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

TECHNICAL REPORT NO. 913

January 27, 1994

Bootstrap Confidence Intervals for Smoothing Splines and their Comparison to Bayesian 'Confidence Intervals' [1]

by

Yuedong Wang and Grace Wahba

[1]Supported by the National Science Foundation under Grant DMS-9121003 and the National Eye Institute under Grant R01 EY09946. e-mail `wahba@stat.wisc.edu`, `wang@stat.wisc.edu`

Bootstrap Confidence Intervals for Smoothing Splines and Their Comparison to Bayesian Confidence Intervals

Yuedong Wang and Grace Wahba [†]

January 27, 1994

University of Wisconsin-Madison
USA

Abstract

We construct bootstrap confidence intervals for smoothing spline and smoothing spline ANOVA estimates based on Gaussian data, and penalized likelihood smoothing spline estimates based on data from exponential families. Several variations of bootstrap confidence intervals are considered and compared. We find that the commonly used bootstrap percentile intervals are inferior to the T intervals and to intervals based on bootstrap estimation of mean squared errors. The best variations of the bootstrap confidence intervals behave similar to the well known Bayesian confidence intervals. These bootstrap confidence intervals have an average coverage probability across the function being estimated, as opposed to a pointwise property.

Keywords: BAYESIAN CONFIDENCE INTERVALS, BOOTSTRAP CONFIDENCE INTERVALS, PENALIZED LOG LIKELIHOOD ESTIMATES, SMOOTHING SPLINES, SMOOTHING SPLINE ANOVA'S.

1 Introduction

Smoothing splines and smoothing spline ANOVAs (SS ANOVAs) have been used successfully in a broad range of applications requiring flexible nonparametric regression models. It is highly desirable to have interpretable confidence intervals for these estimates for various reasons, for example, to decide whether a spline estimate is more suitable than a particular parametric regression. A parametric regression model may be considered not suitable if a large portion of its estimate is outside of the confidence intervals of a smoothing spline estimate.

One way to construct confidence intervals for nonparametric estimates is via the bootstrap. Dikta (1990) constructs pointwise bootstrap confidence intervals for a smoothed nearest neighbor estimate. Hardle and Bowman (1988) and Hardle and Marron (1991) use bootstrap to construct pointwise and simultaneous confidence intervals for a kernel estimate. Kooperberg, Stone and Truong (1993) construct bootstrap confidence intervals for a regression spline estimate of a hazard function. Wahba (1990) suggests the use of an estimate-based bootstrap to construct confidence intervals for a smoothing spline. Meier and Nychka (1993) used bootstrap confidence intervals for spline estimates to obtain the properties of a statistic to test the equality of two rate equations. As far as we know, direct comparisons between smoothing spline bootstrap confidence intervals and the well known Bayesian confidence intervals have not yet been done. In this paper, we provide some evidence that the bootstrap confidence intervals for smoothing splines that we construct have

[†]Address for correspondence: Department of Statistics, University of Wisconsin-Madison, 1210 West Dayton St., Madison, Wisconsin 53706, USA. e-mail wang@stat.wisc.edu, wahba@stat.wisc.edu

an average coverage probability across the function being estimated (as opposed to a pointwise property), similar to the Bayesian confidence intervals. We also propose bootstrap confidence intervals for SS ANOVAs and spline estimates for data from exponential families, which appears to be new.

The so-called Bayesian confidence intervals were proposed by Wahba (1983) for a smoothing spline, where their frequentist properties were discussed. Gu and Wahba (1993b) extended Bayesian confidence intervals to the components of an SS ANOVA, and Gu (1992b) extended them to penalized log likelihood smoothing spline estimates for data from exponential families. Wang (1994) extends the Bayesian confidence intervals to a penalized log likelihood SS ANOVA estimate for data from exponential families. It is well established that these Bayesian confidence intervals have the average coverage probability property, as opposed to a pointwise property (see Nychka (1988)). They have performed well in a number of simulations. See also Abramovich and Steinberg (1993), who generalize the Bayesian intervals to the case of a variable smoothing parameter. In this report, we compare the performance of bootstrap confidence intervals with Bayesian confidence intervals via simulations.

In Section 2, we review Bayesian confidence intervals and bootstrap confidence intervals for smoothing splines with Gaussian data. We show evidence supporting the average coverage probability property of bootstrap confidence intervals. Six variations of bootstrap confidence intervals are considered. We run several simulations to find the best bootstrap confidence intervals and compare them to Bayesian confidence intervals. The parallel comparisons for SS ANOVA are given in Section 3. In Section 4, we run a simulation to compare the performance of Bayesian confidence intervals and bootstrap confidence intervals for a penalized log likelihood smoothing spline estimate based on binary data. We have found that the best variations of the bootstrap intervals behave similar to the Bayesian intervals. Bootstrap intervals have the advantage that they are easy to explain and appear to work better than Bayesian intervals in small sample size experiments with Gaussian data. The disadvantage of the bootstrap intervals is that they are computer intensive.

2 Confidence Intervals for Smoothing Splines

2.1 Smoothing Splines

Consider the model

$$y_i = f(t_i) + \epsilon_i, \quad i = 1, \cdots, n, \quad t_i \in [0,1], \tag{2.1}$$

where $\boldsymbol{\epsilon} = (\epsilon_1, \cdots, \epsilon_n)^T \sim N(0, \sigma^2 I_{n\times n})$, σ^2 unknown and $f \in W_m$ where $W_m = \{f : f, f', \cdots, f^{(m-1)}$ absolutely continuous, $\int_0^1 (f^{(m)})^2 dt < \infty\}$. The smoothing spline $\hat{f}_\lambda$ is the minimizer of

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - f(t_i))^2 + \lambda \int_0^1 (f^{(m)}(t))^2 dt \tag{2.2}$$

over $f \in W_m$. The smoothing parameter λ controls the trade off between the goodness of fit and the roughness. When λ is fixed, $\hat{\boldsymbol{f}}_\lambda = (\hat{f}_\lambda(t_1), \cdots, \hat{f}_\lambda(t_n))^T$ is a linear function of $\boldsymbol{y} = (y_1, \cdots, y_n)^T$: $\hat{\boldsymbol{f}}_\lambda = A(\lambda)\boldsymbol{y}$, where $A(\lambda)$ is the so called "hat" or influence matrix. λ can be selected by a data-based procedure such as generalized cross validation (GCV) or unbiased risk estimation (UBR) (see Wahba, 1990). The GCV estimate of λ is the minimizer of the GCV function

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

The UBR estimate of λ is the minimizer of

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

assuming that σ^2 is known. Denote $\hat{\lambda}$ as an estimate of λ by one of these procedures. Denote $\hat{f}_{\hat{\lambda}}$ as the solution of (2) with $\lambda = \hat{\lambda}$.

2.2 Bayesian Confidence Intervals

Suppose that f in (2.1) is a sample path from the Gaussian process

$$f(t) = \sum_{j=1}^{m} \frac{\tau_j t^{j-1}}{(j-1)!} + b^{\frac{1}{2}} \int_0^t \frac{(t-s)^{m-1}}{(m-1)!} dW(s),$$

where $W(\cdot)$ is a standard Weiner process and $\boldsymbol{\tau} = (\tau_1, \cdots, \tau_m)^T \sim N(0, \xi I_{m\times m})$. Wahba (1978) showed that with $b = \frac{\sigma^2}{n\lambda}$,

$$\hat{f}_\lambda(t) = \lim_{\xi\to\infty} \mathrm{E}(f(t)|\boldsymbol{y}), \qquad \sigma^2 A(\lambda) = \lim_{\xi\to\infty} \mathrm{Cov}(\boldsymbol{f}|\boldsymbol{y}), \tag{2.3}$$

where $\boldsymbol{f} = (f(t_1), \cdots, f(t_n))^T$.

This connection between a smoothing spline and the posterior mean and variance led Wahba (1983) to propose the $(1-\alpha)100\%$ Bayesian confidence intervals for $\{f(t_i)\}_{i=1,n}$ as

$$\hat{f}_{\hat{\lambda}}(t_i) \pm z_{\frac{\alpha}{2}} \sqrt{\hat{\sigma}^2 [A(\hat{\lambda})]_{ii}}, \qquad i = 1, \cdots, n, \tag{2.4}$$

where $\hat{\sigma}^2 = ||I - A(\hat{\lambda})\boldsymbol{y}||^2 / tr(I - A(\hat{\lambda}))$ is an estimate of σ^2. Both simulations (Wahba, 1983) and theory (Nychka (1988), Nychka (1990)) suggest that these Bayesian confidence intervals have good frequentist properties for $f \in W_m$ provided $\hat{\lambda}$ is a good estimate of the λ which minimizes the predictive mean square error. The intervals must be interpreted "across the function", rather than pointwise. More precisely, Nychka defines the *average coverage probability* (ACP) as $\frac{1}{n}\sum_{i=1}^n P(f(t_i) \in C(\alpha, t_i))$ for some $(1-\alpha)100\%$ confidence intervals $\{C(\alpha, t_i)\}_{i=1,n}$. Rather than consider a confidence interval for $f(\tau)$, where $f(\cdot)$ is the realization of a stochastic process and τ is fixed, he considers confidence intervals for $f(\tau_n)$, where f is now a fixed function in W_m and τ_n is a point randomly selected from $\{t_i\}_{i=1,n}$. Then ACP= $P(f(\tau_n) \in C(\alpha, \tau_n))$. Denote $T_n(\lambda) = \frac{1}{n}\sum_{i=1}^n (\hat{f}_\lambda(t_i) - f(t_i))^2$ as the average squared error. Let λ^0 be the value that minimizes $ET_n(\lambda)$. Let $b(t) = E\hat{f}_{\lambda^0}(t) - f(t)$ and $v(t) = \hat{f}_{\lambda^0}(t) - E\hat{f}_{\lambda^0}(t)$. They are the bias term and the variation term of the estimate $\hat{f}_{\lambda^0}(t)$ respectively. Set $b = b(\tau_n)$, $v = v(\tau_n)$ and $\mathcal{U} = (b+v)/(ET_n(\lambda^0))^{1/2}$. Nychka argures that the distribution of $\mathcal{U}$ is close to a standard normal distribution since it is the convolution of two random variables, one normal and the other with a variance that is small relative to the normal component.

We only consider $\{t_i\}_{i=1,n}$ as fixed design points. Let E_n be the empirical distribution for $\{t_i\}_{i=1,n}$. Assume $\sup_{u\in[0,1]} |E_n - u| = O(\frac{1}{n})$.

Assumption 1: $\hat{\lambda}$ is the minimizer of GCV function $V(\lambda)$ over the interval $[\lambda_n, \infty)$, where $\lambda_n \sim n^{-4m/5}$.

Assumption 2: f is such that for some $\gamma > 0$, $\frac{1}{n}\sum_{i=1}^n (Ef_\lambda(t_i) - f(t_i))^2 = \gamma\lambda^2(1 + o(1))$ uniformly for $\lambda \in [\lambda_n, \infty)$.

Lemma 1 (Nychka) *Suppose $\hat{T}_n$ is a consistent estimator of $ET_n(\lambda^0)$. Let $C(\alpha, t) = \hat{f}_{\hat{\lambda}}(t) \pm z_{\frac{\alpha}{2}}\sqrt{\hat{T}_n}$. Then under Assumptions 1 and 2,*

$$\frac{1}{n}\sum_{i=1}^{n} P(f(t_i) \in C(\alpha, t_i)) - P(|\mathcal{U}| \leq z_{\frac{\alpha}{2}}) \longrightarrow 0$$

uniformly in α as $n \longrightarrow \infty$.

Nychka also proves (for $m = 2$) that

$$\frac{\hat{\sigma}^2 trA(\hat{\lambda})/n}{ET_n(\lambda^0)} \xrightarrow{p} \frac{32}{27} \ as \ n \longrightarrow \infty. \tag{2.5}$$

So for large sample size, confidence intervals with $\hat{T}_n$ replaced by $\hat{\sigma}^2 trA(\hat{\lambda})/n$ should have ACP close to or a little bit over the nominal coverage. Bayesian confidence intervals actually use the individual diagonal elements of $A(\hat{\lambda})$ instead of the average $trA(\hat{\lambda})/n$. It is reasonable since most of the diagonal elements are essentially the same (see Nychka (1988)).

