# Improving Sheather and Jones bandwidth selector for difficult densities in kernel density estimation

J.G. Liao

Penn State University, Hershey, PA, USA

jliao@HES.hmc.psu.edu

Yujun Wu

Sanofi-Aventis, Bridgewater, NJ, USA

and Yong Lin

Cancer Institute of New Jersey, New Brunswick, NJ, USA

**ABSTRACT**

Kernel density estimation is a widely used statistical tool and bandwidth selection is critically important. The Sheather and Jones selector (1991) remains the best available data-driven bandwidth selector. It can, however, perform poorly if the true density deviates too much in shape from normal. This paper first develops an alternative selector by following ideas in Park and and Marron (1990) to reduce the impact of the normal reference density. The selector can bring drastic improvement to less smooth densities that SJ selector has difficulty with but may be slightly worse off otherwise. We then propose to combine the alternative selector and SJ selector by using the smaller of the two bandwidths, which has the effect of automatically picking the better one for a particular density. In our extensive simulation study using the 15 benchmark densities in Marron and Wand (1992), the combined selector significantly improves SJ selector for five difficult densities and retains the superior performance of SJ selector for the other ten. A ready to use R function is provided.

Some key words: Mean integrated squared error; Marron and Wand densities; Solve-the-equation plug-in.

# 1. Introduction

Let $X = (x_1, ..., x_n)$ be $n$ independent observations from an unknown one-dimensional density $f$. The kernel density estimator of $f$ is given by

$$\hat{f}(t \mid X, h) = \frac{1}{nh} \sum_{i=1}^{n} K\left( \frac{t - x_i}{h} \right),$$

where $K$ is the kernel function and $h$ is the bandwidth. Kernel density estimation is widely used in science and technology. Textbook and review articles include Silverman (1986), Scott (1992); Wand and Jones (1995); Jones, Marron, & Sheather (1996); and Sheather (2004). It is well known that the density estimator $\hat{f}(\cdot \mid X, h)$ depends critically on the value of bandwidth $h$ but only mildly on the form of kernel $K$. Data-driven bandwidth selectors use data $X$ itself to estimate the optimal bandwidth and are important tools for facilitating wider and more objective use of kernel density estimation. Popular methods include the biased and unbiased cross-validation and various plug-in methods. The Sheather and Jones selector (1991) extended important work of Park and Marron (1990) and others and remains the state of the art for its excellent theoretical properties as well as superior practical performance. The SJ selector, however, can perform poorly if the true $f$ deviates severely from normality because of its use of normal reference density (Park and Marron, 1992; Devroye, 1997). Various researchers have tried to improve SJ selector. In particular, Park and Marron (1992) showed that applying normal reference at a higher stage can reduce the impact of normal reference and thus significantly improve the performance of bandwidth selector for some difficult densities. For smoother densities, however, the strategy can perform worse due to increased variability. We are unaware of any definitive bandwidth selector developed from this research.

This paper picks up the ideas in Park and Marron's paper to improve Sheather and Jones bandwidth selector. We shall first develop an alternative bandwidth selector by reducing the impact of the normal reference density, which can bring drastic improvement to less smooth densities but may be slightly worse off for smoother densities. We then propose to combine the alternative selector with SJ selector by using the minimum of the two, which has the effect of automatically picking the better one for a particular problem. This is based on the insight that SJ selector can serve as a good protective upper bound. In our extensive simulation study using the 15 benchmark densities in Marron and Wand (1992), the combined selector significantly improves SJ selector for five difficult densities (Figure 2) and retains the superior performance of SJ selector for the other ten. This represents the best performance of data-driven bandwidth selectors to our knowledge. A ready to use R function is provided.

# 2. An alternative bandwidth selector

## 2.1 Background

We will first introduce some notation and describe Sheather and Jones's bandwidth selector. Let $\mu_k \equiv \int t^k K(t)dt$ and let

$$R(p) \equiv \int p^2(t)dt$$

for a generic function $p$. For bandwidth $h$, the mean integrated squared error of estimator $\hat{f}(\cdot \mid X, h)$ is given by

$$\text{MISE}(h, f) \equiv E_X\left[\int\left\{\hat{f}(t \mid X, h) - f(t)\right\}^2 dt\right],$$

which, under suitable regularity conditions on $K$ and $f$, has asymptotic expansion (Scott and Terrell, 1987; Wand and Jones, 1995):

$$\text{MISE}(h, f) = \text{AMISE}(h, f) + O(n^{-1} + h^5),$$

where

$$\text{AMISE}(h, f) = \frac{h^4}{4}\mu_2^2 R(f'') + \frac{1}{nh}R(K).$$

The $\text{AMISE}(h, f)$ is minimized by $h_0 = \left[\frac{R(K)}{\mu_2^2 R(f'')n}\right]^{1/5}$, which is a large sample approximation to $h_*$ that minimizes the exact $\text{MISE}(h, f)$. For convenience of notation, define, for an even positive integer $s$, $\Phi_s(f) \equiv (-1)^{s/2} R(f^{(s/2)})$ and

