Confidence Intervals for Nonparametric Curve Estimates: Toward More Uniform Pointwise Coverage

David J. CUMMINS, Tom G. FILLOON, and Douglas NYCHKA

Numerous nonparametric regression methods exist that yield consistent estimators of function curves. Often, one is also interested in constructing confidence intervals for the unknown function. When a function estimate is based on a single global smoothing parameter the resulting confidence intervals may hold their desired confidence level $1-\alpha$ on average but because bias in nonparametric estimation is not uniform, they do not hold the desired level uniformly at all design points. Most research in this area has focused on mean squared error properties of the estimator, for example MISE, itself a global measure. In addition, measures like MISE are one step removed from the practical issue of coverage probability. Recent work that focuses on coverage probability has considered only coverage in an average sense, ignoring the important issue of uniformity of coverage across the design space. To deal with this problem, a new estimator is developed which uses a local cross-validation criterion (LCV) to determine a separate smoothing parameter for each design point. The local smoothing parameters are then used to compute the point estimators of the regression curve and the corresponding pointwise confidence intervals. Incorporation of local information through the new method is shown, via Monte Carlo simulation, to yield more uniformly valid pointwise confidence intervals for nonparametric regression curves. Diagnostic plots are developed (*Breakout Plots*) to visually inspect the degree of uniformity of coverage of the confidence intervals. The approach, here applied to cubic smoothing splines, easily generalizes to many other nonparametric regression estimators. The improved curve estimation is not a solely theoretical improvement such as providing an estimator that has a faster EASE convergence rate but shows its worth empirically by yielding improved coverage probabilities through reliable pointwise confidence intervals.

KEY WORDS: Coverage probability; Local cross-validation; Nonparametric regression; Pointwise confidence intervals, Smoothing splines.

1. INTRODUCTION

A frequent problem in data analysis involves recovering an estimate of a curve (or surface) when data are observed according to a model:

$$Y_k = f(X_k) + e_k. \tag{1}$$

Here (X_k, Y_k), $1 \le k \le n$, are observed, f is a smooth (differentiable) function and e_k are independent random variables with $E(e_k) = 0$ and $\text{var}(e_k) = \sigma^2$. Besides being of interest in its own right this model serves as the simplest case for more general curve fitting problems. Statistical techniques developed for (1) can be extended to cover more complicated situations.

Often in a practical setting, curve estimates without any measure of their reliability are not very useful. For most statistical applications, approximate interval estimates can be constructed by centering the interval at the estimate and taking the width to be some multiple of the standard error based on the normal (or T) distribution. This approach is admittedly approximate but even this fairly crude strategy is difficult to apply in estimating a curve or surface. The problem stems from the fact that standard nonparametric curve estimates deliberately balance the bias and variance of the estimate to minimize the average mean squared error (averaged across design points). This is the typical result of applying a single global smoothing parameter. The net result is that a confidence interval of the form

$$\hat{f}(X_k) \pm Z_{\alpha/2}\sqrt{\text{var}(\hat{f}(X_k))} \tag{2}$$

may not have coverage probability at X_k that is close to $(1-\alpha)$ even if the estimate is normally distributed.

The problem lies in the fact that $\text{Bias}(\hat{f}(X_k))^2/\text{var}(\hat{f}(X_k))$ converges to a single fixed ratio when the curve is constructed to minimize average mean squared error. If this ratio is large, then there will be low coverage because the interval will be consistently centered at the wrong place. The impact of this problem is compounded by the fact that standard confidence intervals will be unreliable precisely where the curve has the most interesting structure. Figure 1 gives an illustration of the fact that coverage probabilities are not uniform across design points for a traditional [minimum EASE; see (4)] nonparametric function estimate. For such methods the estimate tends to be most highly biased at points of sharp curvature; thus the coverage probability is low at points in the function where the bias of the estimate is large.

There are several approaches for dealing with this problem. One strategy is to approximate an unbiased estimate of $f(x)$ by estimating the bias and subtracting it from the original estimate (e.g., Eubank and Speckman 1993). Another strategy is to inflate the standard error to reflect the bias in the estimator. Both of these methods use a fixed amount of smoothing across the range of data. An appealing alternative is to avoid the practice of making complicated adjustments to the estimate. Accordingly, we present a method that varies the amount of smoothing in order to adapt to the different amounts of curvature in f. The idea is to use a slightly different smoothing parameter at each point. This smoothing parameter is determined by a data adaptive method derived from a local version of cross-validation. One advantage of using cross-validation is that there is a minimum of tuning parameters needed.

David J. Cummins is Research Scientist, Statistical and Mathematical Sciences, Eli Lilly & Company Indianapolis, IN 46285. (E-mail: *Cummins_DJ@lilly.com*). Tom G. Filloon is in the Health Care Research Center, Procter & Gamble Co., Mason, OH 45040. Douglas Nychka is Project Leader, Geophysical Statistics Project, National Center for Atmospheric Research, Boulder, CO 80307.

Journal of the American Statistical Association
March 2001, Vol. 96, No. 453, Theory and Methods

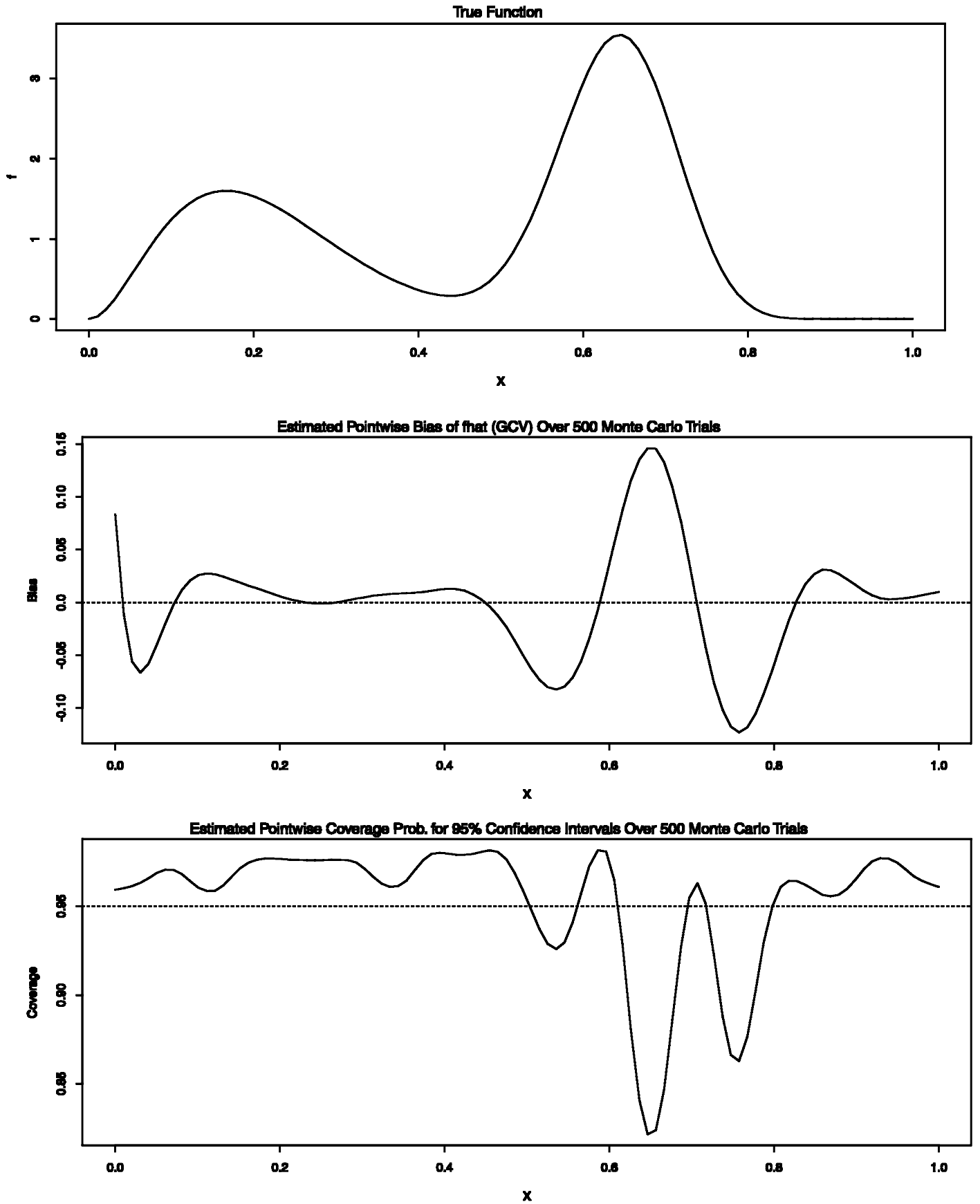

Figure 1. Coverage Probabilities for a Spline Estimate. Changes in curvature cause varying amounts of bias in the estimate. This, in turn, affects coverage probability, so that it does not hold a uniform level across design points. Average coverage (across design points) for this example was .952, which is clearly misleading as a single number.

The idea of a variable bandwidth or smoothing parameter is not a new one but most researchers have only focused on mean squared error properties and not the interval estimates. Some relevant papers include Muller & Stadtmuller (1987), Brockman, Gasser, and Herrmann (1993), and Fan, Hall, Martin, and Patil (1996). Some of the more recent articles include Ruppert and Carroll (2000), which uses pth degree piecewise polynomials, has several tuning parameters and develops an approach similar to this work. A very promising graphical technique for variable smoothing, called SiZer, has been developed by Chaudhuri and Marron (1999). Although it is more subjectively based, SiZer provides an alternative to estimation based methods.