2.3 Bootstrap Confidence Intervals

The following bootstrap method is described in Wahba(1990). Suppose $\{t_i\}_{i=1,n}$ are fixed design points. Let $\hat{f}_{\hat{\lambda}}$ and $\hat{\sigma}^2$ be the estimates of f and σ^2 from the data. Pretending that $\hat{f}_{\hat{\lambda}}$ is the "true" f, generate a bootstrap sample

$$y_i^* = \hat{f}_{\hat{\lambda}}(t_i) + \epsilon_i^*, \quad i = 1, \cdots, n,$$

where $\boldsymbol{\epsilon}^* = (\epsilon_1^*, \cdots, \epsilon_n^*)^T \sim N(0, \hat{\sigma}^2 I_{n\times n})$. Then find the smoothing spline estimate $\hat{f}^*_{\hat{\lambda}^*}$ based on the bootstrap sample. Denote $f^*(t_i)$ as the random variable of bootstrap fit at t_i. Repeat this process B times. So at each point t_i, we have B bootstrap estimates of $\hat{f}_{\hat{\lambda}}(t_i)$. They are B realizations of $f^*(t_i)$. For each fixed t_i, we will use six methods to construct a bootstrap confidence interval for $f(t_i)$:

(A) Percentile-t interval (denoted by T-I). Similar to a Student's t statistic, consider $D_i = (\hat{f}_{\hat{\lambda}}(t_i) - f(t_i))/s_i$, where s_i is an appropriate scale parameter. It is called a pivotal since it is independent of the nuisance parameter σ in certain parametric models. We expect it to reduce the dependence on σ in our case. Denote D_i^* as the bootstrap estimate of D_i, that is, $D_i^* = (f^*_{\hat{\lambda}^*}(t_i) - \hat{f}_{\hat{\lambda}}(t_i))/s_i^*$. Let $x_{\frac{\alpha}{2}}$, $x_{1-\frac{\alpha}{2}}$ be the lower and upper $\alpha/2$ points of the empirical distribution of D_i^*. The $(1-\alpha)100\%$ T bootstrap confidence interval is $(\hat{f}_{\hat{\lambda}}(t_i) - x_{1-\frac{\alpha}{2}}s_i, \hat{f}_{\hat{\lambda}}(t_i) - x_{\frac{\alpha}{2}}s_i)$. The standard deviation of $\hat{f}_{\hat{\lambda}}(t_i) - f(t_i)$ generally equals a constant times σ. So setting $s_i = \hat{\sigma}$, we have the T-I bootstrap confidence intervals.

(B) Another percentile-t interval (denoted by T-II). From the Bayesian model, the exact standard deviation of $\hat{f}_{\hat{\lambda}}(t_i) - f(t_i)$ equals to $\sqrt{\hat{\sigma}^2[A(\hat{\lambda})]_{ii}}$. Setting $s_i = \sqrt{\hat{\sigma}^2[A(\hat{\lambda})]_{ii}}$ in (A), we have the T-II bootstrap confidence intervals.

(C) Normal interval (denoted by Nor). Let $T_i = (\hat{f}_{\hat{\lambda}}(t_i) - f(t_i))^2$ be the squared error at t_i. Denote T_i^* as the bootstrap estimate of T_i. The $(1-\alpha)100\%$ normal bootstrap confidence interval is $(\hat{f}_{\hat{\lambda}}(t_i) + z_{\frac{\alpha}{2}}T_i^*, \hat{f}_{\hat{\lambda}}(t_i) + z_{1-\frac{\alpha}{2}}T_i^*)$, where $z_{\frac{\alpha}{2}}$, $z_{1-\frac{\alpha}{2}}$ are the $\alpha/2$ and $1-\alpha/2$ percentiles of the standard normal distribution. We use the individual squared error estimate instead of average squared error because we want the length of a confidence interval to depend on the distribution of the design points. Generally, the confidence intervals are narrower in a neighborhood with more data.

(D) Percentile interval (denoted by Per) (Efron (1982)). Let $f_L^*(t_i)$, $f_U^*(t_i)$ be the lower and upper $\alpha/2$ points of the empirical distribution of $f^*(t_i)$. The $(1-\alpha)100\%$ confidence interval is $(f_L^*(t_i), f_U^*(t_i))$.

(E) Pivotal method (denoted by Piv) (Efron (1981)). Let $x_{\frac{\alpha}{2}}$ and $x_{1-\frac{\alpha}{2}}$ be the lower and upper $\alpha/2$ points of the empirical distribution of $f^*(t_i) - \hat{f}_{\hat{\lambda}}(t_i)$. Then $x_{\frac{\alpha}{2}} = f_L^*(t_i) - \hat{f}_{\hat{\lambda}}(t_i)$, $x_{1-\frac{\alpha}{2}} = f_U^*(t_i) - \hat{f}_{\hat{\lambda}}(t_i)$. If the empirical distribution of $f^*(t_i) - \hat{f}_{\hat{\lambda}}(t_i)$ approximates the distribution of $\hat{f}_{\hat{\lambda}}(t_i) - f(t_i)$, then $P(x_{\frac{\alpha}{2}} < \hat{f}_{\hat{\lambda}}(t_i) - f(t_i) < x_{1-\frac{\alpha}{2}}) \approx 1-\alpha$. The $(1-\alpha)\%$ pivotal confidence interval for $f(t_i)$ is $(2\hat{f}_{\hat{\lambda}}(t_i) - f_U^*(t_i), 2\hat{f}_{\hat{\lambda}}(t_i) - f_L^*(t_i))$.

(F) Bias corrected percentile interval (denoted by BC) (Efron (1982)). Suppose there exists an increasing function h such that

$$\eta = h(f(t_i)), \quad \hat{\eta} = h(f_\lambda(t_i)), \quad \hat{\eta}^* = h(f^*(t_i)), \quad \hat{\eta} - \eta \sim N(-a, 1), \quad \hat{\eta}^* - \hat{\eta} \sim N(-a, 1),$$

for some constant a. The $(1-\alpha)100\%$ biased corrected confidence interval is $(G_n^{*-1}[\phi(2a + z_{\frac{\alpha}{2}})], G_n^{*-1}[\phi(2a + z_{1-\frac{\alpha}{2}})])$, where G_n^* is the empirical distribution of $f^*(t_i)$, ϕ is the density function of a standard normal distribution, $a = \phi^{-1}(G^*(f_\lambda(t_i)))$.

To study the properties of the bootstrap confidence intervals, rewrite the expected average squared error as

$$E_f T_n(\lambda^0) = \frac{1}{n}\sum_{i=1}^n E_f(\hat{f}_{\lambda^0}(t_i|f) - f(t_i))^2, \tag{2.6}$$

where $\hat{f}_{\lambda^0}(\cdot|f)$ is the smoothing spline estimate of f when f is the true function and the smoothing parameter is equal to λ^0. The bootstrap method replaces f by $\hat{f}_{\hat{\lambda}}$:

$$E_f T_n(\lambda^0) \approx \frac{1}{n}\sum_{i=1}^n E_{\hat{f}_{\hat{\lambda}}}(\hat{f}^*_{\lambda^0}(t_i|\hat{f}_{\hat{\lambda}}) - \hat{f}_{\hat{\lambda}}(t_i))^2, \tag{2.7}$$

where $\hat{f}^*_{\lambda^0}(\cdot|\hat{f}_{\hat{\lambda}})$ is the smoothing spline estimate of $\hat{f}_{\hat{\lambda}}$ when $\hat{f}_{\hat{\lambda}}$ is the true function. We can estimate the right hand side of (2.7) since the true function $\hat{f}_{\hat{\lambda}}$ and true variance $\hat{\sigma}^2$ are known. One way is to generate a bootstrap sample from this true model and fit a smoothing spline with λ^0 chosen by GCV or UBR. Repeat this procedure B times. The average squared error of these B repetitions could be used as an estimate of the expected average squared error. The following theorem proves that for fixed sample size such an estimation is consistent for the right hand side of (2.7). For simplicity of notation, we use f instead of $\hat{f}_{\hat{\lambda}}$ as the true function, σ^2 instead of $\hat{\sigma}^2$ as the true variance.

Theorem 2 *Suppose the true function f and variance σ^2 are known. Denote B bootstrap samples as*

$$\boldsymbol{y}^j = \boldsymbol{f} + \boldsymbol{\epsilon}^j, \quad j = 1, \cdots, B.$$

Let $\hat{\boldsymbol{f}}^j_{\lambda^0}$ be the smoothing spline fit for the jth bootstrap sample. Then for fixed n,

$$\frac{1}{B}\sum_{j=1}^B \frac{1}{n}\sum_{i=1}^n (\hat{f}^j_{\lambda^0}(t_i) - f(t_i))^2 \xrightarrow{a.s.} E\frac{1}{n}\sum_{i=1}^n (\hat{f}_{\lambda^0}(t_i) - f(t_i))^2, \quad B \to \infty.$$

[Proof] Write

$$\hat{\boldsymbol{f}}_{\lambda^0} - \boldsymbol{f} = (A(\lambda^0) - I)\boldsymbol{f} + A(\lambda^0)\boldsymbol{\epsilon}.$$

Then

$$\begin{aligned}
& \frac{1}{n}\sum_{i=1}^{n}(\hat{f}_{\lambda^0}(t_i) - f(t_i))^2 \\
= & \frac{1}{n}(\hat{\boldsymbol{f}}_{\lambda^0} - \boldsymbol{f})^T(\hat{\boldsymbol{f}}_{\lambda^0} - \boldsymbol{f}) \\
= & \frac{1}{n}[\boldsymbol{f}^T(A(\lambda^0) - I)^2\boldsymbol{f} + 2\boldsymbol{f}^T(A(\lambda^0) - I)A(\lambda^0)\boldsymbol{\epsilon} + \boldsymbol{\epsilon}^T A^2(\lambda^0)\boldsymbol{\epsilon}].
\end{aligned}$$

So

$$E\frac{1}{n}\sum_{i=1}^{n}(\hat{f}_{\lambda^0}(t_i) - f(t_i))^2 = \frac{1}{n}[\boldsymbol{f}^T(A(\lambda^0) - I)^2\boldsymbol{f} + \sigma^2 trA^2(\lambda^0)].$$

Similarly, we have

$$\begin{aligned}
& \frac{1}{B}\sum_{j=1}^{B}\frac{1}{n}\sum_{i=1}^{n}(\hat{f}^j_{\lambda^0}(t_i) - f(t_i))^2 \\
= & \frac{1}{B}\sum_{j=1}^{B}\frac{1}{n}[\boldsymbol{f}^T(A(\lambda^0) - I)^2\boldsymbol{f} + 2\boldsymbol{f}^T(A(\lambda^0) - I)A(\lambda^0)\boldsymbol{\epsilon}^j + (\boldsymbol{\epsilon}^j)^T A^2(\lambda^0)\boldsymbol{\epsilon}^j] \\
= & \frac{1}{n}\boldsymbol{f}^T(A(\lambda^0) - I)^2\boldsymbol{f} + 2\boldsymbol{f}^T(A(\lambda^0) - I)A(\lambda^0)\frac{1}{B}\sum_{j=1}^{B}\boldsymbol{\epsilon}^j + \frac{1}{B}\sum_{j=1}^{B}(\boldsymbol{\epsilon}^j)^T A^2(\lambda^0)\boldsymbol{\epsilon}^j \\
\xrightarrow{a.s.} & \frac{1}{n}\boldsymbol{f}^T(A(\lambda^0) - I)^2\boldsymbol{f} + \sigma^2 trA^2(\lambda^0), \qquad B \to \infty.
\end{aligned}$$

So the bootstrap method tries to get an estimate $\hat{T}_n^*$ of $ET_n(\lambda^0)$ directly. From Lemma 1, the bootstrap confidence intervals $C(\alpha, t) = \hat{f}_{\hat{\lambda}}(t) \pm z_{\frac{\alpha}{2}}\sqrt{\hat{T}_n^*}$ should have the ACP property rather than pointwise coverage. The normal intervals use individual squared error estimates at each data point instead of $\hat{T}_n^*$. They should have behave similar to the intervals using $\hat{T}_n^*$ when the pointwise squared errors are not too different from each other. So we expect the normal bootstrap confidence intervals to have the ACP Property, rather than a pointwise property. Actually all these bootstrap confidence intervals will be seen to have the ACP property from our simulations next.

