Available at
www.ElsevierMathematics.com
POWERED BY SCIENCE DIRECT

Statistics & Probability Letters 65 (2003) 139–146

STATISTICS & PROBABILITY LETTERS

www.elsevier.com/locate/stapro

Is C_p an empirical Bayes method for smoothing parameter choice?

S.C. Kou

Department of Statistics, Science Center 6th Floor, Harvard University, Cambridge, MA 02138, USA

Received December 2002; accepted June 2003

Abstract

The C_p selection criterion is a popular method to choose the smoothing parameter in spline regression. Another widely used method is the generalized maximum likelihood (GML) derived from a normal-theory empirical Bayes framework. These two seemingly unrelated methods, have been shown in Efron (Ann. Statist. 29 (2001) 470) and Kou and Efron (J. Amer. Statist. Assoc. 97 (2002) 766) to be actually closely connected. Because of this close relationship, the current paper studies whether C_p could also have an empirical Bayes interpretation for smoothing splines as GML does. It is shown that this is not possible. In addition, necessary conditions for a selection criterion to have an empirical Bayes interpretation are given, using which it is shown that a large class of selection criteria, including Akaike information criterion, Bayesian information criterion and Stein's unbiased risk estimate, does not possess an empirical Bayes explanation.

Keywords: Generalized maximum likelihood; AIC; BIC; SURE; Smoothing spline; MLE; Exponential family; Inverse Gaussian distribution

1. Introduction

Model selection is an important problem in statistics. This paper concerns a particular form of model selection: choosing the smoothing parameter in spline regression. The C_p selection criterion (Mallows, 1973) is a popular method to choose the smoothing parameter (see, for example, Li, 1986, 1987; Hastie and Tibshirani, 1990; Wahba, 1990). Another widely used method is the generalized maximum likelihood (GML) (Wecker and Ansley, 1983; Wahba, 1985; Stein, 1990). These two criteria, from the surface, seem quite different from each other: C_p chooses the smoothing

E-mail address: kou@stat.harvard.edu (S.C. Kou).

0167-7152/$ - see front matter
doi:10.1016/j.spl.2003.06.001

parameter by minimizing an unbiased estimate of the prediction error, while GML is motivated from an empirical Bayes framework. However it is shown in Efron (2001) and subsequently studied in Kou and Efron (2002) that both GML and C_p are actually maximum likelihood estimates with respect to two closely related curved exponential families.

With the close link between C_p and GML being delineated, a question arises naturally: since GML is an empirical Bayes estimate, is that possible that C_p also has some empirical Bayes interpretation? Such an empirical Bayes explanation, if found, may provide further understanding of the C_p criterion. For example, it is well known that although the C_p criterion asymptotically works well under the frequentist setting (see, for example, Li, 1986, 1987; Wahba, 1985; Kou, 2003), finite-sample wise it has the tendency of undersmoothing in that C_p occasionally selects a very wiggly curve even when the true underlying curve is known to be smooth (see, for example, Hurvich et al., 1998). If a Bayesian interpretation for C_p is available, then by looking at the prior distribution (of the underlying regression curve), one may be able to see directly why such a phenomenon is present for C_p—for instance, if the prior puts a lot of weights on wiggly curves, it would be the case. Furthermore, obtaining a Bayesian interpretation also offers the potential to improve the C_p criterion—one might be able to modify or remedy the prior distribution so as to obtain a selection criterion that has stable performance both asymptotically and finite-sample wise.

The current paper investigates this possibility and shows that such an empirical Bayes explanation, unfortunately, is not possible, mainly due to the singularity of the C_p density (a function introduced in Section 2) at zero. In addition, we give necessary conditions for any selection criterion to have an empirical Bayes interpretation, under both Gaussian and non-Gaussian noise. Employing these necessary conditions, we show that a large class of selection criteria, which includes Akaike information criterion (AIC), Bayesian information criterion (BIC) and Stein's unbiased risk estimate (SURE), does not possess empirical Bayes explanation.

