Journal of the Korean Statistical Society (1998), 27: 4, pp 515-529

A Local Linear Kernel Estimator for Sparse Multinomial Data†

Jangsun Baek[1]

ABSTRACT

Burman (1987) and Hall and Titterington (1987) studied kernel smoothing for sparse multinomial data in detail. Both of their estimators for cell probabilities are sparse asymptotic consistent under some restrictive conditions on the true cell probabilities. Dong and Simonoff (1994) adopted boundary kernels to relieve the restrictive conditions. We propose a local linear kernel estimator which is popular in nonparametric regression to estimate cell probabilities. No boundary adjustment is necessary for this estimator since it adapts automatically to estimation at the boundaries. It is shown that our estimator attains the optimal rate of convergence in mean sum of squared error under sparseness. Some simulation results and a real data application are presented to see the performance of the estimator.

Keywords: Sparse multinomial; Kernel estimators; Multinomial smoothing; Local linear regression

1. INTRODUCTION

We often observe that there are many cells with counts 0 and 1 for multinomial data. This happens when the number of cells m is large compared to the total number of observations. Let n_i be the observed frequency in cell i, $i = 1, 2, \cdots, m$ generated from a multinomial probability function with underlying probability vector $\mathbf{p}=(p_1, p_2, \cdots, p_m)^T$, and let $\sum_{i=1}^{m} n_i = N$ be the total number of observations. When N/m is small and there are many cells with small or zero counts, the multinomial data is called to be sparse.

If we are interested in the estimation of p_i, the first candidate among the estimators would be the frequency estimator $p_i{}^* = n_i/N$. $p_i{}^*$ is consistent and efficient only when $N \to \infty$ with m fixed. It is not even consistent under sparse

†The author wishes to acknowledge the financial support of the Korea Research Foundation(1997-003-D00051) made in the program year of 1997.

[1]Department of Statistics, Chonnam National University, Kwangju, 500-757, Korea.

asymptotics where N/m remains small constant as both N and m become infinite. Another type of estimator is obtained by convex combination of the frequency estimator ${p_i}^*$ and prior guess for p_i. Bishop, Fienberg and Holland (1975) showed that this type of estimator is often better than ${p_i}^*$ under squared error loss. It is still not sparse asymptotically consistent. By imposing a smoothness constraint on the cell probabilities, Simonoff (1983) considered an estimator based on a maximum penalized likelihood criterion. Burman (1987), Hall and Titterington (1987) - here after referred to as HT - proposed kernel-based estimators for sparse multinomial data, and Grund and Hall (1993) studied a kernel-type estimator for high-dimensional sparse binary data. HT and Burman (1987) assumed that the true cell probability is determined by a bounded probability density function $f(x)$, $0 < x < 1$, with the following restrictions, respectively:

HT: $f(0) = f^{(1)}(0) = f^{(2)}(0) = f(1) = f^{(1)}(1) = f^{(2)}(1) = 0$,
Burman: $f^{(1)}(0) = f^{(1)}(1) = 0$,

where $f^{(i)}$ is the ith derivative of f. They showed that their second order kernel estimator has the mean sum of squared error ($MSSE$) of $O(m^{-1}N^{-4/5})$, and is sparse asymptotically consistent. Those restrictions are necessary to reduce the boundary bias, but too restrictive for multinomials with ordered categories. Dong and Simonoff (1994) used the boundary kernel that does not require the above conditions. It is well known that the boundary kernel methods are not as simple and as efficient as the automatic boundary correction when using local linear kernel estimator proposed by Fan (1992). Thus we construct a local linear kernel estimator to estimate cell probabilities and investigate its sparse asymptotic consistency in Section 2. Section 3 contains some simulation results and a real data application. A Proof is given in Appendix.

2. A LOCAL LINEAR ESTIMATOR FOR CELL PROBABILITY

Let $Z_1, Z_2, \cdots, Z_N$ be a random sample from a distribution whose cell probability vector is $\mathbf{p} = (p_1, \cdots, p_m)$. HT defined their kernel cell probability estimator as $\tilde{p}_i = N^{-1}\lambda^{-1}\sum_{j=1}^{N} K_\lambda^*((i - Z_j)/\lambda)$, where $K_\lambda^*(\cdot/\lambda) = s(\lambda)K(\cdot/\lambda)$, $s(\lambda) = \{\lambda^{-1}\sum_j K(j/\lambda)\}^{-1}$, K is a kernel and λ is the bandwidth, $i = 1, \cdots, m$. Dong and Simonoff (1994) noticed that HT's estimator is m/N times the Priestley-Chao

kernel regression estimator on the points (j, n_j), $j = 1, \cdots, m$. That is,

$$\tilde{p}_i = \frac{m}{N} \cdot \{\frac{1}{m\lambda} \sum_{j=1}^{m} n_j K_\lambda^*(\frac{i-j}{\lambda})\}.$$

We can, however, reexpress the HT's estimator differently by setting $\lambda^* = \lambda/m$ as follows:

$$\begin{aligned} \tilde{p}_i &= \frac{1}{\lambda} \sum_{j=1}^{m} (\frac{n_j}{N}) K_\lambda^*(\frac{i-j}{\lambda}) \\ &= \frac{1}{\lambda^*} \sum_{j=1}^{m} (\frac{n_j}{N}) K_{\lambda^*}^*(\frac{i/m - j/m}{\lambda^*}) \\ &= \sum_{j=1}^{m} (\frac{n_j}{N}) \frac{K(\frac{i/m-j/m}{\lambda^*})}{\sum K(\frac{i/m-j/m}{\lambda^*})} \end{aligned} \tag{2.1}$$

Therefore, $\tilde{p}_i$ is the Nadaraya-Watson kernel regression estimator on the points $(j/m, n_j/N)$, $j = 1, \cdots, m$, with the new bandwidth λ^*.

