Nonparametric Statistics, Vol. 4, pp. 409–426
Reprints available directly from the publisher
Photocopying permitted by license only

A KERNEL ESTIMATOR FOR DISCRETE DISTRIBUTIONS

BALAJI RAJAGOPALAN and UPMANU LALL*

Utah Water Research Laboratory, Utah State University, Logan, UT-84322-8200

(Received: February 16, 1994; Revised: July 4, 1994; Accepted: July 22, 1994)

We present a discrete kernel estimator appropriate for estimating probability mass functions (p.m.f's) for integer data. Discrete kernel functions analogous to the Beta functions used as kernels in the continuous case are derived for the interior and for the boundary of the domain. An integer bandwidth is considered. Cross validation is used for bandwidth selection. The estimator was motivated by the need to characterize processes (e.g., mixtures of geometric distributions) with long tailed distributions with high mass near the origin, and integer arguments of the random variable. Monte Carlo comparisons with the Hall and Titterington [8] (HT) estimator are offered. An application for estimating the p.m.f.'s of wet and dry spell lengths for a nonparametric renewal model of daily rainfall is also presented. Other possible methods for obtaining discrete weight sequences are also presented.

KEYWORDS: Discrete kernel, multinominal smoothing, boundary kernels

1. BACKGROUND

The problem of nonparametric smoothing of the empirical discrete p.m.f (or multinomial cell proportions) has been of interest in recent years. However, it has not been studied as intensively as nonparametric density estimation, its counterpart in the continuous case. Hall and Titterington [8] mention that smoothing can be beneficial when there are many cells with small or zero frequencies, i.e., the data are sparse. Here we consider that we have a sample $x_1, \dots, x_n$ for n multinomial trials with possible outcomes $1, 2, \dots, k_{\max} \in V$ with probabilities of occurrence $p_1, \dots, p_{k_{\max}}$ that are unknown. Estimates $\hat{p}_i$ of the probabilities p_i may be obtained as sample relative frequencies ($\tilde{p}_i = n_i/n$) or cell proportions, or by smoothing the $\tilde{p}_i$. In the latter case we presume that V is an ordered set and that "distance" between its members is definable through a standard Lebesgue measure. We consider cases where the set V may be bounded or unbounded, and focus on developing an appropriate smoother for the sample relative frequencies that properly deals with the discrete nature of the process.

Our practical interest lay in developing a discrete, nonparametric p.m.f for data on the length (in days) of dry or wet spells of rainfall. The shortest spell considered is 1 day. In general, the longest possible spell is not known a priori. Data suggests long right tailed distributions for dry spell length that may correspond to a mixture of geometric p.m.f.'s (see Rajagopalan *et al.* [10]).

The concept of smoothing in the context of multinomial cell probability estimation was introduced by Good [6 and 7]. This was later studied and improved by Fienberg and Holland [5], Stone [13], Titterington [14], Titterington [15], Aitchison and Aitken [1], Titterington and Bowman [16] among others. Bishop *et al.* [2] show that these estimators are often better than the cell proportion estimate under squared error loss. Hall and Titterington [8] argue that $\tilde{p}_i$ may not be consistent in data sparse situations. The smoothing estimators developed by Wang and Van Ryzin [17], Simonoff [12] and Hall and Titterington [8] formed a starting point for our work.

The general form of smoothing estimators in this context is given by

$$\hat{p}_i = \sum_{j=-\infty}^{j=\infty} K(i,j,h)\,\tilde{p}_j \quad i,j\in\mathrm{I}, \text{ the set of integers} \tag{1}$$

$K(i,j,h)$ is a weight function or kernel, $\tilde{p}_j$ is the relative frequency of cell j and h is called the bandwidth of window width.

Wang and Van Ryzin [17] developed a class of estimators of the form (1), using a Geometric kernel (WV) ($K(i,j,h) = 0.5\,h(1-h)^{|i-j|}$ if $|i-j| \geq 1$; $K(i,j,j,h) = (1-h)$ if $i = j$ and $h \in [0,1]$). The "drop off" of weights associated with the Geometric kernel is rapid. Wang and Van Ryzin [17] estimate h under an approximate (MSE) criterion formed by truncating the Geometric kernel beyond two cells. As a result, very little smoothing is obtained in most cases and not much may be gained for sparse data.

By imposing a smoothness constraint on the cell probabilities, Simonoff [12] obtained relative consistency results for an estimator based on a maximum penalised likelihood criterion (MPLE). In this approach, the estimates $\hat{p}_i$ are solved by minimizing a penalized likelihood function defined as,

$$L = \sum_{i=1}^{k_u} n_i \log(\hat{p}_i) - \beta \sum_{i=1}^{k_u} \{\log(\hat{p}_i/\hat{p}_{i+1})\}^2$$

such that

$$\sum_{i=1}^{k_u} \hat{p}_i = 1 \tag{2}$$

$\beta \geq 0$, is a smoothing parameter, and V: $[1, k_u]$

The estimates from MPLE dependend significantly on the extent of estimation required (i.e., k_u) beyond the maximum observed cell (i.e., $k_{\max}$). This is of concern, because we would prefer a natural extension of the tail of the p.m.f by the method used, rather than a prior specification of its extent.

The estimator developed by Hall and Titterington [8] (here after referred to as HT) is given as,

$$\hat{p}_i = \sum_{j=-\infty}^{j=\infty} W(i,j,h)\,\hat{p}_j \tag{3}$$

where $W(i,j,h) = K((i-j)/h)/s(h)$, $h > 1$ and $s(h) = \sum_{j=-\infty}^{j=\infty} K(j/h)$. $K(\cdot)$ is any suitable continuous univariate kernel function, with compact support satisfying the conditions of positivity, intergration to unity, symmetry, and finite variance which are,

$$\text{(a) } K(u) > 0; \quad \text{(b) } \int K(u)\mathrm{d}u = 1; \quad \text{(c) } \int uK(u)\mathrm{d}u = 0; \quad \text{(d) } \int u^2K(u)\mathrm{d}u = k2 \neq 0 \quad (4)$$

where $u = (i - j)/h$, and $s(h)$ is a multiplicative factor required to normalize the continuous variable kernel function for use with discrete data, such that the desired conditions on $W(\cdot)$ viz. $\sum_{j=-\infty}^{j=\infty} W(i,j,h) = 1$ and $\sum_{j=-\infty}^{j=\infty} j\,W(i,j,\mathrm{h}) = 0$ are satisfied. Hall and Titterington [8] proposed a cross-validatory procedure for selecting h. This was later studied by Dong and Simonoff [3] who extended this estimator to boundary kernels.

It is well known that kernel estimators suffer from increased bias in the boundary region (i.e. $1 \leq i \leq h+1$ in our situation of interest). For the estimates of cells in the boundary there is a lack of full complement of observations on either side of the cell of estimate. As a result, the desired conditions on $W(i,j,h)$ mentioned above will not be preserved. To correct this, special boundary kernels that satisfy the required conditions are used (Müller [9]). Müller [9] formally developed special boundary kernels in the continuous case. Dong and Simonoff [3] developed boundary kernels (condition 4(a) is relaxed) that could be used in the HT estimator for the discrete case. We refer to the HT estimator with the boundary modification of Dong and Simonoff [3] as HT/DS.

We performed comparisons of these three estimators (viz. WV, MPLE and HT/DS) on data generated from long tailed distributions (see Rajagopalan *et al.* [10]) and found HT/DS to be the best. Hence, we compare the relative performance of the estimator we develop later in this paper with HT/DS.

