Smoothers and the C_p, Generalized Maximum Likelihood, and Extended Exponential Criteria: A Geometric Approach

S. C. KOU and Bradley EFRON

Nonparametric regression, often called smoothing, is a widely used data analysis method. The use of a smoother requires the choice of a smoothing parameter that by balancing fidelity and roughness controls how much smoothing is done. Two popular selection criteria for choosing the smoothing parameter are C_p and generalized maximum likelihood (GML). Each of these has its own problems. For C_p, the problem is its high variability, whereas for GML, the problem is its potentially large bias. By studying the geometry of selection criteria, we give an intuitive explanation of the strength and weakness of C_p and GML. The geometry then motivates a new selection method, the extended exponential (EE) criterion, which combines the strength of C_p and GML but mitigates their weaknesses in terms of variability, bias, and undersmoothing.

KEY WORDS: Bias; Curvature; Degrees of freedom; Geometry; Prediction error; Reversal effects; Spline-like smoothers; Variability.

1. INTRODUCTION

Regression is a fundamental problem in statistics; one observes pairs $\{(x_i, y_i),\ i = 1, 2, \ldots, n\}$ and wants to estimate the regression function of y on x. The classical approach fits a polynomial to the data. The alternative method, nonparametric regression, the method considered in this article, approaches the problem under the mild assumption that $f(x)$ is a smooth function of x, without imposing parametric restrictions about the functional dependence (see, e.g., Eubank 1988; Härdle 1990; Hastie and Tibshirani 1990; Wahba 1990; Rosenblatt 1991; Green and Silverman 1994; Simonoff 1996; Bowman and Azzalini 1997).

In practice, the performance of many nonparametric procedures depends critically on the choice of a smoothing parameter that determines how locally the smoothing should be done. This article explores the problem of selecting the appropriate smoothing parameter. There is an impressive literature on choosing smoothing parameters, most of which is written from a very general large-sample perspective (see, e.g., Wahba 1985; Li 1986, 1987; Hall and Johnstone 1992; Jones, Marron, and Sheather 1996; Hurvich, Simonoff, and Tsai 1998). In this article we take a more specialized approach, concentrating on spline-like smoothers and the small-sample properties of selection criteria. At the center of our small-sample study is a geometric interpretation of selection criteria that not only leads to simple formulas that predict the accuracy of competing selection criteria, but also motivates a new selection criterion, the *extended exponential* (EE) criterion, which in some ways has more desirable properties than the two popular criteria C_p and generalized maximum likelihood (GML). Indeed, the term "extended exponential" itself comes from the geometric interpretation; the new criterion can be viewed geometrically as coming from an *exponential* family *extended* from the C_p and GML families. This point will become clear in Section 5 (see Remark 5).

The C_p criterion (Mallows 1973), which chooses the smoothing parameter by minimizing an unbiased estimate of the prediction error, is perhaps the most popular smoothing methodology, if one includes its close cousins such as the generalized cross-validation (GCV) (Craven and Wahba 1979) and the Akaike information criterion (AIC) (Akaike 1974). (The close relationship of GCV and C_p is described in sec. 7 of Efron 1986 and also in sec. 4 of Efron 2001.) Despite its popularity, C_p can be highly variable; it occasionally selects a very wiggly curve even when the true underlying curve is known to be smooth (see, e.g., Hurvich et al. 1998). On the other hand, GML, another selection criterion suggested by Wecker and Ansley (1983) from an empirical Bayes framework and studied by Wahba (1985) and Stein (1990), behaves more stably; rarely would it choose some curve much wigglier than the true underlying curve. GML can have serious problems with bias, however.

Figures 1 and 2 illustrate the results of two simulation experiments. In each experiment, 1000 datasets were generated from a curve shown in panel (a) of each figure (also shown are the generated points $\{(x_i, y_i)\}$ for one particular dataset). The C_p and GML criteria were then applied to these datasets to choose the degrees of freedom (a quantity closely related to the smoothing parameter; see Sec. 2) of the *smoothing-spline* fitted curve. The rightmost two panels in Figures 1(b) and 2(b) show the histograms of the C_p and GML estimated degrees of freedom. In experiment 1 (as discussed in Sec. 2), the ideal degrees of freedom of the true curve are 5.18, and the GML criterion is seen to do a good job; the estimated degrees of freedom are concentrated on 5.18, whereas the C_p estimates of degrees of freedom spread out over a wide range. In experiment 2, the ideal degrees of freedom are 13.42. We see from Figure 2 that although the C_p estimates are still highly variable, they are centered around the right place, whereas the GML estimates are badly biased upward.

The geometric interpretation of C_p and GML provides an explanation of these empirical observations; that is, why the C_p criterion is so variable, whereas the GML criterion can

S. C. Kou is Assistant Professor, Department of Statistics, Harvard University, Cambridge, MA 02138 (E-mail: *kou@stat.harvard.edu*). Bradley Efron is Professor, Department of Statistics, Stanford University, Stanford, CA 94305 (E-mail: *brad@stat.stanford.edu*). The authors thank Iain Johnstone, Trevor Hastie, and Rob Tibshirani for helpful discussions. The authors are also grateful to the editor, the associate editor, and the referees for constructive suggestions that substantially improved the presentation of the article.

Journal of the American Statistical Association
September 2002, Vol. 97, No. 459, Theory and Methods
DOI 10.1198/016214502388618582

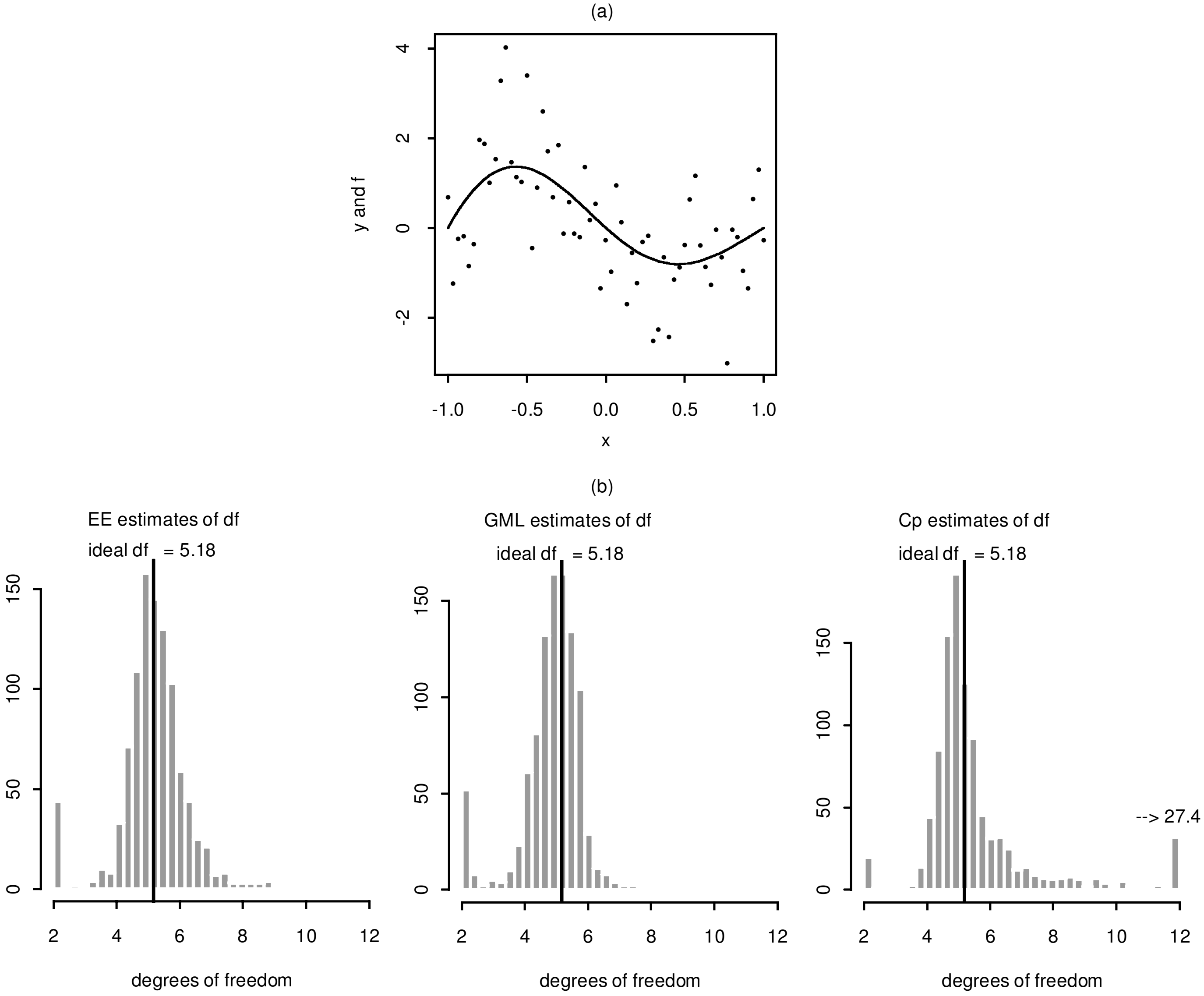

Figure 1. Simulation Experiment 1. (a) True curve and one particular dataset out of 1000 simulations. (b) Histograms of EE, GML, and C_p estimated degrees of freedom. (Histograms truncated at $df = 12$.)

have such a large bias. Roughly speaking, the variability of C_p stems from the fact that it suffers from a geometric instability, whereas the large bias of GML arises from the fact that GML is not Fisher consistent.

With the strength and weakness of C_p and GML delineated, we now propose a new selection criterion, the EE criterion, which combines the strength of the two while mitigating their defects. The new selection criterion, with its root in the geometry, tends to give smaller variance, smaller bias, and smaller tendency toward undersmoothing.

The left panels of Figures 1(b) and 2(b) show the EE estimated degrees of freedom in the two simulation experiments. In both examples, the EE estimates are significantly less variable than the C_p estimates, and the bias of EE is less than that of GML. In other words, the EE criterion behaves more "robustly"; it does not give eccentric estimates as C_p occasionally does, and, unlike GML, its bias stays reasonably in check even in some unfavorable situations.

This article is organized as follows. Section 2 briefly reviews spline-like smoothers and the selection criteria C_p and GML, then discusses the geometric interpretation of these criteria. Section 3 introduces the EE criterion and gives its geometric motivation. Sections 4 and 5 further explore the geometry to provide theoretical approximation of the bias and variance of the selection criteria, as well as theoretical analysis of the stability of selection criteria. Section 6 considers the error of estimating the curve, discussing the connection between estimating the curve and estimating the ideal degrees of freedom. Section 7 gives marginal-Bayesian interpretation of the EE criterion, and Section 8 summaries the results and provides some further discussion.

2. THE GEOMETRY OF C_p AND GENERALIZED MAXIMUM LIKELIHOOD

After giving a brief review of spline-like smoothers and the C_p and GML selection criteria, this section presents a geometric picture of C_p and GML that is used in subsequent analyses concerning variance, bias, and prediction and to motivate the EE selection criterion.

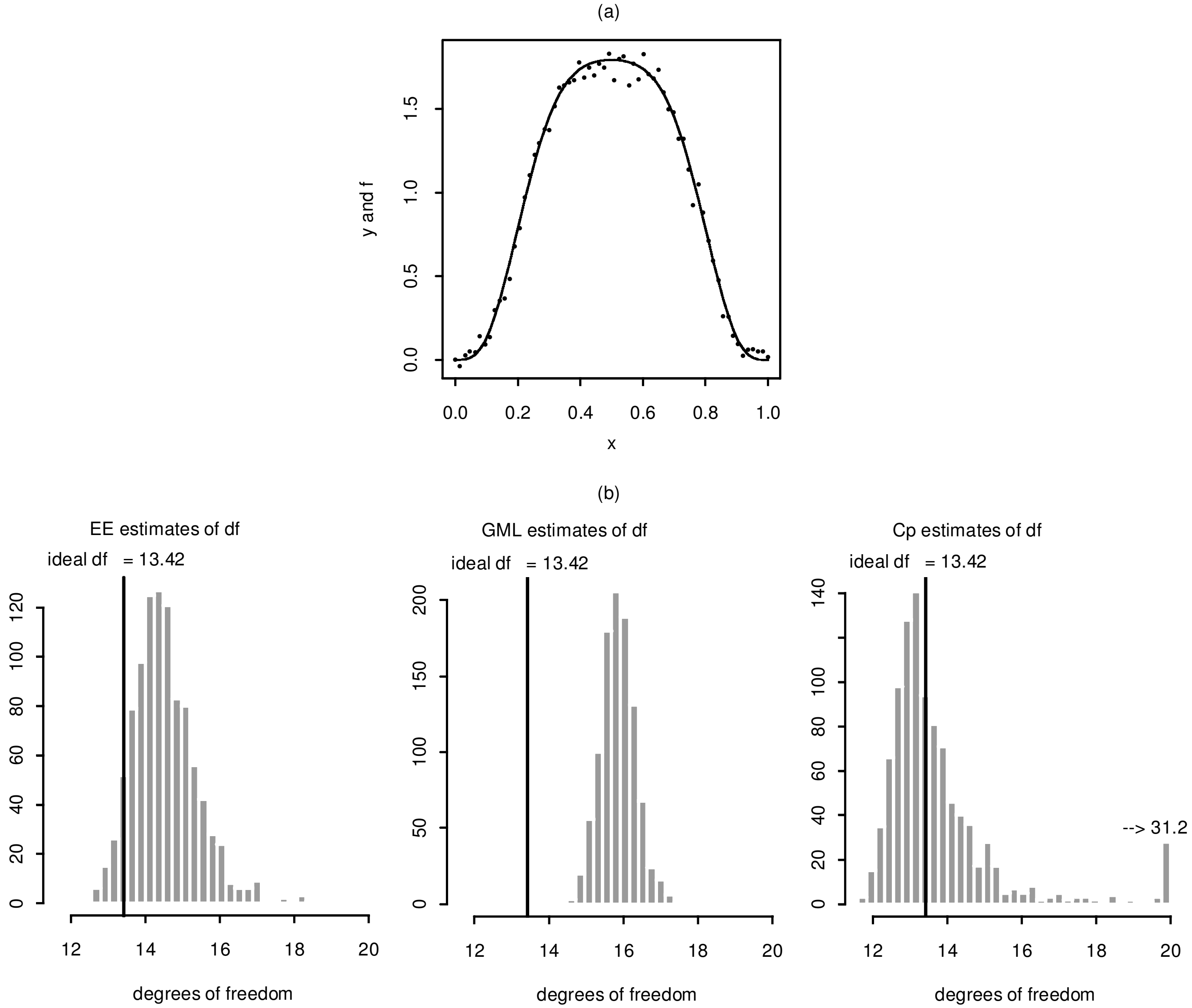

Figure 2. Simulation Experiment 2. (a) True curve and one particular sample out of 1000 simulations. (b) Histograms of EE, GML, and C_p estimated degrees of freedom. (Histograms truncated at $df = 20$.)

2.1 Spline-Like Smoothers and the C_p and Generalized Maximum Likelihood Selection Criteria

Smoothing starts with n observed data points $\{(x_i, y_i)\}_{i=1}^n$ in the plane, the goal being to estimate $f(x) = E(y|x)$, the regression function of y on x, usually nonparametrically. In this article we consider estimation of $f(x)$ at the "design points" x_i, say $f_i = f(x_i)$ using a *linear smoother*,

$$\hat{\mathbf{f}}_\lambda = A_\lambda \mathbf{y}, \qquad (1)$$

with $\mathbf{y} = (y_1, \ldots, y_n)'$ and $\hat{\mathbf{f}}_\lambda = (\hat{f}_{\lambda 1}, \ldots, \hat{f}_{\lambda n})'$, the vector that estimates $\mathbf{f} = (f_1, f_2, \ldots, f_n)' = (f(x_1), f(x_2), \ldots, f(x_n))'$. The entries of the $n \times n$ *smoothing matrix* A_λ depend on $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ and also on a nonnegative *smoothing parameter*, λ. A mnemonically helpful case is the moving average smoother with window width λ, (the "bandwidth"), for which $\hat{f}_{\lambda i}$ is the average of those y_j values having $|x_j - x_i| \le \lambda$. Usually, λ itself must be inferred from the data.

A *selection criterion* is a method of choosing λ on the basis of the data. The C_p *criterion* (Mallows 1973) chooses λ to minimize an unbiased estimate of total squared-error risk. Suppose that the y_i are uncorrelated, with mean f_i and constant variance σ^2, written as

$$\mathbf{y} \sim (\mathbf{f}, \sigma^2 I). \qquad (2)$$

Then the C_p estimate of λ is $\hat{\lambda}^{C_p} = \arg\min_\lambda \{C_\lambda(\mathbf{y})\}$, where the C_p statistic

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

is an unbiased estimate of $E \parallel \hat{\mathbf{f}}_\lambda - \mathbf{f} \parallel^2$, the squared prediction error. The notation $C_\lambda(\mathbf{y})$ assumes that $\mathbf{x}$ is fixed, as is usual in regression problems, and that σ^2 is known. The trace of the smoothing matrix $\mathrm{tr}(A_\lambda)$ is referred to as the *degrees of freedom*,

$$df_\lambda = \mathrm{tr}(A_\lambda),$$

agreeing with the standard definition when A_λ represents polynomial regression; it is a monotonic decreasing function of λ that is usually of more interest than λ itself. We assume that σ^2 is known; see Section 8 for the usual case in which σ^2 must be estimated from the data.

