood given y_i and $\boldsymbol{t}_i$ as

$$l_i(f_i) = \log g(y_i; f_i, \phi) = (y_i h(f_i) - b(f_i))/a(\phi) + c(y_i, \phi). \tag{1}$$

The purpose is to investigate the global relationship between f and $\boldsymbol{t}$.

Let $d\mu_j$ be a probability measure on $\mathcal{T}^{(j)}$ and let $\mathcal{H}^{(j)}$ be a reproducing kernel Hilbert space (RKHS) (Aronszajn, 1950) of functions on $\mathcal{T}^{(j)}$ with $\int_{\mathcal{T}^{(j)}} f_j(t_j) d\mu_j = 0$ for $f_j(t_j) \in \mathcal{H}^{(j)}$. Let $\{1^{(j)}\}$ be the one dimensional space of constant functions on $\mathcal{T}^{(j)}$. Consider the RKHS

$$\mathcal{G} \;=\; \prod_{j=1}^{d} (1^{(j)} \oplus \mathcal{H}^{(j)})$$

$$= \{1\} \oplus \sum_j \mathcal{H}^{(j)} \oplus \sum_{j<k} (\mathcal{H}^{(j)} \otimes \mathcal{H}^{(k)}) \oplus \cdots,$$

where $\{1\}$ denotes the constant functions on $\mathcal{T}$. An element f_j in $\mathcal{H}^{(j)}$ is called a *main effect*, an element f_{jk} in $\mathcal{H}^{(j)} \otimes \mathcal{H}^{(k)}$ is called a *two factor interaction*, and so on.

Similar to the usual ANOVA, model space is a subspace $\mathcal{M}$ of $\mathcal{G}$. By deleting some higher-order interactions, we get less flexible, but more "estimable" models. When a model is chosen, we can regroup and write the model space as

$$\mathcal{M} = \mathcal{H}^0 \oplus \sum_{\beta=1}^{q} \mathcal{H}^\beta, \tag{2}$$

where $\mathcal{H}^0$ is a finite dimensional space containing functions which are not going to be penalized, usually lower order polynomials. See Wahba et al. (1995) for details. A SS ANOVA estimate is the solution to the following variational problem:

$$\min_{f \in \mathcal{M}} \left\{ -\sum_{i=1}^{n} l_i(f_i) + \frac{n}{2} \sum_{\beta=1}^{q} \lambda_\beta \|P_\beta f\|^2 \right\}. \tag{3}$$

The first part in (3) is the negative log likelihood. It measures the goodness of fit. In the second part, P_β is the orthogonal projector in $\mathcal{M}$ onto $\mathcal{H}^\beta$ and $\|P_\beta f\|^2$ is a quadratic roughness penalty. λ_β's are a set of smoothing parameters. They control the trade-off between the goodness of fit and the roughness of the estimate. Writing $\lambda_\beta = \lambda/\theta_\beta$, (3) becomes

$$\min_{f \in \mathcal{M}} \left\{ -\sum_{i=1}^{n} l_i(f_i) + \frac{n}{2} \lambda \|P_* f\|_\Theta^2 \right\}, \tag{4}$$

where $P_* = \sum_{\beta=1}^{q} P_\beta$ is the orthogonal projection in $\mathcal{M}$ onto $\mathcal{H}_* = \sum_{\beta=1}^{q} \mathcal{H}^\beta$ and

$$\|f\|_\Theta^2 = \|P_0 f\|^2 + \sum_{\beta=1}^{q} \theta_\beta^{-1} \|P_\beta f\|^2,$$

is a modified norm indexed by $\Theta = (\theta_1, \cdots, \theta_q)$. We denote R_β as the reproducing kernel (RK) (Aronszajn, 1950) for $\mathcal{H}^\beta$ under the original norm. The RK for $\sum_{\beta=1}^q \mathcal{H}^\beta$ under $\| \cdot \|_\Theta$ is

$$R_\Theta = \sum_{\beta=1}^{q} \theta_\beta R_\beta.$$

The solution to (4) has the form (Wahba, 1990; O'Sullivan et al., 1986)

$$f_{\lambda,\Theta}(\boldsymbol{t}) = \sum_{v=1}^{M} d_v \phi_v(\boldsymbol{t}) + \sum_{i=1}^{n} c_i (\sum_{\beta=1}^{q} \theta_\beta R_\beta(\boldsymbol{t}_i, \boldsymbol{t})) = \boldsymbol{\phi}(\boldsymbol{t})^T \boldsymbol{d} + \boldsymbol{\xi}(\boldsymbol{t})^T \boldsymbol{c}, \tag{5}$$

where $\{\phi_v\}_{v=1}^M$ is a set of basis functions of $\mathcal{H}^0$, $M = \dim(\mathcal{H}^0)$, $\boldsymbol{\phi}^T(\boldsymbol{t}) = (\phi_1(\boldsymbol{t}), \cdots, \phi_M(\boldsymbol{t}))$, $\boldsymbol{\xi}^T(\boldsymbol{t}) = (R_\Theta(\boldsymbol{t}_1, \boldsymbol{t}), \cdots, R_\Theta(\boldsymbol{t}_n, \boldsymbol{t}))$. $\boldsymbol{c}_{n\times 1}$ and $\boldsymbol{d}_{M\times 1}$ are vectors of coefficients to be estimated. Substituting (5) into (4), we can estimate $\boldsymbol{c}$ and $\boldsymbol{d}$ by minimizing