The kernel regression estimator in (2.1) can be replaced with other estimator if it has better properties. The local polynomial kernel estimator has favorable asymptotic properties and boundary behaviour compared with other traditional kernel smoother. In particular, it is design adaptive, and needs no boundary adjustment. The local linear kernel estimator is of particular importance and simplicity among the local polynomial kernel estimators. We propose a local linear kernel estimator to estimate the cell probabilities instead of the Nadaraya-Watson type estimator in (2.1).

We assume that for a bounded probability density function f with its support [0,1] and two bounded continuous derivatives,

$$p_i = \int_{x_i - \frac{1}{2m}}^{x_i + \frac{1}{2m}} f(x)dx \tag{2.2}$$

That is, the ith cell probability p_i equals the chance that a random variable with density f takes a value within the interval of width $1/m$ centered on x_i. Note that this construction implies $sup_i p_i \leq c/m$ for some constant c. Refer to Burman (1987) for other formulations of p_i.

Now let $x_j = j/m$, and $y_j = p_j{}^* = n_j/N$, $j = 1, \cdots, m$. Then we can define the local linear kernel estimator to estimate p_i as follows:

$$\hat{p}_i = \mathbf{e_1}^T(\mathbf{X}_{x_i}{}^T\mathbf{W}_{x_i}\mathbf{X}_{x_i})^{-1}\mathbf{X}_{x_i}{}^T\mathbf{W}_{x_i}\mathbf{y}, \tag{2.3}$$

where $\mathbf{e_1} = (1,0)^T$, $\mathbf{y} = (y_1, \cdots, y_m)^T$,

$$\mathbf{X}_{x_i} = \begin{bmatrix} 1 & x_1 - x_i \\ \vdots & \vdots \\ 1 & x_m - x_i \end{bmatrix},$$

$$\mathbf{W}_{x_i} = diag\{K_\lambda(x_1 - x_i), \cdots, K_\lambda(x_m - x_i)\}, \quad K_\lambda(u) = \lambda^{-1}K(u/\lambda).$$

$K(\cdot)$ is the kernel which is symmetric about zero, and is supported on $[-1, 1]$.

We will first investigate the bias and the variance for the estimator of (2.3) in the following theorem. The proof is given in Appendix.

Theorem 2.1 Let f have two bounded continuous derivatives on $[0, 1]$, and let the kernel K with bandwidth λ be symmetric about zero and supported on $[-1, 1]$. Then for the point x_i, $\lambda < x_i < 1 - \lambda$ at which the estimation of p_i is taking place,

$$E(\hat{p}_i) - p_i = \frac{\lambda^2}{2m} f^{(2)}(x_i)\mu_2(K) + o(\frac{\lambda^2}{m}) + O(\frac{1}{m^2}),$$

$$Var(\hat{p}_i) = \frac{R(K)p_i}{N\lambda m} + o(\frac{1}{N\lambda m}) + O(\frac{1}{Nm}),$$

where

$$\mu_2(K) = \int_{-1}^{1} u^2 K(u)du, \quad R(K) = \int_{-1}^{1} K^2(u)du.$$

The mean squared error (MSE) of the estimator $\hat{p}_i$ at the interior points is obtained from the above theorem:

$$\begin{aligned} E(\hat{p}_i - p_i)^2 &= bias^2 + Variance \\ &\sim \frac{\lambda^4}{4m^2}\{f^{(2)}(x_i)\}^2\mu_2^2(K) + \frac{R(K)p_i}{N\lambda m}. \end{aligned}$$

Now we examine the bias and variance of the estimator at the boundary. When $x_j's$ are random design points, with density g, the MSE of the local polynomial regression estimator at the boundary are already well-known (Fan (1992), Wand and Jones (1995 ; pp 126-130).)

Since our design points $x_j's$ are considered as $x_j = G^{-1}(i/m).+o(1/m)$, where G is the cdf of uniform density g, we have essentially the same result for MSE as in uniform random design case. Suppose that $x_i = \alpha\lambda$ where $0 \leq \alpha < 1$; that is, x_i is within $\alpha\lambda$ of the left boundary of [0, 1]. Let

$$\mu_{l,\alpha}(K) = \int_{-\alpha}^{1} z^l K(z)dz,$$

$$K^*(u,\alpha) = \{|M(u,\alpha)|/|N(\alpha)|\}\ K(u)\ ,\ -\alpha < u < 1\ ,$$

where $N(\alpha)$ is the 2×2 matrix having (i,j) entry equal to $\mu_{i+j-2,\alpha}(K)$ and $M(u,\alpha)$ is the same as $N(\alpha)$, but with the first column replaced by $(1,u)^T$. Then it can be shown that at the left boundary point x_j,

$$E(\hat{p}_i - p_i)^2 \sim \frac{\lambda^4}{4m^2}\{f^{(2)}(x_i)\}^2\{\int_{-\alpha}^{1} u^2 K^*(u,\alpha)du\}^2 + \frac{p_i}{N\lambda m}\left[\int_{-\alpha}^{1}\{K^*(u,\alpha)\}^2 du\right]$$

The result of the right boundary is similar if treated analogously.

