Ann. Inst. Statist. Math.
Vol. 51, No. 4, 679–689 (1999)

ON ADAPTIVE COMBINATION OF REGRESSION ESTIMATORS*

HELGE BLAKER**

Department of Statistics, University of California, Evans Hall 3860, Berkeley, CA 94720-3860, U.S.A.

(Received March 21, 1997; revised January 12, 1998)

Abstract. Consider the problem of choosing between two estimators of the regression function, where one estimator is based on stronger assumptions than the other and thus the rates of convergence are different. We propose a linear combination of the estimators where the weights are estimated by Mallows' C_L. The adaptive estimator retains the optimal rates of convergence and is an extension of Stein-type estimators considered by Li and Hwang (1984, *Ann. Statist.*, **12**, 887–897) and related to an estimator in Burman and Chaudhuri (1999, *Ann. Inst. Statist. Math.* (to appear)).

Key words and phrases: Nested models, nonparametric regression, rates of convergence, adaptive estimator, Mallows' C_L, Stein estimation.

1. Introduction and setup

We consider the problem of estimating the regression function $f(x)$ in the model

$$y_i = f(x_i) + \sigma\varepsilon_i,$$

$i = 1, \ldots, n$, where ε_i is a sequence of independent and identically distributed variables with mean 0 and variance 1, and $x_i \in \mathcal{K}$, a compact set in R^d. Write $\mu_i = f(x_i)$, let $\|\cdot\|$ be the Euclidean norm and let $\|\cdot\|_n^2 = n^{-1}\|\cdot\|^2$, and correspondingly for $\langle\cdot,\cdot\rangle$ and $\langle\cdot,\cdot\rangle_n$. Assume we have available two different linear estimators, based on different assumptions on the regression function, i.e. the parameter space and with different rates of convergence. We are unsure about which estimator to use and hence would like to use the data to aid the selection in an asymptotically consistent manner. Almost all commonly used nonparametric estimators are linear in the data, see e.g. Buja *et al.* (1989) or Kneip (1994). Assume $\hat{\mu}_1 = M_1 y$ is an estimator with optimal rate of convergence for $\mu \in \Theta_1$,

$$\sup_{\Theta_1} n^{-1} E\|\mu - M_1 y\|^2 \sim r_1(n). \tag{1.1}$$

* This research was supported by grant 411.92/001 from the Research Council of Norway and by grant DMS 92-24868 from the National Science Foundation.

** Now at CSIRO Mathematical and Information Sciences, Locked Bag 17, North Ryde, NSW 2113, Australia.

If we are willing to make the more restrictive assumption $\mu \in \Theta_2 \subset \Theta_1$, then a rate-optimal estimator is $\hat{\mu}_2 = M_2 y$ where

$$\sup_{\Theta_2} n^{-1} E\|\mu - M_2 y\|^2 \sim r_2(n), \tag{1.2}$$

and $r_2(n)/r_1(n) \to 0$ as $n \to \infty$, which implies $n^{-1}\|\mu - M_i y\|^2 = O_p(r_i)$, $i = 1, 2$.

Example 1. Let $W_2^p = \{f : \|f^{(p)}\| < \infty\}$ (and f has cont. derivatives of orders up to $p-1$), let $\mathcal{F}_1 = W_2^k$, $\mathcal{F}_2 = W_2^l$, $k < l$. Let $\phi_1 = 1$, $\phi_j(t) = \sqrt{2}\cos 2\pi jt$ for j even and $\sqrt{2}\sin 2\pi jt$ for j odd. Assume every f has the representation $f(t) = \sum_{j=0}^{\infty} a_j \phi_j(t)$ for $t \in [0,1]$, say. Let $\|f\|^2 = \int_0^1 f(t)^2 dt$. Then $\|f^{(k)}\|^2 = \sum_{j=0}^{\infty} (2\pi j)^{2k} a_j^2$. Estimating f is equivalent to estimating the Fourier coefficients a_j, so $\Theta_1 = \{a : \sum_{j=1}^{\infty} (2\pi j)^{2k} a_j^2 < \infty\}$ while $\Theta_2 = \{a : \sum_{j=1}^{\infty} (2\pi j)^{2l} a_j^2 < \infty\}$. Clearly $\Theta_2 \subset \Theta_1$.

