Noname manuscript No.
(will be inserted by the editor)

Tuning selection for two-scale kernel density estimators

**Xinyang Yu [1] · Cheng Wang [2] ·
Zhongqing Yang[1] · Binyan Jiang[1]**

Received: date / Accepted: date

Abstract Reducing the bias of kernel density estimators has been a classical topic in nonparametric statistics. Schucany and Sommers (1977) proposed a two-scale estimator which cancelled the lower order bias by subtracting an additional kernel density estimator with a different scale of bandwidth. Different from existing literatures that treat the scale parameter in the two-scale estimator as a static global parameter, in this paper we consider an adaptive scale (i.e., dependent on the data point) so that the theoretical mean squared error can be further reduced. Practically, both the bandwidth and the scale parameter would require tuning, using for example, cross validation. By minimizing the point-wise mean squared error, we derive an approximate equation for the optimal scale parameter, and correspondingly propose to determine the scale parameter by solving an estimated equation. As a result, the only parameter that requires tuning using cross validation is the bandwidth. Point-wise consistency of the proposed estimator for the optimal scale is established with further discussions. The promising performance of the two-scale estimator based on the adaptive variable scale is illustrated via numerical studies on density functions with different shapes.

Keywords Bias reduction · Kernel density estimation · Point-wise estimator · Tuning parameter selection

✉Binyan Jiang
by.jiang@polyu.edu.hk

[1] Department of Applied Mathematics, Hong Kong Polytechnic University
[2] School of Mathematical Sciences, MOE-LSC, Shanghai Jiao Tong University

Yu and Wang contributed equally to this work and should be considered co-first authors.

1 Introduction

Kernel density estimation which adopts a nonparametric way to estimate the probability density function of a random variable is a fundamental topic in statistics. Let $X_1, \cdots, X_n$ be univariate independent and identically distributed samples drawn from some distribution with an unknown density $f(x)$. The traditional kernel density estimator of $f(x)$ for a given bandwidth h is defined as:

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

where $K(u)$ is a kernel function, and h is a bandwidth parameter that controls the smoothness of the estimator. Based on a Taylor series approximation and some smoothness assumptions, we can show that the bias of this classical kernel density estimator is of order $O(h^2)$. The optimal bandwidth can be correspondingly determined by minimizing the Mean Squared Error (MSE). The application of the classical kernel density estimator in equation (1) and related theories are extensive and more recent applications can be found in the context of bias-reduced local linear regression (Yao, 2012), dynamic networks (Kolar et al., 2010), dynamic graphical model (Chen and Leng, 2016), and Bayes' classifier (Jiang et al., 2020), among others.

Various approaches have been proposed to improve the classical KDE defined in equation (1). One of the classical methods is to use a variable bandwidth, which improves $\hat{f}_h(x)$ by introducing a varying bandwidth instead of treating the bandwidth as a global and static parameter. As pointed out in Terrell and Scott (1992), there are two existing approaches for using a variable bandwidth. One of the approaches is to formulate the bandwidth h as $h(X_i)$ in equation (1) (i.e., the bandwidth is set to be dependent on the samples only), and the resulting estimator is of the form: $\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h(X_i)}K\left(\frac{X_i - x}{h(X_i)}\right)$. Breiman et al. (1977) and Silverman (1986) proposed to set $h(X_i) = hd_i$, where d_i depends on the distance between X_i and its kth nearest neighbors. Abramson (1982) found that the bias can be reduced to $O(h^4)$ by taking $h(X_i)$ proportional to $f(X_i)^{-1/2}$. Hall (1990) presented an easy-to-use variable bandwidth estimator based on a non-variable bandwidth chosen by cross-validation, to eliminate algebraic difficulties. Jones et al. (1995) considered a bias reduction two-stage multiplicative method. The other approach is the so called "balloon estimator" (Loftsgaarden et al., 1965) which formulates h to be dynamic in x. Specifically, the balloon estimator is given as

$$\hat{f}_1(x) = \frac{1}{nh(x)}\sum_{i=1}^{n} K\left(\frac{X_i - x}{h(x)}\right). \tag{2}$$

As suggested by the studies of Mack and Rosenblatt (1979), Tukey and Tukey (1981), and Terrell and Scott (1992), estimators in the form of equation (2) have the advantage of having a straightforward asymptotic analysis and improve the fixed kernel in some cases.

In this paper, we shall focus on the bias-reduced kernel estimator obtained using the higher-order kernel proposed by Schucany and Sommers (1977). Compared with the variable bandwidths methods, the bias-reduced kernel estimator has a simpler form and the introduction of an additional scale parameter brings an extra layer of flexibility in estimation. Specifically, Schucany and Sommers (1977) suggested the use of the following kernel function:

$$K^*_{a,h}(x) = \frac{K_{1,h}(x) - RK_{2,ah}(x)}{1 - R},$$

which reduces the estimation bias to $O(h^4)$ by choosing $R = a^{-2}$. Here K_1 and K_2 are two kernel functions and $K_{i,h}(\cdot) = \frac{1}{h}K_i(\cdot/h)$, $i = 1, 2$. When K_1 and K_2 are set to be the same, i.e., $K_1 = K_2 = K$, the resulting kernel estimator is given as

$$\hat{f}_{a,h}(x) = \frac{\hat{f}_h(x) - a^{-2}\hat{f}_{ah}(x)}{1 - a^{-2}}, \tag{3}$$

where $\hat{f}_h(x)$ and $\hat{f}_{ah}(x)$ are defined as in equation (1) with bandwidths h and ah, respectively. As we can see in the proof of Proposition 1, the lower order bias of $\hat{f}_h(x)$ is equal to a^{-2} times of the bias of $\hat{f}_{ah}(x)$. The reduction of bias for the estimator (3) is essentially achieved by cancelling the lower order bias of the two estimators $\hat{f}_h(x)$ and $\hat{f}_{ah}(x)$, i.e., KDE estimators evaluated at two different scale of bandwidths (i.e., h and ah). Through this paper, we shall call $\hat{f}_{a,h}(x)$ the two-scale estimator with scale parameter a and bandwidth h.

Schucany (1989) propose to choose a by minimizing $a\int(K^*_{a,h}(x))^2dx$ which is proportional to the Mean Integrated Squared Error (MISE). Following this method, when $K_1 = K_2$ is the standard normal density function, Wand and Schucany (1990) showed that the minimum of MISE can be obtained as $a \to 1$. Jones and Foster (1993) greatly expanded the work of Schucany and Sommers (1977) to more generalized jackknifing methods. In all these previous works, the scale parameter a is treated as a static parameter (i.e., independent of the data point), and is estimated based on the kernel functions and the bandwidth.

