The Annals of Statistics
1995, Vol. 23, No. 4, 1143–1159

A GEOMETRIC COMBINATION ESTIMATOR FOR d-DIMENSIONAL ORDINAL SPARSE CONTINGENCY TABLES

BY JIANPING DONG AND JEFFREY S. SIMONOFF

Michigan Technological University and New York University

A geometric combination estimator is proposed for d-dimensional ordinal contingency tables. The proposed estimator is nonnegative. It is shown that, assuming sufficient smoothness and boundary conditions for the underlying probabilities, the rate of convergence of mean summed squared error (MSSE) of this estimator is $O(K^{-1}N^{-8/(d+8)})$ for d-dimensional tables ($d \le 4$) with K cells and sample size N. This rate is optimal under the smoothness assumptions, and is faster than that attained by nonnegative kernel estimates. Boundary kernels for multidimensional tables are also developed for the proposed estimator to relax restrictive boundary conditions, resulting in summed squared error (SSE) being of order $O_p(K^{-1}N^{-8/(d+8)})$ for all $d \ge 1$. The behavior of the new estimator is investigated through simulations and applications to real data. It is shown that even for relatively small tables, these estimators are superior to nonnegative kernel estimators, in sharp contrast to the relatively unimpressive performance of such estimators for continuous data.

1. Introduction. The problem of estimating the cell probabilities of a contingency table has been an important issue in recent years. Let N be the sample size and let K be the number of cells of the table. Suppose that the underlying density possesses rth bounded derivatives. A table is said to be sparse if the number of cells is large, relative to the number of observations; formally, $K^{-1}N^{1/(2r+1)} \to 0$. It is well known that if the data are not sparse, the optimal rate of decrease of mean summed squared error (MSSE) of any estimator to zero is $O(N^{-1})$, which can be achieved by the cell proportions. In practice, we are often faced with large tables. When a table is large, it is very likely to be sparse. The cell proportion estimator can be improved in the sense that MSSE converges to zero at a faster rate for large sparse tables.

Improvement on the cell proportion estimator can come from taking advantage of smoothness in the underlying probabilities. Such smoothness would be reasonable, for example, in a table with ordered categories. Hall and Titterington (1987) (hereafter referred to as HT) proved that the optimal rate of convergence is $O(K^{-1}N^{-2r/(2r+1)})$ for one-dimensional sparse tables. A smooth version f defined on the interval [0, 1] of a discrete density was introduced there so that smoothness conditions could be described using derivatives. HT also developed a kernel estimator that achieves this optimal

Received August 1993; revised January 1994.

AMS 1991 *subject classifications*. Primary 62H17; secondary 62G07.

Key words and phrases. Boundary kernel, contingency table, kernel estimation, large sparse tables, rates of convergence.

rate for one-dimensional tables. See Simonoff (1995) for discussion of these and many other approaches to smoothing categorical data.

In this paper, we propose a new estimator, which is a geometric combination of kernel estimators with different smoothing parameters, for d-dimensional contingency tables. A geometric combination of two kernel estimators was first introduced in the continuous density estimation context by Terrell and Scott (1980) and Koshkin (1988). As noted by Jones and Foster (1993), these estimators are direct nonnegative analogues of the generalized jackknife kernels of Schucany and Sommers (1977). Jones and Foster described other estimators of similar type as well.

The rate of convergence of MSSE of our estimator is $O(K^{-1}N^{-8/(d+8)})$ for d-dimensional tables ($d \leq 4$) assuming appropriate boundary conditions. This rate is optimal under the assumption that the underlying density has bounded fourth partial derivatives. In particular, the convergence rate is $O(K^{-1}N^{-8/9})$ for one-dimensional tables.

Although the HT kernel estimator can also achieve the optimal rate of convergence, it does so at the cost of the possibility of negative probability estimates. The existence of negative probability estimates is particularly unattractive for discrete data, since a negative probability of falling in a particular cell is clearly meaningless.

Burman (1987) also described kernel estimators for categorical data, including multidimensional tables. His estimates are nonnegative, but achieve an MSSE convergence rate of $O(K^{-1}N^{-4/(d+4)})$, which is suboptimal if smoothness at the level of bounded fourth partial derivatives is present.

In the course of deriving the d-dimensional geometric combination estimator, we also generalize the univariate boundary kernel estimators of Dong and Simonoff (1994) to multidimensional tables. These estimators do not require restrictive boundary conditions on the probability matrix to avoid having bias near the boundaries dominate the MSSE of the estimator. They are then used as components of the geometric combination estimator. If this is done, the resultant estimator has summed squared error (SSE) of order $O_p(K^{-1}N^{-8/(d+8)})$, for all d, without any boundary conditions on the underlying probability matrix being necessary.

In the next section, the geometric combination estimator is described, and the statements of its SSE and MSSE convergence rates are given. Section 3 provides discussion of the d-dimensional boundary kernels needed to form the geometric combination estimator. Practical performance of the geometric combination estimator is treated in Section 4, including Monte Carlo examination of its finite sample properties. Applications to real data sets are the focus of Section 5. Proofs of the key results can be found in the Appendix.

2. The geometric combination estimator.

2.1. *Kernel estimation for discrete data.* Consider a d-dimensional table with k_j cells in the jth dimension, $1 \leq j \leq d$. The asymptotics being used here are based on large sparse tables, with the number of cells becoming infinite such that if $k_1 \cdots k_d \equiv K$, then $k_j^{-1} = O(K^{-1/d})$. That is, each dimension has the number of categories growing at the same rate. The cells of the table can be indexed by a sequence of integers $I = (i_1, \dots, i_d)$. Let $p(I)$ be the probability of falling in the Ith cell. Let $X_1, \dots, X_N$ be a random sample from a distribution whose mass function is $\{p(I)\}$. Suppose $X_t = (X_{t1}, \dots, X_{td})$ for $d > 1$. HT defined a kernel estimator for one-dimensional tables,

$$\tilde{p}(i \mid h) = (Nh)^{-1} \sum_{j=1}^{N} w\left(\frac{i - X_j}{h}\right),$$

where h is the smoothing parameter. Note that in HT [and Dong and Simonoff (1994)] the estimator was defined with smoothing parameter equivalent to h^{-1} here, but the present form is more natural, being similar to that for continuous density estimators. To achieve the convergence rate of $O(K^{-1}N^{-2r/(2r+1)})$, HT used a kernel function w such that

$$h^{-1} \sum w\left(\frac{j}{h}\right)\left(\frac{j}{h}\right)^t = 0, \qquad 1 \le t \le r - 1.$$

This kernel is defined to be of order r. In the case of $r = 4$ (four partial derivatives), the above condition forces $w(jh^{-1}) < 0$ for some j and therefore $\tilde{p}(i \mid h)$ can be negative for some i. To avoid negative estimates, the condition $\sum w(jh^{-1})(jh^{-1})^2 = 0$ has to be dropped, thereby restricting the convergence rate to $O(K^{-1}N^{-4/5})$. We propose to form an estimator that is a geometric combination of HT type estimators with different smoothing parameters.

The HT estimator can be easily generalized to d-dimensional tables. Let W_I be a d-dimensional kernel function (we will discuss the properties of W_I later). The HT type estimator for a d-dimensional table can be written as

$$\tilde{p}(I \mid h) = (Nh^d)^{-1} \sum_{t=1}^{N} W_I\left(\frac{i_1 - X_{t1}}{h}, \dots, \frac{i_k - X_{td}}{h}\right).$$

For notational simplicity, we use a single smoothing parameter h. However, different smoothing parameters for each dimension can be used if they are of the same order of magnitude.