The GML criterion (Wecker and Ansley 1983), has a normal-theory empirical Bayesian motivation. If we strengthen (2) to $\mathbf{y} \sim N(\mathbf{f}, \sigma^2 I)$ and put a Gaussian prior on the underlying curve, $\mathbf{f} \sim N(\mathbf{0}, \sigma^2 A_\lambda (I - A_\lambda)^{-1})$, then, according to the Bayes theorem,

$$\mathbf{y} \sim N(\mathbf{0}, \sigma^2 (I - A_\lambda)^{-1}) \quad \text{and} \quad \mathbf{f}|\mathbf{y} \sim N(A_\lambda \mathbf{y}, A_\lambda). \tag{3}$$

The second relationship shows that $\hat{\mathbf{f}}_\lambda = A_\lambda \mathbf{y}$ is the Bayes estimate of $\mathbf{f}$ under squared error loss. The first relationship motivates the GML selection for the smoothing parameter λ, $\hat{\lambda}^{\mathrm{GML}}$ = maximum likelihood estimate (MLE) of λ based on $\mathbf{y} \sim N(\mathbf{0}, \sigma^2 (I - A_\lambda)^{-1})$.

One class of linear smoothers of particular interest in this article is *spline-like* smoothers (Efron 2001), in which the class of smoothing matrices $\{A_\lambda, 0 \le \lambda \le \infty\}$ has the form

$$A_\lambda = U \mathbf{a}_\lambda U', \tag{4}$$

where U is an $n \times n$ orthogonal matrix *not* depending on the smoothing parameter λ and $\mathbf{a}_\lambda = \mathrm{diag}(a_{\lambda i})$, a diagonal matrix with ith diagonal element

$$a_{\lambda i} = 1/(1 + \lambda k_i), \qquad i = 1, 2, \ldots, n, \tag{5}$$

the constants $\mathbf{k} = (k_1, k_2, \ldots, k_n)$, solely determined by $\mathbf{x}$, being nonnegative and nondecreasing. According to (5), spline-like smoothers achieve the goal of smoothing by shrinking the higher-frequency components of the response toward 0, shrinking more for larger values of λ and i.

One popular class of spline-like smoothers is *cubic smoothing splines*, which amount to making a particular choice of U and $\mathbf{k}$ in (4) and (5), (see Green and Silverman 1994, chap. 2). In this case, the first two columns of U span the space of linear functions of x; also, $k_1 = k_2 = 0$, making $a_{\lambda 1} = a_{\lambda 2} = 1$ for all λ, which says that linear functions of x are preserved by the smoother. Note that in this case the degrees of freedom, $df_\lambda = \sum_{i=1}^n a_{\lambda i}$, increase from 2 to n as λ decreases from ∞ to 0 in (5).

Other spline-like smoothers can be fashioned in a variety of ways, including the methods based on orthogonal series. For example, we might take the jth column of U to be the vector x_i^{j+1} Gram–Schmidt orthogonalized with respect to the lower powers. Then the smoother (5) would shrink the higher-powered components of the response toward 0, more so for larger values of λ (see Hastie 1996). The key assumption of spline-like smoothers is that all of the smoothing matrices A_λ are symmetric and have the same eigenvectors (i.e., the columns of U) for all λ. This allows a rotation of coordinates for the model $\mathbf{y} \sim (\mathbf{f}, \sigma^2 I)$, $\hat{\mathbf{f}}_\lambda = A_\lambda \mathbf{y}$ to

$$\mathbf{z} = U'\mathbf{y}/\sigma, \qquad \mathbf{g} = U'\mathbf{f}/\sigma, \qquad \hat{\mathbf{g}}_\lambda = U'\hat{\mathbf{f}}_\lambda/\sigma, \tag{6}$$

putting the smoother family (1) into diagonal form,

$$\mathbf{z} \sim (\mathbf{g}, I), \qquad \hat{\mathbf{g}}_\lambda = \mathbf{a}_\lambda \mathbf{z}. \tag{7}$$

Transformations (6) assume that σ is known; see Section 8 for the case where σ must be estimated from the data. Let $b_{\lambda i} = 1 - a_{\lambda i}$. In the new coordinate system, the C_p statistic can be expressed as a function of $\mathbf{z}^2$,

$$\begin{aligned} C_\lambda(\mathbf{z}^2) &= \|\mathbf{y} - \hat{\mathbf{f}}_\lambda\|^2 + 2\sigma^2 \mathrm{tr}(A_\lambda) - n\sigma^2 \\ &= \sigma^2 \sum_{i=1}^n (b_{\lambda i}^2 z_i^2 - 2b_{\lambda i}) + n\sigma^2. \end{aligned} \tag{8}$$

Define

$$\mathbf{w} = \mathbf{z}^2 = (z_1^2, z_2^2, \ldots, z_n^2)'; \tag{9}$$

then the C_p choice of λ is given by

$$\hat{\lambda}^{C_p} = \arg\min_\lambda \sum_i (b_{\lambda i}^2 w_i - 2b_{\lambda i}). \tag{10}$$

Remark 1. Because $k_i = 0$ in (5) implies $a_{\lambda i} = 1$ and $b_{\lambda i} = 0$ for all values of λ, those w_i with $k_i = 0$ do not contain any information about λ and thus do not enter into the C_p criterion (10). Hereafter, we use the notation "$\sum_i$" to indicate summation over coordinates having $a_{\lambda i} < 1$.

Similarly, the GML selection criterion can be simply described in the $(\mathbf{z}, \mathbf{g})$ coordinate system,

$$\begin{aligned} \hat{\lambda}^{\mathrm{GML}} &= \text{MLE of } \lambda \text{ based on } \mathbf{y} \sim N(\mathbf{0}, \sigma^2 (I - A_\lambda)^{-1}) \\ &= \text{MLE of } \mathbf{z} \sim N(\mathbf{0}, \mathbf{1}/\mathbf{b}_\lambda) = \text{MLE of } \mathbf{w} \sim \chi_1^2/\mathbf{b}_\lambda^2. \end{aligned}$$

This can be further simplified to $\hat{\lambda}^{\mathrm{GML}} = \arg\min_\lambda \sum_i (b_{\lambda i} w_i - \log b_{\lambda i})$, because the density of $\mathbf{w} \sim \chi_1^2/\mathbf{b}_\lambda^2$ is $d_\lambda(\mathbf{w}) = \exp(-\frac{1}{2}\sum_i (b_{\lambda i} w_i - \log b_{\lambda i}))/\prod_i \sqrt{2\pi w_i}$.

2.2 The Geometry of Generalized Maximum Likelihood and C_p

The fact that GML chooses λ as the minimizer of $\sum_i (b_{\lambda i} w_i - \log b_{\lambda i})$ leads to a simple geometric picture. If we define $\eta_{\lambda i}^{\mathrm{GML}} = -b_{\lambda i}$, $\boldsymbol{\eta}_\lambda^{\mathrm{GML}} = (\eta_{\lambda 1}^{\mathrm{GML}}, \eta_{\lambda 2}^{\mathrm{GML}}, \ldots, \eta_{\lambda n}^{\mathrm{GML}})'$, and $\psi_\lambda^{\mathrm{GML}} = -\sum_i \log b_{\lambda i}$, then $\hat{\lambda}^{\mathrm{GML}} = \arg\max_\lambda \{(\boldsymbol{\eta}_\lambda^{\mathrm{GML}})'\mathbf{w} - \psi_\lambda^{\mathrm{GML}}\}$ and $\hat{\lambda}^{\mathrm{GML}}$ must satisfy

$$\frac{\partial}{\partial \lambda}\left\{(\boldsymbol{\eta}_\lambda^{\mathrm{GML}})'\mathbf{w} - \psi_\lambda^{\mathrm{GML}}\right\}\Big|_{\lambda = \hat{\lambda}^{\mathrm{GML}}} = (\dot{\boldsymbol{\eta}}_\lambda^{\mathrm{GML}})'(\mathbf{w} - \boldsymbol{\mu}_\lambda)\Big|_{\lambda = \hat{\lambda}^{\mathrm{GML}}} = 0, \tag{11}$$

where $\boldsymbol{\mu}_\lambda = (\mu_{\lambda 1}, \mu_{\lambda 2}, \ldots, \mu_{\lambda n})'$ with $\mu_{\lambda i} = 1/b_{\lambda i}$, and $\dot{\boldsymbol{\eta}}_\lambda^{\mathrm{GML}}$ is the vector with components $\dot{\eta}_{\lambda i}^{\mathrm{GML}} = \partial \eta_{\lambda i}^{\mathrm{GML}}/\partial \lambda = -\partial/\partial\lambda\, b_{\lambda i}$, which can be further simplified for spline-like smoothers, because under (5),

$$\frac{\partial}{\partial \lambda} b_{\lambda i} = a_{\lambda i} b_{\lambda i}/\lambda \quad \text{and} \quad \frac{\partial^2}{\partial \lambda^2} b_{\lambda i} = -2a_{\lambda i} b_{\lambda i}^2/\lambda^2.$$

The normal equation representation (11) suggests a simple geometric interpretation of the GML criterion. For a given observation $\mathbf{w} = \mathbf{z}^2$, GML chooses λ by projecting $\mathbf{w}$ onto the

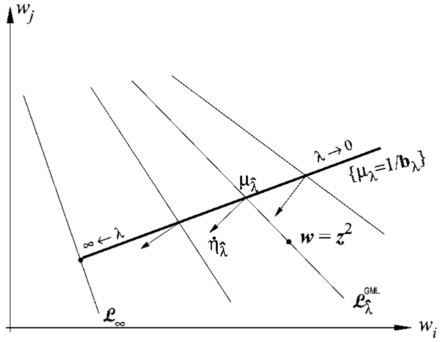

Figure 3. The Geometry of the GML Criterion: $\hat\lambda^{GML}$ is obtained by projecting $\mathbf{w}$ onto $\{\boldsymbol{\mu}_\lambda\}$ orthogonally to the direction $\dot{\boldsymbol{\eta}}_\lambda$. Here, two coordinates, w_i and w_j $(i < j)$, are indicated.

line $\{\boldsymbol{\mu}_\lambda = 1/\mathbf{b}_\lambda : \lambda \geq 0\}$ orthogonally to the direction $\dot{\boldsymbol{\eta}}_\lambda^{\mathrm{GML}}$. We call $\{\boldsymbol{\mu}_\lambda = 1/\mathbf{b}_\lambda : \lambda \geq 0\}$ "the line of expectations," following Efron (2001). Figure 3 diagrams this geometric interpretation two dimensionally.

Let $\mathcal{L}_\lambda^{\mathrm{GML}} = \{\mathbf{w} : (\dot{\boldsymbol{\eta}}_\lambda^{\mathrm{GML}})'(\mathbf{w} - \boldsymbol{\mu}_\lambda) = 0\}$. Solving the normal equation (11) is equivalent to finding the hyperplane $\mathcal{L}_\lambda^{\mathrm{GML}}$ passing through $\mathbf{w}$. It is worth pointing out that because the orthogonal vectors $\dot{\boldsymbol{\eta}}_\lambda^{\mathrm{GML}}$ change direction with λ, the level surfaces $\mathcal{L}_\lambda^{\mathrm{GML}}$ are not parallel to each other (see Fig. 3).

Like GML, C_p also has a simple geometric interpretation. Starting from the C_p formula $\hat\lambda^{C_p} = \arg\min_\lambda \sum_i (b_{\lambda i}^2 w_i - 2b_{\lambda i})$, if we let $\eta_{\lambda i}^{C_p} = -b_{\lambda i}^2$, $\psi_\lambda^{C_p} = -2\sum_i b_{\lambda i}$ then $\hat\lambda^{C_p} = \arg\max_\lambda \{(\boldsymbol{\eta}_\lambda^{C_p})'\mathbf{w} - \psi_\lambda^{C_p}\}$, and the normal equation for $\hat\lambda^{C_p}$ is

$$\frac{\partial}{\partial\lambda}\left\{\left(\boldsymbol{\eta}_\lambda^{C_p}\right)'\mathbf{w} - \psi_\lambda^{C_p}\right\}\Big|_{\lambda=\hat\lambda^{C_p}} = \dot{\boldsymbol{\eta}}_\lambda^{C_p\,\prime}(\mathbf{w} - \boldsymbol{\mu}_\lambda)\Big|_{\lambda=\hat\lambda^{C_p}} = 0, \quad (12)$$

where $\dot\eta_{\lambda i}^{C_p} = \frac{\partial}{\partial\lambda}\eta_{\lambda i}^{C_p} = -2b_{\lambda i}\frac{\partial}{\partial\lambda}b_{\lambda i}$, and $\mu_{\lambda i} = 1/b_{\lambda i}$.

From (12), C_p can be geometrically interpreted as choosing λ by projecting $\mathbf{w}$ onto $\{\boldsymbol{\mu}_\lambda : \lambda \geq 0\}$ orthogonally to the direction $\dot{\boldsymbol{\eta}}_\lambda^{C_p}$, and similarly solving for the C_p estimate $\hat\lambda^{C_p}$ is equivalent to finding the level surface $\mathcal{L}_\lambda^{C_p}$ that contains $\mathbf{w}$, where $\mathcal{L}_\lambda^{C_p} = \{\mathbf{w} : \dot{\boldsymbol{\eta}}_\lambda^{C_p\,\prime}(\mathbf{w} - \boldsymbol{\mu}_\lambda) = 0\}$. Figure 4(a) displays the geometry of C_p together with that of GML. The solid lines represent the orthogonal directions and level surfaces of C_p; the dotted lines, the GML directions and level surfaces. One interesting fact about the geometry is that GML and C_p share the same line of expectations, $\{\boldsymbol{\mu}_\lambda = 1/\mathbf{b}_\lambda : \lambda \geq 0\}$. The difference between GML and C_p is the orientation of the orthogonal direction ($\dot{\boldsymbol{\eta}}_\lambda^{\mathrm{GML}}$ and $\dot{\boldsymbol{\eta}}_\lambda^{C_p}$) or, equivalently, the tilting of the level surface ($\mathcal{L}_\lambda^{\mathrm{GML}}$ and $\mathcal{L}_\lambda^{C_p}$). Figure 4(a) also reveals that as λ varies, the C_p orthogonal direction $\dot{\boldsymbol{\eta}}_\lambda^{C_p}$ rotates faster than its GML counterpart. This is important, and we return to it in Section 5.

2.3 The Ideal Degrees of Freedom

For any smoothing problem of the form (1), there will be an ideal choice of the smoothing parameter λ_0, and a corresponding ideal degrees of freedom, $df_0 = \mathrm{tr}(A_{\lambda_0})$, such that λ_0 minimizes the expected squared error of prediction,

$$\lambda_0 = \arg\min_\lambda E_{\mathbf{f}}\|\hat{\mathbf{f}}_\lambda - \mathbf{f}\|^2. \quad (13)$$

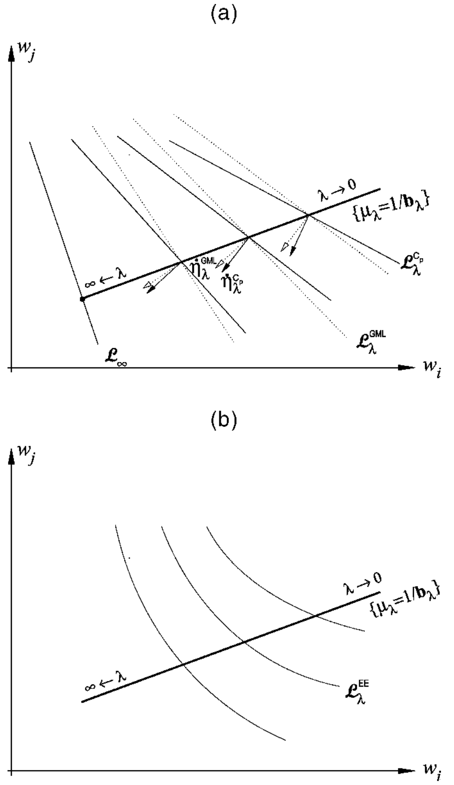
(a)
w_j
w_i
$\lambda \to 0$
$\{\mu_\lambda = 1/\mathbf{b}_\lambda\}$
$\infty \leftarrow \lambda$
$\dot\eta_\lambda^{GML}$
$\dot\eta_\lambda^{C_p}$
$\mathcal{L}_\lambda^{C_p}$
$\mathcal{L}_\lambda^{GML}$
$\mathcal{L}_\infty$
(b)
w_j
w_i
$\lambda \to 0$
$\{\mu_\lambda = 1/\mathbf{b}_\lambda\}$
$\infty \leftarrow \lambda$
$\mathcal{L}_\lambda^{EE}$

Figure 4. The Geometry of Selection Criteria. (a) The geometry of C_p and GML. $\mathcal{L}_\lambda^{GML}$ and $\mathcal{L}_\lambda^{C_p}$ intersect at $\boldsymbol{\mu}_\lambda = 1/\mathbf{b}_\lambda$. The solid lines represent the C_p level surfaces; dotted lines, the GML level surfaces. The thick line is the common line of expectations. (b) The level surfaces of the EE criterion.