$$I(\boldsymbol{c}, \boldsymbol{d}) = -\sum_{i=1}^{n} l_i(\boldsymbol{\phi}^T(\boldsymbol{t}_i)\boldsymbol{d} + \boldsymbol{\xi}^T(\boldsymbol{t}_i)\boldsymbol{c}) + \frac{n}{2}\lambda \boldsymbol{c}^T Q_\Theta \boldsymbol{c}, \tag{6}$$

where $Q_\Theta = \sum_{\beta=1}^q \theta_\beta Q_\beta$ and Q_β's are $n \times n$ matrices with $Q_\beta(i,k) = R_\beta(\boldsymbol{t}_i, \boldsymbol{t}_k)$. Since l_i's are not quadratic, (6) can not be solved directly. But if all $l_i(f_i)$'s are strictly concave, we can use Newton-Raphson procedure to compute $\boldsymbol{c}$ and $\boldsymbol{d}$ for fixed λ and Θ. Let $u_i = -dl_i/df_i$, $w_i = -d^2 l_i/df_i^2$. Let

$$\begin{aligned} \boldsymbol{u}^T &= (u_1, \cdots, u_n), \\ W &= \mathrm{diag}(w_1, \cdots, w_n), \\ S &= (\boldsymbol{\phi}(\boldsymbol{t}_1), \cdots, \boldsymbol{\phi}(\boldsymbol{t}_n))^T. \end{aligned}$$

Then

$$\begin{aligned} \partial I/\partial \boldsymbol{c} &= Q_\Theta \boldsymbol{u} + n\lambda Q_\Theta \boldsymbol{c}, \\ \partial I/\partial \boldsymbol{d} &= S^T \boldsymbol{u}, \\ \partial^2 I/\partial \boldsymbol{c} \partial \boldsymbol{c}^T &= Q_\Theta W Q_\Theta + n\lambda Q_\Theta, \\ \partial^2 I/\partial \boldsymbol{c} \partial \boldsymbol{d}^T &= Q_\Theta W S, \\ \partial^2 I/\partial \boldsymbol{d} \partial \boldsymbol{d}^T &= S^T W S. \end{aligned}$$

The Newton-Raphson iteration satisfies the linear system

$$\begin{pmatrix} Q_\Theta W_- Q_\Theta + n\lambda Q_\Theta & Q_\Theta W_- S \\ S^T W_- Q_\Theta & S^T W_- S \end{pmatrix} \begin{pmatrix} \boldsymbol{c} - \boldsymbol{c}_- \\ \boldsymbol{d} - \boldsymbol{d}_- \end{pmatrix} = \begin{pmatrix} -Q_\Theta \boldsymbol{u}_- - n\lambda Q_\Theta \boldsymbol{c}_- \\ -S^T \boldsymbol{u}_- \end{pmatrix}, \quad (7)$$

where the subscript minus indicates quantities evaluated at the previous Newton-Raphson iteration. Denote

$$\boldsymbol{f} = S\boldsymbol{d} + Q_\Theta \boldsymbol{c} \quad (8)$$

as the vector of estimates of f at design points. Let

$$\begin{aligned} \tilde{Q}_\Theta &= W_-^{1/2} Q_\Theta W_-^{1/2}, \\ \tilde{\boldsymbol{c}} &= W_-^{-1/2} \boldsymbol{c}, \\ \tilde{S} &= W_-^{1/2} S, \\ \tilde{\boldsymbol{d}} &= \boldsymbol{d}, \\ \tilde{\boldsymbol{y}} &= W_-^{-1/2}(W_- \boldsymbol{f}_- - \boldsymbol{u}_-). \end{aligned}$$

(7) can be simplified to

$$\begin{aligned} (\tilde{Q}_\Theta + n\lambda I)\tilde{\boldsymbol{c}} + \tilde{S}\tilde{\boldsymbol{d}} &= \tilde{\boldsymbol{y}}, \\ \tilde{S}^T \tilde{\boldsymbol{c}} &= \boldsymbol{0}. \end{aligned} \quad (9)$$

Choosing appropriate smoothing parameters is crucial for effectively estimating the true function from data by fitting smoothing spline models. The generalized cross validation (GCV) method estimates smoothing parameters by minimizing the GCV score

$$V(\lambda, \Theta) = \frac{1/n \|(I - A(\lambda, \Theta))\tilde{\boldsymbol{y}}\|^2}{[(1/n)tr(I - A(\lambda, \Theta))]^2}, \quad (10)$$

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

$$(w_{1-}^{1/2} f_{\lambda,\Theta}(\boldsymbol{t}_1), \cdots, w_{n-}^{1/2} f_{\lambda,\Theta}(\boldsymbol{t}_n))^T = A(\lambda, \Theta)\tilde{\boldsymbol{y}},$$

and $f_{\lambda,\Theta}(\boldsymbol{t}_i)$'s are computed from the solution of (9).

