THE EFFECT OF NON-NORMALIZATION ON THE RISKS OF THE DENSITY ESTIMATORS

by

Herman Rubin[1]
Purdue University

and

Jeesen Chen
University of Cincinnati

Technical Report #86-50

Department of Statistics
Purdue University

1986

[1] Research supported by the National Science Foundation under Grant DMS-8401996.

THE EFFECT OF NON-NORMALIZATION ON THE RISKS OF THE DENSITY ESTIMATORS

Herman Rubin
Purdue University

and

Jeesen Chen
University of Cincinnati

ABSTRACT

When using $\hat{f}$ to estimate the unknown density function f, losses of the form $H_\alpha(f,\hat{f}) = (\int |f^\beta - \hat{f}^\beta|^\alpha d\mu)^{\frac{1}{\alpha}}$, $\beta = \frac{1}{\alpha}$ are considered. For any f and $\hat{f}$, let $U_\alpha(f,\hat{f}) = \inf\{L_p(f,c\hat{f})$: $c \geq 0\}$. Then if f and $\hat{f}$ are densities, $\alpha \geq 1$, $U_2 = L_2(1-L_2/4)$; $U_\alpha \geq 2^{-\alpha}L_\alpha$. In general, if L_α is small, then U_α closed to L_α. This implies that in the search for good density estimators for losses of the type L_α, normalization can be ignored.

AMS 1980:

Keywords:

THE EFFECT OF NON-NORMALIZATION ON THE RISKS OF THE DENSITY ESTIMATORS

by

Herman Rubin[1]
Purdue University

and

Jeesen Chen
University of Cincinnati

§1. Motivation

A density estimator is a sequence of measurable functions $\{f_n\colon\ n \geq 1\}$ for which $f_n(t) = f_n(t; X_1, X_2, \ldots, X_n)$ being used to estimate the unknown, common density function of a random sample of size n, $\{X_1, X_2, \ldots, X_n\}$. The performance of the density estimator $\{f_n\}$ is measured by a loss function $L(f_n, f)$, which reflects the goodness of fit of f_n to f. With respect to a fixed loss function L, one density estimator is to be considered as better than the other if it has smaller expected loss $EL(f_n, f)$, or, in other words, smaller risk. Many studies focus on how to obtain density estimators $\{f_n\}$ for which the risk converges to zero, as the sample size n approaches to infinite, at very high rate. The common used loss functions are integral square error: $ISE(f_n, f) = \int (f_n(t) - f(t))^2 dt$, and the sup-norm: $L_\infty(f_n, f) = \sup_{-\infty<t<\infty} |f_n(t) - f(t)|$. Other loss functions such as $L_\alpha(f_n, f) = \int ||f_n(t)|^{\frac{1}{\alpha}} - f(t)^{\frac{1}{\alpha}}|^\alpha dt$, for $\alpha \geq 1$, and $H_\alpha(f_n, f) = (L_\alpha(f_n, f))^{\frac{1}{\alpha}}$ gradually received attentions. We deem them as appropriate for measuring the global deviation of f_n from f because, as described in page 255 of Devroye and Györfi (1985), they shame two nice properties: it is always finite and it is invariant under the strictly monotone transformation.

Since being a density function, f is nonnegative and with total mass 1, i.e.

$$\int_{-\infty}^{\infty} f(t)dt = 1,$$

it is unnatural to consider those density estimators which do not meet these two conditions. However, some existing density estimators fall in this category. Some statisticians study this type of density estimators because they focus their attention on the local properties of a density function, for this situation the total mass assumption plays little role. For others, they emphasis on obtaining higher risk convergent rates, and it is known that by relaxing the requirements of density estimators, one might be able to improve the risk convergent rates on those existing density estimators. To illustrate the second situation, let us consider the performance of the kernal density estimators f_n (defined as (1)) under the integral square error loss function. The definition of the kernal density estimator $\{f_n\}$ is

$$f_n(t) = \frac{1}{nh_n} \sum_{i=1}^{n} K\left(\frac{t - X_i}{nh_n}\right),\ t \in \mathbb{R} \tag{1}$$

