Statistica Sinica 2(1992), 255-264

# PENALIZED LIKELIHOOD REGRESSION: A BAYESIAN ANALYSIS

Chong Gu

*Purdue University*

*Abstract:* It was well established by Wahba (1978) that a smoothing spline procedure is equivalent to a Bayesian procedure under a partially improper prior. Based on this interpretation, statistical inference other than point estimation was made possible for smoothing spline estimators of a Gaussian regression function (Wahba (1983)). This article extends Wahba's results to non-Gaussian regression problems via a simple approximation known as Laplace's method. The results make the Bayesian inference tools developed for Gaussian models applicable to non-Gaussian models.

*Key words and phrases:* Bayesian confidence interval, Laplace's method, posterior analysis, smoothing spline.

## 1. Introduction

Smoothing spline technique is widely used as a powerful nonparametric regression tool in data analysis. The method allows the estimator to take a flexible form and seeks an appropriate balance between the goodness-of-fit and the smoothness of the estimator via minimizing the sum of a standard goodness-of-fit criterion and a roughness penalty. For Gaussian sampling likelihood with least squares as the goodness-of-fit, the commonly used quadratic roughness penalty was shown by Wahba (1978) to be equivalent to a partially improper Gaussian prior in the sense that the smoothing spline estimator can be interpreted as the mean of the corresponding Gaussian posterior. When the sampling likelihood is non-Gaussian, a penalized log-likelihood criterion is commonly used in the estimation (O'Sullivan et al. (1986), Gu (1990)). The purpose of this article is to explore a Bayesian interpretation for the penalized likelihood smoothing spline estimator of a non-Gaussian regression function. It is shown that under Wahba's prior, if one approximates the posterior via Laplace's method in a certain way, the smoothing spline estimator is the mean of the approximate posterior. The results allow Wahba's Bayesian confidence intervals (Wahba (1983), Nychka (1988)) and other Bayesian inference tools be applied to non-Gaussian models.

In Section 2, we extend Wahba's posterior calculations to Gaussian sam-

pling likelihood with nondiagonal correlation matrix and introduce notations. Section 3 describes the approximation in the posterior analysis and gives the main results. Section 4 discusses the computation of the posterior covariance, which will be needed in constructing Bayesian confidence intervals. In parallel to Wahba (1983) and Nychka (1988), a Monte-Carlo experiment is presented in Section 5 to illustrate a certain "frequentist" property of Bayesian confidence intervals.

## 2. Extensions of Wahba's Results

We extend Wahba's Gaussian posterior calculations (Wahba (1978, 1983, 1985)) to the case where the sampling errors are non *i.i.d.* We will develop the results in a very general setup. A commonly used specialization is described in Section 5. See Wahba (1990) for other useful specializations.

Suppose on domain $\mathcal{T}$ one observes $y_j = f(t_j) + \epsilon_j$, $j = 1, \ldots, n$, where $t_j \in \mathcal{T}$, and $(\epsilon_1, \ldots, \epsilon_n)^T = \boldsymbol{\epsilon} \sim N(0, \sigma^2 W^{-1})$ with $W$ (positive definite) known. Write $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ and $\boldsymbol{t} = (t_1, \ldots, t_n)^T$. The solution $f_{W,\lambda}$ to the problem

$$\min(\boldsymbol{y} - f(\boldsymbol{t}))^T W(\boldsymbol{y} - f(\boldsymbol{t})) + n\lambda \|P_1 f\|^2, \ f \in \mathcal{H} \tag{2.1}$$

is called a smoothing spline, where $\mathcal{H}$ is a Hilbert space of functions on $\mathcal{T}$ with norm $\|\cdot\|$ in which an evaluation is a continuous linear functional, $\mathcal{H} = \mathcal{H}_0 \oplus \mathcal{H}_1$, $P_1$ is the projector onto $\mathcal{H}_1$ and $\dim(\mathcal{H}_0) = M < \infty$. $\lambda$ is the smoothing parameter which controls the trade-off between the goodness-of-fit and the smoothness. $f_{W,\lambda}$ can be expressed as (Wahba (1990))

