Adv. Appl. Prob. **18,** 388–405 (1986)
Printed in N. Ireland

MUTUAL AND SHARED NEIGHBOR PROBABILITIES: FINITE- AND INFINITE-DIMENSIONAL RESULTS

M. F. SCHILLING,* *California State University, Northridge*

Abstract

Let $X_1, \cdots, X_n$ be i.i.d. random variables defined in $\mathbb{R}^d$ having common continuous density $f(x)$, and let R_{ij} be the rank of X_j in the ordered list of distances from X_i. Both the mutual neighbor probabilities $p_1(r, s) = P(R_{12} = r, R_{21} = s)$ and the neighbor-sharing probabilities $p_2(r, s) = P(R_{13} = r, R_{23} = s)$ are studied from an asymptotic viewpoint. Infinite-dimensional limits are found for both situations and take particularly simple forms. Both cases exhibit considerable stability across dimensions and thus are well approximated by their infinite-dimensional values. Tables are provided to support the results given.

NEAREST NEIGHBORS; GEOMETRIC PROBABILITY

1. Introduction and motivation

The relationships among near neighbors in a collection of events are of interest both in mathematical statistics and in many areas of applied science such as pattern recognition, the social sciences and ecology. As an example of an ecological application, a dominant species in many desert regions of the American Southwest is *Larrea divaricata,* the creosote bush. This plant is known to exude a chemical into the surrounding soil which, along with a widely spreading root system, inhibits additional plant growth (including that of creosote itself) in the immediate vicinity. This results in a markedly regular distribution of the creosote bush across its habitat. Distributions of different species exhibit clustering tendencies, while still others behave as the realization of a two-dimensional Poisson process. Each of these patterns contains entirely different near-neighbor relationships. For results and procedures pertaining to this area of application, see Clark and Evans (1955), Clark (1956), Diggle (1975), Cox and Lewis (1976), Cox (1976), and references therein.

The initial motivation for the current research stemmed from consideration of the following statistical situation. Let $X_1, \cdots, X_n$ be i.i.d. observations from

Received 10 July 1984; revision received 24 May 1985.

* Postal address: Department of Mathematics, School of Science and Mathematics, California State University, Northridge, 18111 Nordhoff St, Northridge, CA 91330, USA.

Research supported in part by National Science Foundation Grants MCS79-19141, MCS80-17103.

some distribution in $\mathbb{R}^d$ possessing continuous density $f(x)$, and let $NN_i(r)$ represent that sample point which is the rth nearest neighbor to X_i. Ties are neglected since their occurrence is an event with zero probability. Then a natural family of statistics which may be considered for inferences about f is based on the information contained in each point and its k nearest neighbors for some $k \geqq 1$ and takes the form

$$T_h(X_1, \cdots, X_n) = \sum_{i=1}^{n} \sum_{r=1}^{k} h(X_i, NN_i(r), r)$$
$$= \sum_{i=1}^{n} \sum_{i'=1}^{n} \sum_{r=1}^{k} h(X_i, X_{i'}, r) I(NN_i(r) = X_{i'})$$

where $I(\cdot)$ represents the indicator function. (For a particular application to non-parametric multidimensional two-sample testing, see Schilling (1986).)

Under conditions ensuring asymptotic normality the limiting distribution of T_h is determined by its first two moments. This requires the computation of covariance terms of the form

$$Eh(X_i, X_{i'}, r)h(X_j, X_{j'}, s)I(NN_i(r) = X_{i'}, NN_j(s) = X_{j'}). \tag{1.1}$$

When $i \neq j$ various nearest-neighbor geometries come into play. There are five mutually exclusive and exhaustive cases:

(i) $NN_i(r) = X_j$, $NN_j(s) = X_i$;
(ii) $NN_i(r) = NN_j(s)$;
(iii) $NN_i(r) = X_j$, $NN_j(s) \neq X_i$;
(iv) $NN_i(r) \neq X_j$, $NN_j(s) = X_i$;
(v) $NN_i(r) \neq X_j$, $NN_j(s) \neq X_i$, $NN_i(r) \neq NN_j(s)$.

(See Figure 1.) X_i and X_j are called mutual neighbors if case (i) occurs, whereas in case (ii) they share a common neighbor. The term 'reflexive nearest neighbors', due to Clark and Evans (1955), has been used by several authors for case (i).

Let the above events and their respective probabilities be denoted by E_m and $p_m(r, s)$, $m = 1, \cdots, 5$, suppressing for brevity the dependence on f, n and d. These probabilities are independent of i and j since the X_i's are exchangeable. Then for $i \neq j$ (1.1) can be written as

$$\sum_{m=1}^{5} \{Eh(X_i, X_{i'}, r)h(X_j, X_{j'}, s) \mid E_m\} p_m(r, s).$$

Using the exchangeability of the X_i's the number of distinct conditional expectations is fairly small and these values will be feasible to compute for certain functions h. Thus the determination of the p_m's is of paramount

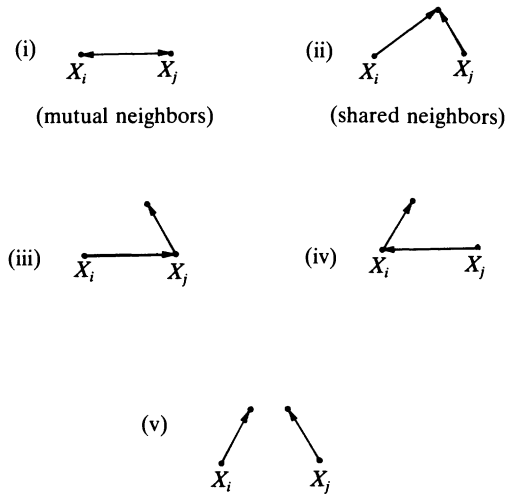

Figure 1. Neighbor configurations. Arrows from X_i and X_j point to $NN_i(r)$ and $NN_j(s)$ respectively

importance. Note however that

$$p_1(r, s) = P(NN_1(r) = X_2 \mid NN_2(s) = X_1)P(NN_2(s) = X_1)$$

$$= P(NN_1(r) = X_2 \mid NN_2(s) = X_1) \cdot \frac{1}{n-1},$$

from which it follows readily that

$$p_3(r, s) = p_4(r, s) = \frac{1}{n-1} - p_1(r, s),$$

$$p_5(r, s) = \frac{n-3}{n-1} + p_1(r, s) - p_2(r, s).$$

Thus it is only necessary to calculate $p_1(r, s)$ and $p_2(r, s)$ for each r and s, the probabilities associated with mutual and shared neighbors, respectively.