Combining the results of MSE in interior and boundary region, we obtain the $MSSE$ of the estimator. Let I be the interior region, i.e., $I = [\lambda, 1-\lambda]$. Then for $\hat{\mathbf{p}} = (\hat{p}_1, \cdots, \hat{p}_m)$,

$$\begin{aligned}
MSSE(\hat{\mathbf{p}}) &= \sum_{i=1}^{m} E(\hat{p}_i - p_i)^2 \\
&= \sum_{x_i \notin I} E(\hat{p}_i - p_i)^2 + \sum_{x_i \in I} E(\hat{p}_i - p_i)^2 \\
&\sim \sum_{x_i \notin I}\{O(\frac{\lambda^4}{m^2}) + O(\frac{1}{N\lambda m})\} \\
&\quad + \sum_{x_i \in I}\left[\frac{\lambda^4}{4m^2}\{f^{(2)}(x_i)\}^2\mu_2^2(K) + \frac{R(K)p_i}{N\lambda m}\right] \\
&\sim O(\lambda m)\{O(\frac{\lambda^4}{m^2}) + O(\frac{1}{N\lambda m})\} \\
&\quad + \{\frac{\lambda^4\mu_2^2(K)}{4m}\}\frac{1}{m}\sum_{x_i \in I}\{f^{(2)}(x_i)\}^2 + \{\frac{R(K)}{N\lambda m}\}\sum_{x_i \in I} p_i
\end{aligned}$$

$$\begin{aligned} &\sim O(\frac{\lambda^5}{m}) + O(\frac{1}{N}) + \frac{\lambda^4 \mu_2^2(K)}{4m} \int_{\lambda}^{1-\lambda} \{f^{(2)}(u)\}^2 du + \frac{R(K)}{N\lambda m} \sum_{x_i \in I} p_i \\ &\sim \frac{\lambda^4}{4m} \mu_2^2(K) \int_0^1 \{f^{(2)}(u)\}^2 du + \frac{R(K)}{N\lambda m} \sum_{i=1}^{m} p_i \\ &= \frac{\lambda^4}{4m} \mu_2^2(K) \int_0^1 \{f^{(2)}(u)\}^2 du + \frac{R(K)}{N\lambda m}. \end{aligned} \tag{2.4}$$

The asymptotically optimal bandwidth minimizing $MSSE$ of (2.4) is

$$\lambda_{opt} \sim \left[\frac{R(K)}{N\mu_2^2(K) \int \{f^{(2)}(x)\}^2 dx} \right]^{1/5},$$

and

$$MSSE(\hat{\mathbf{p}} : \lambda_{opt}) \sim \frac{5\{R(K)\}^{4/5}}{4mN^{4/5}} \left[\mu_2^2(K) \int \{f^{(2)}(x)\}^2 dx \right]^{1/5}.$$

Therefore $MSSE$ of the local linear kernel estimator is $O(m^{-1}N^{-4/5})$, and the estimator is sparse asymptotically consistent.

Remark 2.1: The cell frequency estimator $\mathbf{p}^* = (p_1^*, \cdots, p_m^*)$ has its $MSSE(\mathbf{p}^*)$ $(1/N)(1 + o(1))$. We can compare $MSSE(\hat{\mathbf{p}}; \lambda_{opt})$ with $MSSE(\mathbf{p}^*)$;

$$\frac{MSSE(\hat{\mathbf{p}} : \lambda_{opt})}{MSSE(\mathbf{p}^*)} \sim (\frac{5}{4})(\frac{N^{1/5}}{m})\{R(K)\}^{4/5} \left[\mu_2^2(K) \int \{f^{(2)}(x)\}^2 dx \right]^{1/5}.$$

If $N^{1/5}/m \to \gamma$, $0 < \gamma < \infty$, then $MSSE(\hat{\mathbf{p}}; \lambda_{opt})/MSSE(\mathbf{p}^*) \sim O(1)$, and $\hat{\mathbf{p}}$ attains the same $MSSE$ convergence rate as that of $\mathbf{p}^*$, $O(N^{-1})$. If $N^{1/5}/m \to 0$, then $MSSE(\hat{\mathbf{p}}; \lambda_{opt})/MSSE(\mathbf{p}^*) \sim o(1)$, and $\hat{\mathbf{p}}$ gives smaller $MSSE$ than $\mathbf{p}^*$ does. Thus it is clear that there is an advantage in smoothing.