$$f_{W,\lambda}(\cdot) = \sum_{\nu=1}^{M} \phi_\nu(\cdot) d_\nu + \sum_{j=1}^{n} R(\cdot, t_j) c_j = \boldsymbol{\phi}^T(\cdot)\boldsymbol{d} + R(\cdot, \boldsymbol{t}^T)\boldsymbol{c}, \tag{2.2}$$

where the vector $\boldsymbol{\phi}^T = (\phi_1, \ldots, \phi_M)$ span $\mathcal{H}_0$, $R(\cdot,\cdot)$ has the reproducing property $\langle R(t,\cdot), f\rangle = f(t)$, $\forall f \in \mathcal{H}_1$, $\forall t \in \mathcal{T}$ and $\langle\cdot,\cdot\rangle$ is the inner product in $\mathcal{H}$. By substituting (2.2) in (2.1), one can solve

$$\min(\boldsymbol{y} - Q\boldsymbol{c} - S\boldsymbol{d})^T W(\boldsymbol{y} - Q\boldsymbol{c} - S\boldsymbol{d}) + n\lambda \boldsymbol{c}^T Q \boldsymbol{c}, \tag{2.3}$$

where $(Q)_{i,j} = (R(\boldsymbol{t}, \boldsymbol{t}^T))_{i,j} = R(t_i, t_j)$ and $(S)_{j,\nu} = (\boldsymbol{\phi}^T(\boldsymbol{t}))_{j,\nu} = \phi_\nu(t_j)$.

In the following development, $s$ or $t$ denotes an arbitrary point on the domain $\mathcal{T}$, $\boldsymbol{s}$ denotes a vector of arbitrary points, and $\boldsymbol{t}$ denotes the vector of the sampling (design) points. Following Wahba (1978), we specify a prior for $f$, which is the same as the distribution of the stochastic process $f_\xi(t) = \sum_{i=1}^{M} \theta_i \phi_i(t) + b^{1/2} Z(t)$, where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)^T \sim N(0, \xi I)$, $Z(t)$ is a Gaussian process on $\mathcal{T}$ independent of $\boldsymbol{\theta}$ with $\mathrm{E}Z(t) = 0$ and $\mathrm{E}(Z(s)Z(t)) = R(s,t)$. Let $\boldsymbol{Y} = f_\xi(\boldsymbol{t}) + \boldsymbol{\epsilon}$,

$n\lambda = \sigma^2/b$, and $\mu_\xi(s|y,b,\sigma^2)$ and $V_\xi(s|y,b,\sigma^2)$ be the conditional mean and the conditional covariance respectively of $f_\xi(s)$ given $Y = y$, $b$, and $\sigma^2$.

**Theorem 2.1.** $\mu_\infty(s|y,b,\sigma^2) = \lim_{\xi\to\infty}\mu_\xi(s|y,b,\sigma^2) = f_{W,\lambda}(s)$ *and*

$$\begin{aligned} V_\infty(s|y,b,\sigma^2) &= \lim_{\xi\to\infty} V_\xi(s|y,b,\sigma^2) \\ =& b\{[\phi^T(s) - R(s,t^T)M^{-1}S](S^TM^{-1}S)^{-1}[\phi(s^T) - S^TM^{-1}R(t,s^T)] \\ &+ R(s,s^T) - R(s,t^T)M^{-1}R(t,s^T)\}, \end{aligned}$$

*where* $M = Q + n\lambda W^{-1}$.

We note that the solution of (2.3) satisfies

$$\begin{pmatrix} QWQ + n\lambda Q & QWS \\ S^TWQ & S^TWS \end{pmatrix}\begin{pmatrix} c \\ d \end{pmatrix} = \begin{pmatrix} QWy \\ S^TWy \end{pmatrix},$$

which leads to

$$c = M^{-1}(I - S(S^TM^{-1}S)^{-1}S^TM^{-1})y \text{ and } d = (S^TM^{-1}S)^{-1}S^TM^{-1}y.$$

Thus Theorem 2.1 immediately follows (2.6)-(2.9) of Wahba (1978). As $\xi \to \infty$ the prior on $\theta$ tends to a uniform improper prior. The parameter $\xi$ is essentially introduced as a convenient device for handling this improper prior.

