Ann. Inst. Statist. Math.
Vol. 47, No. 1, 65–80 (1995)

OPTIMIZING THE SMOOTHED BOOTSTRAP

SUOJIN WANG

Department of Statistics, Texas A&M University, College Station, Texas 77843-3143, U.S.A.

(Received October 7, 1993; revised September 30, 1994)

Abstract. In this paper we develop the technique of a generalized rescaling in the smoothed bootstrap, extending Silverman and Young's idea of shrinking. Unlike most existing methods of smoothing, with a proper choice of the rescaling parameter the rescaled smoothed bootstrap method produces estimators that have the asymptotic minimum mean (integrated) squared error, asymptotically improving existing bootstrap methods, both smoothed and unsmoothed. In fact, the new method includes existing smoothed bootstrap methods as special cases. This unified approach is investigated in the problems of estimation of global and local functionals and kernel density estimation. The emphasis of this investigation is on theoretical improvements which in some cases offer practical potential.

Key words and phrases: Bootstrap, functional estimation, kernel density estimation, mean integrated squared error, mean squared error, quantile, rescaling, smoothing.

1. Introduction

The bootstrap introduced by Efron (1979) is a computationally intensive technique that has been shown useful in many statistical problems and applications. Its smoothed version has potential improvements over the standard bootstrap, as is studied by Efron (1979, 1982), Silverman and Young (1987), Hall *et al.* (1989), De Angelis and Young (1992) and others; see Efron and Gong (1983) for an interesting introduction.

Suppose that $X_1, \ldots, X_n$ is a random sample from an unknown continuous distribution F with density f. We are interested in estimating a population functional of interest $\alpha(F)$ with bootstrap estimator $\alpha(\hat{F})$, where $\hat{F}$ is the empirical distribution F_n or its smoothed versions discussed below. Bootstrap resampling may often be required to obtain $\alpha(\hat{F})$ and its statistical properties.

Let K be a symmetric kernel function such that it is itself a density with unit variance. The assumption of unit variance is merely for simplicity in the presentation and could be dropped with some extra notation. The standard kernel

estimator $\hat{f}_h(x)$ of $f(x)$ is given

$$\hat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right), \tag{1.1}$$

where h is the smoothing parameter. The corresponding distribution function estimator is $\hat{F}_h(x) = \int_{-\infty}^{x} \hat{f}_h(t)dt$.

The method of shrinking in the smoothed bootstrap was given in Silverman and Young (1987) as a means of preserving the variance structure, i.e., the resulting kernel density estimate has the same variance structure as the original data. This idea of shrinking was ealier presented in a special case in Silverman (1981). In kernel density estimation, the most complete study to date of correcting an inflated variance is given by Jones (1991), but earlier references date back to Fryer (1976). As in the case of standard smoothing, it has been shown that their shrunk smoothing is beneficial in certain situations in the problems of functional estimation (Silverman and Young (1987)) and density estimation (Jones (1991)).

Their methods may not be applicable in some other general situations as is explained in their papers. It is because the amount of shrinking is fixed no matter what the true underlying distribution might be. That is, the amount of shrinking takes good care of preserving the variance structure, but does not take into account other factors determined by the true distribution that are also important in the estimation procedures, since the behavior of an estimator is affected by other factors such as curvature, besides the variance structure. Another useful reference is Fisher *et al.* (1994) in the case of testing for multimodality.

In this paper we extend Silverman and Young's (1987) method of shrinking and propose the following rescaled version of the smoothed bootstrap. This is an attempt to unify different versions of the smoothed bootstrap and make optimal use of smoothing, at least in asymptotic sense. Let

$$\hat{f}_{h,b}(x) = (1 + r)\hat{f}_h\{x + r(x - \bar{X})\}, \tag{1.2}$$

where $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, $r = (1 + |b|h^2)^{\mathrm{sgn}(b)/2} - 1$, and b is a constant (rescaling parameter) that is independent of n and can be less than zero. However, it is always true that $1 + r > 0$. Moreover, conditional on $X_1, \ldots, X_n$, $\hat{f}_{h,b}(x)$ is a well defined density with the distribution function $\hat{F}_{h,b}(x)$. Let the mean and variance of density $\hat{f}_h(x)$ be $\hat{\mu}_h$ and $\hat{\sigma}_h^2$. Then $\hat{\mu}_h = \bar{X}$ and the mean and variance of $\hat{f}_{h,b}(x)$ are $\bar{X}$ and $(1 - bh^2)\hat{\sigma}_h^2 + O_p(h^4)$. Therefore, $\hat{f}_{h,b}(x)$ is a rescaled form of $\hat{f}_h(x)$. We will call $\alpha(\hat{F}_{h,b})$ the rescaled smoothed bootstrap estimator for $\alpha(F)$. Note that when b^{-1} is taken to be the sample variance S^2, $\alpha(\hat{F}_{h,b})$ is the same as Silverman and Young's shrunk smoothed bootstrap estimator (but centered at $\bar{X}$), and $\hat{f}_{h,b}(x)$ is the variance corrected density estimate studied by Jones (1991). When $b = 0$, $\alpha(\hat{F}_{h,b})$ is the standard smoothed bootstrap estimator $\alpha(\hat{F}_h)$. The flexibility of b enables us to find the best possible amounts of shrinkage for different estimation procedures. The criterion for the selection of b is based on the asymptotic mean (integrated) squared error, abbreviated as MSE (MISE), as the sample size increases to infinity. The optimal b is often outside of interval $[0, S^{-2}]$,

as we will see in later sections. This approach can be alternatively viewed as using variants of S^2 with better MSE or MISE properties. We stress that the theoretical development of the asymptotic properties of the new estimators, rather than their implementations, is our focus in this paper, although some small scale simulation results are also reported.

