ECONOMETRIC REVIEWS, 20(3), 353–367 (2001)

DENSITY ESTIMATION FOR CLUSTERED DATA

Robert V. Breunig

Department of Statistics and Econometrics, The Australian National University, Canberra ACT 0200, Australia

ABSTRACT

The commonly used survey technique of clustering introduces dependence into sample data. Such data is frequently used in economic analysis, though the dependence induced by the sample structure of the data is often ignored. In this paper, the effect of clustering on the non-parametric, kernel estimate of the density, $f(x)$, is examined. The window width commonly used for density estimation for the case of i.i.d. data is shown to no longer be optimal. A new optimal bandwidth using a higher-order kernel is proposed and is shown to give a smaller integrated mean squared error than two window widths which are widely used for the case of i.i.d. data. Several illustrations from simulation are provided.

Key Words: Bandwidth choice; Cluster sampling; Dependent data; Kernel density estimation.

JEL Classification: C14, C42.

1. INTRODUCTION

The technique of non-parametric density estimation using kernel methods for data which is independently and identically distributed (i.i.d.) is well-developed. It is not clear however, how these results are affected when the i.i.d. assumption is violated, although some work has examined density estimation under weakly dependent time series observations. (See, for example, Hall, Lahiri, and Polzehl (1995) and Herrmann, Gasser, and Kneip (1992). In this paper, I consider a

particular deviation from the i.i.d. case—specifically that of non-independence of the data created by clustering of the form frequently found in survey data.

Much of the data used for economic analysis is gathered using survey methods leading to sample data which may violate the i.i.d. assumption. Serial correlation of data is a well-known problem in time-series data, but it is also present in much cross-sectional data, where it is usually ignored by analysts. Most cross-sectional data for economic analysis is gathered through some type of complex survey (see Ullah and Breunig (1998)). Data is usually selected from populations which are stratified and clustered using well-known survey sampling techniques. Clustering, frequently employed to reduce the cost of data collection, generally leads to positive correlation between data points in the same cluster. Much applied cross-sectional, econometric analysis ignores the correlation which is present in such data.

Below, we relax the assumption of independently distributed data and consider the problem of kernel density estimation for clustered data. The choice of window width for kernel density estimation with i.i.d. data has been considered by many authors. Silverman (1986) provides an excellent review. In this paper, we obtain the approximate integrated mean squared error (IMSE) for the kernel density estimation under cluster sampling. An optimal window width is proposed which minimizes the approximate IMSE. This result suggests that the usual optimal window width for i.i.d. data does not hold in the case of clustered data. The combination of a fourth-order kernel and a window width which depends on the degree of correlation in the data turns out to perform well in application and performs better, in an integrated mean-squared error sense, than commonly-used alternatives.

2. DENSITY ESTIMATION IN CLUSTERED DATA

Consider the case of clustered data, where a sample of n units has been drawn from some population. It is assumed that the data is drawn in two stages; a sample of k "clusters" is randomly chosen at the first stage; and in the second stage a sample of n_c elements is chosen from each cluster, $c = 1, \ldots, k$.[1] The total sample size is thus $n = \sum_c n_c$. For simplicity, I assume that in the second stage, $n_c = n/k = m$ elements are chosen from each cluster. Also known as "balanced"

[1]Clustering is frequently found in economic data gathered from surveys. One common example is the income and expenditure survey, where first a sample of villages is chosen and then, within each village, households are randomly selected. Households within the same village (or cluster) can be assumed to face similar conditions—for example we expect heating fuel costs to be correlated for households in the same area. In this paper, I assume that the data has already been gathered and that the analyst has information about the structure of the data. Kish (1965) and Thompson (1992) provide details of how clustered surveys are conducted.

clusters, this assumption will be relaxed below. Assume that the data is characterized by

$$\begin{aligned} &\mathrm{Var}(x_{ci}) = \sigma^2 && \text{for all } i = 1, \ldots, n \text{ and } c = 1, \ldots, k. && \text{(A1)} \\ &\mathrm{Cov}(x_{ci}, x_{c'j}) = \rho\sigma^2 && \text{for all } i \neq j \text{ and } c = c'. && \text{(A2)} \\ &\mathrm{Cov}(x_{ci}, x_{c'j}) = 0 && \text{for all } c \neq c'. \end{aligned}$$

We further assume that the data has a common mean, μ, and that the data are identically, but not independently distributed. The form of dependence is characterized by (A2). Specifically, elements within clusters are correlated, while elements in different clusters are uncorrelated.[2]

The problem of non-parametrically estimating the density for the case of i.i.d. data is well-studied. (See Silverman (1986), Härdle (1990), Pagan and Ullah (1997)). Choosing the optimal window width, h^*, by minimizing the approximate integrated mean squared error (AMISE) of the density estimator provides $h^* \alpha\, n^{-1/5}$ when a second-order kernel is used. (Generally, $h^* = cn^{-1/(2P+1)}$ for a Pth order kernel.) Furthermore, if the underlying true density of the data is normal with variance $\sigma^2$and the kernel is Gaussian, the optimal window width is

$$h^* = 1.06\sigma n^{-1/5}. \tag{1}$$

(See Silverman (1986).)

I will examine how this result changes if the data obey (A1) and (A2) above. Specifically, I will use the same method of minimizing the approximate integrated mean squared error with respect to h, and solving for the optimal window width.