While a is generally treated as a global parameter in the literature, as can be seen from Section 2.2, the MSE evaluated at different data points x is a function of x. Intuitively, similar to the "balloon“ type estimators (Loftsgaarden et al., 1965), the MISE can be further reduced if we set the scale parameter a to be dependent on x. Specifically, in this paper, instead of adopting a global scale a that minimizes the MISE, we shall use different scales for different data points. To emphasize such a dependence, we shall use a_x to denote the scale parameter hereafter, and equation (3) can be rewritten as

$$\hat{f}_{a_x,h}(x) = \frac{\hat{f}_h(x) - a_x^{-2}\hat{f}_{a_xh}(x)}{1 - a_x^{-2}}. \tag{4}$$

Given a data point x and the bandwidth h, we then seek for the optimal a_x such that the MSE evaluated at x is minimized. In this paper, we shall assume that $a_x > 1$ since for any $0 < a_x < 1$, the two bandwidths $(h, a_x h)$ can be rewritten as $(a_x^{-1}h^*, h^*)$ with $h^* := a_x h$.

The rest of this paper is organized as follows. In Section 2, we discuss how to determine (a, h) and (a_x, h) defined in equations (3) and (4). In particular, we derive the point-wise MSE of $\hat{f}_{a_x,h}(x)$ and obtain an approximate equation for the optimal a_x by minimizing the MSE, and further propose to estimate a_x via an estimated equation. In Section 3, we conduct some simulation studies to further illustrate our proposed estimators, with comparison to other classical approaches. Further discussions are provided in Section 4, and all the theoretical proofs are given in the Appendix.

2 Two-scale kernel density estimator: tuning selection

We shall use Hölder classes to capture the smoothness of the density function $f(x)$. Following Tsybakov (2008), the Hölder class $\Sigma(\beta, L)$ defined on a support T is the set of $l := \lfloor \beta \rfloor$ times differentiable functions such that:

$$|f^{(l)}(x) - f^{(l)}(x')| \leq L|x - x'|^{\beta - l}, \quad \forall x, x' \in T.$$

Here $l := \lfloor \beta \rfloor$ denotes the greatest integer strictly less than the real number β. To ensure the validity of the Taylor expansions in this subsection, we shall assume that $f(x) \in \Sigma(5, L)$ for some constant $L > 0$. For the kernel function, we make the following assumptions throughout this paper:

(A1) The kernel function is bounded and symmetric in that $K(u) = K(-u)$, and $\int K(u)u^i du < \infty$ for $i = 1, 2, 3, 4$.
(A2) $\int K^{(4)}(u)^2 du < \infty$.

2.1 Classical method to choose (a, h) in equation (3)

When a_x is x-invariant, the two-scale estimator $\hat{f}_{a_x,h}(x)$ defined by equation (4) reduces to $\hat{f}_{a,h}(x)$ in equation (3). Although it is parametrized by (a, h), the estimator $\hat{f}_{a,h}(x)$ is basically constructed based on the classical kernel density estimator with two different bandwidths: h and ah. Intuitively, the two-scale estimator improves the classical KDE with the extra degree of freedom introduced by the scale parameter a. In particular, when $a \to \infty$, the two-scale estimator reduces to the classical KDE.

For the classical KDE $\hat{f}_h(x)$ defined by equation (1), the bandwidth h is usually determined using based on an unbiased least-squares estimate from leave-one-out cross-validation (UCV) (Rudemo, 1982; Bowman, 1984). Similar to how h is practically determined for $\hat{f}_h(x)$, we can choose (a, h) via UCV too. Specifically, we look for (a, h) such that the following cross-validation estimator of the risk function (up to a constant) is minimized:

$$UCV(a,h) = \int \hat{f}_{a,h}^2(x)dx - \frac{2}{n}\sum_{i=1}^{n} \hat{f}_{a,h}^{(-i)}(x_i).$$

Here $\hat{f}_{a,h}^{(-i)}(x)$ is the density estimator obtained after removing the ith sample.

2.2 Optimal a_x for a given h in equation (4)

In this section, we derive the point-wise optimal choice of a_x for a given h. The bandwidth h for estimator (1) is usually determined by minimizing the MSE. Specifically, we have

$$\text{bias}\left(\hat{f}_h(x)\right) = \text{E}\left[\hat{f}_h(x)\right] - f(x) = h^2 \frac{f^{(2)}(x)}{2!} C_0 + o\left(h^2\right),$$

where $C_0 = \int K(w)w^2 dw$, and

$$\begin{aligned}\text{Var}\left(\hat{f}_h(x)\right) &= \frac{1}{nh^2}\text{E}\left\{K\left(\frac{X_i - x}{h}\right)\right\}^2 + O\left(\frac{1}{n}\right)\\ &= \frac{f(x)\int K(w)^2 dw}{nh} + O\left(\frac{1}{n}\right).\end{aligned}$$

The optimal h is then the minimizer of the dominating terms in the MSE:

$$\begin{aligned}\text{MSE}(\hat{f}_h(x)) &= \text{bias}^2\left(\hat{f}_h(x)\right) + \text{Var}\left(\hat{f}_h(x)\right)\\ &\simeq h^4\left(\frac{f^{(2)}(x)}{2!}\right)^2 + \frac{1}{nh},\end{aligned}$$

where $a \simeq b$ means there exist constants c_1 and c_2 s.t. $c_1 < \frac{a}{b} < c_2$. Adopting the same idea, we first obtain the MSE of the two scale estimator:

Proposition 1. *Suppose assumption* (A1) *hold and assume that* $f(x) \in \Sigma(5, L)$. *We have*

$$\begin{aligned}&\text{MSE}(\hat{f}_{a_x,h}(x))\\ &= a_x^4 h^8 C_1^2 + O\left(h^9\right) + \left(1 - a_x^{-2}\right)^{-2}\left(\frac{C_2}{nh} + \frac{C_2}{na_x^5 h} + O\left(\frac{1}{na_x^3}\right) + O\left(\frac{1}{n}\right)\right),\end{aligned}$$

where $C_1 = \frac{f^{(4)}(x)}{4!}\int K(w)w^4 dw$ *and* $C_2 = f(x)\int K(w)^2 dw$.