For finite samples, some disquieting aspects of the HT estimator become apparent. The non-integer bandwidth leads to an effective kernel that also varies with h in a manner quite different from that prescribed by (4). The effective integer support of $W(i,j,h)$ is $[(i-h^*),(i+h^*)]$, where h^* is the closest integer greater than or equal to h. HT/DS kernels are defined as quadratics or other polynomials over $[i-h, i+h]$. Since this is not the effective integer support of the kernel the effective kernel over the space of integers is not the quadratic defined.

Alternatively, it is possible to develop a kernel that recognizes the data to be in integer space, has an integer bandwidth and satisfies all the required conditions in the integer space. This also obviates the need for normalization of the kernel weights as done in HT/DS. We explored this line of thought and, sought a direct, discrete analog of the continuous kernel density estimator.

The estimator is first presented. Bandwidth estimation is described next. Monte Carlo comparisons with HT/DS are then present. Comparisons with real data sets follow. Discussion of the new estimator and other possible discrete estimators conclude the paper.

2. THE DISCRETE KERNEL ESTIMATOR (DKE)

We define our estimator $\hat{p}_i$ for cell i through a weighted linear combination of the sample relative frequencies, $\tilde{p}_i$ as,

$$\hat{p}_i = \sum_{j=1}^{k_{\max}} K(t_j)\tilde{p}_j \quad (5)$$

where i, j and h are positive integers, $t_j = (i-j)/h, K(t)$ is a kernel function, and V: $[1, \infty]$. In the continuous case, Epanechnikov [4] showed that the MSE optimal kernel of second order, is the quadratic kernel (QK), also known as the Epanechnikov kernel. The general form of the QK is,

$$K(u) = au^2 + b \quad \text{for} \quad |u| \leq 1 \tag{6}$$

In the continuous case, $a = -0.75$, $b = 0.75$. Scott [11], p. 140, Equation 6.25 points out that this corresponds to a Beta density function, defined for $t \in [-1, 1]$. Other members of this class can be used if additional smoothness is desired.

Here, we chose a discrete quadratic (DQ) kernel of the form $K(t_j) = at_j^2 + b$, where $t_j = (i-j)/h$. The main focus then is to specify the constants a and b for the interior $(i > h+1)$ and the boundary region$(1 \leq i \leq h+1)$. The constants a and b are solved to satisfy: (A) the kernel function goes to zero for $|i-j| \geq h$, i.e., $K(t_j) = 0$ for $|t_j| \geq 1$, (B) sum of the weights is unity, i.e., $\sum_{j=i-h}^{j=i+h} K(i-j/h) = 1$ and (C) the first moment of the kernel function is zero, i.e., $\sum_{j=i-h}^{j=i+h} K(i-j/h)t_j = 0$. Note that the above conditions are the discrete versions of the conditions given in Equation (3) for continuous variable kernels. One could choose higher order Beta kernels and derive results similar to these that follow for DQ.

For the interior region $(i > h+1)$ using Conditions (A) and (B) gives Equations (7) and (8),

$$K(t_{i+h}) = K(t_{i-h}) = 0 \tag{7}$$

$$\sum_{j=i-h}^{j=i+h} (at_j^2 + b) = 1, \quad \text{where } t_j = (i-j)/h \tag{8}$$

Condition (C) is satisfied if $a = -b$. The coefficients a and b can now be expressed in terms of the bandwidth h as,

$$a = \frac{-3h}{(1-4h^2)} \quad \text{and} \quad b = \frac{3h}{(1-4h^2)} \tag{9}$$

For the boundary region $(1 < i \leq h+1)$ Condition A is modified as,

$$K(t) = 0 \quad \text{for} \quad t \leq -1 \quad \text{and} \quad t \geq q \quad \text{where} \quad q = (i-1)/h. \tag{10}$$

Applying Conditions (B) and (C) we get Equations (11) and (12).

$$\sum_{j=1}^{j=i+h} (at_j^2 + b) = 1 \tag{11}$$

$$\sum_{j=1}^{j=i+h} t_j(at_j^2 + b) = 0 \tag{12}$$

Solving for a and b we get,

$$a = \frac{-D}{2h(h+i)} \times \frac{1}{\left(\frac{E}{4h^3} - \frac{CD}{12h^3(h+i)}\right)}, \quad b = \left[1 - \frac{aC}{6h^2}\right]\frac{1}{(h+i)} \tag{13}$$

where,

$$C = h(h+1)(2h+1) + (i-2)(i-1)(2i-3)$$
$$D = -h(h+1) + (i-2)(i-1)$$
$$E = (-h(h+1))^2 + ((i-2)(i-1))^2$$

From Equation (10) it can be seen that at the boundary (i.e., $i = 1$) the weight associated with the kernel is zero. This is not desirable because, for long tailed distributions defined on the interval $[1, \infty)$ most of the mass is concentrated right at $i = 1$. Clearly, using the boundary modification in Equation (13) for estimation of p.m.f at the boundary (i.e., $i = 1$) will introduce a large bias in the estimate. Therefore, we need a further modification for estimation at $i = 1$. By not enforcing the $K(t) = 0$ at $i = 1$, we modify (A) to be

$$K(t) = 0 \quad \text{for} \quad t \leq -1 \tag{14}$$

while Equation (11) and (12) remain the same. Solving Equations 14, 11 and 12 for a and b we get,

$$a = \frac{-D}{2h^2} \times \frac{1}{\left(\frac{E}{4h^3} - \frac{CD}{12h^4}\right)}, \quad b = \left[1 - \frac{aC}{6h^2}\right]\frac{1}{h} \tag{15}$$

where,

$$C = h(h-1)(2h-1)$$
$$D = -h(h-1)$$
$$E = -(h(h-1))^2.$$

From Equations (9), (13) and (15) note that the kernel and hence, the estimator $\hat{p}_i$ is expressed strictly in terms of the bandwidth h. An optimal choice of h then completes the definition of the estimator.

Three criterion often used for bandwidth estimation are (1) direct minimization of average mean square error (MSE) (2) Maximum likelihood cross validation (MLCV) and (3) Least squares cross validation (LSCV). These could be optimized over a discrete set of h values.

We tested all the three methods and found LSCV to be the best. Hall and Titterington [8] and Dong and Simonoff [3] also argue in favour of LSCV. The bandwidth is selected by minimizing the LSCV function given as,

$$\mathrm{LSCV}(h) = \sum_{i=1}^{k_{\max}} (\hat{p}_i)^2 - \frac{2}{n} \sum_{i=1}^{k_{\max}} \hat{p}_{-i} n_i \tag{16}$$

where, $\hat{p}_{-i}$ is the estimate of the i^{th} cell, by dropping the i^{th} cell and n. In a related context, Hall and Titterington [8] also show that cross-validation automatically adapts the estimator to an extreme range of sparseness types. If the multinomial is only slightly sparse, cross-validation will produce an estimator which is virtually the same as the cell-proportion estimator. As sparseness increases, cross-validation will automatically supply more and more smoothing, to a degree which is asymptotically optimal.