2.2. *The geometric combination estimator.* Let $\tilde{p}(I \mid jh)$ be HT type estimators with smoothing parameters jh, $j = 1, \dots, d + 1$. Define

$$B_j(h) \equiv h^{-d} \sum_{u_d = i_d - k_d}^{i_d - 1} \cdots \sum_{u_1 = i_1 - k_1}^{i_1 - 1} W_I\left(\frac{u_1}{h}, \dots, \frac{u_d}{h}\right)\left(\frac{u_j}{h}\right)^2.$$

Now, choose $a_1(h), \dots, a_{d+1}(h)$ such that

$$(1) \qquad \sum_{s=1}^{d+1} a_s(h) = 1 \quad \text{and} \quad \sum_{s=1}^{d+1} a_s(h) B_j(sh) s^2 = 0, \qquad 1 \le j \le d.$$

The general form of the geometric combination estimator is then

$$p^*(I \mid h) = \tilde{p}(I \mid h)^{a_1(h)} \tilde{p}(I \mid 2h)^{a_2(h)} \cdots \tilde{p}(I \mid (d+1)h)^{a_{d+1}(h)}.$$

In fact, the estimator can be simplified, if the kernel function satisfies a simple symmetry condition and a common value of h is used for all dimensions. If the kernel is symmetric in its ith and jth arguments (that is,

$$W_I(x_1, \dots, x_i, \dots, x_j, \dots, x_d) = W_I(x_1, \dots, x_j, \dots, x_i, \dots, x_d)),$$

then $B_i(h) = B_j(h)$. Thus, if W_I is a symmetric function [i.e., $W_I(x_1, \dots, x_d) = W_I(x_{i_1}, \dots, x_{i_d})$, where $(i_1, \dots, i_d)$ is a permutation of $(1, \dots, n)$], then $B_1(h) = B_2(h) = \cdots = B_d(h)$. Here, $\sum_{s=1}^{d+1} a_s(h) B_1(sh) s^2 = 0$ implies that $\sum_{s=1}^{d+1} a_s(h) B_j(sh) s^2 = 0$ for all j, and the only constraints for choosing $a_1(h), \dots, a_{d+1}(h)$ are

$$\sum a_s(h) = 1 \quad \text{and} \quad \sum_{s=1}^{2} a_s(h) B_1(sh) s^2 = 0.$$

This implies that, without loss of generality, we can take $a_3(h) = \cdots = a_{d+1}(h) = 0$, giving the simplified geometric combination estimator

$$p^*(I \mid h) = \tilde{p}(I \mid h)^{a_1(h)} \tilde{p}(I \mid 2h)^{a_2(h)},$$

where

$$a_1(h) + a_2(h) = 1 \quad \text{and} \quad a_1(h) B_1(h) + 4a_2(h) B_1(2h) = 0.$$

We are ready to prove the following theorem.

THEOREM 1. *Suppose that f possesses bounded fourth partial derivatives. Let h be of order $K^{1/d} N^{-1/(d+8)}$. Suppose $p(I)$ is of order K^{-1}. Then*

$$SSE(p^*(I \mid h)) = O_p(K^{-1} N^{-8/(d+8)}).$$

A sketch of the proof of the theorem is given in the Appendix. Full details of the proof are available from the authors.

Theorem 1 is not as strong as we would like, in that convergence is in probability, rather than in quadratic mean. Theorem 2 corrects this, but at a cost. The convergence rate of MSSE in O, rather than SSE in O_p, is obtained in Theorem 2. A similar result for continuous density functions was obtained by Koshkin (1988) [see also Cramér (1946)]. The condition $W_I(\theta x) \geq W_I(x)$ for $0 < \theta < 1$ is required in the proof of Theorem 2. Boundary kernels do not satisfy this condition, and therefore cannot be used. For this reason, boundary conditions on the underlying probability matrix have to be assumed. These are stated in terms of an underlying true density f defined on the region $[0, 1]^d$. In addition, the MSSE result is only proved for $d \leq 4$. In fact, the simulations of Section 4 suggest that the MSSE of $p^*(I \mid h)$ converges to zero as expected under the conditions of Theorem 1, despite the weaker result given in that theorem.

THEOREM 2. *Suppose that f has bounded fourth partial derivatives. Suppose that*

$$f(x_1,\dots,x_d) = 0, \qquad \frac{\partial^2 f}{\partial x_{j_1}\,\partial x_{j_2}}(x_1,\dots,x_d) = 0,$$

$$\frac{\partial^3 f}{\partial x_{j_1}\,\partial x_{j_2}\,\partial x_{j_3}}(x_1,\dots,x_d) = 0$$

whenever at least one $x_i = 0$ or 1. Let h be of order $K^{1/d}N^{-1/(d+8)}$. Suppose $p(I)$ is of order K^{-1}. Then for $d \le 4$,

$$MSSE(p^*(I \mid h)) = O(K^{-1}N^{-8/(d+8)}).$$

Since the conditions of Theorem 1 are more general (and realistic) than those of this theorem, we do not present the proof of this theorem here; details are available from the authors. In the proof, an estimate that satisfies $p^*(I \mid h) \le Ch^{-d}$ is used, where C is a constant. Since h is of order $K^{1/d}N^{-1/(d+8)}$,

$$p^*(I \mid h) \le CK^{-1}N^{d/(d+8)}.$$

The larger d is, the worse the upper bound of $p^*(I \mid h)$ becomes, to the point where if $d > 4$, the bound becomes too weak. This is the reason for the requirement $d \le 4$ in the theorem.

Theorem 1 immediately leads to an important corollary:

COROLLARY 3. *For a one-dimensional table under the same conditions,*

$$SSE(p^*) = O_p(K^{-1}N^{-8/9}).$$

The choice of smoothing parameters $h, 2h, \dots, (d+1)h$ in the estimator is arbitrary. We can, in fact, choose any distinct positive real numbers $\alpha_1 h, \dots, \alpha_{d+1}h$ as the smoothing parameters, without changing the convergence rate of the estimator, as long as $a_1(h), \dots, a_{d+1}(h)$ are chosen appropriately.

Note, by the way, that these asymptotic results closely parallel the corresponding results for continuous density estimation, as do results for the kernel estimators themselves.

2.3. *Choosing $a_s(h)$.* Suppose that the d-dimensional kernel function is symmetric. For instance, $W_I(x_1,\dots,x_d) = \Pi_{i=1}^d w(x_i)$, where $w(x_i)$ is a one-dimensional kernel function (that is, W_I is a product kernel). Let $p^*(I \mid h) = \tilde{p}(I \mid h)^{a_1(h)}\tilde{p}(I \mid 2h)^{a_2(h)}$ be the simplified geometric combination estimator, where $a_1(h) = 4/(4 - g(h))$, $a_2(h) = 1 - a_1(h)$ and $g(h) = B_1(h)/B_1(2h)$. If a product of Epanechnikov kernels is used, then

$$g(h) = \left(\frac{4h^2 - 4}{4h^2 - 1}\right)\left(\frac{16h^2 - 4}{16h^2 - 1}\right)^d. \qquad (2)$$

The proof of (2) is given in the Appendix. In practice we can use the above formula to find the exact values of $a_1(h)$ and $a_2(h)$ easily. It should be noted, however, that this formula is only exact when the underlying Epanechnikov kernels are not boundary-corrected. Note that $\lim_{h\to\infty} g(h) = 1$. Thus $\lim_{h\to\infty} a_1(h) = 4/3$ and $\lim_{h\to\infty} a_2(h) = -1/3$, which corresponds to the choices suggested by Terrell and Scott (1980). In fact, $\lim_{h\to\infty} g(h) = 1$ for any kernel (not just the Epanechnikov), including boundary kernels, since both $B_1(h)$ and $B_1(2h)$ converge to the variance of the underlying kernel function. Thus, the simple form

$$p^*(I \mid h) = \tilde{p}(I \mid h)^{4/3}\tilde{p}(I \mid 2h)^{-1/3} \tag{3}$$

approximates the simplified geometric combination estimator for any underlying kernel function, and can be termed the simplified asymptotic geometric combination estimator. We recommend the use of this simple form in practice.