Defining the influence matrix $A(\lambda)$ satisfying $\hat{y} = A(\lambda)y$, it is easy to verify that $A(\lambda) = I - n\lambda W^{-1}[M^{-1} - M^{-1}S(S^TM^{-1}S)^{-1}S^TM^{-1}]$. On applying Theorem 2.1 to $s = t$, we obtain Corollary 2.1.

**Corollary 2.1.** $V_\infty(t|y,b,\sigma^2) = \sigma^2 A(\lambda)W^{-1}$.

The proof of Corollary 2.1 follows tedious algebra.

Assuming a uniform improper prior for $\theta$, the marginal likelihood of $Y$ on parameters $b$ and $\sigma^2$ can be shown to be proportional to $\exp\{-\frac{1}{2b}y^T(M^{-1} - M^{-1}S(S^TM^{-1}S)^{-1}S^TM^{-1})y\}$. Let $W = GG^T$ be the Cholesky decomposition of $W$, $\tilde{S} = G^TS$, $\tilde{Q} = G^TQG$, $\tilde{M} = \tilde{Q} + n\lambda I$, and $\tilde{y} = G^Ty$. The quadratic form remains the same when one replaces everything by the tilded version. Now let $\tilde{S} = FR = (F_1, F_2)\begin{pmatrix} R_1 \\ 0 \end{pmatrix}$ be the QR-decomposition of $\tilde{S}$, with $R_1$ nonsingular. The quadratic form becomes $\tilde{y}^T(\tilde{M}^{-1} - \tilde{M}^{-1}F_1(F_1^T\tilde{M}^{-1}F_1)^{-1}F_1^T\tilde{M}^{-1})\tilde{y}$. By partitioning $(F^T\tilde{M}F)^{-1} = \begin{pmatrix} A & B \\ B^T & D \end{pmatrix}$, and using $F_1^TF = (I,O)$, it can be shown that the quadratic form reduces to $(F^T\tilde{y})^T\begin{pmatrix} O & O \\ O & D - B^TA^{-1}B \end{pmatrix}(F^T\tilde{y})$.

Since $(D - B^T A^{-1} B)^{-1}$ is equal to the bottom right block of $((F^T \tilde{M} F)^{-1})^{-1} = F^T \tilde{M} F$ (Rao (1973), p.33), which is $F_2^T \tilde{Q} F_2 + n\lambda I$, it follows that the likelihood is proportional to $\exp\{-\frac{1}{2b} z^T (F_2^T \tilde{Q} F_2 + n\lambda I)^{-1} z\}$, where $z = F_2^T \tilde{y}$. By including the constant terms involving $b$ and $n\lambda$, one gets $\exp\{-\frac{1}{2b} z^T (F_2^T \tilde{Q} F_2 + n\lambda I)^{-1} z - \frac{n-M}{2} \log b - \frac{1}{2} \log |F_2^T \tilde{Q} F_2 + n\lambda|\}$, which is Wahba's generalized likelihood (Wahba (1985)). Profiling on $\hat{b} = z^T (F_2^T \tilde{Q} F_2 + n\lambda I)^{-1} z/(n-M)$, the profile likelihood of $\lambda$ is proportional to

$$|F_2^T \tilde{Q} F_2 + n\lambda I|^{-1/2} (z^T (F_2^T \tilde{Q} F_2 + n\lambda I)^{-1} z)^{-(n-M)/2},$$

whose mode gives Wahba's generalized maximum likelihood (GML) estimate of $\lambda$.

## 3. Approximate Posterior Analysis

In this section, it is assumed that the sampling likelihood of $y$ is proportional to $\exp\{-\frac{1}{\sigma^2} L(y|\eta)\} = \exp\{-\frac{1}{\sigma^2} L_y(\eta)\}$, where $L_y(\cdot)$ is convex and completely specified, $\eta = f(t)$, and $\sigma^2$ is a "dispersion" parameter possibly unknown. We are interested in the solution $f_{L,\lambda}$ of the penalized log-likelihood problem

$$\min L_y(f(t)) + (n/2)\lambda \|P_1 f\|^2, \quad f \in \mathcal{H}. \tag{3.1}$$