An example application comparing DKE (with DQ kernel) to HT/DS with QK based kernels for four data sets is shown in Figures 1, 2, 3 and 4. The data in Figure 1 was sampled from a Geometric distribution (G1) defined as $G(\pi = 0.2)$. The data in Figure 2 was sampled from a mixture of two Geometric distributions (G2) defined as $(0.3G(\pi = 0.9) + 0.7G(\pi = 0.2))$. The sample sizes for G1 and G2 are 250. Figure 3 shows the p.m.f estimates estimated for the mines data, analysed by Dong and Simonoff [3]. Figure 4 shows the estimated p.m.f from both estimators of dry spell length data, for season 3 (i.e., Jul–Sep) for the station Woodruff, in Utah. The sample size in this case was 539. All four figures indicate that both DKE and HT/DS perform comparably. As both the estimators are similar this is expected. We investigate through Monte Carlo simulations, the behaviour of these estimates for selected situations. The behaviour of

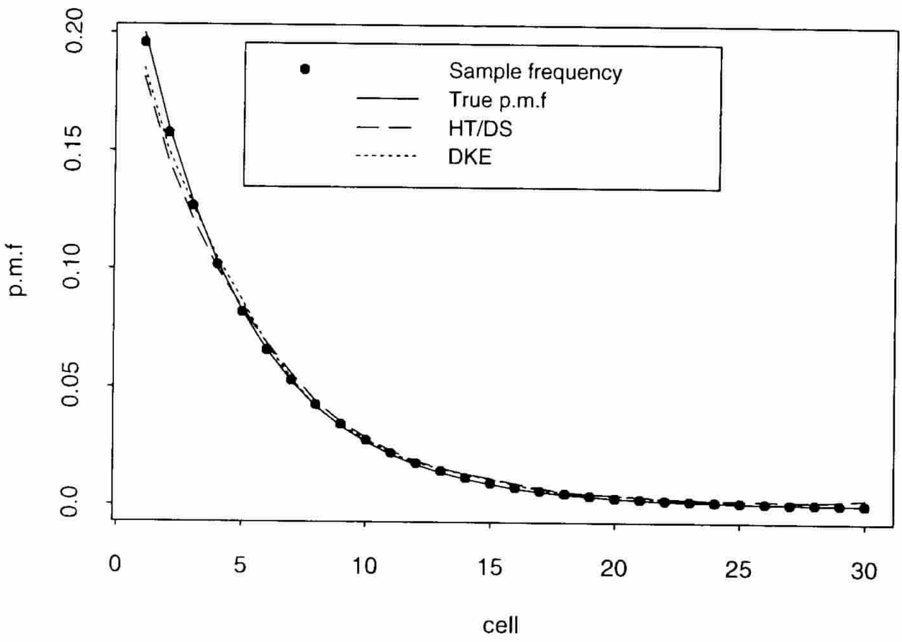

Figure 1. True p.m.f, estimated p.m.f from HT/DS and DKE of a sample of size 250, data generated from Geometric ($\pi = 0.2$), along with the sample frequency.

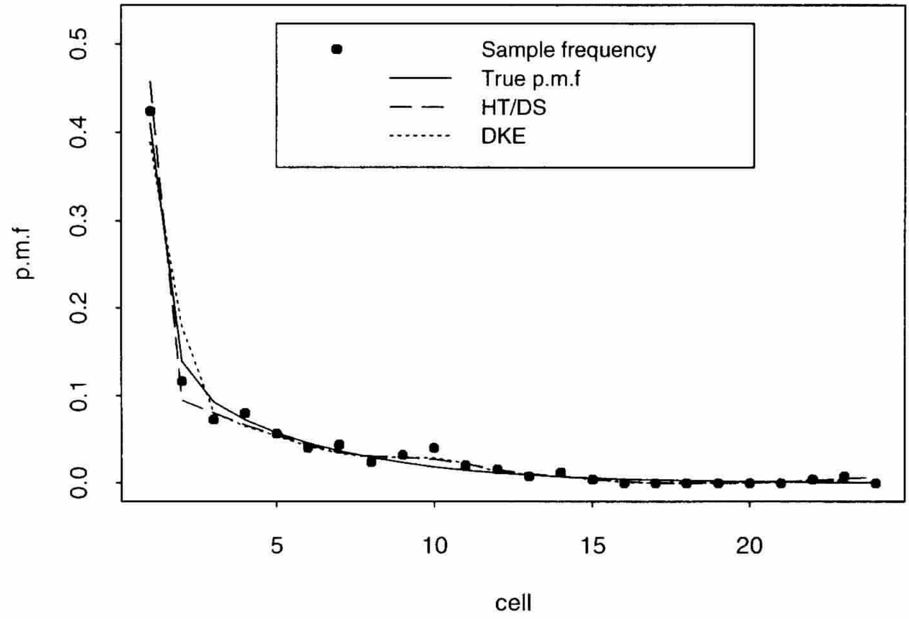

Figure 2. True p.m.f, estimated p.m.f from HT/DS and DKE of a sample of size 250 generated from 0.7* Geometric ($\pi = 0.2$) + 0.3* Geometric ($\pi = 0.9$), along with the sample frequency.

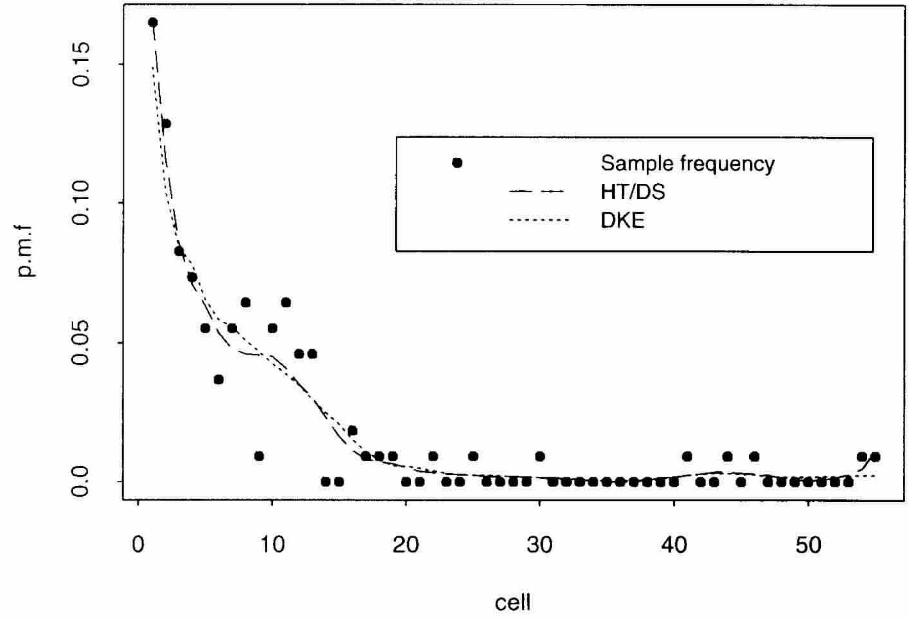

Figure 3. Estimated p.m.f from HT/DS and DKE of the mines data, along with sample frequency.

