DIGITALCOMMONS
@WAYNESTATE

Journal of Modern Applied Statistical Methods

Volume 11 | Issue 1 Article 24

5-1-2012

The Weighted Hellinger Distance for Kernel Distribution Estimator of Function of Observations

Abdel-Razzaq Mugdadi
Jordan University of Science and Technology, Irbid, Jordan

Journal of Modern Applied Statistical Methods
May 2012, Vol. 11, No. 1, 268-273

1538 – 9472/12/$95.00

The Weighted Hellinger Distance for Kernel Distribution Estimator of Function of Observations

Abdel-Razzaq Mugdadi
Jordan University of Science and Technology,
Irbid, Jordan

The asymptotic mean weighted Hellinger distance (AMWHD) is derived for the kernel distribution estimator of a function of observations. In addition, the AMWHD is compared with the asymptotic mean integrated square error (AMISE) of the estimator. A completely data based method is proposed to select the bandwidth in the estimator using the mean weighted Hellinger distance (MWHD).

Key words: Kernel estimation, distribution function estimation, bandwidth, Hellinger distance, mean square error, function of random variables.

Introduction

Given a random sample $X_1, X_2, ..., X_n$ from a distribution $F(x)$ with unknown density function $f(x)$, the kernel density estimator (Rosenblatt, 1956) of $f(x)$ is given by

$$\hat{f}(x) = \frac{1}{nb}\sum_{i=1}^{n} k\left(\frac{x - X_i}{b}\right),$$

where b is the smoothing bandwidth and k is a symmetric function satisfying $\int k(x)dx = 1$. The kernel distribution function estimator (Nadaraya, 1964) of $F(x)$ is given by

$$\hat{F}(x) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x - X_i}{b}\right),$$

where K is the distribution function of the kernel k, $K(x) = \int_{-\infty}^{x} k(u)(du)$, and b is the bandwidth.

Consider the function $g(X_1, X_2, ..., X_m)$ that depends on $m \geq 1$ observations. Assume that g is a real value and is symmetric in its m arguments. Frees (1994) proposed an estimate for the density function $h(t)$ of random variable $g(X_1, X_2, ..., X_m)$ which is given by

$$\hat{h}(t) = \frac{1}{b\binom{n}{m}}\sum_{(n,m)} w\left(\frac{t - g(X_{i_1}, ..., X_{i_m}}{b}\right),$$

where b is the bandwidth, the sum extends over all $1 \leq i_1 < i_2 < ... < i_m \leq n$, and $w(.)$ is a kernel function. If $m = 1$ and $g(x) = x$, then the estimator $\hat{h}(t)$ reduces to the estimator $\hat{f}(x)$.

Abdel-Razzaq Mugdadi is an Associate Professor of Statistics in the Department of Mathematics and Statistics. He worked as an Assistant and Associate Professor with tenure at Southern Illinois University in Carbondale, IL from 2000 to 2010. He has published more than 15 research articles in publications including the *Journal of Statistical Planning and Inference*, *Journal of Nonparametric Statistics*, *Computational Statistics and Data Analysis*, *IEEE transactions on Reliability* and others. Email him at: aamugdadi@just.edu.jo.

The estimator $\hat{h}(t)$ has many applications in real life. For example, in spatial statistics g can be the inter point distance between pairs of objects and in insurance g can be the sum of m claims (Frees, 1994; Ahmad & Fan, 2001; Mugdadi & Ahmad, 2004).

Nadaraya (1964) and Mugdadi and Ghebregiorgis (2005) proposed a kernel distribution estimator of the distribution function of function of observations $H(t)$ as:

$$\hat{H}(t) = \frac{1}{\binom{n}{m}} \sum_{(n,m)} W\left(\frac{t - g(X_{i_1},\ldots,X_{i_m})}{b}\right),$$

where $W(x) = \int_{-\infty}^{x} w(t)dt$, b is the bandwidth and the sum extends over all $1 \le i_1 < i_2 < \ldots < i_m \le n$. Theoretical and simulation analyses show that choice of kernel is not crucial for distribution function estimation in the case of independent and identically (i.i.d) random variables; the most important choice is that of bandwidth. A typical way to select the bandwidth is to minimize one error measure, and the most commonly used is the mean integrated square error (MISE) and its asymptotic (AMISE), where

$$MISE(\hat{H}(t)) = E\int [\hat{H}(t) - H(t)]^2 dt.$$

Another criterion is the mean Hellinger distance (MHD), where

$$MHD(\hat{H}(t)) = E\int [\hat{H}^{\frac{1}{2}}(t) - H^{\frac{1}{2}}(t)]^2 dt.$$

