Penalized Least Squares Regression: Fast Computation via Efficient Approximation

YOUNG-JU KIM and CHONG GU*

Purdue University, USA.

SUMMARY

The penalized least squares method provides versatile and effective nonparametric models for regression with Gaussian responses. The computation of penalized least squares regression is generally of the order $O(n^3)$, n being the sample size, which severely limits its practical applicability. In this article, we study the fast computation of penalized least squares regression via certain low-dimensional approximations that are asymptotically as efficient. A simple algorithm is presented and the Bayes model associated with the approximations is derived, with the latter guiding the porting of Bayesian confidence intervals. The practical choice of the dimension of the approximating space is determined through simulation studies, and empirical comparisons of the approximations with the exact solution are presented. Also explored is a simple modification of the generalized cross-validation method for smoothing parameter selection, which to a large extent fixes the occasional undersmoothing problem suffered by generalized cross-validation.

Keywords: BAYESIAN CONFIDENCE INTERVAL; COMPUTATION; GENERALIZED CROSS-VALIDATION; PENALIZED LEAST SQUARES.

1 Introduction

Consider a regression problem with observations $Y_i = \eta(x_i) + \epsilon_i$, $i = 1, \ldots, n$, where $\epsilon_i \sim N(0, \sigma^2)$, independent. In a classical parametric regression analysis, η is assumed to be of the form $\eta(x, \boldsymbol{\beta})$, known up to the parameters $\boldsymbol{\beta}$, which are to be estimated from the data; when $\eta(x, \boldsymbol{\beta})$ is linear in $\boldsymbol{\beta}$, one has a standard linear model. A parametric model characterizes a set of rigid constraints on η. The dimension of the model space, i.e., the number of unknown parameters, is presumably much smaller than the sample size n.

Parametric models often incur model bias. To avoid this, an alternative approach to estimation is to allow η to vary in a high (possibly infinite) dimensional function space, leading to various

* *Address for correspondence*: Department of Statistics, Purdue University, West Lafayette, IN 47907, USA.

nonparametric or semiparametric estimation methods. A popular approach to the nonparametric estimation of η is via the minimization of a penalized least squares functional,

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \eta(x_i))^2 + \lambda J(\eta), \tag{1.1}$$

where $J(\eta)$ is a quadratic functional measuring the roughness of η. The first term in (1.1) discourages the lack of fit of η to the data, the second term penalizes the roughness of η, and the smoothing parameter λ controls the tradeoff between the two conflicting goals. An example of this is the cubic smoothing spline for univariate regression, with $J(\eta) = \int \ddot{\eta}^2 dx$.

Pioneered by the work of Kimeldorf and Wahba (1970a, 1970b, 1971), penalized least squares regression and generalizations thereof have been studied extensively over the years; see, e.g., Wahba (1990), Green and Silverman (1994), and Gu (2002) for comprehensive treatments.

The minimizer η_λ of (1.1) resides in an n-dimensional space, and despite the existence of $O(n)$ algorithms for univariate regression, the computation in multivariate settings is generally of the order $O(n^3)$. A recent result of Gu and Kim (2001) shows that the minimizers of (1.1) in certain q-dimensional spaces share the same asymptotic convergence rate with η_λ, so long as $q = O(n^{2/(pr+1)+\epsilon})$ for some $p \in [1, 2]$, $r > 1$, and $\forall \epsilon > 0$; the computation of such approximations is of the order $O(nq^2) = O(n^{(pr+5)/(pr+1)+\epsilon})$. For the cubic splines with $J(\eta) = \int \ddot{\eta}^2 dx$, $r = 4$, and for tensor products thereof, $r = 4 + \delta$, $\forall \delta > 0$. The constant p depends on the smoothness of η: for η "barely" satisfying $J(\eta) < \infty$, $p = 1$, and for η satisfying more stringent smoothness conditions, $p > 1$, up to 2. For example, with $J(\eta) = \int \ddot{\eta}^2 dx$, one has $p = 1.5$ for η satisfying $\int (\eta^{(3)})^2 dx < \infty$ and $p = 2$ for η satisfying $\int (\eta^{(4)})^2 dx < \infty$. With $r = 4 + \delta$ and $p \in [1, 2]$, the order of computation range from $O(n^{13/9+\epsilon^*})$ to $O(n^{9/5+\epsilon^*})$, where $\epsilon^* > 0$ is arbitrary. The main purpose of this article is to develop algorithms for the computation of such approximations and to resolve a host of practical and theoretical issues related to the approximations.

The performance of penalized least squares regression hinges on the proper selection of the smoothing parameter λ in (1.1), for which a popular method is the generalized cross-validation (GCV) of Craven and Wahba (1979). Despite its theoretical justification (Li 1986) and adequate practical performance, GCV may yield severe undersmoothing (too small a λ) in up to 10% of the cases. In this article, we also present empirical studies to suggest a simple modification of GCV that may curb undersmoothing for the "bad" cases without sacrificing the generally good performance for the other cases.

The rest of the article is organized as follows. In §2, we shall briefly review some basic facts concerning penalized least squares regression. The basic algorithm is outlined in §3, followed by empirical studies concerning the modification of GCV in §4. The practical choice of q is explored in

§5. The minimizer η_λ of (1.1) is known to be a Bayes estimate with a Gaussian process prior for η, and in §6 we study the Bayes model associated with the q-dimensional approximations. §7 presents further empirical results comparing the exact minimizer and the q-dimensional approximations. Real-data examples are shown in §8. A few remarks in §9 conclude the article.

2 Background

We first review some basic facts concerning penalized least squares regression and set up the notation. For general references concerning these facts, see, e.g., Wahba (1990) and Gu (2002).

2.1 Reproducing Kernel and Solution Expression

The minimization of (1.1) is in a space $\mathcal{H} \subseteq \{\eta : J(\eta) < \infty\}$ in which $J(\eta)$ is a squre (semi) norm, or a subspace therein. The evaluation $[x]f = f(x)$ appears in the first term, which is assumed to be continuous in $\mathcal{H}$. A space $\mathcal{H}$ in which the evaluation is continuous is called a reproducing kernel Hilbert space (RKHS) possessing a reproducing kernel (RK) $R(\cdot,\cdot)$, a non-negative definite function satisfying $\langle R(x,\cdot), f(\cdot)\rangle = f(x)$, $\forall \eta \in \mathcal{H}$, where $\langle\cdot,\cdot\rangle$ is the inner product in $\mathcal{H}$. The norm and the RK determine each other uniquely.

Let $\mathcal{N}_J = \{\eta : J(\eta) = 0\}$ be the null space of $J(\eta)$ and consider the tensor sum decomposition $\mathcal{H} = \mathcal{N}_J \oplus \mathcal{H}_J$. The space $\mathcal{H}_J$ is an RKHS with $J(\eta)$ as the square norm. The minimizer of (1.1) has an expression

$$\eta(x) = \textstyle\sum_{\nu=1}^m d_\nu \phi_\nu(x) + \sum_{i=1}^n c_i R_J(x_i, x),$$

where $\{\phi_\nu\}$ is a basis of $\mathcal{N}_J$ and R_J is the RK in $\mathcal{H}_J$.

For $\mathcal{X}$ a product domain, certain ANOVA decompositions can be built in through the construction of $J(\eta)$. The decomposition can be characterized by $\mathcal{H} = \oplus_{\beta=0}^g \mathcal{H}_\beta$ and $J(\eta) = \sum_{\beta=1}^g \theta_\beta^{-1} J_\beta(\eta_\beta)$, where $\eta_\beta \in \mathcal{H}_\beta$, $0 < \theta_\beta < \infty$, and J_β is the square norm in $\mathcal{H}_\beta$, $\beta > 0$. One has $\mathcal{N}_J = \mathcal{H}_0$, $\mathcal{H}_J = \oplus_{\beta=1}^g \mathcal{H}_\beta$, and $R_J = \sum_{\beta=1}^g \theta_\beta R_\beta$, where R_β is the RK in $\mathcal{H}_\beta$. The θ_β's are an extra set of smoothing parameters to be selected, but they may not appear explicitly in the notation.

2.2 Asymptotic Converegence Rate

Let $f(x)$ be the limiting density of x_i on the covariate domain $\mathcal{X}$, assumed to be bounded from above and below, and define the bilinear form $\tilde{V}(g,h) = \int_{\mathcal{X}} g(x)h(x)f(x)$. The asymptotic convergence rate of η_λ is characterized by an eigenvalue analysis of $J(\eta)$ with respect to $\tilde{V}(\eta) = \tilde{V}(\eta,\eta)$.

Let ψ_ν be the eigenfunctions satisfying $\tilde{V}(\psi_\nu, \psi_\mu) = \delta_{\nu,\mu}$ and $J(\psi_\nu, \psi_\mu) = \rho_\nu \delta_{\nu,\mu}$, where $J(g,h)$ is the bilinear form associated with $J(g)$ and $\delta_{\nu,\mu}$ is Kronecker's delta. Assume $\rho_\nu > C\nu^r$ for some

$C > 0$, $r > 1$, and ν sufficiently large. Assuming $\sum_\nu \rho_\nu^p \eta_\nu^2 < \infty$, where $\eta_\nu = \tilde{V}(\psi_\nu, \eta)$ and $p \in [1, 2]$, it can be shown that as $\lambda \to 0$ and $n\lambda^{2/r} \to \infty$,

$$\tilde{V}(\eta_\lambda - \eta) = O_p(\lambda^p + n^{-1}\lambda^{-1/r}); \tag{2.1}$$

note that $J(\eta) = \sum_\nu \rho_\nu \eta_\nu^2$. If $\mathcal{H}^* \subset \mathcal{H}$ satisfies $\tilde{V}(\eta) = O_p(\lambda J(\eta))$, $\forall h \in \mathcal{H} \ominus \mathcal{H}^*$, then the rate of (2.1) also holds for the minimizer of (1.1) in $\mathcal{H}^*$. The optimal convergence rate is $O_p(n^{-pr/(pr+1)})$, achieved with $\lambda = O(n^{-r/(pr+1)})$. For $\mathcal{H}_q = \mathcal{N}_J \oplus \text{span}\{R_J(z_j, \cdot), j = 1, \ldots, q\}$, where z_j have the limiting density $f(x)$, it can be shown that

$$\tilde{V}(h) = (\tilde{V} + \lambda J)(h) O_p(q^{-1/2}\lambda^{-1/r});$$

random subsets $\{z_j\} \subset \{x_i\}$ have the limiting density $f(x)$. Setting $q = O(\lambda^{-2/r-\tilde{\epsilon}}) = O(n^{2/(pr+1)+\epsilon})$, where $\lambda = O(n^{-r/(pr+1)})$ is optimal and $\epsilon > 0$ is arbitrary, one has $\tilde{V}(\eta) = o_p(\lambda J(\eta))$, $\forall h \in \mathcal{H} \ominus \mathcal{H}_q$. See Gu and Kim (2001).