From Proposition 1 we can obtain the optimal a_x by minimizing the dominating terms:

$$\widetilde{\text{MSE}}(\hat{f}_{a_x,h}(x)) := a_x^4 h^8 C_1^2 + \left(1 - a_x^{-2}\right)^{-2}\left(\frac{C_2}{nh} + \frac{C_2}{na_x^5 h}\right).$$

In particular, we have:

Theorem 1. *For a given h, $\widetilde{\mathrm{MSE}}(\hat{f}_{a_x,h}(x))$ is strictly convex in $a_x > 1$, and the unique minimizer a_x^* can be obtained by the root that is larger than 1 of the following equation:*

$$\frac{4a_x^5 + 5a_x^2 - 1}{(a_x^2 - 1)^3 a_x^5} = \frac{4h^9 n C_1^2}{C_2}. \tag{5}$$

Remark 1: Equation (5) is obtained by setting the first order derivative of $\widetilde{\mathrm{MSE}}(\hat{f}_{a_x,h}(x))$ to be zero. Note that when $a_x \to 1$, the left hand side of equation (5) tends to infinity, and when $a_x \to \infty$, the left hand side of equation (5) tends to zero. By continuity, for any given $n, h, C_1, C_2 > 0$, there will always exist a solution $a_x^* > 1$ such that equation (5) is satisfied. The uniqueness of a_x^* follows directly from the fact that $\widetilde{\mathrm{MSE}}(\hat{f}_{a_x,h}(x))$ is strictly convex.

Notice that C_1 and C_2 are unknown. However, we can replace C_1 and C_2 by their estimators:

$$\hat{C}_1 = \frac{\hat{f}_{h_1}^{(4)}(x)}{4!} \int K(w) w^4 dw \quad \text{and} \quad \hat{C}_2 = \hat{f}_{h_2}(x) \int K(w)^2 dw.$$

Here $\hat{f}_{h_1}^{(4)}(x) = \left(\frac{1}{nh} \sum_{i=1}^n K\left(\frac{X_i - x}{h_1}\right)\right)^{(4)}$, i.e., the 4-th derivative of $\hat{f}_{h_1}(x)$, which can be explicitly computed for a given kernel function $K(\cdot)$. Subsequently for any given h, the optimal a_x can be estimated by finding the root which is bigger than 1 of the following equation

$$\frac{4a_x^5 + 5a_x^2 - 1}{(a_x^2 - 1)^3 a_x^5} = \frac{4h^9 n \hat{C}_1^2}{\hat{C}_2}. \tag{6}$$

To solve equation (6), it is equivalent to solve the following polynomial

$$4\hat{C}_2 a_x^5 + 5\hat{C}_2 a_x^2 - \hat{C}_2 - 4h^9 n \hat{C}_1^2 (a_x^2 - 1)^3 a_x^5 = 0. \tag{7}$$

There are various root-finding algorithms for high order polynomials. In this paper, we use the classical Jenkins-Traub algorithm (Jenkins and Traub, 1972), which can be implemented using the R package "polyroot", to solve equation (7). The real root which is bigger than 1 is outputted as the targeted solution.

Denote the estimator obtained by solving equation (6) as $\hat{a}_x$, we have:

Theorem 2. *Suppose assumptions* (A1) *and* (A2) *hold, and assume that $f(x) \in \Sigma(7, L)$. For any given finite x, we have,*

$$|\hat{a}_x - a_x^*| = O_p\left(\sqrt{\frac{1}{nh_1^9}} + h_1^2 + \sqrt{\frac{1}{nh_2}} + h_2^2\right). \tag{8}$$

Further, by choosing $h \simeq O(n^{-1/9})$, $h_1 \simeq O(n^{-1/13})$ and $h_2 \simeq O(n^{-1/5})$, wehave $|\hat{a}_x - a_x^*| = O_p\left(n^{-2/13}\right)$.

The term $\sqrt{\frac{1}{nh_1^9}}+h_1^2$ in the error bound (8) comes from the estimation error of $\hat{C}_1$, and the term $\sqrt{\frac{1}{nh_2}}+h_2^2$ comes from the estimation error of $\hat{C}_2$. Note that Theorem 1 implies that for a given bandwidth h, the optimal a_x can be explicitly obtained by solving equation (5), and Theorem 2 indicates that the optimal a_x can be estimated by solving equation (6). Consequently, the choice of (a_x, h) can now be obtained via cross-validation over h. Specifically, for a given h we estimate a_x by solving equation (6) and choose h by minimizing

$$UCV(h) = \int \hat{f}^2_{\hat{a}_x,h}(x)dx - \frac{2}{n}\sum_{i=1}^{n} \hat{f}^{(-i)}_{\hat{a}_{x_i},h}(x_i).$$

We remark that the computation complexity is the same as that for the classical kernel density estimation using cross validation. We shall denote the two-scale estimator obtained via this approach as $\hat{f}_{\hat{a},h}(x)$.

3 Simulation

In this section, we conducted some numerical simulations to study the performance of the two scale estimators based on our proposed tuning selection procedures. Specifically we consider:

Method I $\hat{f}_{a,h}$: Cross-validation for both h and a using UCV. We allow $a = 1$ and set $\hat{f}_{1,h} = \hat{f}_h$.

Method II $\hat{f}_{\hat{a},h}$: Estimator of h is given by UCV and estimator of a_x is given following the method (3) above based on h as well as x. Specifically we fit the data with the classical estimator (1) first and use it to obtain $\hat{C}_1, \hat{C}_2$.

Method III $\hat{f}_{\sqrt{2},h}$: Motivated by the formation of bias-corrected Bootstrap estimator (Hall, 2013), we also consider simply setting $a = \sqrt{2}$. As a result we obtain the following estimator:

$$\hat{f}_{\sqrt{2},h}(x) = 2\hat{f}_h(x) - \hat{f}_{\sqrt{2}h}(x).$$

The bandwidth h can then be determined via UCV.

Method IV $\hat{f}_h$: the classical kernel density estimator (1) where h is determined by cross-validation.