The unbiased risk (UBR) method estimates smoothing parameters by minimizing the following unbiased risk estimate

$$\tilde{U}(\lambda, \Theta) = \frac{1}{n}\|(I - A(\lambda, \Theta))\tilde{\boldsymbol{y}}\|^2 + 2\frac{\hat{\sigma}^2}{n} trA(\lambda, \Theta), \tag{11}$$

where $\hat{\sigma}^2 = 1/n \sum_{i=1}^{n} u_{i-}^2/w_{i-}$, an estimate of dispersion parameter. If the dispersion parameter is known to be 1, such as in the case of binary data and Poisson data, a better estimate is

$$U(\lambda, \Theta) = \frac{1}{n}\|(I - A(\lambda, \Theta))\tilde{\boldsymbol{y}}\|^2 + \frac{2}{n} trA(\lambda, \Theta). \tag{12}$$

See Wang (1994) and Wang et al. (1995) for detail discussions about the GCV and UBR methods.

2 The Algorithm

A generic code RKPACK (Gu, 1989; Gu and Wahba, 1991) is available to solve (9) and estimate λ and Θ via GCV (option V) or the UBR method at the same time. When using the UBR method, we can either specify $\sigma^2 = 1$ (option U) or estimate σ^2 (option U˜). This suggests the following algorithm:

Algorithm. Given the matrices S, Q_β's, the response vector $\boldsymbol{y}$ and the starting vector $\boldsymbol{f}_0$:

1. Compute $\boldsymbol{u}_-$ and W_-. Compute the transformations $\tilde{S}$, $\tilde{Q}_\beta = W_-^{1/2} Q_\beta W_-^{1/2}$ and $\tilde{\boldsymbol{y}}$;

2. Call RKPACK with inputs $\tilde{S}$, $\{\tilde{Q}_\beta, \ \ \beta = 1, \cdots, q\}$ and $\tilde{\boldsymbol{y}}$. That is, solve (9) and choose λ and Θ by GCV (option V) or the UBR method (option U or option U˜);

3. Compute the new $\boldsymbol{f}$. Stop if the algorithm converges under some criteria (for example, $\sum_{i=1}^{n} w_{i-}((f_i - f_{i-})/(1 + |f_i|))^2 / \sum_{i=1}^{n} w_{i-} < p$ for some

prespecified $p > 0$ is used in our programs) or the number of iterations exceeds some prespecified number L; otherwise go to step 1.

The starting value $\boldsymbol{f}_0$ may be a constant function, a GLM fit or some other estimates. We usually let $p = 10^{-6}$. The algorithm usually takes 5 to 15 iterations to converge. We believe $L = 30$ is big enough for most applications. Since changing λ and Θ at each iteration means modifying the problem successively, convergence is not guaranteed. Nevertheless, the algorithm converges most of the time.

3 Approximate Bayesian Confidence Intervals for SS ANOVA Estimates

In an observational study, often design is not balanced. It is desirable to construct confidence intervals for the SS ANOVA estimate and decide a region in which the estimates are deemed to be reliable. Confidence intervals for components like main effects and interactions are also useful for model selection. Wahba et al. (1995) derived approximate Bayesian confidence intervals for SS ANOVA estimates and showed how to use them in practice.

Let the prior for $f(\boldsymbol{t})$ be

$$F_\xi(\boldsymbol{t}) = \sum_{\nu=1}^{M} \tau_\nu \phi_\nu(\boldsymbol{t}) + b^{\frac{1}{2}} \sum_{\beta=1}^{q} \sqrt{\theta_\beta} Z_\beta(\boldsymbol{t}), \tag{13}$$

where $\boldsymbol{\tau} = (\tau_1, \cdots, \tau_M)^T \sim N(0, \xi I)$, Z_β are independent, zero mean Gaussian stochastic processes, independent of $\boldsymbol{\tau}$, with $\mathrm{E} Z_\beta(\boldsymbol{t}) Z_\beta(\boldsymbol{z}) = R_\beta(\boldsymbol{t}, \boldsymbol{z})$. With the log likelihood in (1) and $\xi \to \infty$, Wahba et al. (1995) derived the approximate posterior mean and covariance. We list them in the following theorem.

Theorem 1 *Let $\boldsymbol{c}$ and $\boldsymbol{d}$ be a solution to (6). Let λ, Θ be smoothing parameters at convergence and W, Q_Θ be matrix based on converged values. Let*

$n\lambda = \hat{\sigma}^2/b$ and $M = Q_\Theta + n\lambda W^{-1}$. For $g_{0,\nu}(\boldsymbol{t}) = \tau_\nu \phi_\nu(\boldsymbol{t})$, $g_\beta(\boldsymbol{t}) = b^{\frac{1}{2}}\sqrt{\theta_\beta}Z_\beta(\boldsymbol{t})$, $\nu = 1, \cdots, M$, $\beta = 1, \cdots, q$, we have

