Communications for Statistical Applications and Methods
2014, Vol. 21, No. 5, 435–444

DOI: http://dx.doi.org/10.5351/CSAM.2014.21.5.435
Print ISSN 2287-7843 / Online ISSN 2383-4757

The Bandwidth from the Density Power Divergence

Ro Jin Pak[1,a]

[a]Department of Applied Statistics, Dankook University, Korea

Abstract

The most widely used optimal bandwidth is known to minimize the mean integrated squared error(MISE) of a kernel density estimator from a true density. In this article proposes, we propose a bandwidth which asymptotically minimizes the mean integrated density power divergence(MIDPD) between a true density and a corresponding kernel density estimator. An approximated form of the mean integrated density power divergence is derived and a bandwidth is obtained as a product of minimization based on the approximated form. The resulting bandwidth resembles the optimal bandwidth by Parzen (1962), but it reflects the nature of a model density more than the existing optimal bandwidths. We have one more choice of an optimal bandwidth with a firm theoretical background; in addition, an empirical study we show that the bandwidth from the mean integrated density power divergence can produce a density estimator fitting a sample better than the bandwidth from the mean integrated squared error.

Keywords: Density estimator, density power divergence, Kullback-Leibler divergence, L_2 distance, mean integrated square error.

1. Introduction

Suppose that we have a set of random sample $X_1, \ldots, X_n$ of size n from an unknown probability density function f. Then a kernel density estimator $\hat{f}$ is defined by

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

where $K(\cdot)$ is called a kernel function, and h is called a bandwidth, window width or smoothing parameter (Rosenblatt, 1956; Parzen, 1962). The details about many practical aspects of $\hat{f}_n(x)$ can be found in Silverman (1985).

For estimating a density, it is crucial to select the appropriate value of a bandwidth h. To get an appropriate value of a bandwidth, we usually consider a global measure of discrepancy between the kernel density estimator and the density. The most commonly employed measure is the mean integrated squared error(MISE)

$$\mathrm{MISE}(h) = E\int \left\{\hat{f}_n(x) - f(x)\right\}^2 dx,$$

which gives the optimal bandwidth

$$h_{\mathrm{MISE}} = k_2^{-\frac{2}{5}}\left\{\int K(t)^2 dt\right\}^{\frac{1}{5}}\left\{\int f^{(2)}(x)^2 dx\right\}^{-\frac{1}{5}} n^{-\frac{1}{5}},$$

[1] Department of Applied Statistics, Dankook University, 126 Jukjeon-Dong, Suji-Gu, Yongin 448-701,Korea. E-mail: rjpak@dankook.ac.kr

Published 30 September 2014 / journal homepage: http://csam.or.kr

where $k_2 = \int t^2 K(t)dt$.

As the alternative discrepancies, Devroye and Györfi (1985) employed the mean integrated error $E\int|\hat{f} - f|$, and Hall (1987) employed the expected Kullback-Leibler loss $E\int f \log(f/\hat{f})$, and Kanazawa (1993) employed the mean Hellinger distance $E\int\{\hat{f}^{1/2} - f^{1/2}\}^2$, respectively.

In this article, in order to find an optimal bandwidth, we are trying to use the density power divergence between density functions f and g such as

$$d_\alpha(f, g) = \int \left\{ f(x)^{1+\alpha} - \frac{1+\alpha}{\alpha} f(x)^\alpha g(x) + \frac{1}{\alpha} g(x)^{1+\alpha} \right\} dx, \quad (\alpha > 0), \tag{1.1}$$

which was defined by Basu *et al.* (1998). When $\alpha = 0$, the divergence is defined as, so called Kullback-Leibler divergence,

$$d_0(f, g) = \int g(x) \log\left\{\frac{g(x)}{f(x)}\right\} dx.$$

The family of MDPD is indexed by a single parameter, α, which controls the trade-off between the asymptotic efficiency and robustness Of the MDPD estimator (Basu *et al.* 1998). The following sections show how we get an optimal bandwidth utilizing $d_\alpha(\cdot)$ in (1.1) and discuss corresponding properties.

Since Basu *et al.* (1998) have introduced the MDPD estimator, the robustness properties of these estimators have been studied in detail for various areas from fundamental concepts to applications by several authors. For example, Lee and Na (2005) considered the problem of testing for a parameter change based on the cusum test. Durio and Isaia (2011) investigated the MDPD estimation as a practical tool for a parametric regression model building. Most recently, Basu *et al.* (2013) considered parametric hypothesis testing based on the density power divergence in a limited context. However, there has been no study about finding a bandwidth of a density estimator based on the density power divergence. We show how to get an optimal bandwidth by using the MDPD. The resulting bandwidth is like the optimal bandwidth by Parzen (1962), but it reflects the nature of a model density.

