The Annals of Statistics
2001, Vol. 29, No. 2, 470–504

SELECTION CRITERIA FOR SCATTERPLOT SMOOTHERS

BY BRADLEY EFRON

Stanford University

Scatterplot smoothers estimate a regression function $y = f(x)$ by local averaging of the observed data points (x_i, y_i). In using a smoother, the statistician must choose a "window width," a crucial smoothing parameter that says just how locally the averaging is done. This paper concerns the data-based choice of a smoothing parameter for splinelike smoothers, focusing on the comparison of two popular methods, C_p and generalized maximum likelihood. The latter is the MLE within a normal-theory empirical Bayes model. We show that C_p is also maximum likelihood within a closely related nonnormal family, both methods being examples of a class of selection criteria. Each member of the class is the MLE within its own one-parameter curved exponential family. Exponential family theory facilitates a finite-sample nonasymptotic comparison of the criteria. In particular it explains the eccentric behavior of C_p, which even in favorable circumstances can easily select small window widths and wiggly estimates of $f(x)$. The theory leads to simple geometric pictures of both C_p and MLE that are valid whether or not one believes in the probability models.

1. Introduction. Curve fitting is an important statistical task, though not one that has always enjoyed a good scientific reputation. The traditional approach fits a polynomial function of x to the observed data points (x_i, y_i), perhaps a linear, quadratic or cubic curve, immediately raising the question of what is the appropriate degree polynomial to use on the problem at hand.

Scatterplot smoothers operate more locally than polynomial regression methods. For example, a simple smoother might employ local averaging, taking its value at x to be the average of those y_i values for which x_i is within some fixed "window width" λ of x. Choosing the window width is a similar problem, with similar difficulties, to choosing a polynomial degree.

This paper concerns the data-based selection of a smoothing parameter λ. Two selection criteria will be of particular interest: the C_p criterion [Mallows (1973)], which minimizes an unbiased estimate of prediction risk, and generalized maximum likelihood (GML) [Wecker and Ansley (1983), Wahba (1985)], a normal-theory empirical Bayes technique, defined somewhat differently here than in Wahba's work. C_p and GML look quite unlike each other but we will show that they are both maximum likelihood methods, carried out within two closely related curved exponential families, as defined in Efron (1975). These families are points in a continuum, each family of which suggests its own selection criterion.

Received October 1999; revised May 2000.

AMS 2000 *subject classifications*. 62F10, 62L08.

Key words and phrases. Empirical Bayes, C_p, GML, curved exponetial families, choice of smoothing parameter.

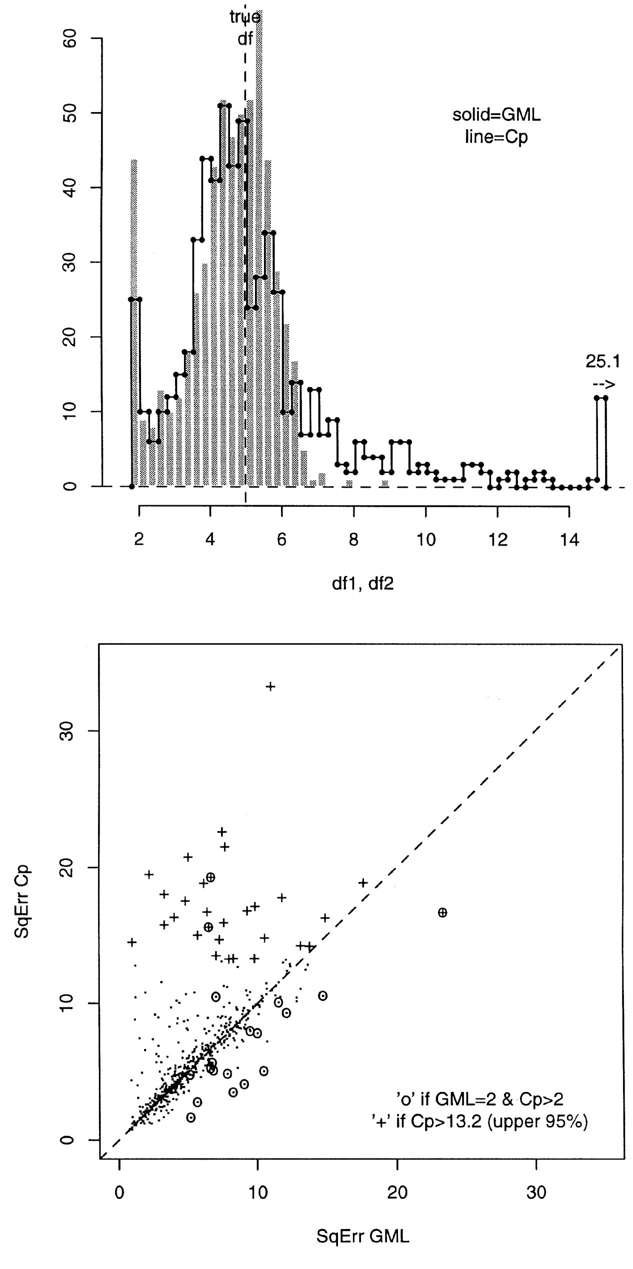

FIG. 1. *Top*: *Sampling experiment comparing GML* (*solid histogram*) *with* C_p (*dotted histogram, truncated at* 15); 600 *trials, true* $df = 5$; *from family in which GML is maximum likelihood. Bottom*: *Total squared error for estimating true regression*; *using GML* (*horizontal axis*) *versus* C_p (*vertical*); +*'s mark trials with* 30 *largest* C_p *df estimates*; *o's mark trials where GML* $\widehat{df} = 2$, C_p $\widehat{df} > 2$. *The sampling experiment is fully described in Section* 3.

Figure 1 shows the results of a sampling experiment comparing GML with C_p. In this experiment, which is fully described in Section 3, 600 data sets $\{(x_i, y_i), i = 1, 2, \ldots, 61\}$, with the x_i's fixed but the y_i's random, were drawn from a probability model in which the true degrees of freedom for regression was $df = 5$. (Roughly speaking, the analogy for polynomial curve fitting would prefer quartics for the true regression functions.) Then the GML and C_p methods were used to estimate the appropriate degrees of freedom for smoothing each data set. The resulting estimates varied from a minimum of two, indicating the prescription of a linear regression, to a maximum 25.1, corresponding to something more than a twenty fourth degree polynomial fit.

The probability model used in the left panel was the curved exponential family for which GML is actually the maximum likelihood estimate. We can see that GML is superior to C_p in this context, the latter having a tendency to produce occasional very large estimates. This tendency resulted in poor squared error estimation of the true regression function, as shown in the right panel. However the leftmost histogram spikes indicate a GML flaw, a greater tendency to oversmooth the data, going all the way to $\widehat{df} = 2$.

The C_p method, defined in Section 4, has a claim to be the most widely used selection criterion and is intimately related to other popular methods: Akaike's information criterion (AIC), Stein's unbiased risk estimate (SURE), and generalized cross-validation (GCV), as discussed in Section 4; see also Stein (1981) and Section 7 of Efron (1986). One of our main goals here is to explain the eccentric performance of C_p, which can give disappointing results even within its own maximum likelihood family. The explanation is given in terms of the geometry of estimation within curved exponential families. The geometry also helps explain the oversmoothing exhibited by both criteria (the tendency to select degrees of freedom smaller than the true value), particularly by GML.

There is a substantial literature on selection criteria for smoothers, most of which is written in a very general nonparametric large-sample asymptotic framework. Some good references include Hall and Johnstone (1992), Wahba (1985) and Chapter 4 of (1990), Eubank (1988) and Li (1986), who provides an impressively general demonstration of asymptotic optimality for C_p selection. This paper takes the opposite point of view, concentrating on parametric inference within finite samples, using splinelike linear smoothers. Our results, hopefully, compensate for their special context with a sharper delineation of the virtues and defects of the various selection criteria. Stein (1990) compares GML and C_p in a framework similar to Figure 1, with interesting results, two of which are quoted in Sections 5 and 10.

After a brief review of splinelike smoothers in Section 2, Section 3 develops the GML criterion as the MLE in a one-parameter curved exponential family, leading to the simple geometric description illustrated in Figure 2. The C_p estimate is similarly described in Section 4, with both C_p and GML belonging to a class of closely related selection criteria. Sections 5 and 6 use the exponential family framework to calculate useful properties of the estimates: standard errors, efficiencies, etc. It is shown that the C_p family's very large curvature destabilizes it as a point estimate. Sections 7–9 concern important aspects of the selection criterion problem, including its nonstandard behavior under repeated sampling. Section 10 discusses the close connection between errors in estimating the appropriate degrees of freedom, of the kind emphasized on the top of Figure 1, and the total squared error in estimating the true regression function, as on the bottom. Some detailed remarks are concentrated into Section 11, which ends with a brief summary of the paper's main ideas.

2. Splinelike smoothers.

We observe n points in the plane,

$$\{(x_i, y_i),\ i = 1, 2, \ldots, n\}, \tag{2.1}$$

and wish to estimate the regression of y on x, $f(x) = E\{y|x\}$, under the vague but important assumption that $f(x)$ is a smooth function of x. In this paper we will consider estimating the regression function only at the "design points" x_i, say $f_i = f(x_i)$, using a linear smoother

$$\hat{\mathbf{f}} = A_\lambda \mathbf{y}. \tag{2.2}$$

Here $\mathbf{y} = (y_1, y_2, \ldots, y_n)'$, $\hat{\mathbf{f}} = (\hat{f}_1, \hat{f}_2, \ldots, \hat{f}_n)'$ the vector of estimates of $\mathbf{f} = (f_1, f_2, \ldots, f_n)'$, and A_λ is an $n \times n$ smoothing matrix whose entries depend on the choice of a nonnegative smoothing parameter λ, as well as the x_i's. The n x_i values are assumed to be distinct in order to avoid some definitional difficulties.

To make our theory as neat as possible we will take the family of smoothing matrices $\{A_\lambda,\ \lambda \geq 0\}$ to be of the form

$$A_\lambda = U\mathbf{a}_\lambda U', \tag{2.3}$$

where U is an $n \times n$ orthogonal matrix, not depending on λ, and $\mathbf{a}_\lambda$ is the diagonal matrix with ith entry

$$a_{\lambda i} = \frac{1}{1 + \lambda k_i}, \qquad i = 1, 2, \ldots, n, \tag{2.4}$$

the constants k_i being a nonnegative, nondecreasing series. In some contexts it will be convenient to consider $\mathbf{a}_\lambda$ to be a vector rather than a diagonal matrix. Situation (2.3), (2.4) will be called "splinelike."

The use of genuine smoothing splines amounts to making a particular choice of U and $\mathbf{k} = (k_1, k_2, k_3, \ldots, k_n)$, this choice depending on $\mathbf{x} = (x_1, x_2, \ldots, x_n)'$ but not on $\mathbf{y}$. References include Green and Silverman (1994), Wahba (1990), Eubank (1988) and Hastie and Tibshirani (1990). For cubic smoothing splines, $\mathbf{k}$ is a nondecreasing sequence beginning with two zeros,

$$0 = k_1 = k_2 < k_3 < k_4 \cdots < k_n, \tag{2.5}$$

so that the first two eigenvalues $a_{\lambda 1}$ and $a_{\lambda 2}$ equal 1 for all λ. The first two columns of the eigenvector matrix U represent linear functions of $\mathbf{x}$, and the jth column behaves much like x_i^{j-1}. Roughly speaking, the smoother $A_\lambda \mathbf{y}$

preserves the part of $\mathbf{y}$ that is linear in $\mathbf{x}$, but shrinks the quadratic, cubic, quartic, etc. components toward zero, the shrinkage getting stronger for higher powers of x and also for larger values of λ.

The advantage of representation (2.3) for our purposes is that a single orthogonal transformation simultaneously diagonalizes all the matrices A_λ. Suppose we begin with the normal sampling model

$$\mathbf{y} \sim N(\mathbf{f}, \sigma^2 I), \tag{2.6}$$

σ^2 known. Then the transformations

$$\mathbf{z} = U'\mathbf{y}/\sigma, \qquad \mathbf{g} = U'\mathbf{f}/\sigma \quad \text{and} \quad \hat{\mathbf{g}}_\lambda = U'\hat{\mathbf{f}}_\lambda/\sigma \tag{2.7}$$

put the smoothing procedure (2.2) into diagonal form,

$$\mathbf{z} \sim N(\mathbf{g}, I) \quad \text{and} \quad \hat{\mathbf{g}}_\lambda = \mathbf{a}_\lambda \mathbf{z}. \tag{2.8}$$

The z_i's are independent unbiased estimates for the g_i's, but the scatterplot smoother uses biased estimates $\hat{g}_{\lambda i} = a_{\lambda i} z_i$ shrunk toward zero, more so for larger values of i and λ, in order to take advantage of the smoothness assumption.

The theory that follows does not depend on U and $\mathbf{k}$ being of the smoothing spline form, though we will assume so in the examples and illustrations. Remark D of Section 12 discusses departing from form (2.4) for the eigenvalues $a_{\lambda i}$. Hastie (1996) describes the construction of "pseudosplines," other families of smoothers of form (2.3).

The *degrees of freedom* of smoother (2.2) is defined to be the trace of A_λ. We will denote degrees of freedom by "ν,"

$$\begin{aligned} \nu &= \text{degrees of freedom} = \operatorname{tr}(A_\lambda) \\ &= \sum_{i=1}^{n} a_{\lambda i} = \sum_{i=1}^{n} \frac{1}{1 + \lambda k_i}; \end{aligned} \tag{2.9}$$

ν is a smoothly monotone function of λ, decreasing from $\nu = n$ at $\lambda = 0$ to $\nu = \#\{k_i = 0\}$ ($= 2$ for cubic smoothing splines) as $\lambda \to \infty$.

3. The generalized maximum likelihood criterion.

GML is a normal-theory empirical Bayes technique, credited by Wahba to Anderssen and Bloomfield (1974) and, specifically for smoothing splines, to Wecker and Ansley (1983). Stein (1990) also credits Patterson and Thompson (1971). Working in the $(\mathbf{g}, \mathbf{z})$ coordinate system (2.7), we assume the Bayesian model

$$\mathbf{g} \sim N(\mathbf{0}, \mathbf{c}_\lambda) \quad \text{and} \quad \mathbf{z}|\mathbf{g} \sim N(\mathbf{g}, I), \tag{3.1}$$

where $\mathbf{c}_\lambda$ is diagonal with ith element, say $c_{\lambda i}$. Defining

$$a_{\lambda i} = \frac{c_{\lambda i}}{c_{\lambda i} + 1} \quad \text{and} \quad b_{\lambda i} = 1 - a_{\lambda i}, \tag{3.2}$$

Bayes theorem allows us to reverse (3.1), obtaining

$$\mathbf{z} \sim N(\mathbf{0}, \mathbf{1}/\mathbf{b}_\lambda) \quad \text{and} \quad \mathbf{g}|\mathbf{z} \sim N(\mathbf{a}_\lambda \mathbf{z}, \mathbf{a}_\lambda). \tag{3.3}$$

Here $\mathbf{a}_\lambda$ and $\mathbf{1}/\mathbf{b}_\lambda$ indicate diagonal matrices with diagonal elements $a_{\lambda i}$ and $1/b_{\lambda i}$, respectively. The second relationship in (3.3) can be used to justify the linear shrinkage estimate $\hat{\mathbf{g}}_\lambda = \mathbf{a}_\lambda \mathbf{z}$ in (2.8): it is the Bayes a posteriori expectation $E_\lambda\{\mathbf{g}|\mathbf{z}\}$ starting from the prior covariance

$$\mathbf{c}_\lambda = \mathbf{a}_\lambda/(\mathbf{1} - \mathbf{a}_\lambda) = \mathrm{diag}(a_{\lambda i}/b_{\lambda i}) \tag{3.4}$$

in (3.1).

The first relationship in (3.3) motivates the GML estimate of λ,

$$\textbf{GML}\text{: } \hat{\lambda} = \mathrm{argmax}\{d_\lambda(\mathbf{z})\}, \tag{3.5}$$

$d_\lambda(\mathbf{z})$ indicating the density of $\mathbf{z}$ as a function of λ. In other words, $\hat{\lambda}$ is the maximum likelihood estimate of λ based on the marginal density, (integrating out $\mathbf{g}$) of $\mathbf{z} \sim N(\mathbf{0}, \mathbf{1}/\mathbf{b}_\lambda)$.