$$\begin{aligned}
E(g_{0,\nu}(\boldsymbol{t})|\boldsymbol{y}) &\approx d_\nu \phi_\nu(\boldsymbol{t}), \\
E(g_\beta(\boldsymbol{t})|\boldsymbol{y}) &\approx \sum_{i=1}^n c_i \theta_\beta R_\beta(\boldsymbol{t}, \boldsymbol{t}_i), \\
\frac{1}{b} Cov(g_{0,\nu}(\boldsymbol{z}), g_{0,\mu}(\boldsymbol{t})|\boldsymbol{y}) &\approx \phi_\nu(\boldsymbol{z})\phi_\mu(\boldsymbol{t}) \boldsymbol{e}_\nu^T (S^T M^{-1} S)^{-1} \boldsymbol{e}_\mu, \\
\frac{1}{b} Cov(g_\beta(\boldsymbol{z}), g_{0,\nu}(\boldsymbol{t})|\boldsymbol{y}) &\approx -d_{\nu,\beta}(\boldsymbol{z})\phi_\nu(\boldsymbol{t}), \\
\frac{1}{b} Cov(g_\beta(\boldsymbol{z}), g_\beta(\boldsymbol{t})|\boldsymbol{y}) &\approx \theta_\beta R_\beta(\boldsymbol{z}, \boldsymbol{t}) - \sum_{i=1}^n c_{i,\beta}(\boldsymbol{z}) \theta_\beta R_\beta(\boldsymbol{t}, \boldsymbol{t}_i), \\
\frac{1}{b} Cov(g_\gamma(\boldsymbol{z}), g_\beta(\boldsymbol{t})|\boldsymbol{y}) &\approx -\sum_{i=1}^n c_{i,\gamma}(\boldsymbol{z}) \theta_\beta R_\beta(\boldsymbol{t}, \boldsymbol{t}_i),
\end{aligned}$$

where $\boldsymbol{e}_\nu$ is the νth unit vector, and $(d_{1,\beta}(\boldsymbol{z}), \cdots, d_{M,\beta}(\boldsymbol{z})) = d_\beta(\boldsymbol{z})^T$ and $(c_{1,\beta}(\boldsymbol{z}), \cdots, c_{n,\beta}(\boldsymbol{z})) = c_\beta(\boldsymbol{z})^T$ are given by

$$d_\beta(\boldsymbol{z}) = (S^T M^{-1} S)^{-1} S^T M^{-1} \begin{pmatrix} \theta_\beta R_\beta(\boldsymbol{z}, \boldsymbol{t}_1) \\ \vdots \\ \theta_\beta R_\beta(\boldsymbol{z}, \boldsymbol{t}_n) \end{pmatrix}, \tag{14}$$

$$c_\beta(\boldsymbol{z}) = [M^{-1} - M^{-1} S (S^T M^{-1} S)^{-1} S^T M^{-1}] \begin{pmatrix} \theta_\beta R_\beta(\boldsymbol{z}, \boldsymbol{t}_1) \\ \vdots \\ \theta_\beta R_\beta(\boldsymbol{z}, \boldsymbol{t}_n) \end{pmatrix}. \tag{15}$$

4 A Special Case: Tensor Product of W_2

To illustrate how to use SS ANOVA method and how to construct Bayesian confidence intervals, consider the special case $\mathcal{T}^{(j)} = [0, 1]$, $j = 1, \cdots, d$ and $d \geq 2$ (we usually transform all continuous variables into $[0, 1]$ for fitting and then transform them back). Take the component space on [0,1] as the RKHS

$$W_2 = \{f : f \text{ and } f^{(1)} \text{ abs. cont.}, \int_0^1 (f^{(2)})^2 < \infty\}$$

with norm

$$f^2 = (\int_0^1 f)^2 + (\int_0^1 f^{(1)})^2 + \int_0^1 (f^{(2)})^2.$$

We decompose $W_2 = \mathcal{N} \oplus \mathcal{L} \oplus \mathcal{S}$, where $\mathcal{N}$ is the space of constants, with the square norm $(\int_0^1 f)^2$; $\mathcal{L}$ is the space of linear functions which integrate to zero, with the square norm $(\int_0^1 f^{(1)})^2$; and $\mathcal{S}$ is the space of functions with square integrable 2nd derivative and satisfy $\int_0^1 f^{(v)} = 0$, $v = 0, 1$, with the square norm $\int_0^1 (f^{(2)})^2$. The RK for subspace $\mathcal{S}$ is

$$R(t, z) = k_2(t)k_2(z) - k_4(t - z),$$

where $k_\nu(\cdot) = B_\nu(\cdot)/v!$ and $B_\nu(\cdot)$ is the νth Bernoulli polynomial. Let $\mathcal{G}$ be the tensor product of component spaces

$$\begin{aligned}
\mathcal{G} &= \otimes_{j=1}^d (\mathcal{N}^j \oplus \mathcal{L}^j \oplus \mathcal{S}^j) \\
&= \{1\} \oplus \{(\oplus_{j=1}^d \mathcal{L}^j) \oplus (\oplus_{j=1}^d \mathcal{S}^j)\} \\
&\quad \oplus \{(\oplus_{j<k} (\mathcal{L}^j \otimes \mathcal{L}^k)) \oplus (\oplus_{j \neq k} (\mathcal{L}^j \otimes \mathcal{S}^k)) \oplus (\oplus_{j<k} (\mathcal{S}^j \otimes \mathcal{S}^k))\} \oplus \cdots,
\end{aligned}$$

where with some abuse of notation, we are omitting factors of the form $\otimes\{\mathcal{N}^j\}$; $\{1\}$ is the space of constant functions on $[0, 1]^d$; $\mathcal{L}^j = \mathcal{N}^1 \otimes \cdots \otimes \mathcal{L}^j \otimes \cdots \otimes \mathcal{N}^d$ is the space of functions that is a constant on t_k, $k \neq j$ and a linear function on t_j. Others have similar interpretation. The terms in the three brackets are the spaces of the constant, the main effects and the 2-interactions respectively. For simplicity of notation (calculations can be easily extended to having more than one interactions), suppose we choose a model space contains the constant, all main effects and the interaction between t_1 and t_2:

$$\begin{aligned}
\mathcal{M} &= \{1\} \oplus \{\oplus_{j=1}^d (\mathcal{L}^j \oplus \mathcal{S}^j)\} \\
&\quad \oplus \{(\mathcal{L}^1 \otimes \mathcal{L}^2) \oplus (\mathcal{S}^1 \otimes \mathcal{L}^2) \oplus (\mathcal{L}^1 \otimes \mathcal{S}^2) \oplus (\mathcal{S}^1 \otimes \mathcal{S}^2)\}.
\end{aligned}$$