$$\hat{\Phi}_s(X, g) \equiv \frac{1}{n^2 g^{s+1}}\sum_{i=1}^{n}\sum_{j=1}^{n} K^{(s)}\left(\frac{x_i - x_j}{g}\right),$$

where $f^{(s)}$ denotes the $s$th derivative of $f$. The asymptotically optimal bandwidth $g$ for using $\hat{\Phi}_s(X, g)$ to estimate $\Phi_s(f)$ is (Jones and Sheather, 1991; Wand and Jones, 1995)

$$g_s \equiv \left[-\frac{2K^{(s)}(0)}{\mu_2 \Phi_{s+2}(f)n}\right]^{\frac{1}{s+3}}.$$

We shall denote $g_4$ by $\gamma_0$ and $g_6$ by $\alpha_0$. An estimator of $h_0$ is then

$$\left[\frac{R(K)}{\mu_2^2 \hat{\Phi}_4(X, \gamma_0)n}\right]^{1/5}.$$

We have, on the other hand, that

$$h_0 = \left[ \frac{\gamma_0}{c\rho(f)} \right]^{7/5},$$

where $c = \left[ \frac{2K^{(4)}(0)\mu_2}{R(K)} \right]^{1/7}$ is a known functional of kernel $K$ and

$$\rho(f) \equiv \left[ -\frac{\Phi_4(f)}{\Phi_6(f)} \right]^{1/7}. \tag{1}$$

Let

$$\hat{\rho}(\gamma, \alpha) \equiv \left[ -\frac{\hat{\Phi}_4(X, \gamma)}{\hat{\Phi}_6(X, \alpha)} \right]^{1/7}$$

be an estimator of $\rho(f)$. It is then natural to estimate $\gamma_0$ by solving

$$\left[ \frac{\gamma_0}{c\hat{\rho}(\gamma_0, \alpha_0)} \right]^{7/5} = \left[ \frac{R(K)}{\mu_2^2 \hat{\Phi}_4(X, \gamma_0) n} \right]^{1/5}.$$

A fundamental difficulty in data-driven bandwidth selection, however, is that $\gamma_0$ and $\alpha_0$ depend on the unknown $\Phi_6(f)$ and $\Phi_8(f)$. The Sheather and Jones's bandwidth selector gets around this problem by using $N(0, \hat{\sigma}^2)$ as the reference density, where $\hat{\sigma}^2$ is a robust estimate of the variance of $f$. It can be shown that

$$\gamma_0^{\text{ref}} = \left[ \frac{96}{\sqrt{2} \times 15n} \right]^{1/7} \hat{\sigma}, \ \alpha_0^{\text{ref}} = \left[ \frac{960}{\sqrt{2} \times 105n} \right]^{1/9} \hat{\sigma}$$

for this normal reference $N\left(0, \hat{\sigma}^2\right)$. Let $\hat{\gamma}_0$ be the solution of

$$\left[ \frac{\gamma_0}{c\hat{\rho}(\gamma_0^{\text{ref}}, \alpha_0^{\text{ref}})} \right]^{7/5} = \left[ \frac{R(K)}{\mu_2^2 \hat{\Phi}_4(X, \gamma_0) n} \right]^{1/5} \tag{2}$$

The Sheather and Jones's bandwidth selector, $\hat{h}_1$, defined as the right side for $\gamma_0 = \hat{\gamma}_0$, converges to $h_0$ at the relative rate $n^{-5/14}$ and generally works well. But it may perform poorly if the true $f$ deviates severely from normality (Park and Marron, 1992; Devroye, 1997).

### 2.2 The alternative bandwidth selector

We now describe our alternative bandwidth selector. For convenience, we shall only consider the normal kernel $K(t) = (2\pi)^{-1/2} e^{-t^2/2}$ because the identity

$$\Phi_s\left(\hat{f}\left(\cdot \mid X, g/\sqrt{2}\right)\right) = \hat{\Phi}_s(X, g) \tag{3}$$

for this kernel simplifies the exposition considerably. Two modifications will be made to (2). First, we will integrate $\gamma_0^{\text{ref}}$ with the rest of $\gamma_0$ by making $\gamma_0^{\text{ref}} = \gamma_0$. It is then only necessary to use the normal reference for ratio $\alpha_0/\gamma_0$, which is $\alpha_0^{\text{ref}}/\gamma_0^{\text{ref}} = d_1 n^{1/7-1/9}$ with $d_1 \approx .992$ for normal reference $N\left(0, \hat{\sigma}^2\right)$. Note, interestingly, that this $\alpha_0^{\text{ref}}/\gamma_0^{\text{ref}}$ does not depend on the specific value of $\hat{\sigma}^2$, which provides additional robustness. Let

$$\alpha_0^{\text{ref}} = d_1 n^{1/7-1/9}\gamma_0 \tag{4}$$

from now on. Sheather and Jones's equation (2) are now modified to be

$$\left[\frac{\gamma_0}{c\hat{\rho}(\gamma_0, \alpha_0^{\text{ref}})}\right]^{7/5} = \left[\frac{R(K)}{\mu_2^2 \hat{\Phi}_4(X, \gamma_0) n}\right]^{1/5} \tag{5}$$