3. *d*-Dimensional boundary-corrected kernel estimators. In this section we derive d-dimensional boundary-corrected second order kernel estimators. These estimators generalize the one-dimensional estimators of Dong and Simonoff (1994). If boundary-corrected estimates are used in the construction of the geometric combination estimator, then its SSE properties will be valid even if the underlying probability matrix does not satisfy any boundary conditions.

It is required that the kernel function W_I satisfy the conditions:

$$h^{-d} \sum_{u_d = i_d - k_d}^{i_d - 1} \cdots \sum_{u_1 = i_1 - k_1}^{i_1 - 1} W_I\left(\frac{u_1}{h}, \dots, \frac{u_d}{h}\right) = 1, \tag{4}$$

$$h^{-d} \sum_{u_d = i_d - k_d}^{i_d - 1} \cdots \sum_{u_1 = i_1 - k_1}^{i_1 - 1} W_I\left(\frac{u_1}{h}, \dots, \frac{u_d}{h}\right)\left(\frac{u_{t_1}}{h}\right)^{\varepsilon_1}\left(\frac{u_{t_2}}{h}\right)^{\varepsilon_2}\left(\frac{u_{t_3}}{h}\right)^{\varepsilon_3} = 0, \tag{5}$$

where $\varepsilon_i = 0$ or 1, $\varepsilon_1 + \varepsilon_2 + \varepsilon_3 = 1$ or 3 and $1 \le t_1 \le t_2 \le t_3 \le d$. If we wish to use the simplified form

$$p^*(I \mid h) = \tilde{p}(I \mid h)^{a_1(h)}\tilde{p}(I \mid 2h)^{a_2(h)},$$

then we also need $B_j(h) = B_1(h)$, $2 \le j \le d$. The kernel function should also satisfy the boundedness conditions

$$\sup_{0<h<1} \left\{ h^{-d} \sum_I \left[\left|\frac{i_1}{h}\right| + \cdots + \left|\frac{i_d}{h}\right|\right]^t \left|W_I\left(\frac{i_1}{h}, \dots, \frac{i_d}{h}\right)\right| \right\} < \infty \quad \text{for } t \le 4,$$

and

$$\sup\left\{\sqrt{x_1^2 + \cdots + x_d^2} : |W_I(x_1, \dots, x_d)| > 0 \quad \text{for some } 1 < h < \infty\right\} < \infty.$$

Recall that $I = (i_1, \dots, i_d)$. If $1 + h \le i_j \le k_j - h$ for all j, then the Ith cell is not a "boundary cell." In this case, let $W_I(X) = W(X)$, where $W(X)$ is a second order symmetric kernel function defined on $[-1, 1]^d$. If the Ith cell is a boundary cell, $i_j \notin (1 + h, k_j - h)$ for some j, a special boundary kernel function is defined to correct the boundary effect. Note that (4) and (5) can be approximated by equations of integrals. For example, (5) can be approximated by

$$(5') \qquad \int_{a_d}^{b_d} \cdots \int_{a_1}^{b_1} W_I(X) x_{t_1}^{\varepsilon_1} x_{t_2}^{\varepsilon_2} x_{t_3}^{\varepsilon_3}\, dx_1 \cdots dx_d = 0, \qquad 1 \le t_1 \le t_2 \le t_3 \le d,$$

where $a_j = \max\{-1, (i_j - k_j)/h\}$ and $b_j = \min\{1, (i_j - 1)/h\}$. Now, let W_I be a polynomial with undetermined coefficients. Solving equations (5′) for the coefficients of the polynomial yields W_I.

The construction of a d-dimensional boundary-corrected kernel function W_I based on using a product of Epanechnikov kernels illustrates the use of these equations. Let

$$W_I(x_1, \dots, x_d) = \begin{cases} \prod_{t=1}^{d} w_t(x_t), & \text{if } a_t \le x_t \le b_t, 1 \le t \le d, \\ 0, & \text{otherwise,} \end{cases}$$

where $w_t(x) = a_{t0} + a_{t1}x + a_{t2}x^2 + a_{t3}x^3$, $1 \le t \le d$. The coefficients $\{a_{ij} \mid 1 \le i \le d,\ 0 \le j \le 3\}$ can be determined by solving d systems of four equations,

$$\int_{a_t}^{b_t} w_t(x)\, dx = 1, \qquad \int_{a_t}^{b_t} w_t(x)\, x\, dx = 0,$$

$$\int_{a_t}^{b_t} w_t(x)\, x^3\, dx = 0, \qquad \int_{a_t}^{b_t} w_t(x)\, x^2\, dx = B,$$

where $t = 1, \dots, d$ and $B = \int_{-1}^{1} (3/4)(1 - x^2)x^2\, dx = 1/5$. Note that these equations are, in fact, linear equations of $\{a_{tj}\}$. The solutions can be easily obtained as follows:

$$\begin{aligned} a_{t0} &= 48\left(a_t^4 + 9a_t^3 b_t + 15a_t^2 b_t^2 + 9a_t b_t^3 + b_t^4\right)(b_t - a_t)^{-7} \\ &\quad + 16\left(a_t^6 - 9a_t^5 b_t + 45a_t^4 b_t^2 + 65a_t^3 b_t^3 + 45a_t^2 b_t^4 + 9a_t b_t^5 + b_t^6\right) \\ &\qquad \times (b_t - a_t)^{-7}, \\ a_{t1} &= 60(a_t + b_t)\left(9a_t^2 + 2a_t^4 + 24a_t b_t + 16a_t^3 b_t + 9b_t^2\right. \\ &\qquad\qquad \left. + 34a_t^2 b_t^2 + 16a_t b_t^3 + 2b_t^4\right)(a_t - b_t)^{-7}, \\ a_{t2} &= 48\left(27a_t^2 + 5a_t^4 + 51a_t b_t + 45a_t^3 b_t + 27b_t^2 + 75a_t^2 b_t^2\right. \\ &\qquad\qquad \left. + 45a_t b_t^3 + 5b_t^4\right)(b_t - a_t)^{-7}, \\ a_{t3} &= 140(a_t + b_t)\left(6 + a_t^2 + 8a_t b_t + b_t^2\right)(a_t - b_t)^{-7}, \end{aligned}$$

where a_t and b_t are defined as above. Note that $\lim_{(a_t, b_t) \to (-1,1)} w_t(x) = \frac{3}{4}(1 - x^2)$, the Epanechnikov kernel. Boundary-corrected kernels for kernel functions other than Epanechnikov kernel can be found in a similar way. It should be noted, however, that [based on the results of Dong and Simonoff (1994) for univariate kernels] it is likely that these kernels will not perform adequately unless the number of categories in each dimension is large enough (say at least 20).

Other approaches to the boundary bias problem are also possible. See Jones (1993) for a thorough discussion of many approaches in the continuous density estimation context.

4. Practical performance and implementation of the estimator. The asymptotics of Section 2 do not address the question of the properties of the geometric combination estimator for finite samples. In this section the results of a small Monte Carlo study designed to investigate those properties are summarized.