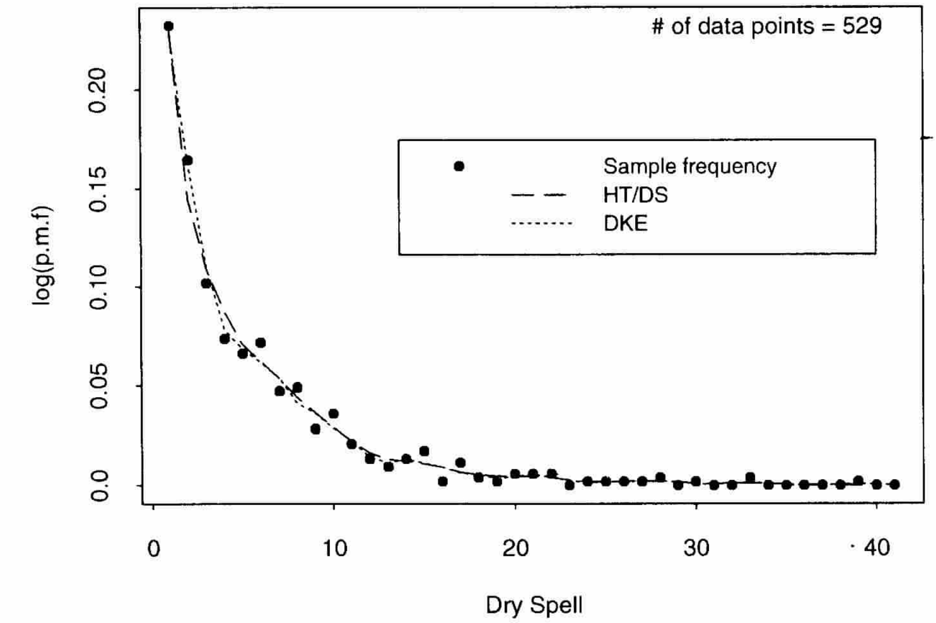

Figure 4. Estimated p.m.f from HT/DS and DKE of the dry spell length data of Woodruff, Utah, along with the sample frequency.

the weight sequence from both the estimators are also probed. The results are discussed in the following section.

3. MONTE CARLO COMPARISONS

We present results from Monte Carlo simulations, comparing our estimator with the HT/DS estimator using QK. Data sets were generated from situations that may be of interest in our particular context (e.g., geometric distribution, with a considerable boundary region). We generated 500 realizations from the two populations G1 and G2. Sample sizes chosen were $n = 50, 100, 200, 300, 500$.

The statistical measures computed to assess the relative performance of DKE and HT/DS estimators are:

1. Average Sum of Squared Errors (ASSE) $(\sum_{j=1}^{j=n\mathrm{sim}}(\sum_{i=1}^{i=k_u}(\hat{p}_{ij}-p_i)^2)/n\mathrm{sim})$ across all realizations for each sample size.
2. Sum of Squared Error (SSE$_j$)$(\sum_{i=1}^{i=k_u}(\hat{p}_{ij}-p_i)^2)$ for each realization $j = 1, \ldots, n\mathrm{sim}$.
3. Average Sum of Absolute Error (ASAE)$(\sum_{j=1}^{j=n\mathrm{sim}}(\sum_{i=1}^{i=k_u} abs(\hat{p}_{ij}-p_i))/n\mathrm{sim})$ across all realizations for each sample size.
4. Cell Root Mean Square Error (CRMSE) $\{\sum_{j=1}^{j=n\mathrm{sim}}((\hat{p}_{ij}-p_i)^2)/n\mathrm{sim}\}^{0.5}$ across all realizations for each sample size and for each cell $i = 1, \ldots, k_u$.
5. Fractional Cell Root Mean Square Error: $\mathrm{FCRMSE}_i = \mathrm{CRMSE}_i/p_i$.

6. Average Cell Bias (CBIAS_i) $\sum_{j=1}^{j=n\mathrm{sim}}((\hat{p}_{ij}-p_i)/n\mathrm{sim})$ across all realizations for each size and for each cell $i = 1, \ldots, k_u$.
7. Fractional Cell Bias: $\mathrm{FCBIAS}_i = \mathrm{CBIAS}_i/p_i$.
8. Coefficient of variation of bandwidth $C_v = s/\bar{h}$ for each sample size. Where s and $\bar{h}$ are the standard deviation and mean of the bandwidths obtained for all the $n$sim realizations.

Note that we chose k_u to be 30 in this case, and p_i's are the true p.m.f's obtained from the known underlying distributions from the samples were generated, $n$sim is the number of simulations, in our case it is 500.

Table 1 shows the ASSE and ASAE for the two estimators for the two populations G1 and G2 considered. It can be observed from Table 1 and Figures 5 and 6 that the performance of the two estimators over these two measures is quite close. Figures 5 and 6 indicate that the ASSE appears to decrease with n at rates -1.03 and -0.86 for HT/DS and -0.85 and -0.9 for DKE, for G1 and G2 respectively. These rates are very similar, and are close to the rate n^{-1} as anticipated in Hall and Titterington's [8] Theorem 2.1. However, the SSE for HT/DS has a larger spread that DKE as can be seen from Figures 7 and 8 for G1 and G2 respectively for a sample size of 50. The results were generally similar for other sample sizes.

As mentioned earlier we are interested in the behaviour of these estimators at the boundary (left boundary) and in the tails. To assess this, CRMSE_i and FCRMSE_i for different sample sizes n were estimated. As an illustration we present the estimates of FCRMSE_i for sample sizes 50 and 500 for G1 in Figures 9a and 9b respectively. Figures 10a and 10b are corresponding figures for G2. These figures suggest that DKE performs better than HT/DS in the tail region for all sample sizes, more so for smaller sample sizes. The results for other sample sizes were intermediate.

From Figures 11 and 12 we see that part of the poorer performance of HT/DS in the tails is due to higher bias.

Table 1. Comparison of ASSE and ASAE

	ASSE			*ASAE*		
	DKE	*PAR*	*HT/DS*	*DKE*	*PAR*	*HT/DS*
Samples generated from G1 (Geometric ($\pi = 0.2$))						
$n = 50$	0.0058	0.0008	0.0084	0.2032	0.0816	0.2737
$n = 100$	0.0032	0.0006	0.0038	0.1558	0.0599	0.1814
$n = 200$	0.0019	0.0003	0.0019	0.1183	0.4250	0.1264
$n = 300$	0.0013	0.0002	0.0012	0.1000	0.0323	0.0987
$n = 500$	0.0008	0.0000	0.0008	0.0780	0.0226	0.0797
Samples generated from G2 (0.7* Geometric ($\pi = 0.2$) + 0.3* Geometric ($\pi = 0.9$))						
$n = 50$	0.0080	–	0.0081	0.2300	–	0.2481
$n = 100$	0.0039	–	0.0038	0.1676	–	0.1638
$n = 200$	0.0021		0.0022	0.1261	–	0.1194
$n = 300$	0.0016	–	0.0016	0.1071	–	0.0978
$n = 500$	0.0010	–	0.0011	0.0855	–	0.0785

NOTE: PNR is the fitted parametric (in this case the fitted Geometric distribution)

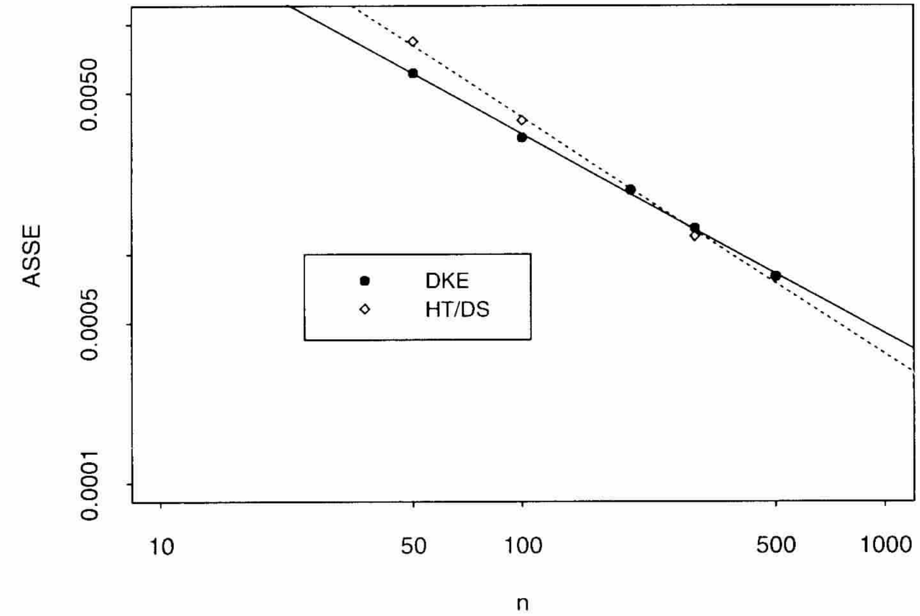

Figure 5. Log-Log plot of ASSE with sample size n, of samples generated from Geometric ($\pi = 0.2$) along with the fitted lines.

