NONPARAMETRIC DENSITY ESTIMATION FOR IMMUNOLOGIC MEASUREMENTS

T. Yang and N. Dubin

New York University Medical Center
341 E. 25th St., 2nd Floor, NYC 10010

KEY WORDS: Local bandwidth, global bandwidth, immunologic measurement, HIV infection

1. Introduction

Suppose that X_1, ..., X_n are independent, identically distributed, real-valued random variables with an unknown probability density $f(x)$. We consider the estimation of the unknown density function $f(x)$. The method of density estimation has generated several considerable literatures over the past few years. The kernel method, adopted here, has proved to be one of the most popular approaches and is well reviewed by Fryer (1977). Briefly, an estimator $f_n(x)$ of the true density function $f(x)$ is constructed by placing a kernel function $K(x)$ over each observation on the data set, $\{X_1, ..., X_n\}$.

In our studies, we use a kernel estimator defined by

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

where K is called *kernel function* and h is called *bandwidth* or *smoothing parameter*. The kernel $K(x)$ has value 1 of integration, *i.e.* $\int K(x)dx = 1$.

An estimator of density function is called *global bandwidth* (GB) *estimator* if the bandwidth h is constant. Moreover, when h is depended upon the sample X_1, ..., X_n, the estimator is named *automatic global bandwidth* (AGB) *estimator*. Various properties of this kind of estimator have been studied since 1956. Rosenblatt (1956) and Parzen (1962) introduced the definition of the kernel estimator. Devroye and Wagner (1980) and Deheuvels and Hominal (1980) proved the uniformly strong convergence of the estimator.

An alternate estimator of the density function is called the *local bandwidth* (LB) *estimator*, for which the bandwidth h is a function of location x. Krieger & Pickands (1981) and Abramson (1982) studied the weak convergence of the LB estimator. Yang (1988) proved the uniformly strong convergency and other properties of the estimator.

The estimator $f_n(x,h)$ contains two unknown conponents, the kernel function K and the bandwidth h. In order to use $f_n(x,h)$ to approach the unknown density $f(x)$, we need a method to choose K and h. Several kernels are presented in the papers by Rosenblatt (1971) and Gasser et al. (1985). After determining the kernel, the selection of the bandwidth h is crucial to the performance of this estimator. If h is too small the estimator gives a curve that is too jumpy, being overly dependent on the particular realization of the data at hand. It is showing features that are not shared by the density f. If h is too large, the estimator creates a bias that can, by oversmoothing, eliminate intrinsic features of f.

One must set up a criterion to choose the optimal bandwidth, that is, to select h to minimize the integrated squared error of f_n,

$$\int [f_n(x,h) - f(x)]^2 \, dx.$$

Unfortunately, this depends on the unknown density f, while any practical method of choosing a bandwidth should depend on the sample. One may write

$$\int [f_n(x,h) - f(x)]^2 dx = \int {f_n}^2(x,h)dx - 2\int f_n(x,h)f(x)dx + \int f^2(x)dx.$$

Since the last summand is independent of h, the goal of minimizing this loss is equivalent to that of minimizing

$$\int {f_n}^2(x,h)dx - 2\int f_n(x,h)f(x)dx.$$

Notice that the second term depends

on the unknown function f. Therefore, this minimization cannot be realized in practice without knowledge of $f(x)$. However, one may write

$$\int f_n(x,h)f(x)dx = E_X[f_n(x,h)].$$

This motivates estimating the second term by

$$n^{-1}\sum_{j=1}^{n} f_{h,j}(X_j),$$

where $f_{h,j}(X_j)$ denotes the kernel density estimator with the jth observation deleted from the sample, *e.g.*

$$f_{h,j}(x) = \frac{1}{(n-1)h}\sum_{i\neq j} K\left(\frac{x-X_i}{h}\right).$$

Now the goal of minimizing the integrated squared error of f_n changes to that of minimizing the cross-validation function of f_n,

