Estimation of Integrated Squared Density Derivatives

by

Peter Hall
Australian National University

J.S. Marron[1]
Australian National University
and University of North Carolina

AMS 1980 subject classification: primary 62G05, secondary 62G20.

Key words and phrases: Integrated squared derivative, kernel estimators, nonparametric estimation, rates of convergence

[1]Research partially supported by NSF Grant DMS-8400602.

Abstract: Kernel density estimators are used for the estimation of integrals of various squared derivatives of a probability density. Rates of convergence in mean squared error are calculated, which show that appropriate values of the smoothing parameter are much smaller than those for ordinary density estimation. The rate of convergence increases with stronger smoothness assumptions, however, unlike ordinary density estimation, the parametric rate of n^{-1} can be achieved even when only a finite amount of differentiability is assumed. The implications for data-driven bandwidth selection in ordinary density estiamation are considered.

Introduction

The estimation of the integral of a squared probability density has long been important in the study of rank-based nonparametric statistics. See Sheather and Hettmansperger (1987)) and section 4.4 of Prakasa Rao (1983) for an account of the literature on this topic. One method of data-driven bandwidth selection for density estimation involves plugging estimates of integrated squared derivatives into an asymptotic representation for the optimal bandwidth.

Under nonparametric assumptions it is natural to form estimates of these quantities based on a kernel estimate of the underlying density. Section 2 describes two methods for doing this, and provides motivation for a slight modification of the estimators.

Section 3 contains rate-of-convergence results in mean squared error of the type developed by Rosenblatt (1956, 1971) and Parzen (1962). As for standard density estimation, the rates become faster when stronger smoothness assumptions are made. An optimality theory is developed in which variance and bias are balanced. Since integration is a smoothing operation, it is not surprising that the optimal bandwidth is much smaller for the integrated squared derivatives of a density than for the ordinary derivatives. A more surprising result is that, unlike the case of standard density estimation, the parametric rate of convergence of n^{-1} may be achieved even when only a finite number of derivatives are assumed to exist for the underlying density.

Section 4 has some remarks, including a discussion of the implications of the convergence rate results for automatic bandwidth selection in density estimation. All proofs are in the appendix.

2. The Estimators

Consider the problem of estimating, for some $m = 0,1,\ldots$, the parameter

$$\theta_m = \int \{f^{(m)}(x)\}^2 dx,$$

using a random sample, $X_1,\ldots,X_n$ from a probability density f. An obvious first attempt at estimation is

$$\tilde{\theta}_m = \int \{\hat{f}^{(m)}(x)\}^2 dx,$$

where $\hat{f}_{(x)}$ is some reasonable estimator of f(x). One candidate is the kernel estimator

$$\hat{f}_h(x) = n^{-1} \sum_{i=1}^{n} K_h(x-X_i) ,$$

where here and in the following a subscript h means a rescaling of the type

$$K_h(\cdot) = h^{-1}K(\cdot/h),$$

K is called the kernel function, and the amount of smoothing is controlled by the bandwidth h. See Prakasa Rao (1983), Devroye and Gyorfi (1984), and Silverman (1986) for access to the large literature concerning $\hat{f}_h$.

The fact that $\tilde{\theta}_m$ can be improved follows from the expansion

$$\tilde{\theta}_m = n^{-1}h^{-2m-1} K^{(m)}*K^{(m)}(0) + n^{-2} \sum\sum_{i\neq j} K_h^{(m)}*K_h^{(m)}(X_i-X_j) , \tag{2.1}$$

where * denotes convolution. Note that the first term does not make use of the data, and hence may be thought of as adding a type of bias in the estimator. This motivates the estimator

$$\hat{\theta}_m = n^{-1}(n-1)^{-1} \sum\sum_{i\neq j} K_h^{(m)}*K_h^{(m)} (X_i-X_j).$$

The convergence rate methods described in Section 3 can be used to show that the bias introduced by the first term in (2.1) can actually dominate the mean squared error, and so only $\hat{\theta}_m$ is treated here. The squared-error rate of convergence of $\hat{\theta}_m$ is never inferior to that of $\tilde{\theta}_m$.