The critical issue in the use of the MDPD estimation is the choice of tuning parameters. Basu *et al.* (1998) made remarks on how to select an α, yet without methodological details. Warwick and Jones (2005) chose an α by minimizing an asymptotic estimator of the mean square error. Fujisawa and Eguchi (2006) proposed an adaptive methods to select an α based on an empirical approximations of the Cramer-Von Mises divergence. The above approaches are theoretically sound but they are make the MDPD estimation harder to use. Durio and Isaia (2011) investigated the use of the MDPD criterion as a practical tool for parametric regression model building. Durio and Isaia (2011) proposed a data-driven way to choose the α using a Monte Carlo Significance test on the similarity between a robust and a classical estimators. In this article we run a simple empirical study to show how to find an appropriate α using distributional adequacy statistics like a Kolmogorov-Smirnov(K-S) test.

2. Main Results

Let $g(x)$ in (1.1) be

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

and let's consider minimize

$$d_\alpha(f, \hat{f}_n) = \int \left\{ f(x)^{1+\alpha} - \frac{1+\alpha}{\alpha} f(x)^\alpha \hat{f}_n(x) + \frac{1}{\alpha} \hat{f}_n(x)^{1+\alpha} \right\} dx$$

w.r.t. h to get an optimal h. Since the first term is free of h, to minimize $d_\alpha(f, \hat{f}_n)$ *w.r.t.* h is equivalent to minimize

$$\frac{1}{\alpha} \int \hat{f}_n(x)^{1+\alpha} dx - \frac{1+\alpha}{\alpha} \int f(x)^\alpha \hat{f}_n(x) dx \tag{2.1}$$

w.r.t. h.

Similar to Silverman (1985), we take an expectation on (2.1) so as to measure the global accuracy of $\hat{f}_n$ to f as

$$\frac{1}{\alpha} E \int \hat{f}_n(x)^{1+\alpha} dx - \frac{1+\alpha}{\alpha} E \int f(x)^\alpha \hat{f}_n(x) dx \tag{2.2}$$

and call (2.2) as the mean integrated density power divergence(MIDPD).

Therefore, the globally optimal bandwidth, h_{MIDPD}, is the solution to

$$E \int \hat{f}_n(x)^\alpha \hat{f}_n^{(1)}(x) dx - E \int f(x)^\alpha \hat{f}_n^{(1)}(x) dx = 0, \tag{2.3}$$

where the number in an upper case parenthesis stands for the order of derivative *w.r.t.* h. We have (2.3) by taking a derivative of (2.2) *w.r.t.* h and setting it to be equal to zero.

The left-hand side of (2.3) can be approximated by the first order Taylor series expansion as follows;

$$E \int \left\{ \hat{f}_n(x)^\alpha - f(x)^\alpha \right\} \hat{f}_n^{(1)}(x) dx \approx E \int \alpha \left\{ \hat{f}_n(x) - f(x) \right\} f(x)^{\alpha-1} \hat{f}_n^{(1)}(x) dx.$$

Hence, we propose to solve the (approximated) mean integrated density power divergence equation,

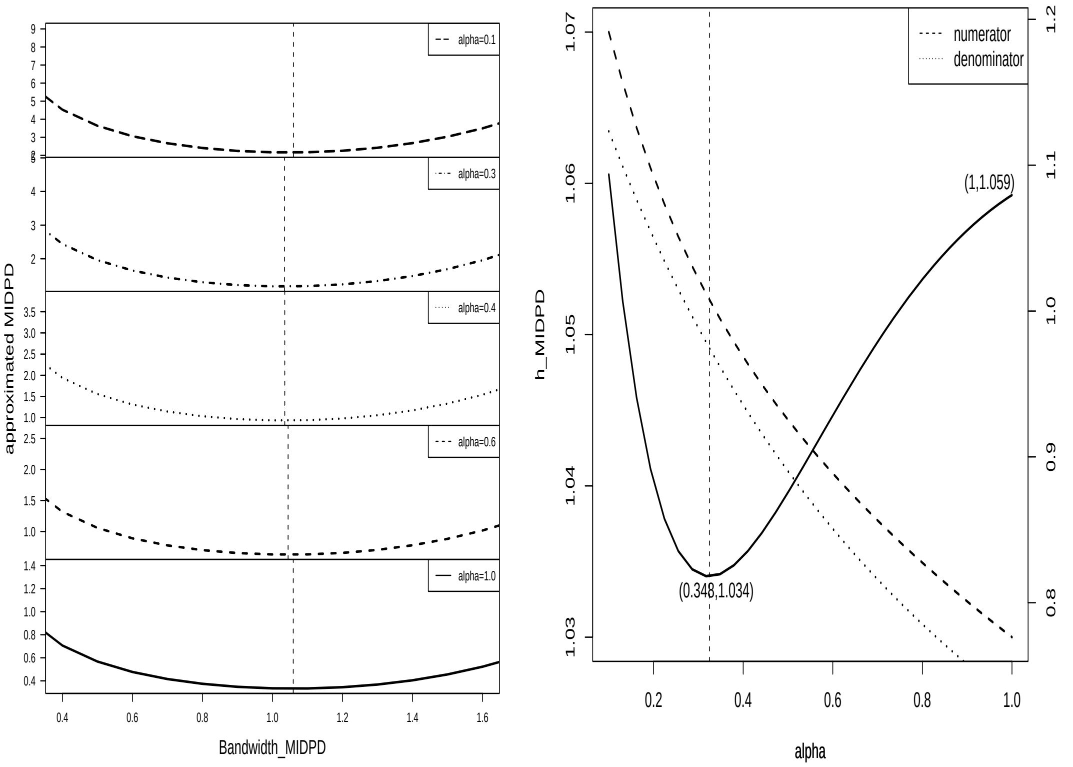
$$E \int \left\{ \hat{f}_n(x) - f(x) \right\} f(x)^{\alpha-1} \hat{f}_n^{(1)}(x) dx = 0, \tag{2.4}$$