3. SIMULATION AND REAL DATA APPLICATION

In this section we compare the performance of the nonparametric kernel estimators under the similar situations as Dong and Simonoff (1994) did. We consider Beta (3,3) and Beta (.6,.6) densities as the underlying probability density function f generating cell probability vector $\mathbf{p}$, respectively. Multinomials were then generated from $\mathbf{p}$ with 100 simulation runs. The tested estimators are the HT's normalized nonboundary-corrected kernel estimator (KNBC), the boundary-corrected kernel estimator (KBC) suggested by Dong and Simonoff

(1994), and the local linear kernel estimator (LL). The bandwidth of the estimators was chosen to minimize its own sum of squared error (SSE), respectively, and the Epanechnikov kernel was used. Values for $m = 10, 20, 50$, and 100 were examined, with N chosen such that $N/m = 1, 2$, and 5; it is quite sparse when $N/m = 1$, but not very sparse when $N/m = 5$. Comparisons were made on the basis of $N \times SSE$.

Figure 3.1 and Figure 3.3 give the boxplots of the difference in $N \times SSE$ between that of KNBC and that of LL, and Figure 3.2 and Figure 3.4 do the same ones between KBC and LL for 100 simulation runs for (a) $m = 10$, (b) $m = 20$, (c) $m = 50$, and (d) $m = 100$. The shaded area on each boxplot is the 95% confidence interval for the median.

Figure 3.1 and Figure 3.2 refers to a Beta(3,3) underlying density, which satisfies HT's boundary condition. We see from Figure 3.1 that LL outperforms KNBC in all cases. Figure 3.2 says KBC and LL are comparable for $m = 10, 20$, but KBC is slightly better than LL for $m = 50, 100$. Note, however, the absolute value of the difference in $N \times SSE$ between KBC and LL is very small ($\ll 0.002$) for $m = 100$. Thus the difference in SSE for both estimators can be considered to be negligible.

Figure 3.3 and Figure 3.4 give the results for a Beta(.6, .6) underlying density, which does not satisfy either HT's or Burman's condition. LL outperforms KNBC except for the very sparse data with small number of cells ($N = 10, 20$ for $m = 10, N = 20$ for $m = 20$). On the other hand, LL clearly performs better than KBC in all cases. In addition LL does better as the multinomial becomes less sparse.

We apply the local linear kernel estimator to a real data which consist of 55 cell multinomial based on the time intervals between explosions in mines in Great Britain that killed more than 10 men in certain time period (Maguire, Pearson, and Wynn (1952), Simonoff (1983)). The data is quite sparse since the sample size is 109.

Figure 3.5 shows the local linear kernel estimates along with the KBC and the frequency estimates. The x axis $(0, 55)$ was rescaled to $(0, 1)$. The data-driven bandwidth was selected by the least squares cross-validation suggested by HT. The bandwidth chosen for LL and KBC was 0.0728 and 0.0632, respectively. LL and KBC estimates are quite similar, and both represent properly the underlying density which is roughly exponential, but of a bulge around the tenth cell, a thicker tail, and a slight rise at the upper extreme (Leonard (1978), Simonoff (1983), Lindsey (1992)), as noted by Dong and Simonoff (1994).

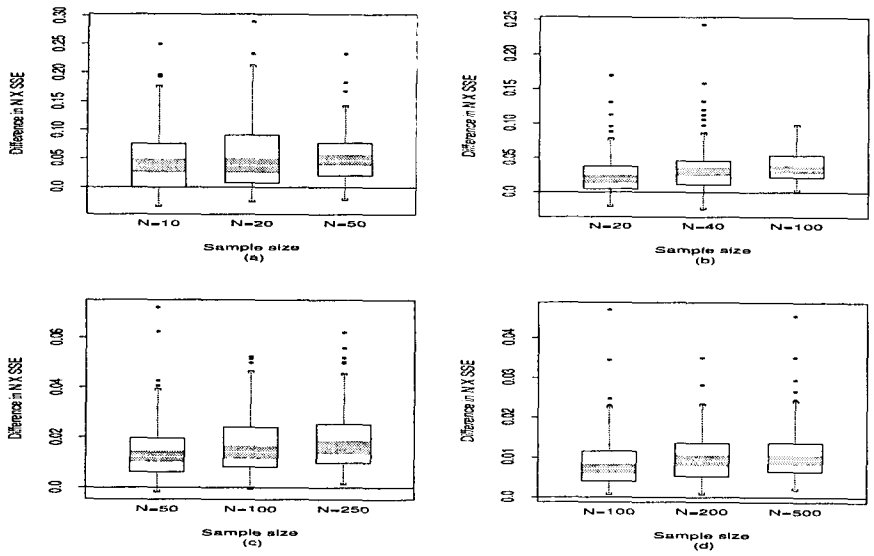

Figure 3.1: Simulation results for Beta(3,3) underlying probability vector. Boxplots represent the difference in $N \times SSE$ for KNBC and LL. (a) = 10 cells, (b) = 20 cells, (c) = 50 cells, and (d) = 100 cells.