Kanzawaa (1993) discussed the relationship between the asymptotic mean Hellinger distance (AMHD) and the AMISE for $\hat{f}(x)$, Ahmad and Mugdadi (2006) examined the relationship between asymptotic mean weighted Hellinger distance (AMWHD) and the AMISE for both $\hat{f}(x)$ and $\hat{F}(x)$, and Mugdadi (2004) studied the AMWHD for $\hat{h}(t)$. This investigation examines the relationship between the AMWHD and the AMISE for $\hat{H}(t)$ and proposes a data method to select the bandwidth for $\hat{H}(t)$ based on the AMWHD $\left(\hat{H}(t)\right)$.

The Asymptotic Mean Weighted Hellinger Distance

One error criterion used to evaluate the estimator is the mean weighted Hellinger distance (MWHD) and its asymptotic (AMWHD), where MWHD$\left(\hat{H}(t)\right)$ is defined by:

$$MWHD(\hat{H}(t)) = E\int [\hat{H}^{\frac{1}{2}}(t) - H^{\frac{1}{2}}(t)]^2 H(t)dt.$$

It can be argued that

$$MWHD(\hat{H}(t)) = \frac{1}{4} MISE(\hat{H}(t)),$$

assuming that $\hat{H}(t) \approx H(t)$ results in:

$$\begin{aligned} MISE(\hat{H}(t)) &= \int E[\hat{H}(t) - H(t)]^2 dt \\ &= \int E[\hat{H}^{\frac{1}{2}}(t) - H^{\frac{1}{2}}(t)]^2 [\hat{H}^{\frac{1}{2}}(t) + H^{\frac{1}{2}}(t)]^2 dt \\ &\approx 4E\int [\hat{H}^{\frac{1}{2}}(t) - H^{\frac{1}{2}}(t)]^2 H(t)dt \\ &= 4MWHD(\hat{H}(t)) \end{aligned}$$

Next, the $AMWHD(\hat{H}(t))$ is derived and compared with the $AMISE(\hat{H}(t))$. Mugdadi and Ghebregiorgis (2005) derived $AMISE(\hat{H}(t))$ as:

$$AMISE(\hat{H}(t)) = \frac{T(H)}{\binom{n}{m}} - \frac{b}{\binom{n}{m}}\rho(w) + \frac{b^4}{4}\mu_2(w)^2 R(H''),$$

where

$$\rho(w) = 2\int uW(u)w(u)du,$$

$$T(H) = \int H(t)[1-H(t)]dt$$

and

$$R(h) = \int h^2(t)dt.$$

Theorem

If the fourth derivative of $H(t)$ exists, then

$$AMISE(\hat{H}(t)) = \frac{T(H)}{4\binom{n}{m}} - \frac{b}{4\binom{n}{m}}\rho(w) + \frac{b^4}{16}\mu_2(w)^2 R(H'').$$

Proof

$$MWHD(\hat{H}(t)) = \int E\hat{H}(t)H(t)dt - 2\int E\hat{H}^{\frac{1}{2}}(t)H^{\frac{3}{2}}(t)dt + \int H^2(t)dt.$$

Using integration by parts and expanding $H(t-bu)$ in a 2nd order Taylor's series about t results in:

$$E(\hat{H}(t)) = E\left(\frac{1}{\binom{n}{m}}\sum_{(n,m)} W\left(\frac{t-g(X_1,X_2,...,X_m)}{b}\right)\right)$$

$$= \int w(u)\begin{pmatrix} H(t) - buH'(t) + \frac{b^2u^2H''(t)}{2} \\ -\frac{b^3u^3H^{(3)}(t)}{6} + \frac{b^4u^4H^{(4)}(t)}{24} \\ +o(b^4) \end{pmatrix}du$$

$$\approx H(t) + \frac{b^2H''(t)}{2}\mu_2(w) + \frac{b^4}{24}\mu_4(w)H^{(4)}(t)$$

therefore,

$$\int E[\hat{H}(t)H(t)]dt \approx \begin{bmatrix} \int H^2(t)dt + \frac{b^2\mu_2(w)}{2}\int H''(t)H(t)dt \\ +\frac{b^4\mu_4(w)}{24}\int H^{(4)}(t)H(t)dt \end{bmatrix}$$