The non-parametric, kernel estimate of the density at any point x, estimated from a sample of size n, is

$$\widehat{f(x)} = \frac{1}{nh}\sum_{j=1}^{n} K\left(\frac{x_j - x}{h}\right) \tag{2}$$

where h is the window width which is assumed to satisfy

$$\begin{aligned} &\text{(i) } h \to 0 \\ &\text{(ii) } nh \to \infty \quad \text{as } n \to \infty \end{aligned} \tag{A3}$$

and the kernel $K(\cdot)$ is a symmetric function which satisfies

$$\begin{aligned} &\text{(i)} && \textstyle\int K(\psi)d\psi = 1 \\ &\text{(ii)} && \textstyle\int \psi K(\psi)d\psi = 0 \\ &\text{(iii)} && \textstyle\int \psi^2 K(\psi)d\psi = \mu_2 < \infty \\ &\text{(iv)} && \textstyle\int \psi^4 K(\psi)d\psi = \mu_4 < \infty. \end{aligned} \tag{A4}$$

[2]The intra-cluster correlation coefficient, ρ, can be surprisingly large in cross-sectional data. Deaton (1997) provides examples using World Bank data where intra-cluster correlation coefficients range from 0.2 to 0.5.

For the case of clustered data, we can re-write the Kernel density estimate as

$$\widehat{f(x)} = \frac{1}{nh}\sum_{c=1}^{k}\sum_{i=1}^{n_c} K\left(\frac{x_{ci}-x}{h}\right) \tag{3}$$

where n_c is the number of observations in the cth cluster. For the derivation which follows, we have assumed $n_c = n/k = m$. It is straightforward to show that the bias of $\widehat{f(x)}$ upto $O(h^2)$ is

$$\text{bias}\,\widehat{f(x)} = \frac{h^2}{2} f''(x)\mu_2 \tag{4}$$

(see Silverman (1986), p. 39).

To find the variance of $\widehat{f(x)}$ under assumptions (A1) and (A2), we first re-write $\widehat{f(x)}$ as

$$\widehat{f(x)} = \frac{1}{n}\sum_{c=1}^{k}\sum_{i=1}^{n_c} W_{ci} \tag{5}$$

where

$$W_{ci} = \frac{1}{h} K\left(\frac{x_{ci}-x}{h}\right). \tag{6}$$

Then

$$\text{Var}(\widehat{f(x)}) = \frac{1}{n^2}\sum_{c=1}^{k}\sum_{i=1}^{n_c}\text{Var}(W_{ci}) + \frac{1}{n^2}\sum_{c=1}^{k}\sum_{c'=1}^{k}\sum_{j=1}^{n_{c'}}\sum_{\substack{i=1 \\ i\neq j \text{ for } c=c'}}^{n_c}\text{Cov}(W_{ci}, W_{c'j}). \tag{7}$$

$\text{Var}(W_{ci}) = EW_{ci}^2 - (EW_{ci})^2$ and since the data are identically distributed, $\text{Var}(W_{ci}) = \text{Var}(W_{11})$. Using this information and the assumption that elements across clusters are uncorrelated

$$\text{Var}(\hat{f}(x)) = \frac{1}{n}EW_{11}^2 - \frac{1}{k}(EW_{11})^2 + \left(\frac{1}{k}-\frac{1}{n}\right)E(W_{11}\cdot W_{12}). \tag{8}$$

First, consider term one

$$\begin{aligned}\frac{1}{n}E(W_{11})^2 &= \frac{1}{n}\int_{x_{11}}\left(\frac{1}{h}K\left(\frac{x_{11}-x}{h}\right)\right)^2 f(x_{11})dx_{11} \\ &= \frac{1}{nh}\int_{\psi_{11}} K^2(\psi_{11})\, f(h\psi_{11}+x)d\psi_{11}.\end{aligned}$$

Expanding $f(x_{11})$ around the point $x_{11} = x$ by the method of Taylor's series,

$$\frac{1}{n}E(W_{11})^2 = \frac{1}{nh}\int_{\psi_{11}} K^2(\psi_{11})\left[f(x) + f'(x)h\psi_{11} + f''(x)(h\psi_{11})^2 + \cdots\right]d\psi_{11}.$$

Keeping terms up to $O(1/nh)$ gives an approximation for the first term of $\mathrm{Var}(\widehat{f(x)})$,

$$\frac{1}{n}E(W_{11})^2 = \frac{1}{nh}\int_{\psi_{11}} K^2(\psi_{11})\, f(x) d\psi_{11}. \tag{9}$$

Now, consider term two, replacing $f(x_{11})$ with a Taylor's series expansion around the point $x_{11} = x$.

$$\begin{aligned}\frac{1}{k}(EW_{11})^2 &= \frac{1}{k}\left\{\int_{\psi_{11}} K(\psi_{11})\left[f(x)+f'(x)h\psi_{11}+\frac{f''(x)}{2}(h\psi_{11})^2\right.\right.\\ &\qquad \left.\left.+\frac{f'''(x)}{6}(h\psi_{11})^3+\frac{f''''(x)}{24}(h\psi_{11})^4+\cdots\right]d\psi_{11}\right\}^2\\ &= \frac{1}{k}\left\{\int_{\psi_{11}}\left[f(x)+hf'(x)\psi_{11}\right]K(\psi_{11})d\psi_{11}+\frac{h^2f''(x)}{2}\int_{\psi_{11}}\psi_{11}K(\psi_{11})d\psi_{11}\right.\\ &\qquad \left.+\frac{h^3}{6}f'''(x)\int_{\psi_{11}}\psi_{11}K(\psi_{11})d\psi_{11}+\frac{h^4}{24}f''''(x)\int_{\psi_{11}}\psi_{11}K(\psi_{11})d\psi_{11}\right\}^2\\ &= \frac{1}{k}\left\{f(x)+\frac{h^2f''(x)\mu_2}{2}+\frac{h^3f'''(x)\mu_3}{6}+\frac{h^4f''''(x)\mu_4}{24}\right\}^2. \end{aligned}\tag{10}$$