2.3 Generalized Cross-Validation

The proper selection of the smoothing parameters λ and θ_β is essential for good practical performance. Evaluating η_λ at the sampling points x_i, one has $\hat{\mathbf{Y}} = A(\lambda)\mathbf{Y}$, where $A(\lambda)$ is known as the smoothing matrix. The popular GCV method selects the λ that minimizes

$$V(\lambda) = \frac{n^{-1}\mathbf{Y}^T(I - A(\lambda))^2\mathbf{Y}}{\{n^{-1}\text{tr}(I - A(\lambda))\}^2}. \tag{2.2}$$

Let $L(\lambda) = n^{-1}\sum_{i=1}^n (\eta_\lambda(x_i) - \eta(x_i))^2$ be the mean square error loss at data points; note that $L(\lambda) \approx \tilde{V}(\eta_\lambda - \eta)$. Under mild conditions, it was shown by Li (1986) that

$$V(\lambda) - L(\lambda) - n^{-1}\boldsymbol{\epsilon}^T\boldsymbol{\epsilon} = o_p(L(\lambda)),$$

where $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^T$; see also Gu (2002, §3.2) for a short proof of the result.

2.4 Bayes Model

The minimizer η_λ of (1.1) is known to be a Bayes estimate. Suppose $\eta = \eta_0 + \tilde{\eta}$, where η_0 has a diffuse prior in $\mathcal{N}_J$ and $\tilde{\eta}$ has a Gaussian process prior with mean 0 and covariance function $E[\tilde{\eta}(x_1)\tilde{\eta}(x_2)] = bR_J(x_1, x_2)$, independent of each other. Observing $Y_i = \eta(x_i) + \epsilon_i$, one has

$$E[\eta(x)|\mathbf{Y}] = \eta_\lambda(x),$$

where η_λ minimizes (1.1) with $\sigma^2/n\lambda = b$. One can also calculate the posterior variance under the Bayes model, which forms the basis for the Bayesian confidence intervals of Wahba (1983).

For $R_J = \sum_{\beta=1}^{g} \theta_\beta R_\beta$, $\tilde{\eta} = \sum_{\beta=1}^{g} \eta_\beta$ decomposes into independent components with covariance functions $b\theta_\beta R_\beta$. Posterior analysis of the components yields the component-wise Bayesian confidence intervals of Gu and Wahba (1993b).

3 Computation

Algorithms for the computation of η_λ were developed by Gu et al. (1989) and Gu and Wahba (1991), which employed a certain numerical structure not shared by the approximation in $\mathcal{H}_q = \mathcal{N}_J \oplus \text{span}\{R_J(z_j, \cdot), j = 1, \ldots, q\}$. In this section, we shall develop an algorithm for the computation of the approximation.

Functions in $\mathcal{H}_q$ has an expression

$$\eta(x) = \sum_{\nu=1}^{m} d_\nu \phi_\nu(x) + \sum_{j=1}^{q} c_j R_J(z_j, x) = \boldsymbol{\phi}^T \mathbf{d} + \boldsymbol{\xi}^T \mathbf{c}, \tag{3.1}$$

where $\boldsymbol{\phi}$ and $\boldsymbol{\xi}$ are vectors of functions and $\mathbf{d}$ and $\mathbf{c}$ are vectors of coefficients. Plugging (3.1) into (1.1), one estimates $\mathbf{d}$ and $\mathbf{c}$ through the minimization of

$$(\mathbf{Y} - S\mathbf{d} - R\mathbf{c})^T (\mathbf{Y} - S\mathbf{d} - R\mathbf{c}) + (n\lambda)\mathbf{c}^T Q \mathbf{c}, \tag{3.2}$$

where S is $n \times m$ with the (i, ν)th entry $\phi_\nu(x_i)$, R is $n \times q$ with the (i, j)th entry $R_J(x_i, z_j)$, and Q is $q \times q$ with the (j, k)th entry $R_J(z_j, z_k)$; it is known that $J(R_J(z_j, \cdot), R_J(z_k, \cdot)) = R_J(z_j, z_k)$, where $J(f, g)$ defines the semi inner product corresponding to the squares semi norm $J(f)$. We shall assume a full column rank for S, which ensures a unique minimizer of (1.1) even though the coefficients $\mathbf{d}$ and $\mathbf{c}$ may not be unique; see, e.g., Gu (2002 , §3.1).

Differentiating (3.2) with respect to $\mathbf{d}$ and $\mathbf{c}$ and setting the derivatives to 0, some algebra yields

$$\begin{pmatrix} S^T S & S^T R \\ R^T S & R^T R + (n\lambda) Q \end{pmatrix} \begin{pmatrix} \mathbf{d} \\ \mathbf{c} \end{pmatrix} = \begin{pmatrix} S^T \mathbf{Y} \\ R^T \mathbf{Y} \end{pmatrix}. \tag{3.3}$$

Fixing the smoothing parameters λ and θ_β (if present, hidden in R and Q) and assuming a full column rank of R, the linear system (3.3) can be easily solved by a Cholesky decomposition of the $(m+q) \times (m+q)$ matrix followed by forward and backward substitutions; see, e.g., Golub and Van Loan (1989, §4.2, §3.1).

Care must be taken when R is singular. Write the Cholesky decomposition

$$\begin{pmatrix} S^T S & S^T R \\ R^T S & R^T R + (n\lambda) Q \end{pmatrix} = \begin{pmatrix} G_1^T & O \\ G_2^T & G_3^T \end{pmatrix} \begin{pmatrix} G_1 & G_2 \\ O & G_3 \end{pmatrix}, \tag{3.4}$$

where $S^TS = G_1^TG_1$, $G_2 = G_1^{-T}S^TR$, and $G_3^TG_3 = R^T(I - S(S^TS)^{-1}S^T)R + (n\lambda)Q$. Possibly with an exchange of indices known as pivoting, one may write

$$G_3 = \begin{pmatrix} H_1 & H_2 \\ O & O \end{pmatrix} = \begin{pmatrix} H \\ O \end{pmatrix},$$

where H_1 is nonsingular. Now define

$$\tilde{G}_3 = \begin{pmatrix} H_1 & H_2 \\ O & \delta I \end{pmatrix}, \quad \tilde{G} = \begin{pmatrix} G_1 & G_2 \\ O & \tilde{G}_3 \end{pmatrix};$$

one has

$$\tilde{G}^{-1} = \begin{pmatrix} G_1^{-1} & -G_1^{-1}G_2\tilde{G}_3^{-1} \\ O & \tilde{G}_3^{-1} \end{pmatrix} \tag{3.5}$$

Premultiplying (3.3) by $\tilde{G}^{-T}$, some algebra yields

$$\begin{pmatrix} I & O \\ O & \tilde{G}_3^{-T}G_3^TG_3\tilde{G}_3^{-1} \end{pmatrix} \begin{pmatrix} \tilde{\mathbf{d}} \\ \tilde{\mathbf{c}} \end{pmatrix} = \begin{pmatrix} G_1^{-T}S^T\mathbf{Y} \\ \tilde{G}_3^{-T}R^T(I - S(S^TS)^{-1}S^T)\mathbf{Y} \end{pmatrix}, \tag{3.6}$$

where $\left(\begin{smallmatrix} \tilde{\mathbf{d}} \\ \tilde{\mathbf{c}} \end{smallmatrix}\right) = \tilde{G}\left(\begin{smallmatrix} \mathbf{d} \\ \mathbf{c} \end{smallmatrix}\right)$. Partitioning $\tilde{G}_3^{-1} = (K, L)$, one has $HK = I$ and $HL = O$, so

$$\tilde{G}_3^{-T}G_3^TG_3\tilde{G}_3^{-1} = \begin{pmatrix} K^T \\ L^T \end{pmatrix} G_3^TG_3(K, L) = \begin{pmatrix} K^T \\ L^T \end{pmatrix} H^TH(K, L) = \begin{pmatrix} I & O \\ O & O \end{pmatrix}.$$

$L^TG_3^TG_3L = O$ implies $L^TR^T(I - S(S^TS)^{-1}S^T)RL = O$, so $L^TR^T(I - S(S^TS)^{-1}S^T)\mathbf{Y} = 0$. The linear system (3.6) is thus of the form

$$\begin{pmatrix} I & O & O \\ O & I & O \\ O & O & O \end{pmatrix} \begin{pmatrix} \tilde{\mathbf{d}} \\ \tilde{\mathbf{c}}_1 \\ \tilde{\mathbf{c}}_2 \end{pmatrix} = \begin{pmatrix} * \\ * \\ 0 \end{pmatrix}, \tag{3.7}$$

which is a solvable system but $\tilde{\mathbf{c}}_2$ can be arbitrary. Replacing the lower-right block O in the matrix on the left-hand side by I, which amounts to replacing G_3 in (3.4) by $\tilde{G}_3$, one sets $\tilde{\mathbf{c}}_2 = 0$ in (3.7). In practice, one may simply perform the Cholesky decomposition of (3.4) with pivoting, replace the trailing O (if present) by δI with an appropriate value of δ, then proceed as if R were of full column rank.

It is easy to see that

$$\hat{\mathbf{Y}} = S\mathbf{d} + R\mathbf{c} = (S, R)\tilde{G}^{-1}\tilde{G}^{-T}\begin{pmatrix} S^T \\ R^T \end{pmatrix}\mathbf{Y} = A(\lambda)\mathbf{Y}.$$

The evaluation of the GCV score of (2.2) is straightforward. For the minimization of the GCV score with respect to λ and θ_β, the quasi-Newton methods of Dennis and Schnabel (1996) can be employed, which uses carefully scaled finite differences to approximate the derivatives.

The formation of the linear system (3.3) takes $O(nq^2)$ floating point operations, or flops. The Cholesky decomposition involves $O(q^3)$ flops, and the forward and backward substitutions need $O(q^2)$ flops. The calculation of $\mathrm{tr}A(\lambda)$ involves n forward substitutions, so is of the order $O(nq^2)$. The overall computation thus takes $O(nq^2)$ flops.