[1] Research supported by the National Science Foundation under Grant DMS-8401996.

where $\{h_n\}$ is a sequence of positive numbers, and the function K is called the kernal function of the estimators. It is easy to see that in order that f_n to be a density, K must satisfy nonnegative assumption and mass 1 assumption. Parzen (1962) proved that, for positive integer r, there are kernals K such that the risk

$$IMSE(n) = E\int (f_n(t) - f(t)^2 dt = 0(n^{-2r/(2r+1)}). \tag{2}$$

However, for $r > 2$, the kernal K which satisfies (2) should assume some negative values. Therefore, for a bona-fide kernal density estimator, the optimal convergent rate is $IMSE(n) = 0(n^{-4/5})$, while giving up the nonnegative assumption on f_n, we are able to obtain higher convergence rate. Terrell and Scott ( ), by relaxing the other requirement (mass 1), obtained a rate, $IMSE(n) = 0(n^{-8/9})$ which is again better than $0(n^{-4/5})$. And Davis ( ), by allowing kernal violating both assumptions, even obtained the rate, $IMSE(n) = 0(\log n/n)$. (Recall the Boyd-Steel ((1978) show that the optimal rate of $IMSE(n)$ for any kind of density estimators could't higher than $\frac{1}{n}$.

From the above discussion, it is interesting to find an operation, μ, on not-bona-fide density estimators f_n that will result a bona-fide density function μf_n such that the risk of μf_n, $EL(\mu f_n, f)$, converges to zero at a rate not more than that of $EL(f_n, f)$. To achieve nonnegative assumption, we may just use $|f_n|$ or f_n^+ to replace the original f_n since many reasonable loss functions satisfy either

(a) $l(f_n^+, f) \leq L(f_n, f)$ or

(b) $L(|f_n|, f) \leq L(f_n, f)$.

Therefore, from now on, we may assume all the density estimators we discuss satisfy the nonnegative assumption.

For a non-negative density estimator f_n, suppose that the total mass

$$\gamma(f_n) = \int_{-\infty}^{\infty} f_n(t)dt$$

is finite (which is true for most cases, see proposition ???), we defined

$$\mu f_n(t) = \gamma(f_n)^{-1} f_n(t), \tag{3}$$

then $\mu f_n(\cdot)$ is a bona-fide density function.

The main result of this paper is:

> "For $\alpha > 0$, if $EL_\alpha(f_n, f) \to 0$ as $n \to \infty$, then $EL_\alpha(\mu f_n, f) \to 0$ as $n \to \infty$. Furthermore, if $EL_\alpha(f_n, f)$ is small, then $EL_\alpha(\mu f_n, f)$ is comparable with $EL_\alpha(f_n, f)$ in the sense that
>
> $$U(\alpha; f_n, f) = EL_\alpha(f_n, f)/EL_\alpha(\mu f_n, f)$$
>
> closes to 1."

We reach the above conclusion by letting $g = \mu f_n$, and studying how much improvement can be if we de-normalized g. More precisely, we compare $\inf_{\lambda>0} L_\alpha(\lambda g, f)$ with $L_\alpha(g, f)$. In section 2, by a simple method establish that a lower bound of $W(\alpha; f, g) = \inf_{\lambda>0} L_\alpha(\lambda g, f)/L_\alpha(g, f)$ is $2^{-\alpha}$. This is good enough to establish the first part of our main results. In section 3 by a more elaborate method, we find the value $V_\alpha = \inf\{W(\alpha; g, f)$: g and f are density functions.$\}$. We also study $V_\alpha(h) = \inf\{W(\alpha; g, f)$: g and f are density functions, and $L_\alpha(g, f) = h\}$, especially for the case $h \to 0$. Since V_α is a crude bound for $L_\alpha(f_n, f)/L_\alpha(\mu f_n, f)$, if V_α close to 1. The value $L_\alpha(f_n, f)/L_\alpha(\mu f_n, f)$ is even more lost to 1 then V_α. This establish our second part of main results.

An application of the main results in density estimator theory is: if the loss function is L_α (or H_α), we may focus our study on those density estimator performed very well locally, then apply the normalization procedure to produce a bona-fide density estimator, and still having very high risk convergence rate.