Table 1 summarizes the results of simulations for one-dimensional tables. The underlying probability vector was generated based on discretizing two underlying probability densities—Beta(3, 3) and Beta(0.6, 0.6)—which were then used to generate multinomial probabilities, using code adapted from Press, Flannery, Teukolsky and Vetterling (1986). The Beta(3, 3) density is one without boundary bias effects, while the Beta(0.6, 0.6) density exhibits such effects.

The table presents results for various values of K, with either $N = K$ or $N = 5K$, for six estimators: a second order kernel estimator without boundary correction, a fourth order kernel estimator without boundary correction, a geometric combination estimator based on second order kernels without boundary correction and versions of each of these three estimators with boundary correction performed. Values of $N \times$ MSSE are given for each estimator. The number of simulation runs in each situation was 500, and differences in $N \times$ MSSE greater than approximately 0.002 were significant at a 0.05 level, based on a pairwise t-test. The kernel estimators were based on an Epanechnikov kernel, and the geometric combination estimator used was of the simplified asymptotic form (3). For each simulation run, the value of h for each estimate was chosen in one of two ways: so as to minimize the true sum of squared error (SSE) of the estimate for that data set (so the values in the table are as small as they could be) or in a data-dependent way. The latter entries are given in parentheses.

We discuss the results based on minimizing SSE first. The most obvious pattern is that boundary effects dominate the performance of all of the estimators. The estimators without boundary correction work best when there are no boundary effects, while those with boundary correction work best when boundary conditions exist. The advantage of non-boundary-corrected estimators when there are no boundary effects generally diminishes as K increases. This is consistent with the results of Dong and Simonoff (1994), and illustrates the difficulties of performing boundary bias correction in small tables.

Given these boundary-related effects, the most striking result is that the higher order estimators (the higher order kernel and geometric combination estimator) clearly outperform the second order kernel. While this is exactly what the asymptotics would suggest, it is still a particularly pleasing result.

TABLE 1
*Results of Monte Carlo simulations comparing the accuracy of various estimators**

	Beta(3, 3) Density		**Beta(0.6, 0.6) Density**	
	$N = K$	$N = 5K$	$N = K$	$N = 5K$
$K = 20$				
Second order (*nbc*)	0.064 (0.144)	0.108 (0.188)	0.271 (0.341)	0.819 (0.868)
Fourth order (*nbc*)	0.053 (0.169)	0.079 (0.195)	0.275 (0.341)	0.616 (0.708)
Geometric (*nbc*)	0.055 (0.155)	0.089 (0.186)	0.270 (0.365)	0.800 (0.881)
Second order (*bc*)	0.095 (0.180)	0.129 (0.219)	0.156 (0.312)	0.352 (0.511)
Fourth order (*bc*)	0.232 (0.364)	0.279 (0.397)	0.150 (0.353)	0.345 (0.557)
Geometric (*bc*)	0.099 (0.203)	0.123 (0.239)	0.143 (0.325)	0.326 (0.496)
$K = 50$				
Second order (*nbc*)	0.034 (0.076)	0.059 (0.098)	0.281 (0.333)	0.770 (0.799)
Fourth order (*nbc*)	0.026 (0.084)	0.042 (0.090)	0.286 (0.367)	0.625 (0.678)
Geometric (*nbc*)	0.028 (0.077)	0.048 (0.093)	0.283 (0.351)	0.768 (0.817)
Second order (*bc*)	0.040 (0.084)	0.058 (0.100)	0.140 (0.242)	0.315 (0.412)
Fourth order (*bc*)	0.099 (0.160)	0.112 (0.168)	0.144 (0.290)	0.273 (0.406)
Geometric (*bc*)	0.042 (0.093)	0.052 (0.096)	0.133 (0.238)	0.314 (0.429)
$K = 100$				
Second order (*nbc*)	0.022 (0.049)	0.036 (0.061)	0.273 (0.307)	0.714 (0.734)
Fourth order (*nbc*)	0.017 (0.047)	0.025 (0.053)	0.280 (0.329)	0.602 (0.640)
Geometric (*nbc*)	0.018 (0.046)	0.030 (0.055)	0.277 (0.319)	0.719 (0.752)
Second order (*bc*)	0.024 (0.052)	0.036 (0.059)	0.133 (0.202)	0.284 (0.347)
Fourth order (*bc*)	0.056 (0.086)	0.060 (0.090)	0.117 (0.197)	0.239 (0.309)
Geometric (*bc*)	0.025 (0.052)	0.034 (0.059)	0.129 (0.203)	0.291 (0.356)
$K = 500$				
Second order (*nbc*)	0.008 (0.011)	0.012 (0.018)	0.236 (0.246)	0.604 (0.612)
Fourth order (*nbc*)	0.008 (0.011)	0.009 (0.015)	0.244 (0.256)	0.564 (0.583)
Geometric (*nbc*)	0.007 (0.010)	0.011 (0.017)	0.241 (0.253)	0.619 (0.631)
Second order (*bc*)	0.008 (0.012)	0.012 (0.017)	0.123 (0.138)	0.233 (0.251)
Fourth order (*bc*)	0.014 (0.016)	0.014 (0.019)	0.099 (0.119)	0.195 (0.223)
Geometric (*bc*)	0.007 (0.010)	0.011 (0.018)	0.122 (0.141)	0.247 (0.266)

*Entries represent $N \times$ MSSE. The smoothing parameter in all cases is chosen to minimize either SSE or the cross-validation criterion (in parentheses). The notation *nbc* refers to estimators without boundary correction, while *bc* refers to estimators with boundary correction.

There is real question about whether density estimators with faster asymptotic convergence rates for continuous data are useful in practice for sample sizes that are not in the hundreds or even thousands; see Marron and Wand (1992) and Scott [(1992), pages 133–138]. Clearly the usefulness of higher order estimators is much more promising in the categorical data context.

If it were known whether boundary effects were going to occur, the best estimator is apparently the fourth order kernel, rather than the geometric combination estimator. There is a (small) cost to this, however, in that typically the probability of the occurrence of a negative probability estimate is about 0.01. This is not necessarily a reasonable way to view these results, however, since typically the existence of boundary bias is not known until after the data are examined. Given this fact, the boundary-corrected geometric combination estimator is much more attractive, in that when it is suboptimal, it is not nearly as bad as the other estimators. If there are no boundary effects, the boundary-corrected geometric combination estimator never has MSSE more than 80% higher than the best choice (it is usually much closer than that). In contrast, the boundary-corrected fourth order kernel estimator can have MSSE more than four times larger than the best value if no boundary effects are present. Use of the estimators without boundary correc-