We are going to show that by allowing the optimal selection of b in (1.2), the rescaled estimators asymptotically improve the standard unsmoothed bootstrap methods, in contrast to the standard and shrunk smoothed bootstrap estimators. In particular, in Section 2 we will investigate the new estimator $\alpha(\hat{F}_{h,b})$ for a functional $\alpha(F)$ that depends on global properties of F, such as those considered in Silverman and Young (1987). Section 3 explores the problem of estimating functionals that depend on local properties of F, such as quantile variances. The rescaling in kernel density estimation is investigated in Section 4. The results of rescaling are appealing in that for suitable choices of b, asymptotically the MSE or MISE is almost always reduced, often significantly in terms of asymptotic order, from that of the existing bootstrap methods.

2. Estimation of global functionals

Suppose that we are interested in the estimation of a global functional (a functional that depends on global properties of the underlying distribution) $\alpha(F)$ that can be written as

$$\alpha(F) = \int_{-\infty}^{\infty} a(t)f(t)dt \tag{2.1}$$

for some function $a(t)$. The same setting has been considered by Silverman and Young (1987). Recall that the standard smoothed bootstrap estimator $\alpha(\hat{F}_h)$ does not affect the first-order asymptotics of the MSE of $\alpha(\hat{F})$. Likewise the rescaling technique, when applied to the problem of estimating $\alpha(F)$ in (2.1), improves the second-order accuracy of the MSE. Therefore, the theory developed in this section is mainly of theoretical interest and the resultant improvement is relatively small in practical terms. However, the technique can improve the first-order accuracy in some other problems such as estimating local functionals and kernel density estimation, as we will see in Sections 3 and 4.

The main result in this section is given in the following theorem which extends Silverman and Young's (1987) results. We first define the following quantities: $\mu = E(X)$, $C_1 = E\{(X-\mu)a'(X)\}$, $C_2 = E\{a''(X)\}$, $C_3 = E[(X-\mu)\{a(X) - \alpha(F)\}a'(X)]$ and $C_4 = E[\{a(X) - \alpha(F)\}a''(X)]$.

THEOREM 2.1. *For any global functional in the form of* (2.1) *with the quantities above well defined, the mean squared error of*

$$\alpha(\hat{F}_{h,b}) = \int_{-\infty}^{\infty} a(t)\hat{f}_{h,b}(t)dt \tag{2.2}$$

can be reduced below that of $\alpha(F_n)$ *by choosing suitable* $h > 0$ *and* b *such that*

$$b\,\mathrm{sgn}(C_3) > \frac{C_4}{|C_3|}, \tag{2.3}$$

provided that $a^{(4)}(x)$ *is continuous and* $C_3 \neq 0$. *In fact,*

$$\text{MSE}\{\alpha(\hat{F}_{h,b})\} = \frac{1}{n}\text{var}\{a(X)\} + \frac{h^2}{n}(C_4 - bC_3) + O(h^4). \tag{2.4}$$

PROOF. From (1.1), (1.2) and (2.2) we have

$$\alpha(\hat{F}_{h,b}) = \frac{1}{n}\sum_{i=1}^{n} z(X_i),$$

where

$$\begin{aligned} z(X_i) &= \frac{1+r}{h}\int_{-\infty}^{\infty} a(t)K[\{t + r(t - \bar{X}) - X_i\}/h]dt \\ &= a(X_i) - r(X_i - \mu)a'(X_i) + \frac{h^2}{2}a''(X_i) + O_p(\varepsilon_n), \end{aligned}$$

and $\varepsilon_n = h^4 + h^2/n^{1/2}$. Note that in the last equation we replaced $\bar{X}$ by μ, resulting in an error of order $O_p(h^2/n^{1/2})$ that is absorbed into $O_p(\varepsilon_n)$. The last equation shows that except for the negligible errors the $z(X_i)$ behave like mutually independent random variables. Hence,

$$E\{\alpha(\hat{F}_{h,b})\} = \alpha(F) - rC_1 + \frac{h^2}{2}C_2 + O(\varepsilon_n)$$

and

$$n\,\text{var}\{\alpha(\hat{F}_{h,b})\} = \text{var}\{a(X)\} - 2rC_3 + h^2C_4 + O(\varepsilon_n),$$

using the fact that $r = bh^2/2 + O(h^4)$. We have thus obtained (2.4). Since $\text{MSE}\{\alpha(F_n)\} = \frac{1}{n}\text{var}\{a(X)\}$, it is seen that the mean squared error of $\alpha(\hat{F}_{h,b})$ will, at least for small h, be smaller than that of $\alpha(F_n)$ as long as

$$C_4 - bC_3 < 0.$$

The inequality above can be warranted by selecting any b satisfying (2.3), as was to be shown. □

In practice, C_3 and C_4 are usually unknown but may be estimated by the corresponding sample means or by a kernel method with bandwidth g (e.g., $g = h$). Here only first-order accuracy of C_3 and C_4 is required since they only appear in the second-order term, so that such estimation results in a higher-order error that is absorbed in $O(\varepsilon_n)$. Noticing the fact that $b = 0$ and $b = S^{-2}$ correspond to the standard and shrunk smoothed bootstrap in Silverman and Young (1987) respectively, the flexibility of choice of b enables the new estimator to have better asymptotic properties. The b is usually no longer selected to be S^{-2}, and the objective here is to reduce the mean squared error instead of preserving the variance structure for the density estimator. Silverman and Young (1987) discussed the

optimal choice of h in the case where $a(t)$ in (2.1) depends on n and converges to zero as $n \to \infty$, by looking ahead at the $O(h^4)$ term in the MSE. Further work is needed for the case where $a(t)$ is a fixed function.