Throughout this study, we have used the Gaussian kernel for computing all the estimators. To better understand the numerical behaviour of the estimators, we consider five different density functions with different shapes: (i) Unimodal: $N(0,1)$; (ii) Bimodal: $0.5N(-1,1)+0.5N(1,1)$; (iii) Unimodal with heavier tail: t-distribution with degree of freedom 10; (iv) $Beta(2,2)$: bounded and relatively flat; (v) $U[-5,5]$: uniform distribution with constant density

$f(x) = 0.1$ for $x \in [-5, 5]$: discrete distribution. We generate $n = 200, 400, 800$ and 1600 samples for training and compute the MSE and bias on 1000 testing points. All the results are summarized in Table 1. To understand how different choices of a and h affect the MSE, we also plot the MSE obtained by comparing with the true density, versus different bandwidths. The results with different sample sizes $100, 200, 500$ are provided in Figures 1-3, and the corresponding minimum of the MSE values under different settings are reported in Table 2. Some conclusions are summarized as fellows: Some conclusions are summarized as fellows:

(i) The two scale estimators (i.e., $\hat{f}_{a,h}$, $\hat{f}_{\hat{a},h}$ and $\hat{f}_{\sqrt{2},h}$) generally outperform the traditional KDE $\hat{f}_h$. In particular, compared with the traditional KDE $\hat{f}_h$, the estimation accuracy can be improved with a simple choice of $a = \sqrt{2}$.

(ii) From table 1, we can see that $\hat{f}_{a,h}$ and $\hat{f}_{\hat{a},h}$ have smaller MSE values under the simulated cases. Compared with the estimator $\hat{f}_{a,h}$, which conducts cross validation for both a and h, the computation of $\hat{f}_{\hat{a},h}$ is much efficient as it only requires the tuning of h. When n increases, the proposed two scale estimator with adaptive scales $\hat{f}_{\hat{a},h}$ provides increasingly improvement on MSE ratio. From the different performance under different models, we can observe that the proposed estimator $\hat{f}_{\hat{a},h}$ with adaptive scales performs better when the targeted density function is relatively far away from zero (such at the beta case). Intuitively, the estimation of the scale parameter a_x relies on the number of observations near the data point x, and hence when $f(x)$ is not too small, we would expect to have smaller estimation error.

(iii) From Figures 1 to 3, we can see that the red line, representing the MSE of $\hat{f}_{\hat{a},h}$ is generally under the curves of other estimators. This to some degree indicates that for a given bandwidth h, the adaptive scale approach does reduce the MSE, comparing with the other estimators. In particular, if a bad choice of bandwidth is chosen, $\hat{f}_{\hat{a},h}$ could be a better choice than the other three estimators. In addition, from Table 2, we can also observe that the minimums of the MSE of the two scale methods (i.e., Methods I, II and III) are comparably smaller than those of Method IV when the density function is smooth.

4 Discussions

In this paper, we propose a point-wise estimator for scale parameter a in the two-scale estimator (3) by solving an estimated equation. Unlike other studies in the literature review which treat a as static and estimate a by minimizing the MISE, our method allows the scale parameter to change adaptively in different data points. A simple estimator has been provided with theoretical results about justifications.

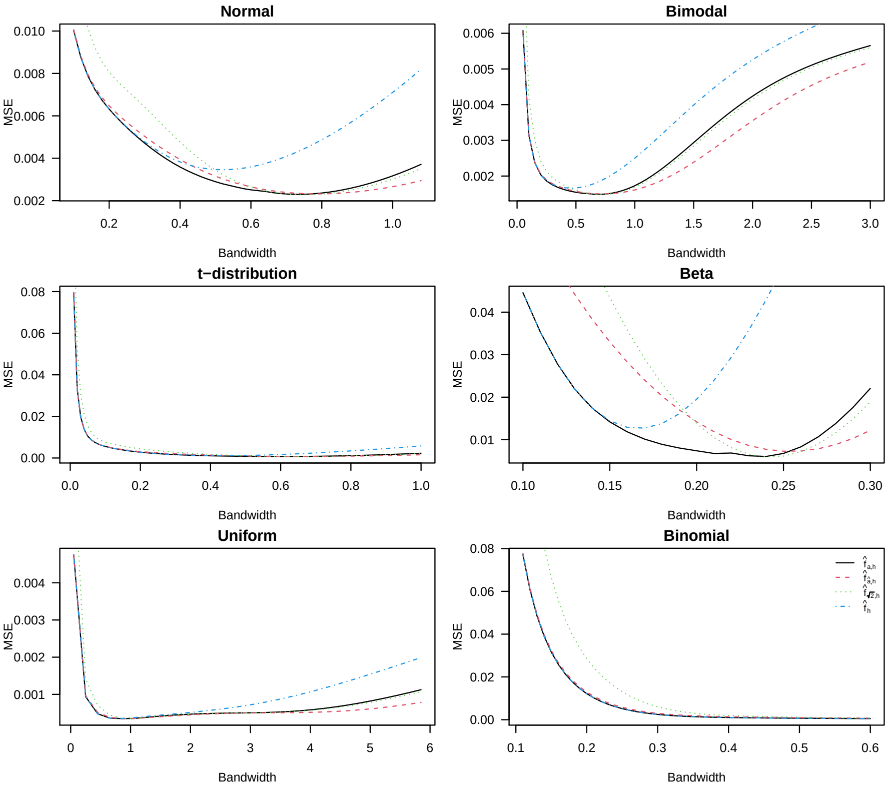

Fig. 1 MSE when sample size is 100 under six different densities.

Our idea of treating a as a function of the data point x is very similar to the balloon estimator in Terrell and Scott (1992). As reported in Terrell and Scott (1992), the balloon-type estimator could potentially outperform other estimators when estimating multivariate density. It would be interesting to explore the extension of the two-scale estimator with adaptive scale parameters to the multivariate case in our future work. Moreover, extensions of our current work to generalized jackknifing methods (Jones and Foster, 1993), asymmetric kernel estimators such as (Igarashi and Kakizawa, 2015), and Nadaraya-Watson type estimators would also be interesting topics for future study.

