Is there a best kernel density estimator?

Kairat Mynbaev[1,5], Saralees Nadarajah[2], Christopher Withers[3], and Aziza Aipenova[4]

[1]Kazakh-British Technical University, Almaty, Kazakhstan

[2]School of Mathematics, University of Manchester, Manchester M13 9PL, UK

[3]Applied Mathematics Group, Industrial Research Limited, Lower Hutt, New Zealand

[4]Kazakh National University, Almaty, Kazakhstan

[5]Corresponding author: Kairat Mynbaev, email: kairat_mynbayev@yahoo.com

Abstract

For the nonparametric density estimators we show that the constant c_1 in the relation $bias = c_1 h^q + o(h^q)$ can be made arbitrarily small, while keeping the variance $var = \frac{1}{nh}(c_2 + o(h))$, as measured by the constant c_2, bounded, provided that the kernels are of order q. We call a *free-lunch effect* the fact that c_1 can be made as small as desired, without increasing the density smoothness requirement or the kernel order. Another problem we consider is testing if a density satisfies a differential equation. This result can be applied to see if a density belongs to a particular family of differential equations.

Keywords: density estimation, bias, local normality test, asymptotic normality, consistency

1. Introduction

Nonparametric density estimation has been a subject of many papers, see Withers and Nadarajah [2012] for the latest references. A lot of research focused on constructing estimators that would reduce bias, relative to the ones suggested earlier. Two common observations are that (1) the smoother the density, the better the rate of convergence and (2) to avail oneself of the higher smoothness of the density, one has to use kernels with special properties, such as higher-order kernels or the ones proposed by Mynbaev and Martins Filho [2010].

Withers and Nadarajah [2012] have explored the procedure of transforming a kernel K into a higher-order kernel $T_a K$ via multiplication of K by a polynomial of order q:

$$(T_a K)(t) = \left(\sum_{i=0}^{q} a_i t^i \right) K(t)$$

with the appropriately chosen vector of coefficients $a = (a_0, ..., a_q)' \in R^{q+1}$. Using incomplete Bell polynomials, they have found a that reduces bias in comparison to that of some of existing estimators and for some main densities. The idea of transforming a kernel by multiplying it by a polynomial is simple and perhaps is a part of the statistical folklore (for example, Lejeune and Sarda [1992] obtained such a transformation from a different concept) but Withers & Nadarajah seem to have been the first to have a closer look at it. As they mention, this method is less calculation-intensive than several other methods.

The first purpose of this paper is to investigate the transformation T_a further. The outcome is surprising: the vector a can be chosen in such a way as to make the bias $bias = c_1 h^q + o(h^q)$, as measured by the constant c_1, arbitrarily small, while keeping the variance $var = \frac{1}{nh}(c_2 + o(h))$, as measured by the constant c_2, bounded, provided that $T_a K$ is of order q. We call a *free-lunch effect* the fact that c_1 can be made as small as desired, without increasing the density smoothness requirement or the kernel order. To put it differently, there is no estimator with the least bias among those which have uniformly bounded variances and are generated by higher-order kernels of fixed order. Note that c_1 can be set to zero, but then $T_a K$ becomes of order higher than q. It is also useful to bear in mind that in a sufficiently large class of estimators there are no unbiased estimators of densities [Rao, 1983, Ch. 1].

The free-lunch effect raises a natural question: as $c_1 \to 0$, could it be that the higher-order terms in h in the decompositions of the bias and variance tend to infinity? We show that this is not the case, that is, the higher-order terms can be controlled not to increase. We do this under two sets of assumptions. The first set is that the density is infinitely differentiable and all moments of K exist (as do Withers & Nadarajah) and the second is that the density has a finite number of derivatives and the kernel and its square possess a finite number of moments. In both cases we provide complete proofs. By extending our proofs for the first set of assumptions, one can justify some formal infinite decompositions from Withers and Nadarajah [2012]. For the second set we give the full proof just because there is a need to control higher-order terms, which does not seem to have been done in the literature.

Another interesting idea suggested by Withers & Nadarajah is to use a linear combination of estimators of lower-order derivatives in order to better estimate a higher-order derivative. Although the outcome (their Theorem 3.1) is weaker than our Theorems 2 and 3, the idea can be used for a different purpose, namely, for testing if a density is of a certain type. Consider, for example, the standard normal density $f(t) = (2\pi)^{-1/2}\exp(-t^2/2)$. It satisfies a differential equation $f'(t) + tf(t) = 0$. The general solution of this equation is $f(t) = c\exp(-t^2/2)$, and if it is to be a density, one has to put $c = (2\pi)^{-1/2}$. We say that a density f is *locally standard normal at point* t if it satisfies the above differential equation at that point. Thus, there is a practical need to test whether a density satisfies a certain differential equation. The second purpose of this paper is to address this need. The testing procedure is accompanied by an asymptotic normality statement. The latest references concerning normality testing include Razali and Wah [2011], Thadewald and Büning [2007], Sürücü [2008], Farrell and Rogers-Stewart [2006] and Szekely and Rizzo [2005]. According to Razali and Wah [2011], the Shapiro-Wilk test has the best power for a given level of significance, followed closely by the Anderson-Darling test. Most existing tests are based on some global properties of normal distributions. Both global and local approaches have their advantages and deficiencies. The main difference between the global and local approaches consists in the amount of calculation: rejecting normality locally is enough to reject it globally.