Which gives, up to order $O(h^4)$,

$$=\frac{1}{k}\left\{f(x)^2+\frac{h^4(f''(x))^2\mu_2^2}{4}+f(x)h^2f''(x)\mu_2+\frac{f(x)h^3f'''(x)\mu_3}{3}+\frac{f(x)h^4f''''(x)\mu_4}{12}\right\}. \tag{11}$$

Term three, $(1/k) - (1/n)\, E(W_{11}W_{12})$, may be written

$$\left(\frac{1}{k}-\frac{1}{n}\right)\int_{x_{11}}\int_{x_{12}}\frac{1}{h^2}K\left(\frac{x_{11}-x}{h}\right)K\left(\frac{x_{12}-x}{h}\right)f(x_{11},x_{12})dx_{11}dx_{12}. \tag{12}$$

We first transform the density $f(x_{11}, x_{12})$ following Rao (1973)

$$f(\psi_{11},\psi_{12}) = f(x_{11},x_{12})\left|\begin{array}{cc}\dfrac{\partial x_{11}}{\partial(\psi_{11})} & \dfrac{\partial x_{12}}{\partial(\psi_{11})}\\ \dfrac{\partial x_{11}}{\partial(\psi_{12})} & \dfrac{\partial x_{12}}{\partial(\psi_{12})}\end{array}\right|. \tag{13}$$

Evaluation of the determinant in equation (13) under assumptions (A1) and (A2) gives $h^2(1-\rho^2)$, thus after replacing x_{11} with $h\psi_{11}+x$ and x_{12} with $h\psi_{12}+x$,

term three becomes

$$\left(\frac{h^2(1-\rho^2)}{h^2}\right)\left(\frac{1}{k}-\frac{1}{n}\right)\int_{\psi_{11}}\int_{\psi_{12}} K(\psi_{11})K(\psi_{12})f(h\psi_{11}+x, h\psi_{12}+x)d\psi_{11}d\psi_{12} \tag{14}$$

Using a bivariate Taylor series expansion of $f(x_{11}, x_{12})$ around the point $x_{11}=x$, $x_{12}=x$, this term becomes

$$\begin{aligned}(1-\rho^2)\left(\frac{1}{k}-\frac{1}{n}\right)\int_{\psi_{11}}\int_{\psi_{12}} K(\psi_{11})K(\psi_{12})\big[f(x,x)+f_1(x,x)h\psi_{11}+f_2(x,x)h\psi_{12}\\ +\tfrac{1}{2}f_{11}(x,x)(h\psi_{11})^2+\tfrac{1}{2}f_{22}(x,x)(h\psi_{12})^2+f_{12}(x,x)(h\psi_{11})(h\psi_{12})+\cdots\big]d\psi_{11}d\psi_{12}\end{aligned} \tag{15}$$

which after simplification, and keeping only terms upto $O(\max\{(1/nh), h^4\})$

$$\begin{aligned}&= \left(\frac{1-\rho^2}{k}\right)\Bigg[f(x,x)+h^2f_{11}(x,x)\mu_2+\frac{h^3}{3}f_{111}(x,x)\mu_3\\ &\quad+\frac{h^4}{12}f_{1111}(x,x)\mu_4+\frac{h^4}{4}f_{1122}(x,x)\mu_2^2\Bigg]\end{aligned} \tag{16}$$

where $f_1(x_1,x_2)=\partial f(x_1,x_2)/\partial x_1, f_{11}(x_1,x_2)=\partial^2 f(x_1,x_2)/(\partial x_1)^2$, etc. and $f_{1122}(x_1,x_2)=\partial^4 f(x_1,x_2)/(\partial x_1)^2(\partial x_2)^2$.

Theorem 1

If the data is characterized by (A1) and (A2), and $\widehat{f(x)}$ is estimated using a kernel which satisfies (A4), then the $\mathrm{Var}(\widehat{f(x)})$ upto $O(\max\{(1/nh), h^4\})$ is

$$\begin{aligned}\mathrm{Var}(\widehat{f(x)})=&\frac{1}{nh}\int_{\psi_{11}}K^2(\psi_{11})f(x)d\psi_{11}+\frac{(1-\rho^2)f(x,x)-f(x)^2}{k}\\ &-\frac{1}{k}\left\{\frac{h^4(f''(x))^2\mu_2^2}{4}+f(x)h^2f''(x)\mu_2+\frac{f(x)h^3f'''(x)\mu_3}{3}+\frac{f(x)h^4f''''(x)\mu_4}{12}\right\}\\ &+\left(\frac{1-\rho^2}{k}\right)\left[\frac{h^4}{4}f_{1122}(x,x)\mu_2^2+h^2f_{11}(x,x)\mu_2+\frac{h^3}{3}f_{111}(x,x)\mu_3+\frac{h^4}{12}\ f_{1111}(x,x)\mu_4\right]\end{aligned} \tag{17}$$

Proof

Combine equations (9), (11), and (16).

The variance in (17) goes to zero if $k\to\infty$ as $n\to\infty$. This is a reasonable assumption, as it characterizes the way that sampling is done for most economic surveys. Average cluster sizes tend to be fairly small (10–12 elements per cluster) while increases in sample size are normally achieved by increasing k, the number of clusters sampled. If m converges to unity as $k\to\infty$ and $n\to\infty$, then the sample should simply be treated as an i.i.d. sample, since the intra-cluster correlation is nonsensical for clusters of size one.

The following corollary shows that the variance expression in (17) will collapse to the variance of the kernel estimate of $f(x)$ for the i.i.d. case when correlation is not present in the data.