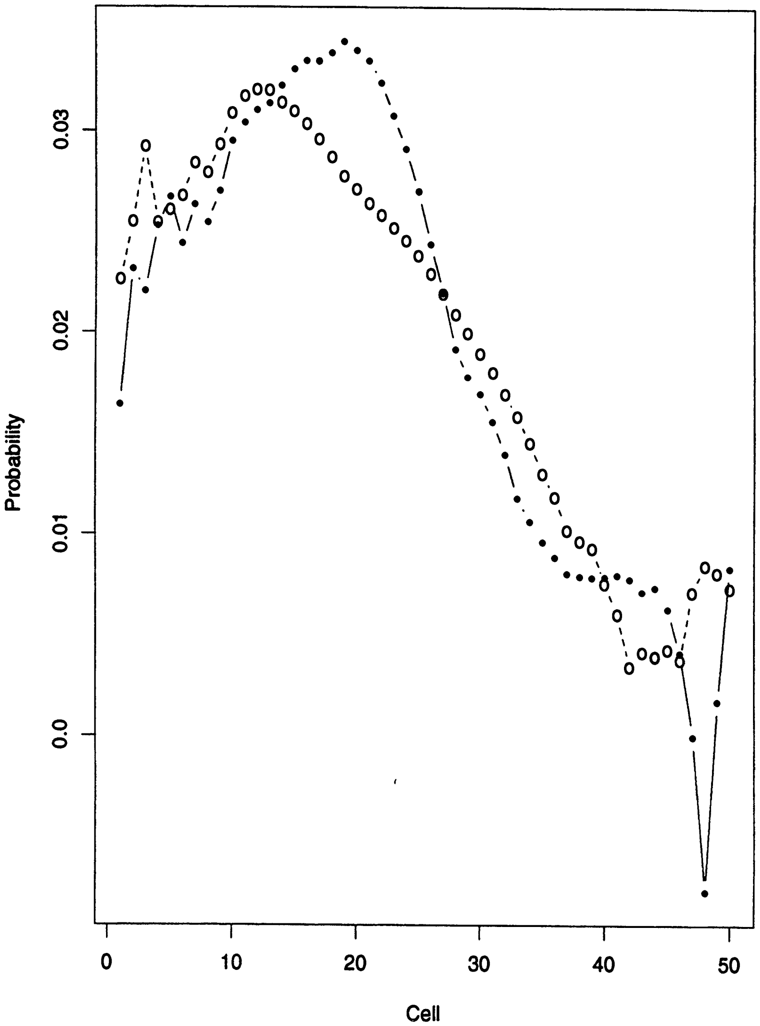


FIG. 1. *Boundary kernel estimates for calcium carbonate data*; $h = 10$ (*solid line and solid circles*) *and* $h = 20$ (*dashed line and open circles*).

tion is not advisable, either, since they have generally twice the MSSE of the boundary-corrected versions when boundary effects exist.

The values discussed thus far are based on knowledge of the true underlying probability vector, and so do not address performance of a practical method. This requires a data-based choice of h. HT and Dong and Simonoff (1994) investigated the use of cross-validation for kernel and boundary kernel estimators, respectively; this method can also be used for the geometric combination estimator. The entries in parentheses are results where h has been chosen by cross-validation for each estimator.

The latter entries are often much larger than those discussed earlier, particularly for small tables, emphasizing that the cross-validated choice of h

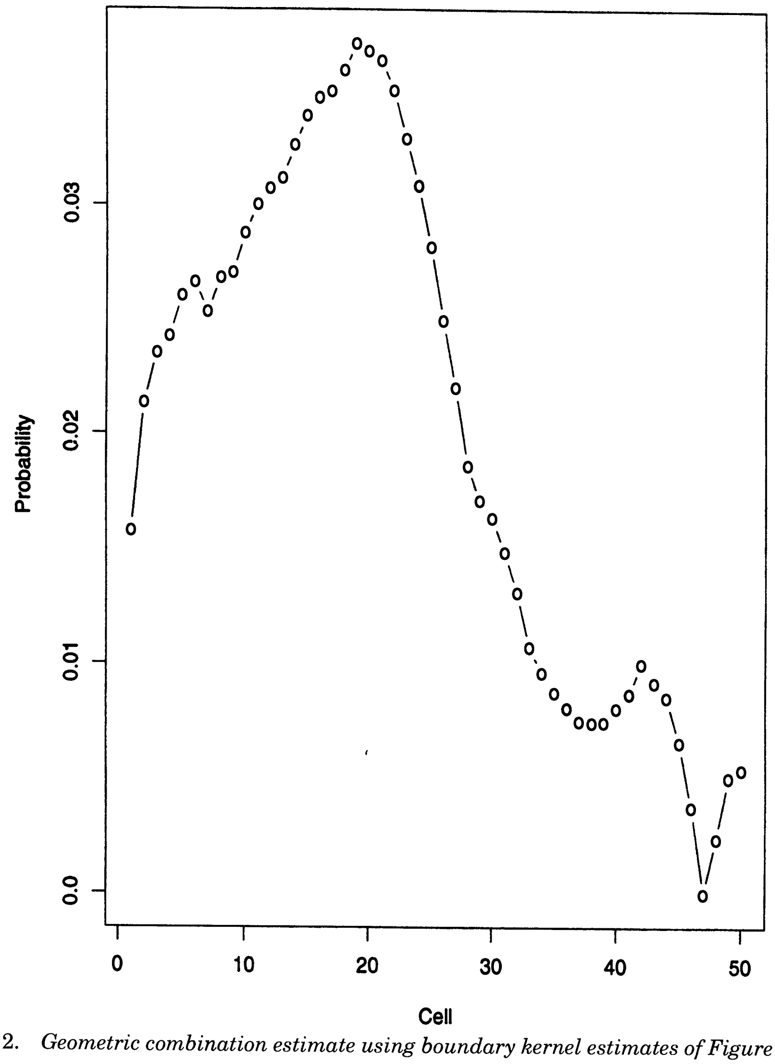


FIG. 2. *Geometric combination estimate using boundary kernel estimates of Figure* 1.

can be far from the optimal choice. Unfortunately, as is typical in smoothing problems, cross-validation sometimes leads to too small a value of h, and undersmoothing. Methods analogous to the "plug-in" method of Sheather and Jones (1991) used for kernel smoothing of continuous data, for example, would be welcome. Still, the geometric combination estimator's performance is not unreasonable. For larger values of K and for the $N = 5K$ case, the higher order estimators still outperform the second order ones. The performance of the fourth order kernel is sometimes much worse than before; apparently it can be quite difficult to choose the smoothing parameter in that case [it can be argued that this is true in the continuous density estimation

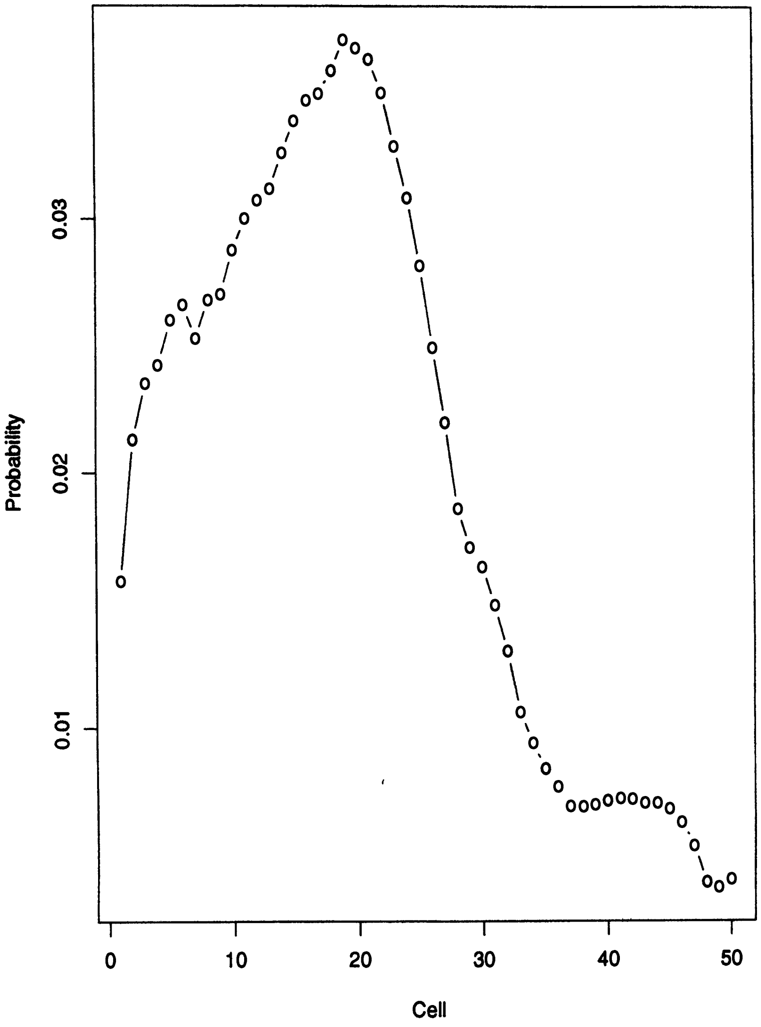


FIG. 3. *Geometric combination estimate using boundary kernels that only correct for bias in the left tail.*

context; see Scott (1992), pages 136–137]. The fourth order kernel is also more likely to be negative when using cross-validation, with a probability of a negative cell being as high as 0.03.