The minimal sufficient statistic for λ is

$$\mathbf{w} = \mathbf{z}^2 = (z_1^2, z_2^2, \dots, z_n^2). \tag{3.6}$$

The w_i are independently distributed as scaled chi-square random variables with one degree of freedom each,

$$w_i \overset{\text{ind}}{\sim} \chi_1^2 \,/\, b_{\lambda i}, \tag{3.7}$$

so that their joint density is

$$d_\lambda(\mathbf{w}) = e^{-\frac{1}{2}\sum(b_{\lambda i} w_i - \log b_{\lambda i})} d_o(\mathbf{w}), \tag{3.8}$$

with $d_o(\mathbf{w}) = 1/\Pi\,[\sqrt{2\pi w_i}\,]$ and $\hat{\lambda} = \mathrm{argmax}\{d_\lambda(\mathbf{w})\}$.

There is an important technical point to note here: in the smoothing spline situation (2.4) and (2.5), cases $i = 1, 2$ have $a_{\lambda i} = 1$, $b_{\lambda i} = 0$ for all λ, so these coordinates contain no information about λ. In defining $\hat{\lambda} = \mathrm{argmax}\{d_\lambda(\mathbf{w})\}$, the sum in the exponent of (3.8) is actually $\Sigma_{i=3}^n$. The same comment applies to all of our other selection criteria, but their computational formulas will turn out to automatically ignore cases $i = 1, 2$. In what follows all of the calculations refer to the $n - 2$-dimensional situation where the first two coordinates have been suppressed, unless specifically noted otherwise.

The family of densities $d_\lambda(\mathbf{w})$ is a one-parameter *curved exponential family*, in the terminology of Efron (1975, 1978). This means that we can write the density as

$$d_\lambda(\mathbf{w}) = e^{\boldsymbol{\eta}'_\lambda \mathbf{w} - \psi(\eta_\lambda)} d_o(\mathbf{w}), \tag{3.9}$$

where η_λ, the natural parameter vector, is a nonlinear function of λ, in this case,

$$\eta_\lambda = -\mathbf{b}_\lambda/2 = (\cdots - b_{\lambda i}/2 \cdots)'. \tag{3.10}$$

The cumulant generating function ψ can be written as

$$\psi(\eta_\lambda) = -\tfrac{1}{2}\sum \log(-\eta_{\lambda i}) \tag{3.11}$$

after shifting a constant factor into $d_o(\mathbf{w})$.

The maximum likelihood estimate (MLE) in a curved exponential family is determined by the expectation vector

$$\boldsymbol{\mu}_\lambda = E_\lambda\{\mathbf{w}\} = \mathbf{1}/\mathbf{b}_\lambda \tag{3.12}$$

and the derivative of $\boldsymbol{\eta}_\lambda$ with respect to λ,

$$\dot{\boldsymbol{\eta}}_\lambda = \left(\cdots \frac{\partial \eta_{\lambda i}}{\partial \lambda} \cdots\right)'. \tag{3.13}$$

The score function $\dot{l}_\lambda(\mathbf{w}) = (\partial/\partial\lambda)\log\{d_\lambda(\mathbf{w})\}$ is

$$\dot{l}_\lambda(\mathbf{w}) = \dot{\boldsymbol{\eta}}'_\lambda(\mathbf{w} - \boldsymbol{\mu}_\lambda), \tag{3.14}$$

so the MLE $\hat{\lambda}$ must satisfy $\dot{\boldsymbol{\eta}}'_{\hat{\lambda}}(\mathbf{w} - \boldsymbol{\mu}_{\hat{\lambda}}) = 0$; see Efron (1975, 1978).

Figure 2 diagrams the GML selection process as it applies to splinelike situations. In this case (2.4) and (3.2) show that the set of possible expectations $\{\boldsymbol{\mu}_\lambda = E_\lambda(\mathbf{w});\ 0 \le \lambda \le \infty\}$ is actually a straight line segment through the $n-2$-dimensional positive orthant, the sample space of $\mathbf{w} = (w_3, w_4, \dots, w_n)$,

$$\{\boldsymbol{\mu}_\lambda\} = \left\{\mathbf{1} + \frac{1}{\lambda}\,\frac{1}{\mathbf{k}};\ 0 \le \lambda \le \infty\right\}. \tag{3.15}$$

Intersecting the *line of expectations* at $\boldsymbol{\mu}_{\hat{\lambda}}$, orthogonally to $\dot{\boldsymbol{\eta}}_{\hat{\lambda}}$, is the flat space

$$\mathscr{L}_{\hat{\lambda}} = \{\mathbf{w}\colon \dot{\boldsymbol{\eta}}'_{\hat{\lambda}}(\mathbf{w} - \boldsymbol{\mu}_{\hat{\lambda}}) = 0\}, \tag{3.16}$$

the set of $\mathbf{w}$ vectors having $\dot{l}_\lambda(\mathbf{w}) = 0$ for λ equal to $\hat{\lambda}$. Solving for the GML estimate $\hat{\lambda}$ amounts to finding the $\mathscr{L}_\lambda$ containing $\mathbf{w}$. This is necessarily an iterative calculation since the orthogonals $\dot{\boldsymbol{\eta}}_\lambda$ change direction with λ [which

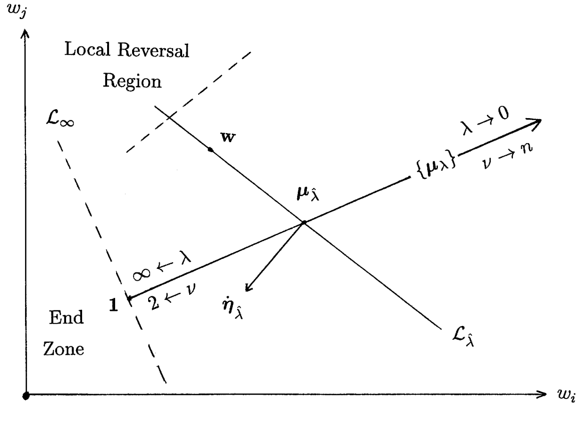

FIG. 2. *A diagram of the GML selection criterion as it applies to splinelike situations; heavy diagonal segment is the line of expectations* $\{\boldsymbol{\mu}_\lambda,\ 0 \le \lambda \le \infty\}$; $\mathscr{L}_{\hat{\lambda}}$ *is the flat surface of* $\mathbf{w}$ *vectors having* $\dot{l}_{\hat{\lambda}}(\mathbf{w}) = 0$; *it passes through* $\boldsymbol{\mu}_{\hat{\lambda}} = \mathbf{1}/\mathbf{b}_{\hat{\lambda}}$ *orthogonal to* $\dot{\boldsymbol{\eta}}_{\hat{\lambda}}$. *Other features in the diagram are explained in the text. For smoothing splines, the GML estimate* $\hat{\lambda}$ *is determined by the* $n-2$ *coordinates* $w_3, w_4, \dots, w_n$; *two coordinates* w_i, w_j *are indicated, with* $i < j$.

is what makes (3.9) a *curved* exponential family instead of a genuine one-parameter exponential family]. The other features in Figure 2 will be discussed later.

With the help of Figure 2 we can now describe the sampling experiment in Figure 1. The vector $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ was taken to be

$$\mathbf{x} = \left(-1, -1 + \frac{2}{60}, -1 + \frac{4}{60}, \ldots, 1\right), \tag{3.17}$$

$n = 61$ points equally spaced from -1 to 1, which determined U and $\mathbf{a}_\lambda$ in (2.3) according to the smoothing spline algorithm. The value λ_5 was found which made degrees of freedom equal 5,

$$\nu = \sum_{i=1}^{61} \frac{1}{1 + \lambda_5 k_i} = 5. \tag{3.18}$$

This determined $\boldsymbol{\mu}_{\lambda_5} = \mathbf{1}/\mathbf{b}_{\lambda_5}$, a point on the line of expectations in Figure 2.

Each of the 600 $\mathbf{w}$ vectors employed in Figure 1 comprised $w_i \overset{\text{ind}}{\sim} \chi_1^2/b_{\lambda_5 i}$ for $i = 1, 2, \ldots, 61$, as in (3.7); $\mathbf{w}$ determined the GML estimate $\hat{\lambda}$ according to the geometry of Figure 2, the C_p estimate of λ as described in Section 4 and finally the corresponding degrees-of-freedom estimates $\hat{\nu} = \sum_{i=1}^n a_{\hat{\lambda}i}$. The $\mathbf{w}$ vectors were actually generated as $\mathbf{w} = \mathbf{z}^2$, where $\mathbf{z} \sim N(\mathbf{0}, \mathbf{1}/\mathbf{b}_{\lambda_5})$ as in (3.3). Each $\mathbf{z}$ was also used to give a single realization $\mathbf{g}|\mathbf{z} \sim N(\mathbf{a}_{\lambda_5}\mathbf{z}, \mathbf{a}_{\lambda_5})$. The squared errors plotted in the right panel of Figure 1 were $\|\mathbf{a}_{\hat{\lambda}}\mathbf{z} - \mathbf{g}\|^2$, calculated separately for $\hat{\lambda}^{(GML)}$ and $\hat{\lambda}^{(C_p)}$.

In what follows the superscripts "1" and "2" will denote GML and C_p estimates, respectively, for example, $\hat{\nu}^{(1)}$ and $\hat{\nu}^{(2)}$ for the two estimates of degrees of freedom. The root mean square deviations from $\nu = 5$ in the left panel of Figure 1 were 1.32 for $\hat{\nu}^{(1)}$ and 2.70 for $\hat{\nu}^{(2)}$. The more robust measures of spread, (90th percentile–10th percentile)/$(2 \cdot 1.28)$, were 1.34 and 2.07, respectively. The superiority of the GML criterion is not surprising since it is the MLE in this sampling experiment. However Wahba (1985), as quoted by Stein (1990), comes to somewhat opposite conclusions, as discussed in Section 11. The discussion in Section 7 indicates that the results of Figure 1 do not much depend on the specific choice of $\mathbf{x}$ in (3.17).

Forty-four of the 600 vectors $\mathbf{w}$ in the trial fell into the "end zone," the region beyond the lower left end of the line of expectations in Figure 2 and gave GML estimates $\hat{\lambda}^{(1)} = \infty$, $\hat{\nu}^{(1)} = 2$, indicating the choice of a linear regression model. Nineteen of the 44 cases for which $\hat{\nu}^{(1)} = 2$ were "saved from the end zone" by the C_p criterion, as discussed further in Sections 4 and 10.

The geometry of Figure 2 applies to the GML selection criterion *whether or not one believes in the normal-theory sampling models (3.1) or (3.7)*. We will, for instance, use the geometry to calculate the biases of the GML criterion under the alternative sampling models of Section 4. The following discussion of influence functions provides another example.

Formula (2.4), $a_{\lambda i} = (1 + \lambda k_i)^{-1}$, yields simple expressions for derivatives with respect to λ, for example for $\mathbf{b}_\lambda = 1 - \mathbf{a}_\lambda$,

$$\dot{b}_{\lambda i} = a_{\lambda i} b_{\lambda i}/\lambda \quad \text{and} \quad \ddot{b}_{\lambda i} = -2a_{\lambda i} b_{\lambda i}^2/\lambda^2. \tag{3.19}$$

Since the natural parameter vector η_λ equals $-\mathbf{b}_\lambda/2$, (3.10), we have

$$\dot{\eta}_\lambda = -\dot{\mathbf{b}}_\lambda/2 = -\frac{1}{2\lambda}(\ldots, a_{\lambda i} b_{\lambda i}, \ldots)'. \tag{3.20}$$

The *influence function* of the MLE in a curved exponential family, that is, the gradient vector $\nabla_{\mathbf{w}}(\hat{\lambda}) = (\cdots \partial\hat{\lambda}/\partial w_i \cdots)'$, is

$$\nabla_{\mathbf{w}}(\hat{\lambda}) = [-\ddot{l}_{\hat{\lambda}}(\mathbf{w})]^{-1}\ \dot{\eta}_{\hat{\lambda}}. \tag{3.21}$$

We see that the influence of w_i on $\hat{\lambda}$ is proportional to $a_{\hat{\lambda} i} b_{\hat{\lambda} i}$, which is maximized at i having $a_{\hat{\lambda} i} b_{\hat{\lambda} i} = 0.5$. For λ_5 as above, $i = 5$ (the “quartic” term) maximized the influence, while the coordinates $i \geq 20$ had negligible influence.

Relationships (3.19) lead to simple expressions for the derivatives of $l_\lambda(\mathbf{w}) = \log(d_\lambda(\mathbf{w}))$, (3.8),

$$\begin{aligned} \dot{l}_\lambda(\mathbf{w}) &= -\frac{1}{2\lambda}\sum a_{\lambda i}(b_{\lambda i} w_i - 1) \quad \text{and} \\ \ddot{l}_\lambda(\mathbf{w}) &= -\frac{1}{2\lambda^2}\sum a_{\lambda i}[a_{\lambda i} - 2b_{\lambda i}(b_{\lambda i} w_i - 1)] \end{aligned} \tag{3.22}$$

(remembering that the sums are for i from 3 to n). We see that GML is the solution in λ of

$$\sum a_{\lambda i}(b_{\lambda i} w_i - 1) = 0, \tag{3.23}$$

which algebraically expresses the geometric solution seen in Figure 2.

All of these calculations assume that σ^2 in model (2.6), $\mathbf{y} \sim N(\mathbf{f}, \sigma^2 I)$, is a known quantity, so that we can calculate $\mathbf{z} = U'\mathbf{y}/\sigma$ in (2.7) and the sufficient vector $\mathbf{w} = \mathbf{z}^2$. Section 8 discusses the estimation of σ^2 when it is unknown and the effect on the accuracy of $\hat{\lambda}$. The example presented in Section 9 of Efron (1999) is more realistic in that the variances of the observations y_i change with x_i. Meanwhile, we will continue to assume σ^2 fixed and known, an assumption which permits sharper comparisons among the selection criteria.

Wahba’s definition of the GML criterion, as given in Section 4.8 of Wahba (1990), proceeds somewhat differently than ours. Let $\mathbf{v} = U'\mathbf{y}$, so $\mathbf{z} = \mathbf{v}/\sigma$. The log likelihood, as a function of both λ and σ^2, is seen to be

$$\log(d_{\lambda,\,\sigma^2}) = -\frac{1}{2}\sum\left[\frac{b_{\lambda i} v_i^2}{\sigma^2} - \log\left(\frac{b_{\lambda i}}{\sigma^2}\right)\right]. \tag{3.24}$$

Wahba jointly maximizes (3.24) over (λ, σ^2), and calls the resulting $\hat{\lambda}$ the GML estimate. In carrying out this maximization, the restricted MLE of σ^2 given λ,

$$\hat{\sigma}_\lambda^2 = \sum b_{\lambda i} v_i^2/(n-2), \tag{3.25}$$

can have a substantial effect on the profile likelihood $d_{\lambda,\,\hat{\sigma}^2_\lambda}$ and the resulting estimate $\hat{\lambda}$. The methods used in this paper employ an estimate $\hat{\sigma}^2$ that does *not* depend on λ; see Section 8.

4. A class of selection criteria. The GML estimate of λ is maximum likelihood in the curved exponential family (3.8). This section defines a class of curved exponential families, each of which gives rise to its own maximum likelihood selection criterion. Included in this class are the GML and C_p criteria.

The GML family is based on the scaled χ^2_1 distribution (3.7). Dropping subscripts, any one component $w \sim \chi^2_1/b$ has a one-parameter exponential family of densities

$$d_b(w) = e^{-\frac{1}{2}[bw-\log(b)]} d_o(w), \tag{4.1}$$

$d_o(w) = 1/\sqrt{2\pi w}$, defined for b and w positive.

We now replace the component densities (4.1) with a different one-parameter exponential family,

$$d_b^{(p)}(w) = e^{-c_o[b^p w-(p/(p-1))b^{p-1}]} d_o^{(p)}(w), \tag{4.2}$$

$p > 1$ a fixed constant, as is $c_o > 0$, both b and w positive. The limiting case as $p \to 1$ is the GML family (4.1); see remark C. In what follows we will use (4.2) instead of (4.1) to form a curved exponential family analogous to (3.8),

$$d_\lambda^{(p)}(\mathbf{w}) = \Pi_i d_{b_{\lambda i}}(\mathbf{w}). \tag{4.3}$$

Applying maximum likelihood estimation to family (4.3) gives MLE "$\hat{\lambda}^{(p)}$" for λ. We will show that $\hat{\lambda}^{(2)}$ is the C_p selection criterion, and that the limiting case as $p \to 1$, $\hat{\lambda}^{(1)}$, is the GML. The important point here is that C_p is also a maximum likelihood criterion, subject to the same kind of geometry seen in Figure 2.