Corollary 1

If the data is characterized by (A1) and $\rho = 0$ in (A2), then the $\mathrm{Var}(\widehat{f(x)})$ upto $O(1/nh)$ is

$$\mathrm{Var}(\widehat{f(x)}) = \frac{1}{nh}\int_{\psi_{11}} K^2(\psi_{11}) f(x) d\psi_{11}. \tag{18}$$

Proof

If the data is independent, then $f(x,x) = f(x)f(x), f_{1122}(x,x) = f''(x)f''(x)$, $f_{11}(x,x) = f''(x)f(x), f_{111}(x,x) = f'''(x)f(x)$, and $f_{1111}(x,x) = f''''(x)f(x)$ and the second, third and fourth terms in equation (17) will cancel out.

The window width which minimizes the AIMSE of $\widehat{f(x)}$ in the i.i.d. case (1) will no longer be optimal in the case of correlated data. If we use the method of minimizing the approximate integrated mean squared error using equations (17) and (4), the resulting solution will be a complex polynomial in h, which will include terms containing μ_2, μ_3, and μ_4. (Hall et al. (1991) consider a similar problem where the AIMSE is also a 7th-degree polynomial in h. They provide an optimal h which is asymptotically equivalent to the implicit minimizer of the 7th-degree polynomial.)

The solution pursued here is to choose a higher-order kernel. Higher-order kernels have been used to reduce bias in kernel density estimation (see Pagan and Ullah (1997), chapter 2, section 4.3). By choosing a fourth-order kernel, terms involving μ_2 and μ_3 will be zero. Minimizing the approximate integrated mean squared error will then yield a simple solution for the optimal h upto the order of approximation considered. The proposed optimal h below is exactly that window width which minimizes the integrated mean squared error (to the order considered) and not simply an asymptotic equivalent.

Replace assumption (A4) with

$$\begin{aligned}
&\text{(i)} \quad \int K(\psi)d\psi = 1\\
&\text{(ii)} \quad \int \psi K(\psi)d\psi = 0\\
&\text{(iii)} \quad \int \psi^2 K(\psi)d\psi = 0\\
&\text{(iv)} \quad \int \psi^3 K(\psi)d\psi = 0\\
&\text{(v)} \quad 0 < \int \psi^4 K(\psi)d\psi = \mu_4 < \infty
\end{aligned} \tag{A4}'$$

Here, the fourth moment of the kernel is required to be positive in addition to the usual assumption that μ_4 be finite. Since μ_4 appears below in the expression for the optimal window width (22) raised to a fractional power, it must be positive in order to give a reasonable (i.e. a positive real number) window width.

We can construct a fourth-order kernel which meets assumption (A4)′ and is based upon the standard normal kernel ($K^*(\psi)$) by assigning a value for μ_4. Setting $\mu_4 = 1$ the resulting kernel is

$$(2 - \tfrac{3}{2}\psi^2 + \tfrac{1}{6}\psi^4)K^*(\psi). \tag{19}$$

Figure 1 presents a graph of this kernel and the standard normal kernel.

Now, the

$$\mathrm{Var}(\widehat{f(x)}) = \frac{1}{nh}\int_{\psi_{11}} K^2(\psi_{11})f(x)d\psi_{11} - \frac{1}{k}\left\{(f(x))^2 + \frac{h^4}{12}f''''(x)f(x)\mu_4\right\} + \left(\frac{1-\rho^2}{k}\right)\left[f(x,x) + \frac{h^4}{12}f_{1111}(x,x)\mu_4\right]. \tag{20}$$

The integrated mean-squared error (IMSE) is

$$\int_x \left\{(\mathrm{bias}(\widehat{f(x)}))^2 + \mathrm{Var}(\widehat{f(x)})\right\}dx. \tag{21}$$

Since we are using a higher-order kernel, this bias will be of $O(h^4)$ instead of having the form in (4). $(\mathrm{bias}\,(\widehat{f(x)}))^2$ will thus be $O(h^8)$, and the approximate integrated mean squared error upto $O(1/nh)$ will be equivalent to the integrated variance.

Corollary 2

Given (A1), (A2), and (A4)′, the optimal window width, h_{opt} will be

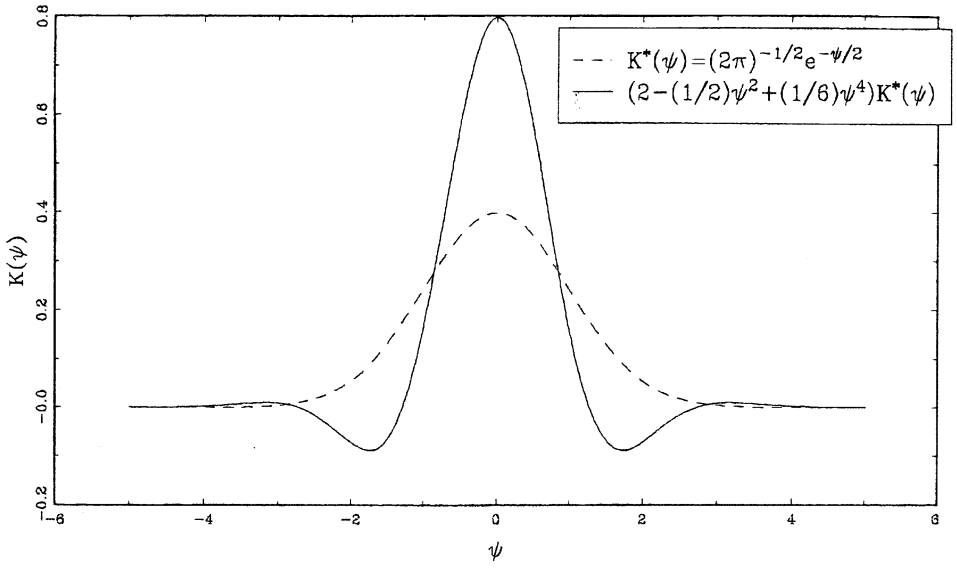

Figure 1. Fourth-order kernel based upon standard normal kernel.

$$h_{opt}=\left[\int_{\psi} K^2(\psi)d\psi\right]^{1/5}\left[\frac{n\mu_4}{3k}\right]^{-1/5}\left[\int_x((1-\rho^2)f_{1111}(x,x)-f''''(x)f(x))dx\right]^{-1/5}. \tag{22}$$

Proof

Use (20) and (21) and minimize the expression for the AIMSE with respect to h and rearrange to solve for h_{opt}.