As an illustrative example, now let us consider the problem of estimating the moments

$$M(k) = \int_{-\infty}^{\infty} t^k f(t)dt$$

for $k = 1, 2, \ldots$. Employing the commonly used normal kernel $K(\cdot) = \phi(\cdot)$ and after some algebra, the estimator defined in (2.2) is

$$\hat{M}_{h,b}(k) = \frac{1}{n(1+r)^k} \sum_{l=0}^{[k/2]} h^{2l} d_{k,l} \sum_{i=1}^{n} (X_i + r\bar{X})^{k-2l},$$

where $d_{k,l} = \binom{k}{2l}(2l-1)(2l-3)\cdots 1$, for $l \geq 1$ and $d_{0,0} = 1$. In the special case of $k = 1$, $\hat{M}_{h,b}(1) = \bar{X}$. This indicates that neither the smoothing nor the rescaling affects the estimation of mean, which is intuitive. It is not the case for other k's, however. For example, when $k = 2$,

$$\begin{aligned}\hat{M}_{h,b}(2) &= \frac{1}{(1+r)^2}\left\{\frac{1}{n}\sum_{i=1}^{n}(X_i + r\bar{X})^2 + h^2\right\} \\ &= \frac{1}{n}\sum_{i=1}^{n} X_i^2 + h^2\left\{1 - b\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2\right)\right\} + O_p(h^4).\end{aligned}$$

By Theorem 2.1, at least for small h, asymptotically $\hat{M}_{h,b}(2)$ has a smaller MSE for any fixed b that has the same sign as C_3 since $C_4 \equiv 0$ in this case.

We have conducted a small scale simulation experiment in the simple case of estimating the second moment. Table 1 reports a summary of the root mean squared errors of the standard and several smoothed bootstrap estimates obtained from the simulation results with 1,000 runs. The selection of h is somewhat arbitrary for different n but in the spirit of the fact that $h \to 0$ as $n \to \infty$. The 'normal' distribution is $N(1,1)$, 'uniform' is rescaled (to have unit variance) $U(0,1)$, 'log-normal' is rescaled *lognormal*$(0,1)$, and 'χ_3^2' is also rescaled. In this particular case, $b = 1.5$ seems to have an overall best performance while $b = 2$ works better for skewed distributions. The shrunk bootstrap ($b = 1$) is seen to be quite good. It appears that smoothing in general is worthwhile mostly for the case of small sample sizes.

As is in Silverman and Young (1987), Theorem 2.1 applies to functionals in the form of (2.1). It is often desirable to consider a more general functional $\alpha(F)$ that does not necessarily have the expression (2.1). Instead, suppose that $\alpha(F)$ admits a first-order von Mises expansion such that for any $\tilde{F}$ satisfying $\sup|\tilde{F}(x) - F(x)| = O_p(n^{-1/2})$,

$$\alpha(\tilde{F}) = \alpha(F) + A(\tilde{F}) - A(F) + O_p(n^{-1}),$$

Table 1. Root mean squared errors of standard and smoothed bootstrap estimates for the second moment; sample sizes n, smoothing parameters h and rescaling parameters b.

n and h	Distribution	Standard	Smoothed				
			$b=0$	$b=.5$	$b=1$	$b=1.5$	$b=2$
	normal	.762	.835	.743	.706	.699	.707
$n=10$	uniform	1.143	1.195	1.144	1.127	1.126	1.133
$h=.6$	χ^2_3	1.351	1.391	1.273	1.202	1.160	1.133
	lognormal	3.155	3.178	2.785	2.504	2.294	2.132
	normal	.527	.583	.522	.496	.494	.504
$n=20$	uniform	.800	.839	.805	.793	.792	.801
$h=.5$	χ^2_3	1.035	1.068	.992	.943	.912	.893
	lognormal	1.863	1.871	1.694	1.559	1.454	1.371
	normal	.343	.377	.343	.329	.330	.342
$n=50$	uniform	.504	.530	.508	.500	.500	.511
$h=.4$	χ^2_3	.632	.655	.615	.589	.574	.566
	lognormal	1.087	1.088	1.018	.963	.919	.884
	normal	.240	.252	.238	.233	.234	.243
$n=100$	uniform	.356	.368	.358	.354	.354	.360
$h=.3$	χ^2_3	.438	.448	.431	.420	.414	.412
	lognormal	.945	.948	.910	.877	.848	.824

where $A(F)$ has the form (2.1) for a smooth function $a(t)$. Then the sampling properties of $\alpha(\tilde{F})$ will be approximately the same as those of $A(\tilde{F})$. Following the arguments of Silverman and Young (1987) and De Angelis and Young (1992) the effect of smoothing on estimation of $\alpha(F)$ may be approximated by the effect on estimation of $A(F)$. The smoothing effects on the latter have been summarized in Theorem 2.1.

We conclude this section by remarking that the discussion here is not giving any global prescription for smoothing (true for all f). Rather, the result shows that for a given f we can find a b for which the smoothed bootstrap has advantages over the unsmoothed bootstrap. This comment applies also to the approaches in the next two sections.