As pointed out by many authors, the bootstrap bias $E(\hat{f}^*_{\hat{\lambda}^*}(t) - \hat{f}_{\hat{\lambda}}(t)|\hat{f}_{\hat{\lambda}})$ generally underestimates the true bias $E(\hat{f}_{\hat{\lambda}}(t) - f(t)|f)$, paticularly at bump points. Hall (1990) suggests using a bootstrap resample of smaller size (say n_1) than the original sample for kernel estimates. He shows that for second-order kernels, the optimal choice of n_1 is of order $n^{1/2}$. It is hard to get a good estimate of n_1 in practice. Furthermore, a bootstrap sample of size n_1 may give a very bad smoothing spline estimate. Dikta (1990) and Hardle and Marron (1991) suggest using an undersmoothed estimate to generate the bootstrap samples. They prove that after the right scaling, for a kernel estimate $\hat{f}_{\hat{\lambda}}$ with $\hat{\lambda}$ as the optimal bandwidth, $\hat{f}^*_{\hat{\lambda}^*}(t) - \hat{f}_{\hat{\lambda}_1}(t)$ and $\hat{f}_{\hat{\lambda}}(t) - f(t)$ have the same limiting distributions as $n \to \infty$, if $\hat{\lambda}_1$ tends to zero at a rate slower than $\hat{\lambda}$. Again, it is difficult to get an estimate of $\hat{\lambda}_1$ in practice. The optimal $\hat{\lambda}_1$ depends on some order of derivative of f. Also, the performance for finite samples may not be satisfactory, which is shown in their simulations. Here we do not intent to construct pointwise confidence intervals. Instead, we only need a decent estimate of $ET(\lambda^0)$. Without trying to estimate the bias, the bootstrap estimates of mean squared error proved satisfactory in our simulations.

2.4 Simulations

In this section, we use some simulations to
(1) study the performance of 6 kinds of bootstrap confidence intervals and find out which are better;
(2) show the ACP property of bootstrap confidence intervals;
(3) compare the performance of bootstrap confidence intervals with the Bayesian confidence intervals.

The experimental design is the same as in Wahba (1983). Three functions are used:

$$\begin{aligned}
&\text{Case 1} \qquad f(t) = \frac{1}{3}\beta_{10,5}(t) + \frac{1}{3}\beta_{7,7}(t) + \frac{1}{3}\beta_{5,10}(t),\\
&\text{Case 2} \qquad f(t) = \frac{6}{10}\beta_{30,17}(t) + \frac{4}{10}\beta_{3,11}(t),\\
&\text{Case 3} \qquad f(t) = \frac{1}{3}\beta_{20,5}(t) + \frac{1}{3}\beta_{12,12}(t) + \frac{1}{3}\beta_{7,30}(t),
\end{aligned}$$

where $\beta_{p,q}(t) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)}t^{p-1}(1-t)^{q-1},\ 0 \le t \le 1.$

	n = 128			n = 64			n = 32		
Case	1	2	3	1	2	3	1	2	3
$\sigma = .0125$	0	0	0	0	4	4	8	70	100
$\sigma = .025$	0	0	0	0	3	4	7	24	57
$\sigma = .05$	0	0	0	0	3	4	7	12	22
$\sigma = .1$	0	0	0	0	0	2	7	8	11
$\sigma = .2$	0	0	0	0	0	0	7	7	8

Table 2.1: Number of replications out of 100 total that have ratio $\hat{\sigma}/\sigma < 0.001$.

Case 1, Case 2 and Case 3 have 1, 2 and 3 bumps respectively (see plots next). They reflect an increasingly complex 'truth'. The experiment consists of $3 \times 3 \times 5 = 45$ combinations of Case 1, 2, 3, n=32, 64, 128 and σ=0.0125, 0.025, 0.05, 0.1 and 0.2. In all cases, $t_i = i/n$. Data are generated for 100 replications of each of these 45 combinations. In all simulations in this report, random numbers are generated using the Fortran routines uni and rnor of the Core Mathematics Library (Cmlib) from the National Bureau of Standards. Spline fits are calculated using RKPACK (see Gu (1989)). Percentiles of a standard normal distribution are calculated using the CDFLIB, developed by B. W. Brown and J. Lovato and available from `statlib`.

We use GCV to select $\hat{\lambda}$ in all simulations in this section. It has been noted that for a small sample size, there is a positive probability that GCV selects $\hat{\lambda} = 0$, especially for small σ^2. In practice, if only σ is known to within a few orders of magnitude, these extreme cases can be readily identified. The numbers of replications out of 100 simulation replications which have ratio $\hat{\sigma}/\sigma < 0.001$ are listed in Table 2.1. Examination of the simulation output reveals that we will have the same numbers if we count cases with ratios smaller than 0.1, that is, there are no cases with ratio between 0.1 and 0.001. The number decreases as sample size increases, σ increases or the number of bumps decreases. All these cases have their $\hat{\lambda}$ smaller than -14 in $\log_{10}$ scale, while others have $\hat{\lambda}$ between -9 to -4 in $\log_{10}$ scale. We do not impose any limitation on the range of λ since we do not want to assume any specific prior knowledge. Instead, we can easily identify these "bad" (interpolation) cases if we know σ within 3 orders of magnitude, which is generally

	n = 128			n = 64			n = 32		
Case	1	2	3	1	2	3	1	2	3
σ = .0125									
T-I	.935	.925	.932	.949	.956	.956	.955	.832	—
T-II	.925	.918	.923	.933	.945	.962	.948	.829	—
Nor	.958	.946	.950	.950	.923	.947	.890	.664	—
Per	.940	.927	.930	.934	.914	.919	.888	.681	—
Piv	.924	.905	.913	.911	.887	.915	.857	.645	—
BC	.919	.897	.905	.898	.878	.878	.849	.645	—
σ = .025									
T-I	.938	.927	.933	.957	.951	.954	.960	.886	.773
T-II	.929	.919	.925	.938	.939	.938	.945	.878	.768
Nor	.960	.949	.952	.953	.937	.932	.911	.765	.634
Per	.940	.924	.931	.940	.924	.924	.907	.774	.640
Piv	.930	.910	.916	.923	.904	.898	.878	.728	.615
BC	.923	.903	.909	.914	.896	.893	.865	.727	.606
σ = .05									
T-I	.939	.929	.932	.955	.950	.948	.962	.940	.907
T-II	.928	.920	.924	.936	.936	.935	.945	.933	.902
Nor	.957	.949	.954	.956	.943	.938	.922	.856	.809
Per	.940	.923	.930	.939	.926	.924	.907	.857	.818
Piv	.932	.914	.918	.923	.905	.907	.893	.814	.766
BC	.923	.908	.909	.915	.901	.901	.884	.804	.762
σ = .1									
T-I	.938	.932	.929	.955	.948	.948	.964	.958	.944
T-II	.926	.923	.923	.933	.931	.935	.948	.950	.936
Nor	.956	.952	.955	.954	.945	.939	.922	.899	.869
Per	.942	.924	.929	.940	.929	.921	.909	.893	.865
Piv	.932	.922	.919	.929	.913	.910	.900	.866	.840
BC	.922	.914	.915	.915	.900	.902	.895	.852	.827
σ = .2									
T-I	.937	.935	.930	.948	.954	.950	.958	.965	.958
T-II	.928	.926	.925	.931	.938	.937	.922	.951	.946
Nor	.959	.954	.952	.947	.950	.942	.906	.918	.904
Per	.944	.923	.925	.931	.930	.921	.896	.902	.889
Piv	.930	.926	.920	.918	.917	.916	.885	.881	.874
BC	.921	.916	.912	.906	.907	.905	.881	.864	.864

Table 2.2: Mean coverages of 95% bootstrap confidence intervals.

	n = 128			n = 64			n = 32		
Case	1	2	3	1	2	3	1	2	3
σ = .0125									
T-I	.050	.038	.038	.077	.070	.395	.123	.220	—
T-II	.052	.038	.039	.078	.072	.081	.124	.219	—
Nor	.039	.041	.040	.060	.086	.083	.150	.214	—
Per	.045	.043	.044	.065	.082	.089	.147	.214	—
Piv	.055	.050	.049	.096	.109	.084	.161	.201	—
BC	.058	.054	.054	.110	.110	.107	.160	.199	—
σ = .025									
T-I	.051	.042	.039	.046	.072	.075	.097	.215	.275
T-II	.054	.042	.041	.054	.074	.077	.104	.214	.273
Nor	.041	.039	.039	.053	.072	.078	.130	.218	.251
Per	.048	.046	.047	.056	.071	.078	.128	.220	.253
Piv	.055	.051	.051	.075	.096	.098	.143	.216	.245
BC	.060	.056	.054	.079	.099	.101	.149	.217	.245
σ = .05									
T-I	.051	.046	.044	.050	.071	.075	.087	.143	.189
T-II	.057	.047	.044	.060	.074	.077	.098	.145	.190
Nor	.044	.042	.040	.055	.066	.074	.122	.171	.193
Per	.051	.047	.049	.059	.066	.071	.125	.171	.188
Piv	.057	.056	.053	.078	.098	.098	.140	.182	.198
BC	.065	.058	.059	.082	.099	.100	.146	.178	.205
σ = .1									
T-I	.058	.051	.050	.052	.074	.075	.088	.111	.142
T-II	.062	.051	.049	.066	.077	.077	.096	.114	.142
Nor	.047	.043	.042	.058	.063	.071	.124	.137	.169
Per	.055	.048	.055	.062	.066	.069	.125	.137	.167
Piv	.062	.054	.053	.080	.094	.095	.139	.149	.170
BC	.071	.056	.054	.091	.101	.100	.139	.160	.172
σ = .2									
T-I	.064	.053	.053	.059	.048	.047	.091	.084	.099
T-II	.067	.054	.053	.072	.055	.052	.110	.091	.100
Nor	.047	.046	.046	.064	.058	.057	.133	.126	.127
Per	.059	.050	.060	.077	.060	.063	.136	.128	.125
Piv	.068	.057	.056	.088	.083	.073	.142	.146	.140
BC	.076	.062	.062	.094	.083	.080	.140	.151	.149

Table 2.3: Standard deviations of coverages of 95% bootstrap confidence intervals.

	n = 128			n = 64			n = 32		
Case	1	2	3	1	2	3	1	2	3
σ = .0125	.967	.956	.960	.955	.930	.924	.897	.669	—
	(.032)	(.040)	(.039)	(.076)	(.088)	(.091)	(.149)	(.207)	—
σ = 0.025	.966	.958	.961	.961	.944	.940	.917	.769	.640
	(.036)	(.041)	(.040)	(.050)	(.076)	(.081)	(.126)	(.218)	(.251)
σ = 0.05	.963	.959	.961	.956	.947	.943	.923	.858	.809
	(.039)	(.044)	(.041)	(.055)	(.076)	(.077)	(.113)	(.170)	(.193)
σ = 0.1	.962	.963	.963	.952	.948	.945	.920	.906	.878
	(.043)	(.039)	(.034)	(.058)	(.076)	(.076)	(.129)	(.134)	(.166)
σ = 0.2	.961	.962	.960	.938	.953	.947	.884	.919	.907
	(.048)	(.041)	(.039)	(.073)	(.059)	(.053)	(.152)	(.128)	(.130)

Table 2.4: Mean coverage and their standard deviations (inside the parentheses) of 95% Bayesian confidence intervals.

true in practice. After identifying the "bad" cases, one can refit by choosing λ in a limited range. For our simulations, we simply drop these cases since they stand for the failure of GCV instead of the confidence intervals and they are "correctable". That is, all summary statistics of coverages are based on the remaining cases. Actually, for σ as small as .0125 or .025 here, the confidence intervals are visually so close to the estimate that they are hard to see. Confidence intervals are unlikely to be very interesting in these cases. We decided to include these cases since we want to know when these confidence intervals fail.

For all bootstrap confidence intervals, B=500. Similar to the above argument, there is a positive probability that the bootstrap sample fit selects an extreme $\hat{\lambda}^* = 0$. Since we know the true function and variance when bootstrapping, we certainly can identify these "bad" cases and should drop them. So these "bad" repetitions are dropped when we calculate the bootstrap confidence intervals. This is not a limitation to bootstrap confidence intervals, but rather a subtle point one needs to be careful when constructing bootstrap confidence intervals. Our results here and in Wahba and Wang (1993) suggest that this phenomena of "bad" cases will not be noticeable in practice for n much bigger than 100.