4 Modification of GCV

The GCV score of (2.2) was proposed by Craven and Wahba (1979) and justified by Li (1986), and is widely used for the selection of smoothing parameters in penalized least squares regression; like other versions of cross-validation in the same and similar settings, GCV may occasionally lead to severe undersmoothing. An alternative method for smoothing parameter selection is the generalized maximum likelihood (GML) derived by Wahba (1985), which never interpolates but, when $\eta(x)$ is "supersmooth" (i.e., $p > 1$) and n is large, consistently undersmoothes to a mild extent. The derivation and computation of GML in the context is discussed in §6.

In this section, we study the empirical performance of a simple modification of GCV and compare with GCV and GML. The modified GCV score is of the form

$$V_\alpha(\lambda) = \frac{n^{-1}\mathbf{Y}^T(I - A(\lambda))^2\mathbf{Y}}{\{n^{-1}\mathrm{tr}(I - \alpha A(\lambda))\}^2}, \tag{4.1}$$

where $\alpha \geq 1$; it reduces to (2.2) when $\alpha = 1$ and yields smoother estimates as α increases.

Two sets of simulations were conducted, one univariate and one multivariate. For the univariate simulation, data were generated from $Y_i = \eta_1(x_i) + \epsilon_i$, $i = 1, \ldots, n$, where $\eta_1(x) = 1 + 3\sin(2\pi x)$, $x_i = (i - 0.5)/n$, and $\epsilon_i \sim N(0, 1)$. For sample sizes $n = 100, 500$, one hundred replicates each were generated and cubic splines were calculated with $q = n$ and λ on a grid $\log_{10}(n\lambda) = (-5)(.05)(-1)$. The mean square error loss $L(\lambda) = n^{-1}\sum_{i=1}^n(\eta_\lambda(x_i) - \eta_1(x_i))^2$ was recorded for all the fits, along with the GCV scores $V_\alpha(\lambda)$ with $\alpha = 1, 1.2, 1.4, 1.6, 1.8$ and the GML score $M(\lambda)$ of (6.9). The smoothing parameter minimizing $L(\lambda)$, $V_\alpha(\lambda)$, and $M(\lambda)$ on the grid were identified and the corresponding losses extracted. Also calculated for the GCV fits was a variance estimate

$$\hat{\sigma}^2 = \frac{n^{-1}\mathbf{Y}^T(I - A(\lambda_v))^2\mathbf{Y}}{n^{-1}\mathrm{tr}(I - A(\lambda_v))} \tag{4.2}$$

proposed by Wahba (1983), where λ_v is the minimizer of $V_\alpha(\lambda)$. The performances of $V_\alpha(\lambda)$ and $M(\lambda)$ in the simulation are summarized in Figure 4.1. As seen in the left and center frames of

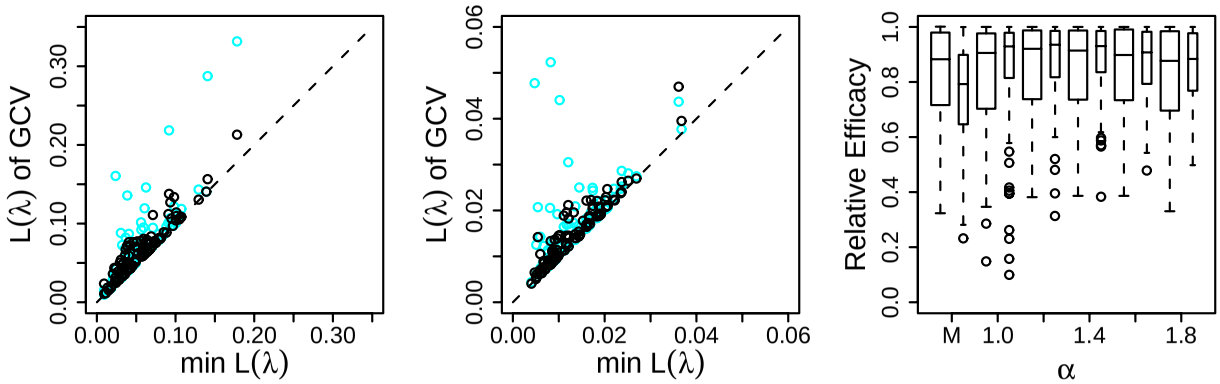


Figure 4.1: Performances of $V_\alpha(\lambda)$ and $M(\lambda)$ in Univariate Simulation. Left: Performances of $V_\alpha(\lambda)$ with $\alpha = 1$ (faded) and $\alpha = 1.4$ for $n = 100$. Center: Performances of $V_\alpha(\lambda)$ with $\alpha = 1$ (faded) and $\alpha = 1.4$ for $n = 500$. Right: min $L(\lambda)/L(\hat{\lambda})$ with $\hat{\lambda}$ minimizing $M(\lambda)$ or $V_\alpha(\lambda)$ at $\alpha = 1, 1.2, 1.4, 1.6, 1.8$, for $n = 100$ (fatter boxes) and $n = 500$ (thinner boxes).

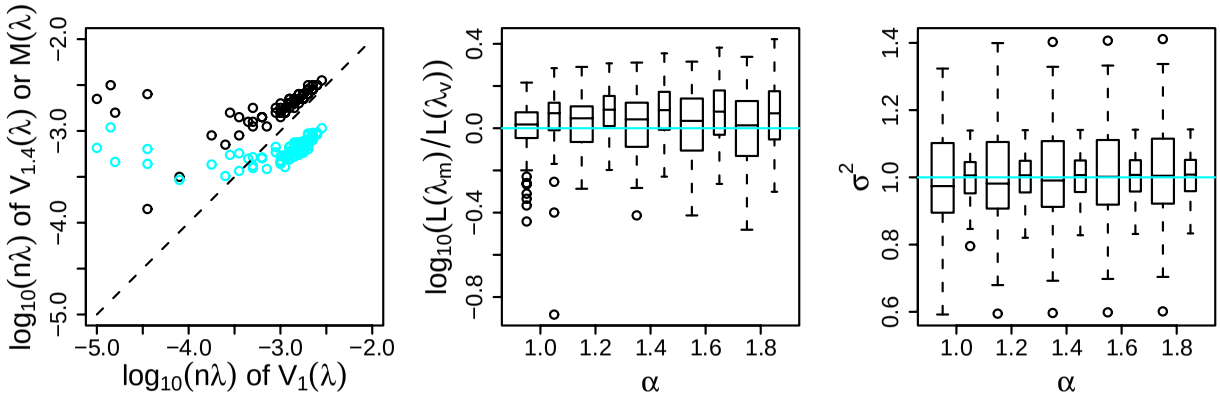


Figure 4.2: Further Results from Univariate Simulation. Left: the λ minimizing $V_1(\lambda)$ versus that minimizing $M(\lambda)$ (faded) or $V_{1.4}(\lambda)$, for $n = 100$. Center: $L(\lambda_m)/L(\lambda_v)$, with λ_m minimizing $M(\lambda)$ and λ_v minimizing $V_\alpha(\lambda)$, for $n = 100$ (fatter boxes) and $n = 500$ (thinner boxes). Right: $\hat{\sigma}^2$ of (4.2), for $n = 100$ (fatter boxes) and $n = 500$ (thinner boxes).

Figure 4.1, the occasional wild failures of GCV were effectively eliminated with $\alpha = 1.4$. The box plots in the right frame of Figure 4.1 show that the best performance of GCV was achieved with α in the range $1.2 \sim 1.4$, and that the performance of GML degraded as n increased, as forecasted by the asymptotic analysis of Wahba (1985). Further empirical results are shown in Figure 4.2, where the left frame compares the minimizers of $V_1(\lambda)$, $V_{1.4}(\lambda)$, and $M(\lambda)$, the center frame compares the performances of $V_\alpha(\lambda)$ and $M(\lambda)$, and the right frame assesses the performance of the variance estimate of (4.2).

In the notation of §2.1 and §3, the formulation of cubic spline on $[0, 1]$ used in the simulation

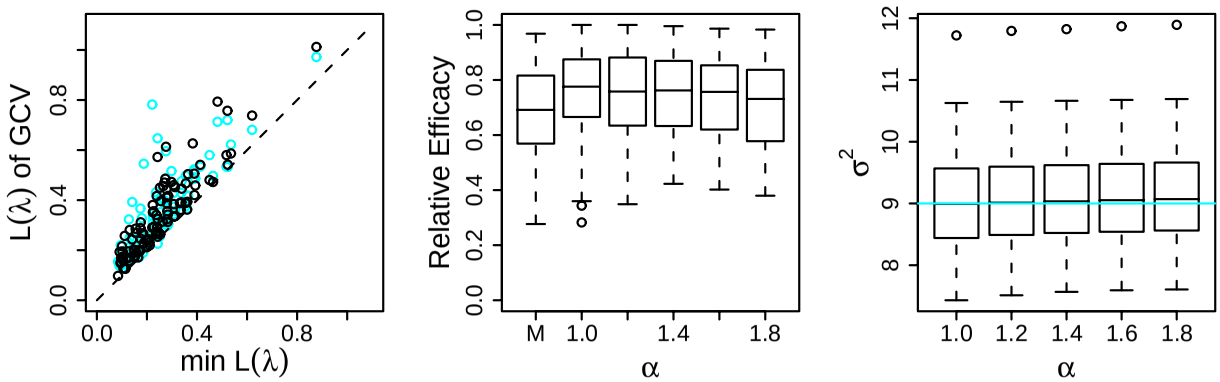


Figure 4.3: Performance of $V_\alpha(\lambda)$, $M(\lambda)$, and $\hat{\sigma}^2$ in Multivariate Simulation. Left: Performances of $V_\alpha(\lambda)$ with $\alpha = 1$ (faded) and $\alpha = 1.4$. Center: min $L(\lambda)/L(\hat{\lambda})$ with $\hat{\lambda}$ minimizing $M(\lambda)$ or $V_\alpha(\lambda)$ at $\alpha = 1, 1.2, 1.4, 1.6, 1.8$. Right: $\hat{\sigma}^2$ of (4.2).

has $m = 2$ with $\phi_1(x) = 1$, $\phi_2(x) = x - 0.5$, and

$$R_J(x_1, x_2) = R_c(x_1, x_2) = k_2(x_1)k_2(x_2) - k_4(|x_1 - x_2|), \tag{4.3}$$

where

$$k_2(x) = \frac{1}{2}\left(k_1^2(x) - \frac{1}{12}\right),$$
$$k_4(x) = \frac{1}{24}\left(k_1^4(x) - \frac{k_1^2(x)}{2} + \frac{7}{240}\right),$$

and $k_1(x) = x - 0.5$. See, e.g., Gu (2002, §2.3.3) for further details.