3. Estimation of local functionals

In Section 2 we have shown that suitable rescaling in (1.2) improves the second-order accuracy of the estimator of a global functional. We now consider the problem of estimating a different type of functionals—local functionals (functionals that depend on local, rather than global, properties of the underlying distribution). The standard smoothed bootstrap estimation of this type of functionals has been considered by Hall *et al.* (1989), and De Angelis and Young (1992); see also Falk and Reiss (1989) who have discussed the benefits of smoothing when bootstrapping the quantile empirical process. Optimal plug-in estimators for non-

parametric local functional estimation have been studied by Goldstein and Messer (1992).

This section will be dedicated to a stronger conclusion that the same kind of rescaling can even generally increase the convergence rate of the mean squared error of the smoothed bootstrap estimator of a local functional in the sense that the mean squared error of $\alpha(\hat{F}_{h,b})$ converges to zero faster than that of $\alpha(\hat{F}_h)$ for properly chosen b. For simplicity, we will focus on the special case of the variance of a sample quantile as in Hall *et al.* (1989) and De Angelis and Young (1992). However, we will present the derivations in a general way so that the idea can be easily seen to apply analogously to other local functionals.

Assume that, for given $0 < p < 1$, the p-th population quantile is uniquely defined and is

$$\xi_p = F^{-1}(p).$$

Let $X_{n,s}$ denote the s-th largest of the sample values $X_1, \ldots, X_n$, where $s = \langle np \rangle + 1$ and $\langle x \rangle$ is the largest integer strictly less than x. Then $\hat{\xi}_p = X_{n,s}$ is the p-th sample quantile. We wish to estimate the variance of $\hat{\xi}_p$ given by

$$\alpha(F) = \int_{-\infty}^{\infty} \{x - \beta(F)\}^2 H\{F(x); n, p\} dF(x), \tag{3.1}$$

where

$$H\{F(x); n, p\} = [n!/\{(s-1)!(n-s)!\}] F(x)^{s-1} \{1 - F(x)\}^{n-s}$$

and

$$\beta(F) = \int_{-\infty}^{\infty} x H\{F(x); n, p\} dF(x).$$

In obtaining a smoothed bootstrap estimate of $\alpha(F)$ we will continue to use second-order kernels since they have many nice properties such as nonnegativity that a higher-order kernel is lacking; see De Angelis and Young (1992). However, we point out that the rescaling technique works exactly in the same manner when a higher-order kernel is employed. Assume that $f''(x)$ exists and is uniformly continuous and bounded, $f(x)$ is bounded away from 0 in a neighborhood of ξ_p and $E(|X|^\varepsilon) < \infty$ for some $\varepsilon > 0$. Then Hall *et al.* (1989) have shown that the relative mean squared error of $\alpha(\hat{F}_h)$, the standard smoothed bootstrap estimator, is of order $O(n^{-4/5})$, in contrast to that of order $O(n^{-1/2})$ for the case of the unsmoothed variance estimation (Hall and Martin (1988)). We are going to show that $\alpha(\hat{F}_{h,b})$, the rescaled version of $\alpha(\hat{F}_h)$, has an even smaller relative mean squared error of order $O(n^{-8/9})$ for a suitably chosen b. This resembles that given by a fourth-order kernel estimate in the standard smoothed bootstrap approach for a reason given at the end of this section.

THEOREM 3.1. *Let* $V_{n,p} = n \operatorname{var}(\hat{\xi}_p)$, $D_1 = -p(1-p) f(\xi_p)^{-4} [f(\xi_p) f''(\xi_p) - \{f'(\xi_p)\}^2]$ *and* $D_2 = 4p^2(1-p)^2 f(\xi_p)^{-5} \int_{-\infty}^{\infty} K(t)^2 dt$, *and assume that the general conditions given above hold. Then*

$$\operatorname{MSE}\{n\alpha(\hat{F}_{h,b})\} = O(n^{-8/9}) \tag{3.2}$$

for the asymptotically optimal choice of

$$b = \frac{D_1}{V_{n,p}} \tag{3.3}$$

and

$$h = cn^{-1/9} \tag{3.4}$$

with fixed $c > 0$, *and thus when* $D_1 \neq 0$,

$$\frac{\mathrm{MSE}\{\alpha(\hat{F}_{h,b})\}}{\mathrm{MSE}\{\alpha(\hat{F}_{h_1})\}} \to 0, \tag{3.5}$$

as $n \to \infty$, *for any choice of* h_1.

PROOF. From the proof of Theorem 3.1 of Hall *et al.* (1989), it is seen that for h satisfying $h \log n \to 0$ and $nh^3/\log n \to \infty$ as $n \to \infty$,

$$E\{n\alpha(\hat{F}_h)\} = V_{n,p} + h^2 D_1 + o\{(nh)^{-1/2}\} + O(h^4) \tag{3.6}$$

and

$$\mathrm{var}\{n\alpha(\hat{F}_h)\} = \frac{D_2}{nh} + o\{(nh)^{-1}\} + O\left\{\left(\frac{h^7}{n}\right)^{1/2} + h^8\right\}. \tag{3.7}$$

By (1.2) and (3.1) we have

$$\begin{aligned}\beta(\hat{F}_{h,b}) &= \int_{-\infty}^{\infty} xH[\hat{F}_h\{x + r(x - \bar{X})\}; n, p]\hat{f}_{h,b}(x)dx \\ &= \frac{\beta(\hat{F}_h) + r\bar{X}}{1 + r},\end{aligned}$$

and therefore the following interesting identity is obtained

$$\begin{aligned}n\alpha(\hat{F}_{h,b}) &= n\int_{-\infty}^{\infty}\left\{x - \frac{\beta(\hat{F}_h) + r\bar{X}}{1 + r}\right\}^2 H[\hat{F}_h\{x + r(x - \bar{X})\}; n, p] \\ &\qquad \cdot \hat{f}_{h,b}(x)dx \\ &= \frac{n\alpha(\hat{F}_h)}{(1 + r)^2}.\end{aligned} \tag{3.8}$$