The paper is organized as follows. Section 2, after reviewing spline regression and the C_p and GML selection criteria, presents the main result, proving the impossibility of C_p's having an empirical Bayes interpretation, as well as giving necessary conditions for a selection criterion to have empirical Bayes explanation. Section 3 extends the result to non-Gaussian case. All the proofs are deferred to the appendix.

2. Main results

2.1. Spline regression and the C_p and GML selection criteria

Suppose we have paired observations, $\{(x_i, y_i),\ i = 1, 2, \ldots, n\}$ and want to estimate $f(x) = E(y|x)$, the regression function of y on x. A linear smoother (Buja et al., 1989) estimates $\mathbf{f} = (f(x_1), f(x_2), \ldots, f(x_n))'$, the value of $f(x)$ at the design points, by $\hat{\mathbf{f}}_\lambda = A_\lambda \mathbf{y}$, where 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. One class of linear smoothers that will be of particular interest in this paper is *spline regression*, in which case the class of smoothing matrices $\{A_\lambda,\ 0 < \lambda < \infty\}$ has the form

$$A_\lambda = U \mathbf{a}_\lambda U' \tag{2.1}$$

with U 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

$$a_{\lambda i} = 1/(1 + \lambda k_i), \quad i = 1, 2, \ldots, n, \tag{2.2}$$

the constants $\mathbf{k} = (k_1, k_2, \ldots, k_n)$, solely determined by $\mathbf{x}$, being nonnegative and nondecreasing. (The cubic smoothing splines $\hat{f}_\lambda = \arg\min_f \left\{ \sum_{i=1}^n [y_i - f(x_i)]^2 + \lambda \int f''(t)^2 \, dt \right\}$ are a special case of (2.1) and (2.2); see Green and Silverman, 1994, Chapter 2.)

To use spline regression in practice, one typically has to infer the value of the smoothing parameter λ from the data. The C_p *criterion* 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. 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 $E\|\hat{\mathbf{f}}_\lambda - \mathbf{f}\|^2$, the squared prediction error. The notation $C_\lambda(\mathbf{y})$ assumes that $\mathbf{x}$ is fixed (as usual in regression problems), and that σ^2 is known. GML, the *Generalized Maximum Likelihood criterion* (Wecker and Ansley, 1983), has a normal-theory empirical Bayes motivation. If one strengthens the likelihood to $\mathbf{y} \sim \mathrm{N}(\mathbf{f}, \sigma^2 I)$, and puts a Gaussian prior on the underlying curve: $\mathbf{f} \sim \mathrm{N}(\mathbf{0}, \sigma^2 A_\lambda (I - A_\lambda)^{-1})$, then according to Bayes theorem,

$$\mathbf{y} \sim \mathrm{N}(\mathbf{0}, \sigma^2 (I - A_\lambda)^{-1}), \quad \mathbf{f}|\mathbf{y} \sim \mathrm{N}(A_\lambda \mathbf{y}, A_\lambda). \tag{2.3}$$

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 choice for the smoothing parameter: $\hat{\lambda}^{\mathrm{GML}}$ is the maximum likelihood estimate of λ based on $\mathbf{y} \sim \mathrm{N}(\mathbf{0}, \sigma^2 (I - A_\lambda)^{-1})$.

The setting of spline regression (2.1) 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

$$\mathbf{z} = U'\mathbf{y}/\sigma, \quad \mathbf{g} = U'\mathbf{f}/\sigma, \quad \hat{\mathbf{g}}_\lambda = U'\hat{\mathbf{f}}_\lambda/\sigma \tag{2.4}$$

putting the smoother family $\hat{\mathbf{f}}_\lambda = A_\lambda \mathbf{y}$ into diagonal form: $\mathbf{z} \sim (\mathbf{g}, I)$, $\hat{\mathbf{g}}_\lambda = \mathbf{a}_\lambda \mathbf{z}$. Let $b_{\lambda i} = 1 - a_{\lambda i}$, $\mathbf{b}_\lambda = (b_{\lambda 1}, b_{\lambda 2}, \ldots, b_{\lambda n})$. In the new coordinate system, the C_p statistic can be expressed as a function of $\mathbf{z}^2$