5. Application to real data sets. Simonoff [(1985), Table 1] examined a 50-cell multinomial adapted from Hald [(1967), page 329], that gives the range in terms of percentage concentrations of calcium carbonate for 52 sets of 5 samples each, taken from a mixing plant of raw metal. Hald [(1967), page 322] presented a theorem stating that these data should follow a normal distribution, but Simonoff noted that a Gamma distribution fits even better [see also Leonard (1978)].

Figure 1 gives boundary kernel estimates for these data using an Epanechnikov kernel and $h = 10$ (solid line and solid circles) and $h = 20$ (dashed line and open circles). Note that we are using the form of the kernel described in Section 2, rather than that of Dong and Simonoff (1994), where the smoothing parameter corresponds to h^{-1}. Although a generally asymmetric shape (consistent with a Gamma density) is apparent in both estimates, the estimate corresponding to $h = 20$ is oversmoothed, while that corresponding to $h = 10$ is too rough in the tails (exhibiting an alarming dip in the right tail).

Figure 2 gives the simplified asymptotic geometric combination estimate based on these boundary kernels; that is, (3) with $h = 10$ [note that for this value of h, (2) implies that the asymptotic form of the simplified estimator and the exact form are virtually identical]. By combining the two boundary kernel estimates, the geometric combination estimate clearly minimizes the weaknesses of both, being smooth in the tails while still identifying the mode around the 20th cell clearly.

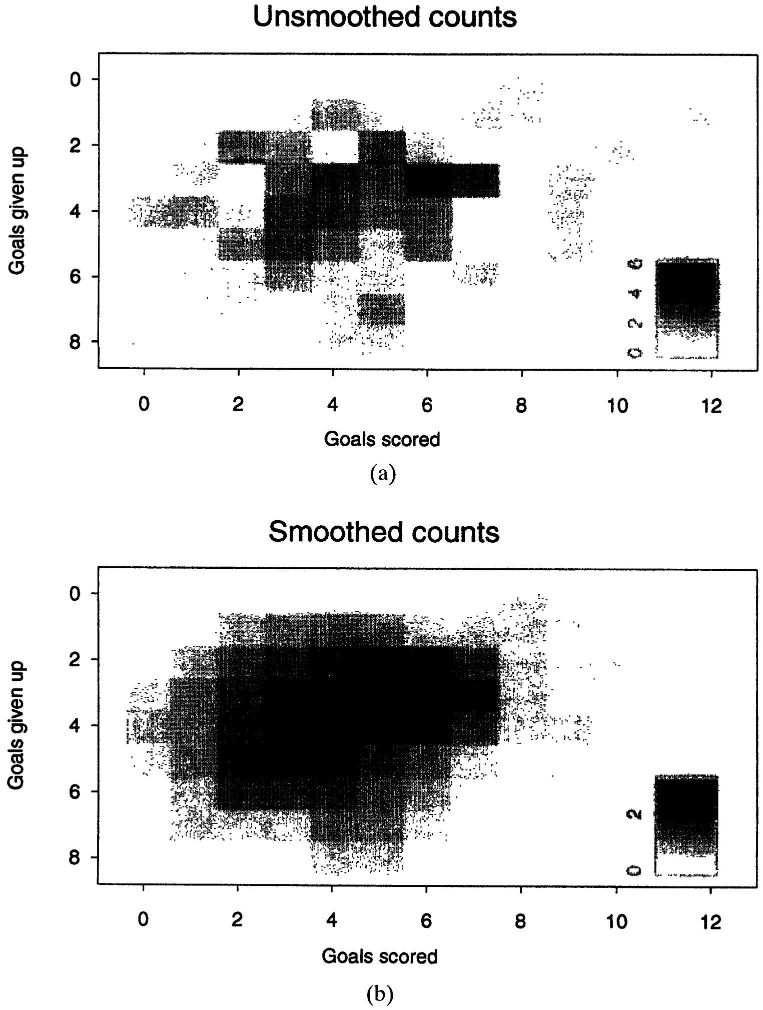
In fact, this estimate can be improved further by recognizing that boundary bias correction is only needed in the left tail. Figure 3 gives the geometric combination estimate based on kernel estimates that only correct for bias in

TABLE 2
Table of observed counts for hockey data

Goals Given Up	**Goals Scored**												
	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
0	0	0	0	0	0	0	0	0	1	0	0	0	0
1	0	0	0	1	2	1	0	1	1	0	0	0	1
2	0	1	3	2	0	3	1	0	0	0	1	0	0
3	0	1	0	2	3	2	6	3	0	1	0	0	0
4	2	2	1	3	3	2	2	0	0	1	0	0	0
5	0	1	2	3	2	1	2	0	0	1	0	0	0
6	0	1	1	2	1	1	0	1	0	0	0	0	0
7	0	1	1	0	1	2	0	0	0	0	0	0	0
8	1	0	0	0	1	1	0	0	0	0	0	0	0

the left tail. The estimate now clearly shows the structure, while avoiding completely the dip in the right tail.

Table 2 and Figure 4 summarize application to a two-dimensional table. Table 2 gives the observed counts for a cross-classification of the 80 games played by the Pittsburgh Penguins of the National Hockey League during the 1991–1992 season, classified by the number of goals scored in the game (columns) by the number of goals given up (rows) [National Hockey League (1992), page 73]. The team was only moderately successful during the season, with 39 wins, 32 losses and 9 ties, despite the fact that it ultimately won the Stanley Cup.



(a)

(b)

FIG. 4. *Shade plots for hockey data*: (a) *unsmoothed counts* (b) *smoothed counts.*

As is typical for a sparse table like this, drawing any conclusions from the table is difficult, past the impression that in most games the team both scored and gave up between 1 and 7 goals. Figure 4 gives a shade plot of the original table and smoothed counts (that is, $n \times p_{ij}^*$), based on $h = 1.7$. The plot represents each value by a shaded box, according to the legend provided. The underlying kernel estimates were not boundary-corrected. The picture is now much clearer. The dominant probability region noted earlier is still apparent, but now a high probability region exhibiting negative correlation is also indicated (note the higher counts in cells corresponding to both wins and losses with scores 5–2 and 5–3), as well as a highest probability region corresponding to 5–3 and 6–3 wins. Thus, it appears that the Penguins had a tendency to be in (slightly) lopsided contests, whether they won or lost, with a slightly greater tendency to win them.