In this paper, asymptotic values for p_1 and p_2 are obtained for arbitrary r and s by exploiting the fact that as n grows and the point density increases, attention can be restricted to small regions of the observation space, within which f is nearly constant; thus the sample behaves locally as a d-dimensional Poisson process. The organization is as follows. Section 2 defines additional notation and presents the result for mutual neighbors. In Section 3 an expression for neighbor-sharing probabilities is obtained for arbitrary $r, s \geqq 1$ and $d > 1$. The case $d = 1$ is treated separately. The limiting behavior of p_1 and p_2 as d varies is studied in Section 4, where letting $d \to \infty$ is shown to produce substantial simplifications. Tables are provided for small r and s and various d.

A discussion of related work and some additional results involving the distribution of the number of points claiming a given point as their nearest neighbor is given in Section 5.

2. Mutual neighbor probabilities

Let $X_1, \cdots, X_n$ be as in the introduction and define

$$R_{ij} = \text{rank of } X_j \text{ among } X_1, \cdots, X_{i-1}, X_{i+1}, \cdots, X_n \text{ when ordered by increasing (Euclidean) distance from } X_i,\ i = 1, \cdots, n;\ j = 1, \cdots, n;\ j \neq i.$$

Thus $R_{ij} = k$ indicates that X_j is the kth nearest neighbor of X_i. By the exchangeability of the X_i's only R_{12}, R_{21}, R_{13} and R_{23} need to be considered in determining $p_1(r, s)$ and $p_2(r, s)$. We clearly have

$$p_1(r, s) = P(R_{12} = r, R_{21} = s);$$
$$p_2(r, s) = (n-2)P(R_{13} = r, R_{23} = s)$$

for the mutual neighbor and neighbor-sharing cases respectively.

In finite samples these probabilities depend on the underlying density f and are extremely difficult to compute exactly. It is shown below, however, that both $np_1(r, s)$ and $np_2(r, s)$ approach computable limits which are independent of f.

The mutual neighbor case has essentially been obtained by Cox (1981), who generalized a formula first given in incorrect form by Clark (1956) and later corrected by Dacey (1969). Recall that $p_1(r, s) = P(NN_1(r) = X_2 \mid NN_2(s) = X_1)(n-1)^{-1}$. Cox determined the values of the conditional probabilities in the case when the X_i's are events in a d-dimensional Poisson process.

Let $\|\cdot\|$ represent the Euclidean norm and for $x_0 \in \mathbb{R}^d$ and $\rho > 0$ write

$$S(x_0, \rho) = \{x \in \mathbb{R}^d : \|x - x_0\| < \rho\}$$

for the sphere centered at x_0 having radius ρ. Let

$$S = S(X_1, \|X_2 - X_1\|); \quad \bar{S} = S(X_2, \|X_2 - X_1\|).$$

Cox's formula applies directly to the current model to yield the following result.

Theorem 2.1.

$$\lim_{n\to\infty} np_1(r, s) = (1 - C_d) \sum_{l=0}^{\min(r', s')} \binom{r' + s' - l}{l, r' - l, s' - l} (1 - 2C_d)^l C_d^{r'+s'-2l},$$

where $r' = r - 1$, $s' = s - 1$, and C_d is the proportion of the volume of $S \cup \bar{S}$ which belongs to S (say) only.

The value of l in Theorem 2.1 arises as the number of points X_i, $i > 2$, falling

in $S \cap \tilde{S}$, which necessitates $r' - l$ and $s' - l$ points in $S \cap \tilde{S}^c$ and $S^c \cap \tilde{S}$ respectively.

3. Neighbor-sharing

The 'three-body problem' associated with the neighbor-sharing probabilities $p_2(r, s)$ is considerably more complex than the case of mutual neighbors. Initially attention is restricted to $d > 1$.

Some further notational definitions are required. Take $\sim$ to mean 'is equivalent in the limit to'. For any sets A and B, let A^1 and A^0 represent A and the complement of A, respectively, and write $B - A$ for $B \cap A^0$. Let $V\{A\} = \int_A dx$ for A measurable, and denote the volume $V\{S(\cdot, 1)\}$ of a d-sphere of radius 1 by K_d. The value of K_d is $\pi^{d/2}/\Gamma(d/2 + 1)$, where $\Gamma(\cdot)$ represents the gamma function. Let $\mathbf{0}$ and $\boldsymbol{e}$ indicate the d-vectors $(0, 0, \cdots, 0)$ and $(1, 0, \cdots, 0)$, respectively.

To determine the asymptotic value of $np_2(r, s)$ for arbitrary r, s and $d > 1$, begin by decomposing according to the identity of the shared neighbor and use exchangeability to obtain

$$np_2(r, s) = n(n-2)P(NN_1(r) = NN_2(s) = X_3).$$

It is necessary to subdivide the space in which X_3 may lie. For given $x_1, x_2 \neq x_1 \in \mathbb{R}^d$ let

$$A_{\alpha,\beta} = S^\alpha \cap \tilde{S}^\beta; \qquad \alpha, \beta = 0, 1.$$

In addition put

$$S_i = S(x_i, \|x_i - x_3\|); \qquad i = 1, 2.$$

(See Figure 2.) Let l now represent the number of points in $S_1 \cap S_2$ rather than in $S \cap \tilde{S}$ as before, and write δ and ε for the number of points X_i, $i > 3$, falling into $S_1 - S_2$ and $S_2 - S_1$, respectively. Note that, given $X_3 \in A_{\alpha,\beta}$, l may now range from 0 to $\bar{l} = \min(r + \alpha - 2, s + \beta - 2)$, and determines δ and $\varepsilon : \delta = r - l + \alpha - 2$, $\varepsilon = s - l + \beta - 2$.