Therefore, we have $M = d + 2$, $q = d + 3$. We usually take

$$\phi_1(\boldsymbol{t}) = 1,$$

$$
\begin{aligned}
\phi_\nu(\boldsymbol{t}) &= t_{\nu-1} - 0.5, \qquad \nu = 2, \cdots, d+1, \\
\phi_M(\boldsymbol{t}) &= (t_1 - 0.5) \times (t_2 - 0.5)
\end{aligned}
$$

as basis functions for $\mathcal{H}^0$ and

$$
\begin{aligned}
R_\beta(\boldsymbol{t}, \boldsymbol{z}) &= R(t_\beta, z_\beta), \qquad \beta = 1, \cdots, d, \\
R_{d+1}(\boldsymbol{t}, \boldsymbol{z}) &= R(t_1, z_1) \times (t_2 - 0.5) \times (z_2 - 0.5), \\
R_{d+2}(\boldsymbol{t}, \boldsymbol{z}) &= (t_1 - 0.5) \times (z_1 - 0.5) \times R(t_2, z_2), \\
R_{d+3}(\boldsymbol{t}, \boldsymbol{z}) &= R(t_1, z_1) \times R(t_2, z_2)
\end{aligned}
$$

as the RK's for $\mathcal{H}^1, \cdots, \mathcal{H}^{d+3}$ respectively.

Write

$$
f(\boldsymbol{t}) = C + f_1(t_1) + \cdots + f_d(t_d) + f_{1,2}(t_1, t_2),
$$

where f_j's are the main effects and $f_{1,2}$ is the interaction between t_1 and t_2. Comparing to Theorem 1, we have

$$
\begin{aligned}
C &= g_{01}, \\
f_j(t_j) &= g_{0,j+1}(\boldsymbol{t}) + g_j(\boldsymbol{t}), \qquad j = 1, \cdots, d, \\
f_{1,2}(t_1, t_2) &= g_{0,M}(\boldsymbol{t}) + g_{d+1}(\boldsymbol{t}) + g_{d+2}(\boldsymbol{t}) + g_{d+3}(\boldsymbol{t}).
\end{aligned}
$$

Suppose we want to calculate posterior means and standard deviations of the main effects, the interaction and the overall function on grid points $\mathcal{Z}_1 \times \cdots \times \mathcal{Z}_d$, where $\mathcal{Z}_j$'s are sets of points in $[0,1]$. For simplicity of notation, suppose $\mathcal{Z}_j = \mathcal{Z} = \{z_1, \cdots, z_K\}$. Calculations for different $\mathcal{Z}_j$'s are the same. Let

$$
\begin{aligned}
\boldsymbol{f}_j^T &= (f_j(z_1), \cdots, f_j(z_K)), \qquad j = 1, \cdots, d, &(16)\\
\boldsymbol{f}_{1,2}^T &= (f_{1,2}(z_1, z_1), \cdots, f_{1,2}(z_K, z_1), \cdots, f_{1,2}(z_1, z_K), \cdots, f_{1,2}(z_K, z_K)), &(17)\\
A &= (S^T M^{-1} S)^{-1}, &(18)
\end{aligned}
$$

$$
\begin{aligned}
\boldsymbol{\phi}_\nu^T &= (\phi_\nu(z_1), \cdots, \phi_\nu(z_K)), \qquad \nu = 2, \cdots, d+1, && (19)\\
\boldsymbol{\phi}_{1,2}^T &= (\phi_M(z_1, z_1), \cdots, \phi_M(z_K, z_1), \cdots, \phi_M(z_1, z_K), \cdots, \phi_M(z_K, z_K)), && (20)\\
\boldsymbol{d}_j^T &= (d_{j+1,j}(z_1), \cdots, d_{j+1,j}(z_K)), \qquad j = 1, \cdots, d, && (21)\\
\boldsymbol{d}_{1,2}^T &= (d_{M,d+1}(z_1, z_1), \cdots, d_{M,d+1}(z_K, z_1), \cdots, d_{M,d+1}(z_1, z_K), \cdots, d_{M,d+1}(z_K, z_K)) \\
&+ (d_{M,d+2}(z_1, z_1), \cdots, d_{M,d+2}(z_K, z_1), \cdots, d_{M,d+2}(z_1, z_K), \cdots, d_{M,d+2}(z_K, z_K)) \\
&+ (d_{M,d+3}(z_1, z_1), \cdots, d_{M,d+3}(z_K, z_1), \cdots, d_{M,d+3}(z_1, z_K), \cdots, d_{M,d+3}(z_K, z_K)), \\
& && (22)\\
C_\beta &= (c_\beta(z_1), \cdots, c_\beta(z_K))_{n\times K}, \qquad \beta = 1, \cdots, d, && (23)\\
C_{1,2} &= (c_{d+1}(z_1, z_1), \cdots, c_{d+1}(z_K, z_1), \cdots, c_{d+1}(z_1, z_K), \cdots, c_{d+1}(z_K, z_K))_{n\times K^2} \\
&+ (c_{d+2}(z_1, z_1), \cdots, c_{d+2}(z_K, z_1), \cdots, c_{d+2}(z_1, z_K), \cdots, c_{d+2}(z_K, z_K))_{n\times K^2} \\
&+ (c_{d+3}(z_1, z_1), \cdots, c_{d+3}(z_K, z_1), \cdots, c_{d+3}(z_1, z_K), \cdots, c_{d+3}(z_K, z_K))_{n\times K^2} && (24)\\
\Lambda_\beta &= (\theta_\beta R_\beta(z_k, t_{\beta i}))_{K\times n}, \qquad \beta = 1, \cdots, d, && (25)\\
\Lambda_{1,2} &= ((\theta_{d+1}R_{d+1} + \theta_{d+2}R_{d+2} + \theta_{d+3}R_{d+3})((z_{k_1}, z_{k_2}), (t_{1i}, t_{2i}))_{K^2\times n}, && (26)\\
\Sigma_\beta &= (\theta_\beta R_\beta(z_i, z_k))_{K\times K}, \qquad \beta = 1, \cdots, d, && (27)\\
\Sigma_{1,2} &= ((\theta_{d+1}R_{d+1} + \theta_{d+2}R_{d+2} + \theta_{d+3}R_{d+3})((z_{i1}, z_{i2}), (z_{k1}, z_{k2})))_{K^2\times K^2}. && (28)
\end{aligned}
$$