$$CV(h) = \int f_n^2(x,h)dx - 2n^{-1}\sum_{j=1}^{n} f_{h,j}(X_j). \qquad (2)$$

This was first suggested by Hall (1983) and Bowman (1984). Several papers (Hall (1983), Stone (1984), Burman (1985), and Marron (1985)) have shown that if h_{CV}^* is a suitably chosen bandwidth based on minimizing (2), then under some mild conditions, h_{CV}^* is asymptotically optimal, in the sense that

$$\Delta(h_{CV}^*, f)/\Delta(h_f^*, f) \rightarrow 1$$

or

$$h_{CV}^*/h_f^* \rightarrow 1,$$

in some mode of convergence, where h_f^* is the minimizer of integrated squared error,

$$\Delta(h,f) = \int [f_n(x,h) - f(x)]^2\, dx.$$

According the above result, we can select an optimal bandwidth based on minimizing the cross-validation instead of on minimizing the integrated squared error of f_n. In practice, we may apply numerical methods to determine the optimal bandwidth, a minimizer of the cross-validation function of f_n.

2. Main Theorem

Before discussing the realization of the kernel estimator, we consider its convergence. The two undetermined components of the kernel estimator in (1) are the kernel function K and the bandwidth h. In order to obtain an estimator of f, we must set some suitable conditions over the kernel K. In other words, we construct the kernel K to satisfy those conditions. Suppose that the kernel K is symmetric, bounded, and has bounded variation. K is a function with order k, that is,

$$\int K(x)x^j dx = \begin{cases} 1, & if\ j = 0; \\ 0, & if\ 1 \le j \le k-1; \\ C, & if\ j = k, \end{cases}$$

where C is a non-zero number, and $K^2(x)$ is integrable. If the density $f(x)$ has up to $(k+1)th$ derivatives and $[f^{(k)}(x)]^2$ is bounded, and integrable, it has been shown (Yang (1988)) that

Theorem 1 *Assume that τ is a function of x and X_1, ..., X_n with values in $[a,b]$, where $0 < a < b < \infty$. If we take the bandwidth h as a proportion of τ, $h = \tau n^{-1/(2k+1)}$, then we have*

$$\sup_x |f_n(x,h) - f(x)| \le O(n^{1/(2k+1)}(loglogn/\,n)^{1/2}) \qquad a.s.$$

According to the result of Theorem 1, we know the LB estimator is uniformly strong convergent. Therefore, in practice we may apply the LB estimator instead of the GB estimator.

Obviously, the computation of the LB estimator is more complicated than that of the GB estimator. However, there are advantages to using the LB estimator. First, we want to choose an optimal bandwidth h to minimize the mean squared error of $f_n(x)$. In other words, one selects an estimator approaching the unknown density in the sense of mean squared error as soon as possible. It has been proven (Yang (1988)) that the LB estimator is closer to $f(x)$ than is the GB estimator in certain cases. Second, one hopes to obtain an estimator (curve) with adaptive smoothness. In general, one would not want an estimator to be grossly under-smooth, even with minimum mean squared error. It is difficult to decide what amount of smoothness is suitable. One wants to achieve a balance between the error value and the smoothness. Applying the LB estimator, it is often possible to obtain a better balance. An example is presented in next section.

3. Application: Immunologic Measurements from Intravenous Drug Users

For our application, we used data collected from a cohort of intravenous drug users from drug detoxification programs and methadone maintenance treatment programs in New York City. This cohort has been described in detail by Des Jarlais et al.(1987). The subset of this cohort to be analyzed here consists of 392 subjects with at least one assessment of immunologic variables and known antibody status for Human Immunodeficiency Virus (HIV). One hundred ninety-one patients (48.7%) were consistently HIV-negative, 185 patients (47.2%) were consistently HIV-positive, and 16 patients (4.1%) converted from being HIV-negative to HIV-positive during the course of follow-up. Subjects were enrolled during 1984-85 and scheduled for yearly visits thereafter, although the actual follow-up intervals were somewhat longer, the mean time between the first and second visit being 15.8 months. Immunologic variables for which data were collected included T-helper (T4) cells, T-suppressor (T8) cells, T4/T8 ratio and total lymphocytes. For this particular illustration we consider T4 cells only.