We feel that the greatest benefit of variable smoothing is in adjusting the local bias of the curve estimate so that it is possible to construct reliable confidence intervals. Another attractive feature is that computing the local smoothing parameters carries little extra computational burden. Section 2 discusses the choice of smoothing parameter for a global estimate and Section 3 presents a version of smoothed, local cross-validation for local smoothing of the data. Section 4 describes the form of the pointwise intervals and Section 5 presents the results of a simulation study. These results are discussed in Section 6. An appendix is included that supports some of the theoretical conjectures about the properties of local smoothing parameter selection. This work is based on the original work in Filloon (1990).

2. CROSS-VALIDATION AND SMOOTHING PARAMETER SELECTION

Based on the additive model (1) this work considers linear estimators of the form

$$\hat{f}_\lambda(\mathbf{x}) = \mathbf{A}(\lambda)\mathbf{y} \tag{3}$$

where $\mathbf{A}(\lambda)$ is an $n \times n$ smoothing matrix. The parameter λ plays the role of a smoothing parameter or bandwidth. Let $\operatorname{tr}\mathbf{A}(\lambda)$ be the effective number of parameters associated with the estimated curve. Although the smoothing matrix can be specified in a number of ways, we focus on cubic smoothing splines. This choice is out of convenience and the following ideas are valid for any reasonable family of smoothing matrices (indexed by λ). These include kernel regression, orthogonal series, regression splines, and locally weighted regression. For the simple one-dimensional curve fitting problem with regularly spaced x values, all these estimators tend to be similar away from the boundaries (see Hastie and Tibshirani 1990, chap. 2). One advantage of smoothing splines is that they can be interpreted as penalized likelihood estimates. Thus it is straightforward to generalize these estimates to more complicated observational models. Another advantage is that computing $\mathbf{A}(\lambda)\mathbf{y}$ is efficient, involving $O(n)$ operations.

The most common measure of statistical accuracy for a curve estimate is the expected average squared error:

$$\mathrm{EASE}(\lambda) = E\left\{(1/n)\sum_{i=1}^{n}\left(f(x_i) - \hat{f}_\lambda(x_i)\right)^2\right\} \tag{4}$$

and a good value for the smoothing parameter is often associated with the value that minimizes this risk function. There is some debate as to whether this risk is a practical measure of fit. However, for our needs, we are more interested in how minimizing this criterion with respect to λ balances the bias and variance in the resulting estimate. For most nonparametric curve estimates there are well-established asymptotic results for this comparison. For example, one can decompose EASE into bias and variance terms so that

$$\mathrm{EASE}(\lambda) = (1/n)\sum_{i=1}^{n} b_i(\lambda)^2 + (1/n)\sum_{i=1}^{n} v_i(\lambda),$$

where

$$b_i(\lambda) = f(x_i) - E(\hat{f}_\lambda(x_i))$$

and

$$v_i(\lambda) = \mathrm{var}(\hat{f}_\lambda(x_i)).$$

Let λ_0 denote the value of λ that gives the global minimum of $\mathrm{EASE}(\lambda)$. If $\hat{f}_\lambda$ is a cubic smoothing spline, and f has 4 continuous derivatives and satisfies certain boundary conditions, one can show that

$$\frac{(1/n)\sum_{i=1}^{n} b_i(\lambda_0)^2}{(1/n)\sum_{i=1}^{n} v_i(\lambda_0)} \to \frac{1}{8}$$

as $n \to \infty$ (Nychka 1990). Thus if the smoothing parameter is chosen to minimize EASE then at least on the average the contribution of the bias is a small fraction ($\frac{1}{9}$) of the total mean squared error in the estimate. It is this fact that justifies the accurate average coverage for a version of pointwise confidence intervals introduced by Wahba, and more is said about this in Section 4.

The problem with a global criterion such as EASE is that it will not guarantee small biases (relative to the variances) uniformly at all points. Ideally, if one knew f, one could choose the smoothing parameter to minimize the pointwise expected squared error:

$$\mathrm{ESE}_i(\lambda) = E(f(x_i) - \hat{f}_\lambda(x_i))^2 \tag{5}$$

Let $\lambda_{0,i}$ denote the minimizer of this function. Then, one can show for cubic smoothing splines $b_i(\lambda_{0,i})^2 / v_i(\lambda_{0,i}) \to \frac{1}{8}$ as $n \to \infty$ (Nychka 1995). This stability in the relative size of the bias and variance suggests that a local smoothing parameter would be useful. The following subsections explain how one can estimate $\lambda_{0,i}$. We begin with a review of how global smoothing parameters are estimated.

2.1 Choice of Global Smoothing Parameter

One choice for λ comes from a generalized criterion based on the idea of cross-validation. The method of cross-validation was introduced by Wahba and Wold (1975) in the context of smoothing splines. Let $\hat{f}_\lambda^i$ denote the curve estimate for f based on the reduced data set where the ith data point has been omitted. The cross-validation function $\mathrm{CV}(\lambda)$ is defined as

$$\mathrm{CV}(\lambda) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}_\lambda^i(x_i))^2. \tag{6}$$

The motivation for minimizing this quantity is that it is a consistent estimate (to within a constant) of EASE(λ). Given the linear form of the estimator, the reduced estimator (called the leave-one-out estimate) is often easy to calculate. For a smoothing spline there is a surprisingly simple relationship: The cross-validation residual is given by

$$e^{(i)}(\lambda) = y_i - \hat{f}_\lambda^i(x_i) = \frac{(y_i - \hat{f}_\lambda(x_i))}{1 - \mathbf{A}_{ii}(\lambda)} = \frac{e_i(\lambda)}{1 - \mathbf{A}_{ii}(\lambda)}. \quad (7)$$

The relationship (7) was shown in Craven and Wahba (1979), and in the same paper an approximation to (6), called the generalized cross-validation function, was proposed:

$$\mathrm{GCV}(\lambda) = \frac{(1/n)\sum_{i=1}^n (y_i - \hat{f}_\lambda(x_i))^2}{(1 - \mathrm{tr}\,\mathbf{A}(\lambda)/n)^2} \quad (8)$$

Under reasonable conditions GCV(λ) is a consistent estimate of EASE(λ) + σ^2. Thus one would expect the minimizer of GCV to converge to the optimal λ_0. Let $\hat{\lambda}$ denote the value of λ that minimizes GCV(λ), and let the resulting curve estimate be denoted $\hat{f}_{\hat{\lambda}}$. The estimate of λ_0 based on minimizing the GCV function has been shown to have good asymptotic properties. Under suitable conditions, $(\hat{\lambda}/\lambda_0) \to 1$ in probability as $n \to \infty$ (Härdle, Hall, and Marron 1988).

For global measures of statistical accuracy, the GCV-based estimate of the smoothing parameter works well. One problem with cross-validation, however, is the large amount of variability associated with smoothing parameter selection. Occasionally GCV will produce estimates of λ that drastically undersmooth the data and produce a very rough curve estimate. Several researchers have suggested alternative procedures to reduce the variance of the smoothing parameter estimate. It should be noted that, in general, smoothing methods work because they exploit higher-order smoothness in f. One simple modification to control undersmoothing is to increase the cost, $\mathcal{C}$, associated with each effective parameter in the curve (see Friedman and Silverman 1989 [6]). This idea leads to

$$\mathrm{GCV}(\lambda, \mathcal{C}) = \frac{(1/n)\sum_{i=1}^n (y_i - \hat{f}_\lambda(x_i))^2}{(1 - \mathcal{C}\,\mathrm{tr}\,\mathbf{A}(\lambda)/n))^2}. \quad (9)$$

This modification produces a pole at the effective number of parameters equal to $n/\mathcal{C}$ and since tr($\mathbf{A}(\lambda)$) is a monotonic function of λ the result is a lower (although data dependent) limit on the minimizer $\hat{\lambda}$ and an upper limit on the effective degrees of freedom of $\hat{f}$. Although this may seem to be an ad hoc device to prevent undersmoothing, there is a simple interpretation in terms of a risk function. Instead of estimating EASE(λ) + σ^2, for $\mathcal{C} > 1$ this criterion estimates a weighted combination of the average squared bias and the average variance. Based on a short derivation in the appendix, we show that in fact

$$E\{\mathrm{GCV}(\lambda, \mathcal{C})\} = \frac{1}{n}\sum_{i=1}^n b_i(\lambda)^2 + \frac{1}{n}\alpha\sum_{i=1}^n v_i(\lambda) + \sigma^2 + O(m_1) \quad (10)$$

where $\alpha = 1 + 2(\mathcal{C} - 1)(\mathrm{tr}\,\mathbf{A}(\lambda)/\mathrm{tr}\,\mathbf{A}^2(\lambda))$ (m_1 is defined in the Appendix). In addition, at the minimum value for this risk one can expect that the ratio of average squared bias to average variance will converge to the fraction $\alpha/4m$, where m is the order of the spline ($m = 2$ for cubic splines). Good choices for $\mathcal{C}$ will be data-dependent. However, we have found that using simulation tools and a scree-plot approach it is possible to choose a value of $\mathcal{C}$ that has minimal effect on the function estimate in most cases and gives great improvement for the few cases where GCV would otherwise drastically undersmooth. For the test functions and datasets examined in this work, a value of $\mathcal{C} = 1.2$ appears optimal. Further discussion of the choice of $\mathcal{C}$ is expected in a forthcoming paper by Cummins and Nychka.