From Theorem 1, the posterior means and standard deviations for the constant, the main effects at points $\mathcal{Z}$ and interaction at points $\mathcal{Z} \times \mathcal{Z}$ are

$$
\begin{aligned}
\mathrm{E}(C|\boldsymbol{y}) &\approx d_1, \\
\mathrm{Cov}(C|\boldsymbol{y}) &\approx bA(1,1), \\
\mathrm{E}(\boldsymbol{f}_j|\boldsymbol{y}) &\approx d_{j+1}\boldsymbol{\phi}_{j+1} + \Lambda_j \boldsymbol{c}, \quad j = 1, \cdots, d, \\
\mathrm{Cov}(\boldsymbol{f}_j|\boldsymbol{y}) &\approx b[A(j+1, j+1)\boldsymbol{\phi}_{j+1}\boldsymbol{\phi}_{j+1}^T - \boldsymbol{d}_j\boldsymbol{\phi}_{j+1}^T - \boldsymbol{\phi}_{j+1}\boldsymbol{d}_j^T + \Sigma_j - \Lambda_j C_j], \\
&\quad j = 1, \cdots, d, \\
\mathrm{E}(\boldsymbol{f}_{1,2}|\boldsymbol{y}) &\approx d_M\boldsymbol{\phi}_{1,2} + \Lambda_{1,2}\boldsymbol{c}, \\
\mathrm{Cov}(\boldsymbol{f}_{1,2}|\boldsymbol{y}) &\approx b[A(M,M)\boldsymbol{\phi}_{1,2}\boldsymbol{\phi}_{1,2}^T - \boldsymbol{d}_{1,2}\boldsymbol{\phi}_{1,2}^T - \boldsymbol{\phi}_{1,2}\boldsymbol{d}_{1,2}^T + \Sigma_{1,2} - \Lambda_{1,2}C_{1,2}].
\end{aligned}
$$

For any point $\boldsymbol{z} \in \mathcal{Z}^d$, the posterior mean of the overall function

$$\mathrm{E}(f(\boldsymbol{z})|\boldsymbol{y}) = \mathrm{E}(C|\boldsymbol{y}) + \sum_{j=1}^{d} \mathrm{E}(f_j(\boldsymbol{z}|\boldsymbol{y})) + \mathrm{E}(f_{1,2}|\boldsymbol{y}).$$

The posterior variance of the overall function can be calculated using the formula

$$\begin{aligned}
&\mathrm{Var}(C + \sum_{j=1}^{d} f_j(\boldsymbol{z}) + f_{1,2}(\boldsymbol{z})|\boldsymbol{y}) = \\
&\quad \mathrm{Var}(C|\boldsymbol{y}) + \sum_{j=1}^{d} \mathrm{Var}(f_j(\boldsymbol{z})|\boldsymbol{y}) + \mathrm{Var}(f_{1,2}(\boldsymbol{z})|\boldsymbol{y}) \\
&\quad +2\sum_{j=1}^{d} \mathrm{Cov}(C, f_j(\boldsymbol{z})|\boldsymbol{y}) + 2\mathrm{Cov}(C, f_{1,2}(\boldsymbol{z})|\boldsymbol{y}) + 2\sum_{j=1}^{d} \mathrm{Cov}(f_j(\boldsymbol{z}), f_{1,2}(\boldsymbol{z})|\boldsymbol{y}).
\end{aligned}$$

Each term above can be read off easily from the calculations for the component posterior variances.