Letting

$$\eta = -c_o b^p \qquad [b = (-\eta/c_o)^{1/p}], \tag{4.4}$$

we can write (4.2) in the standard exponential family form

$$d_\eta(w) = e^{\eta w - \psi(\eta)} d_o(w), \tag{4.5}$$

where η is the natural parameter and

$$\psi(\eta) = -\frac{c_o}{\alpha}\left(\frac{-\eta}{c_o}\right)^\alpha = \frac{c_o}{\alpha} b^{p-1} \quad \left[\alpha \equiv \frac{p-1}{p}\right] \tag{4.6}$$

is the normalizer, or cumulant generating function. Differentiating $\psi(\eta)$ with respect to η, twice, produces the expectation μ and variance V of w,

$$\mu = \frac{1}{b} \quad \text{and} \quad V = \frac{1}{c_o p \cdot b^{p+1}}. \tag{4.7}$$

Density (4.2) has expectation $1/b$ for all values of p, an important point discussed below, though the variance function depends on p. In fact the peculiar-looking density (4.2) is completely determined by three requirements: that w, a nonnegative random variable, is the sufficient statistic, that the natural parameter is proportional to b^p, and that $\mu = 1/b$. Under these requirements the family (4.2) is uniquely determined except for the choice of the positive constant c_o, which will be explored later. Interestingly enough, the "carrier density" $d_o^{(p)}(\mathbf{w})$ must be the positive stable law of order $\alpha = (p-1)/p$; see Remark A, Section 11.

We now proceed as before, letting λ produce $\mathbf{a}_\lambda$ according to (2.4), $\mathbf{b}_\lambda = \mathbf{1} - \mathbf{a}_\lambda$ and finally assuming that the components of $\mathbf{w} = \mathbf{z}^2$, defined as in (2.7), are independently distributed

$$w_i \overset{\text{ind}}{\sim} d_{b_{\lambda i}}^{(p)}(w_i). \tag{4.8}$$

This leads to the curved exponential family (4.3), henceforth called "$\mathscr{F}^{(p)}$",

$$\begin{aligned} \mathscr{F}^{(p)}:\ d_\lambda^{(p)}(\mathbf{w}) &= e^{-c_o \Sigma[b_{\lambda i}^p w_i - b_{\lambda i}^{p-1}/\alpha]} d_o^{(p)}(\mathbf{w}) \\ &\equiv e^{\boldsymbol{\eta}'_\lambda}\mathbf{w} - \psi^{(p)}(\boldsymbol{\eta}_\lambda) d_o^{(p)}(\mathbf{w}), \end{aligned} \tag{4.9}$$

$\alpha = (p-1)/p$, where now the natural parameter vector is

$$\boldsymbol{\eta}_\lambda = -c_o \mathbf{b}_\lambda^p. \tag{4.10}$$

Using (3.19), the crucial derivative vector $\dot{\boldsymbol{\eta}}_\lambda$ is

$$\dot{\boldsymbol{\eta}}_\lambda = -\frac{c_o p}{\lambda} \mathbf{a}_\lambda \mathbf{b}_\lambda^p = -\frac{c_o p}{\lambda}(\ldots, a_{\lambda i} b_{\lambda i}^p, \ldots)'. \tag{4.11}$$

We will use notation such as $\boldsymbol{\eta}_\lambda^{(p)}$ and $\dot{\boldsymbol{\eta}}_\lambda^{(p)}$ when necessary to distinguish different cases. The fact that $\dot{\boldsymbol{\eta}}_\lambda^{(p)} \to \dot{\boldsymbol{\eta}}_\lambda^{(1)}$ as $p \to 1$ shows that $\mathscr{F}^{(p)} \to \mathscr{F}^{(1)}$; see remark C.

The MLE $\hat{\lambda}^{(p)}$ within family $\mathscr{F}^{(p)}$ is the minimizer of the exponent in (4.9),

$$\hat{\lambda}^{(p)} = \underset{\lambda}{\operatorname{argmin}} \sum [b_{\lambda i}^p w_i - b_{\lambda i}^{p-1}/\alpha], \tag{4.12}$$

$\alpha = (p-1)/p$.

THEOREM 1. *$\hat{\lambda}^{(2)}$ is the C_p estimator of λ.*

PROOF. The C_p statistic

$$C_\lambda(\mathbf{y}) = \|\mathbf{y} - \hat{\mathbf{f}}_\lambda\|^2 + 2\sigma^2 \operatorname{tr}(A_\lambda) \tag{4.13}$$

is an unbiased estimator for the prediction error of the linear smoother (2.2), and in this context it is equivalent to Akaikés information criterion and Stein's unbiased risk estimate; see, for example, Section 7 of Efron (1986). The orthogonal transformations (2.7) and (2.8) give

$$\begin{aligned} C_\lambda(\mathbf{y}) &= \sigma^2\{\|\mathbf{z} - \hat{\mathbf{g}}_\lambda\|^2 + 2\Sigma a_{\lambda i}\} \\ &= \sigma^2\{\Sigma(1-a_{\lambda i})^2 z_i^2 + 2\Sigma a_{\lambda i}\} \\ &= \sigma^2 \Sigma(b_{\lambda i}^2 w_i - 2b_{\lambda i}) + 2n\sigma^2. \end{aligned}$$

We see that minimizing (4.13), that is, finding the C_p estimate of λ, is the same as minimizing $\Sigma(b_{\lambda i}^2 w_i - 2b_{\lambda i})$, which gives the MLE $\hat{\lambda}^{(2)}$ according to (4.12). □

The geometry of maximum likelihood estimation in $\mathscr{F}^{(p)}$ is very much like that seen in Figure 2. The sufficient statistic vector $\mathbf{w}$ is the same as before, as is the line of expectations, with the same expectation vector $\boldsymbol{\mu}_\lambda = \mathbf{1}/\mathbf{b}_\lambda$ corresponding to any particular choice of λ. However, there is one important difference: the vector $\dot{\eta}_\lambda^{(p)}$ is rotated counterclockwise relative to $\dot{\eta}_\lambda^{(1)}$, increasingly so as p increases. To see this, consider two coordinates i and j with $i < j$ as in Figure 2, so $b_{\lambda i} < b_{\lambda j}$ in the splinelike situation (2.4), (2.5). Then according to (4.11),

$$\text{(4.14)} \qquad \frac{\dot{\eta}_{\lambda j}^{(p)}/\dot{\eta}_{\lambda i}^{(p)}}{\dot{\eta}_{\lambda j}^{(1)}/\dot{\eta}_{\lambda i}^{(1)}} = \left(\frac{b_{\lambda j}}{b_{\lambda i}}\right)^{p-1},$$

which is an increasing function of p.

Figure 3 illustrates the situation for two coordinates $i < j$. For vectors $\mathbf{w}$ "above" the line of expectations it is easy to see that we obtain $\hat{\lambda}^{(p)} < \hat{\lambda}^{(1)}$ and $\hat{\nu}^{(p)} > \hat{\nu}^{(1)}$ and conversely for $\mathbf{w}$ "below." The quotes are a reminder that this is really an $n-2$-dimensional situation where above and below need to be defined more carefully, as in the reversal region discussion of Section 6. The occasional large values of the C_p estimator $\hat{\nu}^{(2)}$ seen in Figure 1 tended to come from vectors $\mathbf{w}$ cast far above the line of expectations by the long upper trail of the w_i distributions. As partial recompense for this bad behavior, the extra tilt of the $\mathscr{L}_\lambda^{(2)}$ surfaces "saved from the end zone" more than half of the $\mathbf{w}$ vectors falling there, as illustrated by $\mathbf{w}_2$ in Figure 3.

For any value of p, including the GML choice $p = 1$, the vector $\dot{\eta}_\lambda^{(p)}$ rotates counterclockwise as we move toward the $\lambda = 0$, $\nu = n$ end of the line of expectations, and moreover, the counterclockwise rotation is faster for bigger p values. Both of these results follow from (4.11) and (2.4), which for $i < j$ give

$$\text{(4.15)} \qquad -\frac{\partial}{\partial\lambda}\left(\frac{\dot{\eta}_{\lambda j}^{(p)}}{\dot{\eta}_{\lambda i}^{(p)}}\right) = \left(\frac{\dot{\eta}_{\lambda j}^{(p)}}{\dot{\eta}_{\lambda i}^{(p)}}\right)\frac{p+1}{\lambda}(b_{\lambda j} - b_{\lambda i}) > 0.$$

The increased speed of rotation degrades the performance of the C_p method ($p = 2$) as shown in the curvature discussion of Section 6.

The choice of the constant c_o in density $d_o^{(p)}(w)$, (4.2), affects $\text{var}_b^{(p)}(w) = (c_o p b^{p+1})^{-1}$, (4.7). Comparisons between the GML and C_p families $\mathscr{F}^{(1)}$ and

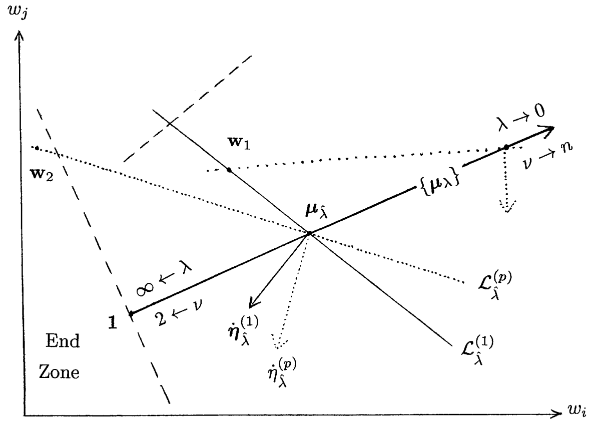
w_j
w_i
$\mathbf{w}_1$
$\mathbf{w}_2$
$\boldsymbol{\mu}_{\hat\lambda}$
$\{\mu_\lambda\}$
$\lambda \to 0$
$\nu \to n$
$\infty \leftarrow \lambda$
$2 \leftarrow \nu$
1
End
Zone
$\dot\eta^{(1)}_{\hat\lambda}$
$\dot\eta^{(p)}_{\hat\lambda}$
$\mathcal{L}^{(p)}_{\hat\lambda}$
$\mathcal{L}^{(1)}_{\hat\lambda}$

FIG. 3. *Geometrical relationship between selection criteria $\hat\lambda^{(1)}$ and $\hat\lambda^{(p)}$. Data vector $\mathbf{w}_1$ has $\hat\lambda^{(1)} = \hat\lambda$; $\mathbf{w}_2$ has $\hat\lambda^{(p)} = \hat\lambda$. Using $\hat\lambda^{(p)}$ instead of $\hat\lambda^{(1)}$ saves $\mathbf{w}_2$ from the end zone, but produces a very large estimate $\hat\nu$ from $\mathbf{w}_1$, as indicated by the uppermost dotted line. This illustration assumes $i < j$.*

$\mathscr{F}^{(2)}$ are more equitable if $\text{var}(w)$ is the same in both cases, which, since $\text{var}^{(1)}_b(w) = 2/b^2$ according to (4.1), is achieved by taking

$$c_o = 1/4b. \tag{4.16}$$

Some of our numerical results use $c_o = 3/8$ for $\mathscr{F}^{(2)}$ which equalizes the variances at $b = 2/3$, the value of b maximizing the $\mathscr{F}^{(2)}$ influence function ab^2, (4.11) or $c_o = 1/4$ which equalizes variances at the limiting value $b = 1$. Family $\mathscr{F}^{(1)}$ also allows a free choice of c_o, but, following (3.7), (4.1), we will always take $c_o = 1/2$ for $p = 1$.

Another interesting choice for c_o in $\mathscr{F}^{(2)}$ is

$$c_o = \Sigma_3^n a^2_{\lambda i}/(4 \cdot \Sigma a^2_{\lambda i} b_{\lambda i}), \tag{4.17}$$

which, as shown in Section 5, equalizes the Fisher information in $\mathscr{F}^{(1)}$ and $\mathscr{F}^{(2)}$ for estimating ν. For $\mathbf{x}$ as in (3.17) and $\nu = 5$ [i.e., $\lambda = \lambda_5$ (3.18)], (4.17) gives $c_o = 1.334$. Taken literally, (4.17) makes c_o depend on λ, which would destroy the exponential family structure (4.9), but we can still use it for a simulation comparison in which λ is fixed. The choice of c_0 is discussed further in remarks A and B.

The simulation in the bottom panel of Figure 4 draws vectors $\mathbf{w}$ from $d^{(2)}_{\lambda_5}(\mathbf{w})$, $c_o = 1.334$ in (4.9), with $\mathbf{x}$ and λ_5 as before, (3.17) and (3.18). In other words it is the same as the Figure 1 simulation, except with the $\mathbf{w} = \mathbf{z}^2$ draws based on family $\mathscr{F}^{(2)}$. The most noticeable feature is the downward bias of the GML estimator: $\text{Prob}\{\hat\nu^{(1)} < 5\} = 0.74$. The C_p estimator $\hat\nu^{(2)}$ performs better here than in Figure 1, as it should, but there are still hints of a long upper tail to its distribution. The curvature discussion of Section 6 provides some

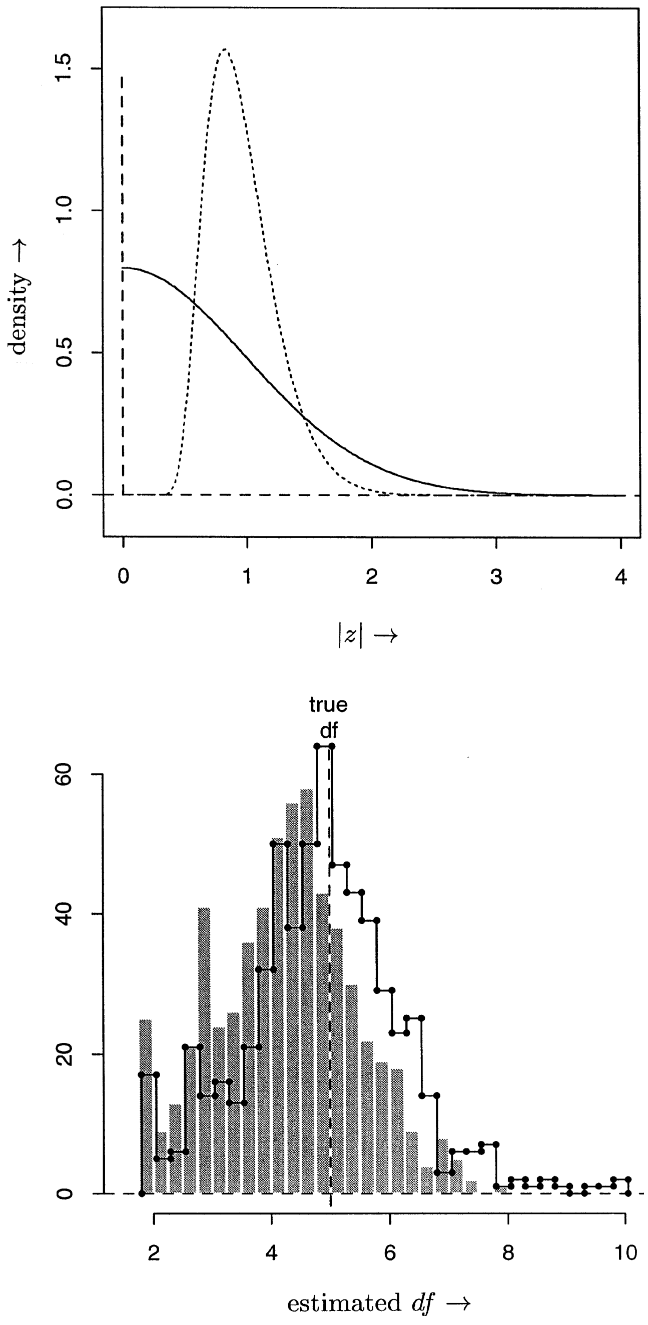

FIG. 4. *Top panel: Density of $|z| = w^{1/2}$ for GML family (solid curve) and C_p family (dashed curve); from $d_b^{(p)}(w)$ (4.2) with $b = 1$ and $p = 1, 2$; $c_0 = 1.334$ for $p = 2$. Bottom panel: Sampling experiment as in Figure* 1 *except that vectors* $\mathbf{w}$ *drawn from* $d_{\lambda_5}^{(2)}(\mathbf{w})$, *the* $\mathcal{F}^{(2)}$ *family with* $\nu = 5$; $c_0 = 1.334$.

explanation for C_p's erratic performance within its own maximum likelihood family $\mathscr{F}^{(2)}$.