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

By defining $\mathbf{w} = \mathbf{z}^2 = (z_1^2, z_2^2, \ldots, z_n^2)'$, the C_p choice of λ is

$$\hat{\lambda}^{C_p} = \arg\min_\lambda \sum_i (b_{\lambda i}^2 w_i - 2b_{\lambda i}). \tag{2.5}$$

Under the coordinate system of $\mathbf{z}$ and $\mathbf{g}$, the GML selection criterion also has a simple form, since (2.3) becomes

$$\mathbf{z} \sim \mathrm{N}(\mathbf{0}, \mathrm{diag}(\mathbf{b}_\lambda^{-1})), \quad \mathbf{g}|\mathbf{z} \sim \mathrm{N}(\mathbf{a}_\lambda \mathbf{z}, \mathbf{a}_\lambda), \tag{2.6}$$

which gives

$$\begin{aligned} \hat{\lambda}^{\mathrm{GML}} &= \mathrm{MLE \ of} \ \mathbf{z} \sim \mathrm{N}(\mathbf{0}, \mathrm{diag}(\mathbf{b}_\lambda^{-1})) = \mathrm{MLE \ of} \ \mathbf{w} \sim \chi_1^2/\mathbf{b}_\lambda^2 \\ &= \arg\min_\lambda \sum_i (b_{\lambda i} w_i - \log b_{\lambda i}), \end{aligned} \tag{2.7}$$

since the density of $\mathbf{w} = \mathbf{z}^2 \sim \chi_1^2/\mathbf{b}_\lambda^2$ is

$$d_\lambda(\mathbf{w}) = \exp\left(-\frac{1}{2}\sum_i (b_{\lambda i} w_i - \log b_{\lambda i})\right) \Big/ \prod_i \sqrt{2\pi w_i}. \tag{2.8}$$

Because of the efficacy of using $\mathbf{z}$ and $\mathbf{g}$ in obtaining simpler expressions, we will be working on them instead of $\mathbf{y}$ and $\mathbf{f}$ whenever possible.

Comparing (2.5) with (2.7) gives an interesting observation: despite the different motivations of C_p and GML, they have similar forms. In addition, if one replaces the GML marginal density (2.8) by a density having the form

$$d_\lambda^{C_p}(\mathbf{w}) = \exp\left(-C\sum_i (b_{\lambda i}^2 w_i - 2b_{\lambda i})\right) \prod_i h(w_i), \tag{2.9}$$

where $h(\cdot)$ is a function not depending on λ, then the MLE of (2.9) leads to the C_p criterion (2.5). Density (2.9), interestingly, forms an exponential family just as (2.8) does, which means that it can be written as $d_\lambda^{C_p}(\mathbf{w}) = \exp(\eta_\lambda' \mathbf{w} - \psi_\lambda) h(\mathbf{w})$, where $\eta_\lambda = -C\mathbf{b}_\lambda^2$ is the natural parameter vector and $\psi_\lambda = -2C\sum_i b_{\lambda i}$ is the cumulant generating function, which, furthermore, implies that $h(\cdot)$ is inverse Gaussian:

$$h(w) = \left(\frac{2C}{w^3}\right)^{1/2} \varphi((2C/w)^{1/2}), \tag{2.10}$$

due to the one-to-one correspondence between a density and its cumulant generating function. ($\varphi(\cdot)$ in (2.10) is the standard normal density.)

2.2. C_p and empirical Bayes

The Bayesian framework (2.3), or equivalently (2.6), provides the empirical Bayes motivation of GML. The similarity between (2.5) and (2.7)–(2.9) naturally raises one question: Can C_p also be interpreted from an empirical Bayes point of view? Such an interpretation, if found, will further our understanding of C_p in that the prior distribution (of the underlying curve) not only directly points out C_p's strength and weakness, but also offers the potential to improve it.