We need to calculate $(S^T M^{-1} S)^{-1}$, $c_\beta(\boldsymbol{t})$ and $d_\beta(\boldsymbol{t})$. Gu and Wahba (1993) discussed how to calculate these quantities when $W = I$. Let $\tilde{Q}_\Theta = W^{1/2} Q_\Theta W^{1/2}$, $\tilde{S} = W^{1/2} S$, and $\tilde{M} = \tilde{Q}_\Theta + n\lambda I$. We can calculate $(\tilde{S}^T \tilde{M}^{-1} \tilde{S})^{-1}$, $\tilde{d}_\beta(\boldsymbol{t})$ and $\tilde{c}_\beta(\boldsymbol{t})$ the same way as Gu and Wahba (1993) with their $R_\beta(\boldsymbol{t}, \boldsymbol{t}_i)$ replaced by $\tilde{R}_\beta(\boldsymbol{t}, \boldsymbol{t}_i) = \sqrt{w_i} R_\beta(\boldsymbol{t}, \boldsymbol{t}_i)$. We than have $(S^T M^{-1} S)^{-1} = (\tilde{S}^T \tilde{M}^{-1} \tilde{S})^{-1}$, $d_\beta(\boldsymbol{t}) = \tilde{d}_\beta(\boldsymbol{t})$ and $c_\beta(\boldsymbol{t}) = W^{1/2} \tilde{c}_\beta(\boldsymbol{t})$. Two utility routines `dcrdr, dsms` in RKPACK can be used to calculate these quantities.

The (1-α)100% Bayesian confidence interval for a function g (g could be the constant C, the main effects f_j's, the interaction $f_{1,2}$ and the overall function f) at a given point $\boldsymbol{z}$ is

$$(\mathrm{E}(g(\boldsymbol{z})|\boldsymbol{y}) - z_{\alpha/2}\sqrt{\mathrm{Var}(g(\boldsymbol{z})|\boldsymbol{y})}, \mathrm{E}(g(\boldsymbol{z})|\boldsymbol{y}) + z_{\alpha/2}\sqrt{\mathrm{Var}(g(\boldsymbol{z})|\boldsymbol{y})}\ ),$$

where $z_{\alpha/2}$ is the upper $\alpha/2$ percentile of the standard Normal distribution.

5 GRKPACK

GRKPACK is a collection of subroutines using the SS ANOVA algorithm in section 2 for binary, binomial, Poisson and Gamma data. It is written in Ratfor (rational Fortran). Users can modify these routines for other distributions in exponential families.

Subroutines in GRKPACK are listed in Table 1. Correspondence of notations between this report and GRKPACK are listed in Table 2.

Table 1: Drivers in GRKPACK

data	single smoothing parameter drivers	multiple smoothing parameter drivers
binary	dbsdr	dbmdr
binomial	dbisdr	dbimdr
Poisson	dpsdr	dpmdr
Gamma	dgsdr	dgmdr

GRKPACK also contains some application and simulation programs we used in Wahba et al. (1994a) and Wang (1994). Correspondence of notations between the report and calculations of Bayesian confidence intervals in these programs are listed in Table 3. Notice that $mainsd(\cdot, j)$ contains the diagonal elements of $\mathrm{Cov}(\boldsymbol{f}_j|\boldsymbol{y})$; $inter$ is a $K \times K$ matrix with $inter(i,j) = \mathrm{E}(\boldsymbol{f}_{1,2}|\boldsymbol{y})((i-1)\times K + j)$; $intersd$ is a $K \times K$ matrix with $intersd(i,j) = \mathrm{Cov}(\boldsymbol{f}_{1,2}|\boldsymbol{y})((i-1)\times K + j, (i-1)\times K + j)$.

All subroutines in RKPACK and GRKPACK are self-documented. RKPACK and GRKPACK are available from `netlib@research.att.com` under the `gcv` directory and `statlib@lib.stat.cmu.edu`. GRKPACK is also available from the author's homepage at `http://www.sph.umich.edu/~yuedong`. or anonymous ftp to `ftp.stat.wisc.edu` under the directory `/pub/wahba`. Any suggestions are appreciated and should be forwarded to the author.

Table 2: GRKPACK notation correspondence

report	code	type	explanation
			options for smoothing parameters
option V	vmu='v'	character	GCV
option U	vmu='u'	character	UBR with dispersion parameter $\sigma^2 = 1$
option U˜	vmu='u˜ '	character	UBR and estimate the dispersion parameter
			model parameters
M	nnull	integer	dimension of the null space $\mathcal{H}^0$
q	nq	integer	number of RKHS's to be penalized
			design, RK's and their transformations
S	s	matrix	the design matrix of $\mathcal{H}^0$
Q_β	q$(\cdot,\cdot,\beta)$	matrix	RK's evaluated at design points
$\tilde{S}$	swk	matrix	$\tilde{S} = W_-^{1/2} S$
$\tilde{Q}_\beta$	qwk$(\cdot,\cdot,\beta)$	matrix	$\tilde{Q}_\beta = W_-^{1/2} Q_\beta W_-^{1/2}$
			precision parameters
p	prec	scale	stop criteria for the algorithm
L	maxiter	integer	maximum number of iterations for the algorithm
			data and their transformations
n	nobs	integer	number of observations
$\boldsymbol{y}$	y	vector	observation vector
$\tilde{\boldsymbol{y}}$	ywk	vector	$\tilde{\boldsymbol{y}} = W_-^{-1/2}(W_- \boldsymbol{f}_- - \boldsymbol{u}_-)$
			coefficients, estimates and devivatives
$\boldsymbol{c}$	c	vector	$\boldsymbol{c}$ vector in (5)
$\boldsymbol{d}$	d	vector	$\boldsymbol{d}$ vector in (5)
$\boldsymbol{f}$	eta	vector	estimate at design points
$\boldsymbol{u}$	u	vector	first derivative vector at design points
W	w	vector	second derivatives, $W = \mathrm{diag}(w(1), \cdots, w(n))$
			summary information
λ	nlaht	scale	main smoothing parameter, nlaht $= \log_{10}(n\lambda)$
Θ	theta	vector	subsidiary smoothing parameters in $\log_{10}$ scale
$V\|U\|\tilde{U}$	score	scale	minimum GCV and UBR score
$\hat{\sigma}^2$	varht	scale	estimate of dispersion parameter