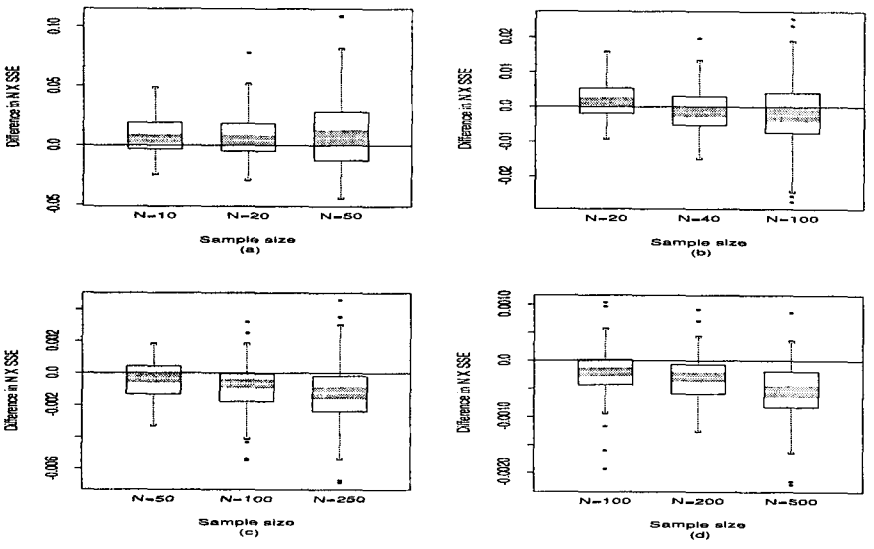

Figure 3.2: Simulation results for Beta(3,3) underlying probability vector. Boxplots represent the difference in $N \times SSE$ for KBC and LL. (a) = 10 cells, (b) = 20 cells, (c) = 50 cells, and (d) = 100 cells.

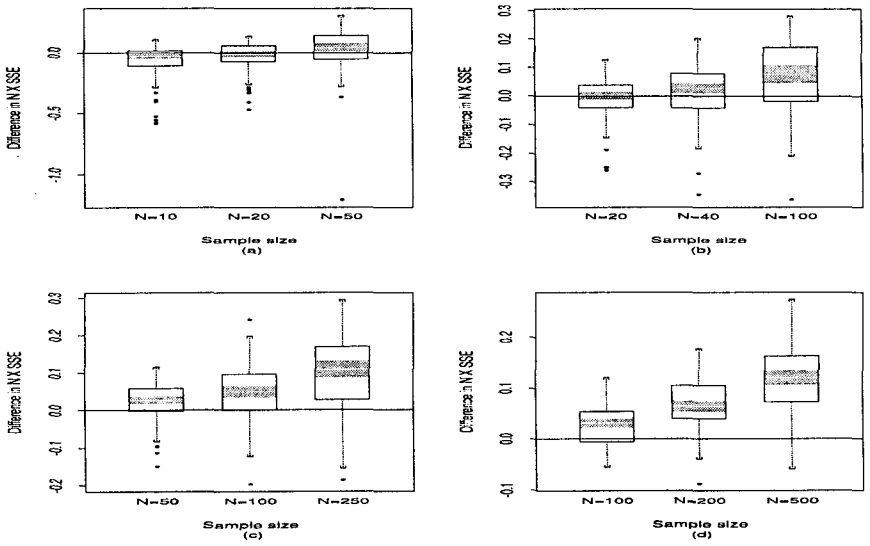

Figure 3.3: Simulation results for Beta(.6,.6) underlying probability vector. Boxplots represent the difference in $N \times SSE$ for KNBC and LL. (a) = 10 cells, (b) = 20 cells, (c) = 50 cells, and (d) = 100 cells.

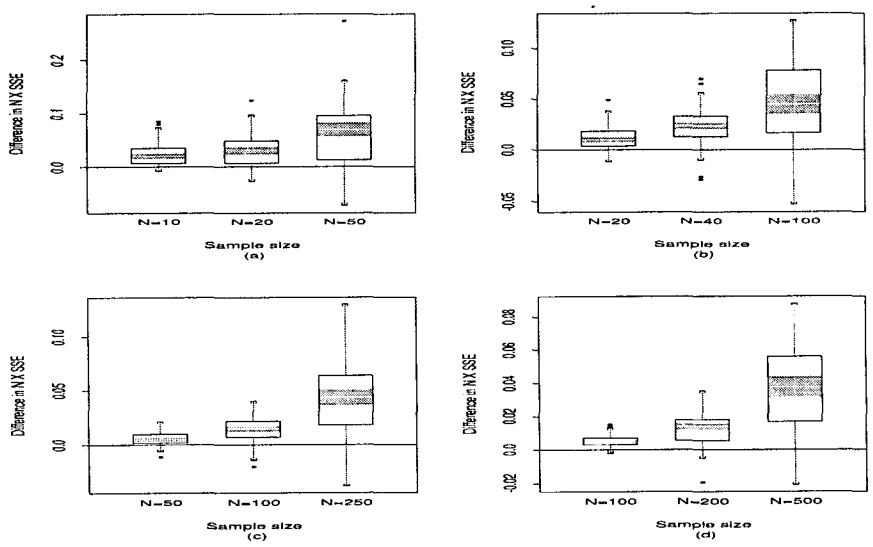

Figure 3.4: Simulation results for Beta(.6,.6) underlying probability vector. Box plots represent the difference in $N \times SSE$ for KBC and LL. (a) = 10 cells, (b) = 20 cells, (c) = 50 cells, and (d) = 100 cells.