As in the i.i.d. case, the optimal window width is proportional to $n^{-1/5}$ and will depend upon both the kernel and the true, underlying density of the data. Analogous to the i.i.d. case, we consider the case where (x_1, x_2) is distributed as a bivariate normal and we choose the fourth-order kernel of (19) allowing us to give an exact value to h_{opt}. For this special case, we have

$$h=\left[\frac{467}{384\sqrt{\pi}}\right]^{1/5}\left[\frac{n}{3k}\right]^{-1/5}\left(\sigma^{-5}\int_z\{(1-\rho^2)\phi^*_{1111}(z_1,z_2;\rho)-\phi''''(z)\cdot\phi(z)\}dz\right)^{-1/5} \tag{23}$$

where ϕ is the standard normal distribution and ϕ^* is the standard normal bivariate with correlation ρ (see Morrison (1976), p. 86). Rearranging, we can write the optimal window width as

$$h = \kappa\sigma n^{-1/5} \tag{24}$$

where

$$\kappa = \left[\frac{467}{128\sqrt{\pi}}\right]^{1/5}[k]^{1/5}\left((\Phi(\rho) - \frac{3}{8\sqrt{\pi}}\right)^{-1/5} \tag{25}$$

and $\Phi(\rho) = \int_z (1 - \rho^2)\phi^*_{1111}(z_1, z_2; \rho)dz$. Unlike the case of i.i.d. data where x is normally distributed, κ will no longer be constant, but will depend upon both the number of clusters, k, and the intra-cluster correlation coefficient, ρ.

As the cluster size increases, the window width will increase for a given size sample, n. If we re-write (22) as

$$h = \tilde{\kappa}\sigma\left(\frac{n}{k}\right)^{-1/5} \tag{26}$$

where

$$\tilde{\kappa} = \left[\frac{467}{128\sqrt{\pi}}\right]^{1/5}\left(\Phi(\rho) - \frac{3}{8\sqrt{\pi}}\right)^{-1/5} \tag{27}$$

we can see that an increase in the number of clusters (or a decrease in the average cluster size) acts as a decrease in the "effective" sample size, (n/k).

It can be shown that $\Phi(\rho)$ is increasing in $\rho$and therefore κ will be decreasing in ρ. Intuitively, as the intra-cluster correlation coefficient increases, data will be clustered more tightly together, and thus a finer window width will be optimal. For the case where $f(x, x)$ is bivariate normal, Table 1 gives values of $\tilde{\kappa}$ for a range of ρ values.

Table 1. Entries in Table Represent $\tilde{\kappa}$ in Equation (27) for Different ρ

ρ	$\tilde{\kappa}$
0.05	2.3561866
0.10	2.0213222
0.15	1.835119
0.20	1.7040834
0.25	1.6011561
0.30	1.5148296
0.35	1.4391234
0.40	1.3704895
0.45	1.3065959
0.50	1.2457628
0.55	1.1866615
0.60	1.1281282
0.65	1.0690206
0.70	1.0080711
0.75	0.94368371
0.80	0.87356777
0.85	0.79392504
0.90	0.69714355
0.95	0.56208067

For $f(x, x)$ bivariate normal with $\sigma = 1$ and $n = 1000$, Figure 2 shows the effect on the optimal window width of changing ρ and k simultaneously.

In practice, ρ can be replaced by a consistent estimator, $\hat{\rho}$:

$$\hat{\rho} = \frac{\sum_{c=1}^{k} \sum_{i=1}^{n_c} \sum_{j \neq i}^{n_c} (x_{ci} - \bar{x})(x_{cj} - \bar{x})}{\hat{\sigma}^2 \sum_{c=1}^{k} n_c(n_c - 1)}$$

and the optimal window width can then be calculated from (24) and (25). A computer program is available from the author for exact calculation of $\Phi(\rho)$. Alternately, the appropriate value from Table 1 could be used to determine $\tilde{\kappa}$ and plugged into (26).

For the case of unbalanced clusters, we can replace $[k]^{1/5}$ in (24) with $[\tilde{k}]^{-1/5}$ where

$$\tilde{k} = \sum \left(\frac{n_c}{n}\right)^2.$$

It is easy to ascertain that $\tilde{k} = 1/k$ when clusters are balanced.

Fan and Marron (1992) posit that higher-order kernels have not seen much use in application because of the unclear meaning of negative weights which higher-order kernels give to some data points and because the gains from using higher-order kernels are negligible for most sample sizes. In the case of clustering, however, higher-order kernels provide a simple way to solve for the optimal