Our second modification is to force $\gamma = \alpha$ in $\hat{\rho}(\gamma, \alpha)$ so that

$$\hat{\rho}(\gamma, \alpha) = \left[-\frac{\hat{\Phi}_4(X, \alpha)}{\hat{\Phi}_6(X, \alpha)}\right]^{1/7} = \rho\left(\hat{f}\left(\cdot \mid X, \alpha/\sqrt{2}\right)\right),$$

where the second equality comes from identity (3). With this, $\rho(f)$ is now estimated by the same functional $\rho$ with true density $f$ replaced by an estimated density $\hat{f}\left(\cdot \mid X, \alpha/\sqrt{2}\right)$. To choose the value for $\gamma = \alpha$, note that $\hat{\rho}(\gamma, \alpha)$ depends much more critically on the value of $\alpha$ in the higher order $\hat{\Phi}_6$ than $\gamma$ in the lower order $\hat{\Phi}_4$. We will therefore make $\gamma = \alpha = \alpha_0^{\text{ref}}$ and modify (5) to

$$\left[\frac{\gamma_0}{c\hat{\rho}(\alpha_0^{\text{ref}}, \alpha_0^{\text{ref}})}\right]^{7/5} = \left[\frac{R(K)}{\mu_2^2 \hat{\Phi}_4(X, \gamma_0) n}\right]^{1/5} \tag{6}$$

where $\alpha_0^{\text{ref}}$ is given in (4). It is shown in the Appendix that $\hat{\rho}(\alpha_0^{\text{ref}}, \alpha_0^{\text{ref}}) - \rho(f) = O_p(n^{-2/9})$. Using this forced functional form $\hat{\rho}\left(\alpha_0^{\text{ref}}, \alpha_0^{\text{ref}}\right)$ proves more stable than using $\hat{\rho}(\gamma_0, \alpha_0^{\text{ref}})$ for smaller $n$ in our experience.

Some bias adjustment is needed on equation (6), however. Note that $\alpha_0^{\text{ref}}$ in (4) is greater than $\gamma_0$ and consequently $\hat{\rho}_2 \equiv \hat{\rho}(\alpha_0^{\text{ref}}, \alpha_0^{\text{ref}}) < \hat{\rho}(\gamma_0, \alpha_0^{\text{ref}}) \equiv \hat{\rho}_1$. The left side of (6) is therefore larger than the left side of (5) it replaces. More precisely, we have

$$\left[\frac{\gamma_0}{c\hat{\rho}_2}\right]^{7/5} - \left[\frac{\gamma_0}{c\hat{\rho}_1}\right]^{7/5} = \left[\frac{\gamma_0}{c\hat{\rho}_2}\right]^{7/5} \frac{7}{5}\frac{\hat{\rho}_1 - \hat{\rho}_2}{\hat{\rho}_1} + o_p\left(n^{-19/45}\right), \tag{7}$$

which is a positive term of order $O_p\left(n^{-19/45}\right)$. To compensate this, we will replace the right side of (6) by something larger. Let $p$ be a generic density and let $h_0(p)$ be the bandwidth that minimizes the asymptotic mean integrated squared error $\mathrm{AMISE}(h, p)$. The right side of (6) is then $h_0\left(\hat{f}\left(\cdot \mid X, \gamma_0/\sqrt{2}\right)\right)$ by identity (3). Now let $h_*(p)$ be the bandwidth that minimizes the exact $\mathrm{MISE}(h, p)$. We shall modify equation (6) to

$$\left[\frac{\gamma_0}{c\hat{\rho}(\alpha_0^{\mathrm{ref}}, \alpha_0^{\mathrm{ref}})}\right]^{7/5} = h_*\left(\hat{f}\left(\cdot \mid X, \gamma_0/\sqrt{2}\right)\right). \tag{8}$$