Another estimate of θ_m is motivated by the fact that, under strong enough conditions,

$$\theta_m = (-1)^m \int f^{(2m)}(x)f(x)dx,$$

which can be estimated by

$$\tilde{\tilde{\theta}}_m = (-1)^m n^{-1} \sum_{i=1}^{n} \hat{f}^{(2m)}(X_i).$$

The same argument used above to motivate $\hat{\theta}_m$ can be employed to show that a better version of $\tilde{\tilde{\theta}}_m$ is

$$\hat{\hat{\theta}}_m = (-1)^m n^{-1}(n-1)^{-1} \sum_{i \neq j} \sum K_h^{(2m)}(X_i - X_j).$$

At first glance it might seem that $\hat{\hat{\theta}}_m$ will be inferior to $\hat{\theta}_m$, since 2m derivatives of f appear to be used in the motivation of $\hat{\hat{\theta}}_m$ while only m derivatives appear in $\hat{\theta}_m$. The fact that this is not the case is demonstrated in Section 3, where it is seen that the two estimators have very similar properties, even when f has fewer than 2m derivatives. Some idea of why this is the case is given by writing

$$\hat{\theta}_m = n^{-1}(n-1)^{-1} h^{-2m-1}(-1)^m \sum_{i \neq j} \sum K^{(2m)} * K\{(X_i - X_j)/h\},$$

$$\hat{\hat{\theta}}_m = n^{-1}(n-1)^{-1} h^{-2m-1}(-1)^m \sum_{i \neq j} \sum K^{(2m)}\{(X_i - X_j)/h\}.$$

3. Rates of Convergence

In ordinary kernel density estimtion, the rate of convergence is typically determined either by the smoothness of the underlying density or by the order of the kernel function.

The density f will be said to have smoothness of order $p > 0$ whenever there is a constant $M > 0$ so that, for all real x and y,

$$|f^{(\ell)}(x)-f^{(\ell)}(y)) \leq M|x-y|^{\alpha}, \tag{3.1}$$

where $p = \ell+\alpha$ and $0 < \alpha \leq 1$.

The kernel function K will be said to have order k when

$$\int x^j K(x)dx = \begin{cases} 1 & j=0 \\ 0 & j=1,\ldots,k-1. \\ C & j=k \end{cases}$$

For simplicity of presentation, we also assume that K is symmetric and has 2m derivatives which vanish at $\pm\infty$. Results similar to those of this paper can be obtained with more effort under weaker smoothness assumptions on k.

The mean squared error of an estimator may be decomposed into variance and squared bias components, which are often treated separately. Asymptotic representations for the various pieces in the present context depend on the relationship between m,p and k, as we now show.

Let $\circledast_m$ denote either $\hat{\theta}_m$ or $\hat{\hat{\theta}}_m$, and put $\kappa_m \equiv \int (K^{(2m)})^2$ and $\kappa(u) \equiv (K*K)(u)$ if $\circledast_m = \hat{\theta}_m$, and $\kappa_m = \int (K^{(2m)}*K)^2$ and $\kappa(u)$ if $\circledast_m = \hat{\hat{\theta}}_m$. Variance and bias of $\circledast_m$ are described by:

Lemma 3.1: If f has smoothness of degree $p > m$ and K has order k, then as $n\to\infty$ and $h\to 0$ with $nh \to \infty$,

(a) for $p > 2m$,

$$\mathrm{var}(\circledast_m) = n^{-2} h^{-4m-1} 2(\int f^2)\kappa_m + 4n^{-1}\{\int (f^{(2m)})^2 - \theta_m^2\}$$

$$+ o(n^{-2} h^{-4m-1} + n^{-1})$$

(b) <u>for</u> $p \leq 2m$,

$$\mathrm{var}(\hat{\theta}_m) = n^{-2} h^{-4m-1} 2(\int f^2)\kappa_m + O(n^{-1} h^{-4m+2p}) + o(n^{-2} h^{-4m-1})$$

(c) <u>for</u> $p > k+m$,

$$\{E(\hat{\theta}_m) - \theta_m\}^2 = h^{2k}(k!)^{-2} \{\int u^k \kappa(u)du\}^2 \{\int f^{(m)} f^{(m+k)}\}^2 + o(h^{2k})$$

(d) <u>for</u> $p \leq k+m$,

$$\{E(\hat{\theta}_m) - \theta_m\}^2 = O(h^{2(p-m)}).$$

The proof of Lemma 3.1 is in the appendix.