If Z is a random variable with a standard normal distribution, then

$$\hat{H}(t) \approx E(\hat{H}(t)) + Z\sqrt{Var(\hat{H}(t))}.$$

Mugdadi and Ghebregiorgis (2005) derived $Var(\hat{H}(t))$, this is given by:

$$Var(\hat{H}(t)) = \begin{bmatrix} O\left(\frac{1}{n}\right) + O\left(\frac{b^2}{n}\right) + \binom{n}{m}^{-1} H(t)(1-H(t)) \\ +\binom{n}{m}^{-1} o(b) - \binom{n}{m}^{-1} bH'(t)\rho(w) \end{bmatrix}.$$

Thus,

$$\hat{H}(t) \approx H(t) + \frac{b^2H''(t)}{2}\mu_2(w) + \frac{b^4H^{(4)}(t)}{24}\mu_4(w) + Z\left[\begin{bmatrix} O\left(\frac{1}{n}\right) + O\left(\frac{b^2}{n}\right) + \binom{n}{m}^{-1} H(t)(1-H(t)) \\ +\binom{n}{m}^{-1} o(b) - \binom{n}{m}^{-1} bH'(t)\rho(w) \end{bmatrix}\right]^{\frac{1}{2}}$$

$$\approx H(t)\begin{bmatrix} 1 + \frac{b^2H''(t)\mu_2(w)}{2H(t)} + \frac{b^4\mu_4(w)H^{(4)}(t)}{24H(t)} \\ +Z\left(\frac{1-H(t)}{\binom{n}{m}H(t)} - \frac{bH'(t)\rho(w)}{\binom{n}{m}H^2(t)} + O\left(\frac{1}{n}\right) + O(b^2)\right)^{\frac{1}{2}} \end{bmatrix}$$

therefore,

$$\hat{H}^{\frac{1}{2}}(t) \approx \begin{bmatrix} 1+\frac{b^2 H''(t)\mu_2(w)}{4H(t)}+\frac{b^4\mu_4(w)H^{(4)}(t)}{48H(t)} \\ +\frac{Z}{2}\begin{bmatrix}\frac{1-H(t)}{\binom{n}{m}H(t)}-\frac{bH'(t)\rho(w)}{\binom{n}{m}H^2(t)} \\ +O\left(\frac{1}{n}\right)+O(b^2)\end{bmatrix}^{\frac{1}{2}} \\ -\frac{b^4}{32}\left(\frac{H''(t)}{H(t)}\right)^2(\mu_2(w))^2 \\ -\frac{Z^2}{8}\begin{bmatrix}\frac{1-H(t)}{\binom{n}{m}H(t)}-\frac{bH'(t)\rho(w)}{\binom{n}{m}H^2(t)} \\ +O\left(\frac{1}{n}\right)+O(b^2)\end{bmatrix}\end{bmatrix}$$

and

$$E(\hat{H}^{\frac{1}{2}}(t)) \approx H^{\frac{1}{2}}(t)\begin{bmatrix} 1+\frac{b^2 H''(t)\mu_2(w)}{4H(t)}+\frac{b^4\mu_4(w)H^{(4)}(t)}{48H(t)} \\ -\frac{b^4}{32}\left(\frac{H''(t)}{H(t)}\right)^2(\mu_2(w))^2 \\ -\frac{1}{8}\left(\frac{1-H(t)}{\binom{n}{m}H(t)}-\frac{bH'(t)\rho(w)}{\binom{n}{m}H^2(t)}\right) \\ +O\left(\frac{1}{n}\right)+O(b^2)\end{bmatrix}$$

thus

$$\int[E\hat{H}^{\frac{1}{2}}(t)H^{\frac{3}{2}}(t)]dt \approx \begin{bmatrix} \int H^2(t)dt+\frac{b^2}{4}\mu_2(w)\int H''(t)H(t)dt \\ +\frac{b^4}{48}\mu_4(w)\int H^{(4)}(t)H(t)dt \\ -\frac{b^4}{32}\mu_2^2(w)\int (H''(t))^2 dt \\ -\frac{1}{8\binom{n}{m}}\int H(t)(1-H(t))dt \\ +\frac{b}{8\binom{n}{m}}\rho(w)+O\left(\frac{1}{n}\right)+O(b^2)\end{bmatrix}$$

therefore,

$$AMWHD(\hat{H}(t)) \approx \begin{bmatrix} \frac{b^4}{16}\mu_2^2(w)R(H'') \\ +\frac{1}{4\binom{n}{m}}\int (H(t) \\ \times[1-H(t)dt-\frac{b}{4\binom{n}{m}}\rho(w)\end{bmatrix}$$