Other definitions of "ideal" are possible, such as those of Hall and Johnstone (1992) and Härdel, Hall, and Marron (1988). As pointed out by a referee, Gu (1998) provided an interesting discussion of the concept of degrees of freedom, where it is argued that degrees of freedom is not replicate-invariant. Here we are interested in (13) because it fits in well with the usual notions of estimation accuracy and the following two theorems, especially Theorem 2 connect the estimation of the curve with the estimation of the ideal degrees of freedom. (See also Theorem 7 in Sec. 6.)

Theorem 1. If $\mathbf{y} \sim (\mathbf{f}, \sigma^2)$ and $\hat{\mathbf{f}}_\lambda = A_\lambda \mathbf{y}$, then the ideal choice of λ is

$$\lambda_0 = \arg\min_\lambda \sum_i (b_{\lambda i}^2(g_i^2 + 1) - 2b_{\lambda i}). \quad (14)$$

Theorem 2. For general λ, the total estimation error

$$E_{\mathbf{f}} \parallel \hat{\mathbf{f}}_\lambda - \mathbf{f} \parallel^2$$
$$= E \parallel \hat{\mathbf{f}}_{\lambda_0} - \mathbf{f} \parallel^2 + d_0 \cdot (df_\lambda - df_{\lambda_0})^2 + o(|df_\lambda - df_{\lambda_0}|^2),$$

where the constant

$$d_0 = \sigma^2 \sum_i \left[\left(\left(\frac{\partial}{\partial\lambda}b_{\lambda_0 i}\right)^2 + \left(b_{\lambda_0 i}\frac{\partial^2}{\partial\lambda^2}b_{\lambda_0 i}\right)\right)(g_i^2 + 1) - \frac{\partial^2}{\partial\lambda^2}b_{\lambda_0 i}\right] \Big/ \left(\sum_i \frac{\partial}{\partial\lambda}a_{\lambda_0 i}\right)^2$$

depends only on the true curve and the design points $x_1, x_2, \ldots, x_n$.

The proofs of Theorems 1 and 2 are deferred to Appendix A. Note that $E(\mathbf{w}) = \mathbf{g}^2 + 1$. Comparing (10) and (14), Theorem 1 reveals that if we replace $\mathbf{w}$ in the C_p criterion by its expectation $\mathbf{g}^2 + 1$, then the resulting choice of λ corresponds exactly with the ideal smoothing parameter, suggesting that C_p is a "natural" criterion. We return to this point in Section 4.1. Theorem 2 says that to second order, the error of estimating the curve $\mathbf{f}$ by using a smoothing parameter λ is proportional to the squared difference between the degrees of freedom corresponding to λ and the ideal degrees of freedom. Thus, comparing different criteria by their performance on estimating the ideal degrees of freedom also provides an $\mathbf{f}$-estimation comparison. For this reason, the next three sections concentrate on estimating the ideal degrees of freedom. Section 6 covers estimating the curve, discussing the relationship between estimating the ideal degrees of freedom and adaptively estimating the curve.

3. THE EXTENDED EXPONENTIAL SELECTION CRITERION

3.1 The Criterion

Figures 1 and 2 reveal some general properties of C_p and GML. In the first example, $\mathbf{x}$ comprises 61 equally spaced points on the interval $[-1, 1]$, $f(x) = \sin(\pi(x+1))/(x/2+1), \sigma = 1$; in the second example, $\mathbf{x}$ is 64 equally spaced points on $[0, 1]$, $f(x) = \frac{1}{3}B_{10,5}(x) + \frac{1}{3}B_{7,7}(x) + \frac{1}{3}B_{5,10}(x)$, $\sigma = .05$, where the beta function $B_{p,q}(x) = (\Gamma(p+q)/\Gamma(p)\Gamma(q))x^{p-1}(1-x)^{q-1}$. (This is case 1 in Wahba 1985.) In experiment 1, where the ideal degrees of freedom are 5.18, the C_p- and GML-estimated degrees of freedom with mean 5.64 and 4.84 and median 5.03 and 5.00 are both nearly unbiased. However, as manifested by the longer tails of the C_p estimates seen in Figure 1, the standard deviation of the C_p-estimated degrees of freedom (2.37) is far larger than the standard deviation of the GML estimates (.94). A more robust measure of spread,

$$(\text{interquartile range})/1.35, \tag{15}$$

gives .81 for C_p and .65 for GML. The larger variability of the C_p estimates undermines it as a competitive estimator in this example.

Figure 2 presents a difficult situation for GML. The ideal degrees of freedom in this case are 13.42. Although the C_p estimate's robust standard error (.91) is still much larger than the corresponding measure for GML (.45), its mean of 13.86 and median of 13.31 show that C_p is almost unbiased. In contrast, GML, with a mean of 15.85 and a median of 15.84, is badly biased upward. C_p outperforms GML in this case. As evidenced in these two examples, the problem for C_p is its instability, whereas the trouble for GML is its potentially large bias. The simulation examples, in a certain sense as pointed out by a referee, provide a small sample reflection of the result of Wahba and Wang (1995), who showed that there is a nonzero probability that C_p (GCV) would choose the smoothing parameter to be 0 (namely, $df = n$); the bar at the right end of the C_p histogram serves as a reminder. At this point, one naturally hopes to find a "robust" selection criterion—one that behaves stably and will not give eccentric estimates as C_p occasionally does, yet at the same time does not suffer from enormous bias.

Our proposed new selection criterion, the EE criterion, satisfies these qualitative requirements to a large extent. Expressed in the coordinate system (6), (9), the EE criterion chooses the smoothing parameter λ according to

$$\hat{\lambda}^{EE} = \arg\min_{\lambda} \sum_i \left[C b_{\lambda i} w_i^{2/3} - 3 b_{\lambda i}^{1/3} \right], \tag{16}$$

where $C = \sqrt{\pi}/[2^{2/3}\Gamma(7/6)] = 1.203$ (see Sec. 4.1 for an explanation of the choice of C) and $w_i = z_i^2$.

Applying the EE formula (16) to the two examples produces the histograms showed in the left panels of Figures 1(b) and 2(b). In the first example, the EE estimates have mean 5.16, median 5.16, and robust deviation .73; in the second example, the mean 14.52, median 14.42, and robust deviation .79. Table 1 summarizes the statistics describing the performance of the EE criterion together with those of C_p and GML.

Comparing the statistics given in Table 1, two properties of the EE criterion are noteworthy: (a) The variability of the EE estimates is significantly smaller than that of C_p, and (b) in the unfavorable situation of experiment 2, although the EE criterion has some bias, the bias is much smaller than that of GML. In short, the EE criterion performs as a reasonably "robust" compromise between C_p and GML.

3.2 The Geometric Motivation

The EE selection criterion is motivated by the geometry of Figures 3 and 4, where C_p and GML share the same line of expectations. If an observation happens to lie on this line, say $\mathbf{w} = \boldsymbol{\mu}_{\lambda^\dagger}$ for some $\lambda^\dagger$, then both GML and C_p would estimate $\lambda^\dagger$ to be the smoothing parameter. In other words, the line of expectations works as a common ground for GML and C_p. What about points away from the line of expectations? Efron (2000) showed that in the case of $E\{\mathbf{w}\} = 1 + \mathbf{g}^2$ lying *below* the line of expectations, C_p tends to work most efficiently, whereas when $E\{\mathbf{w}\}$ is *above* the line of expectations, GML outperforms C_p.

Table 1. Summary Statistics Comparing the EE Estimated Degrees of Freedom With Those of GML and C_p on the Two Simulation Experiments

	Experiment 1 (ideal df = 5.18)			*Experiment 2 (ideal df = 13.42)*		
	C_p	*EE*	GML	C_p	*EE*	GML
Mean	5.64	5.16	4.84	13.86	14.52	15.85
Median	5.03	5.16	5.00	13.31	14.42	15.84
Bias	≈.2	≈.02	≈.2	≈.1	≈ 1.0	≈ 2.4
Standard error	2.37	1.09	.94	2.10	.86	.46
Mean squared error (MSE)	5.78	1.20	1.00	4.62	1.94	6.12
Robust spread	.81	.73	.65	.91	.79	.45

NOTE: The robust measure of spread is the interquartile range divided by 1.35.

The goal of combining the strengths of C_p and GML motivates the EE criterion. Figure 4(b) shows the EE level surface $\mathcal{L}_\lambda^{EE}$, which can be derived from the normal equation of the EE formula (16): $\mathcal{L}_\lambda^{EE} = \{\mathbf{w} : \dot{\boldsymbol{\eta}}_\lambda^{EE\,\prime}(\mathbf{w}^{2/3} - \boldsymbol{\mu}_\lambda^{EE}) = 0\}$, where $\dot{\eta}_{\lambda i}^{EE} = -C^{3/2}\frac{\partial}{\partial\lambda}b_{\lambda i} = -C^{3/2}a_{\lambda i}b_{\lambda i}/\lambda$ and $\mu_{\lambda i}^{EE} = 1/(Cb_{\lambda i}^{2/3})$. Comparing Figure 4(a) and (b), we notice that above the line $\{\boldsymbol{\mu}_\lambda = 1/\mathbf{b}_\lambda : \lambda \geq 0\}$, $\mathcal{L}_\lambda^{EE}$ behaves much like the GML level surface $\mathcal{L}_\lambda^{\mathrm{GML}}$, and below the line, $\mathcal{L}_\lambda^{EE}$ follows $\mathcal{L}_\lambda^{C_p}$. Geometrically, the EE criterion is seen to smoothly combine C_p and GML. The simulation results of Figures 1 and 2 confirm this geometric intuition. Theoretical results are given in Sections 4 and 5. Section 7 discusses the Bayesian properties of the selection criteria, and explains the term "extended exponential."

3.3 A Unified Framework

Geometry motivates the definition of the EE criterion, but geometry alone does not completely explain formula (16). Indeed, there is a class of selection criteria that each have a simple geometric picture. Let $p \geq 1$, and $q \geq 1$ be two fixed constants. Define the function mathtight

$$l_\lambda^{(p,q)}(\mathbf{u}) = \begin{cases} \sum_i \left[(c_q B_{\lambda i})^p u_i - \frac{p}{p-1}((c_q B_{\lambda i})^{p-1} - 1)\right] & \text{if } p > 1 \\ \sum_i (c_q B_{\lambda i} u_i - \log B_{\lambda i}) & \text{if } p = 1, \end{cases} \tag{17}$$

where

$$B_{\lambda i} = b_{\lambda i}^{1/q} \quad \text{and} \quad c_q = \sqrt{\pi}/[2^{1/q}\Gamma(1/2 + 1/q)] \tag{18}$$

(see Sec. 5.1 for the choice of c_q). For each pair (p, q), a selection criterion can be defined as

$$\hat{\lambda}^{(p,q)} = \arg\min_\lambda \{l_\lambda^{(p,q)}(\mathbf{w}^{1/q})\} = \begin{cases} \arg\min_\lambda \sum_i [c_q B_{\lambda i}^p w_i^{1/q} - \frac{p}{p-1} B_{\lambda i}^{p-1}] & \text{if } p > 1 \\ \arg\min_\lambda \sum_i (c_q B_{\lambda i} w_i^{1/q} - \log B_{\lambda i}) & \text{if } p = 1, \end{cases} \tag{19}$$

where $\mathbf{w} = \mathbf{z}^2 = (z_1^2, z_2^2, \ldots, z_n^2)'$ as before. This class of selection criteria, indexed by the pair (p, q), contains the EE, C_p, and GML criteria, as direct calculation replacing (p, q) by $(1, 1)$, $(2, 1)$, and $(\frac{3}{2}, \frac{3}{2})$ gives the following theorem.

Theorem 3. For $p = 1$, and $q = 1$, $\hat{\lambda}^{(1,1)}$ is the same as the GML estimate $\hat{\lambda}^{\mathrm{GML}}$; for $p = 2$, $q = 1$, $\hat{\lambda}^{(2,1)}$ is the C_p estimate $\hat{\lambda}^{C_p}$; $p = q = \frac{3}{2}$ gives the EE selection criterion.

The theorem helps explain the EE formula (16). Because $p = 1$ for GML and $p = 2$ for C_p, to compromise between the two, it is reasonable to take $p = 3/2$. Kou (2001) studied the Bayesian properties of the selection criteria family (19) and showed that indeed, taking $p = 3/2$ is most Bayesian robust. The EE choice of $q = 3/2$ is based on two facts: (a) $q > 1$ is required so that we can have a "curved" level surface to combine the geometry of C_p and GML; (b) Kou (2001) also considered the large-sample properties of the selection criteria and showed that $p \geq q$ is needed for desirable asymptotic properties. In addition to theoretical justifications of Sections 4 and 5, simulations like those of Figures 1 and 2 support choosing EE from class (19).

Like GML and C_p, every member of (19) enjoys a simple geometric picture. The normal equation for $\hat{\lambda}^{(p,q)}$ has the form

$$\dot{l}_{\hat{\lambda}^{(p,q)}}^{(p,q)}(\mathbf{w}^{1/q}) = \frac{\partial}{\partial\lambda} l_{\hat{\lambda}^{(p,q)}}^{(p,q)}(\mathbf{w}^{1/q}) = -\dot{\boldsymbol{\eta}}_\lambda^{(p,q)\prime}\left(\mathbf{w}^{1/q} - \boldsymbol{\mu}_\lambda^{(p,q)}\right)\Big|_{\lambda=\hat{\lambda}^{(p,q)}} = 0, \tag{20}$$

where

$$\dot{\boldsymbol{\eta}}_\lambda^{(p,q)} = \left(\dot{\eta}_{\lambda 1}^{(p,q)}, \dot{\eta}_{\lambda 2}^{(p,q)}, \ldots, \dot{\eta}_{\lambda n}^{(p,q)}\right) \quad \text{with } \dot{\eta}_{\lambda i}^{(p,q)} = -\frac{p}{q\lambda} a_{\lambda i}(c_q B_{\lambda i})^p, \tag{21}$$

and

$$\boldsymbol{\mu}_\lambda^{(p,q)} = \left(\mu_{\lambda 1}^{(p,q)}, \mu_{\lambda 2}^{(p,q)}, \ldots, \mu_{\lambda n}^{(p,q)}\right) \quad \text{with } \mu_{\lambda i}^{(p,q)} = 1/(c_q B_{\lambda i}). \tag{22}$$

Replacing the orthogonal direction $\dot{\boldsymbol{\eta}}_\lambda$ by $\dot{\boldsymbol{\eta}}_\lambda^{(p,q)}$, the line of expectations $\{\boldsymbol{\mu}_\lambda : \lambda > 0\}$ by $\{\boldsymbol{\mu}_\lambda^{(p,q)} = 1/(c_q \mathbf{B}_\lambda) : \lambda > 0\}$, $\mathbf{w}$ by $\mathbf{u} = \mathbf{w}^{1/q}$, and the level surface $\mathcal{L}_\lambda^{\mathrm{GML}}$ by

$$\mathcal{L}_\lambda^{(p,q)} = \left\{\mathbf{u} = \mathbf{w}^{1/q} : \sum_i a_{\lambda i} B_{\lambda i}^p (u_i - 1/(c_q B_{\lambda i})) = 0\right\},$$

Figures 3 and 4 also apply to the geometry of a general estimator $\hat{\lambda}^{(p,q)}$.

We make some remarks about the class (19) and their geometry.

Remark 2. From definition (17), it is seen that $l_\lambda^{(p,q)} \to l_\lambda^{(1,q)}$ as $p \to 1$. Efron (2001) constructed a family $\mathcal{F}^{(p)}$ of selection criteria indexed by a parameter $p \geq 1$, corresponding to taking $q = 1$ in (19). The (p, q) family $\hat{\lambda}^{(p,q)}$ can also be thought of as a two-parameter generalization of that family.