In each case, the number of data points at which the confidence intervals cover the true values are recorded. These numbers are then divided by the corresponding sample sizes to form the coverage percentage of the intervals on the design points. Tables 2.2 and 2.3 list the mean coverages and the standard deviations of the 95% bootstrap confidence intervals. For almost all cases, T-I, T-II, Nor and Per bootstrap confidence intervals are better than Piv and BC bootstrap confidence intervals. T-I intervals work better than T-II's. Nor intervals are a little bit better than Per intervals. T-I intervals are better than Nor intervals in small sample size cases but this is reversed in large sample cases. The average coverages are much improved when the sample size increases from 32 to 64 and improved a little bit when the sample size increases from 64 to 128.

In the remainder of this section, when we mention bootstrap confidence intervals we mean either T-I or Nor bootstrap confidence intervals. To compare with the Bayesian confidence intervals, we use the same data to construct Bayesian confidence intervals. The mean coverages and their standard deviations of 95% confidence intervals are listed in Table 2.4. Comparing Tables 2.2, 2.3 and 2.4. we see that for $n = 32$, bootstrap confidence intervals have better average coverages and

	n = 128		n = 64		n = 32	
	MSE*/MSE	$\hat{\sigma}/\sigma$	MSE*/MSE	$\hat{\sigma}/\sigma$	MSE*/MSE	$\hat{\sigma}/\sigma$
$\sigma = .0125$						
Case 1	1.196(.434)	.979(.072)	1.154(0.457)	.944(.144)	1.018(.604)	.841(.208)
Case 2	.989(.288)	.964(.086)	.991(.415)	.889(.168)	.333(.245)	.502(.194)
Case 3	1.060(.284)	.954(.085)	.956(.395)	.859(.162)	—	—
$\sigma = .025$						
Case 1	1.231(.502)	.983(.071)	1.210(.502)	.964(.121)	1.109(.645)	.879(.195)
Case 2	1.047(.351)	.969(.084)	1.071(.425)	.922(.153)	.647(.237)	.632(.483)
Case 3	1.091(.316)	.963(.084)	1.042(.404)	.904(.154)	.356(.306)	.503(.231)
$\sigma = .05$						
Case 1	1.266(.587)	.985(.071)	1.238(.547)	.967(.120)	1.209(.747)	.905(.185)
Case 2	1.106(.417)	.972(.084)	1.120(.456)	.935(.149)	.897(.612)	.765(.227)
Case 3	1.121(.355)	.969(.084)	1.093(.433)	.928(.152)	.716(.478)	.704(.243)
$\sigma = .1$						
Case 1	1.301(.687)	.987(.070)	1.268(.588)	.971(.119)	1.286(.891)	.926(.184)
Case 2	1.166(.473)	.980(.071)	1.151(.482)	.945(.144)	1.066(.677)	.845(.204)
Case 3	1.152(.388)	.978(.072)	1.124(.470)	.942(.148)	.933(.549)	.816(.235)
$\sigma = .2$						
Case 1	1.326(.735)	.988(.070)	1.236(.636)	.972(.118)	1.256(1.000)	.936(.186)
Case 2	1.203(.537)	.983(.072)	1.182(.518)	.960(.127)	1.149(.715)	.886(.196)
Case 3	1.167(.437)	.982(.072)	1.158(.518)	.964(.124)	1.023(.543)	.892(.205)

Table 2.5: Mean coverage and their standard deviations (inside the parentheses) of the ratios of bootstrap MSE and true MSE and ratios of estimated σ and true σ.

smaller standard deviations than the Bayesian intervals. For $n = 64$, bootstrap confidence intervals and Bayesian confidence intervals are about the same. For $n = 128$, Bayesian confidence intervals have average coverages a little bit over the nominal value, while bootstrap confidence intervals have average coverage a little bit under the nominal value.

For each repetition of the experiment, we calculate the true MSE, bootstrap estimate of MSE (denoted by MSE*) and the estimate of σ (denoted by $\hat{\sigma}$). We then get the ratios: MSE*/MSE and $\hat{\sigma}/\sigma$. The average ratios and their standard deviations are listed in Table 2.5. Notice that $\hat{\sigma}$ underestimates σ on average, which agrees with Carter and Eagleson (1992). Thus the bootstrap samples have smaller variation than they should. This causes the average coverages of bootstrap confidence intervals to be a little bit smaller than the nominal value. On the other hand, underestimation of $\hat{\sigma}^2$ does help the performance of Bayesian confidence intervals since $\hat{\sigma}^2 trA(\hat{\lambda})/n$ overestimates $ET_n(\lambda^0)$ (in theory) by a factor of 32/27. Carter and Eagleson (1992), who studied the same examples used here, found that for these functions, a better choice for the estimation of σ^2 is $\boldsymbol{y}^T(I - A(\hat{\lambda}))^2\boldsymbol{y}/tr[I - A(\hat{\lambda})]^2$. We don't know to what extent these results concerning $\hat{\sigma}^2$ are example-dependent, but we would expect that such a choice of $\hat{\sigma}^2$ would make bootstrap confidence intervals work better relative to the Bayesian intervals in the present experiments. Notice also that even though the bootstrap bias is generally smaller than true bias, MSE* overestimates MSE on average, especially for large sample sizes. The variances of MSE*/MSE's are quite big.

We visually inspect many of the plotted intervals and pointwise coverages. They all give a similar visual impression. Therefore, we just plot some "typical" cases. In Figure 2.1, we plot both bootstrap confidence intervals and Bayesian confidence intervals when $\sigma = 0.2$ for some selected functions and sample sizes. These cases are actually the first replicates in simulations. The pointwise coverages are plotted in Figures 2.2 and 2.3. It is obvious that the pointwise coverages of bootstrap confidence intervals are similar to Bayesian confidence intervals'. That is, the pointwise coverage is smaller than the nominal value at high curvature points, particularly for Per intervals. These plots support the argument that the bootstrap confidence intervals have the ACP property.

3 Component-Wise Confidence Intervals for SS-ANOVA

3.1 SS ANOVAs

Consider the model

$$y_i = f(t_1(i), \cdots, t_d(i)) + \epsilon_i, \quad i = 1, \cdots, n, \quad t_i \in [0, 1], \tag{3.1}$$

where $\boldsymbol{\epsilon} = (\epsilon_1, \cdots, \epsilon_n)^T \sim N(0, \sigma^2 I_{n\times n})$, σ^2 unknown, and $t_k \in \mathcal{T}^{(k)}$, where $\mathcal{T}^{(k)}$ is a measurable space, $k = 1, \cdots, d$. Denote $\mathcal{T} = \mathcal{T}^{(1)} \otimes \cdots \otimes \mathcal{T}^{(d)}$. Assume that f is in some reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ (see Aronszajn (1950)). Suppose that $\mathcal{H}$ admits an ANOVA-like decomposition (see Gu and Wahba (1993a) for details):

$$\mathcal{H} = [1] \oplus \sum_k \mathcal{H}^{(k)} \oplus \sum_{k<l} [\mathcal{H}^{(k)} \otimes \mathcal{H}^{(l)}] \oplus \cdots, \tag{3.2}$$

here, [1] denotes the constant functions on $\mathcal{T}$, and $\mathcal{H}^{(k)}$ consists of functions depending only on t_k, $\mathcal{H}^{(k)} \otimes \mathcal{H}^{(l)}$ consists of functions depending on t_k and t_l, and so forth. An element f_k in $\mathcal{H}^{(k)}$ is called a main effect, an element f_{kl} in $\mathcal{H}^{(k)} \otimes \mathcal{H}^{(l)}$ is called a two factor interaction, and so on. Each $\mathcal{H}^{(k)}$ can be further decomposed: $\mathcal{H}^{(k)} = \mathcal{H}_\pi^{(k)} \oplus \mathcal{H}_s^{(k)}$, where $\mathcal{H}_\pi^{(k)}$ is finite dimensional (the "parametric" part which is not penalized; usually polynomials), and $\mathcal{H}_s^{(k)}$ (the "smooth" part) is the orthocomplement of $\mathcal{H}_\pi^{(k)}$ in $\mathcal{H}^{(k)}$. After deciding on a model (that is, choosing which terms in (3.2) to retain, replacing $\mathcal{H}^{(k)}$, $\mathcal{H}^{(l)}$ and so forth in (3.2) by their decompositions, multiplying out and regrouping, we can write the model space as $\mathcal{M} = \mathcal{H}^0 \oplus \sum_{j=1}^p \mathcal{H}^j$, where $\mathcal{H}^0$ is a finite dimensional space containing functions which are not penalized, and the $\mathcal{H}^j$ are orthogonal. The estimate f_λ of f is obtained by minimizing

$$\frac{1}{n}\sum_{i=1}^n (y_i - f(\boldsymbol{t}(i)))^2 + \lambda \sum_{j=1}^p \theta_j^{-1} \|P^j f\|^2 \tag{3.3}$$

in $\mathcal{M}$, where $\boldsymbol{t}(i) = (t_1(i), \cdots, t_d(i))$. P^j is the orthogonal projection in $\mathcal{M}$ onto $\mathcal{H}^j$. $\|P^j f\|^2$ is a quadratic roughness penalty. λ is the main smoothing parameter, the θ's are subsidiary smoothing parameters, satisfying an appropriate constraint for identifiability (see Wahba (1990), Gu and Wahba (1993b)). This minimization also gives component-wise estimates such as main effects and two factor interactions. We would like to to construct confidence intervals not only for the overall function f, but also for all components in the model. These confidence intervals may, for example, be used as an aid in eliminating unimportant components, or deciding whether certain features are "real".

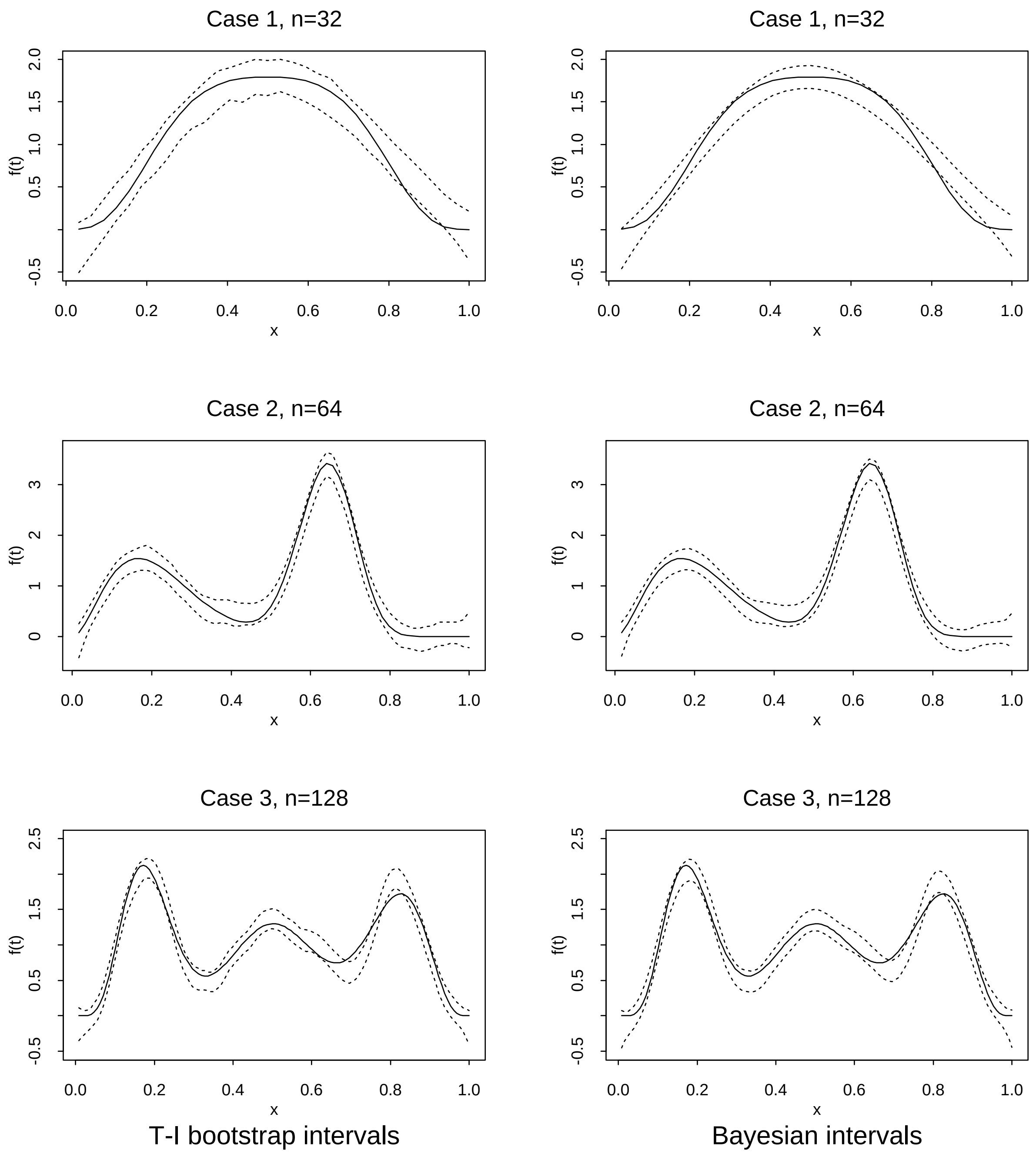

Figure 2.1: Display of the 95% confidence intervals for $\sigma = 0.2$. Solid lines: true function; dotted lines: confidence intervals. Top row: Case 1 and n=32; middle row: Case 2 and n=64; bottom row: Case 3 and n=128. Left column: T-I bootstrap intervals; right column: Bayesian intervals.

