TESTING RANDOMNESS OF SPATIAL POINT PATTERNS
WITH THE RIPLY STATISTIC

GABRIEL LANG and ERIC MARCON

Abstract. Aggregation patterns are often visually detected in sets of location data. These clusters may be the result of interesting dynamics or the effect of pure randomness. We build an asymptotically Gaussian test for the hypothesis of randomness corresponding to a homogeneous Poisson point process. We first compute the exact first and second moment of the Ripley $K$-statistic under the homogeneous Poisson point process model. Then we prove the asymptotic normality of a vector of such statistics for different scales and compute its covariance matrix. From these results, we derive a test statistic that is chi-square distributed. By a Monte-Carlo study, we check that the test is numerically tractable even for large data sets and also correct when only a hundred of points are observed.

Mathematics Subject Classification. 60G55, 60F05, 62F03.

Received June 22, 2011. Revised November 5, 2012.

INTRODUCTION

Analysis of point patterns is relevant in many sciences: cell biology, ecology or spatial economics. The observation of clusters in point locations is considered as a hint for non observable dynamics. For example the clustering of tree locations in a forest may come from better soil conditions or from spreading of seeds of a same mature individual; but clusters are also observed in random distribution as a Poisson point process sample. It is therefore essential to distinguish between clusters resulting from relevant interactions or from complete randomness. Ripley function \cite{20,21} is a widely used tool to quantify the structure of point patterns, especially in ecology, and is well referenced in handbooks \cite{7,8,15,18,23,25}. Up to a renormalization by the intensity of the process, this statistic denoted here $\hat{K}(r)$ estimates the expectation $K(r)$ of the number of neighbors at distance less than $r$ of a point in the sample. The observed $\hat{K}(r)$ is compared to the value of $K(r)$ for a homogeneous Poisson point process chosen as a null hypothesis: the Poisson point process is characterized by an independence of point locations, modelling an absence of interactions between individuals in ecosystems. In this case $K(r)$ is simply the mean number of points in a ball of radius $r$ divided by the intensity, that is $\pi r^2$. If $\hat{K}(r)$ is significantly larger than $\pi r^2$ (respectively smaller), the process is considered as aggregated (respectively over-dispersed) at distance $r$.

Keywords and phrases. Central limit theorem, goodness-of-fit test, Hölling decomposition, $K$-function, point pattern, Poisson process, $U$-statistic.

1 AgroParisTech, UMR 518 Mathématique et Informatique Appliquées, 19 avenue du Maine, 75732 Paris Cedex 15, France. gabriel.lang@agroparistech.fr
2 AgroParisTech, UMR 745 Ecologie des Forêts de Guyane, Campus agronomique BP 316, 97379 Kourou Cedex, France. eric.marcon@agroparistech.fr

Article published by EDP Sciences © EDP Sciences, SMAI 2013
In order to decide if the difference is statistically significant, we build a test of the Poisson process hypothesis; we need information on the distribution of \( \hat{K}(r) \) for this process. But even the variance is not known and statistical methods generally rely on Monte-Carlo simulations. Ripley [22] used them to get confidence intervals. Starting from previous results [24], he also gave critical values for the \( L \) function, a normalized version of \( K \) introduced by [4]. These critical values are valid asymptotically, for a large number of points but low intensity, so that both edge effects and point-pair dependence can be neglected. Further computations of confidence interval bands based on simulation have been proposed in [16] and corrected in [5]. But the simulation is a practical issue for large point patterns, because computation time is roughly proportional to the square of the number of points (one has to calculate the distances between all pairs of points) multiplied by the number of simulations.

We propose here to compute the exact variance of the Ripley statistic. Ward and Ferrandino [30] studied this variance. But they ignored that point pairs are not independent even though points are (Eq. A8, p. 235), thus their derivation of the variance of \( \hat{K}(r) \) was erroneous. A rigorous computation of the variance has been carried out in [27] for a independent sample of uniform variables on the unit square, that is for the Poisson process conditioned by a fixed number of points; for the Poisson process, we compute the exact covariance, considering the Ripley statistic as a \( U \)-statistic as remarked in [22] and using the Höffding decomposition. As the variance is not enough to build a test, we study the distribution of the statistic. We prove its asymptotic normality as the size of the observation window grows. It is then easy to build an asymptotically Gaussian test.