Counts of T4 cells are known to vary enormously among individuals who are immunologically normal, as well as within such individuals over time. On the average, however, we expect the distribution of T4 counts among HIV-negative subjects to be relatively stable over time periods as long as several years. (Over decades one may observe an age-related decline.) Among HIV-positives we expect there to be a downward shift in the T4 count over time, as these target cells are depleted by the virus. Among HIV-converters it would be of interest to know whether, prior to HIV infection, the distribution of T4 counts were similar to that for consistent HIV-negatives. Subsequent to infection, the analogous comparison of HIV-converters to consistent HIV-positives is less direct, because the latter group had been infected for varying (and unknown) lengths of time. For a thorough discussion of such prevalent cohorts, see Brookmeyer and Gail (1987).

Consider first the sample distribution of T4 counts at first visit for the data subset of 191 HIV-negatives, abbreviated as T4NN1 (Table 1). A simple first step is to use parametric estimation. For example, we can choose a Gaussian model given by

$$f(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right], \quad -\infty < x < \infty.$$

Using the sample data, one substitutes the sample mean $\hat{\mu} = 1128.379$ and standard deviation $\hat{\sigma} = 411.9319$ in the above. However, the Shapiro-Wilk (1965) test indicates significant lack of fit ($p < 0.01$) for the normal distribution.

T4 MIDPOINT		FREQ	CUM FREQ	PERCENT	CUM PERCENT
0		0	0	0.00	0.00
120		0	0	0.00	0.00
240		0	0	0.00	0.00
360	**	2	2	1.05	1.05
480	******	6	8	3.14	4.19
600	************	13	21	6.81	10.99
720	********************	19	40	9.95	20.94
840	************************	23	63	12.04	32.98
960	**********************	22	85	11.52	44.50
1080	*****************************	27	112	14.14	58.64
1200	*****************	17	129	8.90	67.54
1320	************	13	142	6.81	74.35
1440	**************	14	156	7.33	81.68
1560	*********	9	165	4.71	86.39
1680	********	8	173	4.19	90.58
1800	******	6	179	3.14	93.72
1920	*****	5	184	2.62	96.34
2040	**	2	186	1.05	97.38
2160	***	3	189	1.57	98.95
2280	**	2	191	1.05	100.00
2400		0	191	0.00	100.00
2520		0	191	0.00	100.00
2640		0	191	0.00	100.00
2760		0	191	0.00	100.00

5 10 15 20 25

FREQUENCY

Table 1. Frequency distribution of T4 counts at the first visit, among subjects who were consistently HIV negative

To obtain f_n for this sample, we use a kernel function given by

$$K(x) = \begin{cases} 3/4(1-x^2), & if\ |x| \leq 1; \\ 0, & elsewhere. \end{cases}$$

This is a second-order kernel function. Also needed is the bandwidth h. To study the effect on the estimators of different choices of bandwidths, we use the cross-validation function to select several possible bandwidths. Figure 1 shows the cross-validation curve as a function of the bandwidth. The horizontal axis denotes the bandwidth h and the vertical axis denotes the cross-validation value. Note that the minimum occurs at approximately h= 230 and h= 350. The estimation of density f_n with h= 230 and h = 350 is presented in Figure 2 (a) and (b), respectively. It would be difficult to decide between these two density curves; nonetheless, both seem to ex-

hibit the essential features of the underlying distribution. In contrast, Figure 2 (c) shows that for h= 110 the estimator gives a density curve which seems undersmoothed.