See, e.g., O'Sullivan et al. (1986) and Gu (1990). The expression (2.2) for the solution depends only on the penalty $\|P_1 f\|^2$, so it still applies here. By substituting (2.2) in (3.1), we solve

$$\min L_y(Qc + Sd) + (n/2)\lambda c^T Q c. \tag{3.2}$$

Under the prior specified in Section 2, letting $\xi \to \infty$, the joint likelihood of $y$, $\eta = f(t)$, $f(s)$, and $\theta$ given $(b, \sigma^2)$ is $p(\eta|y)\tilde{q}(\eta|\theta) r(f(s)|\eta)$, where $p(\eta|y) \propto \exp\{-\frac{1}{\sigma^2} L_y(\eta)\}$, $\tilde{q}(\eta|\theta) \propto \exp\{-\frac{1}{2b}(\eta - S\theta)^T Q^{-1} (\eta - S\theta)\}$, and $r(f(s)|\eta)$ is Gaussian with mean and covariance given in Theorem 2.1 with $\sigma^2 = 0$ and $y = \eta$. Integrating out $\theta$ from $\tilde{q}(\eta|\theta)$ yields

$$q(\eta) \propto \exp\{-\frac{1}{2b} \eta^T (Q^{-1} - Q^{-1} S (S^T Q^{-1} S)^{-1} S^T Q^{-1}) \eta\}.$$

The posterior distribution of interest is $\pi(f(s)|y) \propto \int p(\eta|y) q(\eta) r(f(s)|\eta) d\eta$. To approximate this integral, we adopt Laplace's method which is revisited in recent Bayesian literature. See, e.g., Leonard (1982), Tierney and Kadane (1986), and Leonard et al. (1989) for details about Laplace's method. Here is how it works in our setting. Expanding $\log[p(\eta|y) q(\eta) r(f(s)|\eta)]$ via Taylor series and

neglecting the cubic and higher order terms, one can approximate the integrand by the exponent of the expansion, which is in the form of a Gaussian likelihood. The integration of such an approximate integrand is immediate. In our case, since $q(\eta)$ and $r(f(s)|\eta)$ are Gaussian, we only need to approximate $p(\eta|y)$. Letting $\hat{p}(\eta|y)$ be such an approximation with the expansion centered at the mode $\eta_*$ of $p(\eta|y)q(\eta)$, one gets $\log \hat{p}(\eta|y) = -\frac{1}{2\sigma^2}(\eta - (\eta_* - W^{-1}u))^T W(\eta - (\eta_* - W^{-1}u)) + C$, where $u = (\partial L/\partial\eta)|_{\eta_*}$, $W = (\partial^2 L/\partial\eta\partial\eta^T)|_{\eta_*}$, and $C$ is a constant independent of $\eta$. We approximate the posterior $\pi(f(s)|y)$ via $\hat{\pi}(f(s)|y) \propto \int \hat{p}(\eta|y)q(\eta)r(f(s)|\eta)d\eta$.

**Theorem 3.1.** *The approximate posterior distribution $\hat{\pi}(f(s)|y)$ is Gaussian with mean $f_{L,\lambda}(s)$ and covariance given in Theorem 2.1, where the matrix $W$ is the Hessian defined above.*

**Proof.** Technically, the approximate likelihood $\hat{p}(\eta|y)$ is identical to a Gaussian sampling likelihood with covariance $\sigma^2 W^{-1}$ and observations $Y = \eta_* - W^{-1}u$; hence, the mean and the covariance of $\hat{\pi}(f(s)|y)$ can be calculated via Theorem 2.1. When using Newton iteration to solve (3.2), it can be shown (Gu (1990)) that at the fixed point $\eta_{**}$ the solution satisfies

$$\begin{pmatrix} QWQ + n\lambda Q & QWS \\ S^TWQ & S^TWS \end{pmatrix} \begin{pmatrix} c \\ d \end{pmatrix} = \begin{pmatrix} QW(\eta_{**} - W^{-1}u) \\ S^TW(\eta_{**} - W^{-1}u) \end{pmatrix},$$

where $W$ and $u$ are evaluated at $\eta_{**}$. Hence, it suffices to verify that $\eta_* = \eta_{**}$. By definition, $\eta_*$ is the minimizer of

$$L_y(\eta) + (n/2)\lambda\eta^T(Q^{-1} - Q^{-1}S(S^TQ^{-1}S)^{-1}S^TQ^{-1})\eta.$$

Writing $\eta = Qc + Sd$ with the constraint that $S^Tc = 0$, the problem reduces to solving (3.2) with this constraint. Note that this constraint doesn't change the original problem since $\eta = Qc + Sd$ is an overparameterization of the vector $\eta$. Finally, it is apparent that the minimizers $c$ and $d$ of (3.2), although constraint-free, always satisfy the above constraint (see Gu (1990)). This completes the proof.