to get the optimal bandwidth by the density power divergence and call it as $h_{\text{MIDPD}}(\alpha)$.

Proposition 1. *Under the same conditions of Silverman (1985), we have*

$$h_{\text{MIDPD}}(\alpha) = k_2^{-\frac{2}{5}} \left\{ \int K(t)^2 dt \int f(x)^\alpha dx \right\}^{\frac{1}{5}} \left\{ \int f^{(2)}(x)^2 f(x)^{\alpha-1} dx \right\}^{-\frac{1}{5}} n^{-\frac{1}{5}},$$

where $\alpha > 0$ *and* $k_2 = \int t^2 K(t) dt$. *We can easily figure out that* $h_{\text{MIDPD}}(1) = h_{\text{MISE}}$. *The proof is in the appendix.*



(a) Approx. MIDPD vs. $h_{\rm MIDPD}$

(b) $h_{\rm MIDPD}$ vs. α

Figure 1: *Relation of Approximated MIDPD, $h_{\rm MIDPD}$ and α; a kernel is Gaussian, and a density is $N(0,1)$. A dotted vertical line indicates a minimum.*

3. Discussion

3.1. $h_{\rm MIDPS}$ and α

In real situation we do not know what the true density or distribution would be, so that a very easy and natural choice of a density is the normal density. Similar to Silverman (1985), we have

$$h_{\rm MIDPD}(\alpha) = \left\{ \frac{1}{2\sqrt{\pi}} \frac{\sigma^{1-\alpha}(2\pi)^{\frac{1}{2}-\frac{\alpha}{2}}}{\alpha^{\frac{1}{2}}} \right\}^{\frac{1}{5}} \left\{ \frac{\sigma^{-4-\alpha}(2+\alpha^2)(2\pi)^{-\frac{\alpha}{2}}}{(1+\alpha)^{\frac{5}{2}}} \right\}^{-\frac{1}{5}} n^{-\frac{1}{5}} = \left\{ \frac{(1+\alpha)^5}{2\alpha(2+\alpha)^2} \right\}^{\frac{1}{10}} \sigma n^{-\frac{1}{5}},$$

if a kernel is Gaussian and a true density is the $N(\mu.\sigma^2)$. If $\alpha = 1$, $h_{\rm MIDPD} = (4/3)^{1/5}\sigma n^{-1/5}$, which is as same as $h_{\rm MISE}$ in Silverman (1985). $h_{\rm MIDPD}(\alpha)$ is then minimized around $\alpha = 0.35$.

Figure 1(a) displays the approximated MIDPD for various α when a kernel is Gaussian and a density is $N(0,1)$ and n is assumed to be 1 for illustration. We can observe that $h_{\rm MIDPS}$ differs by α. Each α produces its own optimal h which minimizes corresponding approximated MIDPD.