Another aspect of smoothing parameter selection that is relevant to this article is the form of the GCV criterion for weighted squared error. Suppose that w_i, $1 \le i \le n$ are weights such that $\sum_{i=1}^n w_i = 1$. Given the expected weighted squared error,

$$\mathrm{EWSE}(\lambda) = \sum_{i=1}^n w_i b_i(\lambda)^2 + \sum_{i=1}^n w_i v_i(\lambda), \quad (11)$$

the corresponding GCV function including the modification to prevent undersmoothing is

$$\mathrm{GCV}(\lambda, \mathcal{C}) = \frac{(1/n)\sum_{i=1}^n w_i (y_i - \hat{f}_\lambda(x_i))^2}{(1 - \mathcal{C}\,\mathrm{tr}(\mathbf{WA}(\lambda))/n)^2}, \quad (12)$$

where $\mathbf{W}$ is a diagonal matrix with elements $\mathbf{W}_{ii} = w_i$.

3. LOCAL CROSS-VALIDATION

The idea in this work is to construct an estimate of the pointwise risk (5) by a weighted version of the expected squared error that concentrates the weight on points in a neighborhood of x_i. To specify the weights one could use any nonparametric regression method. With linear estimators of the form (3), the rows of the smoothing matrix provide the weights. For example let $w_j = \mathbf{A}_{ij}(\lambda_G)$ where λ_G is some choice for a global smoothing parameter. The local cross-validation function for design point X_i is defined by

$$\mathrm{LCV}_i(\lambda, \mathcal{C}) = \frac{\sum_{j=1}^n \mathbf{A}_{ij}(\lambda_G)(y_i - \hat{f}_\lambda(x_i))^2}{(1 - \mathcal{C} m_i(\lambda))^2}, \quad (13)$$

where $m_i(\lambda) = \sum_{j=1}^n \mathbf{A}_{ij}(\lambda_G)\mathbf{A}_{jj}(\lambda)$. The estimate of the local smoothing parameter at x_i, $\hat{\lambda}_i$, is obtained by minimizing this function. (It may be useful to note that (13) is just $(1/n)\sum_j \{[w_j(y_i - \hat{f}_\lambda(x_i))^2]/(1 - \mathcal{C}\sum_j w_j \mathbf{A}_{jj})^2\}$ where $w_j = \mathbf{A}_{ij}$ and this is done for each design point x_i. The term $\sum_j w_j$ corresponds to tr($\mathbf{WA}$) in (12) where the weights reflect the distances from the fixed design point x_i.)

One can interpret this criterion as a nonparametric regression estimate of $\mathrm{ESE}_i(\lambda)$. If one squares the cross-validated residuals from (7), then one obtains an unbiased but inconsistent estimate of $\mathrm{ESE}_i(\lambda) + \sigma^2$. To obtain a consistent estimate, one uses a local average of cross-validated residuals for observations in the neighborhood of x_i. This operation is precisely the result when one considers a weighted average based on the

ith row of the smoothing matrix. In this formula, it is important to distinguish between the global smoothing parameter λ_G (chosen by GCV) used to estimate the local cross-validation function and the value of the local smoothing parameter that will be used to estimate f at x_i. The choice of λ_G is an important smoothing parameter in this method. Although one might consider estimating λ_G by GCV for each value of x_i, we use the global value determined from the original fit to the data. It is our opinion that the structure in the ESE curve as a function of x_i should have about as much structure as f. Thus it is reasonable to use the same amount of smoothing applied to the squared residuals.

3.1 Computation of Local Smoothing Parameters

To compute the local smoothing parameter, one can substantially reduce the computational burden by tabulating results for a grid of λ values. Provided that sufficient array storage is available, the computational burden to estimate local estimates will only be marginally more than just finding the global estimate.

Let ρ_ν, $1 \le \nu \le M$, form a grid of smoothing parameter values and define the $n \times M$ matrices $\mathbf{F}$ and $\mathbf{D}$ where $\mathbf{F}_{i,\nu} = \hat{f}_{\rho,\nu}(x_i)$ and $\mathbf{D}_{i,\nu} = \mathbf{A}_{ii}(\rho_\nu)$. These components are intermediate results in computing $\mathrm{GCV}(\lambda, \mathcal{C})$ and thus are necessary even for the global estimator. To calculate LCV_i for the grid of smoothing parameters, it is just a matter of smoothing the squared residuals and diagonal elements for each ρ_ν. For example, let $u_i = (Y_i - \mathbf{F}_{i,\nu})^2$ and $V = \mathbf{A}(\lambda)u$. Now $\sum_j v_{ij}u_i$ will be the numerator in the definition of $\mathrm{LCV}_i(\rho_\nu)$. Thus, to table the values $\mathrm{LCV}_i(\rho_\nu)$ it is necessary to calculate $2M$ additional smooths of length n. For smoothing splines and many other methods, this is an efficient operation. For splines, computing the diagonal elements of $\mathbf{A}$ in order to compute the global estimate of λ dominates the computation time. The extra smooths associated with the local cross-validation functions are a modest fraction of the computational cost.

3.2 The MLCV Modification

Simulation studies suggest that a slight modification of the local smoothing parameter estimates improves the properties of the confidence intervals. The local smoothing parameter is taken to be the minimum of $\hat{\lambda}_i$ and the global estimate, $\hat{\lambda}_G$. For clarity, denote this modified estimate as $\bar{\lambda}_i$. The rationale for this modification is that one expects the local smoothing to selectively decrease the bias at points where the global choice oversmooths. A local value that is larger than $\hat{\lambda}_G$ suggests that the ESE can be decreased by smoothing more. This effect only serves to decrease the width of the resulting pointwise confidence intervals and will not improve its level. In fact, it is our experience that little reduction in the squared error is obtained, and the coverage is significantly degraded.

In summary, a criterion is proposed for determining local smoothing parameters based on cross-validation. The two tuning parameters in this approach are λ_G, which dictates the degree of smoothing to be applied to the cross-validated residuals, and $\mathcal{C}$, which controls the maximum number of effective parameters. This latter modification controls spurious undersmoothing. The resulting method is denoted LCV. Variability that results in oversmoothing is controlled by truncating at the value of the global smoothing parameter (thus, $\bar{\lambda}_i = \min\{\hat{\lambda}_i, \hat{\lambda}_G\}$). The resulting method is denoted MLCV.

4. POINTWISE CONFIDENCE INTERVALS

Wahba (1983) first suggested that pointwise confidence intervals I_λ for the regression curve $f(x)$ at x_i could be constructed using the GCV smoothing spline estimator $\hat{f}_\lambda$ in the form

$$\hat{f}_{\hat{\lambda}}(x_i) \pm z_{\alpha/2}\sqrt{\hat{\sigma}^2 \mathbf{A}_{ii}(\hat{\lambda})} \tag{14}$$

where $\hat{\lambda}$ is the value of λ that minimizes $\mathrm{GCV}(\lambda)$ and $\hat{\sigma}^2$ is the estimator of σ^2 from the smoothing spline estimator $\hat{f}_\lambda$ given by $\mathrm{RSS}/(n - \mathrm{tr}\,\mathbf{A}(\hat{\lambda}))$ where RSS is the residual sum of squares from the estimate. As an aside, the term $\hat{\sigma}^2\mathbf{A}_{ii}$ results from some nice simplifications for splines in terms of expressing the posterior variance of the estimate as a simple function of the smoothing matrix $\mathbf{A}$. For other smoothers what should be included here is the posterior variance for the estimator.

Through simulation results, Wahba found that the average coverage probability, $(1/n)\sum_i \Pr[f(x_i) \in I_{\hat{\lambda}}]$, is close to $1-\alpha$. This result led to the conjecture by Wahba that $\hat{\sigma}^2 \mathrm{tr}\,\mathbf{A}(\hat{\lambda})/n$ was related to $\mathrm{EASE}(\lambda)$.

Nychka (1988) proved this conjecture of Wahba's to be true and a more detailed discussion of these GCV pointwise confidence intervals is given therein. The basic idea for the current approach is to apply the strategy for confidence intervals based on a global smoothing parameter to estimates where the smoothing parameter adapts to the local curvature of the function. Accordingly, the pointwise intervals have the same form as (14) except the estimate and standard error are evaluated at the local smoothing parameter $\hat{\lambda}_i$ rather than $\hat{\lambda}_G$. If the quantity $\hat{\sigma}^2\mathbf{A}_{ii}(\hat{\lambda}_i)$ is close to the pointwise expected squared error then one would expect the confidence interval to hold its level uniformly at all points.

5. SIMULATION STUDY

A Monte Carlo simulation study was performed to compare the performances of the globally cross-validated (GCV) smoothing spline estimator, the locally cross-validated (LCV) smoothing spline estimator, and the modified locally cross-validated (MLCV) smoothing spline estimator. Under a variety of settings, these three estimators and the resulting confidence intervals are compared pointwise with respect to mean squared error, coverage probability, uniformity of coverage across design points, bias, and width.

5.1 Design

The design of this experiment followed a $6 \times 2 \times 2 \times 3$ factorial where the factors are the regression function (see Fig. 2), noise level ($\sigma = \sqrt{.05}, \sqrt{.20}$) sample size ($n = 100, 200$), and the analysis method (GCV, LCV, MLCV). For each simulated dataset, response variables y_i were constructed by adding pseudorandom normal error e_i to a test function $f(x_i)$ at equally spaced design points x_i. Each cell combination of factor levels was simulated 500 times.