Under these conditions, the following corollaries can be proven.

Corollary 1

$$AMWHD(\hat{H}(t)) \approx \frac{AMISE(\hat{H}(t)}{4}.$$

Similar to Powell and Stocker (1996), the optimal bandwidth to minimize the $AMWHD(\hat{H}(t))$ is shown in corollary 2.

Corollary 2

$$b_{opt} = \left[\frac{\rho(w)}{\mu_2^2(w) R(h')} \right]^{\frac{1}{3}} \binom{n}{m}^{\frac{-1}{3}}$$

Bandwidth Selection

The choice of bandwidth is very important in the Kernal density estimator as well as in the Kernal distribution estimator. One of the simplest methods to select bandwidth is based on equation (2.8). Assume that the data is from a normal distribution with mean equal to zero and variance σ_1^2. If s^2 is the variance of the data $g(X_{i1},...,X_{im})$ for all $1 \le i_1 < ... < i_m \le n$, then $R(h') = \frac{1}{4\pi s^3}$, thus

$$b_{opt} = \left[\frac{\rho(w) 4\pi}{\pi_2^2(w)} \right]^{\frac{1}{3}} \binom{n}{m}^{\frac{-1}{3}} s \qquad (3.1)$$

This shows that b_{opt} depends only on the standard deviation and on the Kernel function.

A completely different data based method is proposed to select the bandwidth for a Kernal distribution estimator of the function of observations. The method is based on minimizing the $MWHD(\hat{H}(t))$. The $MWHD(\hat{H}(t)$ is defined as:

$$MWHD(\hat{H}(t)) = E\int [\hat{H}^{\frac{1}{2}}(t) - H^{\frac{1}{2}}(t)^2 H(t)dt$$

$$= \begin{bmatrix} E\int \hat{H}(t)H(t)dt + E\int H^2(t)dt \\ -2E\int \hat{H}^{\frac{1}{2}} H^{\frac{3}{2}}(t)dt \end{bmatrix}$$

$$= \begin{bmatrix} E\int \hat{H}(t)H(t)dt + E\int H^2(t)dt \\ -2\int (E\hat{H}^{\frac{1}{2}}(t)) H^{\frac{3}{2}}(t)dt \end{bmatrix}.$$

Minimizing $MWHD(\hat{H}(t))$ is therefore equivalent to minimizing $MWHD1(\hat{H}(t))$, where,

$$MWHD1(\hat{H}(t)) =$$
$$\int (E\hat{H}(t))H(t)dt - 2\int (E\hat{H}^{\frac{1}{2}}(t))H(t)^{\frac{2}{3}} dt$$

Thus, $MWHD1(\hat{H}(t))$ can be estimated as follows. Let $m_{(1)}$ be a fixed choice of m variables and let $I(A)$ be the indicator function. Also, define

$$H_{n,m_{(1)}}(t) = \frac{1}{\binom{n}{m} - 1} \sum_{(n,m), m \ne m_{(1)}} I(t - g(X_{i_1},...,X_{i_m}))$$

and

$$\hat{H}_{-m_{(1)}}(t) = \frac{1}{\binom{n}{m} - 1} \sum_{(n,m), m \ne m_{(1)}} W\left(\frac{t - g(X_{i_1},...,X_{i_m})}{b}\right)$$

which is the distribution estimated based on a sample with $m_{(1)}$ deleted. Thus, $MWHD1(\hat{H}(t))$ is estimated by

$$MWHD.EST =$$
$$\frac{1}{\binom{n}{m}} \sum_{1 \le i_1 < ... < i_m < n} \begin{bmatrix} \hat{H}_{-m_{(1)}}(g(X_{i_1},...,X_{i_{m(1)}}) \\ \times H_{n,m_{(1)}}(g(X_{i_1},...,X_{i_{m(1)}}) \end{bmatrix}$$