Considering all possible configurations for the location of X_3 and the value of l yields

$$\begin{aligned} np_2(r, s) = n(n-2) \sum_{\alpha,\beta=0}^{1} \sum_{l=0}^{\bar{l}} \binom{n-3}{l, \delta, \varepsilon} \int_{\mathbb{R}^d} f(x_1) \int_{\mathbb{R}^d} f(x_2) \int_{A_{\alpha,\beta}} f(x_3) \\ \cdot \left\{ \int_{S_1 \cap S_2} f(x)\, dx \right\}^l \left\{ \int_{S_1 - S_2} f(x)\, dx \right\}^\delta \left\{ \int_{S_2 - S_1} f(x)\, dx \right\}^\varepsilon \\ \cdot \left\{ 1 - \int_{S_1 \cup S_2} f(x)\, dx \right\}^{n-l-\delta-\varepsilon-3} dx_3\, dx_2\, dx_1. \end{aligned}$$

Terms with $\bar{l} < 0$ (which occur when $\min(r, s) = 1$) are taken to be 0.

Change variables to $y = n^{1/d}(x - x_1)$ and $y_i = n^{1/d}(x_i - x_1)$, $i = 2, 3$, in order

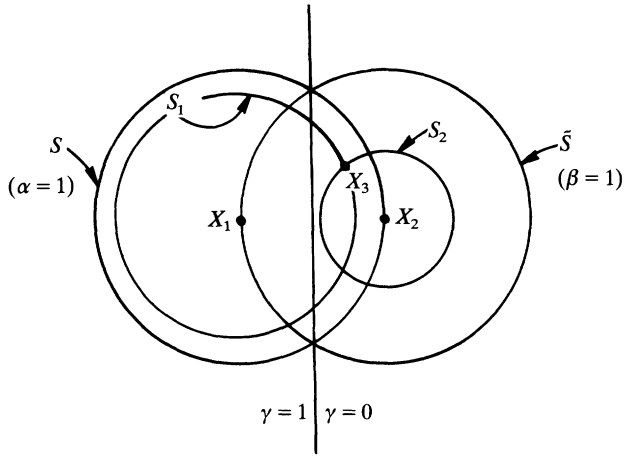

Figure 2. Near-neighbor spheres. Spheres involved in the determination of the probability that X_1 and X_2 share a near neighbor X_3

to center and stabilize the probability contents of the regions of integration. Then the continuity of f allows us to approximate f locally by a d-dimensional Poisson process. We have

$$\int_{S_1 \cap S_2} f(x)\, dx = \frac{1}{n} \int_{S_1' \cap S_2'} f(x_1 + n^{-1/d} y)\, dy,$$

which, for fixed x_1 and y equals

$$\frac{1}{n} f(x_1) V(S_1' \cap S_2') + o\left(\frac{1}{n}\right),$$

where S_1' and S_2' are the images of S_1 and S_2 under the change of variables, namely $S_1' = S(\mathbf{0}, \|y_3\|)$, $S_2' = S(y_2, \|y_3 - y_2\|)$; similar approximations for other terms and dominated convergence lead to

$$\begin{aligned} &np_2(r, s) \\ &= \sum_{\alpha,\beta=0}^{1} \sum_{l=0}^{\bar{l}} \frac{1}{l!\, \delta!\, \varepsilon!} \int_{\mathbb{R}^d} f^{l+\delta+\varepsilon+3}(x_1) \int_{\mathbb{R}^d} \int_{A'_{\alpha,\beta}} V^l\{S_1' \cap S_2'\} \\ &\quad \cdot V^\delta\{S_1' - S_2'\} V^\varepsilon\{S_2' - S_1'\} \exp\left(-f(x_1) V\{S_1' \cup S_2'\}\right) dy_3\, dy_2\, dx_1 + o(1) \end{aligned}$$

where the $A'_{\alpha,\beta}$ are the mappings of the corresponding $A_{\alpha,\beta}$ under the transformation to (x_1, y_2, y_3)-space.

Changing variables once more to $z_i = y_i / \|y_2\|$, $i = 2, 3$, $z = \|y_2\|^d$ and computing volumes by revolution produces

$$\begin{aligned} np_2(r, s) \sim K_d &\sum_{\alpha,\beta=0}^{1} \sum_{l=0}^{\bar{l}} \frac{1}{l!\, \delta!\, \varepsilon!} \int_{\mathbb{R}^d} f^{l+\delta+\varepsilon+3}(x_1) \int_0^\infty z^{l+\delta+\varepsilon+1} \\ &\cdot \int_{A''_{\alpha,\beta}} V^l\{S_1'' \cap S_2''\} V^\delta\{S_1'' - S_2''\} V^\varepsilon\{S_2'' - S_1''\} \exp\left(-f(x_1) V\{S_1'' \cup S_2''\} z\right) dz_3\, dz\, dx_1, \end{aligned}$$

where z_2 is replaced without loss of generality by $\boldsymbol{e}$ in defining $S_1'' = S(\mathbf{0}, \|z_3\|)$, $S_2'' = S(\boldsymbol{e}, \|z_3 - \boldsymbol{e}\|)$ and $A''_{\alpha,\beta} = (S_1'')^\alpha \cap (S_2'')^\beta$ for $\alpha, \beta = 0, 1$.

Fubini's theorem allows z and x_1 to be integrated out successively, giving

$$np_2(r, s) \sim K_d \sum_{\alpha,\beta=0}^{1} \sum_{l=0}^{\bar{l}} \binom{l+\delta+\varepsilon+1}{l, \delta, \varepsilon, 1} \int_{A''_{\alpha,\beta}} V^l\{S_1'' \cap S_2''\} V^\delta\{S_1'' - S_2''\} \cdot V^\varepsilon\{S_2'' - S_1''\} V^{-(l+\delta+\varepsilon+2)}\{S_1'' \cup S_2''\}\, dz_3.$$

Next transform to the polar coordinates $\nu = \|z_3\|$, $\theta =$ angle between the vectors $\boldsymbol{e}$ and z_3, and let $\rho = 1/\nu$. We can again compute volumes by revolution to reduce the above expression to its simplest form, a double integral in the variables ρ and θ. Define

$$D(\rho, \theta) = (1 - 2\rho\cos\theta + \rho^2)^{\frac{1}{2}}$$

$$Q(\rho, \theta) = V\left\{S\left(\mathbf{0}, \frac{1}{\rho}\right) \cap S\left(\boldsymbol{e}, \frac{1}{\rho} D(\rho, \theta)\right)\right\} \Big/ V\left\{S\left(\mathbf{0}, \frac{1}{\rho}\right)\right\},$$

$$H(\rho, \theta) = D^d(\rho, \theta) - \mathbf{0}(\rho, \theta).$$

The quantity $D(\rho, \theta)$ is, by the law of cosines, the third side of a triangle whose other two sides have lengths 1 and ρ and form an angle θ. We have $\|z_3 - e\| = D(\nu, \theta) = \rho^{-1} D(\rho, \theta)$. (See Figure 3.) $Q(\rho, \theta)$ and $H(\rho, \theta)$ represent the volume of intersection and the volume unique to the rightmost sphere, respectively, for the two spheres whose cross-sections are shown in Figure 3, each expressed as a proportion of the volume of the leftmost sphere.