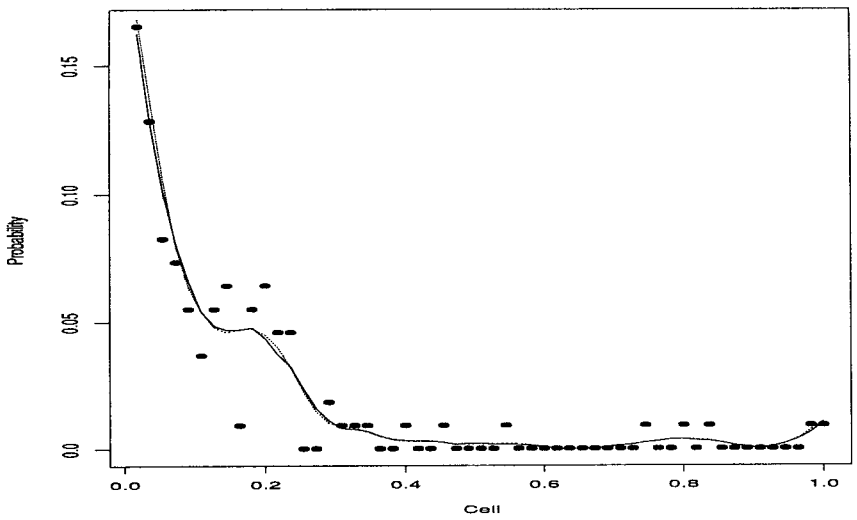

Figure 3.5: Cell probability estimates for mine explosions data; The local linear kernel estimates (solid line), the boundary-corrected kernel estimates (dotted line), and the frequency estimates (filled circles).

APPENDIX

Proof of theorem 2.1 First, we derive the bias of the estimator. Now $E(\mathbf{y}) = \mathbf{p}$, where $\mathbf{p} = (p_1, p_2, \cdots, p_m)^T$ since p_j^* is unbiased.

$$\begin{aligned} E(\hat{p}_i) &= \mathbf{e}_1^T(\mathbf{X}_{x_i}^T\mathbf{W}_{x_i}\mathbf{X}_{x_i})^{-1}\mathbf{X}_{x_i}^T\mathbf{W}_{x_i}E(\mathbf{y}) \\ &= \mathbf{e}_1^T(\mathbf{X}_{x_i}^T\mathbf{W}_{x_i}\mathbf{X}_{x_i})^{-1}\mathbf{X}_{x_i}^T\mathbf{W}_{x_i}\mathbf{p} \end{aligned}$$

Taking Taylor expansions of $f(x)$ in (2.2) around x_j gives

$$p_j = \frac{1}{m}f(x_j) + \frac{1}{24m^3}f^{(2)}(x_j) + o(m^{-3}).$$

Taking Taylor expansions of f and $f^{(2)}$ again at x_i, we obtain

$$\begin{aligned} p_j &= \frac{1}{m}f(x_i) + \frac{1}{24m^3}f^{(2)}(x_i) + \frac{(x_j - x_i)}{m}f^{(1)}(x_i) \\ &\quad + \frac{(x_j - x_i)^2}{2m}f^{(2)}(x_i) + o(m^{-3}). \end{aligned}$$

Thus

$$\mathbf{p} = \mathbf{X}_{x_i}\begin{bmatrix}\frac{1}{m}f(x_i)+\frac{1}{24m^3}f^{(2)}(x_i)\\ \frac{1}{m}f^{(1)}(x_i)\end{bmatrix}+\frac{1}{2m}f^{(2)}(x_i)\begin{bmatrix}(x_1-x_i)^2\\ \vdots\\ (x_m-x_i)^2\end{bmatrix}+\mathbf{o}(m^{-3}).$$

Now the first term in the expansion of $E(\hat{p}_i)$ is therefore,

$$\begin{aligned}&\mathbf{e}_1^T(\mathbf{X}_{x_i}^T\mathbf{W}_{x_i}\mathbf{X}_{x_i})^{-1}(\mathbf{X}_{x_i}^T\mathbf{W}_{x_i}\mathbf{X}_{x_i})\begin{bmatrix}\frac{1}{m}f(x_i)+\frac{1}{24m^3}f^{(2)}(x_i)\\ \frac{1}{m}f^{(1)}(x_i)\end{bmatrix}\\ =\ &\mathbf{e}_1^T\begin{bmatrix}\frac{1}{m}f(x_i)+\frac{1}{24m^3}f^{(2)}(x_i)\\ \frac{1}{m}f^{(1)}(x_i)\end{bmatrix}\\ =\ &\frac{1}{m}f(x_i)+\frac{1}{24m^3}f^{(2)}(x_i)\\ =\ &p_i+O(m^{-3})\end{aligned}$$