To see why this makes sense, let $\hat{h}_0$ denote the right side of (6) and let $\hat{h}_*$ be the right side of (8). Let

$$J_2 = \frac{\mu_4}{20\mu_2}\frac{1}{\hat{\rho}^7(\gamma_0, \gamma_0)},$$

where $\mu_2 = 1$ and $\mu_4 = 3$ for normal kernel. It follows (Hall, et al., 1991) that

$$\hat{h}_* - \hat{h}_0 = J_2\hat{h}_0^{\,3} + o_p(n^{-3/5}), \tag{9}$$

which is a positive term of order $O_p(n^{-3/5})$. Our empirical experience shows this increase in (9) is effective in partially compensating for the increase in (7). For the simulation study below and for smaller sample size $n = 100$, equation (6) has no root for more than 50% of the samples because the left side is consistently larger than the right side. For equation (8), however, a unique root was found for 100% of the samples. Other adjustment may be pursued. But again the proposed adjustment through $h_*(p)$ for a real density $p = \hat{f}\left(\cdot \mid X, \gamma_0/\sqrt{2}\right)$ seems more stable. Let $\hat{\gamma}_0$ be the root of equation (8). Our alternative bandwidth selector $\hat{h}_2$ is defined as the right side of (8) with $\gamma_0 = \hat{\gamma}_0$. It is shown in Appendix that $\hat{h}_2$ has the same $n^{-5/14}$ relative convergence rate as $\hat{h}_1$:

$$\frac{\hat{h}_2 - h_0}{h_0} = O_p\left(n^{-5/14}\right).$$

### 2.3 Performance for 15 benchmark densities

The ultimate judgment of any bandwidth selector is of course its practical performance. For this we conduct an extensive simulation study to compare $\hat{h}_2$ against $\hat{h}_1$. Let $\hat{h}$ be a data-driven bandwidth

selector, $\hat{h}_1$ or $\hat{h}_2$ for our purpose. We shall quantify the performance of $\hat{f}(\cdot \mid X, \hat{h})$ as an estimator of $f$ by the mean integrated squared error

$$\mathrm{MISE}(\hat{h}, f) \equiv E_X\left[\int\left\{\hat{f}(t \mid X, \hat{h}) - f(t)\right\}^2 dt\right],$$

where the expectation is with respect to data vector $X$ with $\hat{h}$ being a function of $X$. Let $h_*$ be the optimal bandwidth that minimizes the exact $\mathrm{MISE}(h, f)$ as before. The relative MISE for selector $\hat{h}$ is then defined as

$$\mathrm{RMISE}(\hat{h}) \equiv \frac{\mathrm{MISE}(\hat{h}, f)}{\mathrm{MISE}(h_*, f)}.$$

The smaller $\mathrm{RMISE}(\hat{h})$ is, the better the performance. We use the 15 benchmark densities in Marron and Wand (1992) which represent a wide variety of potential densities and have been used by a number of authors in comparing different bandwidth selectors. For each of the 15 benchmark densities and for each of $n = 100$, 200, 400, 800, 1600, 3200 and 6400, we generate 1000 datasets of sample size $n$. For every dataset $X = (x_1, ..., x_n)$, $\hat{h}_1$ and $\hat{h}_2$ are computed together with the corresponding integrated squared error

$$\mathrm{ISE}(\hat{h}, f) \equiv \int\left\{\hat{f}(t \mid X, \hat{h}) - f(t)\right\}^2 dt,$$

for $h = \hat{h}_1$ and $h = \hat{h}_2$. These values are used to estimate $\mathrm{RMISE}(\hat{h})$ by direct Monte Carlo method. Figure 1 plots the estimated $\mathrm{RMISE}(\hat{h})$. We see that $\hat{h}_2$ has a much smaller $\mathrm{RMISE}(\hat{h})$ for densities 3, 12, 13, 14 and 15 (very different in shape from normal reference) especially when $n$ is larger but can have a slightly increased $\mathrm{RMISE}(\hat{h})$ for densities 6, 8, 9, and 11 when $n$ is smaller. Our simulation result is consistent with the limited simulation experiment in Park and Marron (1992) that using higher stage of normal reference is beneficial for less smooth density 3 but harmful for more smooth density 6. Note also that $\hat{h}_2$ outperform $\hat{h}_1$ slightly for small size $n = 100$ for the normal benchmark density 1. This could be due to the use of exact $h_*(p)$ in $\hat{h}_2$ as opposed to the asymptotic $h_0(p)$ in $\hat{h}_1$ because this performance advantage disappears as $n$ increases.

## 3. Combining SJ and the alternative estimators

Based on the analysis and simulation study above, it would be ideal if we can automatically choose a bandwidth from $\hat{h}_1$ and $\hat{h}_2$ that is more likely to have a smaller MISE for a particular problem. For this,