APPENDIX

PROOF OF THEOREM 1. Let $f(X)$ be the underlying density function. Then

$$p(T) = \int \cdots \int_R f(X)\,dX = \frac{f(X_T)}{K} + \frac{1}{24}\sum_{j=1}^{d} \frac{\partial^2}{\partial x_j^2}\frac{f(X_T)}{K} k_j^{-2} + O(K^{-(d+4)/d}),$$

where $R = [(t_1 - 1)/k_1,\ t_1/k_1] \times \cdots \times [(t_d - 1)/k_d,\ t_d/k_d]$ and $X_T = ((2t_1 - 1)/2k_1, \dots, (2t_d - 1)/2t_d)$. Thus,

$$E_t = E(\tilde{p}(I \mid th)) = h^{-d} \sum W_I\left(\frac{i_1 - t_1}{th}, \dots, \frac{i_d - t_d}{th}\right) p(T), \qquad 1 \le t \le d.$$

Let $U = I - T$ and $U/K = (u_1/k_1, \dots, u_d/k_d)$. Then

$$E_t = h^{-d} \sum W_I\left(\frac{U}{th}\right)\Bigg[f\left(X_I - \frac{U}{K}\right) K^{-1} + \frac{1}{24} \sum \frac{\partial^2 f}{\partial x_j^2}\left(X_I - \frac{U}{K}\right) K^{-1} k_j^{-2} + O(K^{-(d+4)/d}) \Bigg].$$

Taking Taylor series expansions of $f(X_I - U/K)$ and $(\partial^2 f/\partial x_j^2)(X_I - U/K)$ around X_I yields

$$E_t = p(I) + K^{-1} \sum_{j=1}^{d} \sum_{U} W_I\left(\frac{u_1}{th}, \dots, \frac{u_d}{th}\right)\left(\frac{u_j}{th}\right)^2 (th)^{-d} \left(\frac{th}{2k_j}\right)^2 \frac{\partial^2 f}{\partial x_j^2}(X_I) + O(K^{-(d+4)/d} h^4).$$

Note that (4) and (5) are used here to make the terms associated with (u_j/th) and $(u_j/th)^3$ be equal to zero. Note that

$$\operatorname{Var}(\tilde{p}(I \mid h)) = O\big((KNh^d)^{-1}\big).$$

By Theorem 14.4-1 of Bishop, Fienberg and Holland (1975),

$$\tilde{p}(I \mid h) = p(I) + K^{-1} \sum_{j=1}^{d} \sum_{U} W_I\left(\frac{u_1}{th}, \dots, \frac{u_d}{th}\right)\left(\frac{u_j}{th}\right)^2 (th)^{-d} \left(\frac{th}{2k_j}\right)^2 \frac{\partial^2 f}{\partial x_j^2}(X_I)$$

$$+ O(K^{-(d+4)/d} h^4) + O_p\left((KNh^d)^{-1/2}\right).$$

By the choices of $a_i(h)$,

$$p^*(I \mid h) = p(I)\left(1 + O(K^{-4/d}h^4) + O_p\left(K^{1/2}(Nh^d)^{-1/2}\right) + R\right),$$

$$R = O_p(K^{-4/d}h^4) + O_p\left(K(Nh^d)^{-1}\right).$$

Thus,

$$p^*(I \mid h) = p(I) + O(K^{-(d+4)/d}h^4) + O_p\left((KNh^d)^{-1/2}\right) + O_p\left((Nh^d)^{-1}\right),$$

$$\begin{aligned} \mathrm{SSE}(p^*(I \mid h)) &= K\left(O(K^{-(d+4)/d}h^4) + O_p\left((KNh^d)^{-1/2}\right) + O_p\left((Nh^d)^{-1}\right)\right)^2 \\ &= O(K^{-(d+8)/d}h^8) + O_p(N^{-1}h^{-d}) + O_p(KN^{-2}h^{-2d}). \end{aligned}$$

Let h be of order $K^{1/d}N^{-1/(d+8)}$. Then

$$\mathrm{SSE}(p^*(I \mid h)) = O_p(K^{-1}N^{-8/(d+8)}). \qquad \square$$

Proof of (2).

$$\begin{aligned} B_1(h) &= h^{-d} \sum_{u_d = i_d - k_d}^{i_d - 1} \cdots \sum_{u_1 = i_1 - k_1}^{i_1 - 1} W_I\left(\frac{u_1}{h}, \dots, \frac{u_d}{h}\right)\left(\frac{u_1}{h}\right)^2 \\ &= \left(h^{-3} \sum_{j=-h}^{h} \frac{3}{4}\left(1 - \left(\frac{j}{h}\right)^2\right)(j)^2\right)\left(h^{-1} \sum_{j=-h}^{h} \frac{3}{4}\left(1 - \left(\frac{j}{h}\right)^2\right)\right)^{d-1} \\ &= \left(\frac{1}{4}\left(1 + \frac{1}{h}\right)\left(2 + \frac{1}{h}\right) - \frac{1}{20}\left(1 + \frac{1}{h}\right)\left(2 + \frac{1}{h}\right)\left(3 + \frac{3}{h} - \frac{1}{h^2}\right)\right) \\ &\quad \times \left(\frac{3}{2} + \frac{3}{4h} - \frac{1}{4}\left(1 + \frac{1}{h}\right)\left(2 + \frac{1}{h}\right)\right)^{d-1}. \end{aligned}$$

Taking the ratio of $B_1(h)$ to $B_1(2h)$ yields (2). □

Acknowledgments. This paper is based on the Ph.D. dissertation of the first author, written under the supervision of the second author, at the Department of Statistics and Operations Research of the Leonard N. Stern School of Business at New York University, whose support is gratefully acknowledged. We would like to thank several referees for helpful comments that led to substantial improvements in both the substance and presentation of material in the manuscript. In particular, the simplified form of the geometric combination estimator is based on a specific suggestion of one of the referees.

REFERENCES

BISHOP, Y. M. M., FIENBERG, S. E. and HOLLAND, P. W. (1975). *Discrete Multivariate Analysis.* MIT Press.

BURMAN, P. (1987). Smoothing sparse contingency tables. *Sankhyā Ser. A* **49** 24–36.

CRAMÉR, H. (1946). *Mathematical Methods of Statistics.* Princeton Univ. Press.

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** 57–66.

HALD, A. (1967). *Statistical Theory With Engineering Applications.* Wiley, New York.

HALL, P. and TITTERINGTON, D. M. (1987). On smoothing sparse multinomial data. *Austral. J. Statist.* **29** 19–37.

JONES, M. C. (1993). Simple boundary corrections for kernel density estimation. *Statistics and Computing* **3** 135–146.

JONES, M. C. and FOSTER, P. J. (1993). Generalized jackknifing and higher order kernels. *Journal of Nonparametric Statistics* **3** 81–94.

KOSHKIN, G. M. (1988). Improved non-negative kernel estimate of a density. *Theory Probab. Appl.* **33** 759–764.

LEONARD, T. (1978). Density estimation, stochastic processes and prior information (with discussion). *J. Roy. Statist. Soc. Ser. B* **40** 113–146.

MARRON, J. S. and WAND, M. P. (1992). Exact mean integrated squared error. *Ann. Statist.* **20** 712–736.

NATIONAL HOCKEY LEAGUE (1992). *The National Hockey League Official Guide and Record Book 1992–93.* Triumph Books, Chicago.

PRESS, W. H., FLANNERY, B. P., TEUKOLSKY, S. A. and VETTERLING, W. T. (1986). *Numerical Recipes.* Cambridge Univ. Press.

SCHUCANY, W. R. and SOMMERS, J. P. (1977). Improvement of kernel type density estimators. *J. Amer. Statist. Assoc.* **72** 420–423.