For the multivariate simulation, data were generated from $Y_i = \eta_2(x_i) + \epsilon_i$, $i = 1, \ldots, n$, where $x_i \sim U(0,1)^2$, $\epsilon_i \sim N(0, 3^2)$, and

$$\eta_2(x) = 5 + \exp(3x_{\langle 1\rangle}) + 10^6 x_{\langle 2\rangle}^{11}(1 - x_{\langle 2\rangle})^6 + 10^4 x_{\langle 2\rangle}^3(1 - x_{\langle 2\rangle})^{10} + 5\cos(2\pi(x_{\langle 1\rangle} - x_{\langle 2\rangle})),$$

with the notation $x = (x_{\langle 1\rangle}, x_{\langle 2\rangle}) \in (0,1)^2$. For a sample size $n = 300$, one hundred replicates were generated and tensor product cubic splines were fitted with $q = 150$ and with the smoothing parameters minimizing $L(\lambda)$, $V_\alpha(\lambda)$ with $\alpha = 1, 1.2, 1.4, 1.6, 1.8$, and $M(\lambda)$, respectively. See §5 for the choice of q. The simulation results are summarized in Figure 4.3.

In the notation of §2.1 and §3, the formulation of tensor product cubic spline on $[0,1]^2$ used in the simulation has $m = 4$ with $\phi_1(x) = 1$, $\phi_2(x) = k_1(x_{\langle 1\rangle})$, $\phi_3(x) = k_1(x_{\langle 2\rangle})$, $\phi_4(x) = k_1(x_{\langle 1\rangle})k_1(x_{\langle 2\rangle})$,

where $k_1(u) = u - 0.5$, and

$$\begin{aligned} R_J(x_1, x_2) &= \theta_1 R_c(x_{1\langle 1\rangle}, x_{2\langle 1\rangle}) + \theta_2 R_c(x_{1\langle 2\rangle}, x_{2\langle 2\rangle}) \\ &\quad + \theta_3 R_c(x_{1\langle 1\rangle}, x_{2\langle 1\rangle}) k_1(x_{1\langle 2\rangle}) k_1(x_{2\langle 2\rangle}) + \theta_4 k_1(x_{1\langle 1\rangle}) k_1(x_{2\langle 1\rangle}) R_c(x_{1\langle 2\rangle}, x_{2\langle 2\rangle}) \\ &\quad + \theta_5 R_c(x_{1\langle 1\rangle}, x_{2\langle 1\rangle}) R_c(x_{1\langle 2\rangle}, x_{2\langle 2\rangle}) \\ &= \textstyle\sum_{\beta=1}^{5} \theta_\beta R_\beta(x_1, x_2) \end{aligned} \tag{4.4}$$

with 5 θ_β's, where $R_c(u_1, u_2)$ is given in (4.3) and $x_1 = (x_{1\langle 1\rangle}, x_{1\langle 2\rangle})$, $x_2 = (x_{2\langle 1\rangle}, x_{2\langle 2\rangle})$. See, e.g., Gu (2002, §2.4.3) for further details.

5 Empirical Choice of q

As mentioned in §2.2, a q of the order $O(n^{2/(pr+1)+\epsilon})$, $\forall \epsilon > 0$, is sufficient for asymptotic efficiency. For the cubic spline, $r = 4$, and for the tensor product cubic splines, $r = 4 + \delta$, $\forall \delta > 0$; see, e.g., Gu (2002, Chapter 8). Since $\epsilon, \delta > 0$ can be arbitrarily small, one may use $q = kn^{2/(pr+1)}$ in practice. In this section, we conduct some simulations to suggest adequate values of k for practical use.

Consider the test functions $\eta_1(x)$ and $\eta_2(x)$ as used in §4, which are smooth enough so $p = 2$. Samples of sizes $n = 100, 300, 500$ were drawn from the two simulation settings of §4, respectively. For each of the 6 samples and every k on the grid $k = 5(1)15$, 30 different random subsets $\{z_j\} \subset \{x_i\}$ of size $q = kn^{2/9}$ were generated, and (tensor product) cubic splines were fitted to the data with the smoothing parameters minimizing $V_\alpha(\lambda)$ with $\alpha = 1.4$. The fits with $q = n$ were also calculated. The losses were recorded for all the fits and the results are summarized in Figure 5.1 in box plots. The fact that the box width gradually decreases as k increases indicates that $q = O(n^{2/9})$ is the "correct" scale; similar plots on the $q = O(n^{2/5})$ scale (not shown here) have also been inspected, but the box width shrinks at a much faster rate there. The plots suggest that a k around 10 could be stable enough for practical use.

In practice, we suggest the use of $q = kn^{2/9}$ with k around 10 for (tensor product) cubic splines; examples with "barely" square integrable second derivatives may be artifically constructed but we doubt there are many of such "true" functions out in the real world. Since the computation is so much faster (some timing results can be found in §8), quick checks on the stability can be performed simply by comparing estimates with different random subsets $\{z_j\} \subset \{x_i\}$.

For the multivariate simulation of §4 with $n = 300$, we used $q = 150 \approx 15n^{2/5} \approx 42n^{2/9}$. Our purpose there was to study the behavior of cross-validation, and we chose a q sufficiently large to ensure stability, yet small enough so the experiments were feasible within a reasonable time frame.

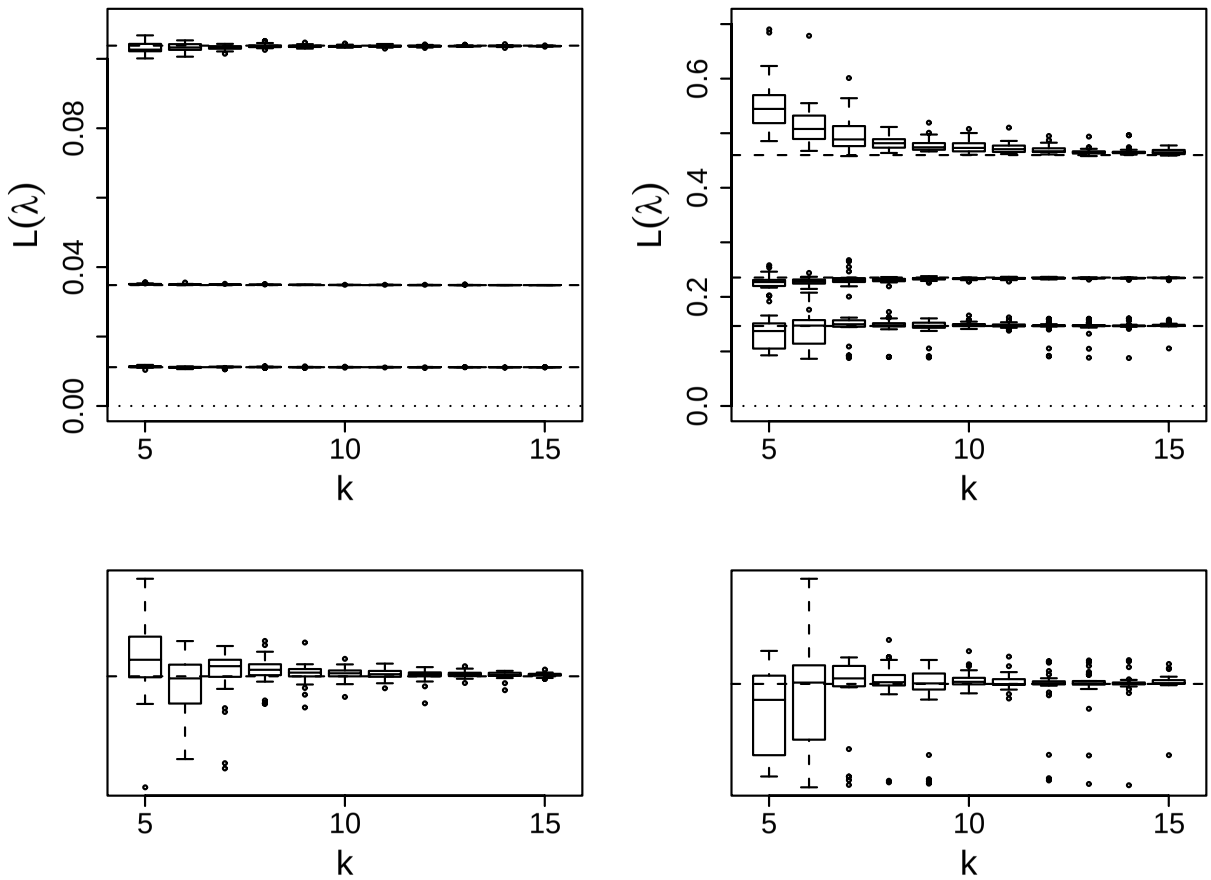


Figure 5.1: Effect of q on Estimation Consistency. Boxplots of $L(\lambda)$ with 30 different random subsets $\{z_j\}$ of size q, for each of $q = kn^{2/9}$. Left: $\eta_1(x)$. Right: $\eta_2(x)$. Top: from high to low, $n = 100, 300, 500$. Bottom: $n = 500$ with better resolution. The dashed lines correspond to $q = n$.

6 Bayes Model

Based on the Bayes model briefly discussed in §2.4, Bayesian confidence intervals were derived by Wahba (1983) and Gu and Wahba (1993b) to provide interval estimates in penalized least squares regression. The Bayes model is also the basis for the GML method for smoothing parameter selection derived by Wahba (1985). In this section, we extend the Bayes model to the approximation in the space $\mathcal{H}_q = \mathcal{N}_J \oplus \text{span}\{R_J(z_j, \cdot), j = 1, \ldots, q\}$.

Similar to §2.4, consider $\eta = \eta_0 + \tilde{\eta}$, where η_0 has a diffuse prior in $\mathcal{N}_J$ and $\tilde{\eta}$ has a mean 0 Gaussian process prior with the covariance function

$$E[\tilde{\eta}(x_1)\tilde{\eta}(x_2)] = bR_J(x_1, \mathbf{z}^T)Q^+R_J(\mathbf{z}, x_2),$$

where Q^+ is the Moore-Penrose inverse of $Q = R_J(\mathbf{z}, \mathbf{z}^T)$. Recall the notation $\xi_j = R_J(z_j, \cdot)$ from

§3 and $R^T = \boldsymbol{\xi}(\mathbf{x}^T)$, where $\mathbf{x} = (x_1, \ldots, x_n)^T$. Since J is the square norm in span$\{\xi_j\}$, $J(\boldsymbol{\xi}^T\mathbf{c}) = \mathbf{c}^T Q\mathbf{c} = 0$ implies $\boldsymbol{\xi}^T\mathbf{c} = 0$, so $\boldsymbol{\xi}(x)$ is in the column space of Q, $\forall x$, and hence $QQ^+R^T = R^T$, where $QQ^+ = Q^+Q$ is the projection matrix in the column space of Q.