A crucial aspect of definition (4.2) for the component densities $d_b^{(p)}(w)$ is that the expected value $\mu = E_b(w)$ equals $1/b$. Linear Bayes theory shows that this is a natural requirement. As a weaker version of the Bayesian assumptions (3.1), suppose that the components of $\mathbf{g}$ and $\mathbf{z}$ satisfy

$$g_i \sim (0, c_i) \quad \text{and} \quad z_i | g_i \sim (g_i, 1), \tag{4.18}$$

with $X \sim (u, v)$ indicating $E\{X\} = u$, $\mathrm{Var}\{X\} = v$. Using this same notation in bivariate form,

$$\begin{pmatrix} g_i \\ z_i \end{pmatrix} \sim \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} c_i \ , & c_i \\ c_i \ , & c_i + 1 \end{pmatrix} \right), \tag{4.19}$$

so that the best linear predictor of g_i is

$$\hat{g}_i = a_i z_i, \qquad a_i = \frac{c_i}{c_i + 1}. \tag{4.20}$$

Then $b_i = 1 - a_i = 1/(c_i + 1)$, so that marginally $z_i \sim (0, 1/b_i)$ and $E\{w_i\} = E\{z_i^2\} = 1/b_i$, as required by our theory. In other words, a linear Bayes justification for the estimate $\hat{g}_i = a_{\lambda i} z_i$ requires marginal expectation $E\{w_i\} = 1/b_{\lambda i}$.

$\mathscr{F}^{(2)}$ gives sampling distributions much different than those obtained from the more familiar family $\mathscr{F}^{(1)}$. Figure 4 compares the density of $|z| = w^{1/2}$ for a single component from $\mathscr{F}^{(1)}$ or $\mathscr{F}^{(2)}$, with $b = 1$ and $c_0 = 1.334$. The $\mathscr{F}^{(1)}$ curve is half-normal, compared to which the $\mathscr{F}^{(2)}$ density is deficient near zero and strongly peaked near 1. This is not very realistic, as the example in Section 9 of Efron (1999) shows, but here we are using $\mathscr{F}^{(2)}$ more as a computational device for the C_p criterion, and as a convenient alternative to $\mathscr{F}^{(1)}$ that illustrates possible weaknesses in the GML criterion, than as a realistic sampling distribution in its own right.

Generalized cross validation (GCV) is a popular selection criterion that looks different from C_p but is actually quite similar. See Remark J, Section 11. Wahba [(1985), equation (1.2)] defines the GCV estimate of λ as the minimizer of

$$\mathrm{GCV}_\lambda = \frac{\|(I - A_\lambda)y\|^2}{[\mathrm{tr}(I - A_\lambda)]^2} = \frac{\Sigma b_{\lambda i}^2 v_i^2}{(\Sigma b_{\lambda i})^2}, \tag{4.21}$$

$\mathbf{v} = U'\mathbf{y}$ as in (3.24). Differentiating with respect to λ shows that $\hat{\lambda}_{\mathrm{GCV}}$ solves

$$\Sigma(b_{\lambda i} v_i^2 - \hat{\sigma}_\lambda^2)\dot{b}_{\lambda i} = 0 \quad \text{where} \quad \hat{\sigma}_\lambda^2 = \frac{\Sigma b_{\lambda j}^2 v_j^2}{\Sigma b_{\lambda j}} \tag{4.22}$$

[rather than the $\hat{\sigma}_\lambda^2$ of (3.25)]. By comparison, differentiating the expression following (4.13) shows that the C_p estimate satisfies

$$\Sigma(b_{\lambda i} v_i^2 - \sigma^2)\dot{b}_{\lambda i} = 0. \tag{4.23}$$

We see that the GCV criterion amounts to minimizing the C_p statistic, with σ^2 replaced by $\hat{\sigma}_\lambda^2$ for every choice of λ. Once again we will proceed differently in Section 8, replacing an unknown σ^2 with an estimate $\hat{\sigma}^2$ that does not depend on λ.

The curved exponential families $\mathscr{F}^{(1)}$ and $\mathscr{F}^{(2)}$ give the GML and C_p selection criterion as maximum likelihood estimates, and they also provide the simple geometric comparisons seen in Figure 3. The geometric relationships remain valid whether or not one believes in the probability models. In particular we will see in Section 6 that the fast rotation of the vectors $\dot{\eta}_\lambda^{(2)}$ causes serious estimation problems for the C_p criterion.

5. Information and efficiency. The curved exponential families $\mathscr{F}^{(p)}$, (4.9), have simple structures, both algebraically and geometrically, which we can use to compute quantities of interest. As a first example we will employ Fisher information calculations to approximate the accuracy of the degrees of freedom estimate $\hat{\nu}$.

The same reasoning that leads from (3.14) and (3.19) to (3.22) gives easy expressions for the derivatives of $l_\lambda(\mathbf{w}) = \log(d^{(p)}(\mathbf{w}))$, (4.9),

$$\begin{aligned} \dot{l}_\lambda(\mathbf{w}) &= \frac{-c_o p}{\lambda} \sum a_{\lambda i} b_{\lambda i}^{p-1} (b_{\lambda i} w_i - 1) \quad \text{and} \\ \ddot{l}_\lambda(\mathbf{w}) &= \frac{-c_o p}{\lambda^2} \sum a_{\lambda i} b_{\lambda i}^{p-1} [a_{\lambda i} + c_{\lambda i}(b_{\lambda i} w - 1)], \end{aligned} \tag{5.1}$$

where

$$c_{\lambda i} \equiv (p-1) a_{\lambda i} - 2 b_{\lambda i}. \tag{5.2}$$

Remembering that $c_o = 1/2$ for $p = 1$, we get (3.22) as the $p = 1$ version of (5.1).

Since $(b_{\lambda i} w_i - 1)$ has mean and variance

$$(b_{\lambda i} w_i - 1) \sim (0,\ 1/[c_o p b_{\lambda i}^{p-1}]), \tag{5.3}$$

independently in i according to (4.7), (4.8), the Fisher information for λ is

$$i_\lambda = E_\lambda [\dot{l}_\lambda(\mathbf{w})]^2 = \frac{c_o p}{\lambda^2} \sum a_{\lambda i}^2 b_{\lambda i}^{p-1}. \tag{5.4}$$

As before, for $p = 1$ it is important to remember that the sum does not include $i = 1, 2$. The Fisher information for degrees of freedom ν is $i_\nu = i_\lambda / (d\nu/d\lambda)^2$, with $d\nu/d\lambda = \Sigma \dot{a}_{\lambda i} = -\Sigma a_{\lambda i} b_{\lambda i}/\lambda$ from (3.19), so

$$i_\nu = c_o p \sum a_{\lambda i}^2 b_{\lambda i}^{p-1} \Big/ \Big(\sum a_{\lambda i} b_{\lambda i}\Big)^2. \tag{5.5}$$

In what follows we will use complete notation like $i_\nu^{(p)}$ only when necessary to differentiate between cases.

The Fisher information approximation $sd\{\hat{\nu}\} \doteq 1/\sqrt{i_{\hat{\nu}}}$ gives reasonably accurate results if we choose the true value of ν large enough to avoid the end-zone spikes seen in Figure 1. Choosing $\nu = 8$ and $\mathbf{x}$ as in (3.17) yielded $sd\{\hat{\nu}^{(1)}\} \doteq 1.20$

for family $\mathscr{F}^{(1)}$, compared with a robustly estimated standard deviation, (nintieth percentile minus tenth percentile)/ $(2 \cdot 1.28)$, of 1.27 for $\hat{\nu}^{(1)}$, based on 600 simulations from $\mathscr{F}^{(1)}$. The same choice of ν and $\mathbf{x}$ gave $sd(\hat{\nu}^{(2)}) \doteq 1.20$ in $\mathscr{F}^{(2)}$, $c_o = 1.334$, compared with simulation value 1.26. Higher order approximations for $sd(\hat{\nu})$ are discussed in Remark I, Section 11.

These calculations assumed we were using $\hat{\nu}^{(p)}$ to estimate ν within family $\mathscr{F}^{(p)}$. What happens if we use $\hat{\nu}^{(p_2)}$ within family $\mathscr{F}^{(p_1)}$? This question makes sense because $\hat{\nu}^{(p_2)}$ and $\hat{\nu}^{(p_1)}$ [or equivalently $\hat{\lambda}^{(p_2)}$ and $\hat{\lambda}^{(p_1)}$] are "estimating the same thing." More precisely, $\hat{\nu}^{(p_2)}$ is *Fisher consistent* for ν within family $\mathscr{F}^{(p_1)}$: if we happen to observe $\mathbf{w} = \boldsymbol{\mu}_\lambda = \mathbf{1}/\mathbf{b}_\lambda$ in Figure 3 then we will correctly estimate $\hat{\lambda}^{(p_2)} = \lambda$ and $\hat{\nu}^{(p_2)} = \nu$. See remark H. Standard considerations, as in Section 6 of Efron (1982), give the first-order asymptotic efficiency of $\hat{\nu}^{(p_2)}$ compared to the MLE $\hat{\nu}^{(p_1)}$, say $E(p_1, p_2)$, to be

$$(5.6)\quad E(p_1, p_2) = \Big[\sum V_{\lambda i}^{(p_1)} \dot{\eta}_{\lambda i}^{(p_1)} \dot{\eta}_{\lambda i}^{(p_2)}\Big]^2 \Big/ \Big\{\Big[\sum V_{\lambda i}^{(p_1)} \dot{\eta}_{\lambda i}^{(p_1)2}\Big]\Big[\sum V_{\lambda i}^{(p_1)} \dot{\eta}_{\lambda i}^{(p_2)2}\Big]\Big\},$$

where $V_{\lambda i}^{(p)}$ is the variance function (4.7) evaluated at $b_{\lambda i}$.

Efficiency refers here to the ratio of asymptotic variances under the true model $\mathscr{F}^{(p_1)}$, yielding the approximation

$$(5.7)\qquad \frac{sd_{p_1}\{\hat{\nu}^{(p_2)}\}}{sd_{p_1}\{\hat{\nu}^{(p_1)}\}} \doteq \frac{1}{\sqrt{E(p_1, p_2)}} = \frac{\{[\sum a_{\lambda i}^2 b_{\lambda i}^{p_1-1}][\sum a_{\lambda i}^2 b_{\lambda i}^{2p_2-p_1-1}]\}^{1/2}}{[\sum a_{\lambda i}^2 b_{\lambda i}^{p_2-1}]}.$$

This last expression results from using (4.11), (4.7) and (3.19) in (5.6). Notice that (5.7) does not involve the scaling constant c_o in (4.2).

Formula (5.7) works reasonably well when $p_1 = 1$, $\nu = 8$, predicting $sd\{\hat{\nu}^{(2)}\}/sd\{\hat{\nu}^{(1)}\} = 1.64$ compared to the simulation ratio 1.79. However it is less successful when the true model is $\mathscr{F}^{(2)}$, $\nu = 8$, predicting $sd\{\hat{\nu}^{(1)}\}/sd\{\hat{\nu}^{(2)} = 3.63$ compared to the observed ratio 1.48. Here, and in other simulations from $\mathscr{F}^{(2)}$, the GML estimate $\hat{\nu}^{(1)}$ outperformed its efficiency predictions.

Sections 2 and 4 of Stein (1990) compare the asymptotic standard errors of $\hat{\nu}^{(1)}$ and $\hat{\nu}^{(2)}$ in the $\mathscr{F}^{(1)}$ context of this paper. For cubic smoothing splines, Stein's development predicts $sd(\hat{\nu}^{(2)})/sd(\hat{\nu}^{(1)}) \to 1.83$, which is quite consonant with the results here.

Further discussion of efficiency appears in Section 5 of Efron (1999). That section also discusses the downward bias of GLM evident in Figure 1 and suggests a simple bias correction.

6. Curvature and the local reversal region. Curved exponential families such as $\mathscr{F}^{(1)}$ and $\mathscr{F}^{(2)}$ enjoy some of the good statistical properties of genuine one-parameter exponential families, more so if the "statistical curvature" γ_λ is small. This is the main point of Efron (1975) where it is shown, for example, that the variance of the MLE is increased proportionally to γ_λ^2.

This section shows that the curvature of $\mathscr{F}^{(p)}$ increases sharply with p, and in particular that $\mathscr{F}^{(2)}$ is highly curved in some of the situations we have been considering. At least part of C_p's eccentric behavior stems from an unfavorable curvature effect: the local reversal region (LRR) of Figure 2 moves closer to $\boldsymbol{\mu}_{\hat{\lambda}}$ as the curvature increases, causing maximum likelihood estimates to go astray.

Forgetting precise definitions for a moment, we can picture the curvature in terms of Figure 2. Consider increasing $\hat{\lambda}$ by a small amount $d\hat{\lambda}$ so that $\mathscr{L}_{\hat{\lambda}+d\hat{\lambda}}$ intersects the line of expectations at $\boldsymbol{\mu}_{\hat{\lambda}+d\hat{\lambda}}$, a little closer to $\mathbf{1}$. In a genuine exponential family (curvature = 0), $\mathscr{L}_{\hat{\lambda}+d\hat{\lambda}}$ parallels $\mathscr{L}_{\hat{\lambda}}$, but in a curved family $\dot{\boldsymbol{\eta}}_{\hat{\lambda}+d\hat{\lambda}}$ will be slightly rotated from $\dot{\boldsymbol{\eta}}_{\hat{\lambda}}$, causing $\mathscr{L}_{\hat{\lambda}+d\hat{\lambda}}$ to intersect $\mathscr{L}_{\hat{\lambda}}$. In Figure 2 the intersection will coincide with the point where $\mathscr{L}_{\hat{\lambda}}$ intersects the local reversal region. The vectors $\dot{\boldsymbol{\eta}}_{\lambda}$ rotate faster in $\mathscr{F}^{(2)}$ than in $\mathscr{F}^{(1)}$, moving the LRR closer to $\boldsymbol{\mu}_{\hat{\lambda}}$ and causing estimation problems.

Now imagine moving $\mathbf{w}$ along $\mathscr{L}_{\hat{\lambda}}$ in the direction away from $\boldsymbol{\mu}_{\hat{\lambda}}$. Efron (1978) shows that the observed Fisher information $-\ddot{l}_{\hat{\lambda}}(\mathbf{w}) = i_{\hat{\lambda}} - \ddot{\boldsymbol{\eta}}'_{\hat{\lambda}}(\mathbf{w} - \boldsymbol{\mu}_{\hat{\lambda}})$ decreases linearly, reaching zero at a critical point $\mathbf{w}_0$ that lies Mahalanobis distance $1/\gamma_{\hat{\lambda}}$ away from $\hat{\mu}_{\hat{\lambda}}$. By definition this point marks the boundary of $\mathrm{LRR}_{\hat{\lambda}}$, the local reversal region for parameter value $\hat{\lambda}$. The term "reversal" reflects an important phenomenon: for points $\mathbf{w}$ in the $\mathrm{LRR}_{\hat{\lambda}}$ portion of $\mathscr{L}_{\hat{\lambda}}$, $\hat{\lambda}$ is a local *minimum* rather than a local *maximum* of the likelihood $l_{\lambda}(\mathbf{w})$.

In particular, suppose that the point $\mathbf{w}$ lies in $\mathscr{L}_{\lambda} \cap \mathrm{LRR}_{\lambda}$ where λ is the true parameter value. For such points, *the estimate $\hat{\lambda}$ must be far from the true value λ, and likewise $\hat{\nu}$ from ν*, since λ is a local minimum of the likelihood.

Figure 5 illustrates the reversal phenomenon. It concerns the 600 C_p estimates $\hat{\nu}^{(2)}$ of Figure 1, in which the $\mathbf{w}$ vectors were drawn from the GML family $\mathscr{F}^{(1)}$, true degrees of freedom $\nu = 5$ as in (3.17) and (3.18). Each estimate $\hat{\nu}^{(2)}$ is plotted versus $T_5^{(2)}$, a measure of distance from $\boldsymbol{\mu}_{\lambda}$ toward LRR_5, defined below at (6.7); points $\mathbf{w}$ with $T_5^{(2)}$ exceeding 1 lie in LRR_5. The empty area of Figure 5 beginning at the arrowed point demonstrates the LRR effect: $\mathbf{w}$ *vectors in* LRR_{λ} *cannot give estimates near the true value* $\nu = 5$. The reversal phenomenon does not affect GML in this case, or in most cases, because very few $\mathbf{w}$ vectors fall into the more distant GML reversal region.