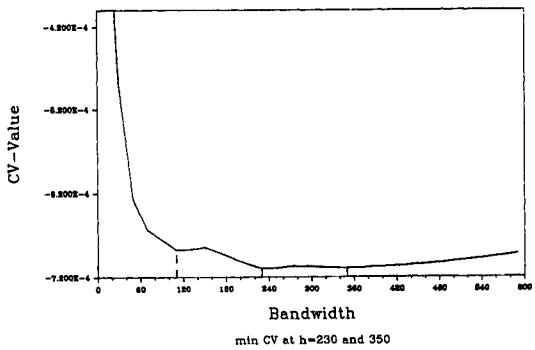

Figure 1. bandwidth v.s. cross-validation value, min CV at h= 230 and 350

Now consider the local bandwidth estimator. The bandwidth function may be defined by

$$h(x; C_1, C_2) = \frac{C_1}{N(C_2)}, \qquad (3)$$

where $C_1 = kR(X_1, ..., X_n)$ is called *scale parameter*, k> 0, and C_2 is called *frequency parameter*. $R(X_1, ..., X_n)$ is range of sample $X_1, ..., X_n$, that is,

$$R(X_1, ..., X_n) = max\{X's\} - min\{X's\}.$$

$N(C_2)$ is frequency of sample $X_1, ..., X_n$ on $[x - C_2, x + C_2]$, that is,

$$N(C_2) = max\{1, \# \text{ of X's on } [x - C_2, x + C_2]\}.$$

Substituting equation (3) into the cross-validation function (2), we can compute CV values for different choices of C_1 and C_2. A two-dimensional cross-validation surface is shown in Figure 3. A minimum occurs at approximately $C_1 = R(X_1, ..., X_n) = 1921.18$ and $C_2 = 30$. Figure 4 shows the resultant local bandwidth density estimator. Notice that this local bandwidth estimator exhibits overall characteristics similar to the global estimators with h= 230 and h= 350 (Figure 2 (a) and (b)).

An objective criterion to decide among these density estimators may be made on the basis of minimizing the cross-validation function (2). For the local bandwidth estimator the $CV = -1.4277E-3$, which is

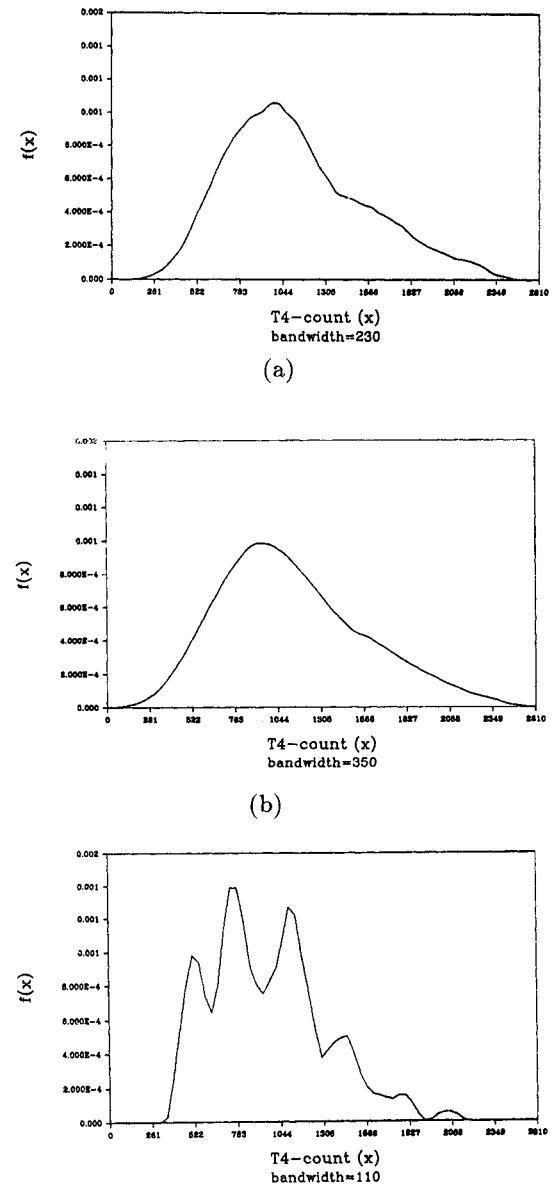

Figure 2. global bandwidth density estimator of T4 with (a) h= 230, (b) h= 350, and (c) h= 110.