The standard application of Laplace's method is to take the Taylor expansion around the mode of the whole integrand, which should be the mode of $p(\eta|y)q(\eta)r(f(s)|\eta)$ rather than that of $p(\eta|y)q(\eta)$. If the posterior $\pi(f(s)|y)$ for certain $s$ needs to be accurately approximated, the mode needs to be searched for every fixed value of $f(s)$, though $\eta_*$ could serve as a good starting guess. However, such an approximation is no longer Gaussian since the mode and $u$, $W$ evaluated at the mode will all depend on $f(s)$. Intuitively, the approximation of

$\hat{\pi}(f(s)|y)$ to $\pi(f(s)|y)$ should be relatively more accurate for lower dimensional $s$ and for $f(s)$ closer to the mean of $r(f(s)|\eta_*)$, because in such cases the mode shift should be smaller. Based on the approximate posterior distribution given in Theorem 3.1, one may construct pointwise approximate Bayesian confidence intervals or simultaneous confidence regions, etc., conditioned on parameters $b$ and $\sigma^2$. Via the same approximation, it is trivial to write down the approximate marginal likelihood for $(b, \sigma^2)$ following the lines of Section 2.

Finally, a few words about the precision of the approximation are in order. For given $(b, \sigma^2)$, $f(s)$ are the parameters of interest in Theorem 3.1. Since at most one sample of the $y_j$'s is "generated" according to a parameter component $f(s)$, which happens when $s = t_j$, the asymptotic results in the literature are not applicable in this circumstance. The precision of the approximation depends, I believe, on the relative strengths of the non-Gaussian sampling likelihood and the Gaussian prior. One can expect a better Gaussian approximation for a larger $n\lambda$. In general no $\pi(f(s)|y)$ is of special interest, and Theorem 3.1 is mainly taken as a useful device for deriving other approximations, e.g., the Bayesian confidence intervals, rather than as an accurate approximation for $\pi(f(s)|y)$. Nevertheless, $f_{L,\lambda}(t)$ is the posterior mode of $\pi(f(t)|y)$, although the statement is not true even for a subvector of $t$. When parameters $b$ and $\sigma^2$ are unknown and are of primary interest, which is the case in certain model selection problems, however, the standard asymptotics in the literature are needed and are likely to be applicable. We shall explore details in further study.

## 4. Computation

To apply the results derived in Sections 2 and 3, it is necessary to compute the quantities involved. The calculation of $f_{L,\lambda}$ is discussed in Gu (1990) with the smoothing parameter $n\lambda = \sigma^2/b$ selected via the generalized cross validation (GCV) method. In this section, we discuss the calculation of the posterior covariance given in Theorem 2.1 assuming $b$ and $\sigma^2$ are known. For the one-parameter exponential family likelihood the dispersion parameter $\sigma^2$ is known, and $b$ could be calculated from the GCV estimate of $n\lambda$.

We rewrite

$$\begin{aligned} V_\infty(s|y, b, \sigma^2)/b = {} & R(s, s^T) + \phi^T(s)(S^T M^{-1} S)^{-1}\phi(s^T) \\ & - \phi^T(s)(S^T M^{-1} S)^{-1} S^T M^{-1} R(t, s^T) \\ & - (\phi^T(s)(S^T M^{-1} S)^{-1} S^T M^{-1} R(t, s^T))^T \\ & - R(s, t^T)(M^{-1} - M^{-1} S(S^T M^{-1} S)^{-1} S^T M^{-1}) R(t, s^T). \end{aligned}$$