Proceeding more carefully now, the formal definition of statistical curvature in Efron (1975) is given in terms of the covariance matrix M_{λ} of $\dot{l}_{\lambda}(\mathbf{w})$ and $\ddot{l}_{\lambda}(\mathbf{w})$, the first two derivatives of the log likelihoods,

$$(6.1)\qquad M_{\lambda} = \begin{pmatrix} \mathrm{var}_{\lambda}\{\dot{l}_{\lambda}\} & \mathrm{cov}_{\lambda}\{\dot{l}_{\lambda}, \ddot{l}_{\lambda}\} \\ \mathrm{cov}_{\lambda}\{\dot{l}_{\lambda}, \ddot{l}_{\lambda}\} & \mathrm{var}_{\lambda}\{\ddot{l}_{\lambda}\} \end{pmatrix} = \begin{pmatrix} \dot{\boldsymbol{\eta}}'_{\lambda} V_{\lambda} \dot{\boldsymbol{\eta}}_{\lambda} & \dot{\boldsymbol{\eta}}'_{\lambda} V_{\lambda} \ddot{\boldsymbol{\eta}}_{\lambda} \\ \dot{\boldsymbol{\eta}}'_{\lambda} V_{\lambda} \ddot{\boldsymbol{\eta}}_{\lambda} & \ddot{\boldsymbol{\eta}}'_{\lambda} V_{\lambda} \ddot{\boldsymbol{\eta}}_{\lambda} \end{pmatrix}.$$

Here we have used $\dot{l}_{\lambda}(\mathbf{w}) = \dot{\boldsymbol{\eta}}'_{\lambda}(\mathbf{w} - \boldsymbol{\mu}_{\lambda})$, $\ddot{l}_{\lambda}(\mathbf{w}) = \ddot{\boldsymbol{\eta}}'_{\lambda}(\mathbf{w} - \boldsymbol{\mu}_{\lambda}) - i_{\lambda}$, $i_{\lambda} = \dot{\boldsymbol{\eta}}'_{\lambda} V_{\lambda} \dot{\boldsymbol{\eta}}_{\lambda}$ the Fisher information for λ, and $\mathrm{cov}_{\lambda}(\mathbf{w}) = V_{\lambda}$. The statistical curvature of $\mathscr{F}$ at λ is defined to be

$$(6.2)\qquad \gamma_{\lambda} = (|M_{\lambda}|/i_{\lambda}^{3})^{1/2} = \frac{1}{i_{\lambda}}[\mathrm{var}_{\lambda}\{\ddot{l}_{\lambda}\} - \mathrm{cov}_{\lambda}\{\dot{l}_{\lambda}, \ddot{l}_{\lambda}\}^2/\mathrm{var}_{\lambda}\{\dot{l}_{\lambda}\}]^{1/2}.$$

Notice that γ_{λ} must be nonnegative since the bracketed term is the residual variance of $\ddot{l}_{\lambda}(\mathbf{w})$ after linear regression on $\ddot{l}_{\lambda}(\mathbf{w})$.

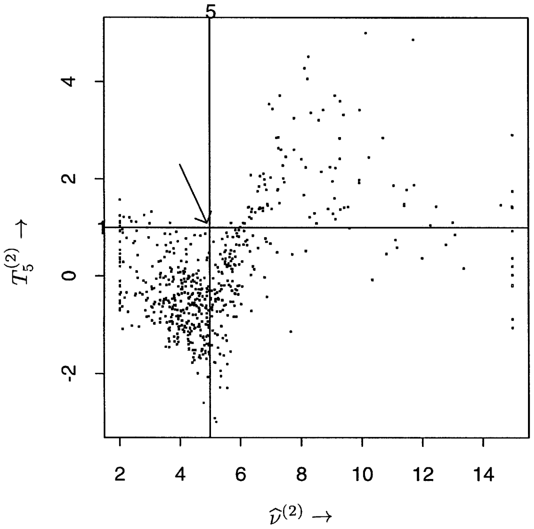

FIG. 5. *An illustration of the reversal region effect, from the simulation of Figure* 1; *the* 600 C_p *estimates* $\hat{\nu}^{(2)}$ *are plotted versus* $T_5^{(2)}$, (6.7); *values of* $T_5^{(2)}$ *exceeding* 1 *are in* $\mathrm{LRR}_5^{(2)}$, *the local reversal region of* $\mathscr{F}^{(2)}$ *for true df* $\nu = 5$. *Empty area beginning at arrowed point shows that* $\hat{\nu}^{(2)}$ *cannot be near* ν *for* $T \geq 1$.

The curvature of family $\mathscr{F}^{(p)}$, (4.9), can be expressed compactly in terms of the following notation: for any function of i, say $c(i)$, define

$$\check{E}\{c\} = \sum_i a_{\lambda i}^2 b_{\lambda i}^{p-1} c(i) \Big/ \sum_i a_{\lambda i}^2 b_{\lambda i}^{p-1}, \tag{6.3}$$

and similarly $\check{\mathrm{Var}}\{c\} = \check{E}\{c - \check{E}(c)\}^2$, the sums taken from 3 to n in the smoothing spline case (2.5).

THEOREM 2. *The squared curvature of* $\mathscr{F}^{(p)}$ *at* λ *is*

$$\gamma_\lambda^2 = \sum a_{\lambda i}^2 b_{\lambda i}^{p-1} \check{c}_{\lambda i}^2 \Big/ c_o p \Big[\sum a_{\lambda i}^2 b_{\lambda i}^{p-1}\Big]^2, \tag{6.4}$$

where

$$c_{\lambda i} = (p-1)a_{\lambda i} - 2b_{\lambda i} \quad \textit{and} \quad \check{c}_{\lambda i} = c_{\lambda i} - \check{E}\{c_{\lambda j}\}. \tag{6.5}$$

We can also write $\gamma_\lambda^2 = \check{\mathrm{V}}\mathrm{ar}\{c_{\lambda i}\}/[c_o p \sum a_{\lambda i}^2 b_{\lambda i}^{p-1}]$.

The theorem follows directly from (5.1), (5.2) and (4.7).

Figure 6 shows the squared curvature of family $\mathscr{F}^{(p)}$ for λ corresponding to degrees of freedom ν from 2 through 14, and $\mathbf{x}$ as in (3.17). The constant c_o used in (4.9) was the Fisher information matching choice $(\Sigma a_{\lambda i}^2)/(2p\Sigma a_{\lambda i}^2 b_{\lambda i}^{p-1})$, as in (4.17). We will see that squared curvatures much exceeding 0.25 degrade the estimation of ν. The C_p family $\mathscr{F}^{(2)}$ greatly exceeds this bound for ν less than 8,

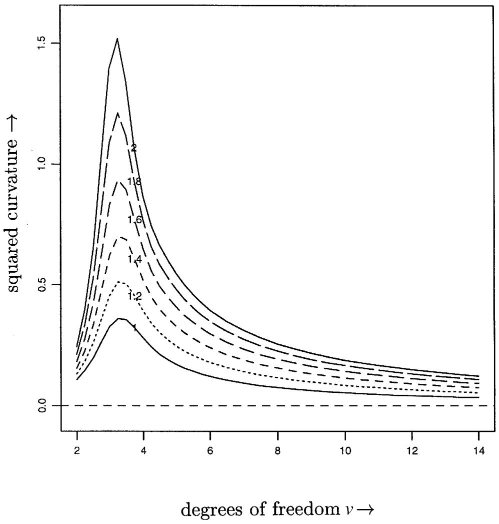

FIG. 6. *Squared curvature of family $\mathscr{F}^{(p)}$ at degrees of freedom ν; formula* (6.4). *Top curve shows that $\mathscr{F}^{(2)}$ has very large curvature when ν is small. Constant c_o in* (4.9) *chosen to match $\mathscr{F}^{(1)}$ Fisher information, as in* (4.17).

while the GML family $\mathscr{F}^{(1)}$ is much less curved. Our other suggested choices of c_o widen the differences: $c_o \leq 3/8$ gives $\mathscr{F}^{(2)}$ a maximum γ^2 exceeding 3. The "cross curvature" calculations below eliminate the choice of c_o from the comparison of $\mathscr{F}^{(1)}$ and $\mathscr{F}^{(2)}$.

The next theorem specifies the location of LRR_λ. Suppressing the superscript "p," let

$$\mathbf{o}_\lambda = \ddot{\eta}_\lambda - \beta_\lambda \dot{\eta}_\lambda, \tag{6.6}$$

where $\beta_\lambda = \dot{\eta}'_\lambda V_\lambda \ddot{\eta}_\lambda / \dot{\eta}_\lambda V_\lambda \dot{\eta}_\lambda$ as in (6.1), so $\mathbf{o}_\lambda$ is the part of $\ddot{\eta}_\lambda$ orthogonal to $\dot{\eta}_\lambda$, using the inner product $\langle u, v\rangle_\lambda = u'V_\lambda v$, with $V_\lambda = \mathrm{cov}_\lambda(\mathbf{w}) = \mathrm{diag}(c_o p b_{\lambda i}^{p+1})^{-1}$ as before.

Define

$$T = T_\lambda^{(p)} = \frac{\mathbf{o}'_\lambda}{i_\lambda}(\mathbf{w} - \boldsymbol{\mu}_\lambda). \tag{6.7}$$

THEOREM 3. (a) *For* $\mathbf{w}$ *in the flat space* $\mathscr{L}_\lambda = \{\dot{l}_\lambda(\mathbf{w}) = 0\}$, *the observed Fisher information is*

$$-\ddot{l}_\lambda(\mathbf{w}) = i_\lambda[1 - T]. \tag{6.8}$$

(b) *The flat space* $\{T = 1\} \cap \mathscr{L}_\lambda$, *of dimension* $n - 4$ *in the cubic smoothing spline case, is the subset of* $\mathscr{L}_\lambda$ *for which* $-\ddot{l}_\lambda(\mathbf{w}) = 0$. *Within this subset the closest point to* $\boldsymbol{\mu}_\lambda$ *in terms of the distance* $\|\mathbf{v}\|_\lambda = [\mathbf{v}'V_\lambda^{-1}\mathbf{v}]^{1/2}$ *is the* critical point $\mathbf{w}_o = \boldsymbol{\mu}_\lambda + V_\lambda[\mathbf{o}_\lambda/(i_\lambda \gamma_\lambda^2)]$, *having distance*

$$\|\mathbf{w}_o - \boldsymbol{\mu}_\lambda\|_\lambda = 1/\gamma_\lambda. \tag{6.9}$$

(c) *If* $\mathbf{w} \sim d_\lambda^{(p_o)}(\cdot)$, (4.9), *then* $T_\lambda^{(p)}$ *has mean and variance*

$$T_\lambda^{(p)} \sim [0, \gamma_\lambda^2(p_0, p)], \tag{6.10}$$

where the squared cross-curvature $\gamma_\lambda^2(p_o, p)$ *equals*

$$\gamma_\lambda^2(p_o, p) = \frac{\mathbf{o}_\lambda^{(p)'} V_\lambda^{(p_o)} \mathbf{o}_\lambda^{(p)}}{i_\lambda^{(p)2}}. \tag{6.11}$$

If $p_o = p$ *then* $\gamma_\lambda(p, p) = \gamma_\lambda^{(p)}$, (6.4).

The theorem's proof, most of which is generalized from Efron (1978), is presented in Remark K, Section 11.

The local reversal region for parameter value λ, LRR_λ or more carefully $\mathrm{LRR}_\lambda^{(p)}$, is defined to be

$$\mathrm{LRR}_\lambda^{(p)} = \{\mathbf{w}\colon T_\lambda^{(p)} \geq 1\} \tag{6.12}$$

The intersection $\mathscr{L}_\lambda \cap \mathrm{LRR}_\lambda$ consists of those points $\mathbf{w}$ having $\dot{l}_\lambda(\mathbf{w}) = 0$ and $-\ddot{l}_\lambda(\mathbf{w}) \geq 0$, leading to erratic estimates $\hat{\nu}$ as seen in Figure 6. The gist of Theorem 3 is that a large curvature puts the LRR close to $\boldsymbol{\mu}_\lambda = E_\lambda\{\mathbf{w}\}$, increasing the probability of a bad estimate. Assuming approximate normality for T, (6.10)–(6.12) give

$$\mathrm{Prob}_\lambda^{(p_o)}\{\mathbf{w} \in \mathrm{LRR}_\lambda^{(p)}\} \doteq 1 - \Phi(1/\gamma_\lambda(p_o, p)). \tag{6.13}$$

In the case of Figure 5 the cross-curvature (6.11) is quite large, $\gamma(1, 2) = 1.19$, indicating a substantial reversal effect, $\mathrm{Prob}^{(1)}\{\mathrm{LRR}^{(2)}\} \doteq 0.20$ from (6.13). The actual observed proportion in Figure 5 was 0.18. By comparison $\gamma(1, 1) = \gamma^{(1)}$ is only 0.415, correctly indicating negligible reversal effects for $\hat{\nu}^{(1)}$. The same analysis shows that $\hat{\nu}^{(1)}$ is less subject to reversal effects than is $\hat{\nu}^{(2)}$ even when sampling from $\mathscr{F}^{(2)}$ as in the right panel of Figure 4, $\gamma(2, 1) = 0.578$ versus $\gamma(2, 2) = 0.739$.

We can separate the poor performance of $\hat{\nu}^{(2)}$ in the Figure 1 simulation into two parts. The efficiency formula (5.7), which depends on the angle between $\mathscr{L}_\lambda^{(1)}$ and $\mathscr{L}_\lambda^{(2)}$, predicts $sd\{\hat{\nu}^{(2)}\}/sd\{\hat{\nu}^{(1)}\} = 1.49$, this calculation being relevant to $\mathbf{w}$ vectors not too far away from $\boldsymbol{\mu}_\lambda$. However, $\hat{\nu}^{(2)}$ also suffers from a nonlocal type of inefficiency not considered in (5.7): for the 18% of the $\mathbf{w}$ vectors in $\mathrm{LRR}^{(2)}$, $\hat{\nu}^{(2)}$ is 2.5 times worse than $\hat{\nu}^{(1)}$, measured in terms of mean absolute deviation from $\nu = 5$.

All of the curvature quantities have simple computational expressions derived from our previous formulas, for example,

$$T_\lambda^{(p)} = -\sum a_{\lambda i} b_{\lambda i}^p \check{c}_{\lambda i}(w_i - 1/b_{\lambda i}) \Big/ \sum a_{\lambda i}^2 b_{\lambda i}^{p-1}, \tag{6.14}$$

using notation (6.5). Notice that (6.14) does not depend on the constant c_o in (4.9); $\mathrm{LRR}_\lambda^{(p)} = \{T_\lambda^{(p)} \geq 1\}$ is defined in purely geometric terms, a specific region in $\mathbf{w}$ space as indicated in Figure 2. The cross-curvature (6.11), which concerns the probability that $\mathbf{w} \sim d_\lambda^{(p_o)}$ falls into $\mathrm{LRR}_\lambda^{(p)}$, does involve $c_o^{(p_o)}$, in notation (6.5),

$$\gamma_\lambda^2(p_o, p) = \frac{\sum a_{\lambda i}^2 b_{\lambda i}^{2p-p_o-1} \breve{c}_{\lambda i}^2}{c_o^{(p_o)} p_o (\sum a_i b_i^{p-1})^2}. \tag{6.15}$$

Definition (6.12) extends $\mathrm{LRR}_\lambda^{(p)}$ outside of $\mathscr{L}_\lambda$ in a manner made obvious in remark K. The best argument for the extension, besides its simplicity, comes from simulation results like those in Figure 5. Points $\mathbf{w}$ in the LRR mitigate toward poor estimates of ν, though Figure 5 shows they are not alone in this regard.

7. Repeated sampling. All of our numerical results have related to situation (3.17), where the design points $\mathbf{x}$ comprised $n = 61$ equally spaced values. Suppose instead we doubled the design point density, taking $n = 121$ equally spaced points

$$\mathbf{x} = \left(-1, -1 + \frac{2}{120} - 1 + \frac{4}{120}, \dots, 1\right). \tag{7.1}$$

It seems plausible that this would greatly change our results, for example making the estimates $\hat{\nu}^{(1)}$ and $\hat{\nu}^{(2)}$ in Figure 1 much more accurate.

In fact this is not true. All of our smoothing spline results—informations, standard deviations, efficiencies, curvatures—*change by less than* 1% *in going from* (3.17) *to* (7.1). Except for simulation error, Figure 1 is hardly affected by the doubled size of n. This seems like a surprising phenomenon but it is easy to understand in terms of repeated sampling within the empirical Bayes model of Section 3.

Going back to the original $(\mathbf{f}, \mathbf{y})$ coordinates of Section 2, the empirical Bayes formulation (3.1) is equivalent to

$$\mathbf{f} \sim N(\mathbf{0}, \sigma^2 C_\lambda) \quad \text{and} \quad \mathbf{y}|\mathbf{f} \sim N(\mathbf{f}, \sigma^2 I), \tag{7.2}$$

with

$$C_\lambda = U \mathbf{c}_\lambda U'$$

as in (2.3). More careful notation would separate out $c_{\lambda 1}$ and $c_{\lambda 2}$, which are infinite in the smoothing spline situation, but as before these coordinates do not affect inferences concerning λ.