However, using the estimator $\hat{\mu}_2$ may result in a large bias if the assumption $\mu \in \Theta_2$ turns out to be wrong. We would therefore like a combination estimator which decides on the basis of data which estimator to use. Define the hybrid estimator

$$\hat{\mu}(\alpha) = \alpha \hat{\mu}_1 + (1 - \alpha)\hat{\mu}_2 \tag{1.3}$$

which can also be viewed as a smoothed version of a pretest-estimator where we test $\mu \in \Theta_2$ vs. $\mu \in \Theta_1/\Theta_2$ and use $\hat{\mu}_2$ ($\alpha = 0$) if we do not reject the null and $\hat{\mu}_1$ ($\alpha = 1$) if we do. The estimated value of α provides insight into the fit of the model $\mu \in \Theta_2$. Clearly one should have $0 \le \alpha \le 1$ but we will ignore this restriction in the technical part since this will hold asymptotically for our estimator $\hat{\alpha}$, see Lemma 2. In order to focus on the main issue we will assume the noise level σ is known. We need a measure of 'how wrong' the null hypothesis is. Let the true mean be μ and define

$$\delta_n = \inf\{\|\mu - \tilde{\mu}\|_n^2 : \tilde{\mu} \in \Theta_2\}$$

and suppose this is attained at the point $\mu_2 \in \Theta_2$. This holds true e.g. if Θ_2 is convex. For technical simplicity we assume the inf is attained, though this assumption can be relaxed at the cost of dealing with a sequence of approximating values.

Burman and Chaudhuri (1999) worked on 'a functional version of the famous James-Stein approach in parameter estimation', meaning a compromise between a nonparametric and a parametric estimator. The techniques used here are very close to their approach. Some advantages in our approach are that there is no need for one estimator to be 'parametric', i.e. have rate n^{-1}, as long as its rate is not faster than this. Estimators with rates faster than n^{-1} are seldom of interest in practice due to their scarcity and may in fact be viewed as pathological, see Li (1986). Moreover, we use Mallows' C_L to estimate the weight to put on each component of the estimator, which enables us to recover the Stein-estimator used by Li (1985, 1987). Burman and Chaudhuri (1999) use a variant of cross-validation

which also requires more assumptions on the original estimators. Their use of leave-one-out cross-validation renders their estimator useless when $\hat{\mu}_1 = y$, i.e. $M_1 = I$ since the leave-one-out estimator is now undefined. Historically, Stein's estimator was developed to improve upon the minimax estimator y (the raw data) for a multivariate normal distribution, and shrinkage towards the origin (or any other point) may be viewed as a crude way of borrowing strength from a lower-dimensional model. This becomes more apparent when one considers the variant shrinking towards the grand mean $\bar{y}$. In more complicated situations like curve estimation the raw data is not an acceptable estimator, even though it may still be minimax. It is desirable to improve a standard estimator by borrowing strength from an estimator appropriate under more restrictive assumptions, and to do so in a manner which does not render the estimator useless if these assumptions do not hold. Instead of (global) minimaxity, we want to retain the optimal rate of convergence. In this sense, our hybrid estimator turns out to be optimal. The rate depends on how wrong the more restrictive assumptions are, quantified by the distance from the actual parameter to the smaller parameter space. This problem was considered by Li and Hwang (1984) for the particular problem that one of the estimators under consideration is the raw data y. It is then desirable to retain global minimaxity while getting optimal rate under the more restrictive model. Our approach extends their estimator to compromises between almost any two nested models.

2. The adaptive estimator

The distance between the true mean and the estimated mean using the hybrid estimator is $\|\mu - \alpha M_1 y - (1-\alpha)M_2 y\|$, which is minimized at

$$\alpha^* = \langle M_1 y - M_2 y, \mu - M_2 y\rangle \|M_1 y - M_2 y\|^{-2}. \tag{2.1}$$

Our approach is to treat this as a pseudoparameter and estimate it by minimizing Mallows' C_L, Mallows (1973), which is an unbiased estimate of the loss. Kneip (1994) discusses choosing smoothing parameters for linear smoothers by this device. Other approaches are possible, e.g. variants of cross-validation (Burman and Chaudhuri (1999)), generalized cross-validation or maximum likelihood. We have $E\hat{L}(\alpha) = En^{-1}\|\mu - \hat{\mu}(\alpha)\|^2$ where

$$\begin{aligned}\hat{L}(\alpha) &= n^{-1}\|y - \hat{\mu}(\alpha)\|^2 + 2\sigma^2 n^{-1}\operatorname{tr}\{\alpha M_1 + (1-\alpha)M_2\} - \sigma^2 \\ &= n^{-1}\|y - M_2 y - \alpha(M_1 y - M_2 y)\|^2 + 2\sigma^2 n^{-1}\{\operatorname{tr} M_2 + \alpha \operatorname{tr}(M_1 - M_2)\} - \sigma^2.\end{aligned}$$