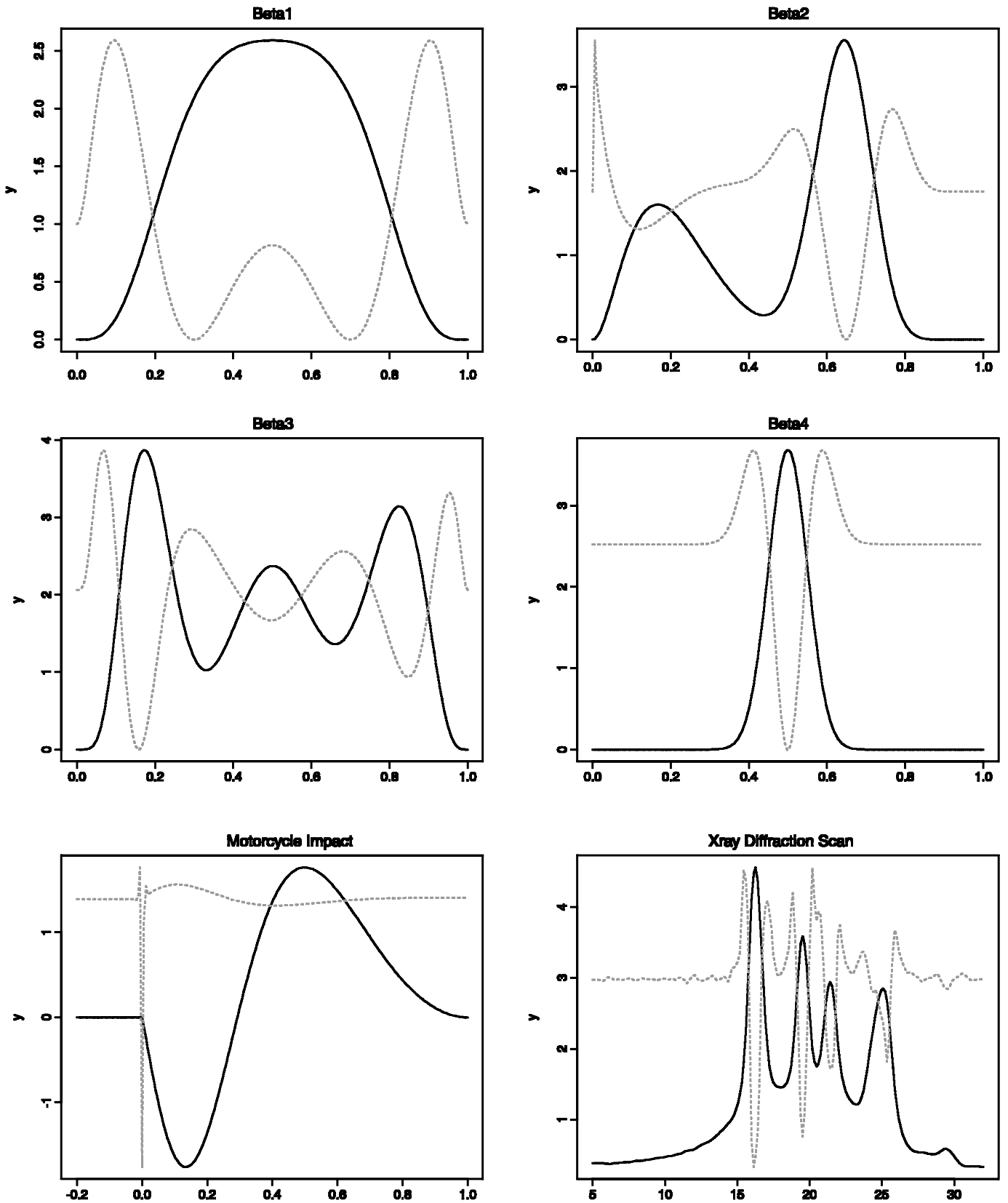

Figure 2. A Suite of Test Functions. The solid line gives the function, the dotted line gives the smoothing spline estimate of the second derivative, rescaled to fit on the same plot. The Motorcycle function violates the cubic spline smoothness assumptions.

The six test functions $f_t(x), t = 1, \ldots, 6$, are shown in Figure 2. The solid line shows the function; the dotted line shows the scaled second derivative.

It should be noted that Nychka (1995) shows that for cubic smoothing splines the term for bias is dominated by the fourth derivative. The reason Figure 2 shows the second derivative is that the results in Nychka (1995) are valid for large samples only. For small samples, lower order terms involving the second derivative are not negligible and in fact usually dominate the expression. Thus in small samples, the second derivative is a better indicator of the places where $\hat{f}$ has large bias.

In Figure 2, the second derivatives have been rescaled to be overlaid with the function values. The actual magnitudes of the second derivatives are not shown. The first four functions are proportional to mixtures of beta densities,

$$f_1(x) = \frac{1}{3}\beta_{10,5}(x) + \frac{1}{3}\beta_{7,7}(x) + \frac{1}{3}\beta_{5,10}(x),$$
$$f_2(x) = \frac{6}{10}\beta_{30,17}(x) + \frac{4}{10}\beta_{3,11}(x),$$
$$f_3(x) = \frac{1}{3}\beta_{20,5}(x) + \frac{1}{3}\beta_{12,12}(x) + \frac{1}{3}\beta_{7,30}(x),$$
$$f_4(x) = \beta_{50,50}(x),$$

where $\beta_{p,q}(x)$ is a beta density function with parameters p and q and $x \in [0, 1]$.

The Motorcycle function is taken from a model of motorcycle impact data in Friedman and Silverman (1989), given here by $f_5(x) = I_{x>0} * \sin\{2\pi * (1-x)^2\}$, where $x \in [-.2, 1]$ and $I_{x>0}$ is 1 for positive x and 0 elsewhere. This function violates the cubic smoothing spline assumptions since at $x = 0$ the second derivative is infinite. It is included to represent a more difficult smoothing problem. The function as presented by Friedman also had nonconstant variance, but that feature has been removed. The methods presented here can be generalized for heterogeneous variance, outliers, nonuniform designs, and so forth.

The sixth function comes from an X-ray diffraction scan, provided by Dr. Perry Haaland of Beckton Dickenson Research Center. Since these are observed data the true "X-ray" function was taken to be the observed data, and triple-smoothed with a running median smoother. Finally, all of the test functions were rescaled to have unit variance so that the same noise levels could be applied to all the test functions and retain a constant signal to noise ratio.

Table 1 gives roughness measures of the six test functions. By way of reminder and to give a reference, a straight-line function is placed in the table in addition to the test functions. Column 2 shows the trace of the smoothing matrix with the optimal value of λ for the case of $n = 200$ and the low level of noise. Column 3 shows the maximum of the absolute value of the second derivative and column four shows the integrated squared second derivative. Table 1 shows dramatically that although all the test functions are scaled to unit variance, they differ widely in degree of curvature and thus the amount of difficulty they will present as smoothing problems.

Table 1. Roughness Measures for the Six Test Functions, With Straight Line Added as a Baseline

Function	Trace	Maximum absolute 2nd derivatives	Integrated squared 2nd derivatives
Straight Line	2.0	0	0
Beta1	11.4	98.8	68.0
Beta2	18.9	714.7	1,447.9
Beta3	22.1	1,009.4	3,687.5
Beta4	22.4	1,444.4	3,513.8
Motorcycle	17.5	∞	∞
X ray	53.7	15.5	162,650.4

5.2 Results

Three sets of tables are presented that show measures of performance for the three methods. For conservation of space, only the low and high levels of noise are shown. Results for the moderate levels of noise are in line with what one might interpolate from the two extreme levels. Table 2 shows average coverage probabilities of the confidence intervals. These are coverage probabilities averaged over the design points for the 500 Monte Carlo trials, consequently the GCV method competes quite well, consistent with Wahba's findings. The problem of uniformity of coverage is seen in results such as shown in Figure 1 in which the pointwise coverage is well above the $1 - \alpha$ level for most of the design points but dismally below the desired level for a few design points. The example in Figure 1 was the Beta2 test function with $n = 100$ and the high level of noise.

For the Beta2 test function, Table 2(b) shows a GCV average coverage of .952, which looks very good but an examination of the pointwise coverages, as seen in Figure 1, shows that the coverages ranged from as low as .824 to as high as .978. Table 3 shows the minimum and maximum pointwise coverages over the design points. This sheds light on the issue of uniformity of coverage, illustrated most dramatically for the Motorcycle test function.

Table 4 shows average widths of the confidence intervals. The table suggests that attaining better uniformity of coverage by local crossvalidation requires slightly wider confidence intervals.

To illustrate the appearance of the MLCV confidence intervals, Figure 3 shows in the top row the GCV and MLCV confidence intervals for one outcome of the Motorcycle function. This outcome was randomly chosen. Although the two estimates look similar, the area from $x = -.2$ to $x = 0$ shows an important difference. Although the true function is flat, the GCV results show a consistent downward trend. The MLCV results show a similar trend but with a bump just before $x = 0$. The bottom row of the figure shows the average of the confidence intervals and function estimates over the 500 Monte Carlo trials. The true function is shown as the dotted line. The mean GCV estimate is "rounding the corner" much more than the mean MLCV estimate. In the case of the GCV method the true function actually breaks into the mean upper confidence interval at $x = 0$, which illustrates the extremely low coverage probability for the GCV method at that design point.