However we will show that this is not possible. In other words, there does not exist a prior distribution $\pi(\cdot)$ on the curve $\mathbf{g}$ such that the Bayesian structure

$$\mathbf{g} \sim \pi(\mathbf{g}), \quad \mathbf{z}|\mathbf{g} \sim \mathrm{N}(\mathbf{g}, I)$$

would give $\mathbf{w} = \mathbf{z}^2$ the marginal distribution $\mathbf{w} = \mathbf{z}^2 \sim d_\lambda^{C_p}(\mathbf{w})$, where $d_\lambda^{C_p}(\mathbf{w})$ is given by (2.9) and (2.10). For convenience we will call $d_\lambda^{C_p}(\mathbf{w})$ the C_p density.

To show the nonexistence of the prior, we first note that the independence of w_i in $d_\lambda^{C_p}(\mathbf{w})$ and the independence structure in the likelihood $\mathbf{z}|\mathbf{g} \sim \mathrm{N}(\mathbf{g}, I)$ make it sufficient to consider only the one-dimensional case—one only needs to show that no density function $\pi(\cdot)$ fulfills these two requirements:

(i) $g \sim \pi(g)$, $z|g \sim \mathrm{N}(g, 1)$ and
(ii) marginally $w = z^2 \sim d_\lambda^{C_p}(w) = \mathrm{e}^{-C(b^2 w - 2b)}(2C/w^3)^{1/2}\varphi((2C/w)^{1/2})$.

Theorem 2.1. *A proper prior $\pi(\cdot)$ on g that satisfies both* (i) *and* (ii) *does not exist.*

The proof of the theorem, shown by contradiction, is deferred to the appendix. The basic idea is that the C_p density has a singularity at zero ($d_\lambda^{C_p}(w) \to 0$, as $w \to 0$), making it impossible to be the marginal density from any prior distribution on the curve $\mathbf{g}$.

At this point, with the hope of C_p's having an empirical Bayes interpretation being rejected, one might wonder: What kind of criterion, then, could have such an explanation? The following theorem generalizes the result of Theorem 2.1, supplying a simple way to check whether a given distribution could be the marginal distribution of z from some prior.

Theorem 2.2. *In order for a density function $p(z)$ to be the marginal density of z from a proper prior with likelihood $z|g \sim \mathrm{N}(g,1)$, it must satisfy*

(a) $\lim_{z\to 0} p(z) > 0$; *and*
(b) $p(z)$ *is infinitely differentiable at* $z = 0$.

Proof. See the appendix. □

AIC, BIC and SURE, besides C_p, are three widely used selection criteria. Since C_p cannot be interpreted from an empirical Bayes angle, it is interesting to ask if AIC, BIC or SURE can be viewed as an empirical Bayes method. In the context of linear smoothers, it can be shown (Efron, 1986) that C_p is identical to AIC and SURE. The BIC chooses the smoothing parameter λ according to

$$\hat{\lambda}^{\mathrm{BIC}} = \arg\min_\lambda \{\|\mathbf{y} - \hat{\mathbf{f}}_\lambda\|^2 + \sigma^2(\log n)\,\mathrm{tr}(A_\lambda)\},$$

where n is the sample size. To incorporate both AIC (thus C_p and SURE) and BIC in a unified framework, we consider a class of selection criteria

$$\hat{\lambda}^{(D)} = \arg\min_\lambda \{\|\mathbf{y} - \hat{\mathbf{f}}_\lambda\|^2 + \sigma^2 D\,\mathrm{tr}(A_\lambda)\}. \tag{2.11}$$

Taking the constant $D = 2$ in (2.11) gives AIC (C_p and SURE), whereas taking $D = \log n$ results in BIC.

Under the coordinate system of $\mathbf{z}$ and $\mathbf{g}$, (2.11) is equivalent to $\hat{\lambda}^{(D)} = \arg\min_\lambda \sum_i (b_{\lambda i}^2 w_i - D b_{\lambda i})$, which, similar to the case of C_p, gives the corresponding density function

$$d_\lambda^{(D)}(\mathbf{w}) = \exp\left(-C\sum_i (b_{\lambda i}^2 w_i - D b_{\lambda i})\right)\prod_i h^{(D)}(w_i). \tag{2.12}$$

Like the C_p density, (2.12) is an exponential family, whose cumulant generating function determines $h^{(D)}$ to be inverse Gaussian:

$$h(w) = \left(\frac{D^2 C}{2w^3}\right)^{1/2} \varphi\left(\sqrt{\frac{D^2 C}{2w}}\right). \tag{2.13}$$

Combining (2.12) with (2.13), we note that $\lim_{\mathbf{w}\to 0} d_\lambda^{(D)}(\mathbf{w}) = 0$. Applying Theorem 2.2, we conclude that entire class (2.11), which includes the popular C_p, AIC, BIC and SURE, cannot be interpreted from an empirical Bayes point of view. This, in certain sense, indicates that the gap between C_p and empirical Bayes is not small.