we propose selector $\hat{h}_3 \equiv \min\left(\hat{h}_1, \hat{h}_2\right)$. We motivate this choice of $\hat{h}_3$ as follows. Jones, Marron and Sheather (1996) found that $\hat{h}_1$ tends to be centered near the optimal bandwidth $h_*$ for smooth densities with relatively small $R^{1/5}(f^{"})$ but is upward biased (over-smoothing ) for densities with larger $R^{1/5}(f^{"})$. This is not surprising as SJ selector uses the very smooth normal density as reference although the technical details of the impact of reference density on bandwidth selector still need to be worked out. Note that $R^{1/5}(f^{"}) \geq \left(\frac{35}{243}\right)^{1/5} \sigma^{-1}$ for any density $f$ with variance $\sigma^2$ (Terrell and Scott, 1985; Terrell, 1990). The value of $R^{1/5}(f^{"})$, for $f$ being $N(0,\sigma^2)$, is $\left(\frac{3}{8\sqrt{\pi}}\right)^{1/5} \sigma^{-1}$, only 8% larger than the minimum. Jones, Marron and Sheather (1996) also found that $\hat{h}_1$ has a much lower variability than other bandwidth selectors that do not use a normal reference. Based on this, we treat $\hat{h}_1$ as an effective approximate upper bound and define $\hat{h}_3$ as $\hat{h}_2$ bounded above by $\hat{h}_1$. For a closer look, consider the following two scenarios. First, $\hat{h}_1$ can severely over-estimate $h_*$ when $R^{1/5}(f^{"})$ is large. In this case, $\hat{h}_3$ automatically chooses the smaller and less biased $\hat{h}_2$. Second, $\hat{h}_1$ becomes the better estimator when $R^{1/5}(f^{"})$ is relatively small. The $\hat{h}_3$ then improves estimator $\hat{h}_2$ by removing some of its upward variability by upper bound $\hat{h}_1$. Terrell (1990) also alluded to the strategy of using the smaller of two bandwidth selectors, one being the stable maximal smoothing and the other being the much more variable but less biased cross-validation. Figure 2 plots the estimated $\mathrm{RMISE}(\hat{h})$ for $\hat{h}_1$ and $\hat{h}_3$ for the simulation study in the last section. The result seems pretty remarkable as $\mathrm{RMISE}(\hat{h}_3)$ achieves the smaller of $\mathrm{RMISE}(\hat{h}_1)$ and $\mathrm{RMISE}(\hat{h}_2)$ for almost all 15 benchmark densities across the range of $n$ from 100 to 6400 with minor exceptions in densities 1 and 5.

The proposed algorithm is implemented in R (The R Development Core Team, 2008). Our R function *bw.liao* takes vector $X$ as input and outputs $\hat{h}_3$. The $\hat{h}_2$ in it is obtained by a binary search on $\gamma_0$ in equation (8). The program successfully processed more 20000 datasets and appears numerically very stable. It takes about 1 to 5 seconds for a single dataset of size $n = 100$ to 6400 on our laptop computer. The code is available at http://geocities.com/jg_liao. The R function *bw.SJ* (Venables and Ripley, 2002) is

used for computing $\hat{h}_1$ and R package *nor1mix* by Martin Mächler used for sampling from the 15 benchmark densities in our simulation study.

## 4. Discussion

We first proposed an alternative bandwidth selector $\hat{h}_2$ by following the idea in Park and Marron (1992) to lessen the impact of the normal reference. In our formulation, normal reference $N(0,\hat{\sigma}^2)$ is only used for ratio $\alpha_0 / \lambda_0$, which happens to be independent of the specific value of $\hat{\sigma}^2$ for additional robustness. The $\hat{h}_3$ then automatically combines the strength of the proposed $\hat{h}_2$ with SJ selector $\hat{h}_1$ by choosing the smaller of the two. The $\hat{h}_3$ significantly improves on $\hat{h}_1$ for five difficult benchmark densities and retains the superior performance of $\hat{h}_1$ for the other ten. We feel confident in recommending $\hat{h}_3$ and its R implementation for general use based on our extensive simulation experience.

One may question if the substantial reduction in $\mathrm{RMISE}(\hat{h}_3)$ for densities 12, 13, 14 and 15 is really very useful since a global bandwidth is not the best choice for densities of such complex shapes anyway. To see why the answer is affirmative, note that there is usually little prior information on the shape of $f$ in nonparametric density estimation. An initial global smoothing is often used to detect the need for variable bandwidth and as the basis for further iteration (Silverman, 1986). Here $\hat{h}_3$ can do a much better job than $\hat{h}_1$. Note also that variable bandwidth requires larger sample size. Global smoothing with $\hat{h}_3$ can in fact be more desirable for small or intermediate $n$.


## Acknowledgement

The authors are extremely grateful to an anonymous reviewer for his guidance throughout the several versions of this paper and for his careful proofreading.

# REFERENCES


Devroye, L. (1997). Universal smoothing factor selection in density estimation: theory and practice (with discussion). *Test*, 6, 223-320.

Hall, P., Sheather, S.J., Jones, M.C. and Marron, J.S. (1991). On Optimal Data-Based Bandwidth Selection in Kernel Density Estimation. *Biometrika* 78, 263-9

Jones, M. C., Marron, J. S., and Sheather, S. J. (1996). A Brief Survey of Bandwidth Selection. *Journal of the American Statistical Association* 91, 401–407.