§2. A Crude Inequality Involving L_α

Recall that if $\alpha > 0$, the α-norm of a function Q is defined as $\|Q\|_\alpha = (\int |Q(t)|^\alpha dt)^{\frac{1}{\alpha}}$. For nonnegative functions f and g,

$$L_\alpha(g, f) = \int |g^{\frac{1}{\alpha}}(t) - f^{\frac{1}{\alpha}}(t) - f^{\frac{1}{\alpha}}(t)^\alpha dt = \|g^{\frac{1}{\alpha}} - f^{\frac{1}{\alpha}}\|_\alpha^\alpha. \tag{4}$$

Proposition 1: If f and g are density functions, $\lambda \geq 0$, then

$$L_\alpha(\lambda g, f) \geq 2^{-\alpha} L_\alpha(g, f). \tag{5}$$

Proof: Notice that $\|g^{\frac{1}{\alpha}}\|_\alpha = \|f^{\frac{1}{\alpha}}\|_\alpha = 1$.

Since

$$\begin{aligned}(L_\alpha(g, f))^{\frac{1}{\alpha}} &= \|g^{\frac{1}{\alpha}} - f^{\frac{1}{\alpha}}\|_\alpha \\ &= \|g^{\frac{1}{\alpha}} - (\lambda g)^{\frac{1}{\alpha}} + (\lambda g)^{\frac{1}{\alpha}} - f^{\frac{1}{\alpha}}\|_\alpha \\ &\leq |1 - \lambda^{\frac{1}{\alpha}}| \|g^{\frac{1}{\alpha}}\|_\alpha + \|(\lambda g)^{\frac{1}{\alpha}} - f^{\frac{1}{\alpha}}\|_\alpha \\ &= |1 - \lambda^{\frac{1}{\alpha}}| + \|(\lambda g)^{\frac{1}{\alpha}} - f^{\frac{1}{\alpha}}\|_\alpha, \end{aligned} \tag{6}$$

and, by the triangle inequality of the norm $\| \ \|_\alpha$,

$$\begin{aligned}\|(\lambda g)^{\frac{1}{\alpha}} - f^{\frac{1}{\alpha}}\|_\alpha geq |\|(\lambda g)^{\frac{1}{\alpha}}\|_\alpha - \|f^{\frac{1}{\alpha}}\|_\alpha| \\ = |\lambda^{\frac{1}{\alpha}} - 1|, \end{aligned} \tag{7}$$

therefore

$$\begin{aligned}(L_\alpha(g, f))^{\frac{1}{\alpha}} &\leq 2\|(\lambda g)^{\frac{1}{\alpha}} - f^{\frac{1}{\alpha}}\|_\alpha \\ &= 2(L_\alpha(\lambda g, f))^{\frac{1}{\alpha}}, \ q.e.d. \end{aligned} \tag{8}$$

Remark: The above proof uses the norm property of $\| \ \|_\alpha$, although we suspect the inequality $\inf_{\lambda>0} \|\lambda g - f\|_\alpha \geq \frac{1}{2}\|g-f\|_\alpha$, is well known, but we couldn't locate the reference.

An interesting application of the above inequality is:

Corollary 2: If f is a density function and $\{g_n\}$ is a sequence of nonnegative functions such that $L_\alpha(f_n, f) \to 0$ as $n \to \infty$, then

$$\gamma(g_n) = \int_{-\infty}^{\infty} g_n(t)dt \to \quad as\ n \to \infty, \tag{9}$$

and

$$L_\alpha(\mu g_n, f) \leq 2^\alpha L_\alpha(g_n, f) \to 0\ as\ n \to 0. \tag{10}$$

Proof: Since

$$\begin{aligned} |\gamma(g_n)^{\frac{1}{\alpha}} - 1| &= \left| \|g_n^{\frac{1}{\alpha}}\|_\alpha - \|f^{\frac{1}{\alpha}}\|_\alpha \right| \\ &\leq \|g_n^{\frac{1}{\alpha}} - f^{\frac{1}{\alpha}}\|_\alpha \\ &= (L_\alpha(g_n, f))^{\frac{1}{\alpha}} \\ &\to 0 \text{ as } n \to \infty, \end{aligned} \tag{11}$$

we prove (9). (10) is a simple consequence of (5), q.e.d.