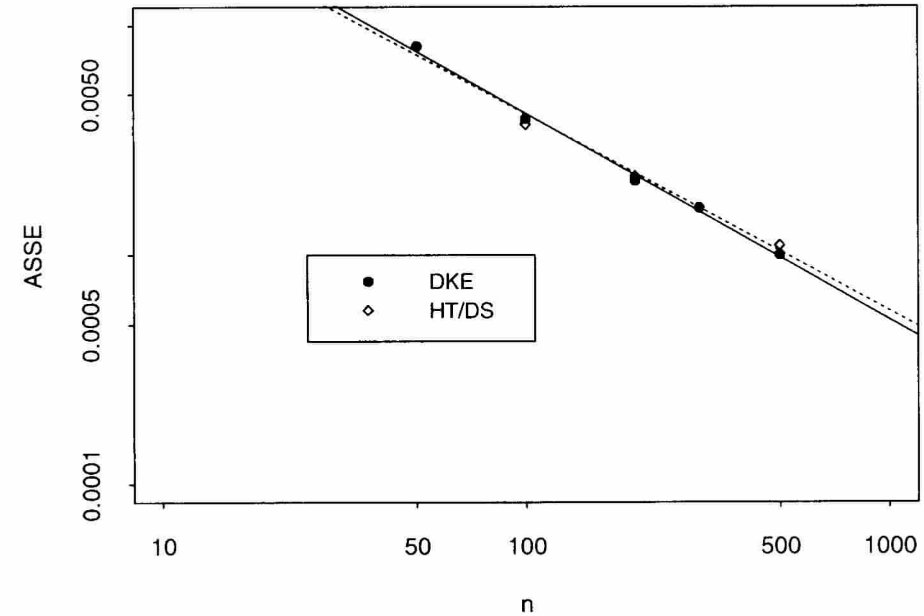

Figure 6. Log-Log plot of ASSE with sample size n, of samples generated from 0.7* Geometric ($\pi = 0.2$) + 0.3* Geometric ($\pi = 0.9$) along with the fitted lines.

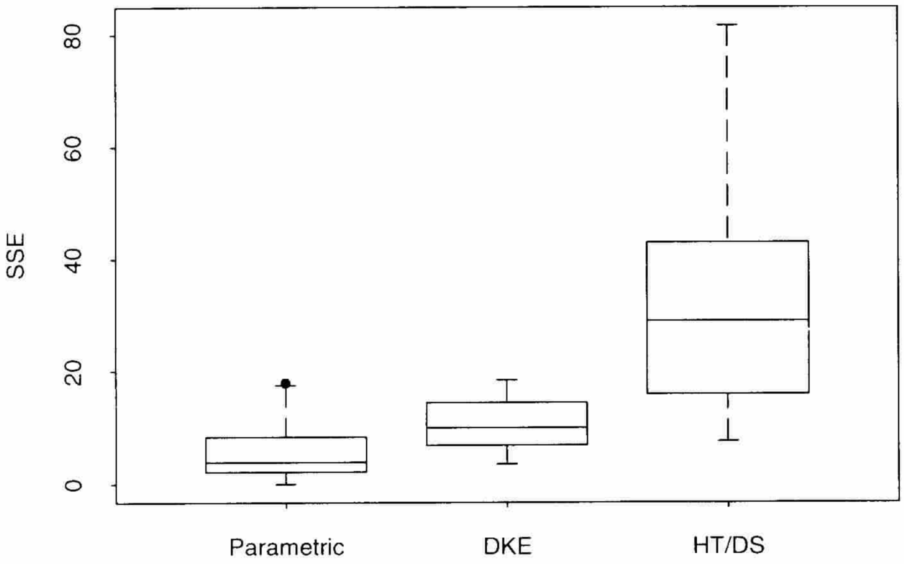

Figure 7. Boxplots of SSE$_j$ from HT/DS, DKE and fitted Parametric distribution, of samples generated from Geometric ($\pi = 0.2$) of sample size 50.

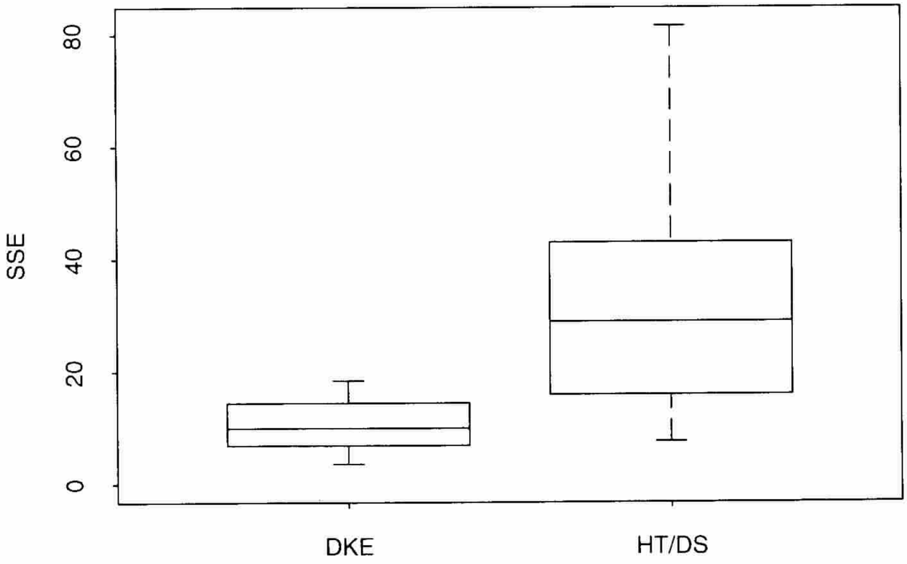

Figure 8. Boxplots of SSE$_j$ from HT/DS and DKE of samples generated from 0.7* Geometric ($\pi = 0.2$) + 0.3* Geometric ($\pi = 0.9$) of sample size 50.

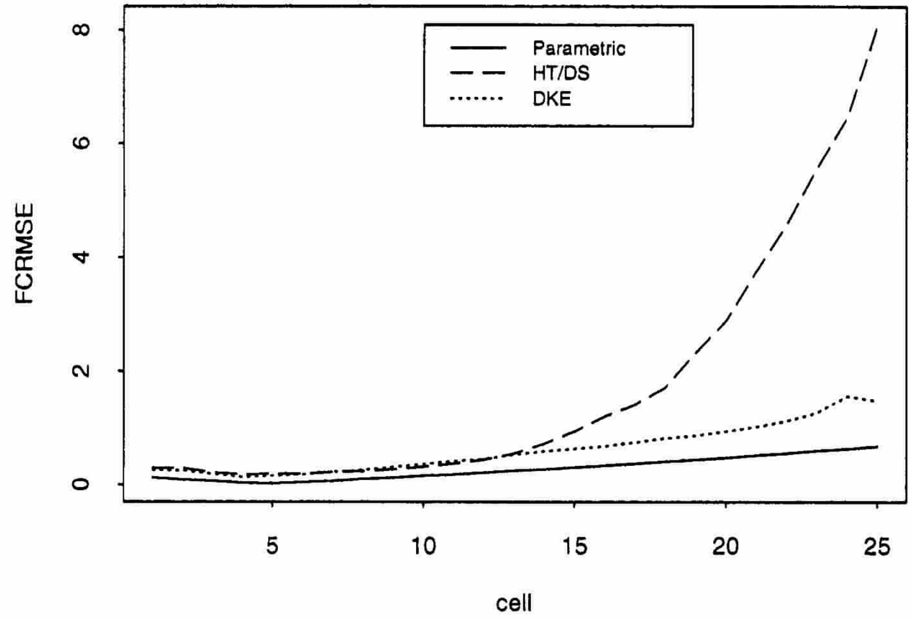

Figure 9(a). FCRMSE$_i$ from HT/DS, DKE and fitted Parametric distribution, of samples generated from Geometric ($\pi = 0.2$) of sample size 50.