Using the notations of Section 2, the calculation utilizes the Cholesky decomposition $W = GG^T$ and the QR-decomposition $\tilde{S} = F_1 R_1$. It can be shown

that

$$\begin{aligned}&(S^TM^{-1}S)^{-1}\\=&(\tilde{S}^T\tilde{M}^{-1}\tilde{S})^{-1}\\=&R_1^{-1}[(F_1^T\tilde{Q}F_1+n\lambda I)-(F_1^T\tilde{Q}F_2)(F_2^T\tilde{Q}F_2+n\lambda I)^{-1}(F_2^T\tilde{Q}F_1)]R_1^{-T}.\end{aligned}$$

Since $\boldsymbol{d} = R_1^{-1}[F_1^T - (F_1^T\tilde{Q}F_2)(F_2^T\tilde{Q}F_2 + n\lambda I)^{-1}F_2^T]\boldsymbol{y}$, (see Gu (1989)), the columns of $(S^TM^{-1}S)^{-1}\boldsymbol{\phi}(\boldsymbol{s}^T)$ can be computed by passing the columns of $\tilde{Q}F_1R_1^{-T}\boldsymbol{\phi}(\boldsymbol{s}^T)$ through the standard procedure for computing $\boldsymbol{d}$. This could be done with $O(n^2)$ extra flops for one dimensional $s$. Since $\boldsymbol{c} = M^{-1}(I - S(S^TM^{-1}S)^{-1}S^TM^{-1})\boldsymbol{y}$ and $\boldsymbol{d} = (S^TM^{-1}S)^{-1}S^TM^{-1}\boldsymbol{y}$, other terms could be calculated similarly. For $\boldsymbol{s} = \boldsymbol{t}$, it can be shown that $A(\lambda)W^{-1} = G^{-T}[I - n\lambda F_2(F_2^T\tilde{Q}F_2 + n\lambda I)^{-1}F_2^T]G^{-1}$, and the computation could be arranged accordingly. To calculate all the diagonal elements $O(n^3)$ flops are needed.

## 5. A Monte-Carlo Experiment

Based on Theorem 2.1 and Corollary 2.1 with $W = I$, Wahba (1983) constructed pointwise Bayesian confidence intervals on the design points and illustrated via simulations that the frequentist coverage of these intervals is rather accurate "on average". Nychka (1988) proved that the "average coverage probability" $\mathrm{ACP}(\alpha) = (1/n)\sum_{j=1}^n P(f(t_j) \in C_\alpha(t_j))$ of the intervals is asymptotically "correct" in the sense that $\mathrm{ACP}(\alpha) \to \alpha$, where the $t_j$'s are equally spaced univariate design points and $C_\alpha(t_j)$'s are the usual pointwise confidence intervals at $t_j$ based on the posterior of Theorem 2.1 with $W = I$. It is not clear if the same result can be proved for more complex covariate structures and for sampling structures other than *i.i.d.* Gaussian errors. In this section, we present a simple Monte-Carlo experiment providing positive evidence that the average coverage of the pointwise Bayesian confidence intervals based on the approximate posterior of Theorem 3.1 is likely to be "correct".

As a specialization of (3.1), consider the smoothing spline logistic regression on $\mathcal{T} = [0,1]$. With independent Bernoulli data, $\sigma^2 = 1$ and $L\boldsymbol{y}(\boldsymbol{\eta}) = \sum_{j=1}^n(y_j\eta_j - \log(1+\exp\eta_j))$, where $\eta_j$ is the logit at design point $t_j$. $\mathcal{H}$ is often taken as $W_2^2[0,1]$ with an inner product $\langle f,g\rangle = (\int f)(\int g) + (\int \dot{f})(\int \dot{g}) + (\int \ddot{f}\ddot{g})$, where $W_2^2$ contains functions with square integrable second derivatives. $\mathcal{H}_0 = \{1, t-.5\}$ with norm $(\int f)^2 + (\int \dot{f})^2$, and $\mathcal{H}_1$ is the complement of $\mathcal{H}_0$ in $W_2^2$. The roughness penalty is $\|P_1f\|^2 = \int \ddot{f}^2$. In this case $M = 2$, and $\phi_1$ and $\phi_2$ can be taken as $k_0(t) = 1$ and $k_1(t) = t - .5$. $R(t,s) = k_2(t)k_2(s) - k_4(|t-s|)$, where $k_2 = (k_1^2 - 1/12)/2$ and $k_4 = (k_1^4 - k_1^2/2 + 7/240)/24$. ($k_\nu = B_\nu/\nu!$ are scaled Bernoulli polynomials.) See, e.g., Wahba (1990), for details and other specializations. What really matters here is the roughness penalty (a semi norm) $\int \ddot{f}^2$

and its null space $\mathcal{H}_0$. Assigning a different norm on $\mathcal{H}_0$ will result in a different $\mathcal{H}_1$ and a different $R$. However, $\|P_1 f\|^2$ always remains the same; so does $f_{L,\lambda}$.