5 Appendix: Technical Lemmas and Proofs

Proof of Proposition 1

Proof. By Taylor expansion,

$$\mathrm{E}\left[\hat{f}_h(x) - f(x)\right]$$

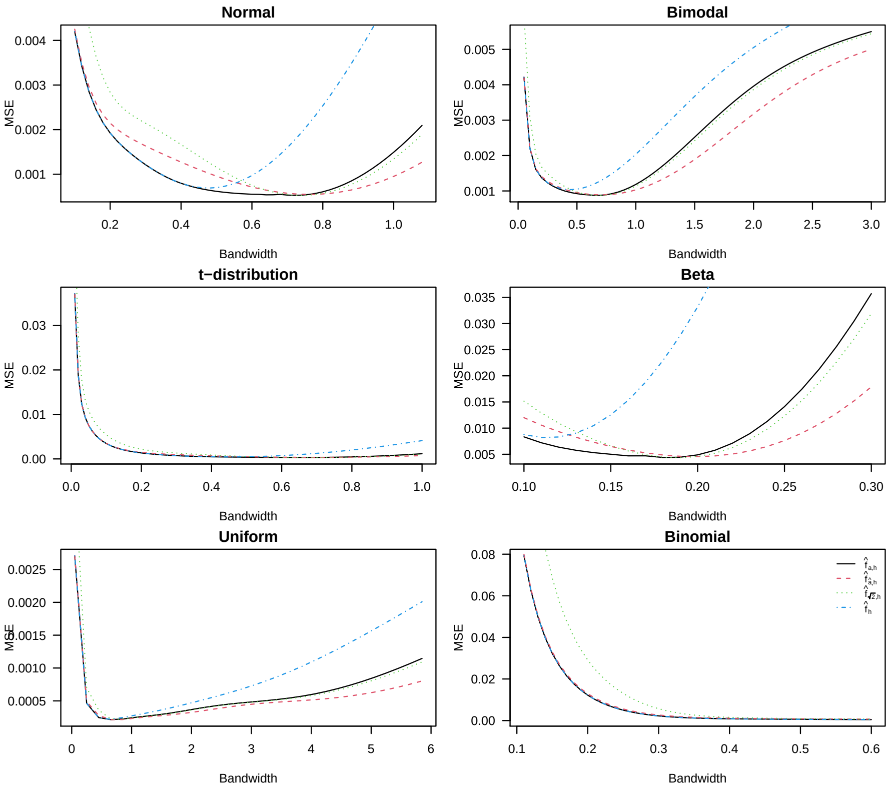

Fig. 2 MSE when sample size is 200 under six different densities.

$$
\begin{aligned}
&= \int K(w) f(wh+x) dw - f(x) \\
&= \int K(w) f^{(1)}(x) whdw + \int K(w) \frac{f^{(2)}(x)}{2!} w^2 h^2 dw \\
&\quad + \int K(w) \frac{f^{(3)}(x)}{3!} w^3 h^3 dw + \int K(w) \frac{f^{(4)}(x)}{4!} w^4 h^4 dw + o\left(h^4\right) \\
&= h^2 \int K(w) \frac{f^{(2)}(x)}{2!} w^2 dw + h^4 \int K(w) \frac{f^{(4)}(x)}{4!} w^4 dw + o\left(h^4\right).
\end{aligned}
$$

Similarly, we have

$$
\begin{aligned}
&E\left[\hat{f}_{ah}(x) - f(x)\right] \\
&= a^2 h^2 \int K(w) \frac{f^{(2)}(x)}{2!} w^2 dw + a^4 h^4 \int K(w) \frac{f^{(4)}(x)}{4!} w^4 dw + o\left(a^4 h^4\right).
\end{aligned}
$$

Consequently, we have

$$
\operatorname{bias}(\hat{f}_{a_x,h}(x)) = E\left[\frac{\hat{f}_h(x) - a^{-2}\hat{f}_{ah}(x)}{1-a^{-2}}\right] - f(x)
$$

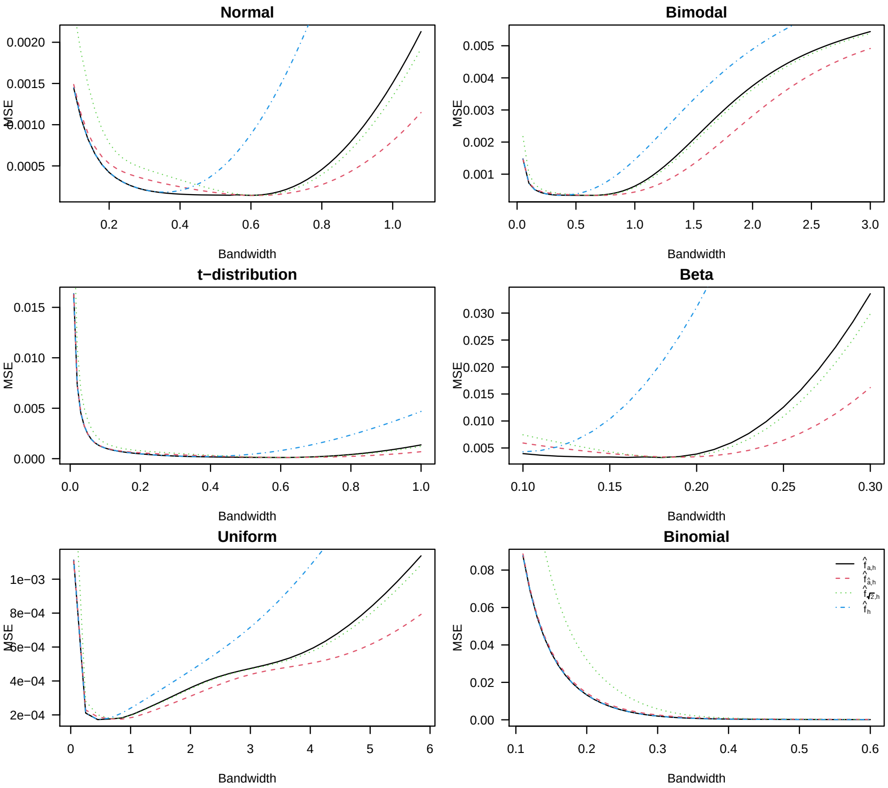

Fig. 3 MSE when sample size is 500 under six different densities.

$$
\begin{aligned}
&= \frac{\int K(w)\frac{f^{(4)}(x)}{4!}w^4h^4dw - a^{-2}\int K(w)\frac{f^{(4)}(x)}{4!}w^4(ah)^4dw}{1-a^{-2}} + O\left(h^5\right) \\
&= -a^2h^4C_1 + O\left(h^5\right),
\end{aligned}
$$

where $C_1 = \frac{f^{(4)}(x)}{4!}\int K(w)w^4dw$. Similarly, for the variance, we have

$$
\begin{aligned}
&\mathrm{Var}(\hat{f}_{a_x,h}(x)) \\
&= \left(1-a^{-2}\right)^{-2}\left(\mathrm{Var}(\hat{f}_h(x)) + a^{-4}\mathrm{Var}(\hat{f}_{ah}(x)) - 2a^{-2}\mathrm{Cov}\left(\hat{f}_h(x), \hat{f}_{ah}(x)\right)\right) \\
&= \left(1-a^{-2}\right)^{-2}\left(\frac{f(x)\int K(w)^2dw}{nh} + \frac{f(x)\int K(w)^2dw}{na^5h}\right. \\
&\left. + 2a^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{1}{n^2}\mathrm{Cov}\left(\frac{1}{h}K\left(\frac{X_i-x}{h}\right), \frac{1}{ah}K\left(\frac{X_j-x}{ah}\right)\right) + O\left(\frac{1}{n}\right)\right).
\end{aligned}
$$