Combining (3.6), (3.7) and (3.8) gives

$$\begin{aligned}\mathrm{MSE}\{n\alpha(\hat{F}_{h,b})\} =& \frac{D_2}{nh} + (h^2 D_1 - 2rV_{n,p})^2 \\ &+ (h^2 D_1 - 2rV_{n,p})[o\{(nh)^{-1/2}\} + O(h^4)] + \rho_n,\end{aligned} \tag{3.9}$$

where $\rho_n = o\{(nh)^{-1}\} + O\{(\frac{h^7}{n})^{1/2} + h^8\}$ is the remainder. Recall that $r = bh^2/2 + O(h^4)$. Thus, for fixed h, $\mathrm{MSE}\{n\alpha(\hat{F}_{h,b})\}$ is minimized by letting

$$b = \frac{D_1}{V_{n,p}},$$

in which case

$$\mathrm{MSE}\{n\alpha(\hat{F}_{h,b})\} = \frac{D_2}{nh} + \rho_n. \tag{3.10}$$

It is readily seen that the fastest rate of convergence of (3.10) is achieved for any $c > 0$ and

$$h = cn^{-1/9},$$

with the rate of convergence being of order $O(n^{-8/9})$. Since the optimal rate of $\mathrm{MSE}\{n\alpha(\hat{F}_h)\}$ is $n^{-4/5}$ (when $D_1 \neq 0$), this verifies the theorem. □

The quantities D_1 and $V_{n,p}$ in (3.3) can be estimated by consistent estimators to obtain an estimator for b in practice. Note that it is possible to find the optimal value of c in (3.4) to minimize the coefficient of the first order term of $\mathrm{MSE}\{n\alpha(\hat{F}_{h,b})\}$, by keeping track of the coefficient of the term with order h^8. The details are lengthy and complicated.

It is worth pointing out that even if a higher-order kernel is used in (1.2) the rescaling technique is still valid in reducing the first-order bias in $\alpha(\hat{F}_h)$. Therefore, Theorem 3.1 is also valid in this case. The proof is very similar to the second-order kernel case and is thus not given here. It is interesting to observe the fact that in this particular variance estimation problem replacing $\bar{X}$ in (1.2) by any constant will not affect any results in Theorem 3.1 or its proof.

Note that the rescaling in (3.8) is to multiply a smoothed estimate by $(1 + \frac{1}{2}bh^2)^{-2}$. This essentially leads to an additive bias correction. One can do so in other settings, such as density estimation itself. A careful examination reveals that the additive bias corrections are asymptotically equivalent to fourth-order kernels. See, for example, Jones and Foster (1993) and references therein for discussion on properties of higher-order kernels. A main difference is that the approach of rescaling preserves the nonnegativity of the kernel we use, while higher-order kernels themselves are not nonnegative.

4. Kernel density estimation

Proper rescaling in (1.2) has been shown to be effective in improving bootstrap estimators for both global and local functionals. We now proceed to prove the fact that such rescaling is also useful, at least in theory, in the problem of kernel density estimation.

There is rich literature concerning kernel density estimation; see for example the excellent monograph by Silverman (1986). However, few authors addressed the approach of rescaling. Jones (1991) considers the problem of correcting for variance inflation and concludes that such correction can be either beneficial or

not, depending on the underlying distribution. A commonly used criterion to measure the performance of a density estimator $\hat{f}$ is the mean integrated squared error defined by

$$\text{MISE}(\hat{f}) = E\int_{-\infty}^{\infty}\{\hat{f}(x) - f(x)\}^2 dx, \tag{4.1}$$

which we will use in the current problem.

Our objective here is to minimize $\text{MISE}(\hat{f}_{h,b})$ by choosing the right b value, say, b_1. Thus, $\hat{f}_{h,b_1}$ is judged to generally perform better than the standard estimate $\hat{f}_h$ (when $b = 0$) and the variance corrected estimate $\hat{f}_{h,S^{-2}}$ (when $b = S^{-2}$) considered by Jones (1991). This criterion is in contrast to that for the variance corrected estimate in Jones (1991) where the density estimate is defined to preserve the sample variance structure.

It is well-known that the mean integrated squared error of $\hat{f}_h$ has the expansion

$$\text{MISE}(\hat{f}_h) = \frac{1}{nh}\int_{-\infty}^{\infty} K^2(t)dt + \frac{h^4}{4}\int_{-\infty}^{\infty}\{f''(x)\}^2 dx + o(\psi_n), \tag{4.2}$$

with the assumption that $\int_{-\infty}^{\infty}\{f''(x)\}^2 dx < \infty$, where $\psi_n = (nh)^{-1} + h^4$. Suppose further that $\mu = E(X)$ exists, $\lim_{|x|\to\infty} f'(x)f(x) = 0$ and $\int_{-\infty}^{\infty}\{(x-\mu)f'(x)\}^2 dx < \infty$. We now establish the following theorem.