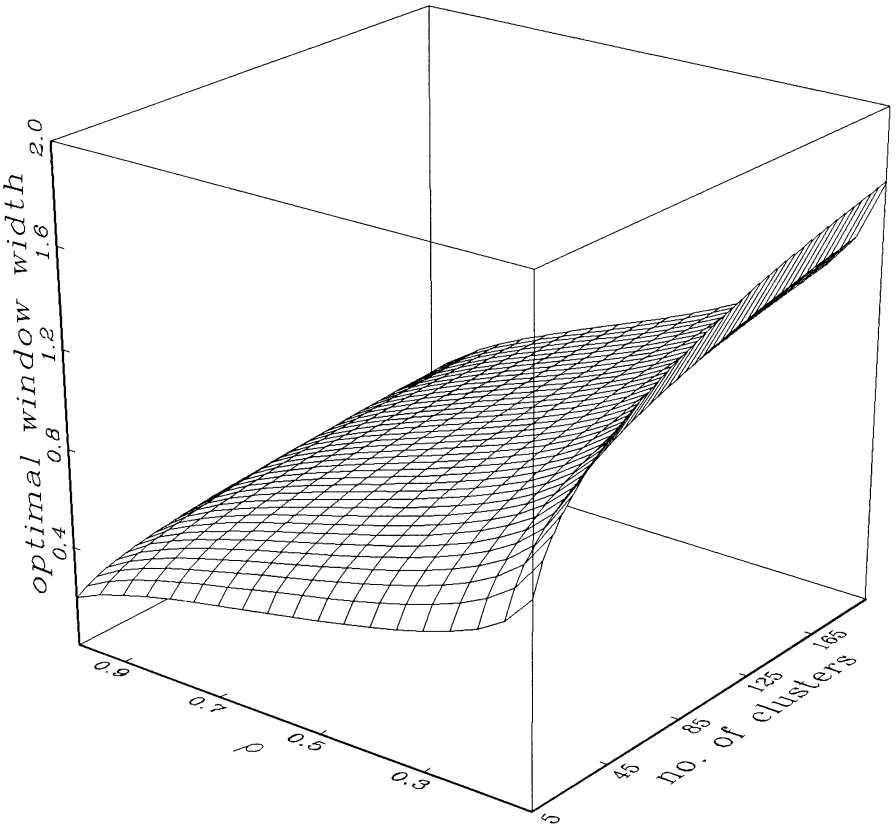

Figure 2. Optimal window width for different cluster sizes and intra-cluster correlation coefficients.

window width and allow for kernel density estimation which is easy to implement and is similar to the i.i.d. case.

3. NUMERICAL PROPERTIES OF h_{opt}

What are the gains in efficiency from using the optimal window width, h_{opt}, of (24)? We compare the mean squared error of $\hat{f}(x)$ using h_{opt} with two alternatives, $h^* = 1.06\sigma n^{-1/5}$ and $h^{**} = cn^{-1/9}$. h^* is the optimal window width when the underlying density is normal and the kernel is Gaussian. h^{**} is the optimal window width in the i.i.d. case when a 4th-order kernel is used. Set $c = 1.44$, which is the optimal proportionality constant given the kernel of (19) and a true, underlying distribution that is normal, i.i.d.[3] (Of course, this ignores the dependence in the data.)

[3]In general in the i.i.d. case, for an r-th order kernel, $h^{**} = (\lambda_2(r!)^2/2r\lambda_{1r})^{(1/2r+1)}n^{-(1/2r+1)}$ where $\lambda_{1r} = \mu_r^2 \int (f^r(x))_x^2 dx$ and $\lambda_2 = \int_\psi K^2(\psi)d\psi$. $\lambda_{1r} = 105/32\sqrt{\pi}$ for this density.

The AIMSE is calculated upto $O(1/nh)$ using the kernel in (19), the standardized bivariate normal distribution, and the three window widths, h_{opt}, h^*, and h^{**}. The AIMSE will be

$$\text{AIMSE} = \frac{1}{nh}\int_{\psi} K^2(\psi)d\psi + \frac{1}{k}\sigma^{-1}\int_z \{(1-\rho^2)\phi^*(z,z) - (\phi(z))^2\}dz + \frac{h^4}{12k}\sigma^{-5}\int_z \{(1-\rho^2)\phi^*_{1111}(z,z) - \phi''''(z)\phi(z)\}dz. \quad (28)$$

Specifying a reference distribution allows exact calculation of the approximate integrated mean squared error. Results are given in Table 2.

h_{opt} outperforms both h^* and h^{**} in the integrated mean squared error sense-not surprising given that it is chosen to minimize the AIMSE. The last two columns of Table 2 show that the gains in approximate IMSE calculated from (28) are quite substantial when compared to h^*—generally on the order of 50%. The gains from using h_{opt} compared to h^{**} are somewhat smaller, but using h_{opt} does give a lower mean squared error. Of course, the gains in integrated mean squared error depend upon the values of n and k chosen here. Gains in IMSE may be greater for other values of n and k.

A detailed simulation study was conducted by the author to compare the various window width options considered above. To conduct the simulation, we first re-write the model as follows