Method	Normal							
	n=200		n=400		n=800		n=1600	
	MSE ratio	Bias	MSE ratio	Bias	MSE ratio	Bias	MSE ratio	Bias
$\hat{f}_{a,h}$	0.925	0.025	0.921	0.018	0.869	0.014	0.793	0.010
$\hat{f}_{\hat{a},h}$	0.904	0.025	0.889	0.018	0.759	0.013	0.715	0.010
$\hat{f}_{\sqrt{2},h}$	0.868	0.024	0.897	0.018	0.722	0.014	0.871	0.011
$\hat{f}_h$	1	0.026	1	0.019	1	0.016	1	0.011
Method	Bimodal							
	n=200		n=400		n=800		n=1600	
	MSE ratio	Bias	MSE ratio	Bias	MSE ratio	Bias	MSE ratio	Bias
$\hat{f}_{a,h}$	0.918	0.016	0.854	0.013	0.875	0.013	0.696	0.009
$\hat{f}_{\hat{a},h}$	0.946	0.016	0.829	0.013	0.917	0.014	0.717	0.010
$\hat{f}_{\sqrt{2},h}$	0.921	0.016	0.807	0.013	0.840	0.013	0.771	0.010
$\hat{f}_h$	1	0.017	1	0.014	1	0.014	1	0.011
Method	Student-t distribution(10 degrees of freedom)							
	n=200		n=400		n=800		n=1600	
	MSE ratio	Bias	MSE ratio	Bias	MSE ratio	Bias	MSE ratio	Bias
$\hat{f}_{a,h}$	0.924	0.025	0.971	0.018	0.939	0.014	0.844	0.006
$\hat{f}_{\hat{a},h}$	0.903	0.025	0.912	0.018	0.871	0.013	0.653	0.005
$\hat{f}_{\sqrt{2},h}$	0.907	0.025	0.914	0.018	0.902	0.013	0.722	0.005
$\hat{f}_h$	1	0.027	1	0.019	1	0.014	1	0.006
Method	Beta(2,2)							
	n=200		n=400		n=800		n=1600	
	MSE ratio	Bias	MSE ratio	Bias	MSE ratio	Bias	MSE ratio	Bias
$\hat{f}_{a,h}$	0.874	0.095	0.789	0.070	0.827	0.067	0.844	0.042
$\hat{f}_{\hat{a},h}$	0.845	0.094	0.787	0.071	0.831	0.067	0.808	0.042
$\hat{f}_{\sqrt{2},h}$	0.856	0.094	0.780	0.070	0.818	0.067	0.857	0.043
$\hat{f}_h$	1	0.103	1	0.081	1	0.075	1	0.049
Method	Uniform							
	n=200		n=400		n=800		n=1600	
	MSE ratio	Bias	MSE ratio	Bias	MSE ratio	Bias	MSE ratio	Bias
$\hat{f}_{a,h}$	1.065	0.014	1.060	0.012	1.052	0.010	0.835	0.008
$\hat{f}_{\hat{a},h}$	1.070	0.014	1.073	0.012	1.044	0.010	0.839	0.007
$\hat{f}_{\sqrt{2},h}$	1.062	0.014	1.039	0.012	1.050	0.010	0.874	0.008
$\hat{f}_h$	1	0.014	1	0.012	1	0.010	1	0.008

Table 1 The MSE ratio and bias of different methods under different settings.

Notice that

$$\begin{aligned}&2a^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{1}{n^2}\mathrm{Cov}\left(\frac{1}{h}K\left(\frac{X_i-x}{h}\right),\frac{1}{ah}K\left(\frac{X_j-x}{ah}\right)\right)\\&=2a^{-2}\frac{1}{n}E\left[\mathrm{Cov}\left(\frac{1}{h}K\left(\frac{X_i-x}{h}\right),\frac{1}{ah}K\left(\frac{X_j-x}{ah}\right)\right)\right].\end{aligned}$$

Since

$$\begin{aligned}&\mathrm{Cov}\left(\frac{1}{h}K\left(\frac{X_i-x}{h}\right),\frac{1}{ah}K\left(\frac{X_j-x}{ah}\right)\right)\\&=\frac{1}{a}\int K(w)K\left(\frac{w}{a}\right)f(wh+x)dw+O(1),\end{aligned}$$

Method	Distribution					
n=100	Normal	Bimodal	t-distribution	Beta	Uniform	Binomial
$\hat{f}_{a,h}$	0.002299	0.001487	0.000737	0.006049	0.000349	0.000529
$\hat{f}_{\hat{a},h}$	0.002321	0.001501	0.000748	0.007290	0.000346	0.000728
$\hat{f}_{\sqrt{2},h}$	0.002277	0.001485	0.000733	0.005935	0.000350	0.000783
$\hat{f}_h$	0.003459	0.001660	0.001132	0.012716	0.000351	0.000536
n=200	Normal	Bimodal	t-distribution	Beta	Uniform	Binomial
$\hat{f}_{a,h}$	0.000527	0.000876	0.000328	0.004390	0.000215	0.000476
$\hat{f}_{\hat{a},h}\cdot$	0.000548	0.000891	0.000352	0.004551	0.000215	0.000687
$\hat{f}_{\sqrt{2},h}$	0.000526	0.000873	0.000326	0.004325	0.000222	0.000715
$\hat{f}_h$	0.000691	0.001037	0.000452	0.008209	0.000218	0.000479
n=500	Normal	Bimodal	t-distribution	Beta	Uniform	Binomial
$\hat{f}_{a,h}$	0.000142	0.000340	0.000121	0.003243	0.000172	0.000112
$\hat{f}_{\hat{a},h}$	0.000141	0.000338	0.000129	0.003299	0.000175	0.000204
$\hat{f}_{\sqrt{2},h}$	0.000143	0.000340	0.000120	0.003262	0.000180	0.000198
$\hat{f}_h$	0.000181	0.000346	0.000209	0.004286	0.000172	0.000112

Table 2 Minimums of MSE for different methods under different settings in Figures 1-3.