This is minimized by

$$\hat{\alpha} = (\sigma^2 \operatorname{tr}(M_2 - M_1) + \langle y - M_2 y, M_1 y - M_2 y\rangle)/\|M_1 y - M_2 y\|^2. \tag{2.2}$$

Comparing with (2.1), we replace $\langle M_1 y - M_2 y, \mu\rangle$ with $\langle y, M_1 y - M_2 y\rangle - \sigma^2 \operatorname{tr}(M_1 - M_2)$, and $E[\langle M_1 y - M_2 y, y - \mu\rangle - \sigma^2 \operatorname{tr}(M_1 - M_2)] = 0$. The corresponding minimum of the estimated loss is

$$\begin{aligned}\hat{L}(\hat{\alpha}) = {} & n^{-1}\|y - M_2 y\|^2 + 2\sigma^2 n^{-1} \operatorname{tr} M_2 - \sigma^2 \\ & - n^{-1}\|M_1 y - M_2 y\|^{-2}\{\sigma^2 \operatorname{tr}(M_2 - M_1) - \langle y - M_2 y, M_1 y - M_2 y\rangle\}^2\end{aligned}$$

Example 2. If $\Theta_1 = R^n$ and $\Theta_2 = \{0\}$, let $\hat{\mu}_1 = y$ and $\hat{\mu}_2 = 0$, so (2.2) becomes $\hat{\alpha} = 1 - n\sigma^2\|y\|^{-2}$ and

$$\hat{\mu}(\hat{\alpha}) = (1 - n\sigma^2/\|y\|^2)y.$$

This is minimax for $n \geq 4$ when $\varepsilon_i \sim N(0,1)$ and for large n, $\hat{\mu}(\hat{\alpha})$ almost coincides with the James-Stein estimator $\hat{\mu}_S = (1 - (n-2)\sigma^2/\|y\|^2)y$, James and Stein (1961). The same observation is made in the original article by Mallows (1973), p. 673, in the setting of linear regression with orthogonal regressors. This is a special case of the next example.

Example 3. If $\Theta_1 = R^n$, let $\hat{\mu}_1 = y$ and (2.2) becomes $\hat{\alpha} = 1 - \sigma^2 \operatorname{tr}(I - M_2)\|y - M_2y\|^{-2}$, or

$$\hat{\mu}(\hat{\alpha}) = y - \sigma^2 \operatorname{tr}(I - M_2)\|y - M_2y\|^{-2}(y - M_2y)$$

and

$$\hat{L}(\hat{\alpha}) = \sigma^2 - \sigma^4\{\operatorname{tr}(I - M_2)\}^2/\{n\|y - M_2y\|^2\}$$

which is the (simplified) Stein estimator that shrinks the raw data towards M_2y and its corresponding (simplified) risk estimator, see Li (1987), p. 967. Li and Hwang (1984) use some identities from Stein (1981) to get exact unbiased risk formulae for an estimator which is very close to $\hat{\mu}(\hat{\alpha})$. Let $A = I - M_2$, where M_2 is symmetric, and assume $2A < (\operatorname{tr} A)I$ in the sense of positive definiteness. Stein (1981) considers (in our notation)

$$\hat{\mu}_S = y - \lambda(y)Ay$$

where $\lambda(y) = \sigma^2/(y'By)$ and $B = ((\operatorname{tr} A)I - 2A)^{-1}A^2$. If $\varepsilon_i \sim N(0,1)$,

$$EL(\hat{\mu}_S, \mu) = \sigma^2 - n^{-1}\sigma^4 Ey'A^2y/(y'By)^2.$$

When n is sufficiently large, the largest eigenvalue of A will be negligible compared to the trace, so $\lambda(y) \approx \sigma^2(\operatorname{tr} A)/\|Ay\|^2$ and $\hat{\mu}_S$ and $\hat{\mu}(\hat{\alpha})$ will be close. The results of Li and Hwang (1984) also hold for $\hat{\mu}(\hat{\alpha})$ which also works for asymmetric M_2. They show that the estimator $\hat{\mu}_S$ is minimax over $\Theta_1 = R^n$ when the errors follow a Gaussian distribution while giving a consistent estimator of μ when $En^{-1}\|\mu - M_2y\|^2 \to 0$. A 'subexample' is $M_2y = \bar{y}\mathbf{1}$, which is appropriate if $\Theta_2 = \{\mu : \mu_i = c, c \in R\}$, with risk

$$\sup_{\Theta_2} n^{-1}E\|\mu - \bar{y}\mathbf{1}\|^2 = n^{-1}\sigma^2,$$

the 'parametric' rate. The resulting estimator is a variant of the 'Lindley' estimator,

$$\hat{\mu}_L = \bar{y}\mathbf{1} - \left(\sigma^2(n-1)/\sum_{i=1}^{n}(y_i - \bar{y})^2\right)(y - \bar{y}\mathbf{1}),$$

where the constant is $n - 1$ instead of $n - 3$ which is the optimal constant in the Gaussian case, see the discussion in Stein (1962). Standard calculations using