$$x_{ci} = \mu + u_c + \varepsilon_{ci} \quad (29)$$

Table 2. Difference in IMSE from Using h_{opt}, h^*, and h^{**}

ρ	h_{opt}	h^*	h^{**}	IMSE (h_{opt})	IMSE (h^*)	IMSE (h^{**})	IMSE (h_{opt})/ IMSE (h^*)	IMSE (h_{opt})/ IMSE (h^{**})
0.05	1.4867	0.2663	0.6686	0.00064	0.00264	0.00110	0.24326	0.58608
0.10	1.2754	0.2663	0.6686	0.00080	0.00270	0.00116	0.29457	0.68606
0.15	1.1579	0.2663	0.6686	0.00091	0.00275	0.00121	0.33148	0.75095
0.20	1.0752	0.2663	0.6686	0.00100	0.00278	0.00126	0.36076	0.79915
0.25	1.0103	0.2663	0.6686	0.00108	0.00281	0.00129	0.38490	0.83750
0.30	0.9558	0.2663	0.6686	0.00115	0.00283	0.00132	0.40513	0.86942
0.35	0.9080	0.2663	0.6686	0.00119	0.00283	0.00133	0.42222	0.89689
0.40	0.8647	0.2663	0.6686	0.00123	0.00282	0.00134	0.43666	0.92113
0.45	0.8244	0.2663	0.6686	0.00125	0.00279	0.00133	0.44879	0.94286
0.50	0.7860	0.2663	0.6686	0.00126	0.00275	0.00131	0.45885	0.96237
0.55	0.7487	0.2663	0.6686	0.00126	0.00269	0.00128	0.46706	0.97942
0.60	0.7118	0.2663	0.6686	0.00124	0.00262	0.00125	0.47359	0.99286
0.65	0.6745	0.2663	0.6686	0.00120	0.00252	0.00120	0.47868	0.99984
0.70	0.6360	0.2663	0.6686	0.00115	0.00239	0.00116	0.48272	0.99392
0.75	0.5954	0.2663	0.6686	0.00109	0.00224	0.00113	0.48654	0.96132
0.80	0.5512	0.2663	0.6686	0.00101	0.00204	0.00115	0.49219	0.87486
0.85	0.5009	0.2663	0.6686	0.00091	0.00181	0.00131	0.50562	0.69489

$n = 1000$, $k = 100$, average cluster size $= 10$, $\sigma = 1$.

where $u_c \sim D(0, \sigma_c^2)$ is an effect common to all elements in cluster c and $\varepsilon_{ci} \sim D(0, \sigma_\varepsilon^2)$ is an idiosyncratic error term with $\text{cov}(u_c, \varepsilon_{ci}) = 0$ for all $i = 1, \ldots, n$ and $c = 1, \ldots, k$ and $\text{cov}(u_c, u_{c'}) = 0$ for all $c \neq c'$. The element variance will then be $\sigma_c^2 + \sigma_\varepsilon^2$, and the intra-cluster correlation coefficient ρ will equal $\sigma_c^2/(\sigma_c^2 + \sigma_\varepsilon^2)$. This provides the structure of (A1) and (A2). Normal distributions were chosen for both the cluster-specific and the idiosyncratic errors. σ_c^2 and σ_ε^2 were fixed so that the total element variance equals one. Thus $\rho = \sigma_c^2$ and $\sigma_\varepsilon^2 = 1 - \sigma_c^2$, allowing different degrees of correlation to be generated in the data. For the simulation, the cluster size, k, was set at 500 and the total sample size, n, at 10,000. (Thus for each simulation, 10,000 numbers were drawn from a $N(0, \sigma_\varepsilon^2)$ and 500 from a $N(0, \sigma_c^2)$. All elements in a cluster share the same draw of the cluster-specific error term.) Average cluster size is 20 and since clusters were chosen to be balanced, the clusters are all the same size.

Figures 3 through 5 are typical realizations of this simulation exercise for $\rho = 0.2$, $\rho = 0.4$, and $\rho = 0.6$. Here we present the comparison between the density estimation using h^* and h_{opt}. The density estimates using h^* tend to under-smooth the data as can be seen by the spurious variation in the density estimates. The standard normal distribution (the marginal distribution of the true density) is shown for reference.

For cluster data with values of ρ greater than 0.15, the suggested kernel (19) and h_{opt} (24) perform very well in simulation. For small values of ρ this combination tends to over-smooth the data, and it is probably best to use the second-order Gaussian kernel and the usual window width. This needs further investigation.

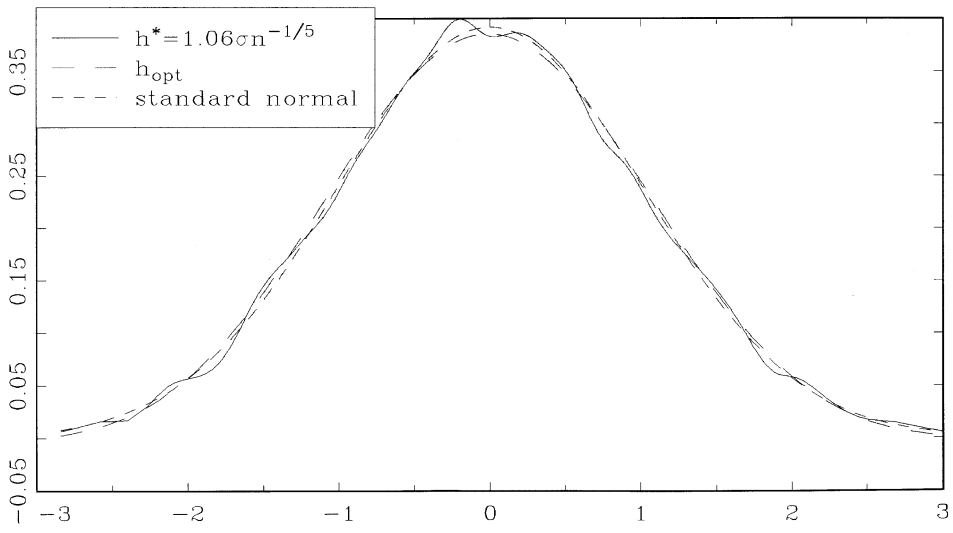

Figure 3. Density estimation for clustered data ($\rho = .2$) h^* vs. h_{opt}.