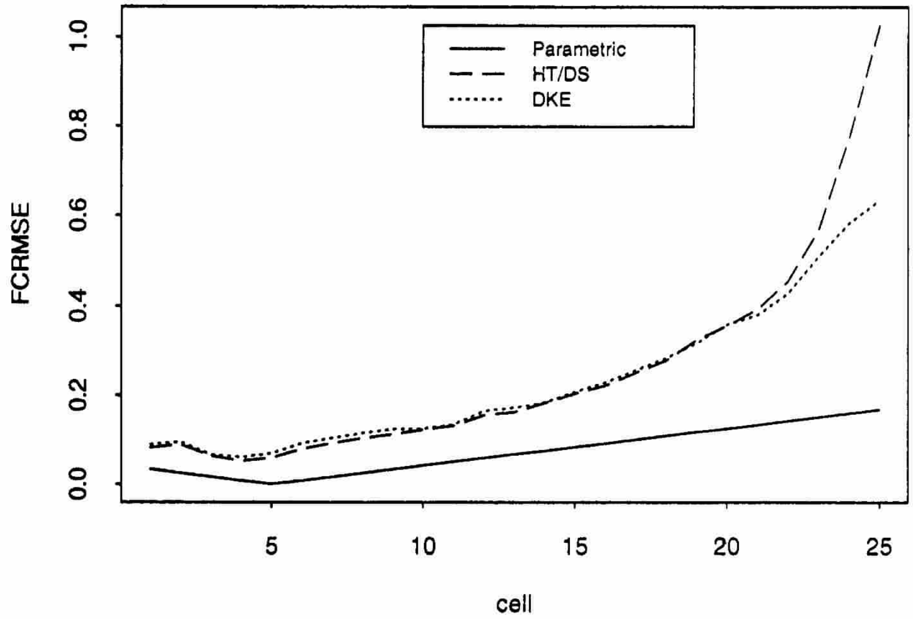

Figure 9(b). FCRMSE$_i$ from HT/DS, DKE and fitted Parametric distribution, of samples generated from Geometric ($\pi = 0.2$) of sample size 500.

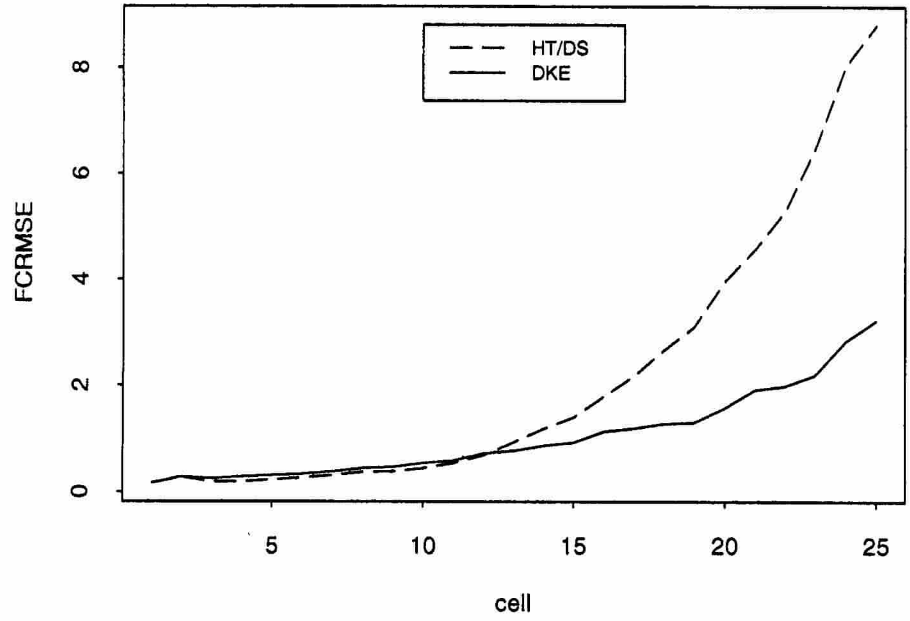

Figure 10(a). FCRMSE$_i$ from HT/DS and DKE, of samples generated from 0.7* Geometric ($\pi = 0.2$) + 0.3* Geometric ($\pi = 0.9$) of sample size 50.

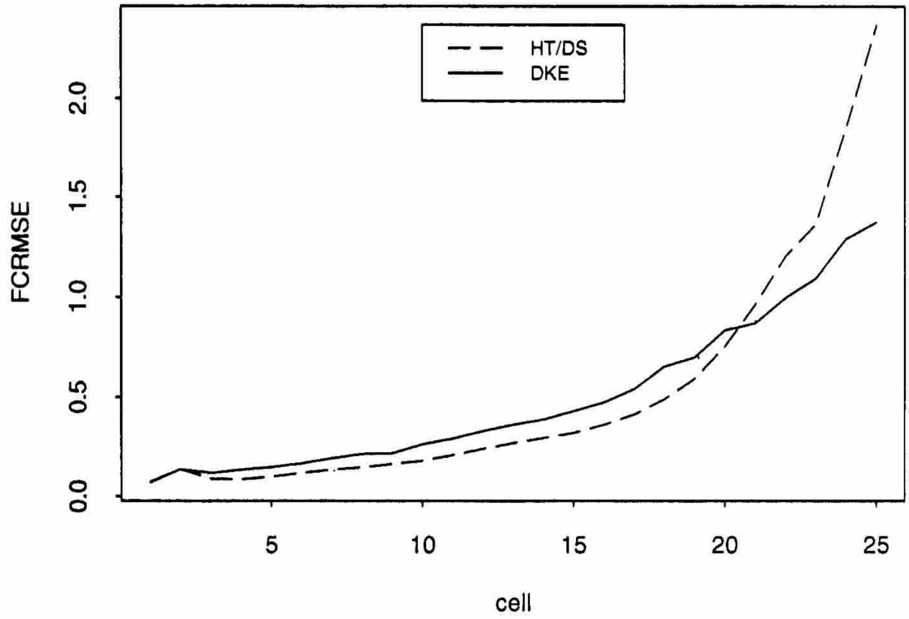

Figure 10(b). FCRMSE$_i$ from HT/DS and DKE, of samples generated from 0.7* Geometric ($\pi = 0.2$) + 0.3* Geometric ($\pi = 0.9$) of sample size 500.