Table 3: Notation correspondence for calculations of posterior means and variances

report	code	type	explanation
			data and prediction points
d	ncov	integer	number of contineous covariates
j	cov(j)	integer	column in data corresponding to t_j
t_{ji}	x(i,cov(j))	scale	jth covariate at the ith design point
K	nplot	integer	number of points in $\mathcal{Z}$
$\mathcal{Z}$	z	vector	points for calculation of predictions
			posterior means and variances
b	b	scale	$b = \hat{\sigma}^2/n\lambda$
$\mathrm{E}(C\|\boldsymbol{y})$	d(1)	scale	posterior mean of the constant
$\mathrm{Var}(C\|\boldsymbol{y})$	b $\times$ sms(1,1)	scale	posterior variance of the constant
$\mathrm{E}(\boldsymbol{f}_j\|\boldsymbol{y})$	main($\cdot$,j)	vector	posterior mean vector of the t_j main effect
$\mathrm{Cov}(\boldsymbol{f}_j\|\boldsymbol{y})$	mainsd($\cdot$,j)	vector	posterior variances of the t_j main effect
$\mathrm{E}(\boldsymbol{f}_{1,2}\|\boldsymbol{y})$	inter	matrix	posterior mean of the (t_1, t_2) interaction
$\mathrm{Cov}(\boldsymbol{f}_{1,2}\|\boldsymbol{y})$	intersd	matrix	posterior variances of the interaction
			matrices for calculations
$(S^TM^{-1}S)^{-1}$	sms	matrix	see (18)
Λ_β^T	r($\cdot,\cdot,\beta$+ncov)	matrix	see (25)
$W^{1/2}\Lambda_\beta^T$	r($\cdot,\cdot,\beta$)	matrix	transform of (25)
$\Lambda_{1,2}^T$	rr($\cdot,\cdot$,2)	matrix	see (26)
$W^{1/2}\Lambda_{1,2}^T$	rr($\cdot,\cdot$,1)	matrix	transform of (26)
C_β	cr($\cdot,\cdot,\beta$)	matrix	see (23)
$C_{1,2}$	crr($\cdot,\cdot$)	matrix	see (24)
$\boldsymbol{d}_j$	dr(j+1,$\cdot$,j)	vector	see (21)
$\boldsymbol{d}_{1,2}$	drr(M,$\cdot$)	vector	see (22)

ACKNOWLEDGEMENTS

The author is indebted to Grace Wahba, Douglas Bates and Chong Gu for their advice and encouragement. The author thanks Dong Xiang, Zhen Luo and Xiwu Lin for testing out the routines and suggestions. This research was supported by the National Institute of Health under Grants R01 EY09946, P60 DK20572 and P30 HD18258.

References

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

Gu, C. (1989). RKPACK and its applications: Fitting smoothing spline models, *Proceedings of the Statistical Computing Section, ASA:* pp. 42–51.

Gu, C. (1990). Adaptive spline smoothing in non-Gaussian regression models, *Journal of the American Statistical Association* **85**: 801–807.

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

Gu, C. and Wahba, G. (1993). Smoothing spline ANOVA with component-wise Bayesian confidence intervals, *Journal of Computational and Graphical Statistics* **2**: 97–117.

Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models*, Chapman and Hall.

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

Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models, *Journal of the Royal Statistical Society, Serr. A* **135**: 370–384.

O'Sullivan, F., Yandell, B. and Raynor, W. (1986). Automatic smoothing of regression functions in generalized linear models, *Journal of the American Statistical Association* **81**: 96–103.

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

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

Wahba, G., Wang, Y., Gu, C., Klein, R. and Klein, B. (1994b). Structured machine learning for 'soft' classification with smoothing spline ANOVA and stacked tuning, testing and evaluation, *Advances in Neural Information Processing 6* pp. 415–422. J. D. Cowan, G. Tesauro and J. Alspector, eds.

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, *The Annals of Statistics* **23**: 1865–1895.

Wang, Y. (1994). Smoothing spline analysis of variance of data from exponential families, Ph.D. Thesis, University of Wisconsin-Madison, Dept. of Statistics. Available by ftp in `ftp.stat.wisc.edu/pub/wahba/theses/ywang.thesis.shar.gz`.

Wang, Y., Wahba, G., Chappell, R. and Gu, C. (1995). Simulation studies of smoothing parameter estimates and Bayesian confidence intervals in Bernoulli SS ANOVA models, *Communications in Statistics: Simulation and Computation* **24**: 1037–1059.

Wang, Y., Wahba, G., Gu, C., Klein, R. and Klein, B. (1996). Using smoothing spline anova to examine the relation of risk factors to the incidence and progression of diabetic retinopathy, Technical Report 956, University of Wisconsin-Madison, Dept. of Statistics. To appear, *Statistics in Medicine*.