It is convenient to decompose z_3-space into the regions $\{z_3 : \|z_3\| < \|z_3 - \boldsymbol{e}\|\}^\gamma$, $\gamma = 0, 1$, which correspond in the original system to the events that the shared neighbor is either closer to ($\gamma = 0$) or farther from ($\gamma = 1$) X_2 than X_1 (see Figure 2). Consider first the contribution to $np_2(r, s)$ from $\gamma = 0$ only.

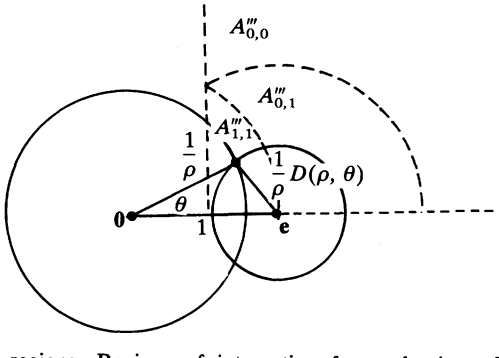

Figure 3. Integration regions. Regions of integration for evaluation of the neighbor-sharing quantity $p_2'(r, s)$. The quadrant between the perpendicular dashed lines yields the contribution of the case $\gamma = 0$

Define the regions

$$A'''_{0,0} = \left\{(\theta, \rho): 0 \leqq \theta < \pi/2,\ 0 < \rho < \min\left(\frac{1}{2\cos\theta}, 2\cos\theta\right)\right\};$$

$$A'''_{0,1} = \left\{(\theta, \rho): 0 \leqq \theta < \pi/3,\ \frac{1}{2\cos\theta} < \rho < 1\right\};$$

$$A'''_{1,0} = \phi;$$

$$A'''_{1,1} = (\theta, \rho): 0 \leqq \theta \leqq \pi/3,\ 1 \leqq \rho < 2\cos\theta\};$$

$$A''' = \bigcup_{\alpha,\beta=0}^{1} A'''_{\alpha,\beta}.$$

The $A'''_{\alpha,\beta}$ are the images under the current transformation of the portions of the respective $A''_{\alpha,\beta}$ regions in which $\gamma = 0$. (See Figure 3.) The advantage of ρ over ν is that the above regions are compact, which facilitates numerical integration.

For the integrand, note that the locus of points $\{z_3\}$ with a particular value of (ν, θ) defines a $(d-1)$-dimensional sphere of radius $\nu \sin\theta$ having surface area $(d-1)K_{d-1}(\nu \sin\theta)^{d-2}$ and that $V(S''_1 \cap S''_2) = K_d \nu^d Q(1/\nu, \theta)$. Then computing volumes by revolution and using $\rho = 1/\nu$ yields the contribution of $\gamma = 0$ to $\lim_{n\to\infty} np_2(r, s)$ given below. Observe that for the $\gamma = 1$ contribution we need only reverse the coordinate system by interchanging $\mathbf{0}$ and $\boldsymbol{e}$ and the result is the same but with δ and ε permuted. Hence we have finally the following result.

Theorem 3.1. For $d > 1$, $\lim_{n\to\infty} np_2(r, s)$ exists and equals $p'_2(r, s; 0) + p'_2(r, s; 1)$, where

$$(3.1)\quad p'_2(r, s; 0) = (d-1)\frac{K_{d-1}}{K_d} \sum_{\alpha,\beta=0}^{1} \sum_{l=0}^{\bar{l}} \binom{l+\delta+\varepsilon+1}{l, \delta, \varepsilon, 1} \iint_{A'''_{\alpha,\beta}} Q^l(\rho, \theta)$$
$$\times \{1 - Q(\rho, \theta)\}^{\delta} H^{\varepsilon}(\rho, \theta)\{1 + H(\rho, \theta)\}^{-(l+\delta+\varepsilon+2)} \rho^{d-1} \sin^{d-2}\theta \, d\rho \, d\theta$$

and $p'_2(r, s; 1)$ has the same form but with δ and ε reversed.

The quantities $p'_2(r, s; 0)$ and $p'_2(r, s; 1)$ represent the contributions of the regions where $\gamma = 0$ and $\gamma = 1$ respectively. It is easy to express $Q(\rho, \theta)$ for $\gamma = 0$ as a one-dimensional integral by revolving around the first coordinate axis:

$$Q(\rho, \theta) = \frac{K_{d-1}}{K_d}\left[\int_0^{\theta} \sin\phi \, d\phi + D^d(\rho, \theta) \int_0^{\cos^{-1}\{(\cos\theta - \rho)/D(\rho, \theta)\}} \sin\phi \, d\phi\right].$$

Thus it is feasible to compute numerical values for $\lim_{n\to\infty} np_2(r, s)$ if r and s are not too large, with no increase in difficulty as the dimension d increases.

Finally, we give attention to the limiting values of $np_2(r, s)$ for the univariate situation, which is not covered by Theorem 3.1.

Theorem 3.2. For $d = 1$,

$$\lim_{n\to\infty} np_2(r, s) = \begin{cases} \frac{3}{2} - 2^{1-r} & (r = s); \\ 1 - \binom{r-1}{s-1} 2^{-r} & (r > s). \end{cases}$$

The proof follows generally along the lines of the proof of Theorem 3.1 and is omitted; however, some of the differences will be mentioned. The event $(\alpha, \beta) = (1, 0)$, which implies that the shared neighbor lies closer to X_1 than to X_2, is clearly possible only if $s > r$. Similarly, $(\alpha, \beta) = (0, 1)$ can occur only when $r < s$, and if $r = s$ then $\alpha = \beta$. The nearest-neighbor spheres S_1, S_2 are either disjoint or nested in the one-dimensional case, which forces the values of l, δ and ε for each region $A_{\alpha,\beta,\gamma}$ into which X_3 can fall.