Instead of (7.2) consider the more general formulation,

$$\mathbf{f} \sim N(\mathbf{0}, \sigma^2 C_\lambda) \quad \text{and} \quad \mathbf{y}\,|\mathbf{f} \sim N(\mathbf{f}, \tilde{\sigma}^2 I) \tag{7.3}$$

with

$$\tilde{\sigma}^2 = \sigma^2/m \tag{7.4}$$

for some positive number m. This model applies to a repeated sampling version of (7.2): having obtained the (unobservable) vector $\mathbf{f} \sim N(\mathbf{0}, \sigma^2 C_\lambda)$, we observe $\mathbf{y}(1), \mathbf{y}(2), \ldots, \mathbf{y}(m) \overset{\text{i.i.d.}}{\sim} N(\mathbf{f}, \sigma^2 I)$ so that the sufficient statistic $\mathbf{y} = \sum_1^m \mathbf{y}(j)/m$ is distributed as in (7.3), (7.4).

Defining

$$\tilde{\lambda} = \lambda/m \quad \text{and} \quad \widetilde{C}_{\tilde{\lambda}} = mC_{m\tilde{\lambda}}, \tag{7.5}$$

model (7.3), (7.4) can be rewritten as

$$\mathbf{f} \sim N(0, \tilde{\sigma}^2 \widetilde{C}_{\tilde{\lambda}}) \quad \text{and} \quad \mathbf{y}|\mathbf{f} \sim N(\mathbf{f}, \tilde{\sigma}^2 I). \tag{7.6}$$

Notice that $c_{\lambda i} = (\lambda k_i)^{-1}$ from (2.4), (3.4), so that $m\ c_{m\tilde{\lambda}, i} = c_{\tilde{\lambda} i}$ and $\widetilde{C}_{\tilde{\lambda}} = C_{\tilde{\lambda}}$. This makes (7.6) identical to our original model (7.2), with a name change for the free variable λ, except that σ^2 has been replaced by $\tilde{\sigma}^2$. However, σ^2 disappears in the transformations (2.7) that bring us to the GML empirical Bayes model (3.1).

In other words, (7.2) and the repeated sampling version (7.3), (7.4) lead to the same model (3.1) for estimating λ or ν. Figure 2 remains exactly the same in both situations, as does any property of the GML estimate, for instance, $sd_\nu\{\hat{\nu}^{(1)}\}$. Remark G in Section 11 describes a different repeated sampling model that gives more intuitive and familiar results.

The effect of changing $\mathbf{x}$ from (3.17) to (7.1) is nearly the same as going from model (7.2) to (7.3), (7.4) with $m = 2$, which explains why the informations, curvatures, etc. changed very little. In our examples the values of $a_{\lambda i}$ for $i \leq 25$ were almost identical in the two situations. Our formulas are insensitive to the higher i terms, so that all the $\hat{\nu}^{(p)}$, not just $\hat{\nu}^{(1)}$, were nearly invariant between (3.17) and (7.1). (This would not have been the case if our examples concerned bigger values of ν, greater than say 25.)

As another test of sensitivity, the design points (3.17) were mapped nonlinearly to give a new $\mathbf{x}$ vector, $x_i \to \sin((\pi/2)x_i)$ for $i = 1, 2, \ldots 61$. The resulting changes in our figures and tables were all less than 3%, again making the point that the numerical comparisons we have been making apply with at least moderate generality.

8. Unknown σ^2.

Our development of the GML estimate in Section 3 assumed that σ^2 was known in the sampling model $\mathbf{y} \sim N(\mathbf{f}, \sigma^2 I)$ so that we could calculate $\mathbf{z} = U'\mathbf{y}/\sigma$ and the sufficient vector $\mathbf{w} = \mathbf{z}^2$, (3.6). Section 4 tacitly makes a similar assumption. In practice σ^2 must itself be estimated, causing some loss of accuracy for $\hat{\nu}$ or $\hat{\lambda}$. This section gives a formula for approximating the accuracy loss.

To begin with, suppose that we have available an estimate $\tilde{\sigma}^2$ for σ^2 that is independent of the vector $\mathbf{w}$ in (4.9), and such that the random variable $R = \sigma^2/\tilde{\sigma}^2$ has mean and variance

$$R \sim (1, \text{var}_R). \tag{8.1}$$

Of particular interest is the chi-squared case where

$$\tilde{\sigma}^2 \sim \frac{\sigma^2 \chi_N^2}{N-2} \qquad \text{with} \quad \text{var}_R = \frac{2}{N-4}. \tag{8.2}$$

The obvious substitute for $\mathbf{w}$ in this situation is based on $\tilde{\mathbf{z}} = U'\mathbf{y}/\tilde{\sigma}$,

$$\tilde{\mathbf{w}} = \tilde{\mathbf{z}}^2 = R\mathbf{w}. \tag{8.3}$$

LEMMA. *If* $\mathbf{w}$ *has mean and covariance* $\mathbf{w} \sim (\boldsymbol{\mu}, V)$ *then*

$$\tilde{\mathbf{w}} \sim (\boldsymbol{\mu}, V + \mathrm{var}_R \ \cdot \ (V + \boldsymbol{\mu}\boldsymbol{\mu}')). \tag{8.4}$$

The proof is immediate by direct calculation of the first and second moments.

We can imagine substituting $\tilde{\mathbf{w}}$ for $\mathbf{w}$ in Figures 2 or 3, while still using the GML or C_p estimates as pictured. The conditional mean and covariance given $\mathbf{w}$ are

$$\tilde{\mathbf{w}}|\mathbf{w} \sim (\mathbf{w}, \mathrm{var}_R \cdot \mathbf{w}\mathbf{w}'), \tag{8.5}$$

so if var_R is small then $\tilde{\mathbf{w}}$ will lie near $\mathbf{w}$ and the estimate of ν or λ will not be much affected. We can use (8.4) to approximate the overall variance increase when $\mathbf{w}$ is replaced by $\tilde{\mathbf{w}}$.

Let

$$\mathbf{I}_\lambda = \dot{\eta}_\lambda / i_\lambda, \tag{8.6}$$

(3.20), (5.4), this being the influence function (3.21) of $\hat{\lambda}$, evaluated at $\mathbf{w} = \boldsymbol{\mu}_\lambda$. A first-order Taylor series expansion gives

$$\hat{\lambda} - \lambda \doteq \mathbf{I}'_\lambda(\mathbf{w} - \boldsymbol{\mu}_\lambda), \tag{8.7}$$

and likewise

$$\tilde{\lambda} - \lambda \doteq \mathbf{I}'_\lambda(\tilde{\mathbf{w}} - \boldsymbol{\mu}_\lambda) \tag{8.8}$$

since the estimate $\tilde{\lambda}$ based on $\tilde{\mathbf{w}}$ is the same function of its vector argument.

To first order the ratio of variances is

$$\frac{\mathrm{var}\{\tilde{\lambda}\}}{\mathrm{var}\{\hat{\lambda}\}} \doteq \frac{\mathbf{I}'_\lambda \widetilde{V}_\lambda \mathbf{I}_\lambda}{\mathbf{I}'_\lambda V_\lambda \mathbf{I}_\lambda}, \tag{8.9}$$

with V_λ and $\widetilde{V}_\lambda$ being the covariance matrices of $\mathbf{w}$ and $\tilde{\mathbf{w}}$.

The same ratio applies to $\mathrm{var}\{\tilde{\nu}\}/\mathrm{var}\{\hat{\nu}\}$ since $\mathbf{I}_\nu = \mathbf{I}_\lambda \cdot (i_\lambda / i_\nu)$, and i_λ / i_ν cancels out of (8.9). The lemma then gives a convenient approximation for the ratio of standard errors.

THEOREM 4.

$$\begin{aligned} \frac{\mathrm{sterr}_\nu(\tilde{\nu})}{\mathrm{sterr}_\nu(\hat{\nu})} &\doteq 1 + \frac{\mathrm{var}_R}{2}\left\{1 + \frac{(\dot{\eta}'_\lambda \boldsymbol{\mu}_\lambda)^2}{\dot{\eta}'_\lambda V_\lambda \dot{\eta}_\lambda}\right\}. \\ &= 1 + \frac{\mathrm{var}_R}{2}\left\{1 + c_o p \frac{(\sum a_{\lambda i} b_{\lambda i}^{p-1})^2}{(\sum a_{\lambda i}^2 b_{\lambda i}^{p-1})}\right\}, \end{aligned} \tag{8.10}$$

this last expression, based on (4.7) *and* (4.11), *applying when* $\hat{\nu}^{(p)}$ *is used to estimate* ν *in family* $\mathscr{F}^{(p)}$, (4.9).

TABLE 1
Ratio sterr($\tilde{\nu}$)/sterr($\hat{\nu}$) *from Theorem* 4, *for* $N = 40$ *in* (8.2), $\mathbf{x}$ *as in* (3.17)

df ν	**$p = 1$,** **$c_o = 0.5$**	**$p = 2$,** **$c_o = 3/8$**	**$p = 2$,** **c_o from (4.17)**
3:	1.06	1.04	1.08
5:	1.09	1.05	1.22
7:	1.13	1.07	1.37
9:	1.16	1.08	1.52

In practical applications of smoothers there are usually ample degrees of freedom for estimating σ^2. The higher numbered coordinates of $\mathbf{v} = U'\mathbf{y}$ are nearly $N(0, \sigma^2)$ distributed because of the smoothness assumption so

$$\tilde{\sigma}^2 = \sum_{n-1-N}^{n} v_i^2/(N-2) \tag{8.11}$$

approximately satisfies (8.1), (8.2). Table 1 shows some results from (8.10), using $N = 40$ and $\mathbf{x}$ as in (3.17), $n = 61$. As a check on the formula, Monte Carlo simulation gave $\text{sterr}_8(\tilde{\nu})/\text{sterr}_8(\hat{\nu}) = 1.14$ in the case $p = 1$, $\nu = 8$, $\mathbf{x}$ as in (3.17), compared to 1.145 from (8.10). This is a favorable situation for Theorem 4. The local approximation (8.8) would not fare as well in the face of substantial end zone or reversal region effects. Estimate (8.11) depends only on the higher coordinates of $\mathbf{v} = U'\mathbf{y}$. Conversely $\hat{\nu}$ and $\tilde{\nu}$ only depend on the lower coordinates when ν is small. In our example we could take them to be the MLEs based on just the first 21 coordinates of $\mathbf{v}$ with virtually no loss of Fisher information, (5.5), and thus achieve perfect independence from $\tilde{\sigma}^2$.

9. End zone calculations.

The GML estimate $\hat{\nu}^{(1)}$ has a tendency toward underestimating ν, that is, oversmoothing, even within its own MLE family $\mathscr{F}^{(1)}$. The spike of 44 cases having $\hat{\nu}^{(1)} = 2$ in the left panel of Figure 1 is a worrisome reminder of this tendency. All 44 cases had $\mathbf{w}$ vectors falling into the "end zone" shown in Figures 2 and 3, a phenomenon examined briefly in this section.

As $\lambda \to \infty$ the vector $\dot{\eta}_\lambda$ becomes parallel to the line of expectations $\{\boldsymbol{\mu}_\lambda = \mathbf{1} + 1/(\lambda\mathbf{k})\}$, (3.15), since (2.4) and (4.11) give

$$\lim_{\lambda\to\infty}\left\{-\frac{\lambda^2}{c_o p}\dot{\eta}_{\lambda i}\right\} = \frac{1}{k_i}. \tag{9.1}$$

For any value of p this makes $\mathscr{L}_\infty$, the limit of the flat spaces $\mathscr{L}_\lambda = \{\mathbf{w}\colon \dot{\eta}'_\lambda(\mathbf{w} - \boldsymbol{\mu}_\lambda) = 0\}$, pass through $\mathbf{1}$ orthogonally to the line of expectations, as indicated in Figures 2 and 3. Since $\mathscr{L}_\lambda$ is the set of $\mathbf{w}$ vectors for which λ is a solution to the MLE equation $\dot{l}_\lambda(\mathbf{w}) = 0$, it is reasonable to suppose that the region beyond $\mathscr{L}_\infty$, called the end zone, will correspond to cases where the MLE occurs at the extreme end of the family, at $\hat{\nu} = 2$ in the smoothing spline case.

In fact, 57 of the 600 **w** vectors in the first simulation experiment of Figure 1 fell into the end zone; 13 of them had $\hat{\nu}^{(1)} > 2$, because of the "reach-back" phenomenon illustrated by $\mathbf{w}_2$ in Figure 3, leaving 44 with $\hat{\nu}^{(1)} = 2$; a further 19 of these 44 had $\hat{\nu}^{(2)} > 2$, showing the greater reach back of $\mathscr{L}_\lambda^{(2)}$ vis-a-vis $\mathscr{L}_\lambda^{(1)}$ indicated in Figure 3.

Define $S_i = (1/b_{\lambda i} - 1)(w_i - 1)$ and

$$S = \sum S_i = \sum\left(\frac{1}{b_{\lambda i}} - 1\right)(w_i - 1). \tag{9.2}$$

Since $(1/b_{\lambda i} - 1) = a_{\lambda i}/b_{\lambda i} = 1/(\lambda k_i)$, (2.4), S is proportional to $\sum(1/k_i)(w_i - 1)$, and Figure 2 makes it easy to see that **w** *is in the end zone if and only if* $S < 0$. For the GML family $\mathscr{F}^{(1)}$, (3.7), we have

$$S_i \overset{\text{ind}}{\sim} \left(\frac{1}{b_{\lambda i}} - 1\right)\left(\frac{\chi_1^2}{b_{\lambda i}} - 1\right), \tag{9.3}$$

a convenient formula for end-zone calculations.

S_3, the first component of S in the smoothing spline case, by itself nearly determined the sign of S (and inclusion in the end zone) in the experiment of Figure 1. Table 2 shows why. In this case the value of $b_{\lambda i}$ is only 0.030, causing S_3 in (9.3) to vary over a much greater range than the other S_i's. All 57 end-zone **w** vectors had $S_3 < 0$.

As a remedy for this overdependence on S_3, the following ad hoc modification to $\hat{\mu}^{(1)}$ was investigated: if $\hat{\nu}^{(1)} = 2$ then the GML estimate was recomputed ignoring w_3, that is, by taking the sum in (3.23) from 4 to n instead of from 3 to n. This had a good effect in Figure 1, reducing the left-hand spike from 44 to 21 and moving $\hat{\nu}$ closer to $\nu = 5$, but of course it worsens estimation if the true ν equals 2. In effect this scheme distorts the level surfaces of estimation from the flat spaces $\mathscr{L}_{\hat{\lambda}}$ seen in Figure 2 to something more complicated, and hopefully more robust, but no general recommendations are possible at this point.

Remark L in Section 11 discusses an extension of the exponential families $\mathscr{F}^{(p)}$ that helps quantify end-zone behavior.

TABLE 2

*Quantiles $S_i^{(\alpha)}$ for the components of S, (9.2); for $\mathscr{F}^{(1)}$, $\nu = 5$, as in Figure 1. S_3 by itself nearly determines the sign of S and the inclusion of **w** in the end zone*

i	**0.025**	**0.05**	**0.10**	**0.50**	**0.90**	**0.95**	**0.975**	$\mathbf{b_{\lambda i}}$
3	−31.8	−28.5	−15.3	473.2	2976.4	4239.8	5555.0	0.030
4	−4.3	−4.2	−4.0	6.1	57.9	84.1	111.2	0.188
5	−1.1	−1.1	−1.1	−0.0	5.3	8.0	10.9	0.471
6	−0.4	−0.4	−0.4	−0.2	1.2	1.8	2.5	0.709
7	−0.2	−0.2	−0.2	−0.1	0.4	0.6	0.9	0.844

10. Estimation of f. Our interpretation of Figure 1 tacitly assumed that an estimator $\hat{\nu}$ should be evaluated in terms of its accuracy in estimating the degrees of freedom ν. However the ultimate goal of smoothing is to have $\hat{\mathbf{f}}_{\hat{\lambda}} = A_{\hat{\lambda}}\mathbf{y}$ be a good estimator of $\mathbf{f}$. One might imagine that $\hat{\nu}^{(2)}$ was really not inferior to $\hat{\nu}^{(1)}$ in the left panel of Figure 1, but was just doing a better job of tracking the different $\mathbf{f}$'s involved in the simulation. This section argues that this is not the case, and that the estimation of ν by $\hat{\nu}$ is a reasonable performance criterion, at least in the GML empirical Bayes context of Section 3. We will continue to use notation such as $\hat{\mathbf{f}}_{\hat{\lambda}}$ in place of $\hat{\mathbf{f}}_{\lambda(\hat{\nu})}$, remembering that λ is a monotone function of ν.