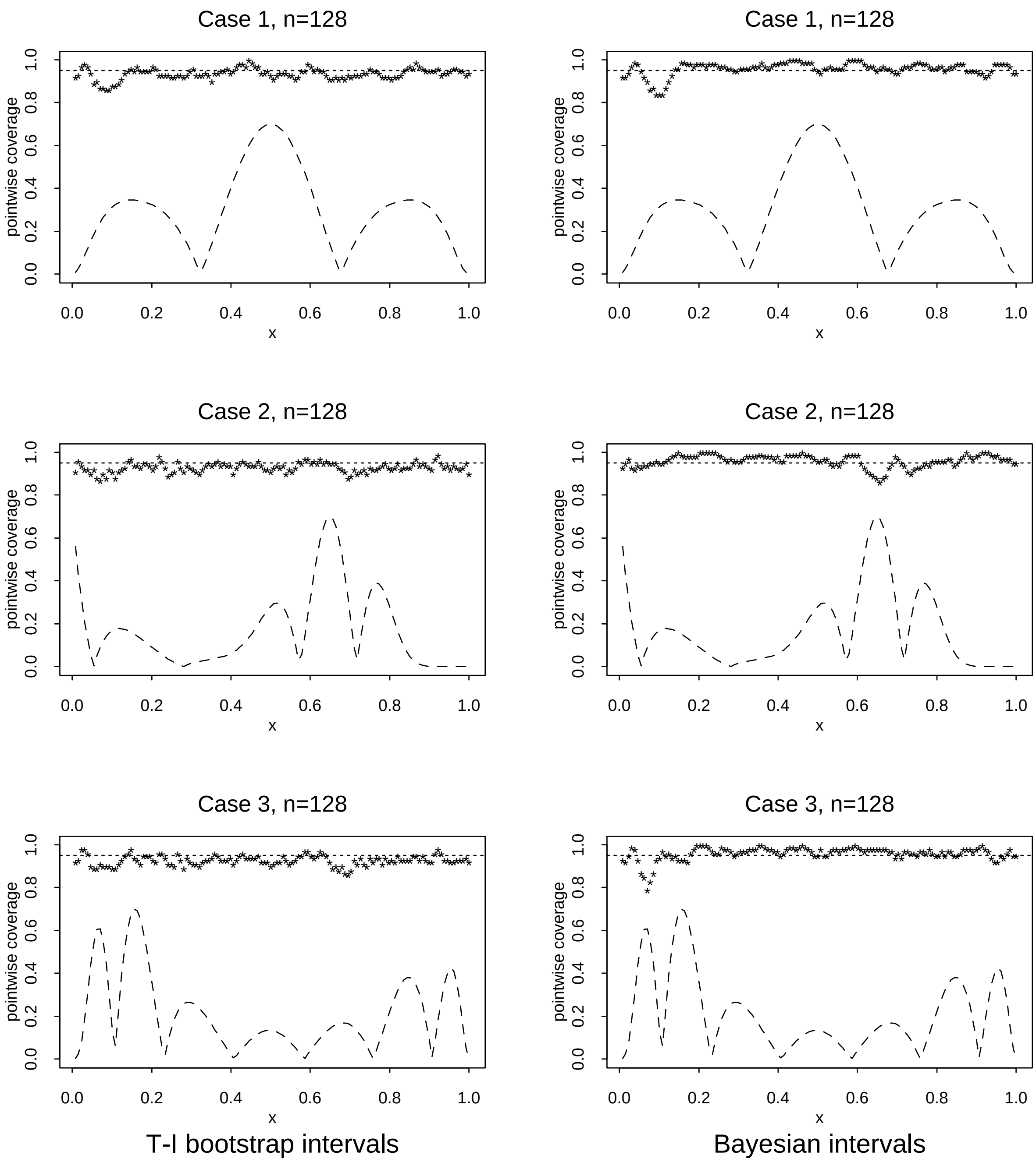

Figure 2.2: Stars are pointwise coverage of 95% confidence intervals when $\sigma = 0.1$ and n=128. Dashed curves are the magnitude of $|\ddot{f}|$. Left column: T-I bootstrap intervals; right column: Bayesian intervals. Top row: Case 1; middle row: Case 2; bottom row: Case 3.

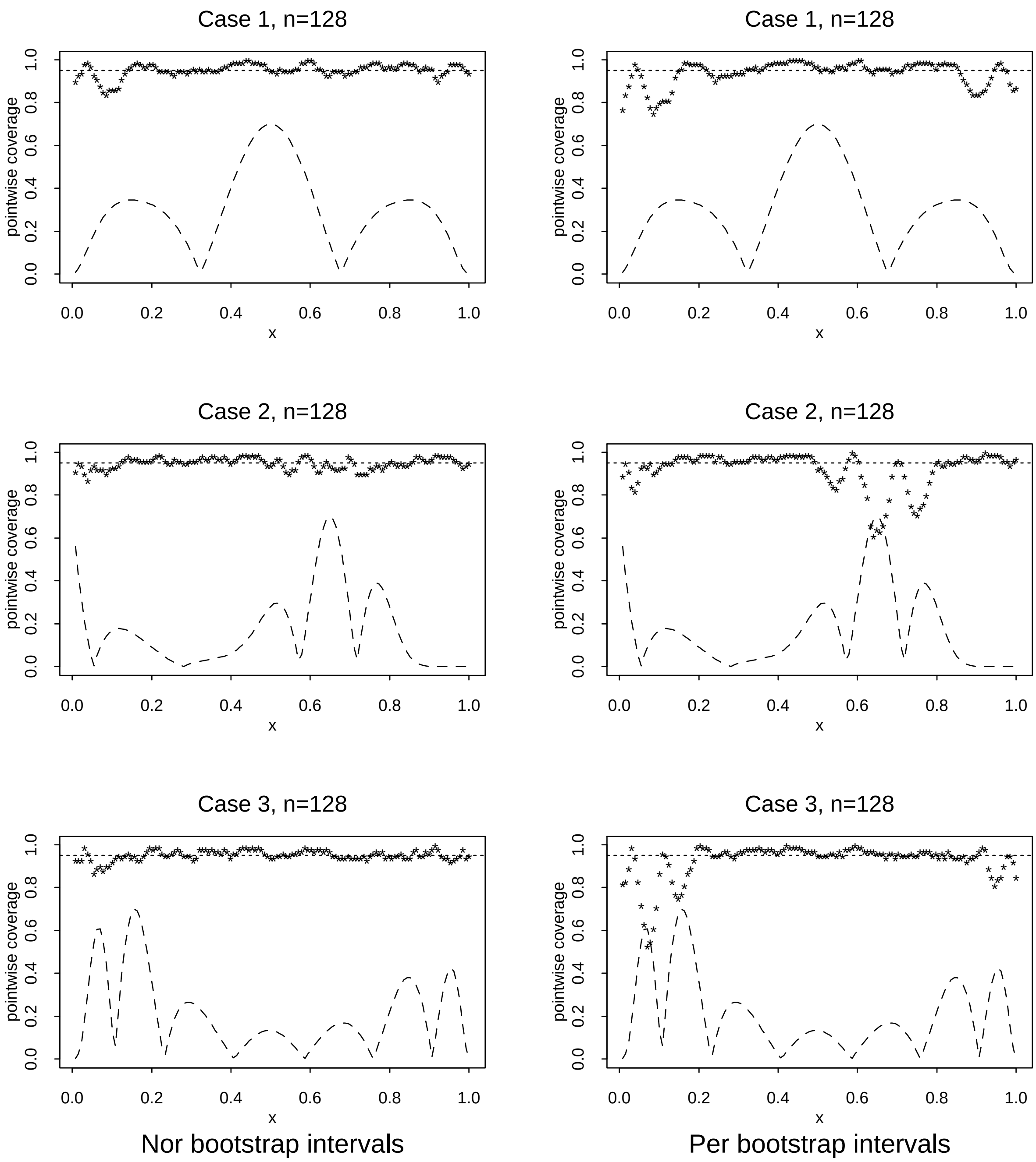

Figure 2.3: Stars are pointwise coverage of 95% confidence intervals when $\sigma = 0.1$ and n=128. Dashed curves are the magnitude of $|\ddot{f}|$. Left column: Nor bootstrap intervals; right column: Per bootstrap intervals. Top row: Case 1; middle row: Case 2; bottom row: Case 3.

3.2 Bayesian Confidence Intervals for Components

Suppose that f in (3.1) is a sample path from the Gaussian process

$$f(\boldsymbol{t}) = \sum_{k=1}^{M} \tau_k \phi_k(\boldsymbol{t}) + b^{\frac{1}{2}} \sum_{j=1}^{p} \sqrt{\theta_j} Z_j(\boldsymbol{t})$$

where $\boldsymbol{\tau} = (\tau_1, \cdots, \tau_m)^T \sim N(0, \xi I_{m\times m})$, $M = \dim(\mathcal{H}^0)$, $\phi_1, \cdots, \phi_M$ span $\mathcal{H}^0$, the Z_j's are independent, zero mean Gaussian stochastic processes, independent of $\boldsymbol{\tau}$, with $EZ_j(\boldsymbol{s})Z_j(\boldsymbol{t}) = R_j(\boldsymbol{s}, \boldsymbol{t})$, where $R_j(\boldsymbol{s}, \boldsymbol{t})$ is the reproducing kernel of H^j. Gu and Wahba (1993b) show that with $b = \frac{\sigma^2}{n\lambda}$, $\hat{f}_\lambda(\boldsymbol{t}) = \lim_{\xi\to\infty} \mathrm{E}(f(\boldsymbol{t})|\boldsymbol{y})$, and that this relation holds with $\hat{f}_\lambda$ and f replaced by $P^j\hat{f}_\lambda$ and $P^j f$. They also get the posterior variances for each component. Then they construct Bayesian confidence intervals for each component similar to the univariate case.

3.3 Bootstrap Confidence Intervals for Components

Similar to the univariate smoothing spline case, a component-wise bootstrap tries to estimate the distribution of an SS ANOVA estimate directly. Following the same procedure as in the univariate case, we first get estimates of f, components of f and σ^2. Then we generate a bootstrap sample and fit with an SS ANOVA model $\hat{f}^*_{\hat{\lambda}^*}$ and collect its components. Repeat this process B times. Treating each component as a single function, we can calculate bootstrap confidence intervals for each component as in the univariate case. We construct T-I (simply denoted as T), Nor, Per, Piv and BC intervals in this case. We do not construct T-II intervals since they are inferior to T-I's.

3.4 Simulations

We use the same example function and the same model space as in Gu and Wahba (1993b). Let $\mathcal{T} = \mathcal{T}^{(1)} \otimes \mathcal{T}^{(2)} \otimes \mathcal{T}^{(3)} = [0,1]^3$, $f(\boldsymbol{t}) = C + f_1(t_1) + f_2(t_2) + f_{12}(t_1, t_2)$, where $C = 5$, $f_1(t_1) = e^{3t_1} - (e^3 - 1)/3$, $f_2(t_2) = 10^6[t_2^{11}(1-t_2)^6 - \beta_{12,7}(t_2)] + 10^4[t_2^3(1-t_2)^{10} - \beta_{4,11}(t_2)]$, and $f_{12}(t_1, t_2) = 5\cos(2\pi(t_1 - t_2))$, where $\beta_{p,q}$ is the Beta function. We fit with a model having three main effects and one two factor interaction: $f(\boldsymbol{t}) = C + f_1(t_1) + f_2(t_2) + f_{12}(t_1, t_2) + f_3(t_3)$.