Following Wahba (1978), we first assume $\eta_0(x) = \mathbf{b}^T\boldsymbol{\phi}(x)$ with $\mathbf{b} \sim N(0, \tau^2 I)$, then let $\tau^2 \to \infty$. It is easy to see that $\mathbf{Y}$ and $\eta(x)$ are jointly normal with mean 0 and covariance matrix

$$\begin{pmatrix} bRQ^+R^T + \tau^2 SS^T + \sigma^2 I & bRQ^+\boldsymbol{\xi} + \tau^2 S\boldsymbol{\phi} \\ b\boldsymbol{\xi}^T Q^+ R^T + \tau^2 \boldsymbol{\phi}^T S^T & b\boldsymbol{\xi}^T Q^+\boldsymbol{\xi} + \tau^2 \boldsymbol{\phi}^T\boldsymbol{\phi} \end{pmatrix} \tag{6.1}$$

where $S^T = \boldsymbol{\phi}(\mathbf{x}^T)$ as given in §3. Standard calculation yields

$$\begin{aligned} E[\eta(x)|Y] &= (b\boldsymbol{\xi}^T Q^+ R^T + \tau^2\boldsymbol{\phi}^T S^T)(bRQ^+R^T + \tau^2 SS^T + \sigma^2 I)^{-1} Y \\ &= \rho\boldsymbol{\phi}^T S^T (M + \rho SS^T)^{-1}\mathbf{Y} + \boldsymbol{\xi}^T Q^+ R^T (M + \rho SS^T)^{-1}\mathbf{Y}, \end{aligned}$$

where $\rho = \tau^2/b$, $n\lambda = \sigma^2/b$, and $M = RQ^+R^T + n\lambda I$. Letting $\rho \to \infty$, by (2.7) and (2.8) of Wahba (1978), that

$$\begin{aligned} &\lim_{\rho\to\infty} (\rho SS^T + M)^{-1} = M^{-1} - M^{-1}S(S^T M^{-1} S)^{-1} S^T M^{-1}, \\ &\lim_{\rho\to\infty} \rho S^T(\rho SS^T + M)^{-1} = (S^T M^{-1} S)^{-1} S^T M^{-1}, \end{aligned}$$

one has

$$\lim_{\tau^2\to\infty} E[\eta(x)|Y] = \boldsymbol{\phi}^T\mathbf{d} + \boldsymbol{\xi}^T\mathbf{c}, \tag{6.2}$$

where

$$\begin{aligned} \mathbf{d} &= (S^T M^{-1} S)^{-1} S^T M^{-1}\mathbf{Y}, \\ \mathbf{c} &= Q^+ R^T (M^{-1} - M^{-1}S(S^T M^{-1} S)^{-1} S^T M^{-1})\mathbf{Y}. \end{aligned} \tag{6.3}$$

It is straightforward to verify that the $\mathbf{d}$ and $\mathbf{c}$ given in (6.3) solve the linear system (3.3); the fact $QQ^+R^T = R^T$ is needed in the verification. Similarly, one has from (6.1),

$$\mathrm{var}[\eta(x)|Y]/b = \boldsymbol{\xi}^T Q^+\boldsymbol{\xi} + \rho\boldsymbol{\phi}^T\boldsymbol{\phi} - (\boldsymbol{\xi}^T Q^+ R^T + \rho\boldsymbol{\phi}^T S^T)(M + \rho SS^T)^{-1}(RQ^+\boldsymbol{\xi} + \rho S\boldsymbol{\phi}),$$

which, as $\rho \to \infty$, yields

$$\begin{aligned} \lim_{\tau^2\to\infty} \mathrm{var}[\eta(x)|Y]/b = {} & \boldsymbol{\xi}^T Q^+\boldsymbol{\xi} + \boldsymbol{\phi}^T (S^T M^{-1} S)^{-1}\boldsymbol{\phi} \\ & - \boldsymbol{\phi}^T (S^T M^{-1} S)^{-1} S^T M^{-1} RQ^+\boldsymbol{\xi} - \boldsymbol{\xi}^T Q^+ R^T M^{-1} S(S^T M^{-1} S)^{-1}\boldsymbol{\phi} \\ & - \boldsymbol{\xi}^T Q^+ R^T (M^{-1} - M^{-1}S(S^T M^{-1} S)^{-1} S^T M^{-1}) RQ^+\boldsymbol{\xi}, \end{aligned} \tag{6.4}$$

where (2.15) of Wahba (1983), that

$$\lim_{\rho\to\infty} \rho I - \rho^2 S^T(\rho SS^T + M)^{-1}S = (S^T M^{-1} S)^{-1},$$

is also needed. From (6.3), it is easy to verify that

$$A(\lambda) = I - n\lambda(M^{-1} - M^{-1}S(S^T M^{-1} S)^{-1} S^T M^{-1}). \tag{6.5}$$

Replacing $\boldsymbol{\xi}$ by R^T and $\boldsymbol{\phi}$ by S^T in (6.4), the matrix reduces to $n\lambda A(\lambda)$ after some algebra, where $A(\lambda)$ is as given in (6.5). The posterior variances at the sampling points are thus given by the diagonals of $\sigma^2 A(\lambda)$.

For the computation of (6.4) away from the sampling points, the second and third lines involve formulas of $\mathbf{d}$ and $\mathbf{c}$ given in (6.3) but with $RQ^+\boldsymbol{\xi}$ replacing $\mathbf{Y}$, which are available through forward and backward substitutions given the Cholesky factor $\tilde{G}$ in §3. It shall be shown below that $(S^T M^{-1} S)^{-1}$ is $n\lambda$ times the upper-left block of $\tilde{G}^{-1}\tilde{G}^{-T}$, where $\tilde{G}^{-1}$ is given in (3.5), so the term $\boldsymbol{\phi}^T(S^T M^{-1} S)^{-1}\boldsymbol{\phi}$ is available through forward substitution.

We now show that

$$\begin{aligned}(S^T M^{-1} S)^{-1} &= (n\lambda)(G_1^{-1}G_1^{-T} + G_1^{-1}G_2\tilde{G}_3^{-1}\tilde{G}_3^{-T}G_2^T G_1^{-T}) \\ &= (n\lambda)\{(S^T S)^{-1} + (S^T S)^{-1}S^T R\tilde{G}_3^{-1}\tilde{G}_3^{-T}R^T S(S^T S)^{-1}\},\end{aligned} \tag{6.6}$$

where the notation follows §3. First note that $M^{-1} = (n\lambda)^{-1}(I - R(n\lambda Q + R^T R)^+ R^T)$; multiply and simplify using the fact that $QQ^+R^T = (n\lambda Q + R^T R)(n\lambda Q + R^T R)^+ R^T = R^T$. Multiplying $S^T(I - R(n\lambda Q + R^T R)^+ R^T)S$ and the right-hand side of (6.6) and using the relations $G_3^T G_3 = R^T(I - S(S^T S)^{-1}S)R + n\lambda Q$ and $G_3^T G_3 \tilde{G}_3^{-1}\tilde{G}_3^{-T}R^T = R^T$, straightforward algebra yields (6.6).

For $R_J = \sum_{\beta=1}^g \theta_\beta R_\beta$, $\tilde{\eta} = \sum_{\beta=1}^g \eta_\beta$ decomposes into multiple components with the prior covariance functions given by

$$E[\eta_\beta(x_1)\eta_\gamma(x_2)] = b(\theta_\beta R_\beta(x_1, \mathbf{z}^T))Q^+(\theta_\gamma R_\gamma(\mathbf{z}, x_2)), \quad \beta, \gamma = 1, \ldots, g.$$

Also decompose the fixed effect $\eta_0 = \sum_{\nu=1}^m \psi_\nu$, where $\psi_\nu = b_\nu \phi_\nu$. The posterior means and variances of arbitrary linear combinations of ψ_ν and η_β can be obtained by simple modifications of (6.2) and (6.4). For example, with $\psi_1 + \eta_1 + \eta_2$, one simply replaces $\boldsymbol{\phi}$ in (6.2) and (6.4) by $(\phi_1(x), 0, \ldots, 0)^T$ and $\boldsymbol{\xi}$ by $\theta_1 R_1(\mathbf{z}, x) + \theta_2 R_2(\mathbf{z}, x)$.

The Bayes model under study can be perceived as a mixed effect model, with $\eta_0 = \boldsymbol{\phi}^T\boldsymbol{\beta}$ being the fixed effect and $\tilde{\eta}$ being the random effect. The GML method of Wahba (1985) is virtually an application of the popular restricted maximum likelihood (REML) method, which is widely used

for the estimation of variance components.

Let $S = F\tilde{T} = (F_1, F_2)\binom{T}{O} = F_1 T$ be the QR-decomposition of S, where F is orthogonal and T is upper triangular. REML maximizes the marginal likelihood of $F_2^T\mathbf{Y}$, which is normal with mean 0 and covariance matrix $bF_2^T RQ^+R^T F_2 + \sigma^2 I = bF_2^T MF_2$. The minus log likelihood is seen to be

$$\frac{1}{2b}\mathbf{Y}^T F_2(F_2^T MF_2)^{-1}F_2^T\mathbf{Y} + \frac{1}{2}\log|F_2^T MF_2| + \frac{n-m}{2}\log b. \tag{6.7}$$

Fixing λ in M and maximizing (6.7) with respect to b, one has

$$\hat{b} = \frac{\mathbf{Y}^T F_2(F_2^T MF_2)^{-1}F_2^T\mathbf{Y}}{n-m},$$

and the profile log likelihood of λ is monotone in

$$\log \mathbf{Y}^T F_2(F_2^T MF_2)^{-1}F_2^T\mathbf{Y} + \frac{1}{n-m}\log|F_2^T MF_2|. \tag{6.8}$$

Partition $U = (F^T MF)^{-1} = F^T M^{-1}F = \begin{pmatrix} F_1^T M^{-1}F_1 & F_1^T M^{-1}F_2 \\ F_2^T M^{-1}F_1 & F_2^T M^{-1}F_2 \end{pmatrix}$. By a standard result in Rao (1973), page 33, the bottom-right block of $U^{-1} = F^T MF$ is given by

$$(F_2^T M^{-1}F_2 - F_2^T M^{-1}F_1(F_1^T M^{-1}F_1)^{-1}F_1^T M^{-1}F_2)^{-1}.$$

Note that (6.5) holds with F_1 replacing S, so one has $(F_2^T MF_2)^{-1} = (n\lambda)^{-1}F_2^T(I - A(\lambda))F_2$. Now since $S^T(I - A(\lambda)) = O$, $F_2(F_2^T MF_2)^{-1}F_2^T = (n\lambda)^{-1}(I - A(\lambda))$, and hence (6.8) is equivalent to

$$M(\lambda) = \frac{\mathbf{Y}^T(I - A(\lambda))\mathbf{Y}}{|I - A(\lambda)|_+^{1/(n-m)}}, \tag{6.9}$$

where $|\cdot|_+$ denotes the product of positive eigenvalues.