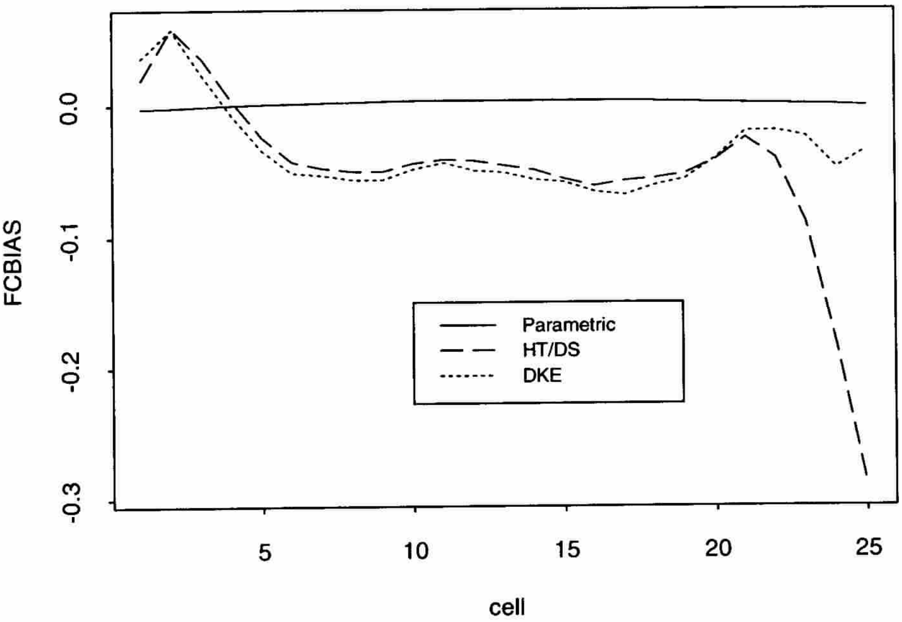

Figure 11. FCBIAS$_i$ from HT/DS and DKE, of samples generated from Geometric ($\pi = 0.2$) of sample size 500.

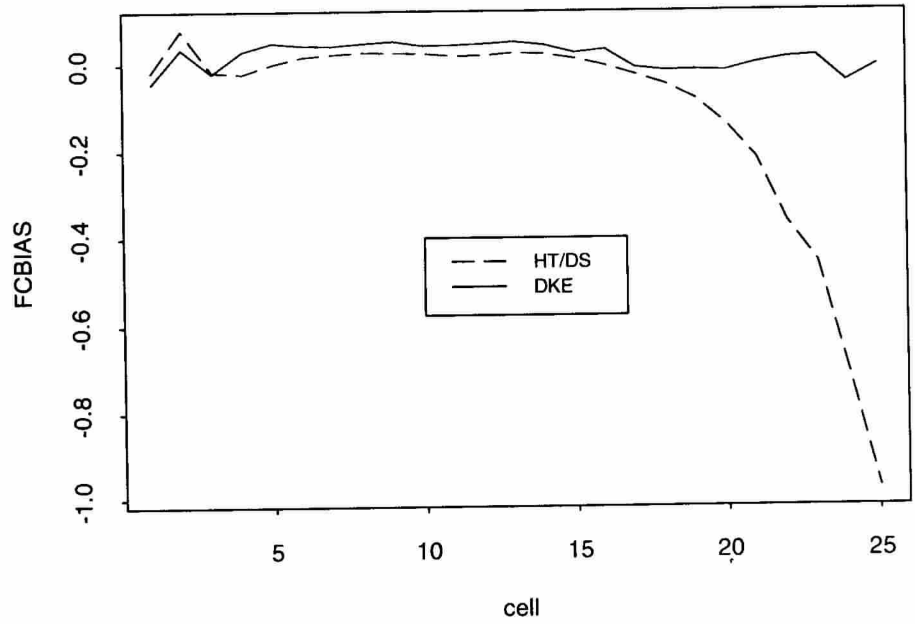

Figure 12. FCBIAS$_i$ from HT/DS, and DKE of samples generated from 0.7* Geometric ($\pi = 0.2$) + 0.3* Geometric ($\pi = 0.9$) of sample size 500.

The MSE expression of the estimate $\hat{p}_i$ as given by Wang and Van Ryzin [17] is,

$$E\left[\sum_{i=1}^{k_{\max}} \{\hat{p}_i - p_i\}^2\right] = \sum_{i=1}^{k_{\max}} \sum_{j=1}^{k_{\max}} W^2(i,j,h)p_j \Big/ n - \sum_{i=1}^{k_{\max}} \left\{ \sum_{j=1}^{k_{\max}} W^2(i,j,h)p_j \right\}^2 \Big/ n + \sum_{i=1}^{k_{\max}} \left\{ \sum_{j=1}^{k_{\max}} W(i,j,h)p_j - p_i \right\}^2 \tag{17}$$

where p_i is the true p.m.f, $W(i,j,h)$ is the weight function, h is the bandwidth and n is the sample size. For the two populations considered viz. G1 and G2 we know the true p.m.f Substituting this for p_i in the above equation, the optimal bandwidth can be determined for various sample sizes. These bandwidth values are then compared with the corresponding average bandwidths obtained from the simulations. These along with the coefficient of variance of bandwidth C_V are summarized in Table 2. It can be observed that C_V is smaller for DKE for all the sample sizes for G1 and G2. Note that DKE smooths the Geometric distribution data (G1) more than HT/DS, and smooths the mixture data (G2) less than HT/DS. Also the average bandwidths from DKE are close to the MSE optimal bandwidths. This suggests that the bandwidth from DKE is more stable than from HT/DS.

The behaviour of HT/DS in these simulations is interesting. There is a tendency to undersmooth relative to the optimal bandwidth. As a result the boundary bias decreases with n, while the tail bias may be high. The higher coefficient of variance of the HT/DS bandwidth suggests a higher degree of adaptation to sample attributes. However, this fails to consistently provide a lower bias on MSE than DKE.

The need to choose a bandwidth in the boundary region that is different from the interior has been recognized by several researchers (e.g., Müller [9]). Generally variation in h across the range of the data, and especially in the tails is needed. The selection of a "local" bandwidth considering boundary kernels and tail regions remains an area of research.

Table 2. Bandwidth statistics.

	Coefficient of Variation		*Average Bandwidth*		*Optimal Bandwidth from MSE Criteria*	
	DKE	HT/DS	DKE	HT/DS	DKE	HT/DS
Sample from G1						
$n=50$	0.349	0.442	6.73	5.48	7.00	8.06
$n=100$	0.305	0.401	6.13	4.97	6.00	8.06
$n=200$	0.361	0.316	4.96	4.36	5.00	7.14
$n=300$	0.290	0.314	4.51	4.21	4.00	6.25
$n=500$	0.275	0.341	4.00	3.47	4.00	5.56
Sample from G2						
$n=50$	0.309	0.291	2.844	3.067	3.00	4.10
$n=100$	0.210	0.220	2.280	2.931	2.00	4.03
$n=200$	0.007	0.213	2.020	2.902	2.00	4.03
$n=300$	0.000	0.212	2.000	2.912	2.00	4.03
$n=500$	0.000	0.214	2.000	2.844	2.00	4.03

4. OTHER POSSIBLE ESTIMATORS

Müller [9] shows how one can develop minimum variance kernels and kernels belonging to different smoothness classes for continuous variates. Extensions of these ideas to the discrete case is also feasible. Here we outline two such extensions.

A discrete, minimum variance (DMV), second order kernel can be developed as the solution to:

$$\text{Minimize} \sum_{j=q}^{i+h} w_j^2 \tag{18}$$

Subject to:

$$w_q = w_{i+h} = 0 \tag{19}$$

$$\sum_{j=q}^{i+h} w_j = 1 \tag{20}$$

$$\sum_{j=q}^{i+h} t_j w_j = 0 \tag{21}$$

where $t_j = (i-j)/h$, i, j, h are integers and $q = \max(i-h, 1)$, recognizes whether we are in the boundary region or the interior.