and $K(u)$ is bounded, we have

$$\int K(w)K\left(\frac{w}{a}\right)f(wh+x)dw = O\Big(\int K(w)f(wh+x)dw\Big) = O(1).$$

With $C_2 = f(x)\int K(w)^2 dw$, the variance of $\hat{f}(x)$ can be simplified as:

$$\mathrm{Var}(\hat{f}(x)) = \left(1-a^{-2}\right)^{-2}\left(\frac{C_2}{nh} + \frac{C_2}{na^5h} + O\left(\frac{1}{na^3}\right) + O\left(\frac{1}{n}\right)\right).$$

Consequently,

$$\begin{aligned}
&\mathrm{MSE}(\hat{f}(x)) \\
&= \mathrm{bias}^2(\hat{f}(x)) + \mathrm{Var}(\hat{f}(x)) \\
&= a^4h^8C_1^2 + O(h^9) + \left(1-a^{-2}\right)^{-2}\left(\frac{C_2}{nh} + \frac{C_2}{na^5h} + O\left(\frac{1}{na^3}\right) + O\left(\frac{1}{n}\right)\right).
\end{aligned}$$

□

Proof of Theorem 1

Proof. For simplicity, we take $C_3 = \frac{C_2}{nh}$. It suffices to show that

$$\log\left\{\left(1-a^{-2}\right)^{-2}\left(C_3 + \frac{C_3}{a^5}\right)\right\},$$

is strictly convex in $a > 1$. Note that

$$\frac{d}{da}\log\left\{\left(1-a^{-2}\right)^{-2}\left(C_3 + \frac{C_3}{a^5}\right)\right\} = \frac{d}{da}\left\{\log\left(1-a^{-2}\right)^{-2} + \log\left\{C_3 + \frac{C_3}{a^5}\right\}\right\}$$

and for any $a > 1$,

$$\frac{d^2}{da^2}\log\left\{\left(1-a^{-2}\right)^{-2}\right\} = \frac{12a^2-4}{a^2(a^2-1)^2} > 0, \quad \frac{d^2}{da^2}\left\{C_3 + \frac{C_3}{a^5}\right\} > 0.$$

We thus conclude that $\widetilde{\text{MSE}}(\hat{f}_{a_x,h}(x))$ is a convex function of a for $a > 1$. By setting $\frac{d}{da}\left(\text{MSE}(\hat{f}(x))\right) = 0$, we obtain the implicit format of the global minimizer:

$$\frac{4h^9 nC_1^2}{C_2} = \frac{4a^5+5a^2-1}{(a^2-1)^3a^5}.$$

□

Proof of Theorem 2

Proof. For any given x, note the fact that a_x^* is the root of

$$\frac{4a_x^5+5a_x^2-1}{(a_x^2-1)^3a_x^5} = \frac{4h^9 nC_1^2}{C_2},$$

while $h \simeq O\left(n^{-1/9}\right)$, we have $a_x^* - 1 > c_0$ for some constant c_0. By Hansen (2009),

$$\begin{aligned}
\text{bias}\left(\hat{f}_{h_1}^{(4)}(x)\right) &= \frac{f^{(6)}(x)h_1^2\int K(u)u^2du}{2} + o\left(h_1^2\right), \\
\text{Var}\left(\hat{f}_{h_1}^{(4)}(x)\right) &= \frac{f(x)\int K^{(4)}(u)^2du}{nh_1^9} + O\left(\frac{1}{n}\right), \\
\text{bias}\left(\hat{f}_{h_2}(x)\right) &= \frac{f^{(2)}(x)h_2^2\int K(u)u^2du}{2} + o\left(h_2^2\right), \\
\text{Var}\left(\hat{f}_{h_2}(x)\right) &= \frac{f(x)\int K(u)^2du}{nh_2} + O\left(\frac{1}{n}\right).
\end{aligned}$$

Let

$$l(a_x, C_1, C_2) = 4(1+a_x^5)C_2 + 5(a_x^2-1)C_2 - 4(a_x^2-1)^3a_x^5nh^9C_1^2,$$

then

$$\begin{aligned}
\frac{\partial l(a_x,C_1,C_2)}{\partial a_x} &= 20a_x^4C_2 + 10a_xC_2 - 4\left(11a_x^2-5\right)a_x^4\left(a_x^2-1\right)^2 nh^9C_1^2, \\
\frac{\partial l(a_x,C_1,C_2)}{\partial C_1} &= -8\left(a_x^2-1\right)^3 a_x^5nh^9C_1, \\
\frac{\partial l(a_x,C_1,C_2)}{\partial C_2} &= 4a_x^5 + 5a_x^2 - 1.
\end{aligned}$$

Under the assumptions of this theorem, there exist positive constants c_1 and c_2 s.t. $\int K(u)u^2du$, $\int K(u)u^4du$, $\int K^{(4)}(u)^2du$ and $\int K(u)^2du$ are all smaller or equal than c_1, and $\max\{\ f^{(6)}(x),\ f^{(2)}(x),\ f(x)\} < c_2$. Then we have

$$
\begin{aligned}
&\operatorname{bias}(\hat{C}_1) \simeq h_1^2 + o\left(h_1^2\right), \quad \operatorname{Var}(\hat{C}_1) \le \frac{c_1^3 c_2}{nh_1^9} + O\left(\frac{1}{n}\right), \\
&\operatorname{bias}(\hat{C}_2) \simeq h_2^2 + o\left(h_2^2\right), \quad \operatorname{Var}(\hat{C}_2) \le \frac{c_1^3 c_2}{nh_2} + O\left(\frac{1}{nh_2}\right).
\end{aligned}
$$

As $n \to \infty$, we have

$$
\left|\hat{C}_1 - \mathrm{E}(\hat{C}_1)\right| = O_p\left(\sqrt{\frac{1}{nh_1^9}}\right), \quad \left|\hat{C}_2 - \mathrm{E}(\hat{C}_2)\right| = O_p\left(\sqrt{\frac{1}{nh_2}}\right),
$$

thus, with $h_1 \simeq O(n^{-1/13})$ and $h_2 \simeq O(n^{-1/5})$, we have that

$$
\begin{aligned}
\left|\hat{C}_1 - C_1\right| &\le \left|\hat{C}_1 - \mathrm{E}(\hat{C}_1)\right| + \left|\mathrm{E}(\hat{C}_1) - C_1\right| = O_p\left(\sqrt{\frac{1}{nh_1^9}} + h_1^2\right), \\
\left|\hat{C}_2 - C_2\right| &\le \left|\hat{C}_2 - \mathrm{E}(\hat{C}_2)\right| + \left|\mathrm{E}(\hat{C}_2) - C_2\right| = O_p\left(\sqrt{\frac{1}{nh_2}} + h_2^2\right).
\end{aligned}
$$

Consequently, as $n \to \infty$, for any given x, $\hat{C}_1$ and $\hat{C}_2$ are in the same order with C_1 and C_2 in probability. Similar to a_x, we can get that $\hat{a}_x - 1 > c_3$ for some constant c_3.