In a Monte-Carlo experiment, Bernoulli responses $y_j$ were generated on $t_j = (j - .5)/100, j = 1, \ldots, 100$, according to a "true" logit function

$$f(t) = 3[10^5 t^{11}(1-t)^6 + 10^3 t^3 (1-t)^{10}] - 2.$$

100 replicates were generated. The minimizers of (3.1) (with the foregoing specialization) were computed using the algorithm of Gu (1990) with $\lambda$ minimizing an appropriate GCV score. There were cases where the GCV score has a minimum at $\lambda = 0$ demanding an interpolation, hence a lower bound for $n\lambda$, $10^{-5}$, was applied in the calculations. Pointwise 90% and 95% Bayesian confidence intervals (symmetric in the logit scale) were computed, using $f_{L,\lambda}(t_j)$'s as the means and the diagonals of $W^{-1/2}(I - n\lambda F_2(F_2^T \tilde{Q} F_2 + n\lambda I)^{-1} F_2^T) W^{-1/2}$ as the variances. The mean average coverages were 89.05% and 93.76% respectively. Plotted in Figure 5.1 are the "true" Bernoulli probability, the estimated Bernoulli probability from the first replicate, and the corresponding 90% pointwise confidence intervals transformed into the probability scale with a "sampled" average coverage 87%. The data $y_j$'s are also plotted in Figure 5.1. The pointwise coverages out of the 100 replicates are plotted in Figure 5.2 where the magnitude of $|\ddot{f}(t)|$ is superimposed. It is rather clear that low coverage is associated with high curvature, which also appears in the simulations of Wahba (1983) and Nychka (1988); see their works for more discussions.

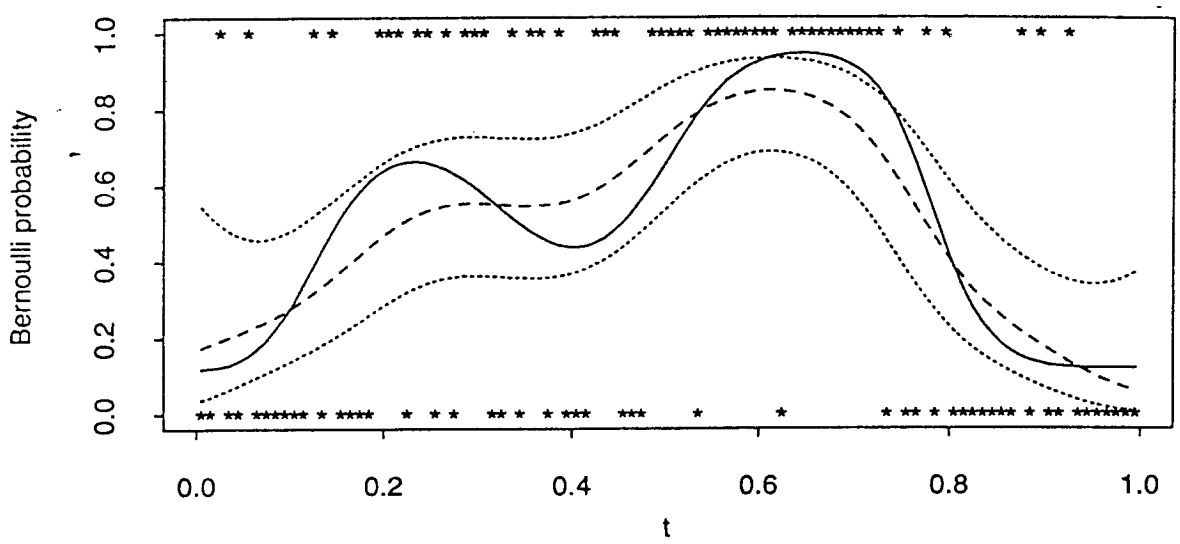


Figure 5.1. Solid line is the "true" probability. Dashed line is an estimated probability. Dotted lines connect bounds of pointwise 90% Bayesian confidence intervals. Stars indicate $y_j$'s.