If we closely look at $h_{\rm MIDPD}$ in Proposition 1,

$$k_2^{-\frac{2}{5}} \left\{ \int K(t)^2 dt \int f(x)^\alpha dx \right\}^{\frac{1}{5}} \left\{ \int f^{(2)}(x)^2 f(x)^{\alpha-1} dx \right\}^{-\frac{1}{5}} n^{-\frac{1}{5}},$$

Table 1: K-S test statistics for various α's

α	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Min	0.070	0.067	0.068	0.068	0.068	0.067	0.070	0.069	0.071	0.071
Med	0.103	0.100	0.099	0.099	0.099	0.100	0.100	0.101	0.102	0.102
Max	0.135	0.133	0.133	0.133	0.132	0.134	0.134	0.134	0.136	0.136

it is just a ratio of the (approximated, integrated) variance to the (approximated, integrated) bias. As α approaches to 1, both the numerator and the denominator of h_{MIDPS} get smaller (Figure 1(b)), but the h_{MIDPS} has 1.034 as a minimum when $\alpha \doteq 0.35$. At $\alpha = 1$, the h_{MIDPS} is 1.059, which turns out to be same as h_{opt} by Silverman (1985, p45). The number 1.034 is the smallest available bandwidth, which properly weighting variance to bias. The α balances the trade-off between variance and bias, and we may consider this as the effect of α.

3.2. An empirical example

There may be no universal way of selecting an appropriate α as Basu *et al.* (1998) pointed out. In this article we propose a simple empirical way to find an appropriate α using a distributional adequacy statistics like Kolmogorov-Smirnov(K-S) test. As an example, we take the Old Faithful Geyser data containing waiting time between eruptions and the duration of the eruption for Old Faithful geyser in Yellowstone National Park, Wyoming, USA. We randomly select 100 observations with replacement from 272 eruption observations.

In this article, we borrow an idea of Durio and Isaia (2011) to select an α so as to test on the similarity between a density estimator and an empirical density but we use the well-known Kolmogorov-Smirnov test for similarity.

We get a density estimator for a random sample with $h_{\text{MIDPD}}(\alpha)$, $\alpha = 0.1, 0.2, \ldots, 1.0$, and then calculate K-S test statistics of a cumulative function of density estimator to an empirical distribution of a sample. We do it for 1000 times. Table 1 shows the K-S test results. The critical value of K-S test at 0.05% level of significance is $0.1360 (= 1.36/\sqrt{100})$. We can observed that $\alpha \in (0.3, 0.5)$, where the medians of K-S test are smallest, provide the statistically best fit for the Old Faithful data. We may choose any $\alpha \in (0.3, 0.5)$ for this particular example. In Section 3.1, we found that the proposed h_{MIDPS} is attained when $\alpha \doteq 0.35$ if a kernel is Gaussian and a true density is the $N(\mu.\sigma^2)$.

3.3. A simulation and computational matters

We run simulations to figure out the effect of α. Random samples (n = 10, 30 and 100) are generated from each density in Table 2. For each random sample, we calculate mean squared errors between a density estimator with $h_{\text{MIDPD}}(\alpha)$ and a true density. We do it 1000 times. Mean squared errors for each density are listed in Table 3. Overall, the numbers when $\alpha = 1$ are small. However, we can find smaller numbers even when α is small like $\alpha \in \{0.3, 0.4, 0.5\}$ and specially n is large. We may say that the h_{MIDPD} with certain α produces better fit than the widely used h_{MISE}. We can find larger numbers where a density is #3, #4 and #5 which have a sharp peak at the center or at the far left of a density. We guess that our computer program cannot cope with data from such densities quite well. Integrations were carried out numerically using trapezoidal rule (Kincaid and Cheney, 1991).

The exact MIDPD, $E[d_\alpha(f, \hat{f}_n)]$, and the approximated MIDPD (A.1) in the appendix are displayed in Figure 2 for various densities which are listed in Table 2. The exact and approximated MIDPD have similar parabolic shapes and the bandwidths that minimize both MIDPD are mostly very close. Overall, the approximated MIDPD is expected to give a bandwidth close to a exact bandwidth. The plots only for $\alpha = 1/3$ are displayed due to space limitations.

Table 2: Densities and corresponding notations; adapted from Marron and Wand (1992)