The various special cases appearing in Lemma 3.1 may be combined into a general mean squared error result if we introduce the notation

$$\nu = \min(p-m, k).$$

Most cases allow statements only about the best exponent of convergence. These are summarized in:

<u>Theorem 3.2</u>: <u>Under the assumptions of Lemma 3.1</u>,

(a) <u>when</u> $\nu \leq 2m + \frac{1}{2}$,

$$E(\hat{\theta}_m - \theta_m)^2 = O(n^{-4\nu/2(2\nu+4m+1)})$$

<u>by taking</u> $h = O(n^{-2/(2\nu+4m+1)})$.

(b) <u>when</u> $\nu > 2m + \frac{1}{2}$,

$$E(\hat{\theta}_m - \theta_m)^2 = O(n^{-1})$$

<u>by taking</u> $h \in [n^{-1/(4m+1)}, n^{-1/2\nu}]$.

When both k and p are sufficiently large, not only the best exponent of convergence, but also the best constants, may be given. First define

$$c_2 = 2(\int f^2)\kappa_m,$$

$$c_2 = (k!)^{-2} \{\int u^k \kappa(u)du\}^2 (\int f^{(m)} f^{(m+k)})^2.$$

Theorem 3.3: Under the assumptions of Lemma 3.1, minimum mean squared errors are achieved as follows:

(a) when $k < 2m + \frac{1}{2}$ and $k < p-m$,

$$E(\hat{\theta}_m - \theta_m)^2 = \frac{(2k+4m+1)C_2}{4m+1} \left\{\frac{(4m+1)C_1 n^{-2}}{2k\, C_2}\right\}^{2k/(4m+2k+1)} + o(n^{-4k/(4m+2k+1)}),$$

by taking

$$h = \left\{\frac{(4m+1)\, C_1 n^{-2}}{2k\, C_2}\right\}^{1/(4m+2k+1)} + o(n^{-1/(4m+2k+1)})$$

(b) when $\nu > 2m + \frac{1}{2}$,

$$E(\hat{\theta}_m - \theta_m)^2 = 4\left\{\int (f^{(2m)})^2 - \theta_m^2\right\} n^{-1} + o(n^{-1}),$$

by taking any h which satisfies

$$h\, n^{1/(4m+1)} \to \infty, \quad h\, n^{1/2\nu} \to 0.$$

The proofs of Theorems 3.2 and 3.3 are immediate from Lemma 3.1. Note that there are a number of "boundary cases", such as $k = 2m + \frac{1}{2}$, that are not explicitly stated here, but may be handled with no additional work.

4. Discussion

Remark 4.1: For rate of convergence results which include some special cases of those presented here, see Schweder (1975) and Sheather and Hettsmansperger (1987). These papers also treat the important problem of how to choose the bandwidth, h.

Remark 4.2: A very important question is: are the rates obtained in Theorem 3.2 the best possible? We conjecture that they are, in the sense of Farrell (1972) and Stone (1980, 1982). In some as yet unpublished work in a closely related setting, L. Goldstein and K. Messer have established some interesting results of this type. Unfortunately that work does not extend to our case.

Remark 4.3: When $\nu > 2m + \frac{1}{2}$, Theorem 3.3 still leaves a good deal of room for choice of h. A slight extension of the expansion of Lemma 3.1 can be used to develop a second order optimality theory of the type sometimes called "deficiency". See Marron and Sheather (1987) for an account of the literature on this subject in the context of quantile estimation.

Remark 4.4: Another natural question is: how do the estimators $\hat{\theta}_m$ and $\hat{\hat{\theta}}_m$ compare? It is easily seen that

$$\int u^k K*K(u)du = 2\int u^k K(u)du$$

$$\int [K^{(2m)}*K]^2 \leq \int [K^{(2m)}]^2.$$

Hence $\hat{\theta}_m$ has smaller variance and $\hat{\hat{\theta}}_m$ has less bias. A means of comparison is to look at the minimum mean square error as given in (a) of Theorem 3.3. Note that C_1 and C_2 appear as a weighted geometric mean, so the question of which of $\hat{\theta}_m$ and $\hat{\hat{\theta}}_m$ is better can only be resolved for each specific K.

Remark 4.4: Lemma 3.1 can also be used to obtain a theory for optimal choice of

K, such as the one studied by Epanechnikov (1969) and Gasser, Müller and Mammitzsch (1985). Note that the answer here is the same as that of Epanechnikov in the case of $\hat{\theta}_0$.