3. Extension to non-Gaussian case

We have been working on the normal likelihood $\mathbf{z}|\mathbf{g} \sim \mathrm{N}(\mathbf{g}, I)$ in the previous section. This section extends our investigation to study whether the results hold if we change the normal assumption to

$$\mathbf{z}|\mathbf{g} = \mathbf{g} + \varepsilon,$$

where $\varepsilon = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)$ and the ε_i are i.i.d. according to some distribution that has zero mean. It turns out that the previous results are qualitatively correct—C_p still cannot have an empirical Bayes explanation even under non-Gaussian case.

Let $f(\cdot)$ denote the density function of ε_i. Again it suffices to consider only the one-dimensional case. The following theorem extends the C_p result of Theorem 2.1 to non-Gaussian situation.

Theorem 3.1. *Suppose the density function $f(\cdot)$ is bounded from above (e.g. normal density, t density, gamma density, etc.). Then there does not exist a proper prior $\pi(\cdot)$ on g that satisfies*

(i) $g \sim \pi(g)$ *and* $z|g = g + \varepsilon$, *with* $\varepsilon \sim f(\varepsilon)$;
(ii) *marginally* $w = z^2 \sim d_\lambda^{C_p}(w) = \mathrm{e}^{-C(b^2 w - 2b)}(2C/w^3)^{1/2}\varphi((2C/w)^{1/2})$.

The proof is deferred to the appendix, which still hinges on the singularity of the C_p density at zero. Complementing Theorem 2.1, we give the necessary condition for a selection criterion to have empirical Bayes interpretation for the non-Gaussian case.

Theorem 3.2. *Suppose $z|g = g + \varepsilon$, where ε has bounded density function $f(\cdot)$. Then in order for a given distribution to be the marginal distribution of z from some prior, its density function $p(z)$ must satisfy* $\lim_{z\to 0} p(z) > 0$.

As we have seen for the selection criteria (2.11) $\lim_{\mathbf{w}\to 0} d_\lambda^{(D)}(\mathbf{w}) = 0$, it follows from Theorem 3.2 that they, including AIC, BIC and SURE, do not have empirical Bayes interpretation even under non-Gaussian noise.

Acknowledgements

The author is grateful to Professor Brad Efron and Professor Jun Liu for helpful discussions. The author also thanks the editor and the referee for very constructive suggestions.

Appendix: proofs

Proof of Theorem 2.1. We prove the theorem by contradiction.

Suppose there does exist a $\pi(\cdot)$ for which both (i) and (ii) are true. Then requirement (i) informs us that w given g has a noncentral χ^2 distribution with 1 degree of freedom and noncentrality parameter g^2. So w given g has density function $f(w|g) = (2\sqrt{w})^{-1}(\varphi(\sqrt{w} - g) + \varphi(-\sqrt{w} - g))$,

where as before $\varphi(\cdot)$ is the standard normal density. Since $\int_{-\infty}^{\infty}\pi(g)f(w|g)\,dg = d_{\lambda}^{C_p}(w)$, simple algebra after rearranging its terms gives

$$\int_{-\infty}^{\infty}(e^{\sqrt{w}g}+e^{-\sqrt{w}g})e^{-1/2g^2}\pi(g)\,dg = \sqrt{8C}\,\frac{1}{w}\exp\left(\left(\frac{1}{2}-Cb^2\right)w+2Cb-\frac{C}{w}\right). \tag{A.1}$$