Density	Notation
#1 gaussian	$N(0,1)$
#2 skewed unimodal	$\frac{1}{5}N(0,1)+\frac{1}{5}N\left(\frac{1}{2},\left(\frac{2}{3}\right)^2\right)+\frac{3}{5}N\left(\frac{13}{12},\left(\frac{5}{6}\right)^2\right)$
#3 strongly skewed	$\sum_{l=0}^{7}\frac{1}{8}N\left(3\left\{\left(\frac{2}{3}\right)^l-1\right\},\left(\frac{2}{3}\right)^{2l}\right)$
#4 kurtotic unimodal	$\frac{2}{3}N(0,1)+\frac{1}{3}N\left(0,\left(\frac{1}{10}\right)^2\right)$
#5 outlier	$\frac{1}{10}N(0,1)+\frac{9}{10}N\left(0,\left(\frac{1}{10}\right)^2\right)$
#6 bimodal	$\frac{1}{2}N\left(-1,\left(\frac{2}{3}\right)^2\right)+\frac{1}{2}N\left(1,\left(\frac{3}{2}\right)^2\right)$
#7 separated bimodal	$\frac{1}{2}N\left(-\frac{3}{2},\left(\frac{1}{2}\right)^2\right)+\frac{1}{2}N\left(\frac{3}{2},\left(\frac{1}{2}\right)^2\right)$
#8 skewed bimodal	$\frac{3}{4}N(0,1)+\frac{1}{4}N\left(\frac{3}{2},\left(\frac{1}{3}\right)^2\right)$
#9 trimodal	$\frac{9}{20}N\left(-\frac{6}{5},\left(\frac{3}{5}\right)^2\right)+\frac{9}{20}N\left(\frac{6}{5},\left(\frac{3}{5}\right)^2\right)+\frac{1}{10}N\left(0,\left(\frac{1}{4}\right)^2\right)$

4. Conclusion

We employ the density power divergence to get an optimal bandwidth for kernel density estimators. The proposed method provides a large class of optimal bandwidths covering well-known bandwidths. We have discovered that h_{MIDPD} with a proper α provides an good density estimator for data. As Basu *et al.* (1998) showed that there may be no universal or general way of selecting an appropriate α. We select an appropriate α empirically; consequently, the density estimators with selected α's seem to fit the data sets better than the existing optimal bandwidth based on MISE. We admit the need to investigate the data-based estimation of the proposed bandwidth to make the results useful in practice. The data-based estimation is another research problem to be considered in future research.

Appendix: Proof of Proposition 1

The principal part of the proof is to find an approximated form of the left hand side of (2.3).

$$\begin{aligned} E\int \hat{f}_n(x)^\alpha \hat{f}_n^{(1)}(x)dx - E\int f(x)^\alpha \hat{f}_n^{(1)}(x)dx &= E\int \left\{\hat{f}_n(x)^\alpha - f(x)^\alpha\right\}\hat{f}_n^{(1)}(x)dx \\ &\approx E\int \alpha\left\{\hat{f}_n(x)-f(x)\right\}f(x)^{\alpha-1}\hat{f}_n^{(1)}(x)dx. \end{aligned}$$

As it has been already known to us,

$$\begin{aligned} E\hat{f}_n(x) &= \frac{1}{n}\sum_{i=1}^{n} E\frac{1}{h}K\left(\frac{x-X_i}{h}\right) = \int \frac{1}{h}K\left(\frac{x-y}{h}\right)f(y)dy \\ &= \int K(t)f(x-ht)dt \quad (y = x-ht). \end{aligned}$$

Table 3: Mean squared errors for various α under various models

Size	Density	α									
		0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
n=10	1	0.0045	0.0046	0.0046	0.0046	0.0046	0.0046	0.0046	0.0045	0.0045	0.0045
	2	0.0100	0.0104	0.0105	0.0105	0.0105	0.0104	0.0103	0.0102	0.0101	0.0100
	3	0.0350	0.0352	0.0353	0.0353	0.0352	0.0352	0.0352	0.0351	0.0350	0.0350
	4	0.0433	0.0437	0.0438	0.0438	0.0437	0.0437	0.0436	0.0435	0.0434	0.0433
	5	0.3839	0.3935	0.3963	0.3964	0.3952	0.3934	0.3913	0.3891	0.3868	0.3846
	6	0.0044	0.0046	0.0046	0.0046	0.0046	0.0045	0.0045	0.0045	0.0045	0.0044
	7	0.0103	0.0103	0.0103	0.0102	0.0103	0.0103	0.0103	0.0103	0.0103	0.0103
	8	0.0103	0.0103	0.0103	0.0102	0.0103	0.0103	0.0103	0.0103	0.0103	0.0103
	9	0.0045	0.0047	0.0047	0.0047	0.0047	0.0047	0.0046	0.0046	0.0046	0.0045
n=30	1	0.0094	0.0098	0.0099	0.0099	0.0098	0.0098	0.0097	0.0096	0.0095	0.0094
	2	0.0035	0.0035	0.0035	0.0035	0.0035	0.0035	0.0035	0.0035	0.0035	0.0035
	3	0.0269	0.0267	0.0267	0.0267	0.0267	0.0267	0.0268	0.0268	0.0268	0.0269
	3	0.0322	0.0321	0.0321	0.0321	0.0321	0.0321	0.0321	0.0322	0.0322	0.0322
	4	0.1641	0.1590	0.1577	0.1576	0.1582	0.1591	0.1601	0.1613	0.1625	0.1637
	5	0.0021	0.0021	0.0021	0.0021	0.0021	0.0021	0.0021	0.0021	0.0021	0.0021
	6	0.0076	0.0074	0.0073	0.0073	0.0074	0.0074	0.0074	0.0075	0.0076	0.0076
	7	0.0076	0.0074	0.0073	0.0073	0.0074	0.0074	0.0074	0.0075	0.0076	0.0076
	8	0.0023	0.0023	0.0023	0.0023	0.0023	0.0023	0.0023	0.0023	0.0023	0.0023
n=100	1	0.0073	0.0075	0.0076	0.0076	0.0076	0.0075	0.0075	0.0074	0.0074	0.0073
	2	0.0014	0.0014	0.0014	0.0014	0.0014	0.0014	0.0014	0.0014	0.0014	0.0014
	3	0.0218	0.0215	0.0215	0.0215	0.0215	0.0215	0.0216	0.0217	0.0217	0.0218
	4	0.0251	0.0247	0.0246	0.0246	0.0246	0.0247	0.0248	0.0249	0.0249	0.0250
	5	0.0845	0.0793	0.0780	0.0779	0.0785	0.0794	0.0805	0.0816	0.0828	0.0841
	6	0.0010	0.0010	0.0010	0.0010	0.0010	0.0010	0.0010	0.0010	0.0010	0.0010
	7	0.0050	0.0048	0.0047	0.0047	0.0047	0.0048	0.0048	0.0049	0.0049	0.0050
	8	0.0050	0.0048	0.0047	0.0047	0.0047	0.0048	0.0048	0.0049	0.0049	0.0050
	9	0.0013	0.0012	0.0012	0.0012	0.0012	0.0012	0.0012	0.0013	0.0013	0.0013