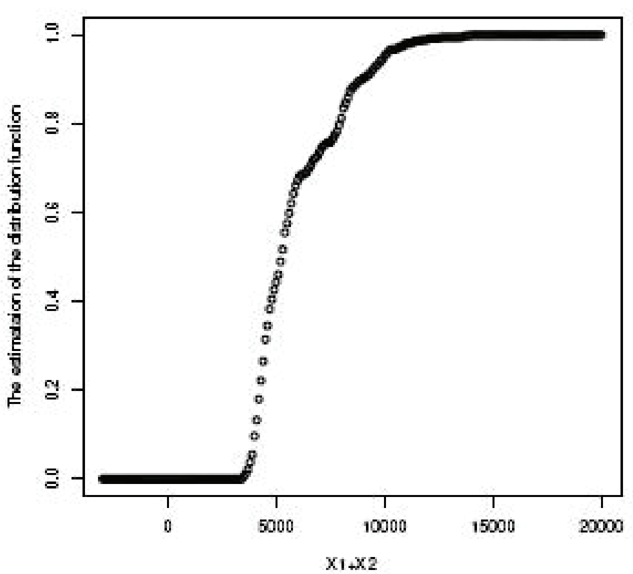
As noted, there are many applications for $\hat{H}(t)$. One example, introduced by Free (1993), regards an insurance claims problem. Table 1 shows total hospital charges (in dollars) in one Wisconsin (USA) hospital for females aged 30-49 in the year 1989.

Consider the case $m = 2$ and the function $g(X_1, X_2) = X_1 + X_2$. By minimizing $MWHD.EST$, the bandwidth is determined to be 0.437. Figure 1 shows the kernel distribution function for $g(X_1, X_2)$ using data in Table 1. It is clear that the kernel estimate is smooth and $\hat{H}(t)$ is 0 when $t \le 0$ because the sum of the charges should be positive; also, $\lim_{t\to\infty} \hat{H}(t) = 1$.

Table 1: Total 1989 Hospital Charges (USD) for Females Aged 30-49

2337	1765	1802
2179	2467	2011
2348	3609	2270
4765	2141	3425
3041	1850	3558
2088	3191	2315
2872	3020	1642
1924	2473	5878
2294	1898	2101
2182	7787	2242
2138	6169	5746

Figure 1: Kernel Distribution Function for $g(X_1, X_2)$



References

Ahmad, I. A., & Fan, Y. (2001). Optimal bandwidth for kernel density estimator of functions of observations. *Statistics and Probability Letters*, *51*(*3*), 245-251.

Ahmad, I. A., & Mugdadi, A. R. (2006). A weighted Hellinger distance as an error criterion for bandwidth selection for kernel estimation. *Journal of Nonparametric Statistics*, *18*, 215-226.

Bolanc, C., Guilln, M., & Nielsen, J. (2010). Transformation kernel estimation of insurance claim cost distribution. In *Mathematical and Statistical Methods for Actuarial Sciences and Finance*, 43-51. Milan, Italy: Springer Italia.

Frees, E. (1994). Estimating densities of functions of observations. *Journal of the American Statistical Association*, *89*, 517-525

Mugdadi, A. R., & Ahmad, I. A. (2004). A bandwidth selection for Kernal density estimation of functions of random variables. *Computational Statistics and Data Analysis*, *47*, 49-62.

Mugdadi, A. R., & Ghebregiorgis, G. (2005). The Kernal distribution estimator of functions of random variables. *Journal of Nonparametric Statistics*, *17*(*7*), 807-818.

Mugdadi, A. R. (2004). Bandwidth selection for the function of observations using Hellinger distance. *Journal of Applied Statistical Science*, *13*(*3*), 231-240.

Nadraya, E. A. (1964). Some new estimates for distribution function. *Theory and Probability Applications*, *9*, 497-500.

Powel, J. L., & Stoker, T. M. (1996). Optimal bandwidth choice for density-weighted averages. *Journal of Econometrics*, *75*, 291-316.

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

Sarda, P. (1993). Smoothing parameter selection for smooth distribution functions. *Journal of Statistical Planning and Inference*, *35*, 65-75.