Since most interesting density estimator f_n satisfies $EL_\alpha(f_n, f) \to 0$ as $n \to \infty$, according to the corollary 2, $\gamma(f_n) \to 0$ in probability as $n \to \infty$.

§3. A Refine Inequality Involving L_α

For density functions f and g, let $U_\alpha(g, f) = \inf\{L_\alpha(\lambda g, f)\colon \lambda > 0\}$. The proposition 1 gives us a lower bound of $U_\alpha(g, f)/L_\alpha(g, f)$; that is $2^{-\alpha}$. In the following, we want to study the best possible C_α such that $U_\alpha(g, f) \geq C_\alpha L_\alpha(g, f)$ for all densities f and g.

First, for the case $\alpha = 2$, i.e. L_α is the square of Kakutani-Hellinger distance, we have the following:

Proposition 3: $U_2(g, f) = L_2(g, f) - \frac{1}{4}(L_2(g, f))^2$.

Proof:

$$\begin{aligned} L_2(\lambda g, f) &= \int_{-\infty}^{\infty} (\sqrt{\lambda g(t)} - \sqrt{f(t)})^2 dt \\ &= \int_{-\infty}^{\infty} \lambda g(t)dt + \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(t)dt - 2\int_{-\infty}^{\infty} \sqrt{\lambda g(t) f(t)}dt \\ &= \lambda + 1 - 2\sqrt{\lambda}\int_{-\infty}^{\infty} f(t)dt\sqrt{g(t)f(t)}dt \\ &= (\sqrt{\lambda} - \int_{-\infty}^{\infty} \sqrt{g(t)f(t)}dt)^2 + (1 - (\int_{-\infty}^{\infty} \sqrt{g(t)dt}dt)^2). \end{aligned}$$

Hence

$$U_2(g,f) = 1 - (\int_{-\infty}^{\infty} \sqrt{g(t)f(t)}dt)^2. \tag{12}$$

Since

$$\begin{aligned} L_2(g,f) &= \int_{-\infty}^{\infty} g(t)dt + \int_{-\infty}^{\infty} f(t)dt - 2\int_{-\infty}^{\infty} \sqrt{g(t)f(t)}dt \\ &= 2 - 2\int_{-\infty}^{\infty} \sqrt{g(t)f(t)}dt, \end{aligned} \tag{13}$$

combining (12) and (13), we have

$$\begin{aligned} U_2(g,f) &= 1 - (1 - \frac{1}{2}L_2(g,f))^2 \\ &= L_2(g,f) - \frac{1}{4}(L_2(g,f))^2 .q.e.d. \end{aligned} \tag{14}$$

For other value α, we need the following two lemmas.

Lemma 4: There exists a nonnegative random variable Z such that

$$E(Z^\alpha) = 1 \tag{15}$$

and

$$E|Z - \lambda|^\alpha = L_\alpha(\lambda^\alpha g, f) \text{ for all } \lambda > 0. \tag{16}$$

Proof: Suppose Y is a random variable with density function g, and m is a positive integer. Define

$$Z_m = \{f(Y)/[g(Y) + m^{-1}]\}^{\frac{1}{\alpha}}. \tag{17}$$

It can be shown that $\{Z_m\colon\ m = 1,2,\ldots\}$ converges in L_α to a random variable Z. For this Z, (15) and (16) are true.q.e.d.

The distribution of Z^α can be represented as an average of two-point, mean 1 distribution i.e. there is a probability space (Z,Σ,m) such that X is an interval, Σ is the Borel subset of X; for each t, G_t represent the distribution of a two-point, mean 1 distribution, and

$$P(Z^\alpha \le x) = \int_{t\in X} G_t(x)m(dt). \tag{18}$$

Proof: This is a simple consequence of Freedman (1971), page 68, the Lemma (108).