2. Main results

For a function K defined on R we denote

$$\alpha_j(K) = \int_R K(t)t^j dt,\ \beta_j(K) = \int_R |K(t)t^j| dt$$

its jth moment and absolute moment, respectively. Recall that K is called a *kernel of order*

q if $\alpha_0(K) = 1$, $\alpha_j(K) = 0$, $j = 2, ..., q-1$, $\alpha_q(K) \neq 0$. Everywhere we assume that the observations $X_1, ..., X_n$ are i.i.d. with density f.

The Rozenblatt-Parzen kernel estimator of $f(x)$ is defined by

$$f_h(x, K) = \frac{1}{n} \sum_{j=1}^{n} \frac{1}{h} K\left(\frac{x - X_j}{h}\right), \ h > 0.$$

If f and K are l times continuously differentiable, then it gives rise to the estimator of $f^{(l)}(x)$

$$f_h^{(l)}(x, K) = \frac{1}{n} \sum_{j=1}^{n} \frac{1}{h^{l+1}} K^{(l)}\left(\frac{x - X_j}{h}\right). \tag{1}$$

In asymptotic statements the sample size n tends to infinity and the bandwidth h depends on n but this dependence usually is not reflected in the notation. Therefore the expression of type $o(h^q)$ presumes that $n \to \infty$.

Theorem 1. *Suppose that f is infinitely differentiable and K has a continuous derivative of order l. Further assume that K and $K^{(l)}$ have absolute moments of all orders,*

$$\limsup_{j \to \infty} \left| \frac{f^{(j)}(x)}{j!} \beta_{j+1}(K) \right|^{1/j} = 0, \ \limsup_{j \to \infty} \left| \frac{f^{(j)}(x)}{j!} \beta_{j+1}(K^{(l)}) \right|^{1/j} = 0,$$

$$\left\| K^{(l)} \right\|_{C(R)} = \sup_{t \in R} \left| K^{(l)}(t) \right| < \infty.$$

Then

$$E f_h^{(l)}(x, K) = \sum_{i=0}^{\infty} \frac{f^{(i+l)}(x)}{i!} (-h)^i \alpha_i(K)$$

and

$$var\left(f_h^{(l)}(x, K)\right) = \frac{1}{n h^{2l+1}} \left\{ \sum_{i=0}^{\infty} \frac{f^{(i)}(x)}{i!} (-h)^i \alpha_i(M) - h \left[h^l E f_h^{(l)}(x, K) \right]^2 \right\}$$

where $M = \left[K^{(l)}\right]^2$ and the series converge for all $h \in R$. Consequently, if K is a kernel of order q, then

$$E f_h^{(l)}(x, K) - f^{(l)}(x) = \frac{f^{(q+l)}(x)}{q!} (-h)^q \alpha_q(K) + O(h^{q+1}), \tag{2}$$

$$var\left(f_h^{(l)}(x, K)\right) = \frac{1}{n h^{2l+1}} \left\{ f(x) \int_R M(t) dt + O(h) \right\}.$$

With the function K we can associate matrices

$$A_q(K) = \begin{pmatrix} \alpha_0(K) & \alpha_1(K) & ... & \alpha_q(K) \\ \alpha_1(K) & \alpha_2(K) & ... & \alpha_{q+1}(K) \\ ... & ... & ... & ... \\ \alpha_q(K) & \alpha_{q+1}(K) & ... & \alpha_{2q}(K) \end{pmatrix}, \ B_q = A_q(K^2).$$

In the next theorem we prove the free-lunch effect, for simplicity limiting ourselves to estimation of $f(x)$.