From the Taylor's expansion, we have

$$f(x-ht) = f(x) - htf^{(1)}(x) + \frac{1}{2}h^2t^2f^{(2)}(x) - \frac{1}{6}h^3t^3f^{(3)}(x) + \frac{1}{24}h^4t^4f^{(4)}(x) + O\left(h^5\right).$$

Then, the expected value of the kernel estimator is written as

$$E\hat{f}_n(x) = f(x)\left[1 + \frac{1}{2}h^2\frac{f^{(2)}(x)}{f(x)}\int t^2K(t)dt + \frac{1}{24}h^4\frac{f^{(4)}(x)}{f(x)}\int t^4K(t)dt + O\left(h^5\right)\right],$$

and

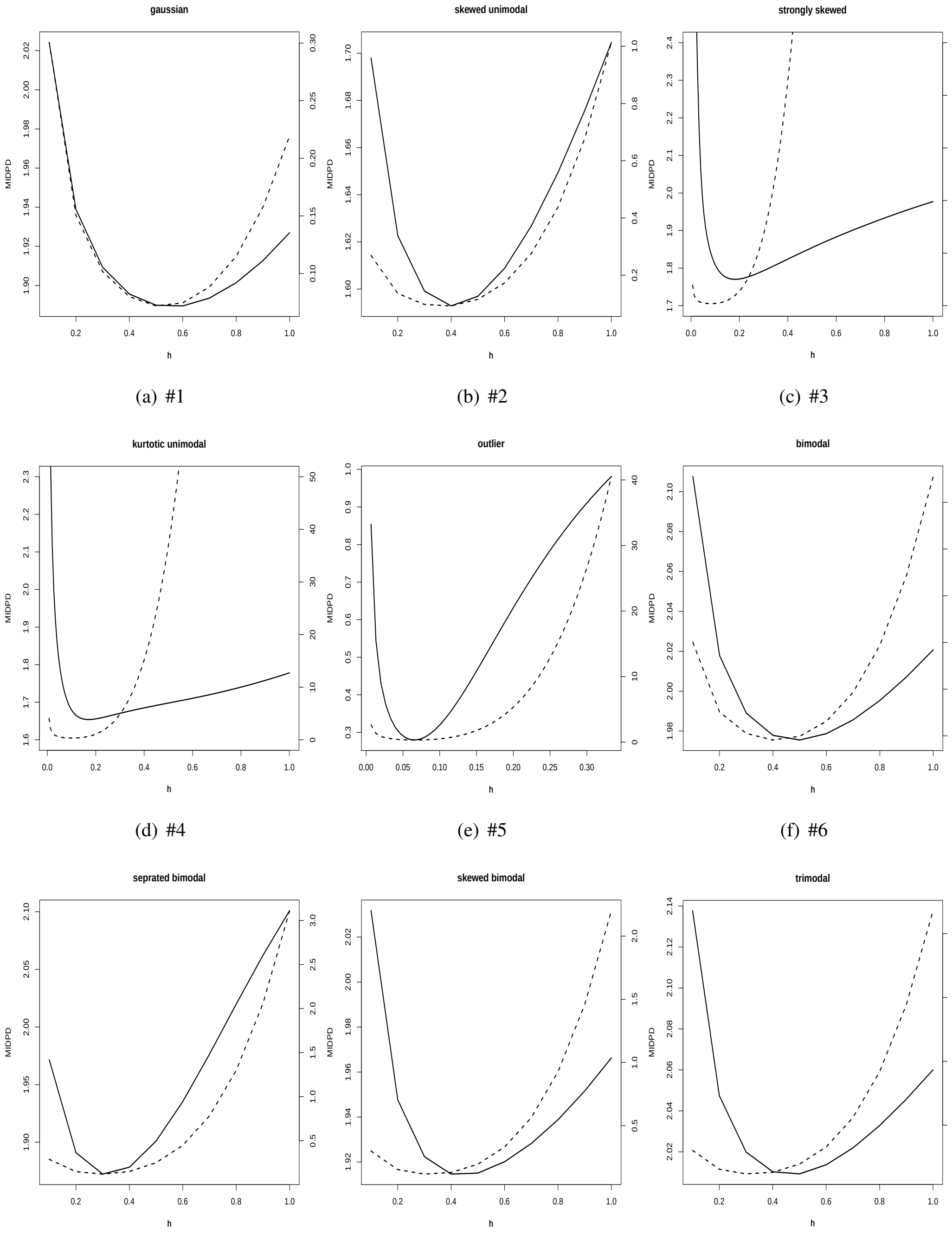
$$\mathrm{Var}\hat{f}_n(x) = \frac{f(x)}{nh}\int K(t)^2dt + O\left(n^{-1}\right).$$