Remark 4.5: It is completely straightforward to extend the results of this paper to the case where f(x) is a density on $\mathbb{R}^d$. For clarity of presentation, this case is not explicitly treated here.

Remark 4.6: Theorem 3.2 has important implications for automatic bandwidth selection of an ordinary kernel density estimator. Hall and Marron (1987a) have shown that, if $\hat{h}_c$ is the bandwidth chosen by least squares cross-validation, then for k=2 and $p \geq 2$,

$$(\hat{h}_c - h_0)/h_0 \sim n^{-1/10},$$

where h_0 is the bandwidth which minimizes mean integrated squared error. Scott and Terrell (1986) have proposed another bandwidth selector which gives similar performance when k=2 and $p \geq 4$.

Hall and Marron (1987b) describe a sense in which the rate $n^{-1/10}$ is the best possible for p essentially no bigger than 2. When k=2 and $p > 2$

$$h_0 \sim h_0^* = n^{-1/5}[\int K^2 \{\int x^2 K(x)dx\}^{-2} \theta_2^{-1}]^{1/5}.$$

See Rosenblatt (1971), for example.

This motivates using the bandwidth

$$\hat{h} = n^{-1/5} [\int K^2 \{\int x^2 K(x)dx\}^{-2} \circledast_2^{-1}]^{1/5},$$

when $\circledast_2$ is either $\hat{\theta}_2$ or $\hat{\hat{\theta}}_2$. To compare this with $\hat{h}_c$, note that, by Theorem 3.2, for properly chosen h and k sufficiently large,

$$(\hat{h} - h_0^*)/h_0^* \sim \begin{cases} n^{-2(p-2)/2p+5} & p \leq 6.5 \\ n^{-\frac{1}{2}} & p > 6.5 \end{cases}.$$

Thus, if we ignore the difference between h_0^* and h_0, $\hat{h}$ is better than $\hat{h}_c$ for $p \geq 2.5$. However, the important feature of this observation is not so much the accuracy confirmed by the faster rate of convergence, but the stability. The plug-in bandwidth,with a relative error of $n^{-\frac{1}{2}}$, is much more robust against sampling fluctuations than is the cross-validatory bandwidth with an error of $n^{-1/10}$.

Appendix

Proof of Lemma 3.1: First consider the bias. Note that

$$E(\hat{\theta}_m) = h^{-2m-1}(-1)^m \int\int K^{(2m)}\{(x-y)/h\}f(x)f(y)dx\,dy$$

$$= h^{-2m}(-1)^m \int\int K^{(2m)}(u)f(x)f(x-hu)dx\,du$$

$$= h^{-m}(-1)^m \int\int K^{(m)}(u)f(x)f^{(m)}(x-hu)dx\,du$$

$$= h^{-m} \int\int K^{(m)}(u)f^{(m)}(x)f(x-hu)dx\,du$$

$$= \int\int K(u)f^{(m)}(x)f^{(m)}(x-hu)dx\,du.$$

Similarly,

$$E(\hat{\theta}_m) = \int\int K*K(u)f^{(m)}(x)f^{(m)}(x-hu)dx\,du.$$

Part (c) now follows by a k-th order Taylor expansion of $f^{(m)}(x-hu)$, together with the fact that both K and K*K are of order k. Part (d) follows by an $(m-\ell)$-th order Taylor expansion of $f^{(m)}(x-hu)$ together with the Lipschitz condition (3.1).

For the variance component, note that

$$\mathrm{var}(\hat{\theta}) = n^{-2}(n-1)^{-2}\, h^{-4m-2} \sum_{i\neq j}\sum \sum_{i'\neq j'}\sum \mathrm{cov}[K^{(2m)}\{(X_i-X_j)/h\},\ K^{(2m)}\{(X_{i'}-X_{j'})/h\}]$$

$$\text{(A.1)} \qquad = \{2n^{-2} + o(n^{-2})\}h^{-4m-2}\ \mathrm{var}[K^{(2m)}\{(X_1-X_2)/h\}]$$

$$+ \{4n^{-1} + o(n^{-1})\}h^{-4m-2}\ \mathrm{cov}[K^{(2m)}\{(X_1-X_2)/h\},\ K^{(2m)}\{(X_2-X_3)/h\}].$$