Curiously, the one-dimensional values of $\lim_{n\to\infty} np_2(r, s)$ tend to $\frac{3}{2}$ as $r = s \to \infty$ while along any sequence of pairs $\{(r, s): r \neq s;\ r, s \to \infty\}$ the value tends to 1. This provides an interesting comparison with the infinite-dimensional result for p_2 obtained in the next section.

4. Infinite-dimensional values

A curious phenomenon of nearest-neighbor processes is their tendency towards simplicity as the spatial dimension becomes infinite. This behavior has been observed elsewhwere (Schilling (1983); Newman et al. (1983)) and it occurs in the present situation as well.

The concept of asymptotics in the dimension is used below to produce simplified versions of the near-neighbor quantities $p_1(r, s)$ and $p_2(r, s)$. The limit in d of $p_2(r, s)$ is particularly simple. Write

$$p_i'(r, s) = \lim_{n\to\infty} np_i(r, s) \quad \text{for} \quad i = 1, 2,$$

having proved the existence of such limits in Theorems 2.1 and 3.1.

4.1. *Mutual neighbors.* The mutual neighbor quantity $p_1(r, s)$ has a nearly binomial form as the dimension d becomes large.

Theorem 4.1.

$$\lim_{d\to\infty} p_1'(r, s) = \binom{r+s-2}{r-1, s-1} 2^{1-r-s}.$$

The theorem follows directly from Theorem 2.1 and Lemma 4.1 below, which exposes the source of the binomial character.

Lemma 4.1. $\lim_{d\to\infty} C_d = \frac{1}{2}$.

Proof. Computing volume by revolution gives

$$1 - 2C_d = 2\frac{K_{d-1}}{K_d}\int_{\frac{1}{2}}^{1}(1-x^2)^{(d-1)/2}\,dx \leqq \frac{K_{d-1}}{K_d}\left(\tfrac{3}{4}\right)^{(d-1)/2}.$$

Now Stirling's approximation yields $K_{d-1}/K_d \sim (d/2\pi)^{\frac{1}{2}}$ and the lemma follows.

Since $C_1 = \frac{1}{3}$, $p_1'(r, s)$ also takes a rather simple form for $d = 1$. From Theorem 2.1 we have, writing $r' = r - 1$, $s' = s - 1$ as before,

$$p_1'(r, s) = 2\sum_{l=0}^{\min(r',s')}\binom{r'+s'-l}{l, r'-l, s'-l}3^{-(r'+s'-l+1)} \quad (d = 1).$$

The convergence of C_d to $\frac{1}{2}$ as $d\to\infty$ is fairly rapid, as shown in Table 1.

Table 2 presents a comparison across dimensions of $p_1'(r, s)$ values, obtained from Theorems 2.1 and 4.1 and Table 1, for $1 \leqq s \leqq r \leqq 3$. The $d = \infty$ values appear to provide reasonable approximations for rather low dimensions.

A natural question concerning nearest neighbors is the following: 'What is the probability that the observation X_1 is *one of* the k nearest neighbors to X_2 given that X_2 is *one of* the k nearest neighbors of X_1?' Write

$$\bar{p}_{1,k} = \frac{1}{k^2}\sum_{r=1}^{k}\sum_{s=1}^{k} p_1(r, s); \quad \bar{p}_{1,k}' = \lim_{n\to\infty} n\bar{p}_{1,k}.$$

Then an expression for the probability in the above query is easily seen to be

$$\left\{\sum_{r=1}^{k}\sum_{s=1}^{k} p_1(r, s)\right\}\Big/\binom{k}{n-1} = k(n-1)\bar{p}_{1,k}.$$

TABLE 1
Values of C_d for selected d

d	C_d
1	0·333
2	0·378
3	0·407
4	0·427
5	0·442
10	0·479
∞	0·500

TABLE 2
Values of $p'_1(r, s)$ for selected r, s and d

	$r=1$	$r=2$	$r=3$	
s				d
	0·667	0·222	0·074	1
	0·621	0·235	0·089	2
	0·593	0·241	0·098	3
1	0·572	0·245	0·105	4
	0·558	0·247	0·109	5
	0·521	0·250	0·119	10
	0·500	0·250	0·125	∞
		0·370	0·222	1
		0·329	0·215	2
		0·306	0·210	3
2		0·292	0·205	4
		0·283	0·202	5
		0·261	0·193	10
		0·250	0·188	∞
			0·272	1
			0·243	2
			0·228	3
3			0·218	4
			0·211	5
			0·196	10
			0·188	∞

The asymptotic value $k\bar{p}'_{1,k}$ has an extremely simple form for $d=\infty$. Applying Theorem 4.1, we have

$$\lim_{d\to\infty} k\bar{p}'_{1,k} = \frac{1}{k}\sum_{r=1}^{k}\sum_{s=1}^{k}\frac{\Gamma(r+s-1)}{\Gamma(r)\Gamma(s)}2^{1-r-s}.$$

Some algebraic manipulations yield the following equivalent form for $k>1$:

$$\begin{aligned}\lim_{d\to\infty} k\bar{p}'_{1,k} &= 1-\frac{1}{2k}\prod_{j=1}^{k-1}\left(1+\frac{1}{2j}\right)\\ &= 1-\Gamma(2k)/(k2^{2k-1}\Gamma(k)^2)\\ &\doteq 1-1/\sqrt{\pi k} \text{ by Stirling's formula.}\end{aligned}$$

Even for k as small as 3 the approximation is excellent—the value obtained is 0·674 whereas the exact figure to three places is 0·688.

Clearly $\lim_{k\to\infty}\lim_{d\to\infty} k\bar{p}'_{1,k}=1$. Theorem 4.2 below shows that the same limit (in k) in fact holds in all finite-dimensional Euclidean spaces.