So the bias of $\hat{p}_i$ is

$$E(\hat{p}_i)-p_i = \{\frac{1}{2m}f^{(2)}(x_i)\}\mathbf{e}_1^T(\mathbf{X}_{x_i}^T\mathbf{W}_{x_i}\mathbf{X}_{x_i})^{-1}\mathbf{X}_{x_i}\mathbf{W}_{x_i}\begin{bmatrix}(x_1-x_i)^2\\ \vdots\\ (x_m-x_i)^2\end{bmatrix}+O(m^{-3})$$

Let $\hat{s}_l(x_i:\lambda)=m^{-1}\sum_{j=1}^m(x_j-x_i)^lK_\lambda(x_j-x_i)$. $l=0,1,2,\cdots$. Then we observe that

$$m^{-1}\mathbf{X}_{x_i}^T\mathbf{W}_{x_i}\mathbf{X}_{x_i}=\begin{bmatrix}\hat{s}_0(x_i:\lambda) & \hat{s}_1(x_i:\lambda)\\ \hat{s}_1(x_i:\lambda) & \hat{s}_2(x_i:\lambda)\end{bmatrix},$$

$$m^{-1}\mathbf{X}_{x_i}^T\mathbf{W}_{x_i}\begin{bmatrix}(x_1-x_i)^2\\ \vdots\\ (x_m-x_i)^2\end{bmatrix}=\begin{bmatrix}\hat{s}_2(x_i:\lambda)\\ \hat{s}_3(x_i:\lambda)\end{bmatrix}.$$

Since $\hat{s}_l(x_i : \lambda) = \lambda^l \int_{-1}^{1} u^l K(u)du + O(m^{-1})$,

$$m^{-1}\mathbf{X}_{x_i}^T \mathbf{W}_{x_i}\mathbf{X}_{x_i} = \begin{bmatrix} 1 + O(m^{-1}) & O(m^{-1}) \\ O(m^{-1}) & \lambda^2\mu_2(K) + O(m^{-1}) \end{bmatrix},$$

$$m^{-1}\mathbf{X}_{x_i}{}^T \mathbf{W}_{x_i} \begin{bmatrix} (x_1 - x_i)^2 \\ \vdots \\ (x_m - x_i)^2 \end{bmatrix} = \begin{bmatrix} \lambda^2\mu_2(K) + O(m^{-1}) \\ O(m^{-1}) \end{bmatrix},$$

where $\mu_l(K) = \int_{-1}^{1} z^l K(z)dz$.
Some simple matrix algebra leads to the expression

$$\begin{aligned} E(\hat{p}_i) - p_i &= \frac{1}{2m} f^{(2)}(x_i)\{\lambda^2\mu_2(K) + o(\lambda^2) + O(m^{-1})\} \\ &= \frac{\lambda^2}{2m} f^{(2)}(x_i)\mu_2(K) + o(\frac{\lambda^2}{m}) + O(\frac{1}{m^2}). \end{aligned}$$

Now we investigate the variance of the estimator. First, note the covariance of $\mathbf{y}$ is

$$Cov(y_i, y_j) = \begin{cases} p_i(1 - p_i)/N, & i = j, \\ -p_i p_j/N, & i \neq j. \end{cases}$$

Then the variance of $\hat{p}_i$ is

$$Var(\hat{p}_i) = (1/N)\mathbf{e}_1^T(\mathbf{X}_{x_i}^T\mathbf{W}_{x_i}\mathbf{X}_{x_i})^{-1}\mathbf{X}_{x_i}^T\mathbf{W}_{x_i}\mathbf{V}\mathbf{W}_{x_i}\mathbf{X}_{x_i}(\mathbf{X}_{x_i}^T\mathbf{W}_{x_i}\mathbf{X}_{x_i})^{-1}\mathbf{e_1},$$

where $\mathbf{V}$ is the N times covariance matrix of $\mathbf{y}$. The (i,i)th diagonal element of $\mathbf{V}$ is $p_i(1 - p_i)$, and (i,j)th off-diagonal element of $\mathbf{V}$ is $-p_i p_j$. Let $v_1(x_j) = p_j$ and $v_2(x_i, x_j) = -p_i p_j$, $i = 1, 2, \cdots, m$, $j = 1, 2, \cdots, m$. Then we can decompose $\mathbf{V}$ into two $m \times m$ matrices as $\mathbf{V}=\mathbf{V_1}+\mathbf{V_2}$, where

$$\mathbf{V_1} = diag\{v_1(x_1), \cdots, v(x_m)\} \quad \text{and} \quad \mathbf{V_2} = (v_2(x_i, x_j)).$$

From the definition of p_j in (2.2), it is easy to see

$$p_j = F(x_i + \frac{1}{2m}) - F(x_i - \frac{1}{2m}),$$

where F is the distribution function of the continuous pdf f. So both v_1 and v_2 are continuous function of $x_j's$. Now

$$m^{-2}\mathbf{X}_{x_i}^T\mathbf{W}_{x_i}\mathbf{V}\mathbf{W}_{x_i}\mathbf{X}_{x_i}$$

$$= m^{-2}\mathbf{X}_{x_i}^T\mathbf{W}_{x_i}\mathbf{V}_1\mathbf{W}_{x_i}\mathbf{X}_{x_i} + m^{-2}\mathbf{X}_{x_i}^T\mathbf{W}_{x_i}\mathbf{V}_2\mathbf{W}_{x_i}\mathbf{X}_{x_i} \qquad \text{(A.1)}$$