By Taylor expansion, there exist $(a_\xi, C_{1,\xi}, C_{2,\xi})$ s.t.

$$
\begin{aligned}
&l(\hat{a}_x, \hat{C}_1, \hat{C}_2) - l(a_x^*, C_1, C_2) \\
&= \left(\hat{C}_1 - C_1\right)\frac{\partial l(a_\xi, C_{1,\xi}, C_{2,\xi})}{\partial C_1} + \left(\hat{C}_2 - C_2\right)\frac{\partial l(a_\xi, C_{1,\xi}, C_{2,\xi})}{\partial C_2} \\
&+ (\hat{a}_x - a_x^*)\frac{\partial l(a_\xi, C_{1,\xi}, C_{2,\xi})}{\partial a_\xi} \\
&= 0. \qquad (9)
\end{aligned}
$$

Note that $a_\xi = O(1)$, $C_{1,\xi} = O(1)$ and $C_{2,\xi} = O(1)$, together with equation (9), we have

$$
\begin{aligned}
|\hat{a}_x - a_x^*| &= \left|\frac{(4a_\xi^5 + 5a_\xi^2 - 1)\left(\hat{C}_2 - C_2\right) - 8\left(a_\xi^2 - 1\right)^3 a_\xi^5 nh^9 C_{1,\xi}\left(\hat{C}_1 - C_1\right)}{20a_\xi^4 C_{2,\xi} + 10a_\xi C_{2,\xi} - 4\left(11a_\xi^2 - 5\right)a_\xi^4\left(a_\xi^2 - 1\right)^2 nh^9 C_{1,\xi}^2}\right| \\
&= O\left(\frac{\left|\hat{C}_2 - C_2\right|}{\left(a_\xi - 1\right)^2 nh^9}\right) + O\left(\left(a_\xi - 1\right)\left|\hat{C}_1 - C_1\right|\right) \\
&= O_P\left(\frac{h_2^2}{\left(a_\xi - 1\right)^2 nh^9}\right) + O_P\left(\left(a_\xi - 1\right)h_1^2\right)
\end{aligned}
$$

$$= O_p \left(\sqrt{\frac{1}{nh_1^9}} + h_1^2 + \sqrt{\frac{1}{nh_2}} + h_2^2 \right).$$

□

Acknowledgments

We thank the Editor, an Associate Editor, and an anonymous reviewer for their insightful comments. Wang' research is supported by the National Natural Science Foundation of China (12031005), and NSF of Shanghai (21ZR1432900). Jiang is partially supported by the National Natural Science Foundation of China (12001459), and HKPolyU Internal Grants.

References

Abramson, I. S. (1982). On bandwidth variation in kernel estimates-a square root law. *The annals of Statistics*, pages 1217–1223.

Bowman, A. W. (1984). An alternative method of cross-validation for the smoothing of density estimates. *Biometrika*, 71(2):353–360.

Breiman, L., Meisel, W., and Purcell, E. (1977). Variable kernel estimates of multivariate densities. *Technometrics*, 19(2):135–144.

Chen, Z. and Leng, C. (2016). Dynamic covariance models. *Journal of the American Statistical Association*, 111(515):1196–1207.

Hall, P. (1990). On the bias of variable bandwidth curve estimators. *Biometrika*, 77(3):529–535.

Hall, P. (2013). *The bootstrap and Edgeworth expansion.* Springer Science & Business Media.

Hansen, B. E. (2009). Lecture notes on nonparametrics. *Lecture notes.*

Igarashi, G. and Kakizawa, Y. (2015). Bias corrections for some asymmetric kernel estimators. *Journal of Statistical Planning and Inference*, 159:37–63.

Jenkins, M. A. and Traub, J. F. (1972). Algorithm 419: zeros of a complex polynomial [c2]. *Communications of the ACM*, 15(2):97–99.

Jiang, B., Chen, Z., Leng, C., et al. (2020). Dynamic linear discriminant analysis in high dimensional space. *Bernoulli*, 26(2):1234–1268.

Jones, M. and Foster, P. (1993). Generalized jackknifing and higher order kernels. *Journal of Nonparametric Statistics*, 3(1):81–94.

Jones, M., Linton, O., and Nielsen, J. (1995). A simple bias reduction method for density estimation. *Biometrika*, 82(2):327–338.

Kolar, M., Song, L., Ahmed, A., Xing, E. P., et al. (2010). Estimating time-varying networks. *The Annals of Applied Statistics*, 4(1):94–123.

Loftsgaarden, D. O., Quesenberry, C. P., et al. (1965). A nonparametric estimate of a multivariate density function. *The Annals of Mathematical Statistics*, 36(3):1049–1051.

Mack, Y. and Rosenblatt, M. (1979). Multivariate k-nearest neighbor density estimates. *Journal of Multivariate Analysis*, 9(1):1–15.

Rudemo, M. (1982). Empirical choice of histograms and kernel density estimators. *Scandinavian Journal of Statistics*, pages 65–78.

Schucany, W. and Sommers, J. P. (1977). Improvement of kernel type density estimators. *Journal of the American Statistical Association*, 72(358):420–423.

Schucany, W. R. (1989). On nonparametric regression with higher-order kernels. *Journal of Statistical Planning and Inference*, 23(2):145–151.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*, volume 26. CRC press.

Terrell, G. R. and Scott, D. W. (1992). Variable kernel density estimation. *The Annals of Statistics*, pages 1236–1265.

Tsybakov, A. B. (2008). *Introduction to nonparametric estimation*. Springer Science & Business Media.

Tukey, P. and Tukey, J. W. (1981). Data driven view selection, agglomeration, and sharpening. *Interpreting multivariate data*, pages 215–243.

Wand, M. P. and Schucany, W. R. (1990). Gaussian-based kernels. *Canadian Journal of Statistics*, 18(3):197–204.

Yao, W. (2012). A bias corrected nonparametric regression estimator. *Statistics & Probability Letters*, 82(2):274–282.