Theorem 4.2. For all $d<\infty$, $\lim_{k\to\infty} k\bar{p}'_{1,k}=1$.

Proof. Regard X_1 and X_2 as fixed and let M_1, M_2 and M_3 be the numbers of points X_i, $i>2$, falling into $S-\bar{S}$, $\bar{S}-S$ and $\bar{S}\cap S$ respectively. For n large and X_1, X_2 close, M_1, M_2, M_3 can be approximated by independent Poisson variables N_1, N_2, N_3 having respective parameters proportional to C_d, C_d and $1-2C_d$. Conditional on $N_1+N_2+N_3=n_0$, $(N_1, N_2, N_3)\sim$ multinomial$(n_0; C_1, C_d, 1-2C_d)$, and from the result of Theorem 2.1 we can write

$$p_1'(r,s)=(1-C_d)\sum_{n_0=0}^{\infty} P(N_1+N_3=r', N_2+N_3=s' \mid N_1+N_2+N_3=n_0).$$

(This result should in fact serve to motivate Theorem 2.1.) Hence

$$k\bar{p}_{1,k}'=\frac{1}{k}(1-C_d)\sum_{n_0=0}^{2k-2} P(\max(N_1+N_3, N_2+N_3)<k \mid N_1+N_2+N_3=n_0).$$

For k large it will be shown that all but a negligible fraction of the terms in the above sum are near either 0 or 1. Note that given $N_1+N_2+N_3=n_0$, N_1+N_3 and N_2+N_3 are positively correlated binomial $(n_0, 1-C_d)$ variables. Choose $\xi\in(\frac{1}{2}, 1)$ arbitrarily. Then for $n_0\leqq k/(1-C_d)-k^{\xi}$ we have

$$\begin{aligned}
&P(\max(N_1+N_3, N_2+N_3)<k \mid N_1+N_2+N_3=n_0)\\
&\quad\geqq P^2(N_1+N_3<k \mid N_1+N_2+N_3=n_0)\\
&\quad\geqq\left[1-\frac{n_0(1-C_d)C_d}{(k-n_0(1-C_d))^2}\right]^2 \text{ using Chebyscheff's inequality;}\\
&\quad\geqq\left[1-\frac{C_d}{(1-C_d)^2}k^{1-2\xi}\right]^2;
\end{aligned}$$

while for $n_0>k/(1-C_d)+k^{\xi}$,

$$\begin{aligned}
&P(\max(N_1+N_3, N_2+N_3)<k \mid N_1+N_2+N_3=n_0)\\
&\quad\leqq P(N_1+N_3<k \mid N_1+N_2+N_3=n_0)\\
&\quad\leqq\frac{n_0(1-C_d)C_d}{(n_0(1-C_d)-k)^2} \text{ again using Chebyscheff;}\\
&\quad\leqq\frac{2}{(1-C_d)^2}k^{1-2\xi}.
\end{aligned}$$

On applying these bounds to the given expression for $k\bar{p}_{1,k}'$, the result follows readily.

4.2. *Neighbor-sharing.* Taking the limit of (3.1) as $d\to\infty$ evokes a surprising result which simplifies $p_2'(r,s)$ completely. The following facts are needed

to produce it:

(i) $(d-1)\dfrac{K_{d-1}}{K_d} \sim (d^3/2\pi)^{\frac{1}{2}}$;

(ii) $0 \leqq Q(\rho, \theta) \leqq D^d(\rho, \theta) \leqq 1$ in A''',

(iii) $Q(\rho, \theta) \to 0$ uniformly in $A'''_{1,1}$ as $d \to \infty$.

The first result is obtainable from Stirling's approximation; the second follows from the geometrical interpretations of Q and D given in Section 3. As for (iii), it is evident from Figure 3 that for any d, $Q(\rho, \theta)$ is never larger for $(\rho, \theta) \in A'''_{1,1}$ than its value at $\theta = \pi/3$, $\rho = 1$, where it equals $1 - 2C_d$. Then (iii) follows from the proof of Lemma 4.1.

The inequality in (ii) implies that the integrand of (3.1) is bounded in absolute value by $|\rho^{d-1} \sin^{d-2} \theta|$. Since $|\rho \sin \theta| < \sqrt{3}/2 < 1$ for all θ and ρ in $A'''_{0,1}$ and $A'''_{0,0}$, the integrals over these regions may be discarded in the limit, as the growth in (3.1) indicated by (i) is dominated by the uniformly bounded geometrical decay of the integrands. Furthermore, by (iii), all terms with $\alpha = \beta = 1$, $l > 0$ become negligible as d tends to ∞. Hence only one term from the sums in (3.1) remains, so that $p'_2(r, s; 0)$ may be replaced in the limit by

$$\lim_{d\to\infty} (d^3/2\pi)^{\frac{1}{2}} \frac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)} \int_0^{\pi/3} \sin^{d-2}\theta \int_1^{2\cos\theta} D^{d(s-1)}(\rho, \theta) \cdot \{1 + D^d(\rho, \theta)\}^{-(r+s)} \rho^{d-1}\, d\rho\, d\theta. \tag{4.1}$$

To evaluate (4.1) put $\rho = \rho(v) = 2\cos\theta - v/d$. Then for given θ, $D^d(\rho(v), \theta) \to \exp(-v\cos\theta)$ and $\rho^{d-1}(v) - (2\cos\theta)^{d-1} \exp(-v/2\cos\theta) \to 0$ as $d \to \infty$, with both limits approached uniformly in v. Thus (4.1) is equivalent to

$$\lim_{d\to\infty} (d/2\pi)^{\frac{1}{2}} \frac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)} \int_0^{\pi/3} \sin^{d-2}\theta (2\cos\theta)^{d-1} \int_0^{\infty} \exp\left[-\{(s-1)\cos\theta + \tfrac{1}{2}\sec\theta\}v\right] \cdot \{1 + \exp(-v\cos\theta)\}^{-(r+s)}\, dv\, d\theta. \tag{4.2}$$