But

$$E[h^{-2m-1} K^{(2m)}\{(X_1-X_2)/h\}]^2 = h^{-4m-1} \iint \{K^{(2m)}(u)\}^2 f(x)f(x-hu)dx\,du$$

$$= h^{-4m-1} (\int f^2)\{\int (K^{(2m)})^2\} + o(h^{-4m-1}),$$

and for $p \geq 2m$,

$$E[h^{-4m-2} K^{(2m)}\{(X_1-X_2)/h\} K^{(2m)}\{(X_2-X_3)/h\}] =$$

$$= \iiint h^{-4m} K^{(2m)}(u)K^{(2m)}(v)f(y+hu)f(y)f(y-hv)du\,dy\,dv \tag{A.2}$$

$$= (-1)^m \iiint K(u)K(v)f^{(2m)}(y+hu)f(y)f^{(2m)}(y-hv)du\,dy\,dv$$

$$= (-1)^m \int (f^{(2m)})^2 f + o(1)$$

Hence, since $E[h^{-2m-1} K^{(2m)}\{(X_1-X_2)/h\}] \to (-1)^m \theta_m$,

$$\text{var}[h^{-2m-1}K^{(2m)}\{(X_1-X_2)/h\}] = h^{-4m-1}(\int f^2)\{\int (K^{(2m)})^2\} + o(h^{-4m-1}),$$

and for $p \geq 2m$,

$$h^{-4m-2} \text{cov}[K^{(2m)}\{(X_1-X_2)/h\}, K^{(2m)}\{(X_2-X_3)/h\}] = \{\int (f^{(2m)})^2 f - \theta_m^2\} + o(1).$$

For the estimator $\hat{\theta}_m$, part (a) now follows from (A.1). To modify this argument for part (b), the only change required is in (A.2), where fewer integrations by parts should be done and the Lipschitz condition (3.1) is again applied. The proof for $\hat{\hat{\theta}}_m$ is entirely similar.

References

Devroye, L. and Györfi, L. (1984). *Nonparametric Density Estimation: The L_1 View*. Wiley, New York.

Farrell, R. H. (1972), "On the best obtainable rates of convergence in estimation of a density function at a point," *Annals of Mathematical Statistics*, 43, 170-180.

Gasser, T., Müller, H. G. and Mammitzsch, V. (1985), "Kernels for nonparametric curve estimation," *Journal of the Royal Statistical Society, Series B*, 47, 238-252.

Hall, P. and Marron, J. S. (1987a), "Extent to which least-squares cross-validation minimises integrated square error in nonparametric density estimation," *Theory of Probability and Related Fields*, to appear.

Hall, P. and Marron, J. S. (1987b), "The amount of noise inherent in bandwidth selection for a kernel density estimator, " *Annals of Statistics*, to appear.

Hall, P. and Marron, J. S. (1986a), "Choice of kernel order in density estimation", unpublished manuscript.

Hall, P. and Marron, J. S. (1986b), "Variable window width kernel estimates of probability densities", unpublished manuscript.

Marron, J. S. and Sheather, S. (1987), "Kernel Quantile Estimators", manuscript in preparation.

Parzen, E. (1962), "On estimation of a probability density function and mode," *Annals of Mathematical Statistics*, 33, 1065-1076.

Prakasa Rao, B. L. S. (1983), *Nonparametric Functional Estimation*, Academic Press, New York.

Rosenblatt, M. (1956), "Remarks on some non-parametric estimates of a density function," *Annals of Mathematical Statistics*, 27, 832-837.

Rosenblatt, M. (1971), "Curve estimates," *Annals of Mathematical Statistics*, 42, 1815-1842.

Schweder, T. (1975). "Window estimation of the asymptotic variance of rank estimators of location", *Scandinavian Journal of Statistics*.

Scott, D. J. and Terrell, G. R. (1986). "Biased and unbiased cross-validation in density estimation," Rice University Tech. Report No. 87-02.

Sheather, S. J. and Hettsmansperger, T. P. (1987) "A data-based algorithm for choosing the window width when estimating the integral of $f^2(x)$", unpublished manuscript.

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

Stone, C. J. (1980), "Optimal convergence rates for nonparametric estimators," *Annals of Statistics*, 8, 1348-1360.

Stone, C. J. (1982), "Optimal global rates of convergence of nonparametric regression," *Annals of Statistics*, 10, 1040-1053.