SCOTT, D. W. (1992). *Multivariate Density Estimation: Theory, Practice and Visualization.* Wiley, New York.

SHEATHER, S. J. and JONES, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *J. Roy. Statist. Soc. Ser. B* **53** 683–690.

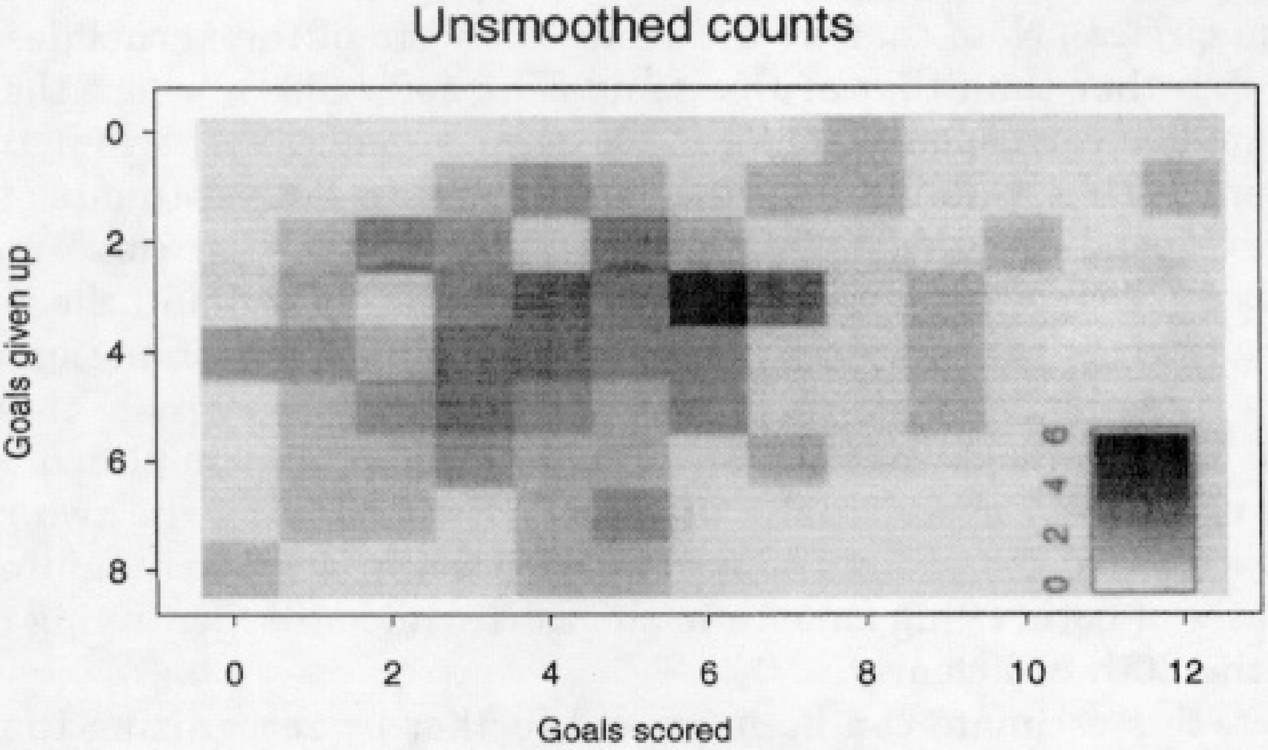
SIMONOFF, J. S. (1985). An improved goodness-of-fit statistic for sparse multinomials. *J. Amer. Statist. Assoc.* **80** 671–677.

SIMONOFF, J. S. (1995). Smoothing categorical data. *J. Statist. Plann. Inference* **40**. To appear.

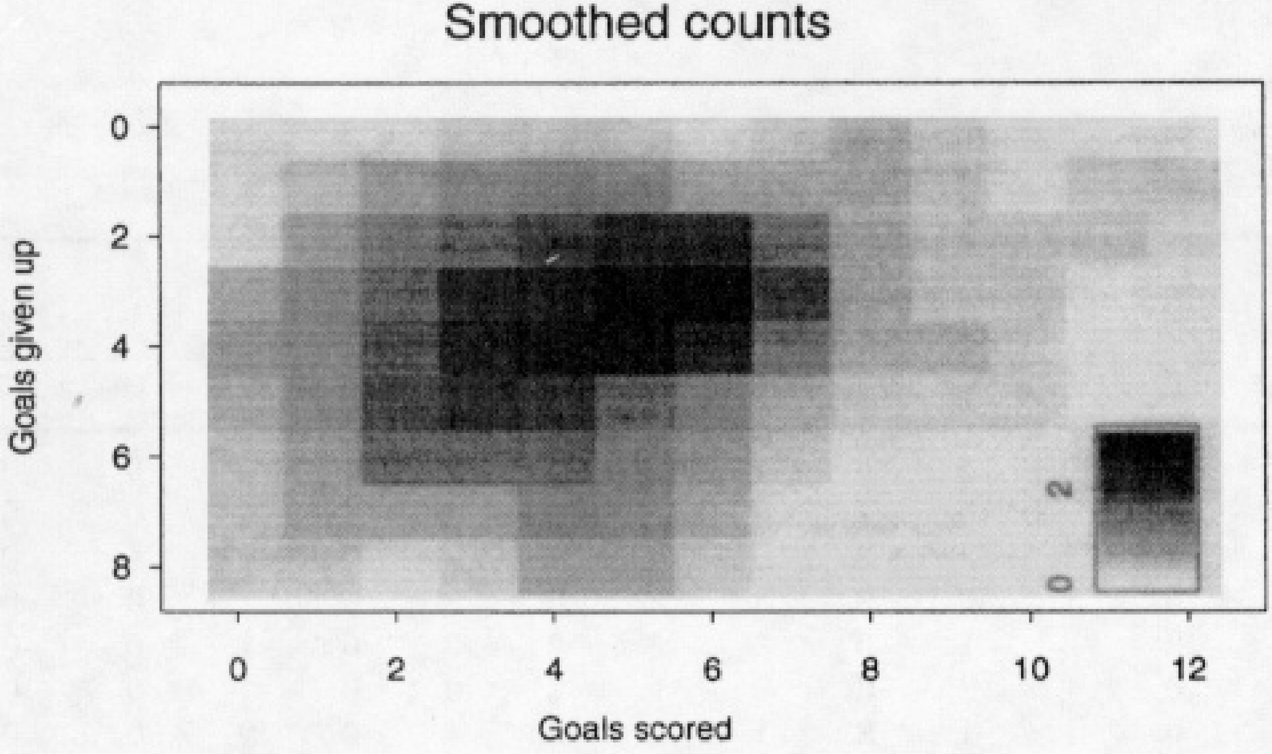
TERRELL, G. R. and SCOTT, D. W. (1980). On improving convergence rates for nonnegative kernel density estimators. *Ann. Statist.* **8** 1160–1163.

DEPARTMENT OF MATHEMATICAL SCIENCE
MICHIGAN TECHNOLOGICAL UNIVERSITY
HOUGHTON, MICHIGAN 49931

DEPARTMENT OF STATISTICS AND
OPERATIONS RESEARCH
NEW YORK UNIVERSITY
44 WEST 4TH STREET, ROOM 8-54
NEW YORK, NEW YORK 10012-1126


Unsmoothed counts
Goals given up
0
2
4
6
8
Goals scored
0
2
4
6
8
10
12
0 2 4 6

(a)


Smoothed counts
Goals given up
0
2
4
6
8
Goals scored
0
2
4
6
8
10
12
0 2

(b)