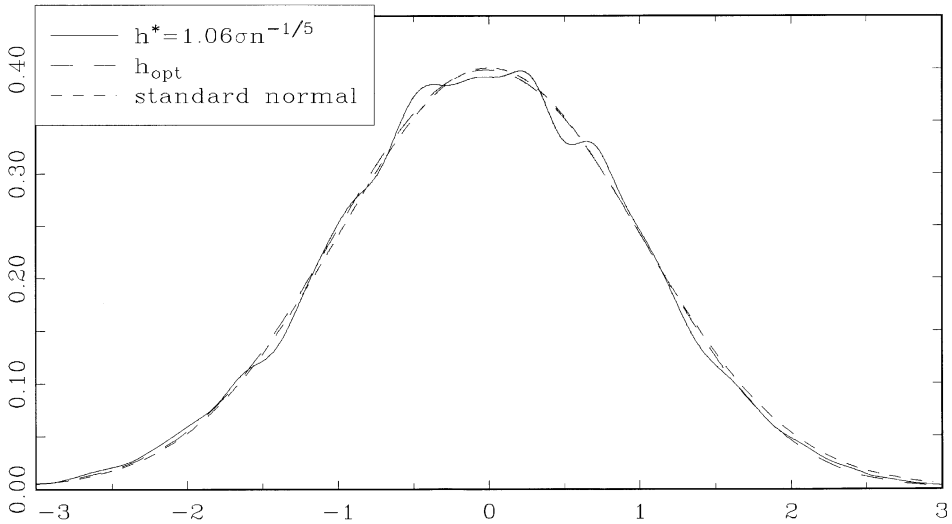

Figure 4. Density estimation for clustered data ($\rho = .4$) h^* vs. h_{opt}.

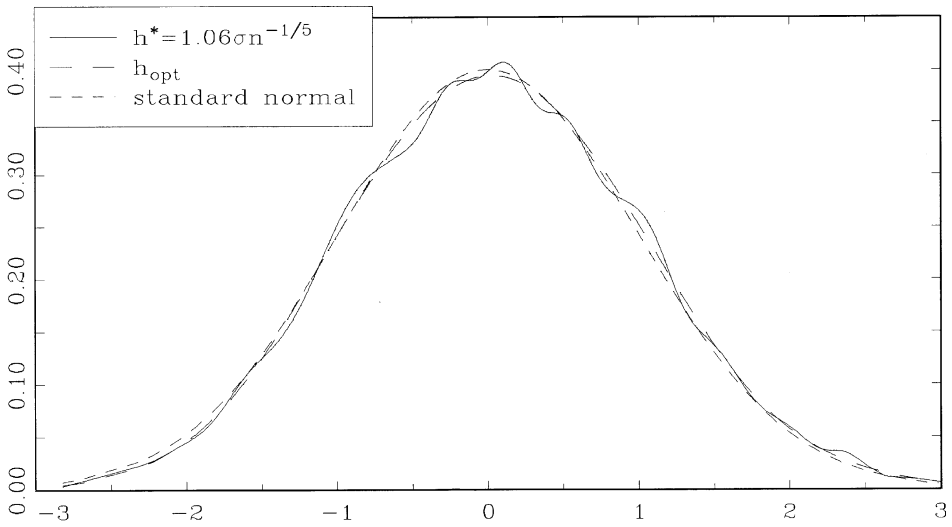

Figure 5. Density estimation for clustered data ($\rho = .6$

4. CONCLUSION

Despite the fact that most data used by economists is gathered in surveys, very little econometric analysis takes the survey structure into account. This paper is a first attempt to bring together the literature on non-parametric density estimation and survey sampling. As demonstrated in the simulation, ignorning the clustering in the data can cause seriously misleading density estimates.

The proposed solution is a higher-order kernel and a window width which takes into account the dependence in the data. As in the i.i.d. case with a second-order kernel, the optimal window width is proportional to $n^{-1/5}$, however a different proportionality constant is now required to minimize the AIMSE. A data-based method is proposed in this paper for choosing the proportionality constant based upon the standard deviation *and* the intra-cluster correlation of the sample data. This optimal window width and fourth-order kernel perform well for the levels of correlation commonly found in cross-sectional, survey data.

ACKNOWLEDGMENTS

I would like to thank Aman Ullah for the initial inspiration for this paper. I also appreciated the comments of two anonymous referees which helped to make the paper clearer.

REFERENCES

Deaton, A. *The Analysis of Household Surveys: A Microeconometric Approach to Development Policy*; Johns Hopkins University Press: Baltimore, MD, 1997, Published for the World Bank.

Fan, J.; Marron, J.S. Best Possible Constant for Bandwidth Selection. *Annals of Statistics* **1992**, *20*(4), 2057–2070.

Hall, P.; Lahiri, S.N.; Polzehl, J. On Bandwidth Choice in Nonparametric Regression with Both Short- and Long-Range Dependent Errors. *Annals of Statistics* **1995**, *23*(6), 1921–1936.

Hall, P.; Sheather, S.J.; Jones, M.C.; Marron, J.S. On Optimal Data-Based Bandwidth Selection in Density Estimation. *Biometrika* **1991**, *78*, 263–271.

Hardle, W. *Applied Nonparametric Regression*; Cambridge University Press: Cambridge, 1990.

Herrmann, E.; Gasser, T.; Kneip, A. Choice of Bandwidth for Kernel Regression when Residuals are Correlated. *Biometrika* **1992**, *79*, 783–795.

Kish, L. *Survey Sampling*; John-Wiley: New York, 1965.

Morrison, D.F. *Multivariate Statistical Methods*; McGraw-Hill: New York, 1976.

Pagan, A.; Ullah, A. *Non-Parametric Econometrics.* Cambridge University Press, 1999, forthcoming.

Rao, C.R. *Linear Statistical Inference and its Applications*; 2nd Ed. John Wiley & Sons: New York, 1973.

Silverman, B.W. *Density Estimation for Statistics and Data Analysis*; Chapman and Hall: London, 1986.

Thompson, S. *Sampling*; John Wiley & Sons: New York, 1992.

Ullah, A.; Breunig, R. Econometric Analysis in Complex Surveys. In *Handbook of Applied Economic Statistics*; Giles, D., Ullah, A., Eds.; Marcel Dekker: New York, 1998.