Stein's unbiased risk formula, see Stein (1981), show that in the Gaussian case, for $n \geq 4$,

$$\sup_{\Theta_2} n^{-1} E\|\mu - \hat{\mu}_L\|^2 = n^{-1}\sigma^2(3n-5)/(n-3) \sim n^{-1}3\sigma^2$$

while $\hat{\mu}_L$ is still minimax over Θ_1. The rate is still optimal (i.e. n^{-1}) for Θ_2 but the constant is inferior. In other words $\hat{\mu}_L$ is inadmissible on Θ_2. Here

$$\delta_n = \inf_{c \in R} n^{-1}\|\mu - c\mathbf{1}\|^2 = n^{-1}\|\mu - \bar{\mu}\mathbf{1}\|^2 = n^{-1}\sum_{i=1}^{n}(\mu_i - \bar{\mu})^2.$$

This quantity enters in asymptotic risk calculations, see Casella and Hwang (1982), e.g.

$$\lim_{n\to\infty} \sup_{\|\mu - \bar{\mu}\mathbf{1}\|^2 \leq nc} n^{-1} E\|\mu - \hat{\mu}_L\|^2 = \sigma^2 c/(\sigma^2 + c).$$

3. Properties of the adaptive estimator

We now study the behavior of the hybrid estimator, both with 'oracle' or ideal metaparameter α^* and estimated metaparameter $\hat{\alpha}$. Lemmas 1 and 2 are similar to Lemmas 5.1, 5.2 and 5.4 in Burman and Chaudhuri (1999) where $r_2(n) = n^{-1}$. The proofs are similar and have been omitted.

LEMMA 1. *We have* $\|M_1 y - M_2 y\|_n^2 = O_p(r_1(n) \vee \delta_n)$. *In addition, if* $\delta_n = 0$ *for large* n, $\|M_1 y - M_2 y\|_n^2$ *stays between* $c_1 r_1(n)$ *and* $c_2 r_1(n)$ *with probability getting arbitrarily close to one for some constants* $0 < c_1 < c_2$.

LEMMA 2. *Assume that* δ_n *tends to zero as* n *tends to infinity. It then holds true that*

$$\alpha^* = \begin{cases} 1 + O_p((r_1(n)/\delta_n)^{1/2}) & \text{if } \delta_n > r_1(n) \\ O_p((\delta_n/r_1(n))^{1/2}) & \text{if } r_2(n) \leq \delta_n \leq r_1(n) \\ O_p((r_2(n)/r_1(n))^{1/2}) & \text{if } \delta_n < r_2(n). \end{cases}$$

First we look at the rate of convergence for 'oracle' choice of metaparameter.

THEOREM 1. *Assume* δ_n *tends to zero as* n *tends to infinity. Then*

$$\|\mu - \hat{\mu}(\alpha^*)\|_n^2 = \begin{cases} O_p(r_1(n)) & \text{if } \delta_n > r_1(n) \\ O_p(\delta_n) & \text{if } r_2(n) \leq \delta_n \leq r_1(n) \\ O_p(r_2(n)) & \text{if } \delta_n < r_2(n). \end{cases}$$

PROOF.

$$\begin{aligned}\|\mu - \hat{\mu}(\alpha^*)\|_n &= \|\alpha^*(\mu - M_1 y) - (1-\alpha^*)(M_2 y - \mu)\|_n \\ &\le |\alpha^*| \cdot \|\mu - M_1 y\|_n + |1-\alpha^*| \cdot \|\mu - M_2 y\|_n \\ &\le |\alpha^*| O_p(r_1(n)^{1/2}) + |1-\alpha^*|\{O_p(r_2(n)^{1/2}) + \delta_n^{1/2}\}\end{aligned}$$