With a random variable $\xi = O_p(1)$ whose expectation is 0 and variance 1, we can write $\hat{f}_n(x)$ as

$$\hat{f}_n(x)=f(x)\left[1+\frac{1}{2}h^2\frac{f^{(2)}(x)}{f(x)}\int t^2K(t)dt+\frac{1}{24}h^4\frac{f^{(4)}(x)}{f(x)}\int t^4K(t)dt+O\left(h^5\right)+\left\{\frac{\int K(t)^2dt}{nhf(x)}\right\}^{\frac{1}{2}}\xi+O_p\left(n^{-\frac{1}{2}}\right)\right],$$

and by taking a derivative of $\hat{f}_n(x)$ *w.r.t.* h we have

$$\hat{f}_n^{(1)}(x)=\left[hf^{(2)}(x)\int t^2K(t)dt+\frac{4}{24}h^3f^{(4)}(x)\int t^4K(t)dt+O\left(h^4\right)-\frac{1}{2}\left\{\frac{\int K(t)^2dt}{nh^3f(x)}\right\}^{\frac{1}{2}}f(x)\xi+O_p\left(n^{-\frac{1}{2}}\right)\right].$$



(a) #1 (b) #2 (c) #3

(d) #4 (e) #5 (f) #6

(g) #7 (h) #8 (i) #9

Figure 2: *Exact MIDPD (solid lines, left axises) and approximated MIDPD (dotted lines, right axises) for h with $\alpha = 1/3$; $n = 30$.*

Therefore,

$$\int \hat{f}_n(x) f(x)^{\alpha-1} \hat{f}_n^{(1)}(x) dx = \int f(x)^{\alpha} \left[h f^{(2)}(x) \int t^2 K(t) dt + \frac{1}{2} h^3 \frac{f^{(2)}(x)^2}{f(x)} \left\{ \int t^2 K(t) dt \right\}^2 \right.$$
$$\left. + \frac{4}{24} h^3 f^{(4)}(x) \int t^4 K(t) dt + O\left(h^4\right) + A\xi - B\xi^2 + O_p\left(n^{-\frac{1}{2}}\right) \right] dx,$$

where A is complex but vanishes upon taking expectation for $E[\xi] = 0$ and we have $B = (2nh^2)^{-1} \int K^2 (t)dt$.

Next,

$$\int f(x)^{\alpha} \hat{f}_n^{(1)}(x) dx = \int f(x)^{\alpha} \left[h f^{(2)}(x) \int t^2 K(t) dt + \frac{4}{24} h^3 f^{(4)}(x) \int t^4 K(t) dt \right.$$
$$\left. + O_p\left(n^{-\frac{1}{2}}\right) + O(h^4) + \frac{1}{2} \left\{ \frac{\int K(t)^2 dt}{nh^3 f(x)} \right\}^{\frac{1}{2}} f(x)\xi \right] dx,$$

where the last term vanishes upon taking expectation for $E[\xi] = 0$.