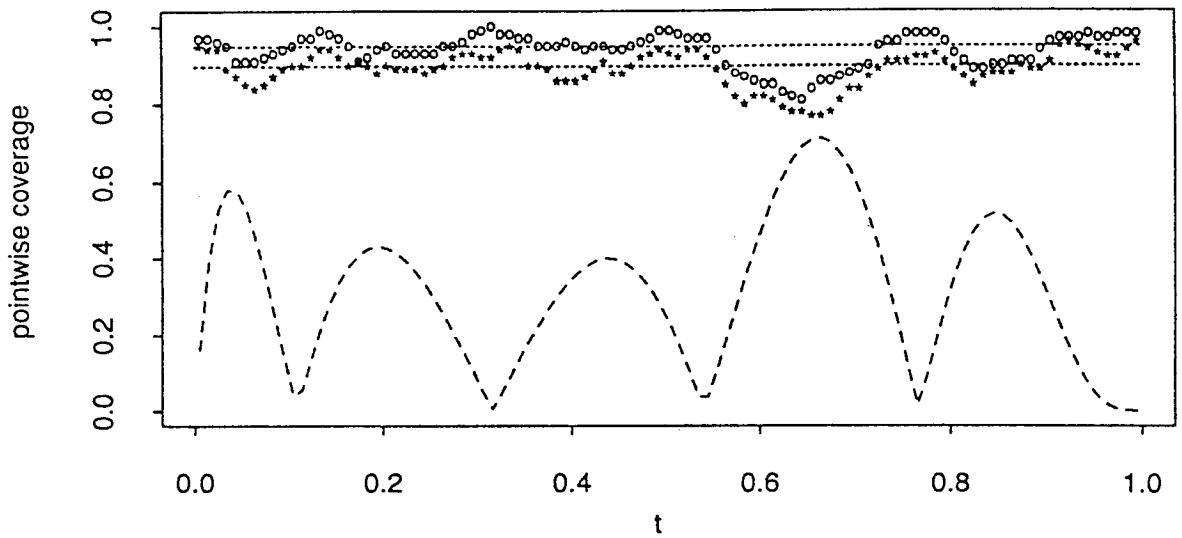


Figure 5.2. Stars are pointwise coverages of 90% intervals. Circles are pointwise coverages of 95% intervals. Dotted lines are nominal values 90% and 95%. Dashed curve is magnitude of $|\ddot{f}|$.

## Acknowledgements

This research was supported in part by AFOSR under grant AFOSR-87-0171 and by NASA under contract NAG5-316 while I was at the University of Wisconsin–Madison, and in part by the NSERC of Canada while I was at the University of British Columbia. I was fortunate to be introduced to the smoothing spline technique by Grace Wahba and to Laplace's method by Tom Leonard at Madison. My thanks also go to a referee for the helpful comments and suggestions.

## References

Gu, C. (1989). Rkpack and its applications: Fitting smoothing spline models. *Proceedings of Statistical Computing Section: American Statistical Association*, 42–51.

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

Leonard, T. (1982). Comment on "A Simple Predictive Density Function", by M. Lejeune and G. D. Faulkenberry. *J. Amer. Statist. Assoc.* **77**, 657–658.

Leonard, T., Hsu, J. and Tsui, K. (1989). Bayesian marginal inference. *J. Amer. Statist. Assoc.* **84**, 1051–1058.

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

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

Rao, C. R. (1973). *Linear Statistical Inference and Its Applications.* John Wiley.

Tierney, L. and Kadane, J. (1986). Accurate approximations for posterior moments and marginal densities. *J. Amer. Statist. Assoc.* **81**, 82–86.

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

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

Wahba, G. (1985). A comparison of GCV and GML for choosing the smoothing parameter in the generalized spline smoothing problem. *Ann. Statist.* **13**, 1378–1402.

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

Department of Statistics, Purdue University, West Lafayette, IN 47907, U.S.A.

(Received November 1989; accepted June 1991)