Remark 3. As $\lambda \downarrow 0$, $\boldsymbol{\mu}_\lambda^{(p,q)}$ moves upward along the line of expectations, the orthogonal direction $\dot{\boldsymbol{\eta}}_\lambda^{(p,q)}$ also changes. For spline-like smoothers, $\dot{\boldsymbol{\eta}}_\lambda^{(p,q)}$ rotates counterclockwise (because $\dot{\eta}_{\lambda i}^{(p,q)}/\dot{\eta}_{\lambda j}^{(p,q)}$ is an increasing function of λ for $i < j$). Furthermore, the speed of rotation of the EE and GML criteria is considerably slower than that of C_p. As explained in Section 5, this slower rotation causes the EE and GML criteria to be substantially more stable than C_p.

4. VARIANCE, BIAS, AND BIAS CORRECTION

The simple geometry of the criteria diagrammed in Figures 3 and 4 helps derive useful theoretical results. We first exploit it to calculate the variance and bias of the (p, q)-estimated degrees of freedom.

4.1 Variance and Bias

In this section we suppress the superscripts for notational ease; for example, we write $\dot{\boldsymbol{\eta}}_\lambda$ instead of $\dot{\boldsymbol{\eta}}_\lambda^{(p,q)}$, and $\dot{l}_\lambda(\mathbf{u})$ instead of $\dot{l}_\lambda^{(p,q)}(\mathbf{u})$, and so on. We denote by $\widehat{df}$ the degrees of freedom associated with $\hat{\lambda}$, $\widehat{df} = df(\hat{\lambda}) = \sum_{i=1}^n a_{\hat{\lambda}i}$ and

$$\mathbf{u} = (u_1, u_2, \ldots, u_n) = (w_1^{1/q}, w_2^{1/q}, \ldots, w_n^{1/q}). \tag{23}$$

Theorem 4. Let df_1 be the degrees of freedom obtained by substituting $\mathbf{u}_0 \equiv E\{\mathbf{u}\}$ into the (p,q)-estimation formula

$$df_1 = df(\lambda_1),\ \lambda_1 = \arg\min_\lambda l_\lambda(\mathbf{u}_0) = \arg\min_\lambda l_\lambda(E\{\mathbf{w}^{1/q}\}). \tag{24}$$

Then a delta method approximation for the standard error of the (p,q) estimator $\widehat{df}^{(p,q)}$ is

$$\begin{aligned} se_{p,q}(\widehat{df}) &\doteq \left[\sum_i \left(\frac{\partial\widehat{df}}{\partial u_i}\bigg|_{\mathbf{u}_0}\right)^2 \operatorname{var} u_i\right]^{1/2} \\ &= \left|\left(\sum_{i=1}^n \dot{a}_{\lambda_1 i}\right)\bigg/ \ddot{l}_{\lambda_1}(\mathbf{u}_0)\right| \left(\sum_i \dot{\eta}^2_{\lambda_1 i} \operatorname{var} u_i\right)^{1/2}, \end{aligned} \tag{25}$$

where a single dot on a variable denotes its first derivative with respect to λ and double dots indicate the second derivatives, that is, $\dot{a}_{\lambda i} = (\partial/\partial\lambda)a_\lambda$ and $\ddot{l}_\lambda(\mathbf{u}) = (\partial^2/\partial\lambda^2)l_\lambda(\mathbf{u}_0)$.

Derivation of Theorem 4. Starting from the normal equation $\dot{l}_{\hat\lambda}(\mathbf{u}) = 0$, standard application of the implicit function theorem gives the delta influence of u_i on $\hat\lambda$, $\partial\hat\lambda/\partial u_i = -(\ddot{l}_{\hat\lambda}(\mathbf{u}))^{-1}(\partial/\partial u_i)\dot{l}_{\lambda}(\mathbf{u})|_{\lambda=\hat\lambda} = \dot{\eta}_{\hat\lambda i}/\ddot{l}_{\hat\lambda}(\mathbf{u})$, from which the delta influence of u_i on $\widehat{df}$ is given by

$$\frac{\partial\widehat{df}}{\partial u_i} = \frac{\partial\widehat{df}}{\partial\hat\lambda}\frac{\partial\hat\lambda}{\partial u_i} = \left(\sum_{i=1}^n \dot{a}_{\hat\lambda i}\right)\dot{\eta}_{\hat\lambda i}/\ddot{l}_{\hat\lambda}(\mathbf{u}). \tag{26}$$

A first-order Taylor expansion on $\widehat{df}$ around $\mathbf{u}_0$ yields

$$\widehat{df} \doteq df_1 + \sum_i \frac{\partial\widehat{df}}{\partial u_i}\bigg|_{\mathbf{u}_0}(u_i - u_{0i}). \tag{27}$$

The desired result follows by applying the delta method calculations on (26) and (27).

The approximation (25) has a simple computational form if the underlying smoothers are spline-like, under which

$$\dot{\eta}^{(p,q)}_{\lambda i} = -\frac{p}{q\lambda}a_{\lambda i}(c_q B_{\lambda i})^p, \tag{28}$$

and

$$\begin{aligned} \ddot{\eta}^{(p,q)}_{\lambda i} &= \frac{p}{q\lambda^2}a_{\lambda i}(c_q B_{\lambda i})^p[2-(1+p/q)a_{\lambda i}], \\ \ddot{l}_\lambda(\mathbf{u}) &= pc_q^{p-1}/(q\lambda^2)\cdot Q_\lambda(\mathbf{u}), \\ \frac{\partial\widehat{df}}{\partial u_i} &= \frac{\partial\widehat{df}}{\partial u_i} = c_q\left(a_{\hat\lambda i}B^p_{\hat\lambda i}\right)\left(\sum_i a_{\hat\lambda i}b_{\hat\lambda i}\right)\bigg/ Q_{\hat\lambda}(\mathbf{u}), \end{aligned} \tag{29}$$

where $Q_\lambda(\mathbf{u}) = \sum_i a_{\lambda i}B^{p-1}_{\lambda i}\{a_{\lambda i}/q + [(1+p/q)a_{\lambda i} - 2]\times(c_q B_{\lambda i}u_i - 1)\}$. Thus, for spline-like smoothers, the delta method approximation simplifies to

$$se_{p,q}(\widehat{df}) \doteq c_q|Q_{\lambda_1}(\mathbf{u}_0)|^{-1}\left(\sum_i a_{\lambda_1 i}b_{\lambda_1 i}\right)\left[\sum_i a^2_{\lambda_1 i}B^{2p}_{\lambda_1 i}\operatorname{var} u_i\right]^{1/2}. \tag{30}$$

Table 2 reports the results of applying approximation (30) to the two cases of experiment 1 and 2. The delta method approximation (30) agrees well with the robust standard error (15) from the simulation results, lending theoretical support to the stable performance of the EE and GML estimates in Figures 1 and 2.

Table 2. Comparison of δ Method Approximation With Simulation Results From the Two Experiments

	Experiment 1		*Experiment 2*	
	Formula (30)	*Empirical se (15)*	*Formula (30)*	*Empirical se (15)*
EE ($p=\frac{3}{2}, q=\frac{3}{2}$)	**.725**	.731	**.787**	.794
C_p ($p=2, q=1$)	**.769**	.812	**.894**	.913
GML ($p=1, q=1$)	**.639**	.653	**.455**	.451

The derivation of Theorem 4, especially the Taylor expansion (27), emphasizes the importance of $\mathbf{u}_0 = E\{\mathbf{w}^{1/q}\}$ and the theoretical degrees of freedom $df_1^{(p,q)}$ (24). Here $df_1^{(p,q)}$ works as the central value of $\widehat{df}^{(p,q)}$. By central value, we mean that the mean and median of $\widehat{df}^{(p,q)}$ are closely located around $df_1^{(p,q)}$, as shown in Table 3, which compares $df_1^{(p,q)}$ with the empirical mean and median of $\widehat{df}^{(p,q)}$. For a (p,q) estimator, we can get an idea of bias by looking at the difference $df_1^{(p,q)} - df_0$, where df_0 is the ideal degrees of freedom. For C_p, $p=2$, $q=1$, $\mathbf{u}_0 = E\{\mathbf{w}\} = \mathbf{g}^2+1$; so $\lambda_1^{C_p} = \arg\min_\lambda \sum_i\{b^2_{\lambda i}(g_i^2+1) - 2b_{\lambda i}\}$ according to (24). But Theorem 1 says that λ_0, the ideal smoothing parameter defined in (13), is also $\lambda_0 = \arg\min_\lambda \sum_i\{b^2_{\lambda i}(g_i^2+1) - 2b_{\lambda i}\}$. Therefore, for C_p, its central value $df_1^{C_p}$ gives the ideal degrees of freedom df_0, echoing our earlier claim for the naturalness of C_p. For general (p,q) estimates, because the central value $\lambda_1^{(p,q)} \neq \lambda_0$, $\widehat{df}^{(p,q)}$ is potentially biased.

Before calculating the bias of $\widehat{df}^{(p,q)}$, we detour to an important technical point. Because $\mathbf{z} \sim N(\mathbf{g}, I)$, it follows that $w_i = z_i^2$ are independent, each having a noncentral chi-squared distribution with 1 degree of freedom and noncentrality parameter g_i^2: $w_i \overset{\text{ind.}}{\sim} \chi^2_1(g_i^2)$. Thus $u_{0i} = E\{u_i\} = E\{w_i^{1/q}\}$ involves the fractional moment of a noncentral χ^2_1 random variable.

Lemma 1. For $w_i \sim \chi^2_1(g_i^2)$, $r \geq 0$, $Ew_i^r = \frac{1}{\sqrt{\pi}}2^r\Gamma(r+1/2)M(-r, 1/2, -g_i^2/2)$, where $M(\cdot,\cdot,\cdot)$ is the confluent hypergeometric function given by $M(c,d,z) = 1 + cz/d + (c)_2z^2/(d)_2 2! + \cdots + (c)_n z^n/(d)_n n! + \cdots$, with $(d)_n$ defined by $(d)_n = d(d+1)\cdots(d+n-1)$. In particular,

$$u_{0i} = Ew_i^{1/q} = \frac{1}{\sqrt{\pi}}2^{1/q}\Gamma(1/q+1/2)M(-1/q, 1/2, -g_i^2/2). \tag{31}$$

Table 3. Comparison of $df_1^{(p,q)}$ With Empirical Mean and Median of $\widehat{df}^{(p,q)}$

		df_1	*Empirical mean*	*Empirical median*
Experiment 1	EE ($p=\frac{3}{2}, q=\frac{3}{2}$)	**5.26**	5.16	5.16
	C_p ($p=2, q=1$)	**5.18**	5.64	5.03
	GML ($p=1, q=1$)	**5.12**	4.84	5.00
Experiment 2	EE ($p=\frac{3}{2}, q=\frac{3}{2}$)	**14.45**	14.52	14.42
	C_p ($p=2, q=1$)	**13.42**	13.86	13.31
	GML ($p=1, q=1$)	**15.85**	15.85	15.84

Proof. See Appendix A.

Lemma 1 not only provides explicit formula for computation, but also helps explain the choice of C in (16) and c_q in (17) and (19). Starting from the normal sampling model $\mathbf{y} \sim N(\mathbf{f}, \sigma^2 I)$, if the true curve $\mathbf{f}$ is linear and a cubic smoothing spline is used to recover $\mathbf{f}$, then the orthogonal transformation (6) would give $g_3 = g_4 = \ldots = g_n = 0$, which, according to (31), forces the central value $\mathbf{u}_0$ to be

$$\begin{aligned} u_{0i} &= \frac{1}{\sqrt{\pi}} 2^{1/q} \Gamma(1/q + 1/2) M(-1/q, 1/2, 0) \\ &= \frac{1}{\sqrt{\pi}} 2^{1/q} \Gamma(1/q + 1/2), \qquad i \geq 3. \end{aligned} \tag{32}$$

Because linear models, although simple, are such an important case, it is natural to require that any reasonable selection criterion preserve linearity when the underlying curve indeed is a straight line. In the case of (p, q) estimators, careful examination of the geometry displayed in Figure 3 indicates that the point corresponding most naturally to 2 degrees of freedom is the left endpoint of the line of expectations. This suggests matching the central value $\mathbf{u}_0 = E\{\mathbf{w}^{1/q}\}$ with the leftend point so that the central degrees of freedom $df_1^{(p,q)} = 2$. By (22), the ith ($i \geq 3$) coordinate of the left endpoint is $1/c_q$, which together with (32) implies $c_q = \sqrt{\pi}/[2^{1/q}\Gamma(1/2 + 1/q)]$.

The geometric picture also leads to a simple way of calculating median bias. Define

$$S = -\dot{\boldsymbol{\eta}}_{\lambda_0}^{(p,q)\prime}\left(\mathbf{u} - \boldsymbol{\mu}_{\lambda_0}^{(p,q)}\right),$$

where, as before, λ_0 is the ideal choice of smoothing parameter. Then, locally (see Fig. 3), $\widehat{df}^{(p,q)} < df_0$ if $\mathbf{u}$ lies to the left of $\mathcal{L}_{\lambda_0}$, the hyperplane associated with the ideal degrees of freedom df_0, and $\widehat{df}^{(p,q)} > df_0$ if $\mathbf{u}$ lies to the right of $\mathcal{L}_{\lambda_0}$. In other words, $\widehat{df}^{(p,q)} - df_0$ locally has the same sign as S, which suggests the approximation

$$P\left(\widehat{df}^{(p,q)} < df_0\right) \doteq P(S < 0). \tag{33}$$

Because $u_i = w_i^{1/q}$ are independently distributed, S has mean $M(S) = -\dot{\boldsymbol{\eta}}_{\lambda_0}^{(p,q)\prime}(\mathbf{u}_0 - \boldsymbol{\mu}_{\lambda_0}^{(p,q)})$, variance $V(S) = \sum_i \dot{\eta}_{\lambda_0 i}^{(p,q)\,2} \operatorname{var} u_i$, and skewness

$$\operatorname{skew}(S) = \frac{E(S - M(S))^3}{V(S)^{3/2}} = -\frac{\sum_i \dot{\eta}_{\lambda_0 i}^{(p,q)\,3} E[(u_i - u_{0i})^3]}{(\sum_i \dot{\eta}_{\lambda_0 i}^{(p,q)\,2} \operatorname{var} u_i)^{3/2}}.$$

An Edgeworth expansion involving the first three moments provides

$$\begin{aligned} P(S < 0) &\doteq \Phi\left(-\frac{M(S)}{\sqrt{V(S)}}\right) \\ &\quad - \frac{1}{6}\operatorname{skew}(S)\left(\frac{M(S)^2}{V(S)} - 1\right)\varphi\left(-\frac{M(S)}{\sqrt{V(S)}}\right), \end{aligned} \tag{34}$$

where Φ and φ are the cdf and density of the standard normal distribution. This also gives the following result.

Theorem 5. The median bias of $\widehat{df}^{(p,q)}$ has Edgeworth approximation

$$\begin{aligned} P\left(\widehat{df}^{(p,q)} < df_0\right) &\doteq \Phi\left(-\frac{M(S)}{\sqrt{V(S)}}\right) \\ &\quad - \frac{1}{6}\operatorname{skew}(S)\left(\frac{M(S)^2}{V(S)} - 1\right)\varphi\left(-\frac{M(S)}{\sqrt{V(S)}}\right). \end{aligned} \tag{35}$$

Formulas (21) and (22) give simple expressions for spline-like smoothers,

$$\begin{aligned} M(S) &= \tfrac{pc_q^p}{q\lambda_0} \textstyle\sum_i a_{\lambda_0 i} B_{\lambda_0 i}^p (u_{0i} - 1/(c_q B_{\lambda_0 i})) \\ V(S) &= \left(\tfrac{pc_q^p}{q\lambda_0}\right)^2 \textstyle\sum_i a_{\lambda_0 i}^2 B_{\lambda_0 i}^{2p} \operatorname{var} u_i \end{aligned}$$

and

$$\operatorname{skew}(S) = \textstyle\sum_i a_{\lambda_0 i}^3 B_{\lambda_0 i}^{3p} E[(u_i - u_{0i})^3] \left(\sum_i a_{\lambda_0 i}^2 B_{\lambda_0 i}^{2p} \operatorname{var} u_i\right)^{-3/2}.$$

Table 4 compares approximation (35) with the empirically observed $P(\widehat{df}^{(p,q)} < df_0)$ for the two simulation experiments. The accuracy of (35) is clearly demonstrated. The theoretical calculation once again reveals that C_p is almost unbiased and that although EE is possibly biased, its potential bias is significantly smaller than the bias of GML.

4.2 Resubstitution and Bias Correction

Theorems 4 and 5 are of more than theoretical value. In real applications where the true curve $\mathbf{g}$ (or, equivalently, $\mathbf{f}$) is unknown, the following resubstitution idea can be used to get estimates of standard error and median bias.