The theorem now follows from Lemma 2. □

By the projection property of $\hat{\mu}(\alpha^*)$, we have

$$\|\hat{\mu}(\hat{\alpha}) - \mu\|_n^2 - \|\hat{\mu}(\alpha^*) - \mu\|_n^2 = \|\hat{\mu}(\hat{\alpha}) - \hat{\mu}(\alpha^*)\|_n^2 = (\hat{\alpha} - \alpha^*)^2 \|\hat{\mu}_1 - \hat{\mu}_2\|_n^2$$

To prove the equivalent of Theorem 1 for $\|\hat{\mu}(\hat{\alpha}) - \mu\|_n^2$, it suffices to show that $(\hat{\alpha} - \alpha^*)^2 \|M_1 y - M_2 y\|_n^2$ tends to zero at least as fast as $\|\hat{\mu}(\alpha^*) - \mu\|_n^2$. It is useful to notice

LEMMA 3. *If $a_n > 0$ and $EX_n^2 = O(a_n^2)$, then $X_n = O_p(a_n)$.*

PROOF. By Chebychev's inequality, for any $\varepsilon > 0$, let $C = \sup_n(a_n^{-2} EX_n^2) < \infty$ and $M = (C/\varepsilon)^{1/2}$. Then $P(a_n^{-1}|X_n| > M) \le a_n^{-2} EX_n^2 / M^2 \le C/M^2 = \varepsilon$ hence $a_n^{-1} X_n = O_p(1)$. □

THEOREM 2. *Assume $E\varepsilon^4 < \infty$ and δ_n tends to zero as n tends to infinity. Assume the largest eigenvalue of $M_2' M_2$, $\lambda_{\max}(M_2' M_2)$, is bounded for all n. If $r_2(n)$ is not faster than n^{-1}, i.e. $\liminf_{n\to\infty} n r_2(n) > 0$, then*

$$\|\mu - \hat{\mu}(\hat{\alpha})\|_n^2 = \begin{cases} O_p(r_1(n)) & \text{if } \delta_n > r_1(n) \\ O_p(\delta_n) & \text{if } r_2(n) \le \delta_n \le r_1(n) \\ O_p(r_2(n)) & \text{if } \delta_n < r_2(n) \end{cases}$$

If $r_2(n)$ is slower than n^{-1}, i.e. $\liminf_{n\to\infty} n r_2(n) = \infty$, then

$$\|\mu - \hat{\mu}(\hat{\alpha})\|_n^2 = \|\mu - \hat{\mu}(\alpha^*)\|_n^2 (1 + o_p(1)).$$

PROOF. In this proof let C denote a generic positive constant whose precise value may change from equation to equation. We have

$$\hat{\alpha} - \alpha^* = \|M_1 y - M_2 y\|^{-2}(\langle y - \mu, M_1 y - M_2 y\rangle - \sigma^2 \operatorname{tr}(M_1 - M_2))$$

and consequently

$$(\hat{\alpha} - \alpha^*)^2 \|\hat{\mu}_1 - \hat{\mu}_2\|_n^2 = \|M_1 y - M_2 y\|_n^{-2}(\langle y - \mu, M_1 y - M_2 y\rangle_n - n^{-1}\sigma^2 \operatorname{tr}(M_1 - M_2))^2.$$

The behavior of $\|M_1 y - M_2 y\|_n^2$ is known from Lemma 1. Define

$$\begin{aligned} A_n &= \langle y - \mu, M_1 y - M_2 y\rangle_n - \sigma^2 n^{-1} \operatorname{tr}(M_1 - M_2) \\ &= \langle \varepsilon, M_1 y\rangle_n - \sigma^2 n^{-1} \operatorname{tr}(M_1) - (\langle \varepsilon, M_2 y\rangle_n - \sigma^2 n^{-1} \operatorname{tr}(M_2)) \\ &= (\langle \varepsilon, M_1 \varepsilon\rangle_n - \sigma^2 n^{-1} \operatorname{tr}(M_1)) - (\langle \varepsilon, M_2 \varepsilon\rangle_n - \sigma^2 n^{-1} \operatorname{tr}(M_2)) \\ &\quad + \langle \varepsilon, M_1 \mu - \mu\rangle_n - \langle \varepsilon, M_2 \mu - \mu_2\rangle_n + \langle \varepsilon, \mu - \mu_2\rangle_n \\ &= A_{n,1} - A_{n,2} + A_{n,3} - A_{n,4} + A_{n,5} \end{aligned}$$

say. By an application of Theorem 2 of Whittle (1960), see Li (1987), p. 970, we have that if $E\varepsilon^4 < \infty$,

$$E[\sigma^2 \operatorname{tr}(M_1) - \langle \varepsilon, M_1 \varepsilon\rangle]^2 \le C \operatorname{tr}(M_1' M_1)$$