Results such as in Table 3 can be illustrated graphically in several ways. Figure 4 shows the pointwise mean squared error for two of the $6 \times 2 \times 2 \times 3$ factorial, namely the Beta2 test function with $n = 100$ and both the low and high levels

Table 2. Coverage Results From 500 Monte Carlo Simulations

	$n = 100$			$n = 200$		
Test Function	GCV Coverage	LCV Coverage	MLCV Coverage	GCV Coverage	LCV Coverage	MLCV Coverage
			(a) Average coverage probabilities: low noise			
Beta1	.949	.901	.945	.960	.905	.950
Beta2	.956	.935	.955	.960	.931	.954
Beta3	.959	.943	.957	.962	.944	.958
Beta4	.956	.906	.957	.961	.892	.960
Cycle	.946	.925	.953	.949	.921	.955
X ray	.954	.930	.960	.962	.926	.967
			(b) Average coverage probabilities: high noise			
Beta1	.931	.870	.932	.951	.871	.943
Beta2	.952	.927	.951	.957	.923	.952
Beta3	.955	.921	.954	.960	.931	.955
Beta4	.951	.910	.954	.957	.896	.957
Cycle	.944	.912	.948	.946	.908	.950
X ray	.947	.925	.959	.957	.923	.964

of noise. The top row of plots in Figure 4 shows this result for the GCV method, the middle row shows the LCV method and the bottom row shows the MLCV method. Where the goal is a flat level of pointwise MSE across the design space, the MLCV method is arguably the superior choice. The systematic deviations for the GCV method are particularly suggestive and this example is representative of the other test functions.

Another way to assess the degree of uniformity of coverage is with a *Breakout Plot* such as in Figure 5, which shows results for the Beta2 test function with $n = 100$ and the high level of noise. For each design point a tally was made of the number of times the true function at that design point was higher than the upper confidence interval at that design point. This is labeled "Breakouts Above." Likewise the tally of the number of times the true function fell below the lower confidence interval is shown in the plot, and labeled "Breakouts Below." The breakout plot is a diagnostic offering an in-depth graphical examination of the pointwise coverage probabilities, decomposing the failures according to whether they occurred at the upper or lower limits. The left column of Figure 5 shows this diagnostic for the GCV, LCV, and MLCV methods.

The right column of Figure 5 shows the standard deviation of pointwise coverages across the 500 trials; for both diagnostics a flat line is the ideal. Both of the diagnostics in Figure 5 clearly indicate the superiority of the MLCV method for giving uniformity of coverage across the design space. The example chosen for Figure 5 is representative of the results for the other test functions.

Table 3. Pointwise Coverage Results From 500 Monte Carlo Simulations

		GCV		LCV		MLCV	
Test Function	n	Min	Max	Min	Max	Min	Max
			(a) Minimum and maximum coverage probabilities: low noise				
Beta1	100	.868	.982	.810	.962	.902	.968
Beta1	200	.908	.982	.818	.956	.920	.974
Beta2	100	.856	.980	.838	.966	.914	.972
Beta2	200	.842	.990	.808	.970	.914	.978
Beta3	100	.858	.984	.902	.974	.918	.980
Beta3	200	.840	.992	.908	.976	.910	.984
Beta4	100	.796	.988	.772	.972	.922	.974
Beta4	200	.786	.988	.784	.978	.900	.978
Cycle	100	.380	.982	.770	.968	.770	.974
Cycle	200	.168	.990	.670	.972	.670	.978
X ray	100	.708	.986	.734	.980	.904	.984
X ray	200	.736	.990	.774	.980	.902	.990
			(b) Minimum and maximum coverage probabilities: high noise				
Beta1	100	.824	.984	.774	.954	.874	.970
Beta1	200	.898	.976	.712	.950	.882	.966
Beta2	100	.824	.978	.800	.964	.912	.976
Beta2	200	.828	.988	.742	.968	.914	.980
Beta3	100	.850	.982	.804	.972	.902	.976
Beta3	200	.840	.988	.870	.972	.892	.980
Beta4	100	.764	.984	.770	.970	.908	.974
Beta4	200	.762	.988	.742	.974	.890	.978
Cycle	100	.500	.986	.784	.958	.782	.972
Cycle	200	.288	.986	.718	.960	.718	.980
X ray	100	.572	.990	.808	.974	.862	.988
X ray	200	.664	.990	.728	.984	.892	.984

Table 4. Width Results From 500 Monte Carlo Simulations

	$n = 100$			$n = 200$		
Test Function	GCV Width	LCV Width	MLCV Width	GCV Width	LCV Width	MLCV Width
			(a) Average width of confidence intervals: low noise			
Beta1	.289	.343	.369	.214	.263	.283
Beta2	.361	.400	.430	.266	.304	.328
Beta3	.402	.439	.465	.296	.335	.355
Beta4	.405	.389	.461	.298	.294	.351
Cycle	.344	.379	.416	.258	.290	.322
X ray	.619	.528	.653	.458	.400	.496
			(b) Average width of confidence intervals: high noise			
Beta1	.511	.625	.680	.388	.483	.529
Beta2	.657	.746	.805	.485	.567	.615
Beta3	.730	.817	.870	.541	.627	.666
Beta4	.732	.716	.856	.543	.540	.654
Cycle	.606	.701	.763	.453	.534	.589
X ray	1.068	.947	1.166	.811	.720	.900

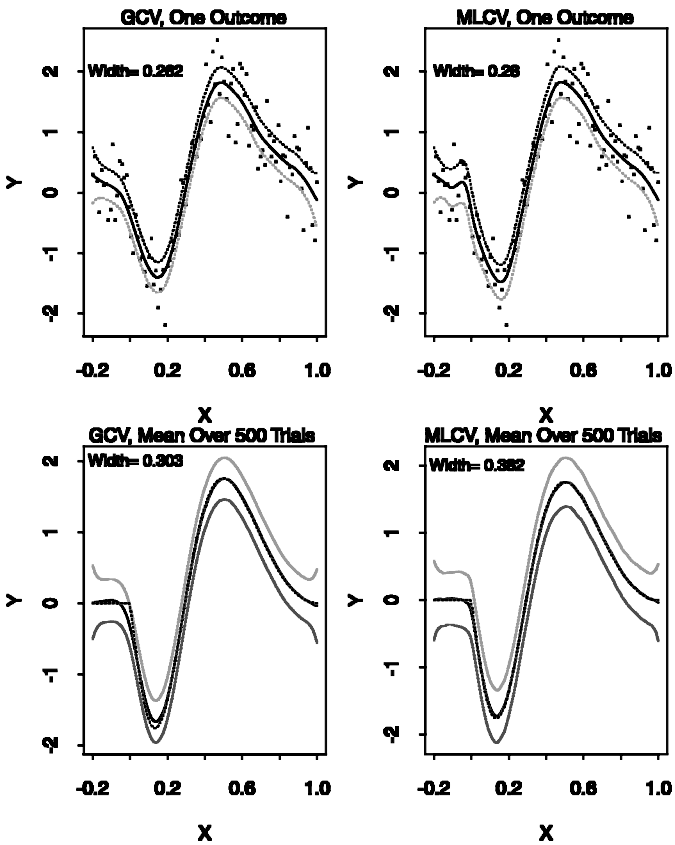

Figure 3. Motorcycle Impact Function, GCV Versus MLCV Methods. Top row shows results for a single Monte Carlo trial (chosen at random). Bottom row shows mean intervals and function estimates over all 500 Monte Carlo trials. Dotted line is the true function.

6. DISCUSSION

Coverage level, width and uniformity of coverage are multiple criteria that should be considered when comparing different methods for constructing confidence intervals. The MLCV method is competitive with the GCV method with respect to average coverage probability (Table 2). The superiority of MLCV with respect to pointwise coverage is clear; for example with the Beta4 test function with low noise and small sample size the GCV method had pointwise coverage as low as .796 but the MLCV method had pointwise coverage of .922 in the worst case (Table 3). The improved uniformity of coverage for the MLCV method comes at the cost of wider intervals (Table 4). Figures 4 and 5 are clear graphical diagnostics showing that MLCV generated confidence intervals have better uniformity of coverage across design points.

6.1 ANOVA Summary

As an aid in summarizing the simulation results, an ANOVA can be performed on the results where the factors are test function, noise level, sample size, and method of computing confidence intervals. The ANOVA sums of squares and LSMEANS can be used to diagnose the sources of variation, to answer questions such as whether some test functions are more challenging than others, and to do valid comparisons of the methods. Table 5 shows the ANOVA sums of squares for the experiment, taking average coverage probability as the response.

Table 5 shows that the sums of squares for *Method* is huge compared to the other effects, and the *Test Function* sums of squares is the next largest (followed closely by *Noise Level*). This leads naturally to an examination of plots of LSMEANS for these effects. Figure 6 shows the LSMEANS for each test function and each analysis method. This shows that the GCV and MLCV methods are competitive but the LCV method is consistently inferior.

Similarly an ANOVA can be performed where the width of the intervals is the response and LSMEANS plots examined. Figure 7 shows that GCV and LCV methods are competitive but the MLCV method produces consistently wider confidence intervals.

6.2 Valid Comparisons

When comparing two methods of computing confidence intervals, the method whose average coverage is closest to the desired level would seem to be the obvious favorite. If the two methods have the same average coverage, then one might compare the widths of the intervals, preferring the method that produces thinner intervals. If the two methods do not have the same average coverage, then it does not make sense to compare the widths of their intervals. Likewise, if two confidence interval methods have the same average coverage, then one may wish to compare some measure of the uniformity of coverage, favoring the method with coverage probabilities that are nearly constant across the design points. To express uniformity of coverage as a single number, we take the standard deviation of the estimated pointwise coverage probabilities. It should be noted that this is different from what was plotted in Figure 5, where for each design point the standard deviation was computed for a binomial proportion across 500 trials. Here we take the estimates of the pointwise coverage probabilities, which are computed for each design point from 500 Bernoulli trials, and compute the standard deviation of this vector of posterior means. The ideal smoother would have pointwise coverage probabilities of $1 - \alpha$ at each and every design point, resulting in a uniformity measure of 0. If the average (across design points) coverage probabilities for two methods are not identical then it does not make sense to compare uniformity of coverage for the one method versus the other.