Theorem 2. *Suppose that f and K are continuous, $\|K\|_{C(R)} < \infty$,*

$$\limsup_{j\to\infty} \left| \frac{f^{(j)}(x)}{j!} \beta_{q+j+1}(K) \right|^{1/j} = 0,$$

$$\det A_q(K) \neq 0. \tag{3}$$

Let a vector $b \in R^{q+1}$ have components $b_0 = 1$, $b_1 = ... = b_{q-1} = 0$, $b_q \neq 0$ and set $a = A_q(K)^{-1}b$. Then

$$Ef_h(x, T_aK) - f(x) = \frac{f^{(q)}(x)}{q!}(-h)^q b_q + O(h^{q+1}), \tag{4}$$

$$var\left(f_h(x, T_aK)\right) = \frac{1}{nh}\left\{f(x)b'C_qb + O(h)\right\} \tag{5}$$

where $C_q = [A_q(K)^{-1}]'B_qA_q(K)^{-1}$ and $b'C_qb > 0$. The terms of higher order in h in (4) *and* (5) *retain their magnitude as $b_q \to 0$.*

Corollary 1. *Denote the elements of $A_q(K)^{-1}$ by A_q^{ij}, $i, j = 0, ..., q$, $c = (1, 0, ..., 0)' \in R^{q+1}$ and $d = (0, ..., 0, b_q)' \in R^{q+1}$. Then $b = c + d$. As $b_q \to 0$,*

$$\begin{aligned} (T_aK)(t) &\to \left(\sum_{i=0}^{q} A_q^{i,1} t^i\right) K(t), \\ b'C_qb &= c'C_qc + O(b_q) \to (C_q)_{11} = \sum_{i,j} A_q^{1i}(B_q)_{ij}A_q^{j1} \end{aligned}$$

It follows that in (4) *b_q can be made as small as desired, while* (5) *retains its magnitude as we do this.*

Remark 1. One can show that B_q is positive definite and (3) holds if K is nonnegative.

In the next theorem we give conditions sufficient for the free-lunch effect when f is not infinitely differentiable and K does not possess moments of all orders.

Theorem 3. *Suppose that* (3) *holds, f is $(q+1)$-times continuously differentiable, $\|f'\|_{C(R)} + \left\|f^{(q+1)}\right\|_{C(R)} < \infty$ and $\beta_{2q+1}(K) + \beta_{2q+1}(K^2) < \infty$. Then* (4) *and* (5) *are true.*

Now we turn to the second subject of this paper: testing for local normality. More generally, consider the expression $F(x) = \sum_{l=0}^{L} g_l(x)f^{(l)}(x)$ where $\{g_l(x)\}$ are given functions and the senior coefficient g_L is different from zero at the given point x. We can test the null hypothesis $H_0 : f$ satisfies the equation $F(x) = 0$ against the alternative hypothesis $H_a : F(x) \neq 0$. It is convenient to use the differential operator D defined by $(Df)(x) = F(x)$. Since the derivative $f^{(l)}(x)$ is estimated by (1), it is natural to estimate $F(x)$ by

$$\hat{F}_h(x) = \sum_{l=0}^{L} g_l(x)f_h^{(l)}(x, K) = \frac{1}{n}\sum_{j=1}^{n}\sum_{l=0}^{L} \frac{g_l(x)}{h^{l+1}} K^{(l)}\left(\frac{x - X_j}{h}\right). \tag{6}$$

As one can see from part (a) of the next theorem, under the null hypothesis it also makes sense to consider the random variable $\hat{G}_h(x) = \hat{F}_h(x)/h$. Provided that

$$f(x)g_L(x) \neq 0 \tag{7}$$

let Ψ denote a normal variable distributed as $N\left(0, f(x)\alpha_0\left(\left[g_L(x)K^{(L)}\right]^2\right)\right)$.

Theorem 4. *Assume that f is infinitely differentiable and K has L continuous derivatives. Suppose that K is of order 1 and that*

$$\max_{l=0,\dots,L} \limsup_{j\to\infty} \left| \frac{f^{(j)}(x)}{j!} \beta_{j+1}(K^{(l)}) \right|^{1/j} = 0, \ \max_{l=0,\dots,L} \left\| K^{(l)} \right\|_{C(R)} < \infty.$$

Then the following statements are true:

(a) The bias of (6) *is given by $E\hat{F}_h(x) - F(x) = -h\alpha_1(K)(Df')(x) + O(h^2)$. Consequently, under H_0*

$$E\hat{F}_h(x) = O(h). \tag{8}$$

If, however, $E\hat{F}_h(x) \to const \neq 0$, as $h \to 0$, then $F(x) \neq 0$ and H_0 can be rejected.