Recall that we produce (30) and (35) via

$$\begin{tikzcd}
 & & & \lambda_1 \arrow[r] & \text{standard error formula} \\
 & \mathbf{u}_0 = E\{\mathbf{w}^{1/q}\} \arrow[urr] \arrow[drr] & & & \\
\mathbf{g} \arrow[ur] \arrow[dr] & & & \text{median bias formula} & \\
 & \mathbf{g}^2 + 1 \arrow[r] & \lambda_0 \arrow[ur] & &
\end{tikzcd}.$$

The resubstitution method replaces the unknown true curve $\mathbf{g}$ with $\hat{\mathbf{g}} = \mathbf{a}_{\hat{\lambda}}\mathbf{z}$, the estimated curve,

$$\begin{tikzcd}
 & & & & \hat{\lambda}_1 \arrow[r] & \text{standard error} \\
 & & & \hat{\mathbf{u}}_0 \arrow[ur] \arrow[drr] & & \\
\mathbf{z} \arrow[r] & \hat{\lambda} \arrow[r] & \hat{\mathbf{g}} = \mathbf{a}_{\hat{\lambda}}\mathbf{z} \arrow[ur] \arrow[dr] & & & \text{estimated median bias} \\
 & & & \hat{\mathbf{g}}^2 + 1 \arrow[r] & \hat{\lambda}_0 \arrow[ur] &
\end{tikzcd}. \tag{36}$$

Equipped with the standard error, we can say more about the estimate. For instance, a naive 90% confidence interval for the degrees of freedom is $[\widehat{df} - 1.65\widehat{se}, \widehat{df} + 1.65\widehat{se}]$. For the particular scatterplot displayed in Figure 2(a), the EE criterion via (36) produces a naive 90% interval [12.65, 15.53], neatly containing 13.42, the ideal degrees of freedom.

The resubstitution idea of (36) also suggests a simple method of bias correction. Suppose that estimate $\widehat{df}$ behaves

Table 4. Comparison of Approximation (35) With Monte Carlo Estimates of $P(\widehat{df}^{(p,q)} < df_0)$

	Experiment 1		Experiment 2	
	Approximation (35)	Observed $P(\widehat{df}^{(p,q)} < df_0)$	Approximation (35)	Observed $P(\widehat{df}^{(p,q)} < df_0)$
EE ($p = \frac{3}{2}, q = \frac{3}{2}$)	**.483**	.507	**.068**	.071
C_p ($p = 2, q = 1$)	**.573**	.583	**.544**	.540
GML ($p = 1, q = 1$)	**.584**	.618	**< 0**	0

approximately normally around its center of distribution. Then a bias correction is given by

$$\widetilde{df} = \widehat{df} + \Phi^{-1}\left(P(\widehat{df} < df_0)\right)\widehat{se}, \qquad (37)$$

with $P(\widehat{df} < df_0)$ approximated by (35) through resubstitution (36). Figure 5 shows the effect of bias correction on the EE and C_p estimates $\widehat{df}^{EE}$ and $\widehat{df}^{C_p}$ for the second experiment (Fig. 2). The unshaded bars present the histogram of $\widehat{df}$; the shaded bars, the histogram of the bias-corrected estimate $\widetilde{df}$. Formula (37) is seen to work well for the EE estimates but essentially has no effect on the C_p estimates, the reason being that for C_p, the problem is its large variability, not bias. Formula (37) has also been tried on the GML estimates, however without satisfactory result. The problem is that the bias of the GML estimates is so big that Edgeworth approximation (35) starts breaking down. More than 90% of the GML estimates in experiment 2 assign negative values to the Edgeworth approximation of $P(\widehat{df} < df_0)$, which in turn invalidates (37).

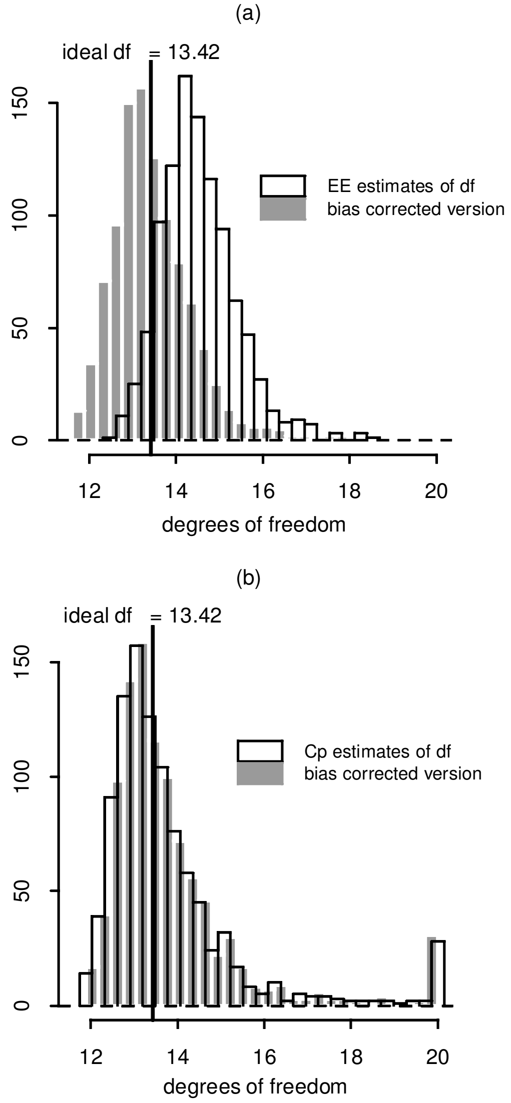

Figure 5. Bias Correction (37) on EE and C_p Estimates. (a) Histograms of EE and the corresponding bias corrected estimates. (b) Histograms of C_p estimates and its bias-corrected version.

5. REVERSAL EFFECT AND VARIABILITY

Tables 1 and 2, through simulations and theoretical approximation, exhibit the high variability of the C_p estimates compared with the EE or GML estimates, as seen in Figures 1 and 2, where the troublesome long right tails of the C_p estimates are particularly worrisome. In this section we show that this is due to the *reversal effect* (Efron 2001) caused by the rotation of the direction $\dot{\boldsymbol{\eta}}_\lambda^{(p,q)}$ (21). In addition, we also show that the reversal effect of the EE estimator is much smaller than that of C_p, which underlines the stable behavior of the EE estimates.

Going back to the geometry of Figure 3, points on $\mathcal{L}_\lambda$, the level surface passing through $\boldsymbol{\mu}_\lambda^{(p,q)}$ orthogonally to $\dot{\boldsymbol{\eta}}_\lambda^{(p,q)}$, satisfy the normal equation $\dot{l}_\lambda^{(p,q)}(\mathbf{u}) = -\dot{\boldsymbol{\eta}}_\lambda^{(p,q)\prime}(\mathbf{u} - \boldsymbol{\mu}_\lambda^{(p,q)}) = 0$. Because the directions $\dot{\boldsymbol{\eta}}_\lambda^{(p,q)}$ are not parallel to each other, different hyperplanes intersect each other, with points on the intersection of two hyperplanes satisfying both normal equations. One direct consequence of nonparallelness is numerical instability, because the solution of a given normal equation might not be unique. But this is not the only trouble, however.

Figure 6 illustrates this "crossover" phenomenon, showing $\mathcal{L}_{\lambda_0}$, the flat space corresponding to the ideal degrees of freedom df_0, intersecting $\mathcal{L}_{\lambda_0+d\lambda}$ for some small $d\lambda$. From $\dot{l}_{\lambda_0}^{(p,q)}(\mathbf{u}) = 0$ and $\dot{l}_{\lambda_0+d\lambda}^{(p,q)}(\mathbf{u}) = 0$, we see that as $d\lambda \to 0$, the limiting intersection point $\mathbf{u}^\dagger$ satisfies

$$\dot{l}_{\lambda_0}^{(p,q)}(\mathbf{u}^\dagger) = 0 \quad \text{and} \quad \ddot{l}_{\lambda_0}^{(p,q)}(\mathbf{u}^\dagger) = 0. \qquad (38)$$

Now imagine moving a point $\mathbf{u}$ along $\mathcal{L}_{\lambda_0}$ away from $\boldsymbol{\mu}_{\lambda_0}^{(p,q)}$ toward $\mathbf{u}^\dagger$. Because the second derivative $\ddot{l}_{\lambda_0}^{(p,q)}(\mathbf{u})$ is linear in $\mathbf{u}$ according to (29), it changes sign from positive to 0 at the critical point $\mathbf{u}^\dagger$, and then negative beyond it. This means that as $\mathbf{u}$ moves upward, λ_0 switches from a local minimum of $l_\lambda^{(p,q)}$ to a local *maximum*. Therefore, for points lying on $\mathcal{L}_{\lambda_0}$ beyond $\mathbf{u}^\dagger$, the (p, q) estimate $\hat{\lambda}^{(p,q)}$, which by definition is the global minimizer of $l_\lambda^{(p,q)}$, must be far away from λ_0!

This phenomenon was called the *reversal effect* by Efron (2001). Figure 6 also shows that the faster the rotation of

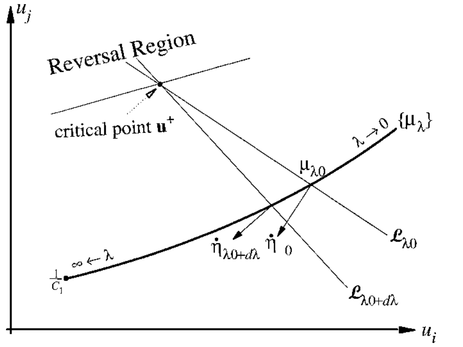
u_j
Reversal Region
critical point $\mathbf{u}^+$
$\lambda \to 0$ $\{\mu_\lambda\}$
$\mu_{\lambda 0}$
$\mathcal{L}_{\lambda 0}$
$\dot{\eta}_{\lambda 0+d\lambda}$
$\dot{\eta}\ _0$
$\infty \leftarrow \lambda$
$\frac{1}{C_1}$
$\mathcal{L}_{\lambda 0+d\lambda}$
u_i

Figure 6. Illustration of the Reversal Effect. The rotation of $\dot{\boldsymbol{\eta}}_\lambda$ causes different hyperplanes $\mathcal{L}_\lambda$ to intersect. The limiting intersecting point defines the beginning of the reversal region. The faster the rotation of $\dot{\boldsymbol{\eta}}_\lambda$, the closer the reversal region to the line of expectations and the more severe the reversal effect.

$\dot{\boldsymbol{\eta}}_\lambda^{(p,q)}$, the closer the critical point $\mathbf{u}^\dagger$ to the line of expectations, and hence the greater the chance that a point falls beyond $\mathbf{u}^\dagger$, that is, the more severe the reversal effect. The EE and GML directions $\dot{\boldsymbol{\eta}}_\lambda^{(\frac{3}{2},\frac{3}{2})}$ and $\dot{\boldsymbol{\eta}}_\lambda^{(1,1)}$ rotate more slowly than the C_p direction $\dot{\boldsymbol{\eta}}_\lambda^{(2,1)}$, implying that EE and GML suffer less than C_p from the reversal effect. This is the main factor behind the stable behavior of the EE criterion. Table 5 measures the speed of rotation in terms of curvature.

The following theorem, modified slightly from theorem 3 of Efron (2001), formally connects the distance from $\mathbf{u}^\dagger$ to $\boldsymbol{\mu}_{\lambda_0}^{(p,q)}$ with the speed of rotation of $\dot{\boldsymbol{\eta}}_\lambda^{(p,q)}$.

Theorem 6. The minimum distance between the critical point $\mathbf{u}^\dagger$ (38), and $\boldsymbol{\mu}_{\lambda_0}^{(p,q)}$ is given by

$$\min_{\mathbf{u}^\dagger}\Big[\big(\mathbf{u}^\dagger - \boldsymbol{\mu}_{\lambda_0}^{(p,q)}\big)'\Sigma^{-1}\big(\mathbf{u}^\dagger - \boldsymbol{\mu}_{\lambda_0}^{(p,q)}\big)\Big]^{1/2}$$

$$= \frac{1}{\gamma_{\lambda_0}(\Sigma)} \frac{\dot{\boldsymbol{\eta}}_{\lambda_0}^{(p,q)\prime}\dot{\boldsymbol{\mu}}_{\lambda_0}}{\dot{\boldsymbol{\eta}}_{\lambda_0}^{(p,q)\prime}\Sigma\,\dot{\boldsymbol{\eta}}_{\lambda_0}^{(p,q)}}.$$

Here Σ, a symmetric nonnegative definite matrix discussed later, defines $\gamma_\lambda(\Sigma)$, the *curvature* of $\boldsymbol{\eta}_\lambda^{(p,q)}$ with respect to Σ,

$$\gamma_\lambda(\Sigma) = \left(\frac{\det(M_\lambda)}{\big(\dot{\boldsymbol{\eta}}_\lambda^{(p,q)\prime}\Sigma\,\dot{\boldsymbol{\eta}}_\lambda^{(p,q)}\big)^3}\right)^{1/2}$$

$$\text{with } M_\lambda = \begin{pmatrix} \dot{\boldsymbol{\eta}}_\lambda^{(p,q)\prime}\Sigma\,\dot{\boldsymbol{\eta}}_\lambda^{(p,q)} & \dot{\boldsymbol{\eta}}_\lambda^{(p,q)\prime}\Sigma\,\ddot{\boldsymbol{\eta}}_\lambda^{(p,q)} \\ \dot{\boldsymbol{\eta}}_\lambda^{(p,q)\prime}\Sigma\,\ddot{\boldsymbol{\eta}}_\lambda^{(p,q)} & \ddot{\boldsymbol{\eta}}_\lambda^{(p,q)\prime}\Sigma\,\ddot{\boldsymbol{\eta}}_\lambda^{(p,q)} \end{pmatrix}.$$

Proof. See Appendix B.

Remark 4. Two matrices are of particular interest: $\Sigma = I$ and $\Sigma = V_{\lambda_0} = \partial\boldsymbol{\mu}_\lambda/\partial\boldsymbol{\eta}_\lambda = \operatorname{diag}((c_q B_{\lambda_0 i})^{-(p+1)}/p)$. When $\Sigma = I$, $\gamma_\lambda(I)$ is the usual Euclidean curvature, and the theorem says that the minimum Euclidean distance from $\mathbf{u}^\dagger$ to $\boldsymbol{\mu}_{\lambda_0}^{(p,q)}$ is proportional to the radius of Euclidean curvature of $\boldsymbol{\eta}_\lambda^{(p,q)}$ at λ_0; when $\Sigma = V_{\lambda_0}$, $\gamma_\lambda(V_\lambda)$ is the statistical curvature defined by Efron (1975) and used by Efron (2001) to compare C_p and GML, and because $\frac{\partial}{\partial\lambda}\boldsymbol{\mu}_{\lambda_0} = V_{\lambda_0}\dot{\boldsymbol{\eta}}_{\lambda_0}^{(p,q)}$, the theorem says that the minimum Mahalanobis distance from $\mathbf{u}^\dagger$ to $\boldsymbol{\mu}_{\lambda_0}^{(p,q)}$ is equal to the radius of statistical curvature.

Table 5. Squared Statistical Curvature of the EE, C_p, and GML Criteria

	EE	*C_p*	GML
Experiment 1	.29	.70	.08
Experiment 2	.09	.23	.02

The theorem suggests that the curvature is a good measure of the reversal effect. Table 5 reports the squared statistical curvatures of the EE, C_p, and GML criteria on the two simulation experiments. The curvature of the EE criterion is much smaller than that of C_p.

Motivated by the reversal phenomenon on $\mathcal{L}_{\lambda_0}$, we extend it by defining the *reversal region* to be the region beyond the critical point $\mathbf{u}^\dagger$ (see Fig. 6),

$$\text{reversal region} = \big\{\mathbf{u} : \ddot{l}_{\lambda_0}^{(p,q)}(\mathbf{u}) - \beta_{\lambda_0}\dot{l}_{\lambda_0}^{(p,q)}(\mathbf{u}) < 0\big\},$$

where $\beta_{\lambda_0} = \ddot{\boldsymbol{\eta}}_{\lambda_0}^{(p,q)\prime}V_{\lambda_0}^2\dot{\boldsymbol{\eta}}_{\lambda_0}^{(p,q)}/(\dot{\boldsymbol{\eta}}_{\lambda_0}^{(p,q)\prime}V_{\lambda_0}^2\dot{\boldsymbol{\eta}}_{\lambda_0}^{(p,q)})$. For any point that falls into the reversal region, λ_0 is not even a local minimum of $l_\lambda^{(p,q)}$, implying that $\hat{\lambda}$ must be far from λ_0. Figure 7 plots

$$R_0 = \big[\ddot{l}_{\lambda_0}^{(p,q)}(\mathbf{u}) - \beta_{\lambda_0}\dot{l}_{\lambda_0}^{(p,q)}(\mathbf{u})\big]/\lambda_0^2 \tag{39}$$

(horizontal axis) versus the EE- and C_p-estimated degrees of freedom (vertical axis) for the 1000 simulations of Figure 1. It is quite evident that observations with a negative value of R_0 are likely to produce wild curves (i.e., curves with very large degrees of freedom) and vice versa.