One may feel that a method with slightly better coverage but much wider intervals may not be preferable, if the wider intervals seems too high a price to pay for the improved coverage. To put the methods on an equal footing for comparison we propose the concept of $Width^{1-\alpha}$ and $Uniformity^{1-\alpha}$. For the simulation study, let $\alpha = .05$ and consider $Width^{95}$ and $Uniformity^{95}$. The ANOVA model from Section 6.1 can be used to predict the width of the confidence intervals if the average coverage were exactly .95; this is $Width^{95}$. Similarly, one can predict the standard deviation of the pointwise coverage probabilities across the design points if the average coverage across the design points were exactly .95; this is $Uniformity^{95}$. LSMEANS for $Uniformity^{95}$ are shown in Figure 8.

From the simulation study results, the differences between the LSMEANS for *Width* and $Width^{95}$ are so subtle that

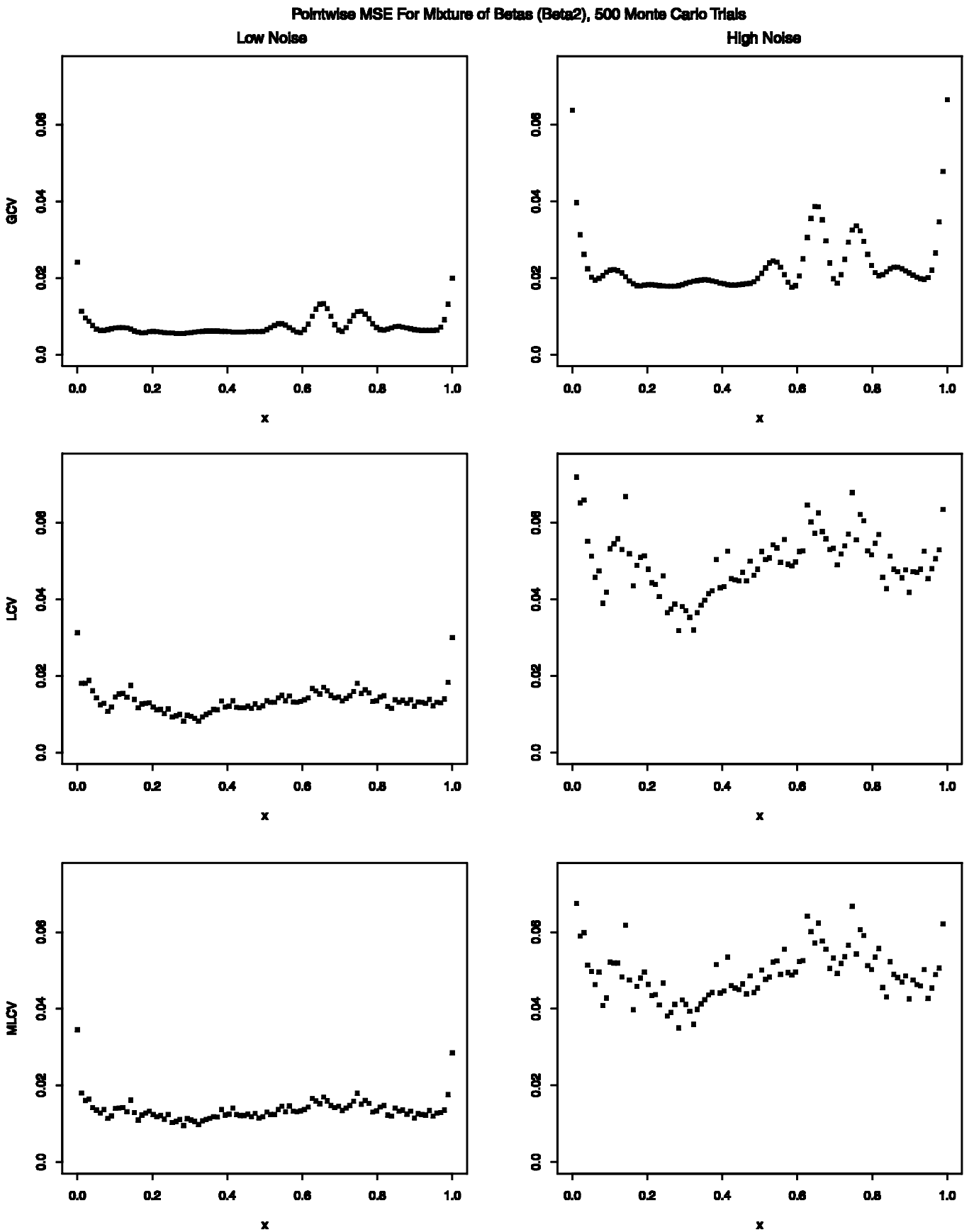

Figure 4. Pointwise Mean Squared Error Of Curve Estimates for Mixture of Betas (Beta2). Comparison of GCV, LCV, and MLCV methods in first, second, and third rows, respectively.

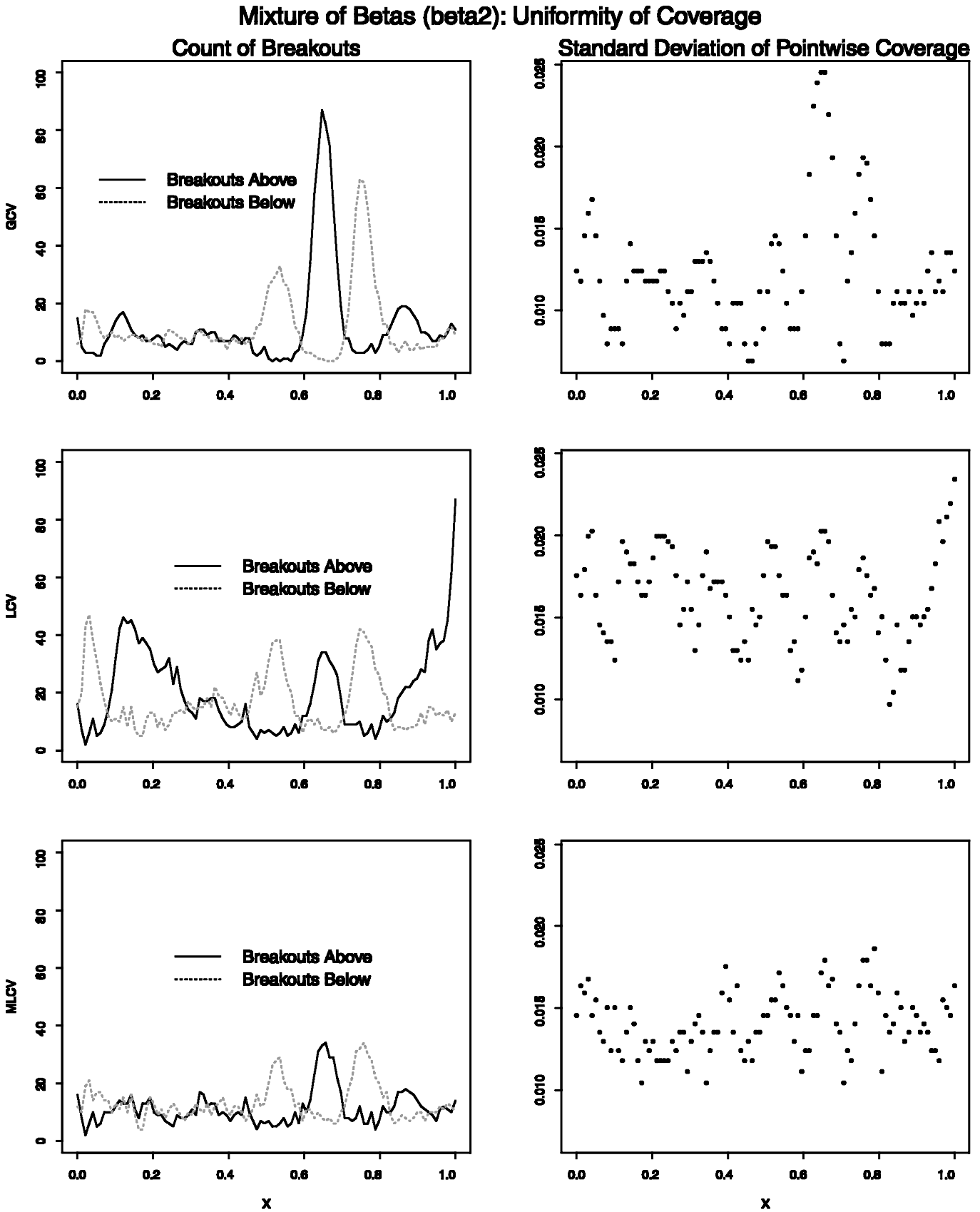

Figure 5. Uniformity of Coverage for Mixture of Betas (beta2). Comparison of GCV, LCV, and MLCV Methods.

Table 5. ANOVA Model for Coverage Probability Results From 500 Monte Carlo Simulations

ANOVA summary for analysis of coverage			
Source	*DF*	*Sum of Squares*	*F Statistic*
Method	2	.02300	736.30
Test Function	5	.00485	62.09
Noise Level	1	.00094	60.10
Sample Size	1	.00011	6.89
Method*Test Function	10	.00338	21.66
Method*Sample Size	2	.00029	9.20
Method*Noise Level	2	.00020	6.26
Test Function*Sample Size	5	.00021	2.72
Test Function*Noise Level	5	.00055	7.08

NOTE: *indicates interaction term.