THEOREM 4.1. *Under the above general conditions, the asymptotic mean integrated squared error of $\hat{f}_{h,b}$ is minimized at*

$$b = b_1 = \frac{3\int_{-\infty}^{\infty}\{f'(x)\}^2 dx}{2\int_{-\infty}^{\infty}\{(x-\mu)f'(x)\}^2 dx}, \tag{4.3}$$

with

$$\text{MISE}(\hat{f}_{h,b_1}) = \frac{1}{nh}A_1 + \frac{h^4}{4}A_2 + o(\psi_n), \tag{4.4}$$

where

$$A_1 = \int_{-\infty}^{\infty} K^2(t)dt, \qquad A_2 = \int_{-\infty}^{\infty}\{f''(x)\}^2 dx - \frac{9}{4}\frac{[\int_{-\infty}^{\infty}\{f'(x)\}^2 dx]^2}{\int_{-\infty}^{\infty}\{(x-\mu)f'(x)\}^2 dx},$$

and $\psi_n = (nh)^{-1} + h^4$. Furthermore, when $A_2 \neq 0$ the optimal h is

$$h = h_1 = \left(\frac{A_1}{A_2 n}\right)^{1/5}.$$

PROOF. By the definition in (1.2), we have

$$E\{\hat{f}_{h,b}(x)\} = (1+r)E[\hat{f}\{x + r(x - \bar{X})\}] \tag{4.5}$$
$$= f(x) + \frac{h^2}{2}[f''(x) + b\{f(x) + (x-\mu)f'(x)\}] + o(h^2),$$

and

$$\operatorname{var}\{\hat{f}_{h,b}(x)\} = \operatorname{var}\{\hat{f}_h(x)\} + o(\psi_n). \tag{4.6}$$

Thus, from equations (4.5) and (4.6) and the regularity conditions, it is easily seen that

$$\begin{aligned}\text{(4.7)}\quad &\operatorname{MISE}\{\hat{f}_{h,b}(x)\} \\ &= \int_{-\infty}^{\infty} \left(\operatorname{var}\{\hat{f}_h(x\} + \frac{h^4}{4}[f''(x) + b\{f(x) + (x-\mu)f'(x)\}]^2\right) dx \\ &\quad + o(\psi_n) \\ &= \frac{1}{nh}\int_{-\infty}^{\infty} K^2(t)dt \\ &\quad + \frac{h^4}{4}\int_{-\infty}^{\infty} [\{f''(x)\}^2 - 3b\{f'(x)\}^2 + b^2\{(x-\mu)f'(x)\}^2]dx \\ &\quad + o(\psi_n).\end{aligned}$$

The last equation above was obtained by using the identities

$$\int_{-\infty}^{\infty} f''(x)f(x)dx = -\int_{-\infty}^{\infty} \{f'(x)\}^2 dx \tag{4.8}$$

and

$$\int_{-\infty}^{\infty} (x-\mu)f^{(i)}(x)f^{(i+1)}(x)dx = -\frac{1}{2}\int_{-\infty}^{\infty} \{f^{(i)}(x)\}^2 dx,$$

for $i = 0, 1$. Therefore, equation (4.7) is minimized at

$$b = \frac{3\int_{-\infty}^{\infty}\{f'(x)\}^2 dx}{2\int_{-\infty}^{\infty}\{(x-\mu)f'(x)\}^2 dx}.$$

The asymptotic minimum value of $\operatorname{MISE}(\hat{F}_{h,b})$ is readily seen to be that in (4.4). The optimal choice of h is obvious, completing the proof. □

Theorem 4.1 indicates that the rescaled density estimator in (1.2) for a properly chosen b always has a smaller mean integrated squared error than that of the standard estimator $\hat{f}_h$. This is different from the case of the variance corrected density estimation where the mean integrated squared error may be decreased or increased depending on the underlying distribution. Furthermore, the new method

asymptotically improves the variance corrected estimator even when the underlying distribution favors variance correction, since it is in the optimized form.

Note that the $O(h^4)$ term in (4.4) is nonnegative. This can be checked by using the identities above and the Cauchy-Schwartz inequality. Moreover, for the conventional kernel estimator the term $\int_{-\infty}^{\infty}\{f''(x)\}^2dx$ in the bias in (4.2) indicates the curvature and when f is "more curved" greater bias would show up. When the rescaling adjustment in Theorem 4.1 is allowed some of the curvature is eliminated. In fact, $[\int_{-\infty}^{\infty}\{f'(x)\}^2dx]^2 = [E\{f''(X)\}]^2$ reflects the significance of the average curvature and $\int_{-\infty}^{\infty}\{(x-\mu)f'(x)\}^2dx$ takes care of the scale.

To appreciate the result in Theorem 4.1 we now consider the problem of estimating the following normal mean mixture distribution:

$$f(x) = p\phi(x-a) + (1-p)\phi(x+a).$$

For simplicity, let $p = 1/2$. Then it is easy to obtain that b_1 in (4.3) is

$$b_1 = \frac{3g_1(a)}{2g_2(a)},$$

and the bias reduction (the second term of A_2 in (4.4)) is

$$BR = -\frac{9g_1^2(a)}{32\sqrt{\pi}g_2(a)},$$

where $g_1(a) = 1+e^{-a^2}(1-2a^2)$ and $g_2(a) = 3/2+a^2+e^{-a^2}(3/2-a^2)$. When $a = 0$, $f(x)$ is the standard Gaussian density. Then $b_1 = 1$ and $BR = -3/(8\sqrt{\pi})$, which is the same as in the Jones (1991) method. This indicates that the Jones method is optimal in this special case. Jones (1991) has demonstrated good effects of the rescaling in this case. The two methods are different when $a \neq 0$. For example, for $a = 1$ we have S^2 approximately equal to 2, $b_1 = .5589$ and $BR = -.0236$, which has some improvement over $-.0195$ obtained by the Jones method. Thus it is possible that $b_1 > S^{-2}$ in the case of density estimation. Algebraically more complicated results for $p \neq 0$ have also been obtained with similar conclusions. Since the new method is applicable to a general class of densities, including those not satisfying equation (14) of Jones (1991), the asymptotic gain could be more dramatic than this example shows. However, our criterion used here, the MISE, is an overall measure. It does not appear to perfectly measure the performance of density estimates if our main interest is the tails of the distribution.