We only run simulations for n=100 (n=200 is too computer intensive) with three levels of σ (1, 3 and 10). GCV method is used to choose smoothing parameters for all simulations. B=100 for all simulations. 100 replicates are generated for each experiment, and data for the 95% confidence intervals are collected. In each case, the number of data points at which the confidence intervals cover the true values of f, f_1, f_2, f_{12} and f_3 are recored. These numbers are then divided by the sample size to form the coverage percentage of the intervals on design points. We summarize these coverage percentages using box-plots (Figures 3.1, 3.2 and 3.3). We only plot the box-plots for T, Nor, Per and Piv intervals since they are uniformly better than BC intervals. Similar to the smoothing spline case, the GCV criterion selects $\hat{\lambda} \approx 0$ (nearly interpolates the data) in one of the one hundred $\sigma = 1$ replicates, in one of the $\sigma = 3$ replicates and in two of the $\sigma = 10$ replicates. Again, these cases can be readily detected by examining estimates of σ^2 which are orders of magnitude smaller than the true values. We exclude these four cases.

From these box-plots, we see that T, Nor, Per and Piv intervals work well. T and Nor intervals are a little bit better than Per and Piv intervals. The good performance of Nor intervals suggests that Lemma 1 might be true for component-wise Nor intervals. Comparing with the box-plots of Gu and Wahba (1993b) in the top row of their Figure 1, we can see that T bootstrap confidence intervals have somewhat better mean coverages and smaller variability than the Bayesian confidence

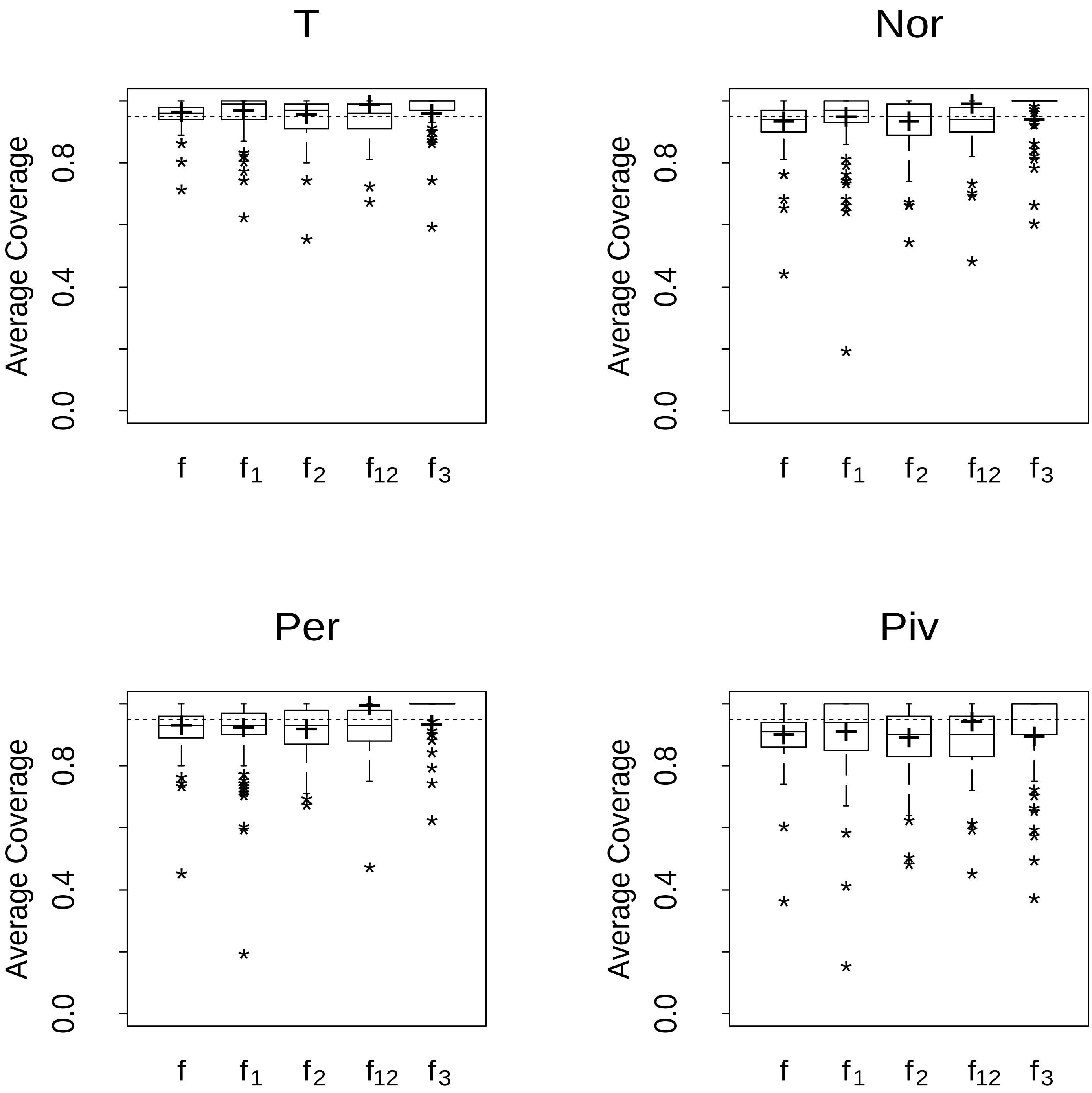

Figure 3.1: Coverage percentages of 95% bootstrap intervals when $\sigma = 1$, for T, Nor, Per and Piv methods. Plusses: sample means; dotted lines: nominal coverage.

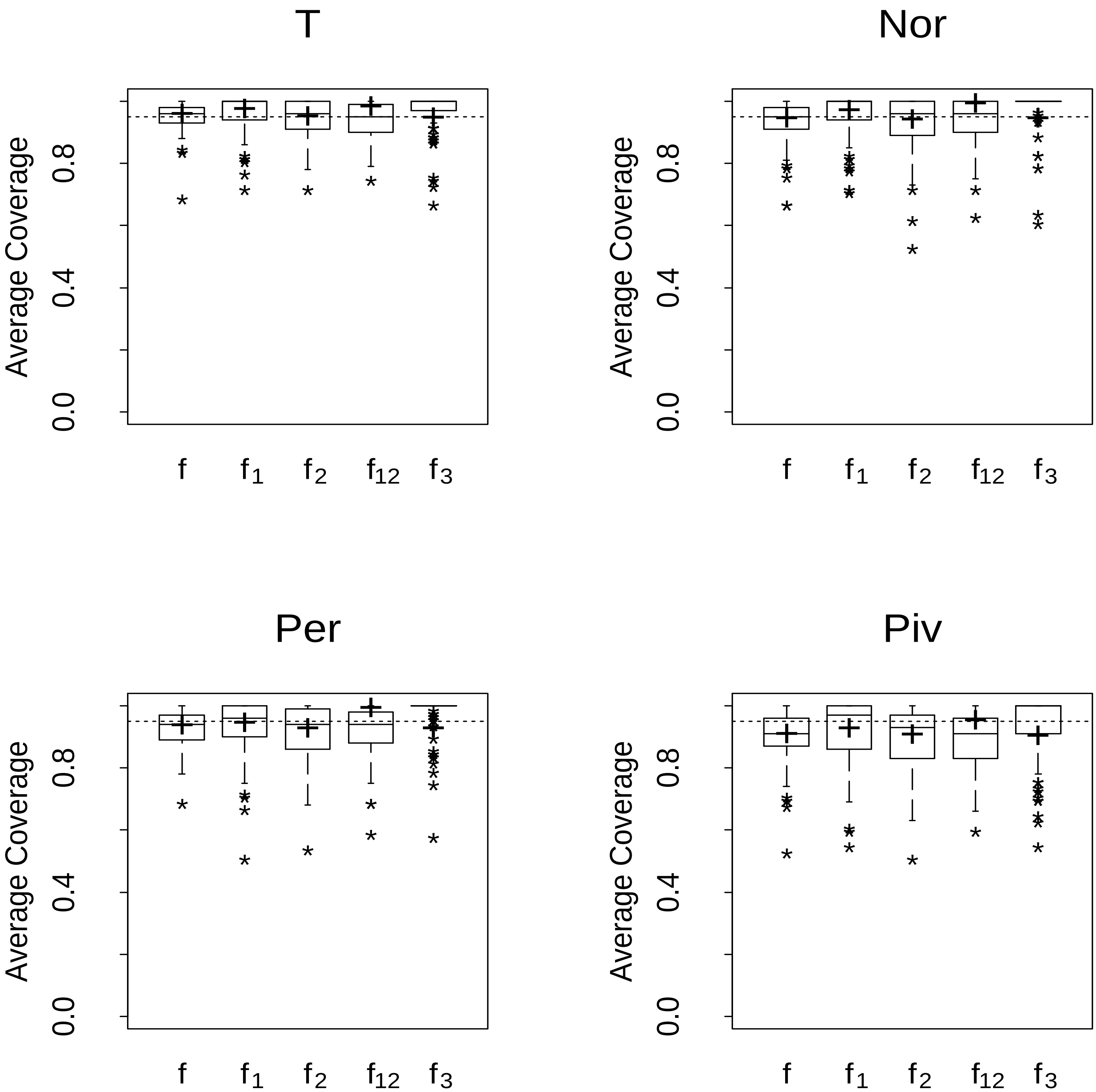

Figure 3.2: Coverage percentages of 95% bootstrap intervals when $\sigma = 3$, for T, Nor, Per and Piv methods. Plusses: sample means; dotted lines: nominal coverage.

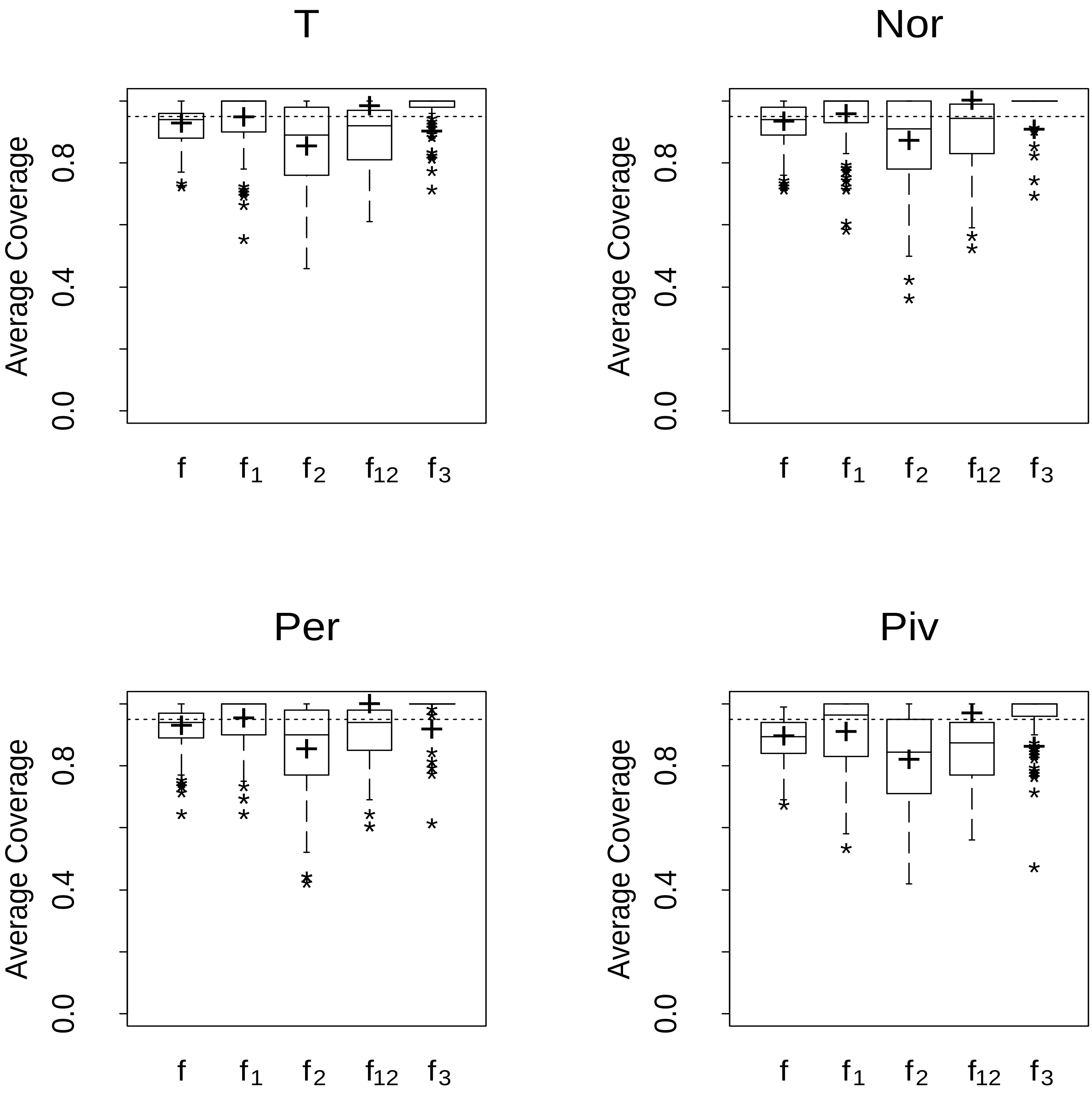

Figure 3.3: Coverage percentages of 95% bootstrap intervals when $\sigma = 10$, for T, Nor, Per and Piv methods. Plusses: sample means; dotted lines: nominal coverage.