The numerator of (6.9) is readily available. For the denominator, note that

$$|I - A(\lambda)|_+^{-1} = |(n\lambda)^{-1}F_2 MF_2| = |(n\lambda)^{-1}F_2^T RQ^+R^T F_2 + I| = |(n\lambda)^{-1}Q^+R^T F_2 F_2^T R + I|.$$

Let $Q^+ = (P_1, P_2)\begin{pmatrix} D_Q^{-1} & O \\ O & O \end{pmatrix}\begin{pmatrix} P_1^T \\ P_2^T \end{pmatrix} = P_1 D_Q^{-1}P_1^T$ be the eigenvalue decomposition of Q^+, where D_Q is diagonal with the positive eigen values of Q. Noting that $P_2^T R^T = O$, one has

$$|(n\lambda)^{-1}Q^+R^T F_2 F_2^T R + I| = |D_Q^{-1}(n\lambda)^{-1}P_1^T R^T F_2 F_2^T RP_1 + I| = |Q + (n\lambda)^{-1}R^T F_2 F_2^T R|_+/|Q|_+.$$

The formation of $R^T F_2 F_2^T R$ is $O(nq^2)$ and the eigenvalue problem is $O(q^3)$.

		0%	1%	5%	25%	50%	75%	95%	99%	100%
$\lvert\tilde{\eta}-\hat{\eta}\rvert/e$:	$n=100$	0.0000	0.0000	0.0004	0.0022	0.0050	0.0102	0.0287	0.0665	0.5175
	$n=300$	0.0000	0.0001	0.0003	0.0017	0.0040	0.0081	0.0209	0.0425	0.4868
$s_{\tilde{\eta}}/s_{\hat{\eta}}$:	$n=100$	0.9169	0.9757	0.9863	0.9959	0.9991	1.0003	1.0020	1.0055	1.0469
	$n=300$	0.9466	0.9791	0.9871	0.9964	0.9992	1.0004	1.0019	1.0041	1.0132

Table 7.1: Quantiles of $|\tilde{\eta}(x_i)-\hat{\eta}(x_i)|/\sqrt{L}$ and $s_{\tilde{\eta}}(x_i)/s_{\hat{\eta}}(x_i)$ in $\eta_1(x)$ Simulation: $n=100,300$.

7 Numerical Accuracy

Based on the posterior mean and posterior standard deviation under the Bayes model for the exact solution with $q=n$, Wahba (1983) constructed the so-called Bayesian confidence intervals. Despite the derivation under the Bayes model, the intervals demonstrate a certain frequentist across-the-function coverage property with the smoothing parameter selected by GCV; see, e.g., Wahba (1983) and Nychka (1988). Component-wise intervals were studied by Gu and Wahba (1993a).

Under the Bayes model specified in §6 for $q<n$, the formulas of the posterior mean and posterior standard deviation on the sampling points, when expressed in terms of the smoothing matrix $A(\lambda)$, match those of Wahba (1978, 1983) for $q=n$; the formula of $M(\lambda)$ also matches that in Wahba (1985). The analysis of §6 further allows the approximation of posterior mean and posterior standard deviation away from the sampling points and in component-wise calculations. We now assess the numerical accuracy of the approximation through simulation studies.

Consider again the simulation settings of §4 and §5. For sample size $n=100$, one hundred replicates were drawn from the $\eta_1(x)$ simulation, fits were calculated with $q=n$, and the posterior means $\hat{\eta}(x_i)$ and the posterior standard deviations $s_{\hat{\eta}}(x_i)$ were calculated on the sampling points. For each of the replicates, 10 different random subsets $\{z_j\}\subset\{x_i\}$ of size $q=10n^{2/9}$ were generated and fits were calculated, and the posterior means $\tilde{\eta}(x_i)$ and the posterior standard deviations $s_{\tilde{\eta}}(x_i)$ were calculated. All fits were calculated with λ minimizing $V_\alpha(\lambda)$ for $\alpha=1.4$. The standardized differences $|\tilde{\eta}(x_i)-\hat{\eta}(x_i)|/\sqrt{L}$ in posterior mean and the ratios $s_{\tilde{\eta}}(x_i)/s_{\hat{\eta}}(x_i)$ in posterior standard deviation were recorded, where $L=e^2=n^{-1}\sum_{i=1}^n(\hat{\eta}(x_i)-\eta(x_i))^2$ was the mean square error loss of the fit with $q=n$. This yielded $100(10)(100)=10^5$ entries of differences and ratios. The experiment was repeated for sample size $n=300$ on 30 replicates, yielding $30(10)(300)=9\times10^4$ entries of differences and ratios. These results are summarized in Table 7.1.

For the $\eta_2(x)$ simulation, the same experiment was conducted on ten replicates for sample size $n=300$. Functions on $[0,1]^2$ can be decomposed as

$$\eta(x)=\eta_\emptyset+\eta_{\langle1\rangle}(x_{\langle1\rangle})+\eta_{\langle2\rangle}(x_{\langle2\rangle})+\eta_{\langle1,2\rangle}(x_{\langle1\rangle},x_{\langle2\rangle}), \tag{7.1}$$

		0%	1%	5%	25%	50%	75%	95%	99%	100%
$\lvert\tilde{\eta}-\hat{\eta}\rvert/e$:	η	0.0000	0.0005	0.0024	0.0094	0.0271	0.0544	0.1678	0.3272	1.1193
	$\eta_{\langle 1\rangle}$	0.0000	0.0005	0.0021	0.0094	0.0388	0.1061	0.4626	1.0486	2.0120
	$\eta_{\langle 2\rangle}$	0.0000	0.0004	0.0025	0.0128	0.0503	0.1264	0.4468	1.5058	2.6709
	$\eta_{\langle 1,2\rangle}$	0.0000	0.0002	0.0008	0.0041	0.0237	0.0769	0.3954	0.9532	2.2575
$s_{\tilde{\eta}}/s_{\hat{\eta}}$:	η	0.2954	0.8724	0.9419	0.9868	0.9990	1.0086	1.1264	2.0861	7.4497
	$\eta_{\langle 1\rangle}$	0.1471	0.7544	0.9756	0.9947	1.0026	1.0099	1.0798	1.3310	9.7389
	$\eta_{\langle 2\rangle}$	0.0004	0.9131	0.9710	0.9981	1.0050	1.0157	1.0860	1.7120	826.33
	$\eta_{\langle 1,2\rangle}$	0.0057	0.4279	0.8866	0.9995	1.0243	1.1019	2.0818	7.7034	378.96

Table 7.2: Quantiles of $|\tilde{\eta}(x_i)-\hat{\eta}(x_i)|/\sqrt{L}$ and $s_{\tilde{\eta}}(x_i)/s_{\hat{\eta}}(x_i)$ in $\eta_2(x)$ Simulation: $n=300$.

where $\eta_{\langle 1\rangle}$, $\eta_{\langle 2\rangle}$, and $\eta_{\langle 1,2\rangle}$ satisfy certain side conditions to ensure identifiability; this is an ANOVA decomposition with $\eta_{\langle 1\rangle}$ and $\eta_{\langle 2\rangle}$ being the main effects and $\eta_{\langle 1,2\rangle}$ being the interaction. For the formulation of tensor product cubic spline as described in §4, one has the ANOVA decomposition of the posterior mean

$$\begin{aligned}
\eta(x) &= \textstyle\sum_{\nu=1}^{4} d_\nu \phi_\nu(x) + \sum_{j=1}^{q} c_j \big(\sum_{\beta=1}^{5} \theta_\beta R_\beta(z_j,x)\big) \\
&= \big[d_1\phi_1(x)\big] + \big[d_2\phi_2(x) + \textstyle\sum_{j=1}^{q} c_j\theta_1 R_1(z_j,x)\big] \\
&\quad + \big[d_3\phi_3(x) + \textstyle\sum_{j=1}^{q} c_j\theta_2 R_2(z_j,x)\big] + \big[d_4\phi_4(x) + \textstyle\sum_{j=1}^{q} c_j\big(\sum_{\beta=3}^{5}\theta_\beta R_\beta(z_j,x)\big)\big] \\
&= \eta_\emptyset + \eta_{\langle 1\rangle} + \eta_{\langle 2\rangle} + \eta_{\langle 1,2\rangle},
\end{aligned}$$

with the side conditions

$$\int_0^1 \eta_{\langle 1\rangle} dx_{\langle 1\rangle} = \int_0^1 \eta_{\langle 2\rangle} dx_{\langle 2\rangle} = \int_0^1 \eta_{\langle 1,2\rangle} dx_{\langle 1\rangle} = \int_0^1 \eta_{\langle 1,2\rangle} dx_{\langle 2\rangle} = 0;$$

see (4.4) and the text prior to that for the notation of ϕ_ν and R_β. Besides those of the overall function $\eta(x)$, component-wise differences $|\tilde{\eta}(x_i)-\hat{\eta}(x_i)|/\sqrt{L}$ and ratios $s_{\tilde{\eta}}(x_i)/s_{\hat{\eta}}(x_i)$ were also calculated, where the loss $L=e^2$ was taken as the respective component-wise versions. The results are summarized in Table 7.2. For η, the range of $e=\sqrt{L}$ for the 10 replicates was $(0.486, 0.771)$, but for $\eta_{\langle 1\rangle}$, $\eta_{\langle 2\rangle}$, and $\eta_{\langle 1,2\rangle}$, the e ranges were $(1.932, 9.165)$, $(1.139, 7.795)$, and $(3.207, 10.724)$, respectively; note that the identifiability of the components are defined through integrations over the domain but the comparison of fits and the calculation of L were done on the sampling points. Things were not as favorable as in the $\eta_1(x)$ simulation but the overall accuracy appears to be okey given the moderate signal to noise ratio.