Another concern is to test simultaneously the aggregation/dispersion at different scales. This is rarely correctly achieved in practical computations with Monte-Carlo simulations. The confidence bands or test rejection zone are often determined without taking the dependence between the numbers of neighbors at different scales into account. Heinrich [14] proposed the first multiscale goodness-of-fit tests based on the Ripley function for Poisson processes. He considered a set of scales \( (r_1, \ldots, r_d) \), computed the covariance matrix of the estimates \( \hat{K}(r_i) \) and proved the asymptotic normality for the vector \( (\hat{K}(r_1), \ldots, \hat{K}(r_d)) \). He derived Kolmogorov–Smirnov, Cramer von Mises and chi-square goodness-of-fit tests from these results. Grabarnik and Chiu [10] proposed a similar test based on the \( k \) first neighbors of a point, that is more difficult to use in practice because the number of neighbours is an additional parameter to tune. The test that we propose is very similar to the chi-square test of Heinrich; the only difference lies in the correction of the bias due to edge effects. Our method of correction allows us to compute the exact value of the covariance matrix and not only its asymptotical value, as for the Heinrich test. This is a major improvement in practice because the level of the test is very sensitive to approximations in the computation of the covariance matrix. A similar exact computation of the variance matrix is untractable for the Heinrich test: only an estimation method based on subsampling of the data may be proposed as done in [12] for the inhomogeneous case.

The paper is built as follows: Section 1 introduces the precise definition of \( K(r) \) and the current definition of \( \hat{K}(r) \). In Section 2, after the definition of our statistics (no edge effects correction, known or unknown intensity), we list the main results of the paper: exact bias due to the edge effects and exact variance of \( \hat{K}(r) \) for a homogeneous Poisson process with known or unknown intensity; covariance between \( \hat{K}(r) \) and \( \hat{K}(r') \) for two different distances \( r \) and \( r' \). The main theorem contains the convergence of the vector \( (\hat{K}(r_1), \ldots, \hat{K}(r_d)) \) to a Gaussian distribution with explicit covariance in the following asymptotic framework: data from the same process are collected on growing squares of observation. These results allow a simple, multiscale and efficient test procedure of the Poisson process assumption. Section 3 provides a Monte-Carlo comparison of the tests and Section 4 gives our conclusions. The last section contains the proofs.

1. Definition of the Ripley \( K \)-function

We recall the characterizations of the dependence of the locations for a general point process \( X \) over \( \mathbb{R}^2 \). We refer to the presentation of [18].
1.1. Definitions

For a point process $X$, define the point process $X^{(2)}$ on $\mathbb{R}^2 \times \mathbb{R}^2$ of all the couples of two different points of the original process. The intensity of this new process gives information on the simultaneous presence of points in the original process. Denote $\rho^{(2)}(x, y)$ its density (called the second-order product density). The Poisson process of density $\rho(x)$ is such that $\rho^{(2)}(x, y) = \rho(x)\rho(y)$.

The Ripley statistic is a way to estimate the density $\rho^{(2)}(x, y)$. Precisely it is an estimate of the integral on test sets of the ratio $g(x, y) = \rho^{(2)}(x, y)/\rho(x)\rho(y)$. The function $g(x, y)$ characterizes the fact that the points $x$ and $y$ appear simultaneously in the samples of $X$. If $g(x, y) = 1$, the points appear independently. If $g(x, y) < 1$, they tend to exclude each other; if $g(x, y) > 1$, they appear more frequently together.