for some constant $C > 0$. Since $\sigma^2 n^{-1} \operatorname{tr}(M_1' M_1) \le \sup_{\Theta_1} n^{-1} E\|\mu - M_1 y\|^2$, and

$$E[A_{n,1}^2] \le C n^{-1} (n^{-1} \operatorname{tr}(M_1' M_1)),$$

we find that $A_{n,1} = O_p(n^{-1/2} r_1^{1/2})$. The same argument gives that $A_{n,2} = O_p(n^{-1/2} r_2^{1/2})$. Furthermore, $EA_{n,3}^2 \le Cn^{-2}\|M_1\mu - \mu\|^2 \le Cn^{-1} r_1$ so $A_{n,3} = O_p(n^{-1/2} r_1^{1/2})$ and

$$\begin{aligned} EA_{n,4}^2 &\le Cn^{-2}\|M_2\mu - \mu_2\|^2 \le Cn^{-2}(\|M_2\mu - M_2\mu_2\|^2 + \|M_2\mu_2 - \mu_2\|^2) \\ &\le Cn^{-2}\left(\lambda_{\max}(M_2' M_2)\|\mu - \mu_2\|^2 + \sup_{\Theta_2} \|\mu - M_2\mu\|^2\right) \\ &= O(n^{-1}\delta_n) + O(n^{-1} r_2(n)) \end{aligned}$$

so $A_{n,4} = O_p(n^{-1/2}(\delta_n^{1/2} \vee r_2(n)^{1/2}))$. Finally, $A_{n,5} = O_p(n^{-1/2}\delta_n^{1/2})$ since $EA_{n,5}^2 \le Cn^{-2}\|\mu - \mu_2\|^2 = Cn^{-1}\delta_n$ and all together $A_n = O_p(n^{-1/2} r_1^{1/2}) + O_p(n^{-1/2}\delta_n^{1/2})$. Therefore,

$$A_n^2 \|M_1 y - M_2 y\|_n^{-2} = \begin{cases} O_p(n^{-1} r_1/\delta_n) + O_p(n^{-1}) & \text{when } \delta_n > r_1 \\ O_p(n^{-1}\delta_n/r_1) + O_p(n^{-1}) & \text{when } \delta_n \le r_1 \end{cases}$$

which implies $(\hat{\alpha} - \alpha^*)^2 \|M_1 y - M_2 y\|_n^2 = O_p(n^{-1})$. This concludes the proof since $\|\hat{\mu}(\alpha^*) - \mu\|_n^2$ does not go to zero faster than this by assumption. For the last statement of the theorem, recall that $A_n = O_p(a_n)$ implies $A_n = o_p(b_n)$ whenever $a_n/b_n \to 0$, $a_n, b_n \to 0$. □

4. A numerical example

In this section we report on a numerical experiment. Let

$$\Theta_1 = \{f : \|f^{(2)}\| < \infty, f \text{ continuously differentiable}\} \text{ and } \Theta_2 = \{f : f \text{ is linear}\}.$$

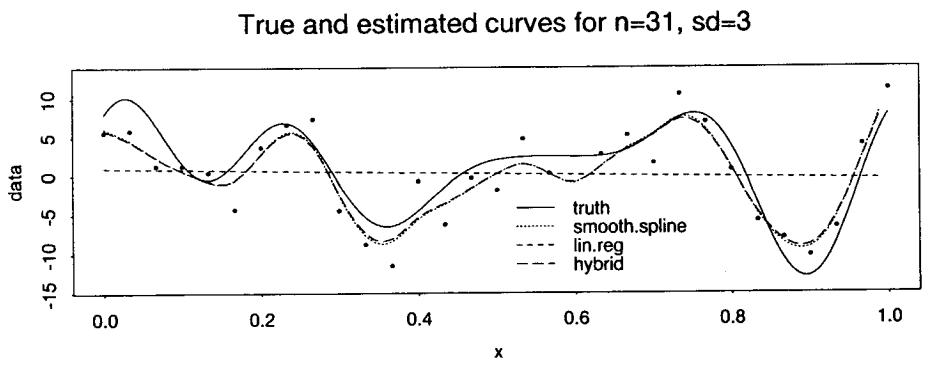

Fig. 1. Smoothing spline, linear regression and hybrid estimator when f_1 is truth.