8 Examples

We now apply the techniques developed to a couple of real data sets. The primary goal here is not data analysis, however; the purpose is to visually compare the approximation with the exact solution with $q = n$. Also of interest are timing results in real-data applications.

All fits presented below were calculated with the smoothing parameters minimizing $V_\alpha(\lambda)$ for $\alpha = 1.4$. The timing results were obtained on a workstation with Athlon MP1800+ and 2GB RAMS, running FreeBSD 4.4 and R 1.4.0.

8.1 Ozone in Los Angeles Basin

Daily measurements of ozone concentration and eight meteorological quantities in the Los Angeles basin were recorded for 330 days of 1976. The data were used by Breiman and Friedman (1985) to illustrate their ACE algorithm (alternating conditional expectation) and by Buja et al. (1989) to illustrate nonparametric additive models through the back-fitting algorithm. An analysis of the data using penalized least squares regression with $q = n$ can be found in Gu (2002, §3.7.2).

Following Gu (2002, §3.7.2), a tensor product cubic spline model of the form

$$Y = \eta_\emptyset + \eta_1(x_{\langle 1\rangle}) + \eta_2(x_{\langle 2\rangle}) + \eta_3(x_{\langle 3\rangle}) + \eta_{1,3}(x_{\langle 1\rangle}, x_{\langle 3\rangle}) + \epsilon$$

was fitted to the data, where Y was $\log_{10}$(ozone concentration) (ppm), $x_{\langle 1\rangle}$ was inversion base temperature (^oF), $x_{\langle 2\rangle}$ was Dagget pressure gradient (mmHg), and $x_{\langle 3\rangle}$ was visibility (miles). The observed x_i were mapped into $(0,1)^3$, on which the tensor product cubic spline was formulated with $\phi_1(x) = 1$, $\phi_2(x) = k_1(x_{\langle 1\rangle})$, $\phi_3(x) = k_1(x_{\langle 2\rangle})$, $\phi_4(x) = k_1(x_{\langle 3\rangle})$, and $\phi_5(x) = k_1(x_{\langle 1\rangle})k_1(x_{\langle 3\rangle})$, where $k_1(u) = u - 0.5$, and

$$\begin{aligned} R_J(x_1, x_2) = {} & \theta_1 R_c(x_{1\langle 1\rangle}, x_{2\langle 1\rangle}) + \theta_2 R_c(x_{1\langle 2\rangle}, x_{2\langle 2\rangle}) + \theta_3 R_c(x_{1\langle 3\rangle}, x_{2\langle 3\rangle}) \\ & + \theta_4 R_c(x_{1\langle 1\rangle}, x_{2\langle 1\rangle}) k_1(x_{1\langle 3\rangle}) k_1(x_{2\langle 3\rangle}) + \theta_5 k_1(x_{1\langle 1\rangle}) k_1(x_{2\langle 1\rangle}) R_c(x_{1\langle 3\rangle}, x_{2\langle 3\rangle}) \\ & + \theta_6 R_c(x_{1\langle 1\rangle}, x_{2\langle 1\rangle}) R_c(x_{1\langle 3\rangle}, x_{2\langle 3\rangle}), \end{aligned}$$

where $R_c(u_1, u_2)$ is given in (4.3). The model was fitted with $q = n = 330$ and $q = 37 \approx 10(330)^{2/9}$, respectively. The main effects of the fits are plotted in Figure 8.1. The fitting for $q = 330$ took about 1506 CPU seconds and that for $q = 37$ took about 29 CPU seconds.

8.2 Global Temperature Map

Maps of meteorological quantities constructed from records registered at weather stations are valuable tools in various applications such as climate change studies. A data set involving 690 weather

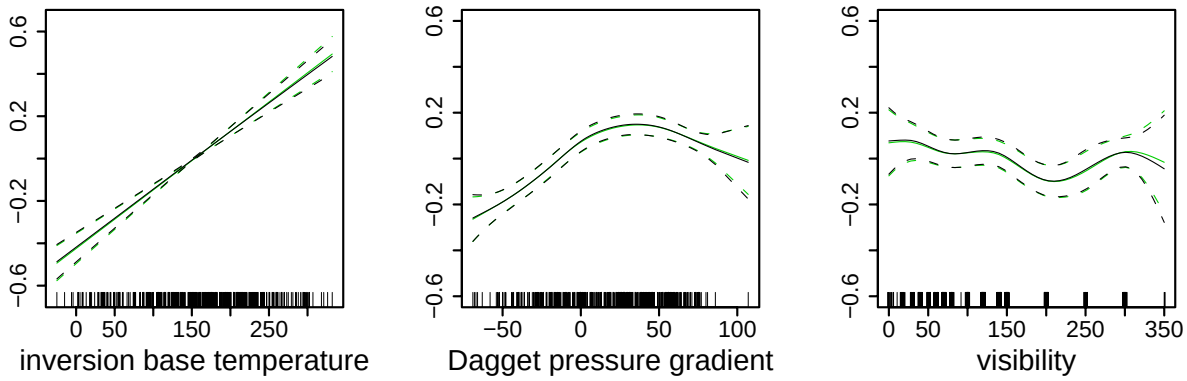


Figure 8.1: Main Effects of the Ozone Fits. Plotted are the fitted terms and 95% Bayesian confidence intervals, with $q = n$ (faded) and $q = 10n^{2/9}$. The rugs on the bottom of the plots mark the sampling points, with the visibility jittered.

stations over the globe was derived by Wang and Ke (2002, §8.2), which contained the locations of the stations (x) and the average temperatures from December 1980 to February 1981 ($Y\,^oC$). To illustrate their Splus package `assist`, Wang and Ke (2002) fitted a global temperature map to the data using the spherical spline constructed by Wahba (1981); a similar illustration based on 725 stations can be found in Luo and Wahba (1997).

For points x_1 and x_2 on the sphere, write $w = \cos\gamma(x_1, x_2)$, where $\gamma(x_1, x_2)$ is the angle between x_1 and x_2, and $W = (1 - w)/2$. The spherical spline of order 2 has $m = 1$ with $\phi(x) = 1$ and

$$R_J(x_1, x_2) = \log(1 + W^{-1/2})(12W^2 - 4W) - 12W^{3/2} + 6W + 1;$$

see Wahba (1981) for details. The formulation is apparently invariant of the coordinate system used on the sphere. With the limiting distribution $f(x)$ of x_i bounded from above and below, the eigenvalues of the corresponding $J(\eta)$ with respect to $V(\eta) = \int_{\mathcal{S}} f(x)\eta^2(x)dx$ grow at a rate $\rho_\nu = O(\nu^4)$, where $\mathcal{S}$ denotes the sphere, so $r = 4$; see §2.2 for the notation and Wahba (1981) for technical details.

The majority of the weather stations in the data set are on or near the continents, so the distribution density $f(x)$ is not bounded from below. The distribution is actually much denser in Europe and Japan, making the upper bound also a shaky one. With the highly non-uniform distribution of x_i, the asymptotics is not even remotely plausible, so it is no surprise that the empirical formula $q = 10n^{2/9}$ does not work here.

In an effort to understand how one may achieve accurate approximation in the situation, the following experiment was conducted. First, the fit $\hat{\eta}$ with $q = n = 690$ was computed and evaluated on the sampling points x_i. For each of the subset sizes $q = 50, 100, 150, 200, 250$, twenty pairs of

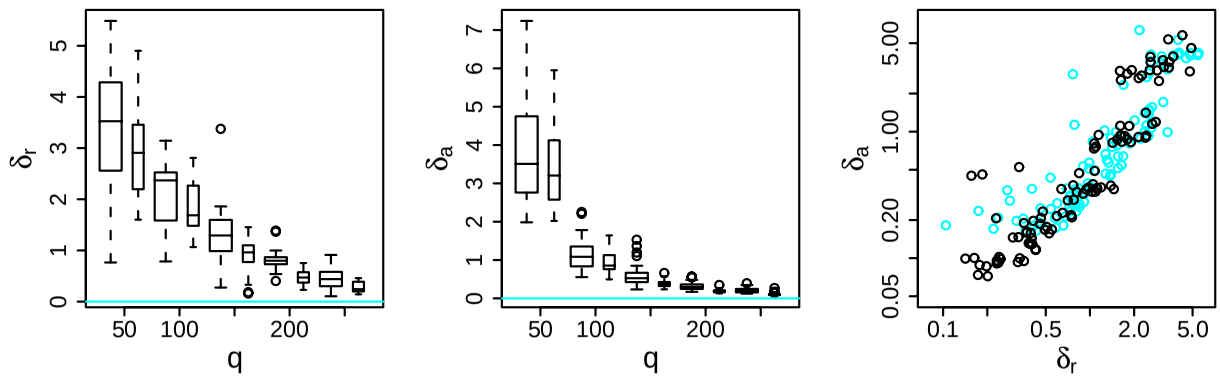


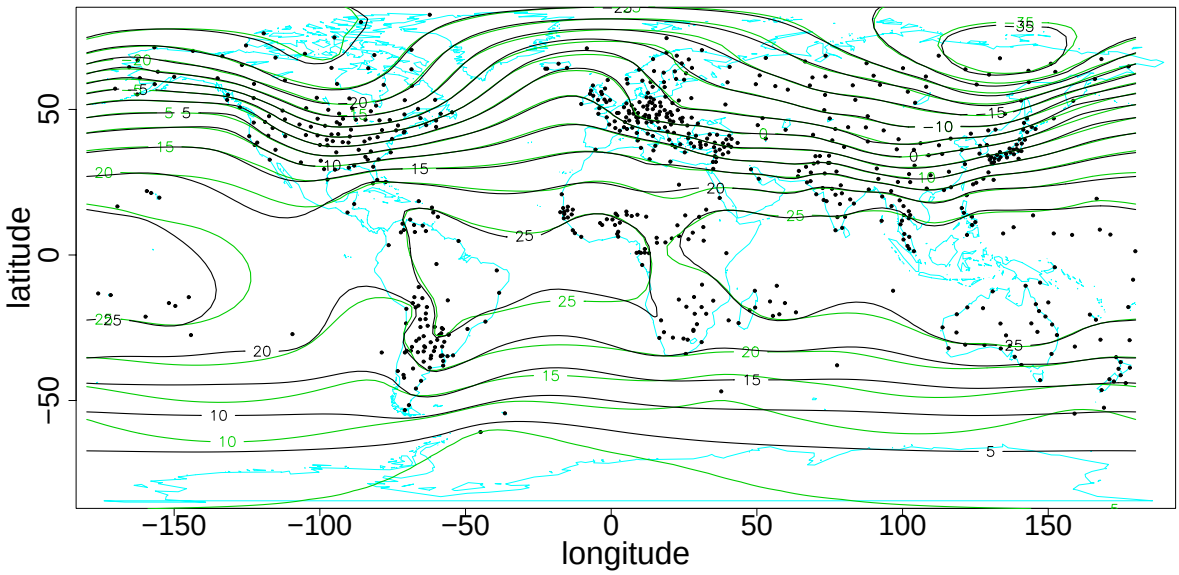
Figure 8.2: Approximation Accuracy of Spherical Spline. Left and Center: The effect of subset size q on δ_r and δ_a, for simple random placement of z_j (fatter boxes) and "space-filling" random placement (thinner boxes). Right: δ_a (pair average) versus δ_r, for simple random placement of z_j (faded) and "space-filling" random placement.