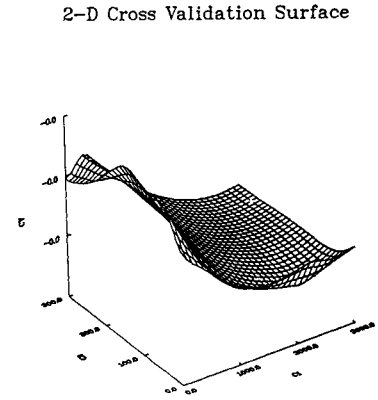

Figure 3. C_1 and C_2 v.s. cross-validation values

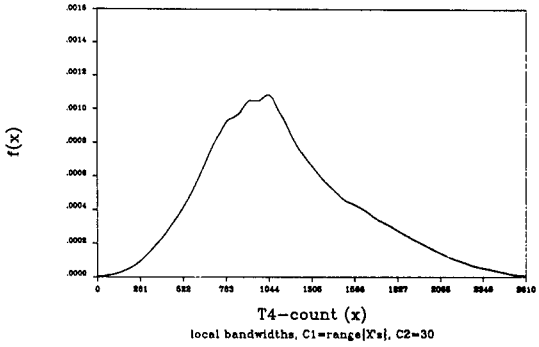

Figure 4. local bandwidth density estimator with $C_1 = range\{X's\}$ and $C_2 = 30$

smaller than the -0.7097E-3 minimum CV value for the global bandwidth estimator. In general, the local bandwidth estimator was superior for all the data subsets considered in our application, as may be determined from Table 2.

Table 2. Comparison of the CV value for the GB and LB estimators for the various data subsets

Data Subsets	GB	LB	Ratio
T4NN1	-0.7097E-3	-1.4277E-3	50:100
T4NN2	-0.9081E-3	-1.6516E-3	51:100
T4PP1	-0.9086E-3	-1.8170E-3	50:100
T4PP2	-1.0457E-3	-2.0487E-3	55:100

In Table 2, the acronym T4NN1 refers to data from the first visit for consistently HIV-negative patients, T4NN2 refers to data from HIV-negatives at their second visit, T4PP1 refers to data from consistently HIV-positive patients at their first visit, and T4PP2 refers to data from HIV-positives at their second visit. The ratio is the absolute CV value of the GB estimator divided by the absolute CV value of the LB estimator. A ratio less than one indicates that the LB estimator is superior in terms of the minimum CV criterion.

In spite of the superiority of the LB estimator with respect to the minimum CV criterion, occasionally one obtains an LB estimator which is undersmoothed, as was the case for data subset T4NN2. In addition, for HIV converters, the data were too sparse to successfully apply the LB estimator. For these reasons and also to consistently use one type of density estimator in the graphical comparisons which follow, we chose to illustrate our substantive findings with the GB estimator.