Jones, M.C. and Sheather, S.J. (1991). Using non-stochastic terms to advantage in kernel-based estimation of integrated squared density derivatives. *Statistics and Probability letters* 11, 511-514.

Marron, J.S. and Wand, M. P. (1992). Exact mean integrated squared error. *The Annals of Statistics* 20, 712-736.

Park, B.U. and Marron, J.S. (1990). Comparison of data-driven bandwidth selectors. *Journal of the American Statistical Association* 85, 66-72.

Park, B.U. and Marron, J.S. (1992). On the use of pilot estimators in bandwidth selection. *Nonparametric statistics* 1, 231-240.

Raykar, V. C. and Duraiswami, R. (2006). Fast optimal bandwidth selection for kernel density estimation. In *Proceedings of the sixth SIAM International Conference on Data Mining* (pp. 524-528). Bethesda.

Scott, D. W. (1992). Multivariate Density Estimation: theory, practice and visualization. Wiley.

Sheather, S. (2004). Density Estimation. *Statistical Science* 19, 588–597.

Sheather, S. J. and Jones, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society series B*, 53, 683–690.

Scott, D.W. and Terrell G.R. (1987) Biased and Unbiased Cross-Validation in Density Estimation. *Journal of the American Statistical Association* 82, 1131-46

Silverman, B.W. (1986). Density estimation for statistics and data analysis. Chapman & Hall.

Terrell, G. R. (1990). The maximal smoothing principle in density estimation. *Journal of the American Statistical Association* 85 470–477.

Terrell, G. R. and Scott, D. W. (1985). Oversmoothed nonparametric density estimates. *Journal of the American Statistical Association* 80, 209–214.

The R Development Core Team. (2008). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing. URL http://www.R-project.org.

Van der Vaart, A. (1998). *Asymptotic Statistics.* Cambridge University Press.

Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S* (4th ed.). Springer.
Wand, M. P. and Jones, M. C. (1995). *Kernel Smoothing.* Chapman and Hall.

## Appendix: Convergence rate of $\hat{h}_2$

We now show that $\hat{h}_2$ has the same $n^{-5/14}$ relative convergence rate as $\hat{h}_1$. We will rely on the *M*- and Z-estimation theory such as in Chapter 5 of van der Vaart (1998). Note that $\alpha_0^{\text{ref}}$ in (4) is of order $n^{-1/9}$. It follows from the asymptotic bias and variance formula in Jones and Sheather (1991) or in Appendix 4 of Raykar and Duraiswami (2006) that

$$\hat{\Phi}_s\left(X, \alpha_0^{\text{ref}}\right) - \Phi_s\left(f\right) = O_p(n^{-2/9}) \tag{10}$$

for both $s = 4$ and $s = 6$. It then follows that

$$\hat{\rho}\left(\alpha_0^{\text{ref}}, \alpha_0^{\text{ref}}\right) - \rho\left(f\right) = O_p(n^{-2/9}). \tag{11}$$

Let

$$V(h, f) \equiv \frac{\partial \text{MISE}(h, f)}{\partial h},$$

where the differentiation is with respect to $h$ with density $f$ as fixed. Let

$$A(\gamma, X) \equiv V\left(h, \hat{f}\left(\cdot \mid X, \gamma / \sqrt{2}\right)\right),$$

be a univariate function of $\gamma$ by substituting $h$ by

$$h = \left[\frac{\gamma}{c \cdot \hat{\rho}\left(\alpha^{\text{ref}}, \alpha^{\text{ref}}\right)}\right]^{7/5} \quad \text{with } \alpha^{\text{ref}} = d_1 n^{1/7 - 1/9}\gamma.$$

It follows from the definition of $h_*(p)$ above that $\hat{\gamma}_0$ is the root of $A(\gamma, X) = 0$. For proper scaling, let $\theta = n^{1/7}\gamma$ and $\theta_0 = n^{1/7}\gamma_0$ Define

$$B(\theta, X) \equiv nh^2 A\left(n^{-1/7}\theta, X\right).$$

It follows that

$$B(\theta, X) = nh^5 {\mu_2}^2 \hat{\Phi}_4\left(X, \gamma\right) - R(K) + O(h^2 + nh^6)$$

$$= \frac{\hat{\Phi}_4(X, \gamma)}{\Phi_4(f)}\left[\frac{\rho(f)}{\hat{\rho}(\alpha^{\text{ref}}, \alpha^{\text{ref}})}\right]^7 \left(\frac{\theta}{\theta_0}\right)^7 R(K) - R(K) + O(h^2 + nh^6).$$