Dominated convergence theorem says the left-hand side of (A.1) satisfies

$$\lim_{w\to 0+}\int_{-\infty}^{\infty}(e^{\sqrt{w}g}+e^{-\sqrt{w}g})e^{-1/2g^2}\pi(g)\,dg = 2\int_{-\infty}^{\infty}e^{-1/2g^2}\pi(g)\,dg > 0, \tag{A.2}$$

where the last inequality is a direct consequence of the fact $\int_{-\infty}^{\infty}\pi(g)\,dg=1$. However letting $w\to 0$ on the right-hand side of (A.1) yields

$$\lim_{w\to 0+}\sqrt{8C}\,\frac{1}{w}\exp\left(\left(\frac{1}{2}-Cb^2\right)w+2Cb-\frac{C}{w}\right)=0. \tag{A.3}$$

The proof is completed by noting the clear conflict between (A.2) and (A.3). □

Proof of Theorem 2.2. Let $\pi(g)$ denote the prior for g. Then we must have

$$\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\exp\left(-\frac{1}{2}(z-g)^2\right)\pi(g)\,dg = p(z).$$

Rearranging the terms yields

$$\int_{-\infty}^{\infty}\exp(zg)\left[\exp\left(-\frac{1}{2}g^2\right)\pi(g)\right]dg = \sqrt{2\pi}\exp\left(\frac{1}{2}z^2\right)p(z). \tag{A.4}$$

It is easy to see that

$$0 < c = \int_{-\infty}^{\infty}e^{-1/2g^2}\pi(g)\,dg < \infty.$$

Therefore the left-hand side of (A.4) is the moment generating function of the distribution $(1/c)e^{-1/2g^2}\pi(g)$ up to a normalizing constant. (a) and (b) are now immediate consequences of the basic properties of moment generating functions. □

Proof of Theorem 3.1. Suppose there does exist such a prior $\pi(\cdot)$. Then from (i), w given g has density $f(w|g)=(2\sqrt{w})^{-1}(f(\sqrt{w}-g)+f(-\sqrt{w}-g))$. So we must have $\int_{-\infty}^{\infty}\pi(g)f(w|g)\,dg=d_{\lambda}^{C_p}(w)$, which is equivalent to

$$\int_{-\infty}^{\infty}\pi(g)[f(\sqrt{w}-g)+f(-\sqrt{w}-g)]\,dg = \left(\frac{4C}{\pi}\right)^{1/2}\frac{1}{w}\exp\left(-\frac{Cb^2}{w}\left(w-\frac{1}{b}\right)^2\right).$$

Letting $w\to 0$, the limit of the left-hand side is positive, while the right-hand side goes to 0. This contradiction concludes the proof. □

The proof of Theorem 3.2 is almost identical to that of Theorem 3.1, and is hence omitted.

References

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

Efron, B., 1986. How biased is the apparent error rate of a prediction rule? J. Amer. Statist. Assoc. 81, 461–470.

Efron, B., 2001. Selection criteria for scatterplot smoothers. Ann. Statist. 29, 470–504.

Green, P., Silverman, B., 1994. Nonparametric Regression and Generalized Linear Models. Chapman & Hall, London.

Hastie, T., Tibshirani, R., 1990. Generalized Additive Models. Chapman & Hall, London.

Hurvich, C., Simonoff, J., Tsai, C., 1998. Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. J. Roy. Statist. Soc. C 60, 271–294.

Kou, S.C., 2003. On the efficiency of selection criteria in spline regression. Probab. Theory Related Fields 127, 153–176.

Kou, S.C., Efron, B., 2002. Smoothers and the C_p, GML and EE criteria: a geometric approach. J. Amer. Statist. Assoc. 97, 766–782.

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

Li, K.-C., 1987. Asymptotic optimality for C_p, C_L, cross-validation and generalized cross-validation: discrete index set. Ann. Statist. 15, 958–975.

Mallows, C., 1973. Some comments on C_p. Technometrics 15, 661–675.

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

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

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

Wecker, W., Ansley, C., 1983. The signal extraction approach to nonlinear regression and spline smoothing. J. Amer. Statist. Assoc. 78, 81–89.