The mean, variance, and skewness of R_0 are

$$M(R_0) = \lambda_0^{-2}\big[i_{\lambda_0} + \big(\beta_{\lambda_0}\dot{\boldsymbol{\eta}}_{\lambda_0}^{(p,q)} - \ddot{\boldsymbol{\eta}}_{\lambda_0}^{(p,q)}\big)'\big(\mathbf{u}_0 - \boldsymbol{\mu}_{\lambda_0}^{(p,q)}\big)\big], \tag{40}$$

$$V(R_0) = \lambda_0^{-4}\textstyle\sum_i\big(\beta_{\lambda_0}\dot{\eta}_{\lambda_0 i}^{(p,q)} - \ddot{\eta}_{\lambda_0 i}^{(p,q)}\big)^2 \operatorname{var} u_i, \tag{41}$$

and

$$S(R_0) = \frac{\sum_i(\beta_{\lambda_0}\dot{\eta}_{\lambda_0 i}^{(p,q)} - \ddot{\eta}_{\lambda_0 i}^{(p,q)})^3 E(u_i - u_{0i})^3}{[\sum_i(\beta_{\lambda_0}\dot{\eta}_{\lambda_0 i}^{(p,q)} - \ddot{\eta}_{\lambda_0 i}^{(p,q)})^2 \operatorname{var} u_i]^{3/2}} \tag{42}$$

A three-term Edgeworth expansion thus gives

$$P(\mathbf{u} \in RR) = P(R_0 < 0) \doteq \Phi\left(-\frac{M(R_0)}{\sqrt{V(R_0)}}\right) - \frac{1}{6}S(R_0)\left(\frac{M(R_0)^2}{V(R_0)} - 1\right)\varphi\left(-\frac{M(R_0)}{\sqrt{V(R_0)}}\right). \tag{43}$$

Table 6 records approximation (43) along with the Monte Carlo estimates of $P(\mathbf{u} \in RR)$. The EE criterion is seen to have smaller reversal region effect [in experiment 1, $P(\mathbf{u} \in RR) \doteq .088$ for EE versus .20 for C_p; in experiment 2, $P(\mathbf{u} \in RR) \doteq .002$ for EE versus .067 for C_p] and thus less variability. Computational details in the case of spline-like smoothers are provided in Appendix B.

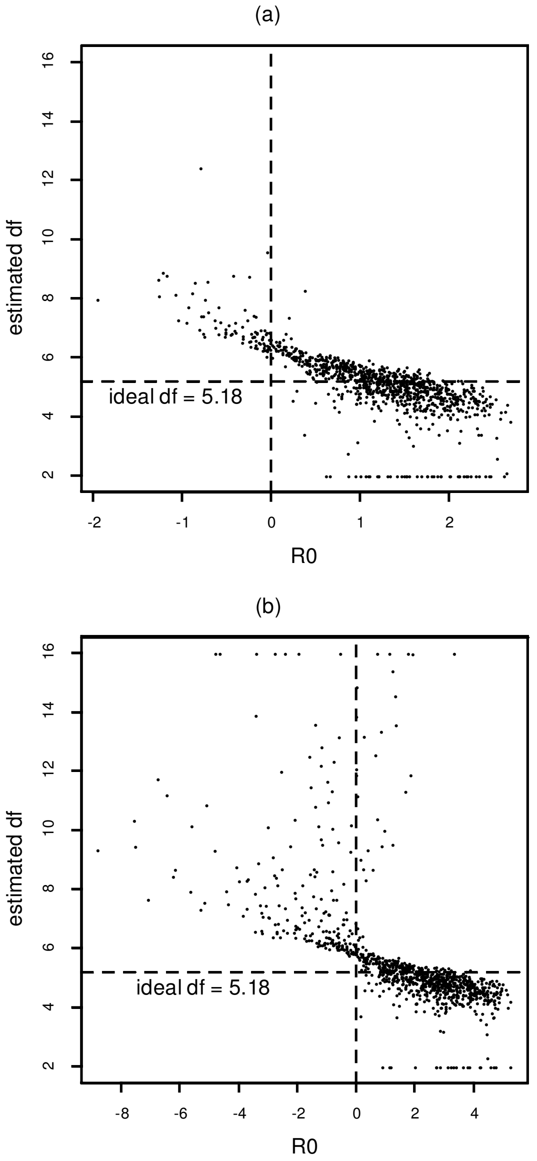

Figure 7. Plots Showing the Reversal Region Effect. (a) The EE estimated degrees of freedom against R_0 for the 1000 simulations of Figure 1. (b) The C_p estimates of Figure 1.

6. ADAPTIVE ESTIMATION OF f

In this section we consider the effect of using an estimated smoothing parameter on the subsequent estimation of $\mathbf{f}$, complementing our previous investigation of estimating df_0, the ideal degrees of freedom. The link between the two has been extensively studied by Efron (2001); also see the discussion of Section 2.3. Here, in the context of spline-like smoothers, we propose an approximation for the prediction error $E\|\hat{\mathbf{f}}_{\hat\lambda}-\mathbf{f}\|^2=\sigma^2E\|\hat{\mathbf{g}}_{\hat\lambda}-\mathbf{g}\|^2$, where $\hat{\mathbf{f}}_{\hat\lambda}=A_{\hat\lambda}\mathbf{y}, \hat{\mathbf{g}}_{\hat\lambda}=\mathbf{a}_{\hat\lambda}\mathbf{z}$.

Table 6. Approximation (43) and Empirically Observed P($\mathbf{u}\in RR$), the Probability of an Observation Falling Into the Reversal Region

	Experiment 1		*Experiment 2*	
	Approximation (43)	*Observed P($\mathbf{u}\in RR$)*	*Approximation (43)*	*Observed P($\mathbf{u}\in RR$)*
EE	.088	.084	.002	.002
C_p	.201	.195	.067	.065
GML	.019	.016	1.92×10^{-9}	0

The approximation starts from the decomposition of $E\|\hat{\mathbf{g}}_{\hat\lambda}-\mathbf{g}\|^2$,

$$\begin{aligned}E\|\hat{\mathbf{g}}_{\hat\lambda}-\mathbf{g}\|^2&=E\|(\hat{\mathbf{g}}_{\hat\lambda}-\hat{\mathbf{g}}_{\lambda_0})+(\hat{\mathbf{g}}_{\lambda_0}-\mathbf{g})\|^2\\&=E\|\hat{\mathbf{g}}_{\lambda_0}-\mathbf{g}\|^2+2E\{(\hat{\mathbf{g}}_{\lambda_0}-\mathbf{g})'(\hat{\mathbf{g}}_{\hat\lambda}-\hat{\mathbf{g}}_{\lambda_0})\}\\&\quad+E\|\hat{\mathbf{g}}_{\hat\lambda}-\hat{\mathbf{g}}_{\lambda_0}\|^2\\&=E\|\mathbf{a}_{\lambda_0}\mathbf{z}-\mathbf{g}\|^2+2E\{(\mathbf{a}_{\lambda_0}\mathbf{z}-\mathbf{g})'(\mathbf{a}_{\hat\lambda}\mathbf{z}-\mathbf{a}_{\lambda_0}\mathbf{z})\}\\&\quad+E\|(\mathbf{a}_{\hat\lambda}-\mathbf{a}_{\lambda_0})\mathbf{z}\|^2.\end{aligned}\tag{44}$$

Theorem 7. The squared estimation error $E\|\hat{\mathbf{g}}_{\hat\lambda}-\mathbf{g}\|^2$ has approximation

$$\begin{aligned}E\|\hat{\mathbf{g}}_{\hat\lambda}-\mathbf{g}\|^2\doteq E\|\hat{\mathbf{g}}_{\lambda_0}-\mathbf{g}\|^2+\xi_0E(\widehat{df}-df_0)^2\\+\kappa_0E(\widehat{df}-df_0)+R,\end{aligned}\tag{45}$$

where

$$\xi_0=\left(\sum_i a_{\lambda_0 i}b_{\lambda_0 i}\right)^{-2}\sum_i a_{\lambda_0 i}^2b_{\lambda_0 i}^2(g_i^2+1),$$

$$\kappa_0=2\left(\sum_i a_{\lambda_0 i}b_{\lambda_0 i}\right)^{-1}\sum_i a_{\lambda_0 i}b_{\lambda_0 i}[1-b_{\lambda_0 i}(g_i^2+1)],$$

and

$$R=\frac{\sum_i a_{\lambda_0 i}^2b_{\lambda_0 i}^2\operatorname{cov}((\widehat{df}-df_0)^2,z_i^2)}{(\sum_i a_{\lambda_0 i}b_{\lambda_0 i})^2}+\frac{2\sum_i a_{\lambda_0 i}b_{\lambda_0 i}\operatorname{cov}(\widehat{df},a_{\lambda_0 i}z_i^2-g_iz_i)}{\sum_i a_{\lambda_0 i}b_{\lambda_0 i}}.$$

The derivation of Theorem 7 is deferred to Appendix A. Notice that $\kappa_0/2$ is the weighted average of $1-b_{\lambda_0 i}(g_i^2+1)$, which is likely to be close to 0 for large i, because we expect $g_i\to0$ and $b_{\lambda_0 i}\to1$. The foregoing approximation hence suggests that the cost of adaptively estimating $\mathbf{f}$, besides the unavoidable error $E\|\hat{\mathbf{g}}_{\lambda_0}-\mathbf{g}\|^2$, comes mainly from $E(\widehat{df}-df_0)^2$, the MSE of estimating df_0, and the covariances between $\widehat{df}$ and $\mathbf{z}$. Approximation (45) also indicates that the excess risk due to adaptation (i.e., estimating λ from the data rather than using a fixed λ) arises from the variance of $\widehat{df}$ and the correlation between $\widehat{df}$ and $\mathbf{z}$.

Approximation (45) is not yet immediately useful because of the unknown covariances. Expansion (27) provides simple approximations for them,

$$\begin{aligned}\operatorname{cov}(\widehat{df},z_i^2)&\doteq\frac{\partial\widehat{df}}{\partial u_i}\Big|_{\mathbf{u}_0}\operatorname{cov}(u_i,z_i^2),\\\operatorname{cov}(\widehat{df},z_i)&\doteq\frac{\partial\widehat{df}}{\partial u_i}\Big|_{\mathbf{u}_0}\operatorname{cov}(u_i,z_i),\\\operatorname{cov}\big((\widehat{df}-df_0)^2,z_i^2\big)&\doteq2(df_1-df_0)\left(\frac{\partial\widehat{df}}{\partial u_i}\Big|_{\mathbf{u}_0}\right)\operatorname{cov}(u_i,z_i^2)\\&\quad+\left(\frac{\partial\widehat{df}}{\partial u_i}\Big|_{\mathbf{u}_0}\right)^2\operatorname{cov}\big((u_i-u_{0i})^2,z_i^2\big),\end{aligned}\tag{46}$$

Table 7. The Error of Estimating the Curve **g**

		Empirical (SE)	*Approximation (45)*	*Sum II*
Experiment 1	EE	5.75 (.12)	5.49	4.94
	C_p	6.10 (.14)	5.60	5.10
	GML	5.81 (.12)	5.46	4.93
Experiment 2	EE	12.26 (.15)	12.32	11.97
	C_p	12.47 (.16)	12.16	11.84
	GML	12.72 (.15)	12.98	12.68

NOTE: Empirical mean (and standard error) from simulations together with approximation (45); sum II is the sum of the first two terms of (45).

where the formula for $\partial\widehat{df}/\partial u_i|_{\mathbf{u}_0}$ is given by (29) and the covariances between $\mathbf{u}$ and $\mathbf{z}$ can be calculated by Lemma 1 or the method used to derive it (see App. A); for example,

$$\begin{aligned} \operatorname{cov}(u_i, z_i) &= g_i 2^{1+1/q}\Gamma(1/q+1/2) \\ &\quad \times M(1-1/q, 3/2, -g_i^2/2)/(q\sqrt{\pi}), \\ E(u_i z_i^2) &= 2^{1+1/q}\Gamma(1/q+3/2) \\ &\quad \times M(-1-1/q, 1/2, -g_i^2/2)/\sqrt{\pi}, \\ E(u_i^2 z_i^2) &= 2^{1+2/q}\Gamma(2/q+3/2) \\ &\quad \times M(-1-2/q, 1/2, -g_i^2/2)/\sqrt{\pi}. \end{aligned} \qquad (47)$$

Table 7 reports the performance of approximation (45) on the two simulations, where, in addition to (46), $E(\widehat{df}-df_0)$ is estimated by $df_1 - df_0$ and $E(\widehat{df}-df_0)^2$ is estimated by $(df_1-df_0)^2+\operatorname{var}(\widehat{df})$ through (30). It is worth pointing out that in real applications, where the true curve $\mathbf{g}$ is unknown, (45) can be used together with the resubstitution method of Section 4.2 to estimate the total estimation error. The approximation is seen to work reasonably well on our two cases. Table 7 also shows that the first two terms in (45), $E\|\hat{\mathbf{g}}_{\lambda_0}-\mathbf{g}\|^2+\xi_0 E(\widehat{df}-df_0)^2$, account for more than 90% of the total estimation error, suggesting their role as a moderate indicator of total estimation error. Not surprisingly, the EE criterion works "robustly" in terms of estimating the curve, consistent with our previous results.

7. A MARGINAL BAYESIAN INTERPRETATION FOR THE EXTENDED EXPONENTIAL CRITERION

We have been working mainly under the frequentist model $\mathbf{y} \sim (\mathbf{f}, \sigma^2 I)$, although the GML criterion was derived in the Bayesian framework (3). Actually, like GML, the EE criterion has a marginal Bayesian interpretation: instead of the GML marginal density $\mathbf{w} \sim \exp(-\frac{1}{2}\sum_i(b_{\lambda i}w_i - \log b_{\lambda i}))/\prod_i\sqrt{2\pi w_i}$, if we assume the density

$$u_i = w_i^{2/3} \overset{\text{ind.}}{\sim} \exp\Bigl(-C_0\bigl(c_q^{3/2}b_{\lambda i}u_i - 3c_q^{1/2}b_{\lambda i}^{1/3}\bigr)\Bigr)d_0^{EE}(u_i), \qquad (48)$$

then calculating the MLE yields the EE criterion (16), because $\sum_i[c_q b_{\lambda i}w_i^{2/3} - 3b_{\lambda i}^{1/3}]$ is the negative log-likelihood up to a constant. [The properties of the "carrier" $d_0^{EE}(u_i)$ are discussed later.] More generally, every member of the (p,q) family (19) has a marginal Bayesian interpretation; if we assume that $\mathbf{u} = \mathbf{w}^{1/q}$ comes from the marginal density

$$\mathbf{u} = \mathbf{w}^{1/q} \sim \begin{cases} \exp\Bigl(-C_0\sum_i\bigl[(c_qB_{\lambda i})^p u_i - \frac{p}{p-1}(c_qB_{\lambda i})^{p-1}\bigr]\Bigr)d_0^{(p,q)}(\mathbf{u}) \\ \quad \text{if } p>1 \\ \exp\Bigl(-C_0\sum_i\bigl[c_qB_{\lambda i}u_i - \log(c_qB_{\lambda i})\bigr]\Bigr)d_0^{(1,q)}(\mathbf{u}) \\ \quad \text{if } p=1 \end{cases} \qquad (49)$$

rather than the GML density, then the MLE gives the (p,q) criterion. The density (49) is a curved exponential family as a function of the parameter λ, which means that it can be written as

$$\mathbf{u} = \mathbf{w}^{1/q} \sim \exp(\boldsymbol{\eta}_\lambda'\mathbf{u} - \psi_\lambda)d_0^{(p,q)}(\mathbf{u}),$$

where $\boldsymbol{\eta}_\lambda = -C_0(c_q\mathbf{B}_\lambda)^p$ is the natural parameter vector and ψ_λ is the cumulant generating function given by

$$\psi_\lambda = \psi(\boldsymbol{\eta}_\lambda) = \begin{cases} -C_0\sum_i\frac{p}{p-1}(c_qB_{\lambda i})^{p-1} & \text{if } p>1 \\ -C_0\sum_i\log B_{\lambda i} & \text{if } p=1 \end{cases}. \qquad (50)$$

Because of the one-to-one correspondence between a density and its cumulant generating function, (50) in fact completely determines the distribution. For instance, the C_p marginal density is inverse Gaussian, which was shown by Efron (2001), and in the EE marginal distribution (48), the normalizing density $d_0^{EE}(u_i)$ in (48) follows a positive stable law with order 1/3, and, consequently, its exponential tilts give the EE marginal density. (For reference of stable laws, see Feller 1971.) Compared with the GML density, which is based on the assumption that $w_i \overset{\text{ind.}}{\sim} \chi_1^2/b_{\lambda i}$ (see Sec. 2.1), the stable law with order 1/3 has a heavier tail. This suggests that from a marginal Bayesian standpoint, the EE criterion is more robust, intuitively agreeing with our earlier analysis.

