Journal of Statistical Planning and Inference 141 (2011) 445–452

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

An improved C_p criterion for spline smoothing

Chun-Shu Chen [a,*], Hsin-Cheng Huang [b,c]

[a] Institute of Statistics and Information Science, National Changhua University of Education, Changhua, Taiwan
[b] Institute of Statistical Science, Academia Sinica, Taipei, Taiwan
[c] Institute of Statistics, National Chiao Tung University, Hsinchu, Taiwan

ARTICLE INFO

Article history:
Received 29 January 2010
Received in revised form
10 June 2010
Accepted 16 June 2010
Available online 22 June 2010

Keywords:
Effective degrees of freedom
Selection effect
Smoothing spline
Stein's unbiased risk estimate

ABSTRACT

Spline smoothing is a popular technique for curve fitting, in which selection of the smoothing parameter is crucial. Many methods such as Mallows' C_p, generalized maximum likelihood (GML), and the extended exponential (EE) criterion have been proposed to select this parameter. Although C_p is shown to be asymptotically optimal, it is usually outperformed by other selection criteria for small to moderate sample sizes due to its high variability. On the other hand, GML and EE are more stable than C_p, but they do not possess the same asymptotic optimality as C_p. Instead of selecting this smoothing parameter directly using C_p, we propose to select among a small class of selection criteria based on Stein's unbiased risk estimate (SURE). Due to the selection effect, the spline estimate obtained from a criterion in this class is nonlinear. Thus, the effective degrees of freedom in SURE contains an adjustment term in addition to the trace of the smoothing matrix, which cannot be ignored in small to moderate sample sizes. The resulting criterion, which we call adaptive C_p, is shown to have an analytic expression, and hence can be efficiently computed. Moreover, adaptive C_p is not only demonstrated to be superior and more stable than commonly used selection criteria in a simulation study, but also shown to possess the same asymptotic optimality as C_p.

1. Introduction

Consider noisy data $\boldsymbol{Y} \equiv (Y_1,\ldots,Y_n)'$ observed at design points $t_1,\ldots,t_n$ according to the following nonparametric regression model:

$$Y_i = f(t_i)+\varepsilon(t_i); \quad i=1,\ldots,n, \tag{1}$$

where $f(\cdot)$ is an unknown function of interest and $\boldsymbol{\varepsilon} \equiv (\varepsilon(t_1),\ldots,\varepsilon(t_n))' \sim \mathrm{N}(\mathbf{0},\sigma^2\boldsymbol{I})$ are errors.

Many approaches such as kernel smoothing (Wand and Jones, 1995), local polynomials (Cleveland, 1979), wavelets (Donoho and Johnstone, 1994), regression splines (Wand, 2000), and smoothing splines (Craven and Wahba, 1979) have been proposed for estimating $f(\cdot)$. Some asymptotic properties of these approaches can also be found in Stone (1977, 1980, 1982), Wahba (1990), and Eubank (1999). In this article, we concentrate on the smoothing spline approach, where the estimate of $f(\cdot)$ is obtained by minimizing the penalized least squares criterion:

$$\sum_{i=1}^{n}(y_i-g(t_i))^2+\lambda\int_0^1 (g''(t))^2\,dt, \tag{2}$$

* Corresponding author. Tel.: +886 4 7232105x3236.
E-mail address: cschen@cc.ncue.edu.tw (C.-S. Chen).

0378-3758/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.jspi.2010.06.017

over $\{g(\cdot) : g$ is twice differentiable in [0,1] with $\int_0^1 (g''(t))^2\,dt < \infty\}$, where $\lambda \geq 0$ is a smoothing parameter. The first term in (2) corresponds to the lack of fit, and the second term in (2) penalizes the roughness of the curve with a smaller (larger) λ corresponding to a wigglier (smoother) estimate of $f(\cdot)$. For a given λ, the smoothing spline estimate is linear and can be expressed as

$$\hat{\boldsymbol{f}}_\lambda \equiv (\hat{f}_\lambda(t_1), \ldots, \hat{f}_\lambda(t_n))' = \boldsymbol{B}_\lambda \boldsymbol{Y}, \tag{3}$$

where $\boldsymbol{B}_\lambda$ is an $n \times n$ matrix depending only on λ and $\{t_1,\ldots,t_n\}$. Furthermore, $\boldsymbol{B}_\lambda$ has an eigen-decomposition of the form

$$\boldsymbol{B}_\lambda = \boldsymbol{U}\boldsymbol{G}_\lambda \boldsymbol{U}', \tag{4}$$

where $\boldsymbol{U}$ does not depend on λ, $\boldsymbol{G}_\lambda$ is a diagonal matrix with the ith diagonal element $G_i(\lambda) \equiv (1+\lambda k_i)^{-1}$, and $0 \leq k_1 \leq k_2 \leq \cdots \leq k_n$ are the eigenvalues of $\boldsymbol{W}\boldsymbol{M}^{-1}\boldsymbol{W}'$. Here $\boldsymbol{W}$ is an $n \times (n-2)$ band matrix and $\boldsymbol{M}$ is an $(n-2) \times (n-2)$ symmetric band matrix with the (i,j)th entry given respectively by