random subsets $\{z_j\} \subset \{x_i\}$ were generated, and for each pair, fits $\tilde{\eta}_1$ and $\tilde{\eta}_2$ were calculated and the quantities

$$\delta_r = \frac{\sum_{i=1}^n (\tilde{\eta}_1(x_i) - \tilde{\eta}_2(x_i))^2}{\sum_{i=1}^n (Y_i - \tilde{\eta}_1(x_i))^2 + \sum_{i=1}^n (Y_i - \tilde{\eta}_2(x_i))^2},$$
$$\delta_a = \frac{1}{n}\sum_{i=1}^n (\tilde{\eta}_j(x_i) - \hat{\eta}(x_i))^2, \quad j = 1, 2,$$

were recorded; δ_r assesses the relative discrepancy between the pair and δ_a measures the accuracy of $\tilde{\eta}$ as approximation of $\hat{\eta}$. Following a hunch that a "uniform" distribution of z_j over the covered area may yield better approximation, we also tried a pseudo "space-filling" filter in the random selection of $\{z_j\}$ by disallowing mutual distances that were less that 3 angular degrees, and repeated the above experiment with such a filter in place. The results of the experiments are summarized in Figure 8.2. It can be seen that the "space-filling" placement of z_j generally leads to better accuracy, and that δ_r is roughly monotone in δ_a. The fitting for $q = 690$ took about 384 seconds, and that for $q = 50, 100, 150, 200, 250$ took about 3, 10, 20, 33, and 50 CPU seconds, respectively.

Plotted in Figure 8.3 are the contours of the $q = 690$ fit and a $q = 200$ fit with a "space-filling" random placement of z_j; the rough appearance of the standard error is largely due to its overall flatness and the resulting fine scale. Densely sampled areas fetch smaller standard errors, whereas places like the southern Pacific Ocean, the southern Indian Ocean, and the Antarctic fare worse. The evaluation of the posterior mean from the fits took little time compared to the fitting, but the evaluation of posterior standard deviation was a major undertaking. For the $q = 200$ fit, the exaluation of standard error took a bit less than 1 CPU second per point plus a negligible



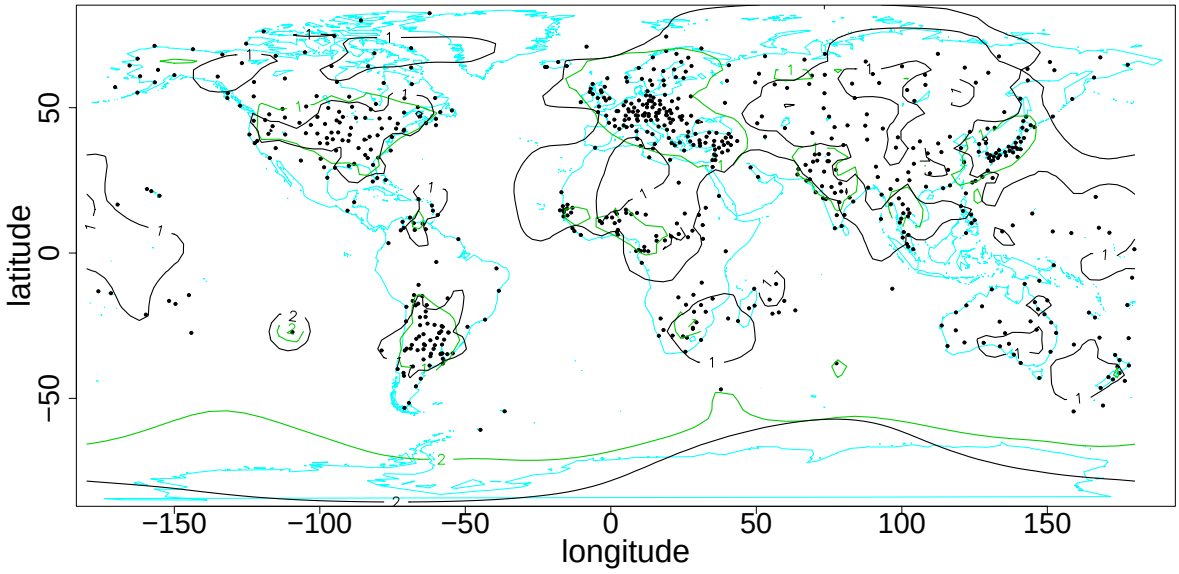


Figure 8.3: Globa Temperature Map. Top: Posterior means. Bottom: Posterior standard deviations. The fit $\hat{\eta}$ with $q = 690$ is in faded lines and the fit $\tilde{\eta}$ with $q = 200$ is in solid lines. The stations are superimposed as the dots.

overhead; for the $q = 690$ fit, it took over 12 CPU seconds per point, plus a overhead of about 30 CPU seconds. It takes a few thousands of grid points to produce a decent looking map, so the difference is practically significant.

9 Discussion

In this article, we studied the fast computation of penalized least squares regression through asymptotically efficient low-dimensional approximations. Algorithms were developed, Bayesian confidence intervals derived, empirical rules proposed, and numerical accuracies assessed. Also proposed and investigated was a simple modification of generalized cross-validation, which compared favorably against standard GCV and GML for smoothing parameter selection.

The idea of fast computation through low-dimensional approximation is an old one: it is simply a version of penalized regression splines. The quantification of the adequate dimension q through a combination of asymptotic analysis and numerical simulation is new, however; see also Gu and Wang (2002). Through more delicate placement of the "knots" z_j, it should be possible to achieve asymptotic efficiency with q smaller than what we prescribe here, but our random placement of "knots" is simple to operate and is "universally" applicable. The empirical formula $q = 10n^{2/9}$ is certainly not universally applicable, with a counter example already seen in §8.2, but the theoretical consideration and numerical experiments leading to it may serve as a model for the discovery of similar empirical formulas in specific types of application settings.

On product domains permitting ANOVA decompositions, a feature of our approach inherited from smoothing splines but unusual for regression splines is the terms in different function spaces bundled together in the basis $R_J(z_j, x) = \sum_{\beta=1}^{g} \theta_\beta R_\beta(z_j, x)$. For better or worse, we are able to accommodate multiple terms with no increase in q yet with (hopefully) little loss in flexibility, which is impossible with separate terms, especially when interactions are included. Also, because of the presence of smoothing parameters θ_β in the basis $R_J(z_j, x)$, the algorithm of Wood (2000) does not apply as the analytical derivatives of $V_\alpha(\lambda)$ and $M(\lambda)$ are not available here.

Most of the calculations reported in this article were performed in R (Ihaka and Gentleman 1996), an open-source clone of S/Splus. The code is being polished, and will replace the $O(n^3)$ numerical engine currently powering the `ssanova` function in the R package `gss` by the second author. Work is also under way for the development of fast algorithms for penalized likelihood non-Gaussian regression through asymptotically efficient approximations.

Acknowledgements

This research was supported by U.S. National Institutes of Health under Grant 1R33HL68515-01, and by Purdue Research Foundation under a research grant.

References

Breiman, L. and J. H. Friedman (1985). Estimating optimal transformations for multiple regression and correlation. *J. Amer. Statist. Assoc. 80*, 580–598 (with discussions).

Buja, A., T. Hastie, and R. Tibshirani (1989). Linear smoothers and additive models. *Ann. Statist. 17*, 453–555 (with discussions).

Craven, P. and G. Wahba (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.

Dennis, J. E. and R. B. Schnabel (1996). *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*. Philadelphia: SIAM. Corrected reprint of the 1983 original.

Golub, G. and C. Van Loan (1989). *Matrix Computations* (2nd ed.). Baltimore, MD: The Johns Hopkins University Press.

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

Gu, C. (2002). *Smoothing Spline ANOVA Models*. New York: Springer-Verlag.

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

Gu, C. and Y.-J. Kim (2001). Penalized likelihood regression: General formulation and efficient approximation. Technical report, Department of Statistics, Purdue University, West Lafayette, IN.

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

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

Gu, C. and G. Wahba (1993b). Smoothing spline ANOVA with component-wise Bayesian "confidence intervals". *J. Comput. Graph. Statist. 2*, 97–117.

Gu, C. and J. Wang (2002). Penalized likelihood density estimation: Direct cross-validation and scalable approximation. Technical report, Department of Statistics, Purdue University, West Lafayette, IN.

Ihaka, R. and R. Gentleman (1996). R: A language for data analysis and graphics. *J. Comput. Graph. Statist. 5*, 299–314.

Kimeldorf, G. and G. Wahba (1970a). A correspondence between Bayesian estimation of stochastic processes and smoothing by splines. *Ann. Math. Statist. 41*, 495–502.

Kimeldorf, G. and G. Wahba (1970b). Spline functions and stochastic processes. *Sankhya Ser. A 32*, 173–180.

Kimeldorf, G. and G. Wahba (1971). Some results on Tchebycheffian spline functions. *J. Math. Anal. Applic. 33*, 82–85.

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

Luo, Z. and G. Wahba (1997). Hybrid adaptive splines. *J. Amer. Statist. Assoc. 92*, 107–116.

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

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

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

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*, Volume 59 of CBMS-NSF Regional Conference Series in Applied Mathematics. Philadelphia: SIAM.

Wang, Y. and C. Ke (2002). *ASSIST: A Suite of S-plus Functions Implementing Spline Smoothing Techniques*.

Wood, S. N. (2000). Modelling and smoothing parameter estimation with multiple quadratic penalties. *J. Roy. Statist. Soc. Ser. B 62*, 413–428.