Remark 5. The term "extended exponential" comes from the fact that the EE density is an exponential family, which can be viewed as extended from the C_p and GML densities.

The fact that EE, C_p, and GML all have a Bayesian interpretation suggests performing a Bayesian simulation to supplement the frequentist simulation of Figures 1 and 2. We take $\mathbf{x}$ to be 61 equally spaced points on the $[-1,1]$ interval and take the true degrees of freedom to be 8, which determines the smoothing parameter λ through $\sum_{i=1}^{61}a_{\lambda i} = \sum_{i=1}^{61}1/(1+\lambda k_i) = 8$. Then we draw 1000 samples of $\mathbf{w}$ from the GML and C_p marginal densities. For each of these 1000 samples, we apply the EE, C_p, and GML criteria to choose the degrees of freedom.

Figure 8 displays the simulation results. The top row (a) shows the result of applying the three criteria to the 1000 samples from the GML density. Not surprisingly, GML works very well, whereas C_p suffers from instability. The bottom row (b) shows the performance of the three criteria on the 1000 samples from the C_p density. Compared to the previous case, C_p performs much better, because it is working in its own family, whereas the performance of GML deteriorates. In both cases, the EE criterion works well, even though neither

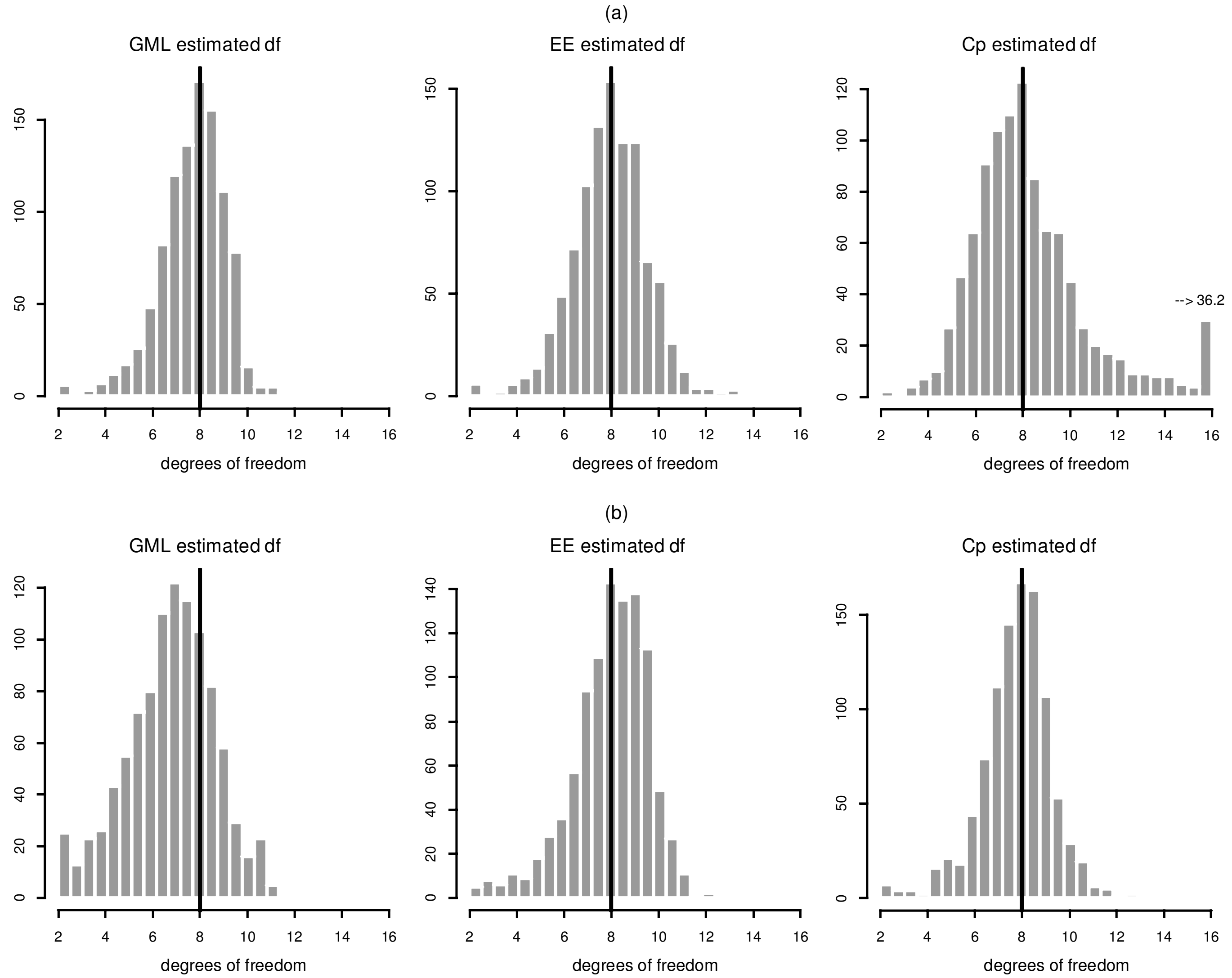

Figure 8. Bayesian Simulation. (a) The performance of EE, C_p, and GML on the 1000 samples from the GML density; (b) the performance of EE, C_p, and GML on the 1000 samples from the C_p marginal density. The true degrees of freedom are 8 in both cases.

of the cases comes from the EE marginal distribution. Figure 8 thus gives the EE criterion some Bayesian support.

Remark 6. In contrast to its frequentist counterpart in which the curve is fixed, we can think of the Bayesian simulation as first drawing 1000 curves from some (possibly implicit) prior, then generating 1 sample from each curve. Figure 8 also reflects the variability of the underlying curves.

8. DISCUSSION

Working on spline-like smoothers, in this article we have studied the geometry of selection criteria. This not only gives an intuitive explanation of the weakness and strength of the C_p and GML criteria, but also motivates the EE criterion. EE smoothly and robustly combines the strengths of GML and C_p, giving smaller bias, smaller variance, and smaller reversal effects. The simulation as well as theoretical results based on the geometry support EE (particularly its bias-corrected version) as a competitive criterion in both the frequentist and Bayesian sense. An interesting open problem is to generalize the EE criterion to general linear smoothers. Although the class of spline-like smoothers is large, such a generalization would certainly be desirable. We conclude this article by discussing the case in which the variance σ^2 must be estimated from the data.

If σ^2 is unknown in the model $\mathbf{y} \sim (\mathbf{f}, \sigma^2 I)$, then we can replace it with an estimate $\tilde\sigma^2$. Definitions (6), (7), and (9) give $\tilde{\mathbf{z}} \equiv U'y/\tilde\sigma = \mathbf{z}\cdot(\sigma/\tilde\sigma)$ and $\tilde{\mathbf{w}} \equiv \tilde{\mathbf{z}}^2$,

$$\tilde{\mathbf{u}} = \tilde{\mathbf{w}}^{1/q} = \mathbf{u}\cdot R, \quad \text{where } R = (\sigma^2/\tilde\sigma^2)^{1/q},$$

leading to estimators $\tilde\lambda^{(p,q)} = \arg\min_\lambda\{l_\lambda^{(p,q)}(\tilde{\mathbf{u}})\}$, and likewise $\widetilde{df}^{(p,q)}$; see (19). For $p = 2$ and $q = 1$ (the C_p case), this amounts to substituting $\tilde\sigma^2$ for σ^2 in (8). The close connection between $\tilde\lambda^{(2,1)}$ and Wahba's (1985) GCV criterion is explained at the end of section 4 of Efron (2001), complementing a similar discussion there for GML at the end of Section 3.

If $R \sim (1, \operatorname{var}_R)$ independently of $\mathbf{u}$, then it is easy to see that

$$\tilde{\mathbf{u}} \sim (\mathbf{u}_0, \operatorname{var}\mathbf{u} + \operatorname{var}_R \cdot (\mathbf{u}_0\mathbf{u}_0' + \operatorname{var}\mathbf{u})), \qquad (51)$$

where $\text{var}\,\mathbf{u} = \text{diag}(\text{var}\,u_i)$. The influence function calculations in Section 4 lead from (51) to the useful approximation

$$\frac{\text{var}\{\widetilde{df}^{(p,q)}\}}{\text{var}\{\widehat{df}^{(p,q)}\}} \doteq 1 + \text{var}_R \cdot \left[1 + \frac{(\sum_i a_{\lambda_1 i} B_{\lambda_1 i}^{p-1}/c_q)^2}{\sum_i a_{\lambda_1 i}^2 B_{\lambda_1 i}^{2p}\,\text{var}\,u_i}\right], \qquad (52)$$

using the identity $\sum_i a_{\lambda_1 i} B_{\lambda_1 i}^{p} u_{0i} = \sum_i a_{\lambda_1 i} B_{\lambda_1 i}^{p-1}/c_q$, which can be derived from the normal equation (20). In practice, the estimate $\tilde{\sigma}^2$ can be based on the higher components of $U'\mathbf{y} \sim (\sigma\mathbf{g}, \sigma^2 I)$, as in (8.11) of Efron (2001),

$$\tilde{\sigma}^2 = \sum_{n-1-M}^{n} (U'\mathbf{y})_i^2/(M-2),$$

because the assumed smoothness of $\mathbf{f}$ implies that $g_i \doteq 0$ for i large and that $\tilde{\sigma}^2$ and $\mathbf{u}$ are nearly independent. The Gaussian model $\mathbf{y} \sim N(\mathbf{f}, \sigma^2 I)$ then has $\text{var}\{\sigma^2/\tilde{\sigma}^2\} = 2/(M-4)$. Applied with $M = 40$, this together with (52) gives $se(\widetilde{df})/se(\widehat{df}) = 1.17$ for C_p and 1.18 for GML, compared with a simulation value of 1.22 (with $se = .03$) for the latter.

Formula (52) shows the loss of precision in $\widehat{df}^{(p,q)}$ from having to estimate σ^2. Results like those in Sections 4–6 can be developed similarly for the unknown σ^2 case, at the expense of complications due to the $\tilde{\mathbf{u}}$ nondiagonal covariance matrix in (51).

APPENDIX A: PROOFS

Proof of Theorem 1

According to (6), $\| \hat{\mathbf{f}}_\lambda - \mathbf{f} \|^2 = \sigma^2 \| \hat{\mathbf{g}}_\lambda - \mathbf{g} \|^2$. It follows that

$$E\|\hat{\mathbf{f}}_\lambda - \mathbf{f}\|^2 = \sigma^2 E\|\hat{\mathbf{g}}_\lambda - \mathbf{g}\|^2 = \sigma^2\left[\sum_{i=1}^{n} E(a_{\lambda i} z_i - g_i)^2\right]$$
$$= \sigma^2\left[\sum_{i=1}^{n}\big(b_{\lambda i}^2(g_i^2+1) - 2b_{\lambda i}\big) + n\right].$$

Consequently, $\lambda_0 = \arg\min_\lambda E\|\hat{\mathbf{f}}_\lambda - \mathbf{f}\|^2 = \arg\min_\lambda \sum_i \{b_{\lambda i}^2(g_i^2+1) - 2b_{\lambda i}\}$.

Proof of Theorem 2

Applying a Taylor expansion on λ around λ_0 yields

$$E \| \hat{\mathbf{f}}_\lambda - \mathbf{f} \|^2 = E \| \hat{\mathbf{f}}_{\lambda_0} - \mathbf{f} \|^2 + \frac{1}{2}\left(\frac{\partial^2}{\partial\lambda^2} E \| \hat{\mathbf{f}}_\lambda - \mathbf{f} \|^2\right)\bigg|_{\lambda=\lambda_0} (\lambda - \lambda_0)^2 + o(|\lambda-\lambda_0|^2). \qquad (A.1)$$

Notice that the first derivative term vanishes because λ_0 is the minimizer of $E \| \hat{\mathbf{f}}_\lambda - \mathbf{f} \|^2$. Because λ and df_λ have a one-to-one correspondence, we have

$$\lambda - \lambda_0 = \left(\frac{\partial df_\lambda}{\partial\lambda}\bigg|_{\lambda=\lambda_0}\right)^{-1} (df_\lambda - df_{\lambda_0}) + o(|df_\lambda - df_{\lambda_0}|). \qquad (A.2)$$

Substituting this into (A.1) yields $E_{\mathbf{f}} \| \hat{\mathbf{f}}_\lambda - \mathbf{f} \|^2 = E \| \hat{\mathbf{f}}_{\lambda_0} - \mathbf{f} \|^2 + d_0 \cdot (df_\lambda - df_{\lambda_0})^2 + o(|df_\lambda - df_{\lambda_0}|^2)$, where the constant

$$d_0 = \frac{1}{2}\left[\frac{\partial^2}{\partial\lambda^2} E \| \hat{\mathbf{f}}_\lambda - \mathbf{f} \|^2 \Big/ \left(\frac{\partial df_\lambda}{\partial\lambda}\right)^2\right]\Bigg|_{\lambda=\lambda_0}$$
$$= \sigma^2 \sum_i \left[\left(\left(\frac{\partial}{\partial\lambda} b_{\lambda_0 i}\right)^2 + \left(b_{\lambda_0 i}\frac{\partial^2}{\partial\lambda^2} b_{\lambda_0 i}\right)\right)(g_i^2+1) - \frac{\partial^2}{\partial\lambda^2} b_{\lambda_0 i}\right]$$
$$\times \left(\sum_i \frac{\partial}{\partial\lambda} a_{\lambda_0 i}\right)^{-2}.$$

The relationships $(\partial/\partial\lambda) b_{\lambda i} = a_{\lambda i} b_{\lambda i}/\lambda$ and $(\partial^2/\partial\lambda^2) b_{\lambda i} = -2 a_{\lambda i} b_{\lambda i}^2/\lambda^2$ come directly from the fact that $a_{\lambda i} = 1/(1+\lambda k_i)$.

Proof of Lemma 1

$$E w_i^r = E|z_i|^{2r} = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} |x|^{2r} \exp\left(-\frac{(x-g_i)^2}{2}\right) dx$$
$$= \frac{1}{\sqrt{2\pi}}\int_{g_i}^{\infty} (x-g_i)^{2r} e^{-x^2/2}\, dx$$
$$+ \frac{1}{\sqrt{2\pi}}\int_{-g_i}^{\infty} (x+g_i)^{2r} e^{-x^2/2}\, dx \qquad (A.3)$$

Borrowing a special function, *Hh* function, from the mathematical physics literature, which, for example, Abramowitz and Stegun (1972, p. 691) define as

$$\begin{cases} \text{Hh}_r(x) = \dfrac{1}{\Gamma(r+1)}\displaystyle\int_x^{\infty} (t-x)^r e^{-\frac{1}{2}t^2}\, dt & \text{for } r \ge 0 \\ \text{Hh}_{-1}(x) = e^{-\frac{1}{2}x^2} \end{cases}.$$

(A.3) can be concisely expressed as $E w_i^r = (\Gamma(2r+1)/\sqrt{2\pi})[\text{Hh}_{2r}(g_i) + \text{Hh}_{2r}(-g_i)]$. An interesting connection between the *Hh* function and another special function, the parabolic cylinder function U, which is widely used in mathematical physics, $\text{Hh}_r(x) = e^{-x^2/4} U(r+\frac{1}{2}, x)$, gives $E w_i^r = (\Gamma(2r+1)/\sqrt{2\pi}) e^{-g_i^2/4}[U(2r+\frac{1}{2}, g_i) + U(2r+\frac{1}{2}, -g_i)]$. The parabolic cylinder function U itself (e.g., according to formula 19.12.3 of Abramowitz and Stegun 1972), satisfies $U(a,x) + U(a,-x) = (2\sqrt{\pi}2^{-\frac{1}{4}-\frac{1}{2}a} e^{-x^2/4})/(\Gamma(\frac{3}{4}+\frac{1}{2}a))M(\frac{1}{2}a+\frac{1}{4}, \frac{1}{2}, \frac{1}{2}x^2)$, where $M(\cdot,\cdot,\cdot)$ is the confluent hypergeometric function. It finally follows that

$$E w_i^r = \frac{\Gamma(2r+1)}{\Gamma(r+1)} 2^{-r} e^{-\frac{1}{2}g_i^2} M\left(\frac{1}{2}+r, \frac{1}{2}, \frac{1}{2}g_i^2\right)$$
$$= \frac{1}{\sqrt{\pi}} 2^r \Gamma\left(r+\frac{1}{2}\right) M\left(-r, \frac{1}{2}, -\frac{1}{2}g_i^2\right).$$

The last step uses two identities: $M(a,b,z) = e^z M(b-a, b, -z)$ and $\sqrt{\pi}\Gamma(2x) = 2^{2x-1}\Gamma(x)\Gamma(x+\frac{1}{2})$.