The orthogonal transformations in (2.7) show that

$$\|\mathbf{f} - \hat{\mathbf{f}}_{\hat{\lambda}}\|^2/\sigma^2 = \|\mathbf{g} - \hat{\mathbf{g}}_{\hat{\lambda}}\|^2, \tag{10.1}$$

$\hat{\mathbf{g}}_{\hat{\lambda}} = \mathbf{a}_{\hat{\lambda}}\mathbf{z}$, while

$$\begin{aligned} E_{\lambda}\{\|\mathbf{g} - \hat{\mathbf{g}}_{\hat{\lambda}}\|^2 \ |\mathbf{z}\} &= \|(\mathbf{a}_{\lambda} - \mathbf{a}_{\hat{\lambda}})\mathbf{z}\|^2 + \operatorname{tr}(\mathbf{a}_{\lambda}) \\ &= \sum (b_{\hat{\lambda}i} - b_{\lambda i})^2 w_i + \nu \end{aligned} \tag{10.2}$$

according to (3.2), (3.3). Moreover, a Taylor expansion of $\hat{\lambda}(\mathbf{w})$ around $\mathbf{w} = \boldsymbol{\mu}_{\lambda}$ yields

$$b_{\hat{\lambda}i} - b_{\lambda i} \doteq \frac{a_{\lambda i} b_{\lambda i}}{\sum a_{\lambda j} b_{\lambda j}}(\hat{\nu} - \nu). \tag{10.3}$$

Here we have used (3.19) and Fisher consistency $\hat{\nu}(\boldsymbol{\mu}_{\lambda}) = \nu$.

Combining (10.1)–(10.3) gives the useful approximation

$$E_{\lambda}\{\|\mathbf{f} - \hat{\mathbf{f}}_{\hat{\lambda}}\|^2/\sigma^2 \ |\mathbf{z}\} \doteq \nu + (\hat{\nu} - \nu)^2 Q(\mathbf{w}), \tag{10.4}$$

where $Q(\mathbf{w}) = (\sum a_{\lambda i}^2 b_{\lambda i}^2 w_i)/(\sum a_{\lambda i} b_{\lambda i})^2$. If $\hat{\nu}^{(1)}(\mathbf{w})$ and $\hat{\nu}^{(2)}(\mathbf{w})$ are competing estimates of ν, we can rewrite (11.4) as

$$R(\mathbf{w}) \equiv \frac{E_{\lambda}\{\|\mathbf{f} - \hat{\mathbf{f}}_{\hat{\lambda}^{(2)}}\|^2|\mathbf{z}\} - \nu\sigma^2}{E_{\lambda} \ \|\mathbf{f} - \hat{\mathbf{f}}_{\hat{\lambda}^{(1)}}\|^2|\mathbf{z}\} - \nu\sigma^2} \doteq \frac{(\hat{\nu}^{(2)} - \nu)^2}{(\hat{\nu}^{(1)} - \nu)^2}. \tag{10.5}$$

$R(\mathbf{w})$ is the ratio of a posteriori "excess risk", the increase in the conditional squared estimation error of $\mathbf{f}$ that comes from having to estimate ν. The point here is that $R(\mathbf{w})$, a measure of estimation efficiency for $\mathbf{f}$, depends directly on how well $\hat{\nu}^{(1)}$ and $\hat{\nu}^{(2)}$ estimate ν. In the simulation experiment of Figure 1, hereforth called "Experiment 1," the median of $(\hat{\nu}^{(2)} - \nu)^2/(\hat{\nu}^{(1)} - \nu)^2$ equalled 1.58, agreeing reasonably well with our previous efficiency comparisons.

Formula (10.5) was checked for Experiment 1. Let $R(\mathbf{w})$ and $\widehat{R}(\mathbf{w})$ be the exact and approximate ratios in (10.5), the exact value being evaluated from (10.2); their logs had mean and standard deviation

$$\log\{R(\mathbf{w})\} = 0.78 \pm 2.47, \qquad \log \widehat{R}(\mathbf{w}) = 0.72 \pm 2.59 \tag{10.6}$$

and correlation 0.986. $R(\mathbf{w})$ exceeded 1, indicating a preference for GML over C_p, in 405 of the 600 trials.

A different, unconditional, comparison is obtained by taking the expectation over $\mathbf{z}$ in (10.7) but ignoring the correlations between $\hat{\lambda}(\mathbf{w})$ and the individual w_i,

$$
\begin{aligned}
E_\lambda\left\{\frac{\|\mathbf{f}-\hat{\mathbf{f}}_{\hat{\lambda}^{(2)}}\|^2}{\sigma^2}-\frac{\|\mathbf{f}-\hat{\mathbf{f}}_{\hat{\lambda}^{(1)}}\|^2}{\sigma^2}\right\} &\doteq E_\lambda\sum\left\{\frac{(b_{\hat{\lambda}^{(2)}i}-b_{\lambda i})^2}{b_{\lambda i}}-\frac{(b_{\hat{\lambda}^{(1)}i}-b_{\lambda i})^2}{b_{\lambda i}}\right\}\\
&\doteq q_\lambda E_\lambda\{(\hat{\nu}^{(2)}-\nu)^2-(\hat{\nu}^{(1)}-\nu)^2\}\\
&\doteq \frac{q_\lambda}{i_\nu}\left[\frac{1}{E(p_1,p_2)^2}-1\right],
\end{aligned}
\tag{10.7}
$$

with $q_\lambda=(\sum a_{\lambda i}^2 b_{\lambda i})/(\sum a_{\lambda i}b_{\lambda i})^2$. Here we have used (11.3), (5.5), (5.7) and $E_\lambda\{w_i\}=1/b_{\lambda i}$. For Experiment 1 the last expression in (10.7) yields 0.46 as the estimated excess squared error risk for C_p compared to GML.

An exact version of (10.7) that takes into account the correlation between $\hat{\lambda}(\mathbf{w})$ and w_i is based on the following lemma, taken from Section 2 of Efron and Morris (1973).

Lemma.

$$
E_\lambda\{(b_{\hat{\lambda}(\mathbf{w})i}-b_{\lambda i})^2 w_i\}=E_\lambda^{(i)}\left\{\frac{(b_{\hat{\lambda}(\mathbf{w})i}-b_{\lambda i})^2}{b_{\lambda i}}\right\},
\tag{10.8}
$$

where $E_\lambda^{(i)}$ *indicates expectation with respect to*

$$
w_i\sim\chi_3^2/b_{\lambda i}\ \textit{independently of}\ w_j\overset{\text{ind}}{\sim}\chi_1^2/b_{\lambda j}\ \textit{for}\ j\neq i.
\tag{10.9}
$$

Combined with (10.1), (10.2), *the lemma gives*

Theorem 5.

$$
E_\lambda\left\{\frac{\|\mathbf{f}-\hat{\mathbf{f}}_{\hat{\lambda}(\mathbf{w})}\|^2}{\sigma^2}\right\}=\nu+\sum E_\lambda^{(i)}\left\{\frac{(b_{\hat{\lambda}(\mathbf{w})i}-b_{\lambda i})^2}{b_{\lambda i}}\right\}.
\tag{10.10}
$$

A decision-theoretic approach to minimizing $E_\lambda\|\mathbf{f}-\hat{\mathbf{f}}_{\hat{\lambda}(\mathbf{w})}\|^2$ would find estimators $\hat{\lambda}_{(i)}(\mathbf{w})$ that were optimal in some sense with respect to the loss function $(b_{\hat{\lambda}i}-b_{\lambda i})^2/b_{\lambda i}$. [Notice that we could use different estimators $\hat{\lambda}_{(i)}(\mathbf{w})$ for different components i.] A simpler expedient is to estimate λ in the ith case by the MLE with respect to (10.9). The same calculation as in (3.22) shows that the ith MLE $\hat{\lambda}_{(i)}$ satisfies (3.23) for the $(n+2)$-vector $\mathbf{w}_{(i)}$,

$$
\mathbf{w}_{(i)}=(w_1,w_2,\dots,w_{i-1},w_i/3,w_i/3,w_i/3,w_{i+1},\dots,w_n).
\tag{10.11}
$$

Dividing w_i into three smaller parts usually makes $\hat{\lambda}_{(i)}$ bigger than the original GML estimate $\hat{\lambda}$, and $\hat{\nu}_{(i)}$ smaller than $\hat{\nu}$. A Taylor series expansion gives an approximation based on $\ddot{l}_{\hat\lambda}(\mathbf{w})$, (3.22),

$$\hat\lambda_{(i)} \doteq \hat\lambda \cdot [1 + a_{\hat\lambda i}/(-\ddot{l}_{\hat\lambda}(\mathbf{w}))] \tag{10.12}$$

Since the GML estimator $\hat\nu$ itself tends to underestimate ν there is not much incentive for actually using $\hat\nu_{(i)}$ in place of $\hat\nu$.

We can think of Experiment 1 as an efficient way to compare the performance of GML and C_p in estimating $\mathbf{f}$ over a wide variety of cases, namely for 600 independent $\mathbf{f}$ vectors selected according to $\mathbf{f} \sim N(0, \sigma^2 C_\lambda)$, (7.2). The superiority of GML in this comparison is almost guaranteed by its status as the empirical Bayes MLE. This does not mean that GML is better for every $\mathbf{f}$. Averaging over different ensembles of $\mathbf{f}$'s could easily tip the comparison in favor of C_p, as suggested by the simulation in Figure 4.

Section 3 of Stein (1990) gives a different, and elegant, approach to results like (10.5) and Theorem 5. Working in the $\mathscr{F}^{(1)}$ context, with the sampling points x_i equally spaced, Stein shows that in a precise sense the asymptotic excess prediction risk using C_p is twice that for GML. (This result applies to linear rather than cubic splines.)

Wahba (1985) presents nine small sample simulations as a supplement to her asymptotic arguments favoring C_p over GML [actually GCV (4.21) over GML (3.28).] Starting with a known $\mathbf{f}$, she generates $\mathbf{y} \sim N(\mathbf{f}, \sigma^2 I)$, computes $\hat{\mathbf{f}}_{\hat\lambda(1)}$ and $\hat{\mathbf{f}}_{\hat\lambda(2)}$ and compares them as estimates of $\mathbf{f}$. Ten $\mathbf{y}$'s are generated for each of three different $\mathbf{f}$'s and three different σ's, but it is the choice of $\mathbf{f}$ that dominates the results, from the point of view of Figure 1. Wahba's experiment concerns three vectors $\mathbf{f}$ rather than 600. The three $\mathbf{f}$'s are interesting, representing unimodal, bimodal and trimodal linear combinations of beta functions. All three favor C_p over GML, but not more decisively than many of the realizations of Experiment 1.

It is important to notice that *the results in this paper do not depend on* "$\mathbf{f}$." Everything, from the geometric characterization of Figures 2 and 3 to the standard errors, efficiencies and curvatures of Sections 5 and 6, is determined by U and $\mathbf{a}_\lambda$ in (2.3). In the smoothing spline case U and $\mathbf{a}_\lambda$ are themselves determined by the covariate vector $\mathbf{x}$. Moreover, as discussed in Section 7, our numerical results are quite forgiving of substantial changes in $\mathbf{x}$.

This point of view has a great advantage: it isolates the estimation of the smoothing parameter λ, eliminating $\mathbf{f}$ as a nuisance parameter. In some ways, though, this simplification is an oversimplification. Current work by the author [Efron (2000)] compares C_p and GML in the standard frequentist model where $\mathbf{f}$ is fixed. In this context the possible biases of GML play a major role, making C_p more attractive, though it is still badly flawed by the reversal instabilities of Section 6.

11. Remarks, details and summary. Remarks on our results, including some proofs, details, and technical points, appear in this section, which concludes with a summary of the paper's main ideas.

A. *Positive stable laws.* An exponential family $d_\eta(w) = e^{\eta w - \psi(\eta)} d_o(w)$ having $\eta = -c_o b^p$ and $\mu = 1/b$, must satisfy $\psi'(\eta) = 1/b = (-\eta/c_o)^{-1/p}$, and so

$$\psi(\eta) = -\frac{c_o}{\alpha}\left(\frac{-\eta}{c_o}\right)^{\alpha} \quad [\alpha = (p-1)/p] \tag{11.1}$$

as in (4.4)–(4.7). If w is a positive variate we have

$$\int_0^\infty e^{\eta w} d_o(w) = e^{\psi(\eta)} = e^{-(c_o^{1-\alpha}/\alpha)\cdot(-\eta)^\alpha}. \tag{11.2}$$

According to Feller [(1971), Theorem 1 of Section XIII.6], $d_o(w)$ must then be the density of a positive stable law of order $\alpha = (p-1)/p$.

For $c_o = 1$, (12.2) becomes $\int_0^\infty e^{\eta w} d_o(w) = e^{-(-\eta)^\alpha/\alpha}$. Changing variables to $\tilde{w} = c_1 w$ and $\tilde{\eta} = \eta/c_1$, with $c_1 = c_o^{(1-\alpha)/\alpha}$, gives

$$\int_0^\infty e^{\tilde{\eta}\tilde{w}} d_o(\tilde{w}/c_1)/c_1 = e^{-(c_o^{1-\alpha}/\alpha)\cdot(-\tilde{\eta})^\alpha}. \tag{11.3}$$

Laplace transform theory says that $d_o(\tilde{w}/c_1)/c_1$ in (11.3) must represent the same density as $d_o(w)$ in (11.2). In other words, *the choice of c_o in* (4.2) *or* (4.9) *scales the carrier density $d_o(\cdot)$ by a multiplicative factor of $c_o^{(1-\alpha)/\alpha}$.*

For $p = 2$, $\alpha = 1/2$, $d_o(w)$ is the "inverse Gaussian" density [Feller (1971), Section XIII.3] defined in terms of $\phi(x) = e^{-x^2/2}/\sqrt{2\pi}$ by

$$d_o(w) = (2c_o/w^3)^{1/2}\phi\left(\left(\frac{2c_o}{w}\right)^{1/2}\right); \tag{11.4}$$

$d_b^{(2)}(w) = e^{-c_o[b^2 w - 2b]} d_o(w)$ is the density (4.2) appearing in Figure 4, used with $c_o = 1.334$ to generate the $\mathbf{w}$ vectors in the simulation of Figure 4.

B. *Repeated sampling interpretation of c_o.* The choice of the constant c_o in exponential family (4.2) affects the variance but not the expectation of $\mathbf{w}$, (4.7), which suggests that c_o has a sample size interpretation. The following result is easy to verify for positive integer c_o: if $\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_{c_o}$ are independently distributed according to density (4.2) with $c_o = 1$, then $\bar{\mathbf{w}} = \sum \mathbf{w}_l/c_o$ has density (4.2) with constant c_o.

C. $\lim\{\mathscr{F}^{(p)}\} = \mathscr{F}^{(1)}$. Writing $d_b^{(p)}(w)$ in (4.2) as

$$d_b^{(p)}(w) = e^{-c_o[b^p w - (b^{p-1}-1)\alpha]} e^{c_o/\alpha} d_o^{(p)}(w) \tag{11.5}$$

and letting $p \to 1$, $c_o = 1/2$, the exponent in (12.5) goes to $-\frac{1}{2}[bw - \log(b)]$ as in the $\mathscr{F}^{(1)}$ density (4.1).

D. *Computational methods for smoothing splines.* The Splus command smoothspline(x, y, spar $= \lambda$) produces $\hat{\mathbf{f}} = A_\lambda \mathbf{y}$ but not the matrix A_λ. However, setting $\mathbf{y} = \mathbf{e}_i$, the ith coordinate vector, yields $\hat{\mathbf{f}}$ equal to the ith column of A_λ, giving A_λ by successive use of $\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n$. A standard eigendecomposition of A_λ then gives U and $\mathbf{a}_\lambda$ in (2.3). This need only be done for one value of λ, say λ_o, since

$$\mathbf{a}_\lambda = \mathbf{a}_{\lambda_o} \Big/ \left[\mathbf{a}_{\lambda_o} + \frac{\lambda}{\lambda_o}(\mathbf{1} - \mathbf{a}_{\lambda_o})\right], \tag{11.6}$$

and U does not depend on λ.

E. *Fixed-frame smoothers.* The fact that U in (2.3) does not depend on λ, forming a "fixed frame" for the transformations (2.7), is essential to our methods. Less important is the splinelike choice of eigenvalues $a_{\lambda i} = [1 + \lambda k_i]^{-1}$, (2.4). This produces convenient algebraic expressions such as (3.19) and makes the line of expectations in Figure 2 and 3 genuinely linear, but is not crucial to the computations. We can proceed more generally, for example, with the equation $\sum b_{\lambda i}(w_i - 1/b_{\lambda i}) = 0$ replacing (3.23), and still carry out the calculations, albeit with increased numerical difficulties.