LSMEANS plots of $Width^{95}$ look the same as LSMEANS plots of width (Fig. 7). This is because average coverage across the design points is not difficult to attain. However, lack of uniformity of coverage has dire effects on simultaneous inference, where the hypothesis is about the entire curve, making computing valid confidence bands a much more difficult problem. For example, for the motorcycle impact test function with high noise and sample size of 100, the GCV-based confidence intervals have average coverage of .944. This is not significantly different (at the .05 level) from .95, but Table 3b reveals that at some design points the coverage probability is as low as .50. This value occurs at the point of discontinuity of that function. Coverage probabilities on either side of that point are (.93, .86, .71, .50, .53, .79, and .94) and simultaneous bands based on the global GCV-driven smoothing parameter with a Bonferroni adjustment factor have simultaneous coverage probability of .892 for this test function which is significantly different from .95. A forthcoming paper by Cummins and Nychka will deal with the problem of confidence bands in-depth.

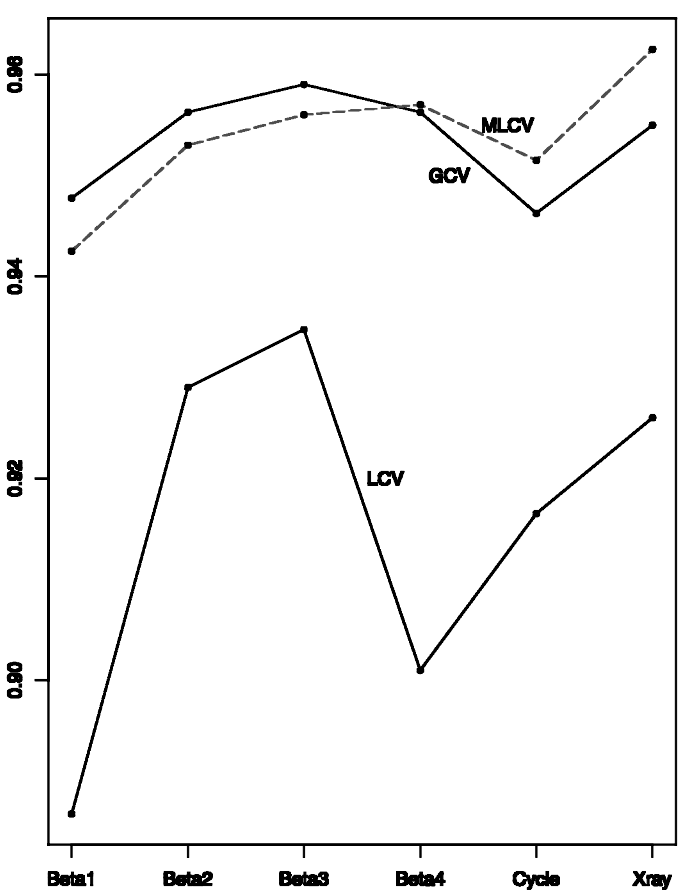

Figure 6. LSMEANS for Average Coverage by Test Function and Confidence Interval Method.

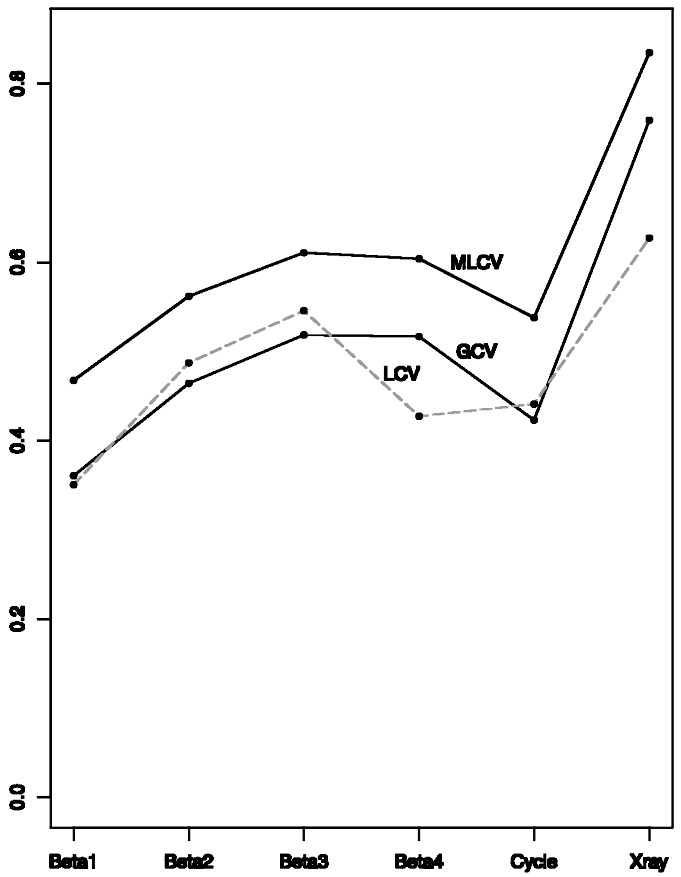

Figure 7. LSMEANS for Average Halfwidth of Confidence Intervals by Test Function and Confidence Interval Method.

It has been noted that $1-\alpha$ coverage is not the only desirable trait. The width of the confidence intervals and the uniformity of coverage should not be ignored. Which method is best in practical settings depends on what traits are considered most important. Here we offer a few specific scenarios as examples. For a given application, it may be more desirable to have a method which may give 90% rather than 95% coverage but has better uniformity of coverage, spreading the uncertainty uniformly throughout the design space. Alternatively, a compromise may be sought that results in confidence intervals that are wider than GCV-based intervals but thinner than MLCV-based intervals. This compromise may result in uniformity of coverage that is superior to the GCV-based but inferior to the MLCV-based confidence intervals.

Table 6 gives an example of such a compromise. To obtain thinner confidence intervals, the MLCV algorithm was modified by simply increasing the roughness penalty at the "inner" smoothing step where the squared residuals are smoothed. In this MLCV variant, a value of $\mathcal{C} = 20$ was used in (9) and the

Table 6. MLCV Variant with $\mathcal{C} = 20$: Pilot Study of Coverage, Half-width, and Uniformity From 500 Monte Carlo Simulations for Three Test Functions

Pilot study of MLCV variant with $\mathcal{C} = 20$: coverage, width, and uniformity comparisons									
	Beta2 test function			*Cycle test function*			*X ray test function*		
Method	*Coverage*	*Width*	*Uniformity*	*Coverage*	*Width*	*Uniformity*	*Coverage*	*Width*	*Uniformity*
GCV	.952	.329	.0324	.944	.303	.0725	.947	.534	.0600
MLCV Variant	.954	.349	.0225	.952	.325	.0432	.956	.554	.0351
MLCV	.950	.401	.0134	.947	.380	.0263	.958	.579	.0198

smoother was applied to the squared residuals in computing the local λ values in the MLCV algorithm. This was done in a small pilot study involving three of the test functions at the high level of noise and sample size of 100. The results gave confidence intervals that were thinner than what the standard MLCV produced but wider than what GCV produced. On the other hand, the resulting intervals had uniformity of coverage that was superior to GCV but inferior to MLCV. Table 6 presents the pilot study results. (For such a small study there is not enough data to build a model to compute $Width^{95}$ and $Uniformity^{95}$.) A value of $\mathcal{C} = 20$ is rather extreme and was chosen from an experimental design point of view. Our current hypothesis is that, for the types of biological data that we see in the pharmaceutical industry, a value of $\mathcal{C} = 2$ is a good default for the inner smoothing of the residuals, and a value of $\mathcal{C} = 1.2$ is a good default value for the purposes of estimating the global smoothing parameter λ_G. A forthcoming paper will explore this issue in more detail.

As a final comparison, Figure 9 places the results for the GCV method shown in Figure 1 side-by-side with the same results obtained for the MLCV method. The superior uniformity of coverage for the MLCV method clearly stems from its stabilizing effect on pointwise bias.

The performance results for the MLCV method should be viewed in a broad context, and are not specific solely to smoothing splines or cross-validation. They should hold for any smoothing procedure where the smoothing parameter is determined by a local loss criterion. In addition, the improved curve estimation advantage here is not a solely theoretical improvement such as providing an estimator that has a faster EASE convergence rate but shows its worth empirically

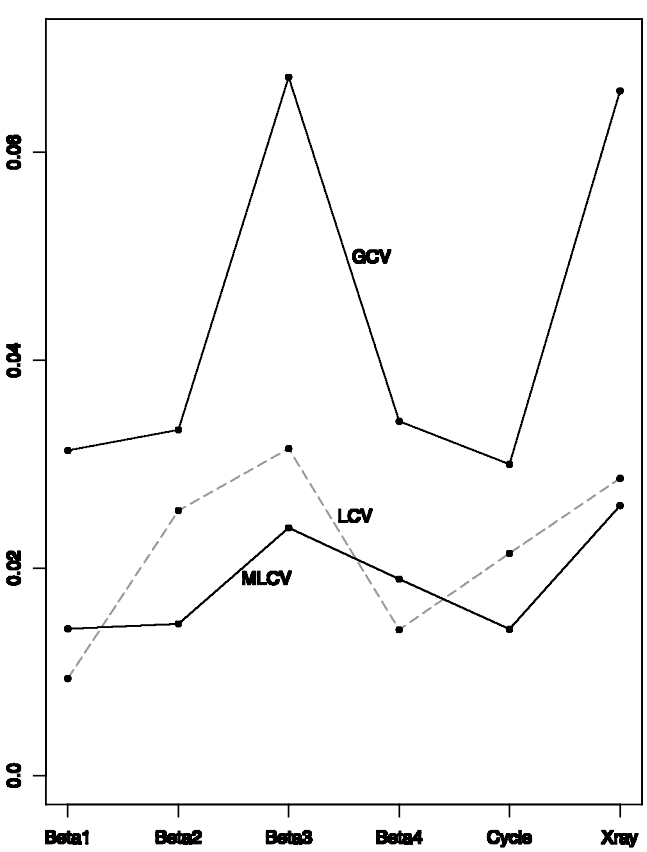

Figure 8. LSMEANS for $Uniformity^{95}$ by Test Function and Confidence Interval Method.