If we assume that h is small and n is large, we have

$$E \int \left\{ \hat{f}_n(x) - f(x) \right\} f(x)^{\alpha-1} \hat{f}_n^{(1)}(x) dx \approx \frac{1}{2} h^3 \left\{ \int t^2 K(t) dt \right\}^2 \int f^{(2)}(x)^2 f(x)^{\alpha-1} dx - \frac{1}{2nh^2} \int K(t)^2 dt \int f(x)^{\alpha} dx.$$

Solving the above equation for h gives $h_{\text{MIDPD}}(\alpha)$, which is

$$k_2^{-\frac{2}{5}} \left\{ \int K(t)^2 dt \int f(x)^{\alpha} dx \right\}^{\frac{1}{5}} \left\{ \int f^{(2)}(x)^2 f(x)^{\alpha-1} dx \right\}^{-\frac{1}{5}} n^{-\frac{1}{5}}.$$

Appendix: An Alternative Proof of Proposition 1

The Equation (2.4) can be rewritten as

$$E \int \alpha \left\{ \hat{f}_n(x) - f(x) \right\} f(x)^{\alpha-1} \hat{f}_n^{(1)}(x) dx = \frac{\alpha}{2} \frac{d}{dh} \int E\left[\left\{ \hat{f}_n(x) - f(x) \right\}^2 \right] f(x)^{\alpha-1} dx = 0.$$

The solution to the above equation gives $h_{\text{MIDPD}}(\alpha)$. According to Silverman (1985) we have the approximated equation

$$E\left[\left\{ \hat{f}_n(x) - f(x) \right\}^2 \right] \approx \frac{1}{4} h^4 k_2^2 f^{(2)}(x)^2 + \frac{1}{nh} f(x) \int K(t)^2 dt,$$

and then we have

$$\int E\left[\left\{ \hat{f}_n(x) - f(x) \right\}^2 \right] f(x)^{\alpha-1} dx \approx \frac{1}{4} h^4 k_2^2 \int f^{(2)}(x)^2 f(x)^{\alpha-1} dx + \frac{1}{nh} \int f(x)^{\alpha} dx \int K(t)^2 dt. \quad \text{(A.1)}$$

Take a derivative of (A.1) in the appendix w.r.t. h and set it equal to zero. The solution to that equation for h turns out

$$k_2^{-\frac{2}{5}} \left\{ \int K(t)^2 dt \int f(x)^{\alpha} dx \right\}^{\frac{1}{5}} \left\{ \int f^{(2)}(x)^2 f(x)^{\alpha-1} dx \right\}^{-\frac{1}{5}} n^{-\frac{1}{5}},$$

which is h_{MIDPD} and is equal to the optimal h of Parzen (1962) when $\alpha = 1$.

References

Basu, A., Harris, I. R., Hjort, N. L. and Jones, M.C. (1998). Robust and efficient estimation by minimizing a density power divergence, *Biometrika*, **85**, 549–559.

Basu, A., Mandal, A., Martin, N. and Pardo, L. (2013). Testing statistical hypotheses based on the density power divergence, *Annals of the Institute of Statistical Mathematics*, **65**, 319–348.

Devroye, L. and Györfi, L. (1985). *Nonparametric Density Estimation: The L_1 View*, Wiley, New York.

Durio, A. and Isaia, E. D. (2011). The Minimum density power divergence approach in building robust regression models, *Informatica*, **22**, 43–56.

Fujisawa, H. and, Eguchi, F. (2006). Robust estimation in the normal mixture model, *Journal of Statistical Planning and Inference*, **136**, 3989–4011.

Hall, P. (1987). On Kullback-Leibler loss and density estimation, *The Annals of Statistics*, **15**, 1491–1519.

Kanazawa, Y. (1993). Hellinger distance and Kullback-Leiber loss for the kernel density estimator, *Statistics and Probability Letters*, **18**, 315–321.

Kincaid, D. and Cheney, W. (1991). *Numerical Analysis: Mathematics of Scientific Computing*, Brooks/Cole, New York.

Lee, S. and Na, O. (2005). Test for parameter change based on the estimator minimizing density-based divergence measures, *Annals of the Institute of Statistical Mathematics*, **57**, 553–573.

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

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

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

Silverman, B. W. (1985). *Density Estimation for Statistics and Data Analysis*, Chapman and Hall\CRC, New York.

Warwick, J. and Jones, M. C. J. (2005). Choosing a robustness tuning parameter, *Journal of Statistical Computation and Simulation*, **75**, 581-588.

Received July 8, 2014; Revised August 13, 2014; Accepted August 17, 2014