Lemma 6: Let

$$c_\alpha \equiv \inf\{U_\alpha(g,f)/L_\alpha(g,f): \quad f \text{ and } g \text{ are densities}\},$$

and

$$d_\alpha \equiv \inf\{E(|W-\lambda|^\alpha)/E(|W-1|^\alpha):$$
$$W \text{ is a two-point, mean 1 random variable with } EW^\alpha = 1;\ \lambda > 0\}.$$

Then $c_\alpha = d_\alpha$.

Proof: From the definition of d_α, for any two-point, mean 1 random variable W,

$$E(|W-\lambda|^\alpha) \geq d_\alpha E(|W-1|^\alpha). \tag{19}$$

Hence, for all $t \in \mathcal{X}$,

$$\int |w-\lambda|^\alpha G_t(dw) \geq d_\alpha \int |w-1|^\alpha g_t(dw).$$

We have

$$\begin{aligned} L_\alpha(\lambda^\alpha g, f) &= E|Z-\lambda|^\alpha \\ &= \int_{t\in\mathcal{X}} \int |z-\lambda|^\alpha G\prime_t(z^\alpha)\alpha z^{\alpha-1} dz m(dt) \\ &\geq \int_{t\in\mathcal{X}} d_\alpha \int |z-1|^\alpha G\prime_t(z^\alpha)\alpha z^{\alpha-1} dz m(dt) \\ &= d_\alpha E|Z-1|^\alpha \\ &= d_\alpha L_\alpha(g,f). \end{aligned}$$

According to the definition of c_α,

$$c_\alpha \geq d_\alpha. \tag{20}$$

Since $c_\alpha \leq d_\alpha$ is trivial due to the Lemma 4, we conclude

$$c_\alpha = d_\alpha . q.e.d. \tag{21}$$

In the following, let

$$\begin{aligned} \Psi(\alpha;h) = \inf\{&E|W-\lambda|^\alpha: \\ &W \text{ is a two point, mean 1 random variable with} \\ &EW^\alpha = 1 \text{ and } E|W-1|^\alpha = h\}, \end{aligned} \tag{22}$$

and let $W(p;\xi,\eta)$ denote a two point random variable

$$\begin{cases} P(W(p;\xi,\eta)=\xi) = p \\ P(W(p;\xi,\eta)=\eta) = 1-p \equiv q. \end{cases} \tag{23}$$

Obviously,

$$c_\alpha = d_\alpha = \inf_{h>0}(\Psi(\alpha;h)/n). \tag{24}$$

For fixed h and fixed pp let us find λ^α which minimize $E|W-\lambda|^\alpha$ for $W = W(p;\xi,\eta)$ with $EW^\alpha = 1$ and $E|W-1|^\alpha = h$. Notice that

$$P\xi^\alpha + q\eta^\alpha = 1, \tag{25}$$

$$P(1-\xi)^\alpha + q(\eta-1)^\alpha = h, \tag{26}$$

and

$$E|W-\lambda|^\alpha = p|\lambda-\xi|^\alpha + q|\eta-\lambda|^\alpha. \tag{27}$$

We may assume $\xi \le 1 \le \eta$ without loss generality. From (27), it is easy to show that

$$\lambda^* = (q^{\frac{1}{\beta}}\eta + p^{\frac{1}{\beta}}\xi)/(p^{\frac{1}{\beta}} + q^{\frac{1}{\beta}}) \text{ for } \beta = \alpha - 1. \tag{28}$$

Hence

$$\lambda^* - \xi = q^{\frac{1}{\beta}}(\eta-\xi)/(p^{\frac{1}{\beta}} + q^{\frac{1}{\beta}}), \tag{29}$$

and

$$\eta - \lambda^* = p^{\frac{1}{\beta}}(\eta-\xi)/(p^{\frac{1}{\beta}} + q^{\frac{1}{\beta}}). \tag{30}$$