$$w_{ij} = \begin{cases} d_j^{-1} & \text{if } i=j, \\ -d_j^{-1} - d_{j+1}^{-1} & \text{if } i=j+1, \\ d_{j+1}^{-1} & \text{if } i=j+2, \\ 0 & \text{otherwise,} \end{cases}$$

and

$$m_{ij} = \begin{cases} (d_j + d_{j+1})/3 & \text{if } i=j, \\ d_j/6 & \text{if } i=j-1 \text{ or } i=j+1, \\ 0 & \text{otherwise,} \end{cases}$$

where $d_i = t_{i+1} - t_i > 0$; $i=1,\ldots,n-1$. Note that k_i can also be well approximated by $\pi^4(i-1.5)^4/n$ for $i \geq 3$ (Silverman, 1984). The readers are referred to Silverman (1984), Culpin (1986), and Green and Silverman (1994) for more details on computing $\{k_i\}$.

Clearly, it is essential to select an appropriate λ in (2). Many selection procedures have been proposed, such as C_p (Mallows, 1973), generalized cross validation (GCV) (Craven and Wahba, 1979), generalized maximum likelihood (GML) (Wecker and Ansley, 1983; Wahba, 1985), and the extended exponential (EE) criterion (Kou and Efron, 2002). Although C_p and GCV tend to perform similarly (Efron, 2001) and have also been shown to be asymptotically optimal under the squared error loss (Li, 1986), they are known to be unstable for small to moderate sample sizes (e.g., Kohn et al., 1991; Hurvich et al., 1998; Kou and Efron, 2002; Kou, 2003, 2004). For example, Kou and Efron (2002) and Kou (2003) show by using a geometric argument that C_p may sometimes select a very oscillatory curve even when the underlying curve is smooth due to its geometric instability. In fact, it is well known that C_p tends to overfit particularly when the collection of candidate methods is large (e.g., Birgé and Massart, 2007). In contrast, GML and EE are more stable than C_p, but they do not possess the same asymptotic optimality as C_p. Some comparisons between GCV and GML can also be found in Reiss and Ogden (2009). To our knowledge, there does not exist a criterion in the literature that simultaneously achieves finite sample stability and asymptotic optimality.

Kou (2003) introduced a class of selection criteria $\{\gamma_{p,q} : p \geq 1, q \geq 1\}$ for selecting λ in (2):

$$\gamma_{p,q}(\lambda) = \begin{cases} \displaystyle\sum_{i=1}^n \left(c_q H_i(\lambda)^{p/q} Z_i^{2/q} - \frac{p}{p-1} H_i(\lambda)^{(p-1)/q} \right) & \text{if } p > 1, \\ \displaystyle\sum_{i=1}^n \left(c_q H_i(\lambda)^{1/q} Z_i^{2/q} - \frac{1}{q} \log(H_i(\lambda)) \right) & \text{if } p = 1, \end{cases} \tag{5}$$

where $H_i(\lambda) = \lambda k_i/(1+\lambda k_i)$, $\boldsymbol{Z} = (Z_1,\ldots,Z_n)' \equiv \boldsymbol{U}'\boldsymbol{Y}/\sigma$ with $\boldsymbol{U}$ defined in (4), and $c_q \equiv \sqrt{\pi}/(2^{1/q}\Gamma(1/2+1/q))$. Here σ is assumed known. When it is unknown, it can be replaced by an estimate (e.g., Efron, 2001). For a given pair of (p,q), the criterion $\gamma_{p,q}$ selects $\hat{\lambda}_{p,q}$ that minimizes $\gamma_{p,q}(\lambda)$ over $\lambda \geq 0$. As shown in Kou (2003), the class of (5) includes C_p (corresponding to $p=2$ and $q=1$), GML (corresponding to $p=q=1$), and EE (corresponding to $p=q=1.5$) as special cases.

Our goal in this paper is to develop a selection procedure that performs well in small sample sizes while retaining the large sample optimality. So instead of selecting λ directly, our idea is to select among a small subset of candidate criteria $\{\gamma_{p,q} : p \geq 1, q \geq 1\}$ based on Stein's (1981) unbiased risk estimate, from which we obtain the estimated smoothing parameter $\hat{\lambda}_{\hat{p},\hat{q}}$ according to the selected criterion. Note that the estimate of $\boldsymbol{f} = (f(t_1), \ldots, f(t_n))'$ obtained from the criterion $\gamma_{p,q}$ can be written as

$$\hat{\boldsymbol{f}}_{\hat{\lambda}_{p,q}} = (\hat{f}_{\hat{\lambda}_{p,q}}(t_1), \ldots, \hat{f}_{\hat{\lambda}_{p,q}}(t_n))' = \boldsymbol{B}_{\hat{\lambda}_{p,q}} \boldsymbol{Y}, \tag{6}$$

which is nonlinear because of the selection effect, making $\boldsymbol{B}_{\hat{\lambda}_{p,q}}$ depend on $\boldsymbol{Y}$ through $\hat{\lambda}_{p,q}$. Consequently, the commonly used effective degrees of freedom $\mathrm{tr}(\boldsymbol{B}_{\hat{\lambda}_{p,q}})$ (e.g., Buja et al., 1989) which fails to incorporate selection uncertainty, is not appropriate for criterion $\gamma_{p,q}$ and is subject to adjustment.