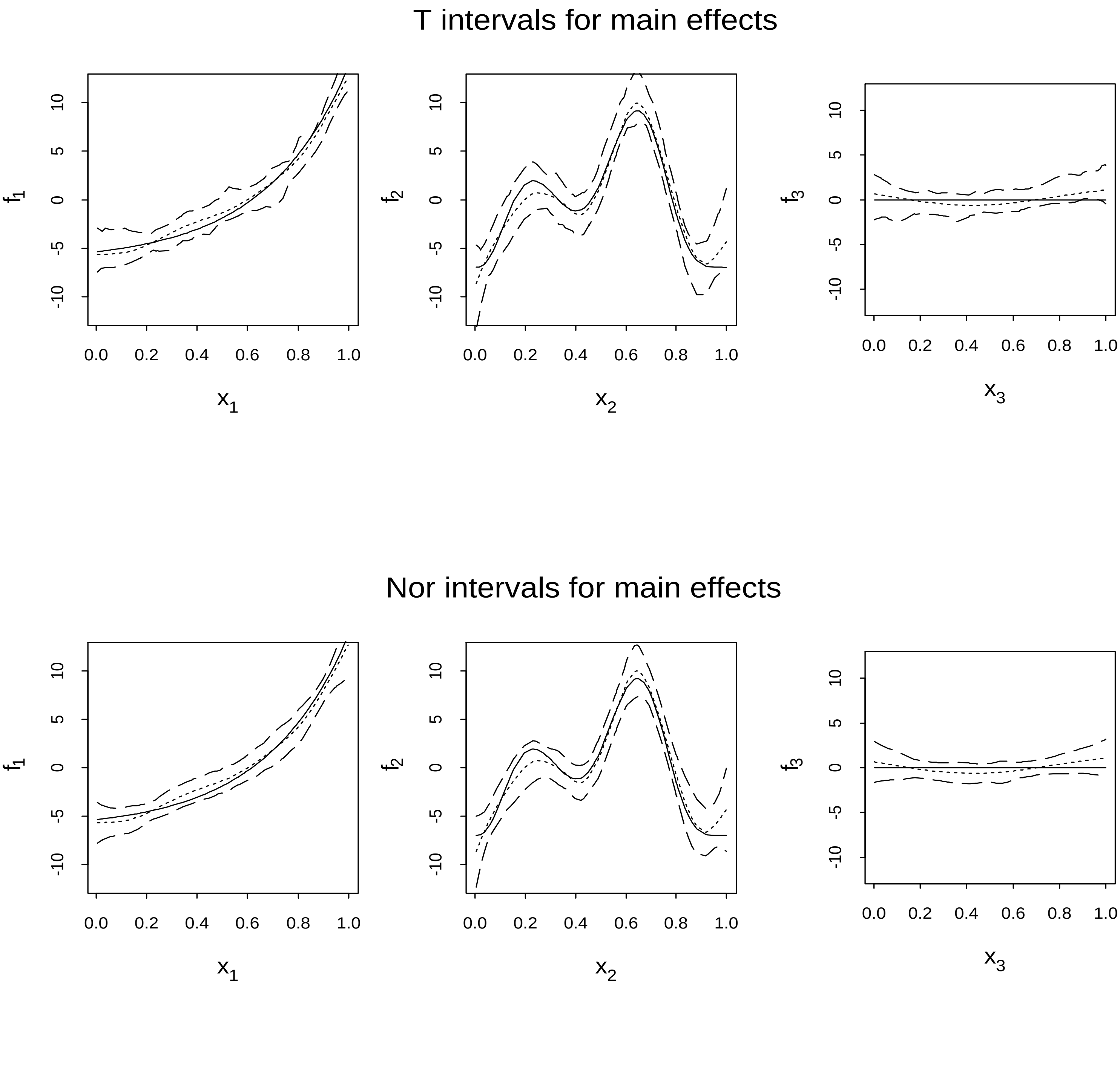

Figure 3.4: Display of the 95% intervals in a "typical" n=100 and $\sigma = 3$ case. Solid lines: true function; dotted lines: SS ANOVA fit; dashed lines: confidence intervals. Top row: T intervals for 3 main effects; bottom row: Nor intervals. Left column: f_1; middle column: f_2; right column: f_3.

intervals. A point worth noting is that Bayesian confidence intervals for f_3 are actually simultaneous confidence intervals. This is not true for bootstrap confidence intervals.

We visually inspected many of the plotted intervals and (with the above four exceptions) they all look similar. A "typical" case for $\sigma = 3$ is plotted in Figure 3.4. We can see that bootstrap confidence intervals are not very smooth. This is because $B = 100$ is not big enough. We expect that with $B \geq 500$, the bootstrap confidence intervals will look smoother.

4 Confidence Intervals for Penalized Log Likelihood Estimation for Data from Exponential Families

4.1 The Model

Nelder and Weddlerburn (1972) introduce a collection of statistical regression models known as generalized linear models (GLIM's) for analysis of data from exponential families (see McCullagh and Nelder (1989)). Data have the form $(y_i, \boldsymbol{t}(i))$, $i = 1, \cdots, n$, where y_i are independent observations, each from an exponential family with density $\exp((y_i h(f_i) - b(f_i))/a(\omega) + c(y_i, \omega))$, where $f_i = f(\boldsymbol{t}(i))$ is the parameter of interest and depends on the covariate $\boldsymbol{t}(i)$, $\boldsymbol{t}(i) \in \mathcal{T}$. $h(f_i)$ is a monotone transformation of f_i known as the canonical parameter. ω is an unknown scale parameter. GLIM model assumes that f is a linear or other simple parametric function of the components of $\boldsymbol{t}$. To achieve greater flexibility, O'Sullivan (1983), O'Sullivan, Yandell and Raynor (1986) and Gu (1990) only assume f is in a RKHS $\mathcal{H}$ on $\mathcal{T}$. See also Wahba (1990). In what follows we will only consider the univariate case $d = 1$, $p = 1$. The estimate of f_λ is then the solution of the following penalized log likelihood problem

$$\min L_{\boldsymbol{y}}(\boldsymbol{f}) + \frac{n}{2}\lambda\|P_1 f\|^2, \qquad f \in \mathcal{H}, \tag{4.1}$$

where $L_{\boldsymbol{y}}(\boldsymbol{f})$ denotes the minus log likelihood, $\mathcal{H} = \mathcal{H}^0 \oplus \mathcal{H}^1$, P_1 is the projector onto $\mathcal{H}^1$ and $\dim(\mathcal{H}^0) = M < \infty$. λ is the smoothing parameter which can be estimated by an iterative GCV or UBR method (see Gu (1992a)). For penalized log likelihood estimation with smoothing spline ANOVA, see Wahba, Wang, Gu, Klein and Klein (1994).

4.2 Approximate Bayesian Confidence Intervals

Considering only the univariate case here, and setting $\boldsymbol{t} = t$, suppose f is a sample path from the Gaussian process

$$f(t) = \sum_{k=1}^{M} \tau_k \phi_k(t) + b^{\frac{1}{2}} Z(t),$$

where $\boldsymbol{\tau} = (\tau_1, \cdots, \tau_m)^T \sim N(0, \xi I_{m\times m})$, $\phi_1, \cdots, \phi_M$ span $\mathcal{H}^0$, $Z(t)$ is a zero mean Gaussian process and is independent of $\boldsymbol{\tau}$, with $EZ(s)Z(t) = R(s,t)$, where $R(s,t)$ is the reproducing kernel of $\mathcal{H}^1$. Gu (1992b) sets $b = \frac{\sigma^2}{n\lambda}$, and obtained the approximate posterior distribution of $\boldsymbol{f}$ given $\boldsymbol{y}$ as Gaussian with $\hat{f}_\lambda(t) \approx \lim_{\xi\to\infty} \mathrm{E}(f(t)|\boldsymbol{y})$. He found the posterior coveriance $\lim_{\xi\to\infty} \mathrm{Cov}(\boldsymbol{f}|\boldsymbol{y})$, in terms of the relevant "hat" or influence matrix for the problem and the Hessian of the log likelihood with respect to $\boldsymbol{f}$, evaluated at the fixed point of the Newton iteration for the minimizer of (4.1). See Gu (1992b) for details. Wang (1994) proves that these Bayesian confidence intervals approximately have the ACP property.

4.3 Bootstrap Confidence Intervals

The process is the same as in Section 2. The only difference is now the bootstrap samples are non-Gaussian. No approximation is involved after we get a spline fit, so we might expect that the bootstrap confidence intervals will work better than the Bayesian confidence intervals. We construct Nor, Per, Piv and BC bootstrap confidence intervals. Notice that in the case of Bernoulli data, there is no unknown scale parameter. Therefore the Piv intervals are the same as T-I intervals.

4.4 A Simulation

We use the same experimental design as Gu (1992b). Bernoulli responses y_i are generated on $t_i = (i - 0.5)/100$, $i = 1, \cdots, 100$, 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 are generated. $B = 100$. The iterative unbiased risk (UBR) method is used to select λ (U in Gu (1992a)). We also repeat Gu's (1992b) experiment for Bayesian confidence intervals, using UBR to select λ, which will allow direct comparison with the bootstrap intervals here.

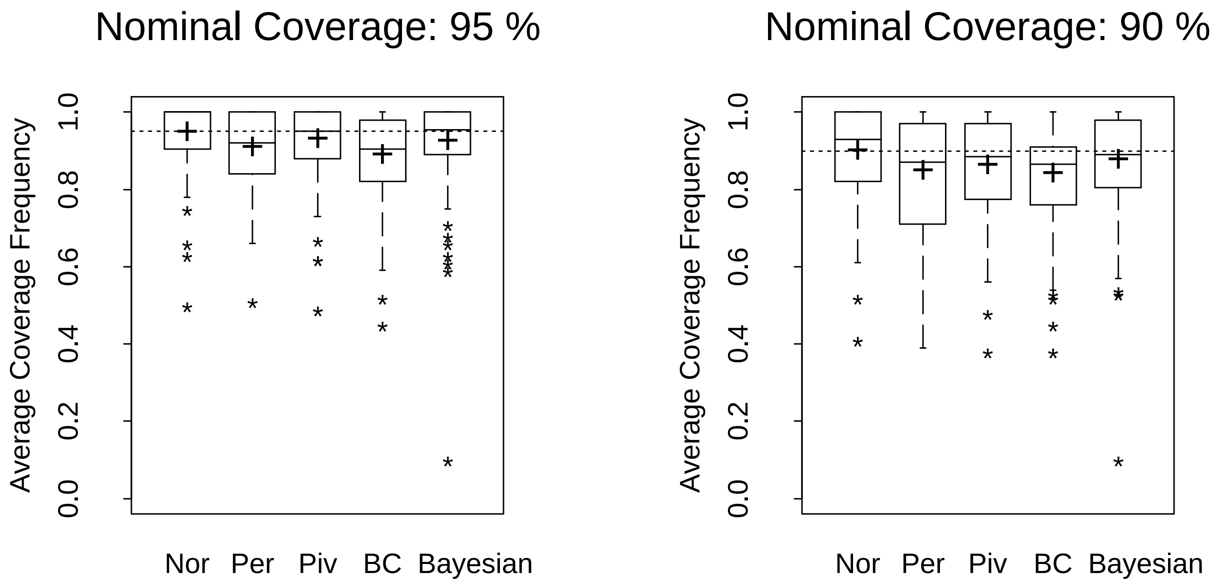

Figure 4.1: Coverage percentages bootstrap intervals. Plusses: sample means; dotted lines: nominal coverage.