Practically, a consistent estimator of b_1 may be needed in the estimation procedure. One can first estimate $f'(x)$ by a kernel method

$$\hat{f}'_g(x) = \frac{1}{ng^2}\sum_{i=1}^{n} K'\left(\frac{x-X_i}{g}\right),$$

where g is a new smoothing parameter. Then the functionals in (4.3) may be estimated via numerical intergration. Alternatively, due to the nature of the functionals, by (4.8) $a_1 = \int_{-\infty}^{\infty}\{f'(x)\}^2dx$ may be estimated with

$$\hat{a}_1 = \frac{-1}{n}\sum_{i=1}^{n}\hat{f}''_g(X_i),$$

and $a_2 = \int_{-\infty}^{\infty}\{f'(x)(x-\mu)\}^2 dx = -\int_{-\infty}^{\infty} f''(x)(x-\mu)^2 f(x)dx - 2\int_{-\infty}^{\infty} f'(x)(x-\mu)f(x)dx$ may be estimated with

$$\hat{a}_2 = \frac{-1}{n}\sum_{i=1}^{n}\{\hat{f}_g''(X_i)(X_i-\bar{X})^2 + 2\hat{f}_g'(X_i)(X_i-\bar{X})\},$$

where $\hat{f}_g''(x) = \frac{1}{ng^3}\sum_{i=1}^{n} K''(\frac{x-X_i}{g})$; see Jones and Sheather (1991). We may use h or a somewhat smaller value for g.

Now suppose that we want to estimate a density with a data set of size $n = 100$ which is in fact sampled from a t_4 distribution. Figures 1 and 2 give three estimates, with $\hat{b}_1 = 3\hat{a}_1/2\hat{a}_2 = .90$ (using bandwidth $g = .7$). The true b_1 is 1.12 and $S^{-2} = .53$. The normal kernel was used for K. All the computing was done with a short $S+$ program. Since the distribution is so heavy tailed, if we use h smaller than .7 it is likely to have bad estimates in the tails. In both figures, both rescaled estimates improve over the standard method, and the optimal method appears to perform the best. This is especially true when the standard method is oversmoothed (see Jones (1991) for a similar comparison in the case of a normal density).

For heavy tailed distributions we might even want to purposely use a larger h in the optimal rescaled method since this will help smooth out unwanted bumps in the tails, and the method is otherwise rather robust to oversmoothing. It seems to be true at least for symmetric distributions such as t distributions.

Sometimes we may want to measure the performance of density estimators at each fixed x value. In such a case, we commonly employ the measure of the mean squared error $\mathrm{MSE}\{\hat{f}(x)\}$. When $f(x) + (x-\mu)f'(x) \neq 0$ the bias in (4.5) can be reduced to a higher-order error by selecting a sensible b, now dependent on x, so that $\mathrm{MSE}\{\hat{f}(x)\}$ can also be reduced to a higher-order error.

Finally we present the following corollary of Theorem 4.1.

COROLLARY 4.1. *Assume that $f(x)$ is twice differentiable for all x. Then under the same conditions as in Theorem* 4.1, *the quantity A_2 in* (4.4) *is zero if and only if (iff) $f(x)$ is a normal density with variance b_1^{-1}.*

PROOF. First it is seen that $A_2 = 0$ iff

$$f''(x) + b_1\{f(x) + (x-\mu)f'(x)\} = 0 \tag{4.9}$$

for all x. This is equivalent to

$$f''(x) = -b_1\{(x-\mu)f(x)\}',$$

i.e., $f'(x) = -b_1(x-\mu)f(x) + c_0$ for some c_0. But $c_0 = 0$ by letting $x \to \infty$. Thus, equivalently $[\log\{f(x)\}]' = -b_1(x-\mu)$, or

$$f(x) = c_1 \exp\{-b_1(x-\mu)^2/2\}.$$

That is, $f(x)$ is a normal density with variance b_1^{-1}. □

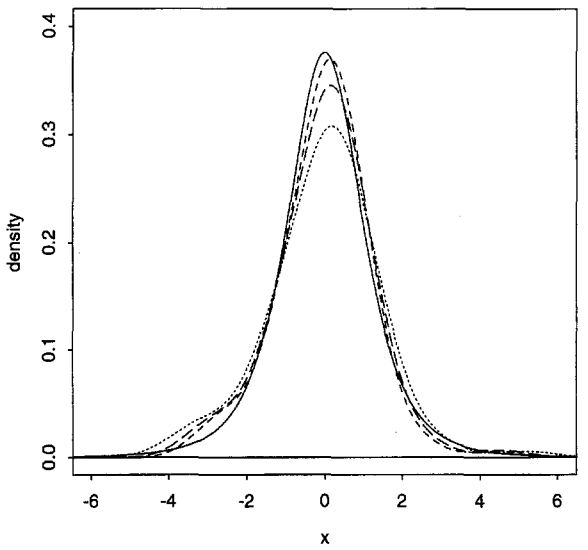

Fig. 1. True t_4 density (solid line) and estimated densities $\hat{f}_h$ (dotted line), $\hat{f}_{h,S^{-2}}$ (long-dashed line) and $\hat{f}_{h,\hat{b}_1}$ (short-dashed line); $h = .7$, $\alpha = .7$.