Hastie's 1996 paper on pseudosplines begins with any convenient fixed frame ("seed basis"), for example orthogonal polynomials in $\mathbf{x}$, and goes on to construct approximations of form (2.3) for a general family of smoothers.

F. *The estimator $\hat{\nu}^{(1.5)}$.* In an obvious sense $\hat{\nu}^{(1.5)}$ is a compromise estimator, conceivably more robust than either $\hat{\nu}^{(1)}$ or $\hat{\nu}^{(2)}$. Both experiments in Figure 1 were also analyzed using $\hat{\nu}^{(1.5)}$, but with inconclusive results: $\hat{\nu}^{(1.5)}$ outperformed $\hat{\nu}^{(2)}$ modestly in both simulations, but still gave eccentric results in Experiment 1. The cross-curvature $\gamma(1, 1.5) = 0.78$ is less than $\gamma(1, 2) = 1.19$ but still big enough to suggest serious LRR effects in (6.13).

G. *Complete repeated sampling.* A complete repeated sampling version of (7.2) is based on the model

$$\begin{aligned} &\mathbf{f}_l \sim N(\mathbf{0}, \sigma^2 C_\lambda) \quad \text{and} \\ &\mathbf{y}_l | \mathbf{f}_l \sim N(\mathbf{f}, \sigma^2 I) \text{ independently for } l = 1, 2, \dots, m. \end{aligned} \tag{11.7}$$

We observe $\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_m$ and wish to infer λ. This differs from (7.3), (7.4) in that there are repeated (unobservable) realizations of $\mathbf{f}$ as well as of $\mathbf{y}$. Each $\mathbf{y}_l$ gives $\mathbf{z}_l = U'\mathbf{y}_l/\sigma$ and $\mathbf{w}_l = \mathbf{z}_l^2$, where the $\mathbf{w}_l$ vectors are independently distributed according to (3.8). The average vector $\bar{\mathbf{w}} = \sum \mathbf{w}_l/m$ is sufficient for λ. It follows essentially the same curved exponential family as before, (3.8), except that instead of (3.7) we have $\bar{w}_i \overset{\text{ind}}{\sim} \chi^2_m/(m b_{\lambda i})$.

Figure 2, with $\mathbf{w}$ replaced by $\bar{\mathbf{w}}$, applies exactly as drawn. As m gets large, $\bar{\mathbf{w}}$ moves closer to the line of expectations since

$$\bar{\mathbf{w}} \sim (\boldsymbol{\mu}_\lambda,\ \operatorname{diag}(2/m b^2_{\lambda i})). \tag{11.8}$$

In other words we obtain the nice "local" behavior of maximum likelihood estimation, leading to the usual asymptotic optimality properties, obviating concerns about nonlocal pathology due to the local reversal region or the end zone.

The trouble with model (11.7) is that it does not apply to the usual smoothing situation, where more data collection can provide more data vectors $\mathbf{y}_l$, but all of which still refer to the same unknown $\mathbf{f}$. This leads to the partial repeated sampling model of Section 7, which does *not* enjoy the usual asymptotic properties of maximum likelihood estimation.

H. *Estimating equation unbiasedness.* The estimators $\hat{\lambda}^{(p)}$, (4.12), and $\hat{\nu}^{(p)}$ enjoy a form of unbiasedness relating to estimating equations. For any vector $\mathbf{b}$ having positive components, define

$$Q_{\mathbf{b}}^{(p)}(\mathbf{w}) = \sum \left[b_i^p w_i - \frac{p}{p-1} b_i^{p-1} \right], \tag{11.9}$$

so $\hat{\lambda}^{(p)} = \operatorname{argmin}_\lambda \{Q_{\mathbf{b}_\lambda}^{(p)}(\mathrm{w})\}$. Suppose that the true expectation of $\mathbf{w}$ is $1/\mathbf{b}_o$ for some vector $\mathbf{b}_o$, as in (3.12) for instance. Then the expectation of Q is $E_{\mathbf{b}_o}\{Q_{\mathbf{b}}^{(p)}(\mathbf{w})\} = \sum(b_i^p/b_{oi} - (p/(p-1))b_i^{p-1})$, having gradient vector

$$\nabla_{\mathbf{b}} E_{\mathbf{b}_o}\{Q_{\mathbf{b}}^{(p)}(\mathbf{w})\} = p\left(\ldots, b_i^{p-2}\left[\frac{b_i}{b_{oi}} - 1\right], \ldots\right)'. \tag{11.10}$$

It is easy to see from (12.10) that $E_{\mathbf{b}_o}\{Q_{\mathbf{b}}^{(p)}(\mathbf{w})\}$ is minimized at $\mathbf{b} = \mathbf{b}_o$.

This means that the estimating equation $Q_{\mathbf{b}_\lambda}^{(p)}(\mathbf{w})$, whose minimization gives $\hat{\lambda}^{(p)}$, has its minimum expected value at $\mathbf{b}_{\lambda_o}$ when λ_o is the true parameter value, no matter which family $\mathscr{F}^{(p_o)}$ is giving $\mathbf{w}$. This kind of "estimating equation unbiasedness" bolsters our contention that all the estimators $\hat{\lambda}^{(p)}$ are estimating the same thing. Estimating equation unbiasedness is not a universal property; it fails for instance if we change the C_p criterion (4.13) to $\|\mathbf{y} - \hat{\mathbf{f}}_\lambda\|^2 + c\sigma^2 \operatorname{tr}(A_\lambda)$ for some value other than $c = 2$.

I. *Better estimates of standard error.* The first-order approximation for the standard deviation of an MLE, $sd(\hat{\nu}) \doteq 1/i_\nu^{1/2}$, (5.5), is extended to higher order in formula (10.1) of Efron (1975),

$$sd(\hat{\nu}) \doteq \frac{1}{i_\nu^{1/2}}[1 + \gamma_\nu^2 + 4\Gamma_\nu^2/i_\nu]^{1/2}, \tag{11.11}$$

where γ_ν^2 is the squared curvature $\gamma_{\lambda(\nu)}^2$, (6.4), and Γ_ν is a type of "naming curvature" having to do with the relationship between ν and the optimum local representation of $\mathscr{F}^{(p)}$. [Formula (12.11) does not include a bias term discussed in Efron (1975).] It can be shown that for the MLE $\hat{\nu}^{(p)}$ in $\mathscr{F}^{(p)}$,

$$\Gamma_\nu^2 = \frac{(\sum a_{\lambda i}^2 b_{\lambda i})^2}{2(\sum a_{\lambda i} b_{\lambda i})^4}, \tag{11.12}$$

not depending on p. In the case $p = 1$, $\nu = 8$, (12.11) gives

$$sd(\hat{\nu}) = 1.195 \cdot [1 + 0.0760 + 0.1331]^{1/2} = 1.32,$$

compared to the simulation estimate 1.37.

J. *Cross-Validation.* Ordinary cross-validation, as opposed to GCV, (8.21), is often advocated as a selection criterion; see Section 3.2 of Green and Silverman (1994). Section 7 of Efron (1986) discusses why this is appropriate when the pairs (x_i, y_i) are thought of as randomly sampled from a bivariate distribution, but not in the regression context of this paper where $\mathbf{x}$ is considered fixed.

K. *Proof of Theorem 3.* For a given value of λ we can standardize the curved exponential family (4.5) to have $\mu_\lambda = 0$ and $V_\lambda = I$ via the linear transformations

$$\mathbf{w} \to V_\lambda^{-1/2}(\mathbf{w} - \mu_\lambda) \quad \text{and} \quad \eta \to V_\lambda^{1/2}\eta \tag{11.13}$$

without changing the likelihood as a function of λ, or values of i_λ, M_λ, γ_λ or $\gamma_\lambda^2(p_0, p)$. In this coordinate system $i_\lambda = \|\dot{\eta}_\lambda\|^2$ and (6.1) becomes

$$M_\lambda = \begin{pmatrix} \|\dot{\eta}_\lambda\|^2 & \|\dot{\eta}_\lambda\|\|\ddot{\eta}_\lambda\|\mathrm{Cos} \\ \|\dot{\eta}_\lambda\|\|\ddot{\eta}_\lambda\|\mathrm{Cos} & \|\ddot{\eta}_\lambda\|^2 \end{pmatrix} \quad \text{where } \mathrm{Cos} = \frac{\dot{\eta}_\lambda'\ddot{\eta}_\lambda}{\|\dot{\eta}_\lambda\|\|\ddot{\eta}_\lambda\|}. \tag{11.14}$$

Moreover, comparing (6.2) and (6.6) gives

$$\gamma_\lambda^2 = \frac{\|\dot{\eta}_\lambda\|^2\|\ddot{\eta}_\lambda\|^2(1 - \mathrm{Cos}^2)}{\|\dot{\eta}_\lambda\|^6} = \frac{\|\mathbf{o}_\lambda\|^2}{i_\lambda^2}. \tag{11.15}$$

The observed Fisher information in a curved exponential family is $-\ddot{l}_\lambda(\mathbf{w}) = i_\lambda - \ddot{\eta}_\lambda'(\mathbf{w} - \mu_\lambda)$, equaling $i_\lambda - \ddot{\eta}_\lambda'\mathbf{w}$ now that $\mu_\lambda = 0$. For $\mathbf{w}$ in $\mathscr{L}_\lambda$, $\ddot{\eta}_\lambda'\mathbf{w} = \mathbf{o}_\lambda'\mathbf{w}$ by orthogonality so, since T in (6.7) is invariant under transformations (11.13),

$$-\ddot{l}_\lambda(\mathbf{w}) = i_\lambda\left[1 - \frac{\mathbf{o}_\lambda'}{i_\lambda}\mathbf{w}\right] = i_\lambda[1 - T], \tag{11.16}$$

verifying (6.8). It is clear from (11.16) that the shortest vector $\mathbf{w}$ in $\mathscr{L}_\lambda$ satisfying $-\ddot{l}_\lambda(\mathbf{w}) = 0$ must be

$$\mathbf{w}_0 = \frac{i_\lambda}{\|\mathbf{o}_\lambda\|^2}\,\mathbf{o}_\lambda = \frac{1}{\gamma_\lambda}\,\frac{\mathbf{o}_\lambda}{\|\mathbf{o}_\lambda\|}, \tag{11.17}$$

using (11.15), which verifies (6.9). Finally, (6.10) follows directly from definition (6.7) and $\mathrm{cov}_\lambda^{(p_0)}(\mathbf{w}) \equiv V_\lambda^{(p_0)}$.

In the standardized coordinates (11.13), the extension (6.12) of $\mathrm{LRR}_\lambda^{(p)}$ outside of $\mathscr{L}_\lambda$ is made parallel to $\dot{\eta}_\lambda$,

$$\mathrm{LRR}_\lambda^{(p)} = \{\dot{l}_\lambda(\mathbf{w}) = 0 \text{ and } T \geq 1\} \oplus c\,\dot{\eta}_\lambda, \tag{11.18}$$

$c \in (-\infty, \infty)$. This gives good results in simple situations such as Fisher's circle model [Efron (1978)], but is justified here mainly by simulations like those in Figure 6.

L. *Extending the Families* $\mathscr{F}^{(p)}$. The curved exponential families $\mathscr{F}^{(p)}$, (4.9), can be extended in a natural way into the end zone. Letting $\theta \equiv 1/\lambda$ in (2.4), we can write the components of $\mathbf{a}_\lambda$ as

$$a_{\theta i} = \frac{\theta}{\theta + k_i} \text{ for } i = 1, 2, \ldots, n. \tag{11.19}$$

The components of $\mathbf{b}_\theta = \mathbf{1} - \mathbf{a}_\theta = \mathbf{k}/(\theta + \mathbf{k})$ are positive for

$$\theta \in (-k_{\min}, \infty), \tag{11.20}$$

where $k_{\min}$ is the minimum positive k_i value, $k_{\min} = k_3$ for smoothing splines. Using $b_{\theta i}^p$ in (4.9), the curved family $\mathscr{F}^{(p)}$ is now defined for θ in $(-k_{\min}, \infty)$, which extends the line of expectations $\{\boldsymbol{\mu}_\theta = \mathbf{1} + \theta/\mathbf{k}\}$ into the end zone.

We can now have estimates $\hat{\theta}$ less than zero in Figure 2, corresponding to $\mathbf{w}$ vectors in the end zone. This is useful way to quantify how *far* in the end zone $\mathbf{w}$ may lie.

Summary. This paper compares the C_p and GML criteria for selecting the smoothing parameter of a scatterplot smoother. The use of curved exponential families permits a finite-sample analysis, avoiding the mathematical and definitional difficulties of asymptotics for the smoother problem, at the expense of considerably less generality.

Some main conclusions.

1. Both C_p and GML have simple geometric descriptions, as shown in Figures 2 and 3, that are independent of any probability models for the data.
2. Both C_p and GML are maximum likelihood estimates, each within its own one-parameter curved exponential family. These two are members of a class of curved exponential families described in Section 4, each member of which defines its own selection criterion.
3. Exponential family theory gives simple approximations for the standard errors and efficiencies of the C_p and GML estimators; see Section 5.
4. The GML curved family, described in Section 3, is based on a familiar normal-theory hierarchical (or empirical Bayesian) model. GML performs well within this family as shown in the left panel of Figure 1 and is its recommended selection criterion, perhaps after augmentation by bias corrections.
5. GML can be badly biased toward oversmoothing when the data is sampled from the curved family for which C_p is the MLE, as shown in the right panel of Figure 4.
6. The C_p exponential family, which offers the only maximum likelihood justification for the C_p estimate, is useful for challenging GML but not really believable in its own right as a sampling model for the data in a smoothing problem, as seen in the top panel of Figure 4, for instance.
7. The C_p estimator performs erratically even within its own exponential family. Section 6 shows that the trouble lies in that family's very large curvature, which produces the bad estimation effects seen in Figure 6.

REFERENCES

ANDERSSEN, R. and BLOOMFIELD, P. (1974). A time series approach to numerical differentiation. *Technometrics* **16** 69–75.

EFRON, B. (1975). Defining the curvature of a statistical problem (with applications to second order efficiency). *Ann. Statist.* **3** 1189–1242.

EFRON, B. (1978). The geometry of exponential families. *Ann. Statist.* **6** 362–376.

EFRON, B. (1982). Maximum likelihood and decision theory. *Ann. Statist.* **10** 340–356.

EFRON, B. (1986). How biased is the apparent rate of a prediction rule? *J. Amer. Statist. Assoc.* **8** 461–470.

EFRON, B. (1999). Selection criteria for scatterplot smoothers. Technical Report 207, Stanford Univ.

EFRON, B. (2000). Smoothers and the cost of model selection. Technical Report 209, Stanford Univ.

EFRON, B. and MORRIS, C. (1973). Stein's estimation rule and its competitors—an empirical Bayes approach. *J. Amer. Statist. Assoc.* **68** 117–130.

EFRON, B. and TIBSHIRANI, R. (1998). The problem of regions. *Ann. Statist.* **26** 1687–1718.

EUBANK, R. (1988). *Spline Smoothing and Nonparametric Regression*. Dekker, New York.

FELLER, W. (1971). *An Introduction To Probability Theory and Its Applications* **2**. Wiley, New York.

GREEN, P. and SILVERMAN, B. (1994). *Nonparametric Regression and Generalized Linear Models*. Chapman and Hall, London.

HALL, P. and JOHNSTONE, I. (1992). Empirical functionals and efficient smoothing parameter selection. *J. Roy. Statist. Soc. Ser. B* **54** 475–530.

HASTIE, T. (1996). Pseudosplines. *J. Roy. Statist. Soc. Ser. B* **58** 379–396.

HASTIE, T. and TIBSHIRANI, R. (1990). *Generalized Additive Models*. Chapman and Hall, London.

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

MALLOWS, C. (1973). Some comments on C_p. *Technometrics* **15** 661–667.

PATTERSON, H. and THOMPSON, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika* **58** 545–554.

STEIN, C. (1981). Estimating the mean of a multivariate normal distribution. *Ann. Statist.* **9** 1135–1151.

STEIN, M. (1990). A comparison of generalized cross-validation and modified maximum likelihood for estimating the parameters of a stochastic process. *Ann. Statist.* **18** 1139–1157.

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

WAHBA, G. (1990). *Spline Models For Observational Data*. SIAM, Philadelphia.

WECKER, W. and ANSLEY, C. (1983). The signal extraction approach to non-linear regression and spline smoothing. *J. Amer. Statist. Assoc.* **78** 81–89.

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
SEQUOIA HALL
STANFORD, CALIFORNIA 94305-4065
E-MAIL: brad@stat.stanford.edu