The coverage percentage of 95% and 90% intervals are plotted in Figure 4.1. Nor and Piv intervals work better than Per and BC intervals, and are similar to Bayesian intervals. Nor has smaller variance. The pointwise coverage coverages are plotted in Figure 4.2. The bootstrap intervals are similar to Bayesian confidence intervals in the sense that the pointwise coverage is smaller than the nominal value at high curvature points. Nor intervals are a little better than Piv's in terms of dropping less than Piv's at high curvature points. Nor or Bayesian intervals would be

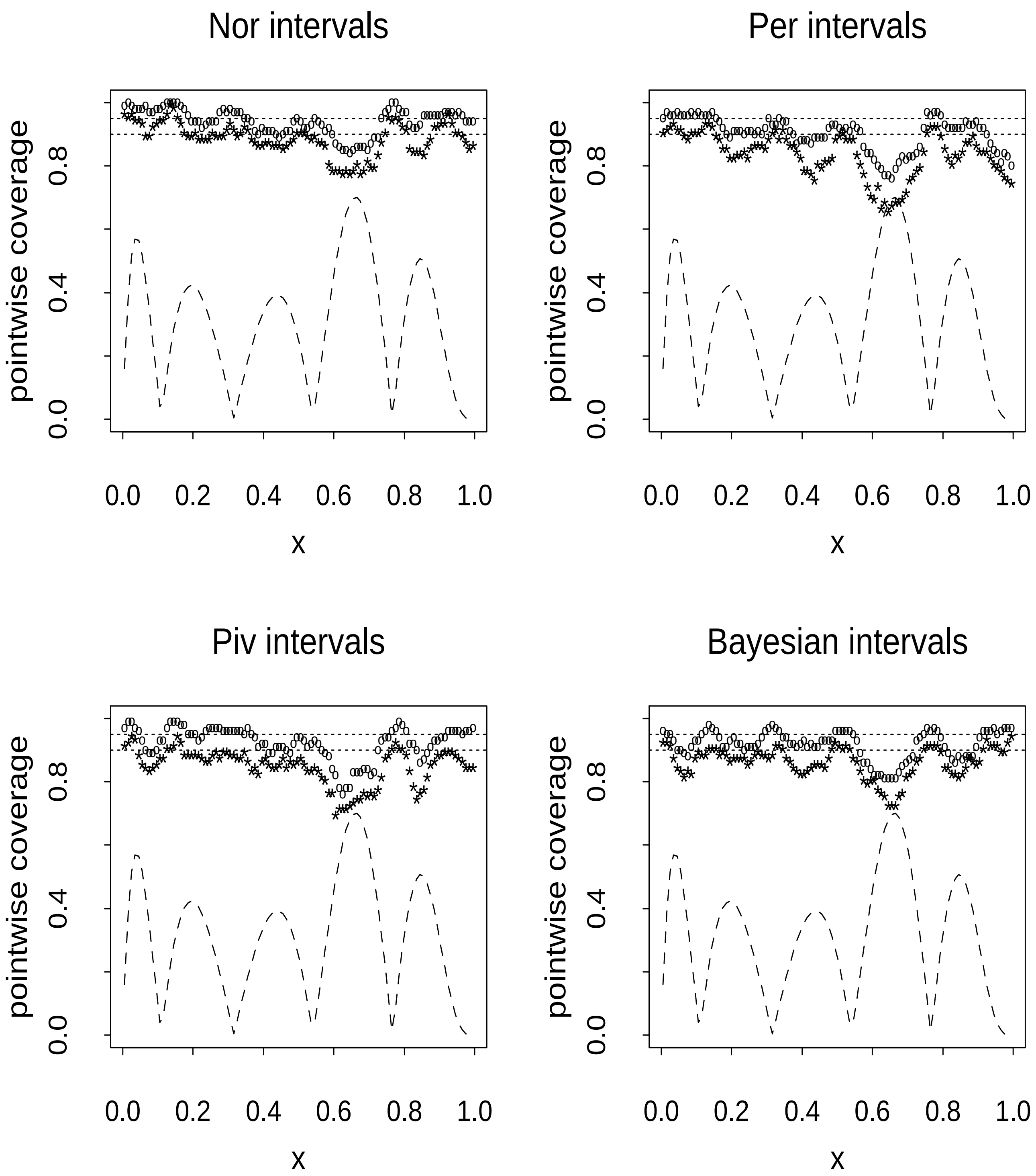

Figure 4.2: Stars are pointwise coverage of 90% intervals. Circles are pointwise coverage of 95% intervals. Dotted lines are nominal values 90% and 95%. Dashed curves are the magnitude of $|\ddot{f}|$.

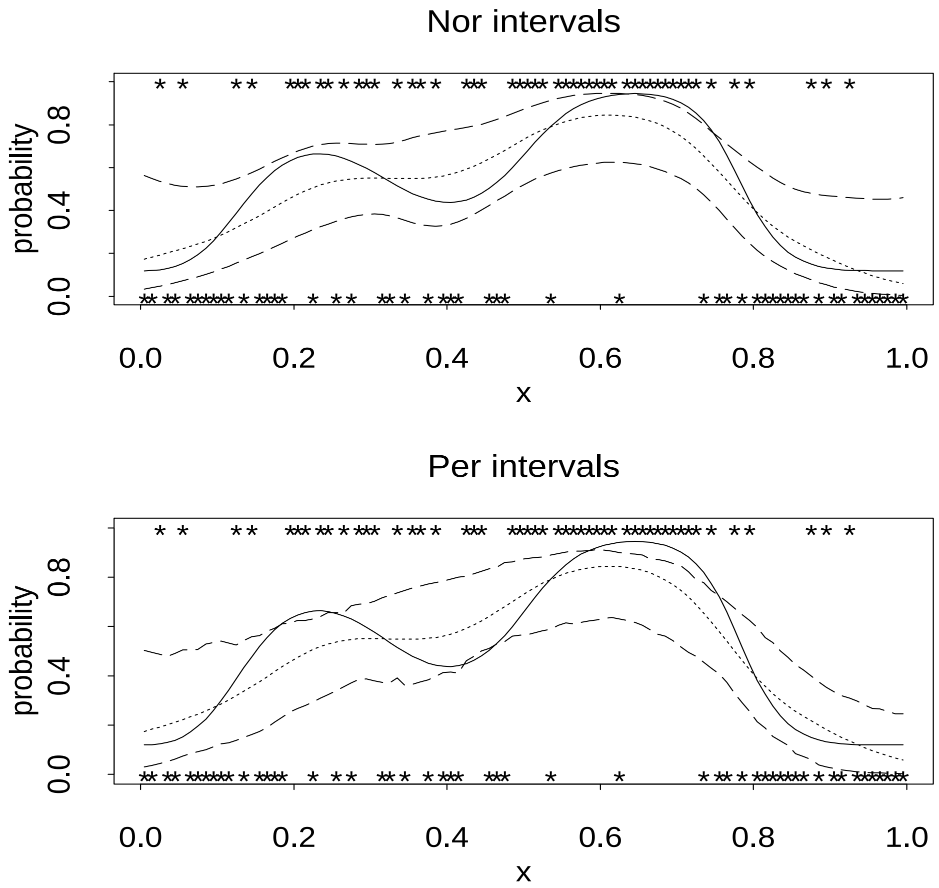

Figure 4.3: Display of the 90% intervals in a "typical" case. Stars: data; solid lines: true function; dashed lines: spline fit; dotted lines: confidence intervals. Top: Nor intervals; bottom: Per intervals.

recommended on the basis of this particular experiment. A "typical" case is plotted in Figure 4.3.

5 Conclusions

We have compared the performance of several versions of bootstrap confidence intervals with themselves and with Bayesian confidence intervals. Bootstrap confidence intervals work as well as Bayesian intervals from an ACP point of view and appear to be better for small sample sizes. We find it reassuring that the best variations of bootstrap confidence intervals and the Bayesian confidence intervals give such similar results. This similarity lends credence to both methods. The advantages of bootstrap confidence intervals are:

1) They are easy to understand, even by an unsophisticated user. They can be used easily with any distribution;

2) They appear to have better coverage in small samples in the examples tried.

The disadvantage of bootstrap confidence intervals is that computing them is very computer intensive, especially for SS ANOVA and non-Gaussian data. But compared to typical data collection costs, the cost of several minutes or even several hours of CPU time is small.

Just like Bayesian intervals, these bootstrap confidence intervals should be interpreted as across the curve, instead of pointwise.

Even though the bootstrap confidence intervals are essentially an automatic method, they should be implemented carefully. If the bootstrap method is used, we recommend using either T-I or Nor intervals for Gaussian data, and Nor intervals for Non-Gaussian data. The commonly used Per intervals work well, but are inferior to T-I or Nor intervals in our simulations. When bootstrapping for small sample sizes and using GCV to select smoothing parameter(s), one should exclude interpolating cases, especially when using T intervals.

6 Acknowledgments

This research was supported by the National Science Foundation under Grant DMS-9121003 and the National Eye Institute under Grant R01 EY09946. We thank Douglas Bates for his invaluable work in setting up the computing resources used in this project. Y. Wang would like to acknowledge a helpful conversation with W. Y. Loh concerning the bootstrap.

References

Abramovich, F. and Steinberg, D. (1993). Improved inference in nonparametric regression using L_k-smoothing splines, manuscript, Tel Aviv University.

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

Carter, C. K. and Eagleson, G. K. (1992). A comparison of variance estimations in nonparametric regression, *Journal of the Royal Statistical Society B* **54**: 773–780.

Dikta, G. (1990). Bootstrap approximation of nearest neighbor regression function estimates, *Journal of Multivariate Analysis* **32**: 213–229.

Efron, B. (1981). Nonparametric standard errors and confidence intervals, *Canadian Journal of Statistics* **9**: 139–172.

Efron, B. (1982). *The Jackknife, the bootstrap, and Other Resampling Plans*, CBMS 38, SIAM-NSF.

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. (1992a). Cross-validating non Gaussian data, *Journal of Computational and Graphical Statistics* **2**: 169–179.

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

Gu, C. and Wahba, G. (1993a). Semiparametric ANOVA with tensor product thin plate spline, *Journal of the Royal Statistical Society B* **55**: 353–368.

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

Hall, P. (1990). Using the bootstrap to estimate mean squared error and select smoothing parameter in nonparametric problems, *Journal of Multivariate Analysis* **32**: 177–203.

Hardle, W. and Bowman, W. (1988). Bootstrapping in nonparametric gression: Local adaptive smoothing and confidence bands, *Journal of the American Statistical Association* **83**: 102–110.

Hardle, W. and Marron, J. S. (1991). Bootstrap simultaneous error bars for nonparametric regression, *The Annals of Statistics* **19**: 778–796.

Kooperberg, C., Stone, C. and Truong, Y. K. (1993). Hazard regression, Technical Report No. 389, University of California-Berkeley, Dept. of Statistics.

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

Meier, K. and Nychka, D. (1993). Nonparametric estimation of rate equations for nutrient uptake, *Journal of the American Statistical Association* **88**: 602 –614.

Nychka, D. (1988). Bayesian confidence intervals for smoothing splines, *Journal of the American Statistical Association* **83**: 1134–1143.

Nychka, D. (1990). The average posterior variance of a smoothing spline and a consistent estimate of the average squared error, *The Annals of Statistics* **18**: 415–428.

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

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

Wahba, G. (1983). Bayesian confidence intervals for the cross-validated smoothing spline, *Journal of the Royal Statistical Society B* **45**: 133–150.

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

Wahba, G. and Wang, Y. (1993). Behavior near zero of the distribution of GCV smoothing parameter estimates for splines, TR 910, University of Wisconsin-Madison, Dept. of Statistics, submitted.

Wahba, G., Wang, Y., Gu, C., Klein, R. and Klein, B. (1994). Structured machine learning for 'soft' classification with smoothing spline ANOVA and stacked tuning, testing and evaluation, University of Wisconsin-Madison Statistics Dept. TR 909, to appear in " Advances in Neural Information Processing Systems 6", J. Cowan, G. Tesauro, and J. Alspector, eds, Morgan Kauffman.

Wang, Y. (1994). Smoothing spline analysis of variance of data from exponential families, Ph.D. Thesis, University of Wisconsin-Madison, Dept of Statistics, in preparation.