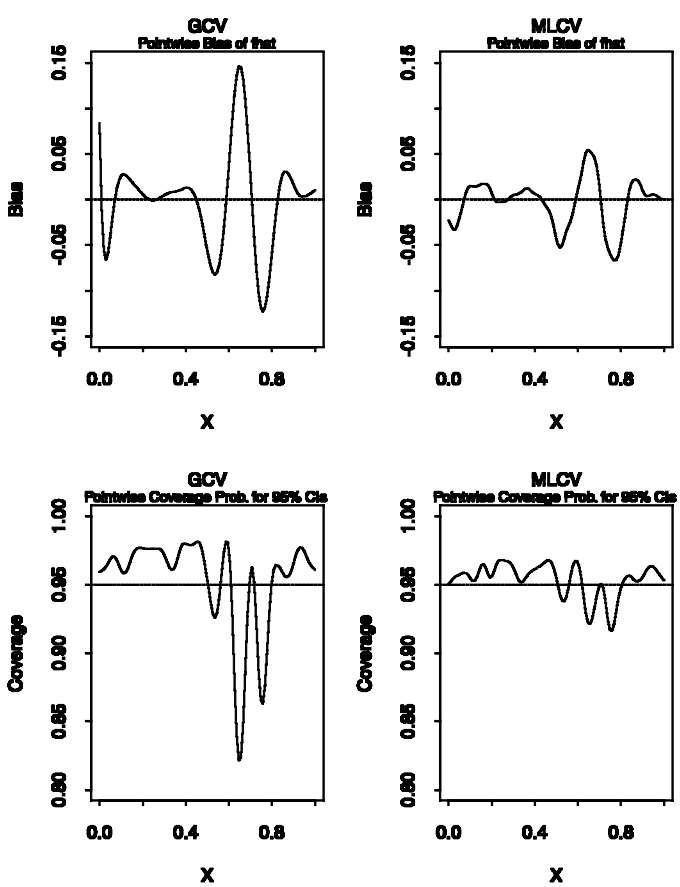

Figure 9. Bias and Pointwise Coverage Probability for the Beta2 Test Function, GCV Versus MLCV Methods.

by yielding improved coverage probabilities through reliable pointwise confidence intervals.

The MLCV method was shown to be superior to the standard use of generalized cross-validation for constructing confidence intervals. The simulation study explored six test functions at two levels of noise and two levels of sample size, and the results demonstrated that the MLCV method can give improved coverage at areas of relatively high curvature or rapid change in second derivative. The extension to more than one dimension is straightforward because of similarities of thin plate splines to the one-dimensional estimator. It is still an open question how effective local cross-validation ideas are in higher-dimensional contexts. Locally adaptive estimation is widely acknowledged as a difficult problem. While the MLCV method is an improvement, a perfectly functioning locally adaptive estimator remains an elusive goal. Although not emphasized in this paper, MLCV can be computed rapidly with little extra overhead compared to GCV. The encouraging simulation results suggest it should be incorporated as a routine tool for data analysis.

APPENDIX: EXPECTED VALUE OF COST-MODIFIED GCV (9)

This appendix evaluates the expected value of the modified GCV criterion. To simplify notation let $b^2 = (1/n)\sum_{i=1}^n b_i(\lambda)^2$ and $m_1 = (1/n)\mathrm{tr}\,\mathbf{A}(\lambda)$ and $m_2 = (1/n)\mathrm{tr}\,\mathbf{A}(\lambda)^2$. Then $(1/n)\sum_{i=1}^n V_i(\lambda) = \sigma^2 m_2$ and

$$E(\mathrm{GCV}(\lambda, \mathcal{C})) = \frac{b^2 + \sigma^2(1 - 2m_1 + m_2)}{(1 - \mathcal{C}m_1)^2}.$$

Theorem A.1 If $\mathcal{C}m_1 < 1$ and $(m_1/m_2) < \infty$ then

$$\frac{E(\mathrm{GCV}(\lambda, \mathcal{C})) - \sigma^2}{b^2 + \alpha\sigma^2 m_2} = 1 + O(m_1)$$

where $\alpha = 1 + 2(\mathcal{C} - 1)(m_1/m_2)$.

Proof. Using the fact that $(1 - \varepsilon)^{-2} = 1 + 2\varepsilon + o(\varepsilon^2)$ for $|\varepsilon| < 1$,

$$E(\mathrm{GCV}(\lambda, \mathcal{C})) = (b^2 + \sigma^2(1 - 2m_1 + m_2))(1 + 2\mathcal{C}m_1) + O(m_1^2)$$

and so

$$E(\mathrm{GCV}(\lambda, \mathcal{C})) - \sigma^2 - (b^2 + \alpha\sigma^2 m_2) = 2\sigma^2(\mathcal{C} - 1)m_1 + (1 - \alpha)\sigma^2 m_2 + 2\mathcal{C}m_1 b^2 + 2\sigma^2\mathcal{C}m_1(-2m_1 + m_2) + O(m_1^2)$$

By the choice of α the first two terms on the right-hand side cancel (because $\alpha\sigma^2 m_2 = \sigma^2[2\mathcal{C}m_1 - 2m_1 + m_2]$):

$$E(\mathrm{GCV}(\lambda, \mathcal{C})) - \sigma^2 - (b^2 + \alpha\sigma^2 m_2) = 2\mathcal{C}m_1 b^2 + 2\sigma^2\mathcal{C}m_1(-2m_1 + m_2) + O(m_1^2)$$

Dividing through by $b^2 + \alpha\sigma^2 m_2$ and changing the sign on $-2m_1$ to form an inequality gives

$$\left|\frac{E(\mathrm{GCV}(\lambda, \mathcal{C})) - \sigma^2}{b^2 + \alpha\sigma^2 m_2} - 1\right| < \frac{2\sigma^2\mathcal{C}m_1(2m_1 + m_2) + K_1 m_1^2}{b^2 + \alpha\sigma^2 m_2}$$

for some $K_1 < \infty$.

Dropping the b^2 in the denominator maintains the inequality, and collecting and canceling terms the right-hand side of the inequality becomes

$$m_1\left(K_2\left(\frac{m_1}{m_2}\right) + K_3\right)$$

for some $K_2, K_3 < \infty$. Based on this estimate and the hypotheses on (m_1/m_2), the theorem now follows.

[Received November 1999. Revised March 2000.]

REFERENCES

Brockman, M., Gasser, T., and Herrmann, E. (1993), "Locally Adaptive Bandwidth Choice for Kernel Regression Estimators," *Journal of the American Statistical Association*, 88, 1302–1309.

Chaudhuri, P., and Marron, J. S. (1999), "Sizer for Exploration of Structures in Curves," *Journal of the American Statistical Association*, 94, 807–823.

Craven, P., and Wahba, G. (1979), "Smoothing Noisy Data With Spline Functions," *Numerische Mathematik*, 31, 377–403.

Eubank, R. L., and Speckman, P. L. (1993), " Confidence Bands in Nonparametric Regression," *Journal of the American Statistical Association*, 88, 1287–1301.

Fan, J., Hall, P., Martin, M., and Patil, P. (1996), "On Local Smoothing of Nonparametric Curve Estimators," *Journal of the American Statistical Association*, 91, 258–266.

Filloon, T. G. (1990), *Improved Curve Estimation With Smoothing Splines Through Local Cross-Validation*, unpublished doctoral dissertation, North Carolina State University.

Friedman, J. H., and Silverman, B. W. (1989), "Flexible Parsimonious Smoothing and Additive Modeling," *Technometrics*, 31, 3–21.

Härdle, W., Hall, P., and Marron, J. S. (1988), "How far are Automatically Chosen Regression Smoothing Parameters From Their Optimum?" *Journal of the American Statistical Association*, 83, 86–95.

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

Müller, H. G., and Stadtmüller, U. (1987), "Variable Bandwidth Kernel Estimators of Regression Curves," *The Annals of Statistics*, 15, 182–201.

Nychka, D. (1988), "Bayesian Confidence Intervals for Smoothing Splines," *Journal of the American Statistical Association*, 83, 1134–1143.

——— (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.

——— (1995), "Splines as Local Smoothers," *The Annals of Statistics*, 23, 1175–1197.

Ruppert, D., and Carroll, R. J. (2000). "Spatially-adaptive Penalties for Spline Fitting," *Australian and New Zealand Journal of Statistics*, 42, 205–223.

Wahba, G. (1983), "Bayesian 'Confidence Intervals' for the Cross-validated Smoothing Spline," *Journal of the Royal Statistical Society, Ser. B*, 45, 133–150.

Wahba, G., and Wold, "A Completely Automatic French Curve: Fitting Spline Functions By Cross Validation," *Communications in Statistics*, 4, 1–17.