Consider the comparison of distributions of T4 counts for HIV-negatives at their first (T4NN1) and second (T4NN2) visits (Figure 5). The distributions are remarkably similar, in spite of the relatively low first-visit-to-second-visit correlation for individual subjects (Pearson correlation coefficient = 0.11). Notice, however, that the distribution at second visit is moderately shifted to the left. In Figure 6 we compare first-visit T4 counts for HIV-positives (T4PP1) to HIV-negatives (T4NN1). As expected, one sees a dramatically leftward-shifted distribution for the HIV-positives, reflecting substantially decreased T4 counts in this subset. Among HIV-positives, one also sees a leftward shift over time when comparing their first (T4PP1) to their second visit (T4PP2) T4 counts (Figure 7). Although there were relatively few ($n = 16$) HIV-converters available to study, they represent a valuable opportunity to assess T4 counts before and after HIV infection in the same group of subjects. Remarkably, in spite of the small sample size, among HIV-converters the distribution of T4 counts at the visit prior to infection with HIV (T4NP1) was very similar to that for HIV-negatives at first visit (T4NN1) (Figure 8), although shifted somewhat to the left. Also remarkably, among HIV-converters the distribution of T4 counts subsequent to infection (T4NP2) had already dramatically shifted to the left (Figure 9) and was comparable to that for HIV-positives at first visit (T4PP1) (Figure 10).