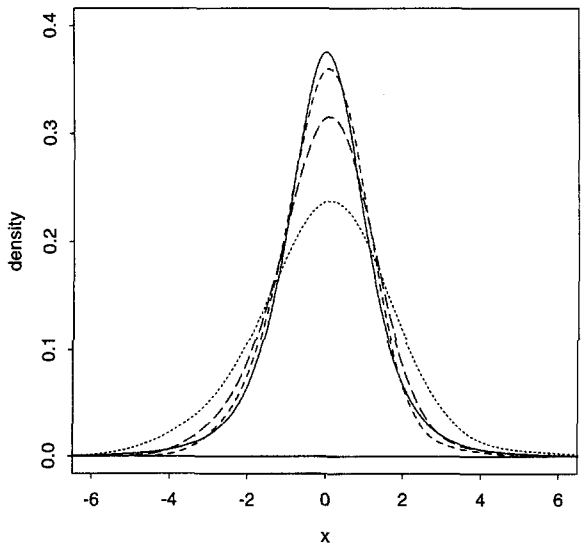

Fig. 2. True t_4 density (solid line) and estimated densities $\hat{f}_h$ (dotted line), $\hat{f}_{h,S^{-2}}$ (long-dashed line) and $\hat{f}_{h,\hat{b}_1}$ (short-dashed line); $h = 1.2$, $\alpha = .7$.

Note that if we assume the weaker condition that $f(x)$ is twice differentiable in the open intervals (x_{i-1}, x_i) for $i = 1, \ldots, k$ (k may be ∞), where $x_0 = -\infty < x_1 < \cdots < x_k = \infty$, then $A_2 = 0$ iff $f(x)$ is a density produced by k functions of form $f_i(x) = d_i \exp\{-b_1(x-\mu)^2/2\}$ (with $d_i \geq 0$) defined on segment (x_{i-1}, x_i)

for $i = 1, \ldots, k$. This conclusion can be shown following the proof of Corollary 4.1.

5. Conclusions

In this paper we have taken a unified approach and developed the technique of a generalized rescaling that has potential applications in variety of statistical problems. In particular, we have shown that the asymptotic performance of the smoothed bootstrap estimators for both global and local functionals can be generally improved by optimally choosing the rescaling parameter. In the case of estimating a local functional, the application of the technique even eliminates the first-order of mean squared error of the smoothed bootstrap estimator. This new technique is also proved to make asymptotic improvements in the problem of kernel density estimation. Our limited numerical experience suggests, however, that the shrunk smoothed bootstrap is often nearly optimal.

It is worth mentioning that in this paper we have discussed only the univariate case, but it is in principle possible to extend the results developed here to the multivariate case.

Acknowledgements

The author is grateful to a referee for many insightful comments and suggestions, including one leading to the result in Corollary 4.1. This material is based in part upon work supported by the Texas Advanced Research Program under Grant No. 160802 and the National Science Foundation under Grant DMS-9200610. A part of this research was performed while the author was visiting the U. S. Bureau of Labor Statistics on an ASA/NSF/BLS research fellowship.

References

De Angelis, D. and Young, G. A. (1992). Smoothing the bootstrap, *Internat. Statist. Rev.*, **60**, 45–56.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife, *Ann. Statist.*, **7**, 1–26.

Efron, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans*, SIAM, Philadelphia.

Efron, B. and Gong, G. (1983). A leisurely look at the bootstrap, the jackknife and cross-validation, *Amer. Statist.*, **37**, 36–48.

Falk, M. and Reiss, R.-D. (1989). Weak convergence of smoothed and unsmoothed bootstrap quantile estimates, *Ann. Probab.*, **17**, 362–371.

Fisher, N. I., Mammen, E. and Marron, J. S. (1994). Testing for multimodality, *Comput. Statist. Data Anal.*, **18**, 499–512.

Fryer, M. J. (1976). Some errors associated with the non-parametric estimation of density functions, *Journal of the Institute of Mathematics and its Applications*, **18**, 371–380.

Goldstein, L. and Messer, K. (1992). Optimal plug-in estimators for nonparametric functional estimation, *Ann. Statist.*, **20**, 1306–1328.

Hall, P. and Martin, M. A. (1988). Exact convergence rate of bootstrap quantile variance estimator, *Probab. Theory Related Fields*, **80**, 261–268.

Hall, P., DiCiccio, T. J. and Romano, J. P. (1989). On smoothing and the bootstrap, *Ann. Statist.*, **17**, 692–704.

Jones, M. C. (1991). On correcting for variance inflation in kernel density estimation, *Comput. Statist. Data Anal.*, **11**, 3–15.

Jones, M. C. and Foster, P. J. (1993). Generalized jackknifing and higher order kernels, *Journal of Nonparametric Statistics*, **3**, 81–94.

Jones, M. C. and Sheather, S. J. (1991). Using non-stochastic terms to advantage in estimating integrated squared density derivatives, *Statist. Probab. Lett.*, **11**, 511–514.

Silverman, B. W. (1981). Using kernel density estimates to investigate multimodality, *J. Roy. Statist. Soc. Ser. B*, **43**, 97–99.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, London.

Silverman, B. W. and Young, G. A. (1987). The bootstrap: to smooth or not to smooth? *Biometrika*, **74**, 469–479.