A smooth, disrete (DSμ) kernel of smoothness μ can be defined by solving the problem: Minimize $\sum_{j=q}^{i+h-\mu}(w_{j+\mu} - w_j)^2$, subject to the conditions (19) through (21) above. Solutions to the two problems defined above can be readily obtained by defining the associated Lagrangian problems and solving them for the weights w_j that define the kernel sequence over the appropriate span of integers.

The weight sequences resulting for DMV and DS1 ($\mu = 1$) for selected values of h, and i are compared with the DQ and HT/DS weight sequences in Table 3. In the interior, the HT/DS, DQ and DS1 weight sequences coincide. This is to be expected since they all converge to the quadratic kernel. The DMV sequence degenerates to uniform weights as expected. An examination of the weight sequences in the boundary region shows that the DQ sequences stay closer to the DS1 sequences than the HT/DS ones. Thus if a computationally fast approximation to the DS1 sequences was desired in the boundary region, DQ would be preferred. Note that the DMV sequences in the boundary region are still generally closer to the DS1 than the HT/DS.

An interesting aspect of the HT/DS sequence is the adaptation of the weight sequence as h varies between two integers. We observe that the weight sequences at the intermediate h value are not strictly in between the weight sequences at the end points. While this may lead to a high degree of adaptability of the HT/DS procedure, it makes it rather difficult to assess its impact on the estimation procedure. The high coefficient of variation of the bandwidth selected by HT/DS may be related to the nature of the resulting weight sequence.

Table 3. Comparison of weight sequences.

	$h = 2$	$h = 2.5$	$h = 3$
Interior			
DQ	0,.3,.4,.3,0	–	0,.14.23,.26,.23,.14,0
HT/DS	0,.3,.4,.3,0	0,.11,.25,.29,.25,.11,0	0,.14.23,. 26,.23,.14,0
DMV	0,.33,.33,.33,0		0,.2,.2,.2,.2,.2,.0
DS1	0,.28,.44,.28,0		0,.14.23,. 26,.23,.14,0
Boundary			
$i = 1$			
DQ	1,0,0	–	.75,.5,-.25,0
HT/DS	0,1,0	0,1.7,-7,0	0,.5,-.25,0
$i = 2$			
DQ	0,1,0,0	–	0,.75,.5,-.25,0
HT/DS	0,.63,.37,0	0,.62,.45,-.07,0	0,.5,.4,.1,0
DMV	0,1,0,0		0,.83,.33,-.16,0
DS1	0,1,0,0		0,.8,.4,-.2,0
$i = 3$			
DQ		–	0,.3,.4,.3,0,0
HT/DS		0,.28,.35,.28,.08,0	0,.28,.32,.28,.12,0
DMV			0,.4,.3,.2,.1,0
DS1			0,.34,.37,.23,.06,0

Notes: i is the point of estimate, on which the kernel is placed, h is the bandwidth. DQ, DMV and DS1 do not admit non integer bandwidths. The HT/DS weights correspond to a quadratic kernel, and admits non-interger h

The boundary kernels developed by Dong and Simonoff [13] do not correspond to the ones presented by Müller [9] for the continuous case. It may be interesting to try the Müller [9] boundary kernels, possible with a floating boundary value, directly with the HT procedure.

Computational considerations have restricted our Monte Carlo investigations thus far to DQ and HT/DS. The relative utility of DMV and DS may be investigated subsequently. Except in the boundary region, our limited investigations show that differences between the different kernels may not be large. Consequently, kernels that are easier to compute are expedient. In this respect the DQ kernels are to be preferred.

5. SUMMARY AND CONCLUSIONS

The estimator presented here was motivated by practical considerations. We offer this work in the hope that it will stimulate interest and theoretical development. We show that the discrete kernel procedure advocated can give results comparable to those from the HT/DS procedure. Computational advantages of the DKE procedure and the similarity of its properties to kernel sequences based on smoothness criteria were demonstrated. The relative stability of the bandwidth selection procedure and the DQ weight sequence also recommend it as an alternative to the HT/DS method.

We present only one special case (a quadratic kernel in the interior and in the boundary region). Clearly other similar higher order kernels can be derived. However, as it is typical in the kernel smoothing literature, bandwidth selection is likely to be a more tenuous issue than kernel specification. The LSCV choice of h appears to perform quite satisfactorily for the test cases. Extensions to the multivariate case are being investigated.

Acknowledgements

Partial support of this work by the U.S Forest Service under contract notes, INT-915550-RJVA and INT-92660-RJVA, Amend #1 is acknowledged. The principal investigator of the project is D. S. Bowles. We are grateful to J. Simonoff and J. Dong for discussions, provision of relevant manuscripts and computer code for p.m.f estimation using HT/DS estimator.

References

1. J. Aitchison and C. G. Aitken (1976). Multivariate binary discrimination by the kernel method, *Biometrika*, **63**, 413–420.
2. Y. M. Bishop, S. E. Fienberg and P. W. Holland (1975). *Discrete multivariate analysis: Theory and Practice*, MIT Press, Cambridge, Mass.
3. J. Dong and J. S. Simonoff (1994). The construction and properties of boundary kernels for sparse multinomials, *Journal of Computational and Graphical Statistics*, **3** (1), 57–66.
4. V. A. Epanechnikov (1969). Nonparametric estimations of a multivariate probability density, *Theor. Probab. Appl.*, **14**, 153–158.
5. S. E. Fienberg and P. W. Holland (1973). Simultaneous estimation of multinomial cell probabilities, *J. Amer. Statist. Assoc.*, **68**, 683–691.
6. I. J. Good (1965). *The estimation of probabilities*, MIT Press, Cambridge, Mass.
7. I. J. Good (1967). A Bayesian significance test for multinomial distributions (with discussion), *J. Roy. Statist. Soc., Ser. B.*, **29**, 399–431.
8. P. Hall, and D. M. Titterington (1989). On smoothing sparse multinomial data, *Australian Journal of Statistics*, **29**, 19–37.
9. H. G. Müller (1991). Smooth optimum kernel estimators near endpoints, *Biometrika*, **78** (3), 521–530.
10. B. Rajagopalan, U. Lall and D. G. Tarboton (1993). Simulation of daily precipitation from a nonparametric renewal model. Working Paper WP-93-HWR-UL/003. In Utah Water Research Laboratory, Utah State University, Logan, UT.
11. D. W, Scott (1992). *Multivariate density estimation*, Wiley series in probability and mathematical statistics, John Wiley and Sons, New York.
12. J. S. Simonoff (1983). A penalty function approach to smoothing large sparse contingency tables, *Ann. Statist.*, **11**, 208–218.
13. M. Stone (1974). Cross-validation and multinomial prediction, *Biometrika*, **61**, 509–515.
14. D. M. Titterington (1976). Updating a diagnostic system using unconfirmed cases, *Appl. Statist.*, **25**, 238–247.
15. D. M. Titterington (1980). A comparative study of kernel-based density estimates for categorical data, *Technometrics*, **22**, 259–268.
16. D. M. Titterington and A. W. Bowman (1985). A comparative study of smoothing procedures for ordered categorical data, *J. Statist. Comput. Simul.*, **21**, 291–312.
17. M C. Wang and J. Van Ryzin (1981). A class of smooth estimators for discrete distributions, *Biometrika*, **68** (1), 301–309.