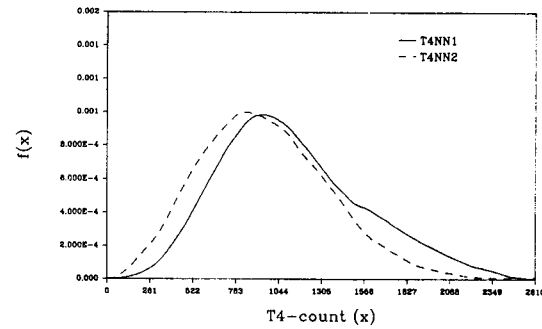

Figure 5. Comparison of GB estimators for T4NN1 and T4NN2

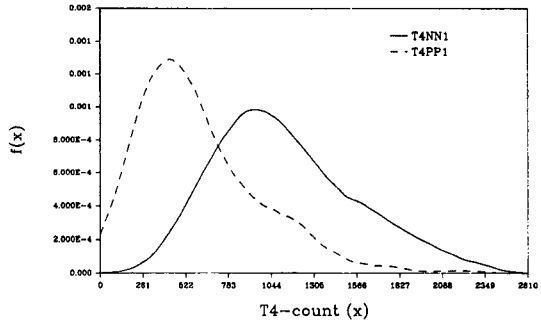

Figure 6. Comparison of GB estimators for T4PP1 and T4NN1

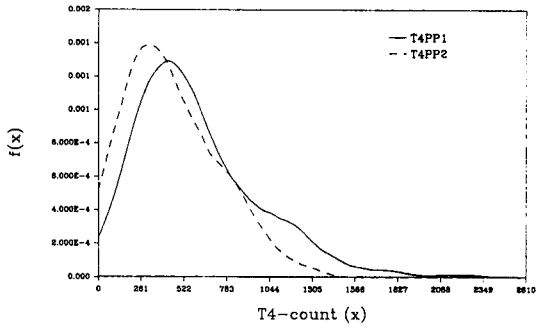

Figure 7. Comparison of GB estimators for T4PP1 and T4PP2

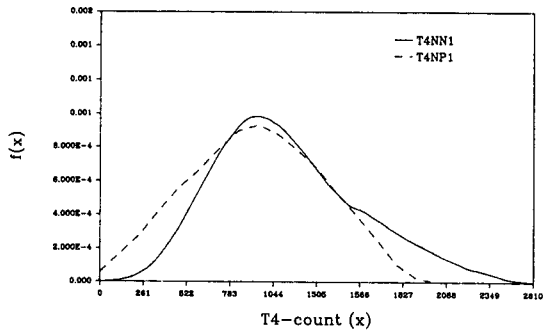

Figure 8. Comparison of GB estimators for T4NP1 and T4NN1

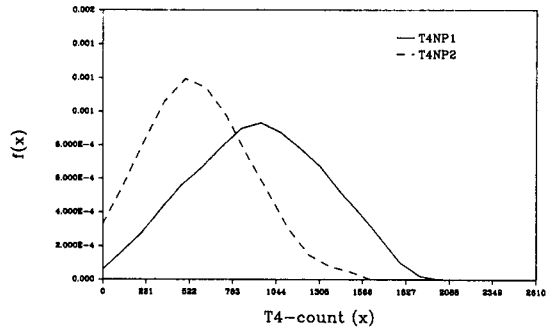

Figure 9. Comparison of GB estimators for T4NP1 and T4NP2

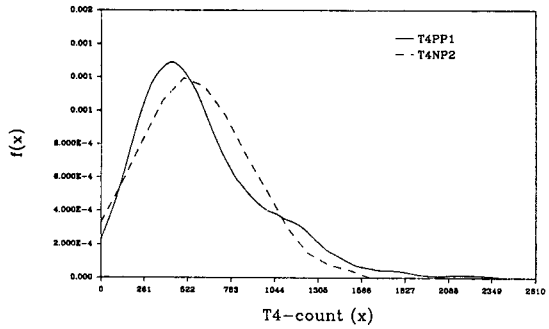

Figure 10. Comparison of GB estimators for T4PP1 and T4NP2

Acknowledgment

This research was supported by grants DA04722 and DA03574 from the National Institute on Drug Abuse and center grant CA-16087 from the National Cancer Institute. We would like to thank Drs. Don Des Jarlais and Michael Marmor for providing us with the sample data.

References.

Abramson, I. (1982). Arbitrariness of the pilot estimator in adaptive kernel methods. *J. Multi. Anal.* **12** 562-567

Brookmeyer, R. and Gail, M.H. (1987). Biases in prevalent cohorts. *Biometrics* **3** 739-749

Burman, P. (1985). A data dependent approach to density estimation. *Z. Wahrsch. verw. Gebiete.* **69** 609-628

Deheuvels, P. and Hominal, P. (1980). Estimation automatique de la density. *Revue de Statistique,* **25** 25-55

Des Jarlais, D. C., Friedman, S. R., Marmor, M., Cohen, H., Mildvan, D., Yancovitz, S., Mathur, U., El-Sadr, W., Spira, T. J., Garber, J., Beatrice, S. T., Abdul-Quader, A. S., and Sotheran, J. L. (1987). Development of AIDS, HIV seroconversion, and potential cofactors for T4 cell loss in a cohort of intravenous drug users.

Devroye, L. and Wagner, T. J. (1980). The strong uniform consistency of kernel density estimates. In *Multivariate Analysis,* **V59-77**, P. R. Krishndiah Edit, New York: North-Holland.

Fryer, M. J. (1977). A review of some non-parametric methods of density estimation. *J. Inst. Math. Applic.* **20** 335-354

Gasser, T., *Müller*, H-G., and Mammitzsch, V. (1985). Kernels for nonparametric curve estimation. *J. R. Statist.* **B 47** 238-252

Hall, P. (1983). Large sample optimality of least squares cross-validation in density estimation. *Ann. Statist.* **11** 1156-1174

Krieger, A. M. and Pickands, J. (1981). Weak convergence and efficient density estimation at a point. *Ann. Statist.* **9** *1066-1078*

Marron, J. S. (1985). An asymptotically efficient solution to the bandwidth problem of kernel density estimation. *Ann. Statist.* **13** 1011-1023

Parzen, E. (1962). One estimation of a probability density function and mode. *Ann. Math. Statist.* **33** 1065-1076

Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *Ann. Math. Statist.* **27** 832-837

Rosenblatt, M. (1971). Curve estimates. *Ann. Math. Statist.* **42** 1815-1842

Shapiro, S. S. and Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika,* **52** 591-611

Stone, C. J. (1984). An asymptotically optimal window selection rule for kernel density estimates. *Ann. Statist.* **12** 1285-1297

Yang, T. (1988). On the nonparametric curve estimations with local bandwidth selections. *Ph.D. Dissertation, Dept. of Math. Sci., Univ. of Cincinnati.*