The 'action' in the above expression occurs at $\theta = \pi/4$ since $(2\sin\theta\cos\theta)^d$ is constant and equal to 1 there but decreases geometrically otherwise. Since the inner integral is continuous in θ and does not involve d, θ may be replaced by $\pi/4$ with no effect on the limit. This gives

$$\int_0^{\infty} \exp(-sv/\sqrt{2})(1 + \exp(-v/\sqrt{2}))^{-(r+s)}\, dv$$

for the inner integral in (4.2). Letting $t = \exp(-v/\sqrt{2})/(1 + \exp(-v/\sqrt{2}))$

reduces this to

$$\sqrt{2}\int_0^{\frac{1}{2}} t^{s-1}(1-t)^{r-1}\,dt. \tag{4.3}$$

While this in itself has no closed form, one may be obtained by now including the contribution of the term $p_2'(r, s; 1)$. By symmetry this term will be the same for the outer integral and equal to

$$\sqrt{2}\int_0^{\frac{1}{2}} t_1^{r-1}(1-t_1)^{s-1}\,dt_1 \tag{4.4}$$

for the inner integral. The substitution $t = 1 - t_1$ allows (4.3) and (4.4) to be combined into the single integral

$$\sqrt{2}\int_0^1 t^{s-1}(1-t)^{r-1}\,dt = \sqrt{2}\frac{\Gamma(r)\Gamma(s)}{\Gamma(r+s)}.$$

Thus we have reduced to

$$\begin{aligned} p_2'(r, s) &\sim (d/2\pi)^{\frac{1}{2}}\frac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)}\int_0^{\pi/3}\sin^{d-2}\theta(2\cos\theta)^{d-1}\left\{2^{\frac{1}{2}}\frac{\Gamma(r)\Gamma(s)}{\Gamma(r+s)}\right\}d\theta \\ &= (d/\pi)^{\frac{1}{2}}\int_0^{\pi/3}\sin^{d-2}\theta(2\cos\theta)^{d-1}\,d\theta. \end{aligned}$$

The upper limit may be replaced by $\pi/2$ without effect, by the argument following (4.2). Then letting $u = \sin^2\theta$ gives finally

$$\begin{aligned} p_2'(r, s) &\sim (d/\pi)^{\frac{1}{2}}2^{d-2}\int_0^1 u^{(d-3)/2}(1-u)^{(d-2)/2}\,du \\ &\sim (d/\pi)^{\frac{1}{2}}2^{d-2}\Gamma((d-1)/2)\Gamma(d/2)/\Gamma(d-1/2) \\ &\to 1. \end{aligned}$$

Thus we have proven the following result.

Theorem 4.3. For all positive integers r and s, $\lim_{d\to\infty} p_2'(r, s)$ exists and equals 1.

Note that from Theorem 4.3, $p_2(r, s) = (n-2)P(NN_1(r) = NN_2(s) = X_3) \doteq 1/n$ when d and n are large, whereas $P(NN_1(r) = X_3) = P(NN_2(s) = X_3) = 1/(n-1)$; thus the implication of Theorem 4.3 is that in high-dimensional spaces the identities of the k-nearest neighbors of one sample point in a Poisson ensemble are approximately independent of those of any other randomly selected sample point.

We may also be interested in the probability that two points, say X_1 and X_2, have *any* of their respective k nearest neighbors in common for specified k.

This corresponds to the mutual neighbor problem treated in and prior to Theorem 4.2. Writing analogously

$$\bar{p}_{2,k} = \frac{1}{k^2} \sum_{r=1}^{k} \sum_{s=1}^{k} p_2(r, s),$$

$$\bar{p}'_{2,k} = \lim_{n \to \infty} n\bar{p}_{2,k},$$

we have $P(X_1 \text{ and } X_2 \text{ share a } k\text{-nearest neighbor}) = k^2\bar{p}_{2,k}$. There are simple closed forms for $\bar{p}'_{2,k}$ for both $d = 1$ and $d = \infty$. From Theorem 3.2, we have for $d = 1$

$$\bar{p}'_{2,k} = \frac{1}{k^2}\left[\sum_{r=1}^{k} (\tfrac{3}{2} - 2^{1-r}) + 2\sum_{r=2}^{k}\sum_{s=1}^{r-1}\left(1 - \binom{r-1}{s-1}2^{-r}\right)\right] = 1 - \frac{1}{2k};$$

while Theorem 4.3 trivially yields $\lim_{d\to\infty} \bar{p}'_{2,k} = 1$ for all k. Unless k is very small the difference between the univariate and infinite-dimensional values is again seen to be slight.

In order to assess the rates of convergence of the asymptotic neighbor-sharing values to their infinite-dimensional limits, a numerical integration

TABLE 3
Values of $p'_2(r, s)$ for selected r, s and d

	$r = 1$	$r = 2$	$r = 3$	
s				d
	0·50	0·75	0·88	1
	0·63	0·78	0·86	2
	0·71	0·81	0·87	3
1	0·76	0·84	0·89	4
	0·80	0·87	0·91	5
	0·92	0·96	0·96	10
	1·00	1·00	1·00	∞
		1·00	0·75	1
		0·96	0·88	2
		0·96	0·92	3
2		0·97	0·95	4
		0·97	0·96	5
		0·98	0·98	10
		1·00	1·00	∞
			1·25	1
			1·01	2
			0·99	3
3			0·99	4
			0·99	5
			0·99	10
			1·00	∞

program was used to evaluate $p_2'(r, s)$ for $r, s \leqq 3$ and various $d > 1$. Computations were performed on an IBM 3033 using the IMSL library routine DCADRE. Table 3 provides the results along with values for $d = 1$, computed easily from Theorem 3.2.

Table 2 and 3 indicate that near-neighbor geometric probabilities, when the observations are locally Poisson-like, vary slowly with d in a smooth, frequently monotone fashion, with the infinite-dimensional values providing good approximations for quite low-dimensional situations. Parallel results have been found for nearest neighbor goodness-of-fit tests (Schilling (1983)).

5. Related work and additional results

As mentioned previously, the probability that an arbitrary point in a Poisson ensemble is the rth nearest neighbor of its own sth nearest neighbor was first obtained by Clark and Evans (1955) in the two-dimensional case for $r = s = 1$ and later generalized, incorrectly, to $r = s > 1$, by Clark (1956). Correct values for the latter case were provided by Dacey (1969), and the generalization to d dimensions with r and s not necessarily equal was obtained by Cox (1981).