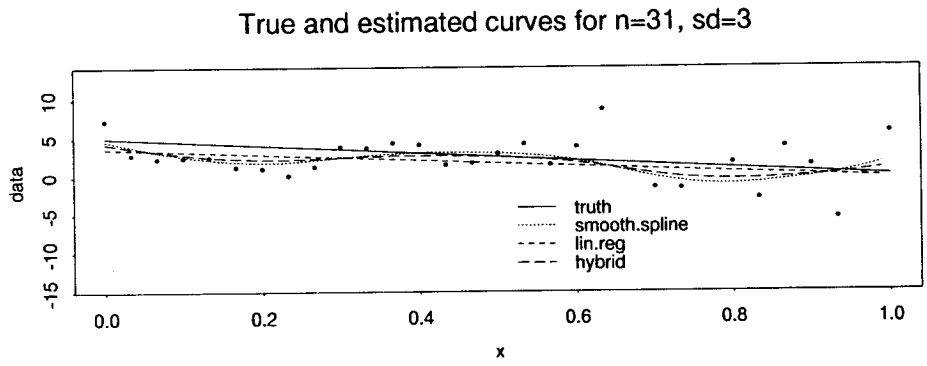

Fig. 2. Smoothing spline, linear regression and hybrid estimator when f_2 is truth.

The true regression function is taken to be either

$$f_1(x) = a_0 + 2\sum_{j=1}^{4}\{a_j \cos(2\pi jx) + b_j \sin(2\pi jx)\} \quad \text{or} \quad f_2(x) = 5 - 5x$$

where $0 \leq x \leq 1$, $a_0 = 1.0$, $a = (-0.5, 0.5, 2.5, 1.0)'$ and $b = (2.5, 1.0, 0.5, 0.5)'$. We take $\hat{\mu}_1 = M_1 y$ to be the cubic smoothing spline estimator with the smoothing parameter chosen by cross-validation, computed by the S-plus function `smooth.spline`. Strictly speaking, this data-dependent choice of smoothing parameter makes $\hat{\mu}_2$ nonlinear in y. The estimator $\hat{\mu}_2$ is the linear regression estimate of y on x. We take x_j to be equispaced on $[0, 1]$ and use Gaussian errors from the S-plus function `rnorm` with $\sigma = 1$. It is well known that $r_1(n) = n^{-4/5}$ and $r_2(n) = n^{-1}$. The three estimates were computed for $n = 21$, $n = 41$ and $n = 101$, each by 10000 replications, and the mean squared errors together with the estimated values of α were computed. Figures 1 and 2 show typical situations, the first picture for f_1 the true regression function and the second picture for f_2 the true function. The mean squared error (MSE) for the three estimators is tabulated

Table 1. Mean squared error for different estimators of regression function, f_1 true regression function, based on 10000 simulations. Standard deviation in ( ), $\sigma = 1$.

Sample size	21	41	101
Smooth. spline	23.81(0.47)	21.92(0.08)	22.20(0.06)
Lin. regr.	607.57(0.02)	1146.25(0.02)	2766.35(0.02)
Hybrid	23.47(0.48)	21.81(0.08)	22.19(0.06)
mean($\hat{\alpha}$)	0.97	0.99	1.00
sd($\hat{\alpha}$)	0.01	0.01	0.00

Table 2. Mean squared error for different estimators of regression function, f_2 true regression function, based on 10000 simulations. Standard deviation in ( ), $\sigma = 1$.

Sample size	21	41	101
Smooth. spline	3.28(0.04)	3.46(0.04)	3.42(0.04)
Lin. regr.	2.01(0.02)	2.02(0.02)	2.02(0.02)
Hybrid	2.49(0.02)	2.63(0.03)	2.72(0.03)
mean($\hat{\alpha}$)	0.27	0.30	0.33
sd($\hat{\alpha}$)	0.40	0.41	0.42

Table 3. Mean squared error for different estimators of regression function, f_3 true regression function, based on 10000 simulations. Standard deviation in ( ), $\sigma = 1$.

Sample size	21	41	101
Smooth. spline	3.86(0.04)	3.94(0.04)	3.90(0.04)
Lin. regr.	2.91(0.02)	2.90(0.02)	2.93(0.02)
Hybrid	3.07(0.02)	3.16(0.02)	3.24(0.02)
mean($\hat{\alpha}$)	0.41	0.46	0.49
sd($\hat{\alpha}$)	0.44	0.45	0.45

in Tables 1, 2 and 3. Here $\hat{\alpha}$ was truncated to be in $[0, 1]$. It appears that the hybrid estimator does a very good job when f_1 is the true regression function and an acceptable job when f_2 is the truth.