(b) If $nh^{2L+1} \to \infty$ and (7) *holds, then under the null $\operatorname{plim} \hat{F}_h(x) = 0$ (this equation is preferable to* (8) *because in practice $E\hat{F}_h(x)$ is unknown).*

(c) If $nh \to \infty$ and (7) *holds, then $(nh^{2L+1})^{1/2} \left[\hat{F}_h(x) - E\hat{F}_h(x) \right] \xrightarrow{d} \Psi$. If, in addition, $nh^{2L+3} \to 0$ then $(nh^{2L+1})^{1/2} \left[\hat{F}_h(x) - F(x) \right] \xrightarrow{d} \Psi$.*

(d) If $nh \to \infty$, $nh^{2L+3} \to 0$ and (7) *holds, then under the null $(nh^{2L+3})^{1/2} \hat{G}_h(x) \xrightarrow{d} \Psi$.*

3. Monte Carlo simulations

The main point of Theorem 2 is that the asymptotic bias is regulated by the constant b_q in (4). This point has been sufficiently illustrated by simulations in Withers and Nadarajah [2012], although the construction of their estimator is more complex than ours. Controlling the variance of their estimator was not their direct purpose but they showed that the mean squared errors did not increase when n increased. Owing to the general identity $E(X - c)^2 = var(X) + [Bias(X)]^2$, this confirms that the variance in (5) retains its magnitude. Thus we do not need to illustrate the free lunch effect on the computer.

Now we report the results of simulations for Theorem 4. We have chosen two densities to test: the standard normal, which satisfies the equation $f'(t) + tf(t) = 0$, and the Cauchy density, which does not satisfy that equation. We have selected the kernel $K \sim N(0.1, 1)$. This kernel is of order 1, as required in Theorem 4 where $L = 1$. A nonzero first moment $\alpha_1(K)$ increases bias, as seen from (2), and our simulations show that this increase is significant when, for example, the kernel $K \sim N(1, 1)$ is used. Using values $\alpha_1(K)$ positive and smaller than 0.1 did not improve the estimation.

We have experimented with sample sizes ranging from 1000 to 200,000. Increasing the sample size beyond 100,000 had little effect on the fit, as measured by the Root Average Squared Error (RASE). The RASE decreased from 0.025 to 0.006, as the sample size increased, for both densities considered. Since the idea behind estimating a differential equation is to combine the estimates of the density and its derivatives, the bandwidth should provide a good estimation of the density and its derivatives in the first place. This is how the bandwidths were chosen. They ranged between 0.15 and 0.35, slightly decreased when the sample size increased and were a little smaller for the Cauchy density than for the standard normal.

In case of the standard normal density the estimator from Theorem 4 estimates $g(t) = f'(t) + tf(t)$, which is zero. As the sample size increases, the maximum absolute value of the estimator decreases from 0.039 to 0.017, with little change thereafter. The main sign that zero is being estimated is that the estimator behaves erratically, without approaching

any particular shape with the increase in the sample size. On the other hand, in case of the Cauchy density the function $g(t)$ is not identically zero. Naturally, the estimator approaches its shape. Overall, from our simulations we derive the following conclusions. In practice, one often cannot experiment with sample sizes. As usual, it is easier to reject the null rather than accept it. If the estimator seems to approach a particular shape, then it cannot tend to zero and the null hypothesis should be rejected.

Overall, from our simulations we derive the following conclusions. The bias is indeed regulated by the moment $\alpha_1(K)$. In theory it can be made as small as desired but in practice, starting from some point, the noise in the data dominates this effect. As usual, it is easier to reject the null rather than accept it. If the estimator seems to approach a particular shape, then it cannot tend to zero and the null hypothesis should be rejected.

References

Withers, C. S. and Nadarajah, S. (2012) "Density estimates of low bias," *Metrika*. Published online: 31 March 2012. DOI 10.1007/s00184-012-0392-x

Mynbaev, K. T. and Martins Filho, C. (2010) "Bias reduction in kernel density estimation via Lipschitz conditions," *Journal of Nonparametric Statistics*, 22, 219-235.

Lejeune, M. and Sarda, P. (1992) "Smooth estimators of distribution and density functions," *Computational Statistics and Data Analysis*, 14, 457–471.

Razali, N. and Wah, Y. B. (2011) "Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests," *Journal of Statistical Modeling and Analytics*, 2 (1), 21–33.

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

Thadewald, T. and Büning, H. (2007) "Jarque–Bera Test and its Competitors for Testing Normality – A Power Comparison," *Journal of Applied Statistics*, 34 (1), 87–105.

Sürücü, Barış (2008) "A power comparison and simulation study of goodness-of-fit tests," *Computers and Mathematics with Applications*, 56 (6), 1617–1625.

Farrell, P. J. and Rogers-Stewart, K. (2006) "Comprehensive study of tests for normality and symmetry: extending the Spiegelhalter test," *Journal of Statistical Computation and Simulation*, 76(9), 803 – 816.

Szekely, G. J. and Rizzo, M. L. (2005) "A new test for multivariate normality," *Journal of Multivariate Analysis*, 93, 58–80.