We now have

$$\lim_{n\to\infty} B(\theta, X) = R(k)\left(\frac{\theta^7}{\theta_0^7} - 1\right)$$

in the neighborhood of $\theta_0$ and the limit function on the right side is monotonic on $\theta$ with value zero at $\theta = \theta_0$. There therefore exists a solution $\theta_n$ for Equation $B(\theta, X) = 0$ with the property of $\theta_n \to \theta_0$ as $n \to \infty$. To derive the convergence rate for such $\theta_n$, do a standard Taylor expansion

$$\theta_n - \theta_0 = \frac{B(\theta_0, X)}{B'(\theta_0, X) + O(|\theta_n - \theta_0|)}.$$

It follows from (10) and (11) that $B(\theta_0, X) = O(n^{-2/9})$ and therefore $\hat{\theta} = \theta_0 + O(n^{-2/9})$. Jones and Sheather (1991) and Park and Marron (1992) showed that it is only necessary to estimate $\theta_0$ to the order of $o(n^{-1/14})$ for $\hat{h}_1$ to converge to $h_0$ at rate of $O(n^{-5/14})$. We have, similarly for $\hat{h}_2$, that

$$\frac{\hat{h}_2 - h_0}{h_0} = O_p(n^{-5/14}).$$

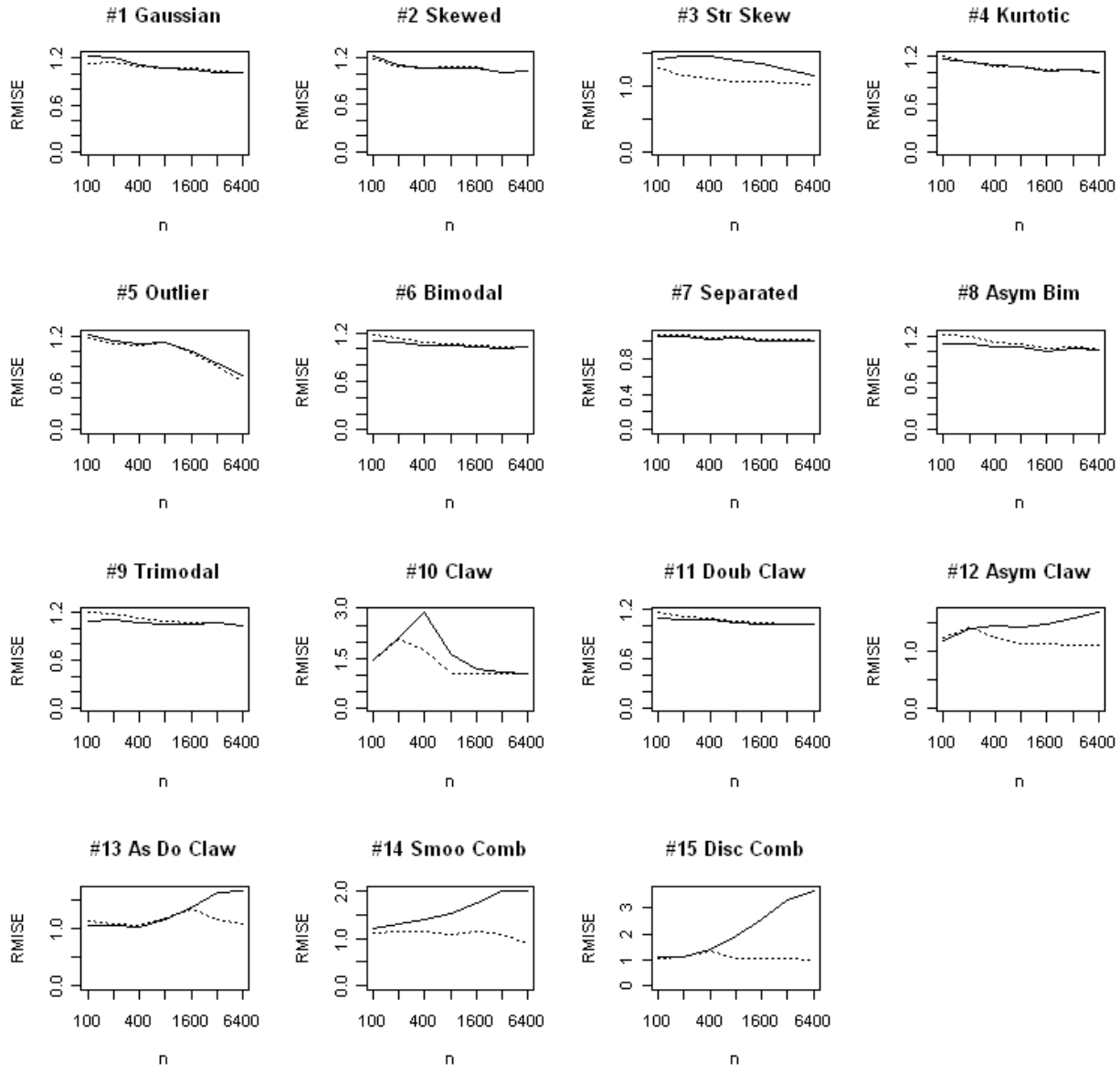


Figure 1. The estimated $\mathrm{RMISE}(\hat{h})$ for the 15 benchmark densities in Marron and Wand (1992) and for sample size $n = 100$, 200, 400, 800, 1600, 3200 and 6400 based on Monte Carlo simulation of 1000 replications. The solid line is for Sheather and Jones's selector $\hat{h}_1$ and the dashed line is for the proposed selector $\hat{h}_2$ A smaller value represents better performance in estimating $f$.