Schwarz and Tversky (1980) defined the R value of any member of a list of elements as its rank in the proximity order from its own nearest neighbor and obtained the distribution of R under several models. They showed that this distribution approaches a geometric distribution for the model treated in the preceding—i.e., locally approximately Poisson samples in Euclidean d-space—and noted the stability as d changes. Note that their results yield the values of $p_1'(1, s)$ for $s = 1, 2, 3, \cdots$.

Pickard (1982) considered mutual nearest neighbors for $r = s = 1$ in Poisson processes and obtained their frequency of occurrence ($p_1'(1, 1)$) and the probability distribution for the relative distance to the next closest point.

Several authors have studied the distribution of the number of points N in a Poisson process which claim an arbitrary point as a nearest neighbor. The earliest of these were again Clark and Evans (1955), who provided only Monte Carlo results in $\mathbb{R}^2$. Roberts (1969) found bounds for the individual probabilities for N, and in the process obtained the values of $p_1'(1, s)$ for $s = 1, 2, 3, \cdots$ in two dimensions. Roberts also furnished Monte Carlo results in two and three dimensions.

Recently, Newman et al. (1983) have studied expressions for the d-dimensional distributions of N, which appear to be intractable, and determined the corresponding large-dimensional limits in several Poisson-type models. The distribution of N as $d \to \infty$ is found to approach a Poisson distribution with parameter 1 in most of the cases considered.

The results given for neighbor-sharing (Theorems 3.1 and 4.3) can be used

to shed some light on the finite-dimensional distributions of N as well: let N_n be the number of points claiming X_1 as their nearest neighbor when $X_1, \cdots, X_n$ are distributed as before. As $n \to \infty$ the distribution of N_n will approach that of N defined above. Applying inclusion–exclusion and exchangeability we have

$$P(N_n \geqq 1) \geqq \sum_{i=2}^{n} P(NN_i(1) = X_1) - \sum\sum_{i \neq j} P(NN_i(1) = NN_j(1) = X_1)$$

$$= 1 - \binom{n-1}{2} \frac{p_2(1, 1)}{n-2},$$

$$P(N_n \geqq 2) \leqq \sum\sum_{i \neq j} P(NN_i(1) = NN_j(1) = X_1) = \binom{n-1}{2} \frac{p_2(1, 1)}{n-2};$$

combining these results with the formula $P(N_n = 1) = P(N_n \geqq 1) - P(N_n \geqq 2)$ yields the following simple asympotic bounds on $P(N_n = k)$ for $k = 0, 1, 2$:

$$\lim_{n \to \infty} P(N_n = 0) \leqq \tfrac{1}{2} p_2'(1, 1); \tag{5.1}$$

$$\lim_{n \to \infty} P(N_n = 1) \geqq 1 - p_2'(1, 1); \tag{5.2}$$

$$\lim_{n \to \infty} P(N_n \geqq 2) \leqq \tfrac{1}{2} p_2'(1, 1). \tag{5.3}$$

For $d = 1$ these bounds are achieved exactly since N_n is always $\leqq 2$ with probability 1.

The distribution of N_n was estimated for $d = 3$ and 5 via a Monte Carlo experiment involving in each case $n = 10\,000$ points uniformly distributed on the 'wrapped-around' d-cube, or torus, in order to eliminate boundary effects. The results are shown in Table 4 along with the values of the bounds given

TABLE 4
Monte Carlo and theoretical values, bounds for $P(N_n = k)$*

	$k=0$	1	2	3	4	$\geqq 2$
d						
1	0·25 (0·25)	0·50 (0·50)	0·25			0·25 (0·25)
3	0·30 (0·35)	0·44 (0·29)	0·21	0·04		0·26 (0·35)
5	0·33 (0·40)	0·41 (0·20)	0·20	0·05	0·01	0·26 (0·40)
∞	0·37 (0·50)	0·37 (0·00)	0·18	0·06	0·02	0·26 (0·50)

* Results for $d = 3$ and 5 are based on $n = 10\,000$ points uniformly distributed on a d-dimensional torus; values for $d = 1$ and ∞ are exact for $n = \infty$. Numbers in parentheses are upper/lower bounds obtained from (5.1)–(5.3).

above. Also included are the theoretical values for $d = 1$ and $d = \infty$, the latter obtained from the aforementioned result of Newman et al. (1983) on the limiting Poisson distribution of N_n. Once again, the stability across dimension is notable; it is particularly striking for $P(N_n \geqq 2)$.

Acknowledgement

The careful reading and helpful comments of the referee are greatly appreciated.

References

CLARK, P. J. (1956) Grouping in spatial distributions. *Science* **123,** 373–374.

CLARK, P. J. AND EVANS, F. C. (1955) On some aspects of spatial pattern in biological populations. *Science* **121,** 397–398.

COX, T. F. (1976) The robust estimation of the density of a forest stand using a new conditioned distance method. *Biometrika* **63,** 493–499.

COX, T. F. (1981) Reflexive nearest neighbours. *Biometrics* **37,** 367–369.

COX, T. F. AND LEWIS, T. (1976) A conditioned distance ratio method for analyzing spatial patterns. *Biometrika* **73,** 483–491.

DACEY, M. F. (1969) Proportion of reflexive *n*th order neighbours in a spatial distribution. *Geographical Analysis* **1,** 385–388.

DIGGLE, P. J. (1975) Robust density estimation using distance methods. *Biometrika* **62,** 39–48.

NEWMAN, C. M., RINOTT, Y. AND TVERSKY, A. (1983) Nearest neighbors and Voronoi regions in certain point processes. *Adv. Appl. Prob.* **15,** 726–751.

PICKARD, D. K. (1982) Isolated nearest neighbors. *J. Appl. Prob.* **19,** 444–449.

ROBERTS, F. D. K. (1969) Nearest neighbours in a Poisson ensemble. *Biometrika* **56,** 401–406.

SCHILLING, M. F. (1983) An infinite-dimensional approximation for nearest neighbor goodness of fit tests. *Ann. Statist.* **11,** 13–24.

SCHILLING, M. F. (1986) Multivariate two-sample tests based on nearest neighbors. *J. Amer. Statist. Assoc.* To appear.

SCHWARZ, G. AND TVERSKY, A. (1980) On the reciprocity of proximity relations. *J. Math. Psychol.* **22,** 157–175.