To study the intermediate case, let $f_3(x) = \gamma(n)(x^2 - x)$. It is easy to show that $\delta_n \sim \gamma(n)^2/5 = O(\gamma(n)^2)$. The intermediate case in Theorem 2 is when $n^{-1/2} < \gamma(n) < n^{-2/5}$. The behavior of the estimators in this case is similar to the case $\mu \in \Theta_2$, as can be seen from Table 1. In the simulations, $\gamma(n) = 10 \cdot n^{-9/20}$ was used so $\delta = O(n^{-9/10})$. A typical situation is Fig. 3.

As far as rates of convergence are concerned, we can safely use the adaptive hybrid estimator as long as we do not use an estimator with rate faster than n^{-1}. This is not a problem in practice. It might look as if all estimators are

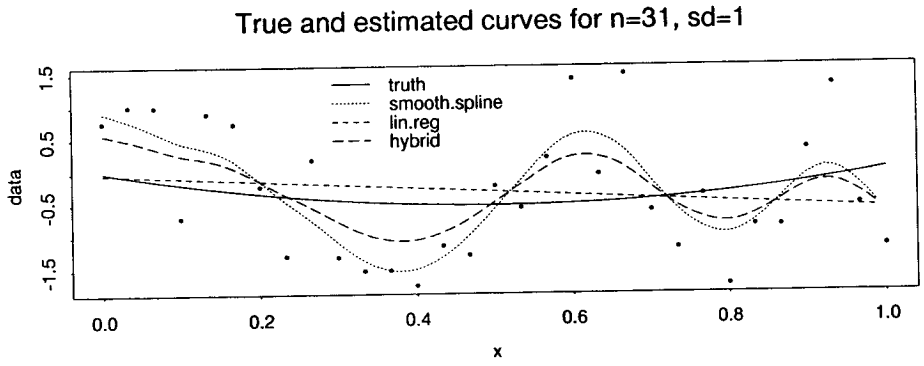

Fig. 3. Smoothing spline, linear regression and hybrid estimator when f_3 is truth.

asymptotically inadmissible, since we could use the hybrid estimator together with some other estimator, getting still better rates in some part of the parameter space etc. However, what we really would like is $\sup_{\Theta_i} n^{-1}E\|\mu - \hat{\mu}(\hat{\alpha})\|^2$ for $i = 1, 2$ in order to compare exact asymptotic risk. It is reasonable to conjecture that the supremum risk gets larger if more weight factors need to be estimated, recall the example involving the Lindley estimator. So even though the rate is still optimal, the constant in the asymptotic risk will increase. Computing the exact asymptotic risk is only possible in very special cases, see Golubev and Nussbaum (1990). It is also clear that replacing σ^2 by a consistent estimator will not invalidate the results of Theorem 2. In practice, most nonparametric regression estimators have smoothing parameters which when estimated from the data will destroy their linearity. This could also be taken into account at the cost of increased detail.

Acknowledgements

The author would like to thank Rudy Beran for many stimulating discussions and Kjell Doksum for bringing the paper by Burman and Chaudhuri to his attention.

References

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

Burman, P. and Chaudhuri, P. (1999). A hybrid approach to parametric and nonparametric regression, *Ann. Inst. Statist. Math.* (to appear).

Casella, G. and Hwang, J. T. (1982). Limit expressions for the risk of James-Stein estimators, *Canad. J. Statist.*, **10**, 305–309.

Golubev, G. K. and Nussbaum, M. (1990). A risk bound in Sobolev class regression, *Ann. Statist.*, **18**, 758–778.

James, W. and Stein, C. M. (1961). Estimating with quadratic loss, *Proc. 4th Berkeley Symp. on Math. Statist. Probab.*, Vol. 1, 361–380, University of California Press.

Kneip, A. (1994). Ordered linear smoothers, *Ann. Statist.*, **22**, 835–866.

Li, K.-C. (1985). From Stein's unbiased risk estimates to the method of generalized cross-validation, *Ann. Statist.*, **13**, 1352–1377.

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**, 958–976.

Li, K.-C. and Hwang, J. T. (1984). The data-smoothing aspect of Stein estimates, *Ann. Statist.*, **12**, 887–897.

Mallows, C. L. (1973). Some comments on C_P, *Technometrics*, **15**, 661–675.

Stein, C. M. (1962). Confidence sets for the mean of a multivariate normal distribution (with discussion), *J. Roy. Statist. Soc. Ser. B*, **24**, 265–296.

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

Whittle, P. (1960). Bounds for the moments of linear and quadratic forms in independent variables, *Theory Probab. Appl.*, **5**, 302–305.