Therefore

$$\begin{aligned}
&\inf_{\lambda>0} E|W(p;\xi,\eta)-\lambda|^\alpha \\
&\quad = (pq^{\frac{\alpha}{\beta}} + qp^{\frac{\alpha}{\beta}})(\eta-\xi)^\alpha(p^{\frac{1}{\beta}} + q^{\frac{1}{\beta}})^{-\alpha} \\
&\quad pq(\eta-\xi)^\alpha(p^{\frac{1}{\beta}} + q^{\frac{1}{\beta}})^{-\beta} \\
&\quad pq((\eta-1)+(1-\xi))^\alpha(p^{\frac{1}{\beta}} + q^{\frac{1}{\beta}})^{-\beta},\ \beta = \alpha-1.
\end{aligned} \tag{31}$$

Combining (22), (24), (25), (26), and (31), we have

$$\begin{aligned}
c_\alpha = \inf\{&pq((\eta-1)+(1-\xi))^\alpha(p^{\frac{1}{\beta}} + q^{\frac{1}{\beta}})^{-\beta}h^{-1} : \\
&h>0,\ \beta=\alpha-1,\ 0\le p\le 1,\ p\xi^\alpha + q\eta^\alpha = 1,\ \text{and} \\
&P(1-\xi)^\alpha + q(\eta-1)^\alpha = h\}.
\end{aligned} \tag{32}$$

We are unable to provide a simple formula for c_α. But, since for the most interesting case is that $h = E|W-1|^\alpha$ being small, instead of considering c_α (defined in (24)), we compute

$$\tilde{c}_\alpha = \lim_{h\to 0}\frac{\Psi(\alpha;h)}{h}. \tag{33}$$

For fixed p, and let $h \to 0$, from (26), we have $\xi \to 1-$ and $\eta \to 1+$. Since

$$p(1-\xi) \sim q(\eta-1) \text{ as } n \to 0, \tag{34}$$

and hence

$$h \sim p[1 + (p/q)^{\alpha+1}](1-\xi)^{\alpha} \text{ as } n \to 0. \tag{35}$$

By (31), (34) and (35), we have

$$\inf_{\lambda>0} E|W(p;\xi,\eta) - \lambda|^{\alpha}/h \to (p^{\frac{1}{\beta}} + q^{\frac{1}{\beta}})^{-\beta}(p^{\beta} + q^{\beta})^{-1} \text{ as } h \to 0. \tag{36}$$

We conclude that:

$$\lim_{h\to 0} \frac{\Psi(\alpha;h)}{h} = \tilde{c}_{\alpha} = \inf\{(p^{\frac{1}{\beta}} + q^{\frac{1}{\beta}})^{-\beta}(p^{\beta} + q^{\beta})^{-1} : \ \beta = \alpha - 1,\ 0 \le p \le 1,\ q = 1-p\}. \tag{37}$$

In terms of $L_{\alpha}(g,f)$, we have:

Conclusion: If $h = L_{\alpha}(g,f)$ is small, then

$$\inf_{\lambda>0} L_{\alpha}(\lambda g, f) = \tilde{c}_{\alpha} h + o(h). \tag{39}$$

The constant c_{α} has a lower bound

$$\begin{aligned} \tilde{c}_{\alpha} &\ge \inf_{\substack{p+q=1 \\ 0\le p\le 1}} (p^{\frac{1}{\beta}} + q^{\frac{1}{\beta}})^{-\beta} \inf_{\substack{p+q=1 \\ 0\le p\le 1}} (p^{\beta} + q^{\beta})^{-1} \\ &= 2^{-(d-2)} \end{aligned} \tag{40}$$

Numerically study shows:

α	$\tilde{c}_{\alpha}$	$2^{-\alpha}$
1	0.5	0.5
1.5	0.8718	0.3535
2	1	0.25
3	0.7602	0.125
4	0.4641	0.0625
5	0.2666	0.03125

Reference

Davis, K.B. (1977) Mean integrated square error properties of density estimates, *Annals of Statistics* **5**, pp. 530-535.

Devroye, L. and Györfi, L. (1985) *Nonparametric Density Estimation; the L_1 View.* John Wiley & Sons, Inc., New York.

Freedman, D. (1971) *Brownian Motion and Diffusion.* Holden-Day, Inc.

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

Terrell, G.R. and Scott, D.W. (1980) On improving convergence rates for nonnegative kernel density estimators, *Annals of Statistics* **8**, pp. 1160-1163.