Proof of Theorem 7

Viewing $\mathbf{a}_\lambda$ as a function of df_λ and applying a first-order Taylor expansion on $a_{\hat{\lambda} i}$ around $a_{\lambda_0 i}$ yields

$$a_{\hat{\lambda} i} \doteq a_{\lambda_0 i} + \frac{\partial a_{\lambda i}}{\partial df}\bigg|_{\lambda=\lambda_0} (\widehat{df} - df_0)$$
$$= a_{\lambda_0 i} + \frac{a_{\lambda_0 i} b_{\lambda_0 i}}{\sum_j a_{\lambda_0 j} b_{\lambda_0 j}} (\widehat{df} - df_0).$$

Applying it on the decomposition (44) gives

$$E\|\hat{\mathbf{g}}_{\lambda_0}-\mathbf{g}\|^2=\sum_{i=1}^{n}\left[b_{\lambda_0 i}^2(g_i^2+1)-2b_{\lambda_0 i}+1\right],$$

$$2E(\hat{\mathbf{g}}_{\lambda_0}-\mathbf{g})'(\hat{\mathbf{g}}_{\hat{\lambda}}-\hat{\mathbf{g}}_{\lambda_0})\doteq 2\left(\sum_i a_{\lambda_0 i}b_{\lambda_0 i}\right)^{-1}\sum_i a_{\lambda_0 i}b_{\lambda_0 i}\times E\left\{(\widehat{df}-df_0)(a_{\lambda_0 i}z_i^2-g_iz_i)\right\},$$

and

$$E\|\hat{\mathbf{g}}_{\hat{\lambda}}-\hat{\mathbf{g}}_{\lambda_0}\|^2\doteq\left(\sum_i a_{\lambda_0 i}b_{\lambda_0 i}\right)^{-2}\sum_i a_{\lambda_0 i}^2b_{\lambda_0 i}^2E\left\{(\widehat{df}-df_0)^2z_i^2\right\}.$$

Because

$$\begin{aligned}E&\left\{(\widehat{df}-df_0)(a_{\lambda_0 i}z_i^2-g_iz_i)\right\}\\&=E(\widehat{df}-df_0)\cdot E(a_{\lambda_0 i}z_i^2-g_iz_i)+\mathrm{cov}(\widehat{df},a_{\lambda_0 i}z_i^2-g_iz_i)\\&=[1-b_{\lambda_0 i}(g_i^2+1)]E(\widehat{df}-df_0)+\mathrm{cov}(\widehat{df},a_{\lambda_0 i}z_i^2-g_iz_i)\end{aligned}$$

and

$$\begin{aligned}E\{(\widehat{df}-df_0)^2z_i^2\}&=E(\widehat{df}-df_0)^2\cdot E\{z_i^2\}+\mathrm{cov}((\widehat{df}-df_0)^2,z_i^2)\\&=(g_i^2+1)E(\widehat{df}-df_0)^2+\mathrm{cov}((\widehat{df}-df_0)^2,z_i^2),\end{aligned}$$

we have $E\|\hat{\mathbf{g}}_{\hat{\lambda}}-\mathbf{g}\|^2\doteq E\|\hat{\mathbf{g}}_{\lambda_0}-\mathbf{g}\|^2+\xi_0E(\widehat{df}-df_0)^2+\kappa_0E(\widehat{df}-df_0)+R$, as in (45).

APPENDIX B: DETAILS OF THE REVERSAL EFFECT

Proof of Theorem 6

Suppressing the scripts, we have a quadratic optimization problem,

$$\min_{\mathbf{u}}(\mathbf{u}-\boldsymbol{\mu}_\lambda)'\Sigma^{-1}(\mathbf{u}-\boldsymbol{\mu}_\lambda),\qquad\text{(B.1)}$$

subject to $\dot{l}_\lambda(\mathbf{u})=0$ and $\ddot{l}_\lambda(\mathbf{u})=0$. The Lagrangian is $h(\mathbf{u},\alpha,\beta)=(\mathbf{u}-\boldsymbol{\mu}_\lambda)'\Sigma^{-1}(\mathbf{u}-\boldsymbol{\mu}_\lambda)-\alpha\dot{l}_\lambda(\mathbf{u})-\beta\ddot{l}_\lambda(\mathbf{u})$. Setting $\frac{\partial h}{\partial\mathbf{u}}=\frac{\partial h}{\partial\alpha}=\frac{\partial h}{\partial\beta}=0$ and solving the resulting linear equations gives the minimum distance,

$$\begin{aligned}\min_{\mathbf{u}}&\left[(\mathbf{u}-\boldsymbol{\mu}_\lambda)'\Sigma^{-1}(\mathbf{u}-\boldsymbol{\mu}_\lambda)\right]^{1/2}\\&=\left(\frac{\ddot{\boldsymbol{\eta}}_\lambda'\Sigma\ddot{\boldsymbol{\eta}}_\lambda}{(\dot{\boldsymbol{\eta}}_\lambda'\Sigma\dot{\boldsymbol{\eta}}_\lambda)^2}-\frac{(\ddot{\boldsymbol{\eta}}_\lambda'\Sigma\dot{\boldsymbol{\eta}}_\lambda)^2}{(\dot{\boldsymbol{\eta}}_\lambda'\Sigma\dot{\boldsymbol{\eta}}_\lambda)^3}\right)^{-\frac{1}{2}}\frac{i_\lambda}{\dot{\boldsymbol{\eta}}_\lambda'\Sigma\dot{\boldsymbol{\eta}}_\lambda}\\&=\frac{1}{\gamma_\lambda}\frac{i_\lambda}{\dot{\boldsymbol{\eta}}_\lambda'\Sigma\dot{\boldsymbol{\eta}}_\lambda},\end{aligned}$$

which is achieved at

$$\mathbf{u}(\lambda)=\boldsymbol{\mu}_\lambda+\frac{1}{\gamma_\lambda^2}\frac{i_\lambda}{(\dot{\boldsymbol{\eta}}_\lambda'\Sigma\dot{\boldsymbol{\eta}}_\lambda)^2}\Sigma\left(\ddot{\boldsymbol{\eta}}_\lambda-\frac{\ddot{\boldsymbol{\eta}}_\lambda'\Sigma\dot{\boldsymbol{\eta}}_\lambda}{\dot{\boldsymbol{\eta}}_\lambda'\Sigma\dot{\boldsymbol{\eta}}_\lambda}\dot{\boldsymbol{\eta}}_\lambda\right).$$

For $\Sigma=V_\lambda$, the minimum distance is simplified to the radius of statistical curvature, $1/\gamma_\lambda$. The proof also says that the trajectory of the point minimizing (B.1) is given by $\mathbf{u}(\lambda)=\boldsymbol{\mu}_\lambda+o_\lambda$, where

$$o_\lambda=\frac{1}{i_\lambda\gamma_\lambda^2}V_\lambda\left(\ddot{\boldsymbol{\eta}}_\lambda-\frac{\ddot{\boldsymbol{\eta}}_\lambda'V_\lambda\dot{\boldsymbol{\eta}}_\lambda}{\dot{\boldsymbol{\eta}}_\lambda'V_\lambda\dot{\boldsymbol{\eta}}_\lambda}\dot{\boldsymbol{\eta}}_\lambda\right).$$

So ideally, the reversal region should be defined as the region consisting of the points lying above the hyperplane that passes through the intersection of $\{\mathbf{u}:\dot{l}_\lambda(\mathbf{u})=0\}$ and $\{\mathbf{u}:\ddot{l}_\lambda(\mathbf{u})=0\}$ and is tangent to the curve $\{\mathbf{u}(\lambda)\}$. Any hyperplane passing through the intersection of $\{\mathbf{u}:\dot{l}_\lambda(\mathbf{u})=0\}$ and $\{\mathbf{u}:\ddot{l}_\lambda(\mathbf{u})=0\}$, except $\mathcal{L}_\lambda$ itself, has the form $\{\mathbf{u}:\ddot{l}_\lambda(\mathbf{u})-\beta\dot{l}_\lambda(\mathbf{u})=0\}$ for some $\beta\in\mathrm{R}$. It then follows that β should be taken as

$$\frac{\ddot{\boldsymbol{\eta}}_\lambda'V_\lambda\dot{\mathbf{u}}(\lambda)}{\dot{\boldsymbol{\eta}}_\lambda'V_\lambda\dot{\mathbf{u}}(\lambda)}=\frac{\ddot{\boldsymbol{\eta}}_\lambda'V_\lambda\dot{\boldsymbol{\mu}}_\lambda+\ddot{\boldsymbol{\eta}}_\lambda'V_\lambda\dot{o}_\lambda}{\dot{\boldsymbol{\eta}}_\lambda'V_\lambda\dot{\boldsymbol{\mu}}_\lambda+\dot{\boldsymbol{\eta}}_\lambda'V_\lambda\dot{o}_\lambda}.$$

A simple approximation for β is $\ddot{\boldsymbol{\eta}}_\lambda'V_\lambda^2\dot{\boldsymbol{\eta}}_\lambda/\dot{\boldsymbol{\eta}}_\lambda'V_\lambda^2\dot{\boldsymbol{\eta}}_\lambda$, which gives the reversal region defined in Section 5. Besides being simple, simulation results like those in Figure 6 give the best support for this definition.

Computational Details in the Case of Spline-Like Smoothers

$$\gamma_\lambda^2(I)=\frac{(p+q)^2}{p^2c_q^{2p}}\left\{\frac{\sum_i a_{\lambda i}^4B_{\lambda i}^{2p}}{(\sum_i a_{\lambda i}^2B_{\lambda i}^{2p})^2}-\frac{(\sum_i a_{\lambda i}^3B_{\lambda i}^{2p})^2}{(\sum_i a_{\lambda i}^2B_{\lambda i}^{2p})^3}\right\},$$

$$\gamma_\lambda^2(V_\lambda)=\frac{(p+q)^2}{pc_q^{p-1}}\left\{\frac{\sum_i a_{\lambda i}^4B_{\lambda i}^{p-1}}{(\sum_i a_{\lambda i}^2B_{\lambda i}^{p-1})^2}-\frac{(\sum_i a_{\lambda i}^3B_{\lambda i}^{p-1})^2}{(\sum_i a_{\lambda i}^2B_{\lambda i}^{p-1})^3}\right\},$$

$$\beta_{\lambda_0}=-\frac{1}{\lambda_0}\left[2-\left(1+\frac{p}{q}\right)\frac{\sum_i a_{\lambda_0 i}^3B_{\lambda_0 i}^{-2}}{\sum_i a_{\lambda_0 i}^2B_{\lambda_0 i}^{-2}}\right],$$

$$\begin{aligned}M(R_0)=\frac{p(p+q)c_q^{p-1}}{q^2\lambda_0^4}&\left\{\frac{1}{p+q}\left(\sum_i a_{\lambda_0 i}^2B_{\lambda_0 i}^{p-1}\right)\right.\\&\left.+\sum_i\left[a_{\lambda_0 i}B_{\lambda_0 i}^{p-1}\left(a_{\lambda_0 i}-\frac{\sum_i a_{\lambda_0 i}^3B_{\lambda_0 i}^{-2}}{\sum_i a_{\lambda_0 i}^2B_{\lambda_0 i}^{-2}}\right)(c_qB_{\lambda_0 i}u_{0i}-1)\right]\right\},\end{aligned}$$

$$V(R_0)=\frac{p^2(p+q)^2c_q^{2p}}{q^4\lambda_0^8}\sum_i\left[a_{\lambda_0 i}^2B_{\lambda_0 i}^{2p}\left(a_{\lambda_0 i}-\frac{\sum_i a_{\lambda_0 i}^3B_{\lambda_0 i}^{-2}}{\sum_i a_{\lambda_0 i}^2B_{\lambda_0 i}^{-2}}\right)^2\mathrm{var}\,u_i\right],$$

and

$$S(R_0)=\frac{\sum_i a_{\lambda_0 i}^3B_{\lambda_0 i}^{3p}\left(a_{\lambda_0 i}-\frac{\sum_i a_{\lambda_0 i}^3B_{\lambda_0 i}^{-2}}{\sum_i a_{\lambda_0 i}^2B_{\lambda_0 i}^{-2}}\right)^3E(u_i-u_{0i})^3}{\left(\sum_i a_{\lambda_0 i}^2B_{\lambda_0 i}^{2p}\left(a_{\lambda_0 i}-\frac{\sum_i a_{\lambda_0 i}^3B_{\lambda_0 i}^{-2}}{\sum_i a_{\lambda_0 i}^2B_{\lambda_0 i}^{-2}}\right)^2\mathrm{var}\,u_i\right)^{3/2}}.$$

[Received February 2001. Revised November 2001.]

REFERENCES

Abramowitz, M., and Stegun, I. (1972), *Handbook of Mathematical Function*, Washington, DC: U.S. National Bureau of Standards.

Akaike, H. (1974), "A New Look at Statistical Model Identification," *IEEE Transactions on Automatic Control*, AU-19, 716–722.

Bowman, A., and Azzalini, A. (1997), *Applied Smoothing Techniques for Data Analysis: The Kernel Approach With S-PLUS Illustrations*, New York: Oxford University Press.

Craven, P., and Wahba, G. (1979), "Smoothing Noisy Data With Spline Functions: Estimating the Correct Degree of Smoothing by the Method of Generalized Cross-Validation," *Numerical Mathematics*, 31, 377–403.

Efron, B. (1975), "Defining the Curvature of a Statistical Problem (With Application to Second-Order Efficiency)" (with discussion), *The Annals of Statistics*, 3, 1189–1242.

——— (1986), "How Biased is the Apparent Error Rate of a Prediction Rule?," *Journal of the American Statistical Association*, 81, 461–470.

——— (2000), "Smoothers and the Cost of Model Selection," Technical Report 209, Stanford University, Dept. Statistics.

——— (2001), "Selection Criteria for Scatterplot Smoothers," *The Annals of Statistics*, 29, 470–504.

Eubank, R. (1988), *Spline Smoothing and Nonparametric Regression*, New York: Marcel Dekker.

Feller, W. (1971), *An Introduction to Probability Theory and Its Applications*, Vol. II, New York: Wiley.

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

Gu, C. (1998), "Model Indexing and Smoothing Parameter Selection in Nonparametric Function Estimation" (with discussion), *Statistica Sinica*, 8, 607–646.

Hall, P., and Johnstone, I. (1992), "Empirical Functionals and Efficient Smoothing Parameter Selection" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 54, 475–530.

Härdle, W. (1990), *Applied Nonparametric Regression*, Cambridge, U.K.: Cambridge University Press.

Härdle, W., Hall, P., and Marron, J. (1988), "How Far are Automatically Chosen Regression Smoothing Parameters From the Optimum" (with discussion), *Journal of the American Statistical Association*, 83, 86–101.

Hastie, T. (1996), "Pseudosplines," *Journal of the Royal Statistical Society*, Ser. B, 58, 379–396.

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

Hurvich, C., Simonoff, J., and Tsai, C. (1998), "Smoothing Parameter Selection in Nonparametric Regression Using an Improved Akaike Information Criterion," *Journal of the Royal Statistical Society*, Ser. C, 60, 271–294.

Jones, M., Marron, J., and Sheather, S. (1996), "A Brief Survey of Bandwidth Selection for Density Estimation," *Journal of the American Statistical Association*, 91, 401–407.

Kou, S. C. (2001), "Extended Exponential Criterion: A New Selection Procedure for Scatterplot Smoothers," unpublished doctoral thesis Stanford University, Department of Statistics.

Li, K.-C. (1986), "Asymptotic Optimality of C_L and Generalized Cross-Validation in Ridge Regression With Application to Spline Smoothing," *Annals of Statistics*, 14, 1101–1112.

——— (1987), "Asymptotic Optimality for C_p, C_L, Cross-Validation and Generalized Cross-Validation: Discrete Index Set," *Annals of Statistics*, 15, 958–975.

Mallows, C. (1973), "Some Comments on C_p," *Technometrics*, 15, 661–675.

Rosenblatt, M. (1991), *Stochastic Curve Estimation*, NSF-CBMS Regional Conference Series in Probability and Statistics, Vol. 3, Hayward, CA: IMS.

Simonoff, J. (1996), *Smoothing Methods in Statistics*, New York: Springer-Verlag.

Stein, M. (1990), "A Comparison of Generalized Cross-Validation and Modified Maximum Likelihood for Estimating the Parameters of a Stochastic Process," *Annals of Statistics*, 18, 1139–1157.

Wahba, G. (1985), "A Comparison of GCV and GML for Choosing the Smoothing Parameter in the Generalized Spline Smoothing Problem," *Annals of Statistics*, 13, 1378–1402.

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

Wahba, G., and Wang, Y. (1995), "Behavior Near Zero of the Distribution of GCV Smoothing Parameter Estimates," *Statistics & Probability Letters*, 25, 105–111.

Wecker, W., and Ansley, C. (1983), "The Signal Extraction Approach to Nonlinear Regression and Spline Smoothing," *Journal of the American Statistical Association*, 78, 81–89.