Referring to the result in Wand and Jones (1995, p122), the first part of (A.1) is

$$\begin{bmatrix} (\lambda m)^{-1}R(K)v_1(x_i) + o((\lambda m)^{-1}) & O(m^{-2}) \\ O(m^{-2}) & \lambda m^{-1}\mu_2(K^2)v_1(x_i) + O(m^{-2}) \end{bmatrix}. \qquad \text{(A.2)}$$

After some matrix algebra, we get the following result of the second part of (A.1).

$$m^{-2}\mathbf{X}_{x_i}^T\mathbf{W}_{x_i}\mathbf{V}_2\mathbf{W}_{x_i}\mathbf{X}_{x_i} = \begin{bmatrix} A & B \\ B & C \end{bmatrix},$$

where

$$\begin{aligned} A &= m^{-2}\sum_{k=1}^{m}\sum_{l=1}^{m} K_\lambda(x_k - x_i)K_\lambda(x_l - x_i)v_2(x_k, x_l), \\ B &= m^{-2}\sum_{k=1}^{m}\sum_{l=1}^{m} (x_l - x_i)K_\lambda(x_k - x_i)K_\lambda(x_l - x_i)v_2(x_k, x_l), \\ C &= m^{-2}\sum_{k=1}^{m}\sum_{l=1}^{m} (x_k - x_i)(x_l - x_i)K_\lambda(x_k - x_i)K_\lambda(x_l - x_i)v_2(x_k, x_l). \end{aligned}$$

It is easy to see that $A = O(m^{-1})$, $B = O(m^{-1})$, $C = O(m^{-2})$. Thus

$$m^{-2}\mathbf{X}_{x_i}^T\mathbf{W}_{x_i}\mathbf{V}_2\mathbf{W}_{x_i}\mathbf{X}_{x_i} = \begin{bmatrix} O(m^{-1}) & O(m^{-1}) \\ O(m^{-1}) & O(m^{-2}) \end{bmatrix}. \qquad \text{(A.3)}$$

Combining (A.2) and (A.3), we get

$$m^{-2}\mathbf{X}_{x_i}^T\mathbf{W}_{x_i}\mathbf{V}\mathbf{W}_{x_i}\mathbf{X}_{x_i}$$

$$= \begin{bmatrix} (\lambda m)^{-1}R(K)v_1(x_i) + o((\lambda m)^{-1}) + O(m^{-1}) & O(m^{-1}) \\ O(m^{-1}) & \lambda m^{-1}\mu_2(K^2)v_1(x_i) + O(m^{-2}) \end{bmatrix}$$

where $R(K) = \int K^2(z)dz$. These can be combined to obtain

$$Var(\hat{p}_i) = \frac{R(K)p_i}{N\lambda m} + o(\frac{1}{N\lambda m}) + O(\frac{1}{Nm}).$$

Note: This research was completed for Korea Research Foundation in March 1997. During the referring period of this paper we found a related work, Aerts, M., Augustyns, I. and Janssen, P. (1997).

REFERENCES

Aerts, M., Augustyns, I. and Janssen, P. (1997). "Smoothing Sparse Multinomial Data Using Local Polynomial Fitting," *Nonparametric Statistics*, Vol. **8**, 127-147.

Bishop, Y. M. M., Fienberg, S. E. and Holland, P. W. (1975). *Discrete Multivariate Analysis: Theory and Practice*, Cambridge, MIT Press.

Burman, P. (1987). "Smoothing Sparse Contingency Tables," *Sankhyā*, Ser. A, **49**, 24-36.

Dong, J. and Simonoff, J. S. (1994). "The Construction and Properties of Boundary Kernels for Smoothing Sparse Multinomials," *Journal of Computational and Graphical Statistics*, **3**, 1, 57-66.

Fan, J. (1992). "Design-Adaptive Nonparametric Regression," *Journal of the American Statistical Association*, **87**, 998-1004.

Grund, B. and Hall, P. (1993). "On the Performance of Kernel Estimators for High-Dimensional, Sparse Binary Data," *Journal of Multivariate Analysis*, **44**, 321-344.

Hall, P. and Titterington, D. M. (1987). "On Smoothing Sparse Multinomial Data," *Australian Journal of Statistics*, **29**, 19-37.

Leonard, T. (1978). "Density Estimation, Stochastic Processes and Prior Information (with discussion)," *Journal of the Royal Statistical Society*, Ser. B, **40**, 113-146.

Lindsey, J. K. (1992). *The Analysis of Stochastic Processes using GLIM*, Berlin, Springer-Verlag.

Maguire, B. A., Pearson, E. S. and Wynn, A. H. A. (1952). "The Time Intervals Between Industrial Accidents," *Biometrika*, **39**, 168-180.

Simonoff, J. S. (1983). "A Penalty Function Approach to Smoothing Large Sparse Contingency Tables," *The Annals of Statistics*, **11**, 208-218.

Wand, M. P. and Jones, M. C. (1995). *Kernel Smoothing*, London, Chapman & Hall.