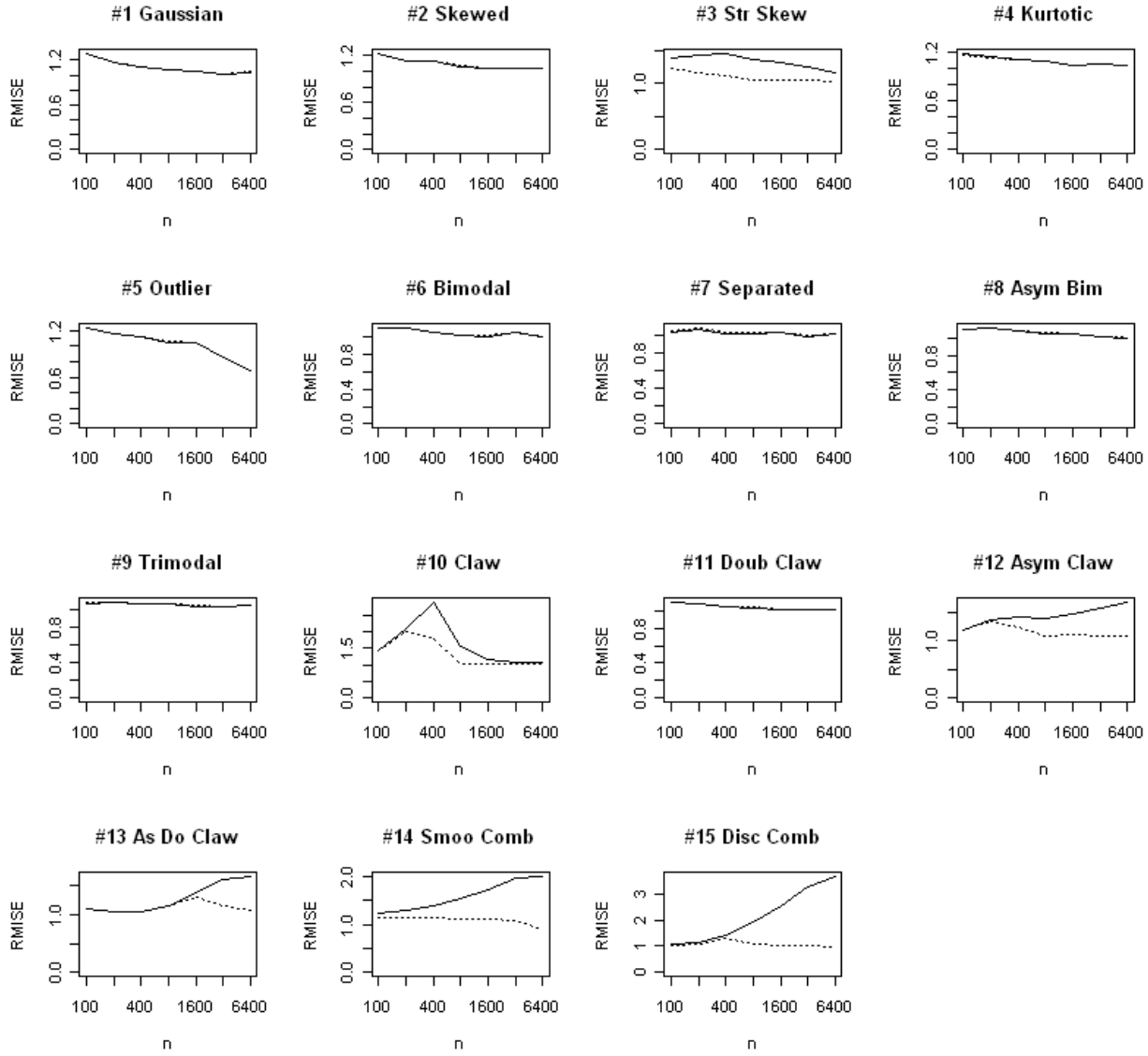


Figure 2. The estimated $\mathrm{RMISE}(\hat{h})$ for the 15 benchmark densities in Marron and Wand (1992) and for sample size $n = 100$, 200, 400, 800, 1600, 3200 and 6400 based on Monte Carlo simulation of 1000 replications. The solid line is for Sheather and Jones's selector $\hat{h}_1$ and the dotted line is for the combined selector $\hat{h}_3$. A smaller value represents better performance in estimating $f$.