We assume the translation invariance of the point process: $g(x, y) = g(x - y)$. In order to estimate the function $g$, we define its integral as the set function $K$. Let $A$ be a Borel set:

$$K(A) = \int_A g(x)dx.$$

If we also assume that the point process is isotropic, we define the Ripley $K$-function as

$$K(r) = K(B(x, r)),$$

where $B(x, r)$ is the closed ball with center $x$ and radius $r$. The translation invariance implies that $K(B(x, r))$ does not depend on $x$. For example, if the process is a Poisson process then $g(x) = 1$ and $K(r) = \pi r^2$. We define the Ripley statistic that estimates the $K$-function. Let $A$ be a bounded Borel set of the plane $\mathbb{R}^2$, $m$ the Lebesgue measure, $\hat{\rho}(x)$ an estimator of the local intensity of the process and $\{\cdot\}$ denotes the indicator function of a set; for a realization $S$ of the point process $X$, $S = \{X_1, \ldots, X_N\}$, a general form of the Ripley statistic is

$$\hat{K}_A(r) = \frac{1}{m(A)} \sum_{X_i \neq X_j \in S} \frac{\{d(X_i, X_j) \leq r\}}{\hat{\rho}(X_i) \hat{\rho}(X_j)}.$$

Note that estimator $\hat{K}_A(r)$ refers to a preexistent estimator of the local intensity $\hat{\rho}(x)$, to make it unsensible to the inhomogeneity of the intensity. In practice, $\hat{\rho}(x)$ is a local kernel estimator, that uses the only available local information contained in the locations of neighbors in a fixed ball around the considered point of the sample. This estimator is then very much dependent of the indicator function in the numerator, because they are based on the same information. It cannot be considered as a constant close to the true value of the local intensity. This is why we do not manage to compute the exact value of the two first moments of this statistic. We only address the problem of testing homogeneous Poisson processes and the Ripley statistic has simplified expressions given below.

2. Main results

This section presents the theoretical results on the Ripley statistic and the resulting test.

2.1. Definitions and assumptions

Throughout the paper, we refer to the expectation $e_{r,n}$, the centered indicator function $h$ and its conditional expectation $h_1$. We gather here these definitions.

Let $n$ be an integer; $A_n$ denotes the square $[0, n]^2$; $U$ is a random location in $A_n$ with an uniform random distribution; its density is $1/n^2$ with respect to the Lebesgue measure $d\xi_1 d\xi_2$ over $A_n$. $V$ is an independent copy of $U$. We denote $d(x, y)$ the Euclidean distance between $x$ and $y$ in the plane. We define $e_{r,n} = E(\{d(U, V) \leq r\})$, $h(x, y, r) = I\{d(x, y) \leq r\} - e_{r,n}$, and $h_1(x, r) = E(h(U, V, r)|V = x)$.

We assume that $X$ is a homogeneous Poisson process on $\mathbb{R}^2$ with intensity $\rho$. We consider that the data are available on the square $A_n$. The setting of the asymptotics was suggested by practitioners in ecological modeling.
and forestry: the accumulation of tree location data comes from measuring wider and wider sets of land and the inter-tree distances \( r \) do not vary with \( n \).

Let \( N \) denote the random number of observed points and \( S = \{X_1, \ldots, X_N\} \) denote the sample of observed points. We consider two cases:

1. If the intensity \( \rho \) is known, the Ripley statistic is expressed as
   \[
   \hat{K}_{1,n}(r) = \frac{1}{n^2 \rho^2} \sum_{X_i \neq X_j \in S} I\{d(X_i, X_j) \leq r\}.
   \]

2. If the intensity \( \rho \) is unknown, we use the unbiased estimator \( \hat{\rho}^2 = N(N-1)/n^4 \) (see [26]) and define
   \[
   \hat{K}_{2,n}(r) = \frac{n^2}{N(N-1)} \sum_{X_i \neq X_j \in S} I\{d(X_i, X_j) \leq r\}.
   \]

### 2.2. Bias

It is known that a large number of neighbors of the points located near the edges of \( A_n \) may lie outside \( A_n \) causing a bias in the estimation. We compute the bias due to this edge effect.

**Proposition 2.1.** Assume that \( r/n < 1/2 \).

\[
\begin{align*}
\mathbb{E}\hat{K}_{1,n}(r) - K(r) &= r^2 \left( \frac{8r}{3n} + \frac{r^2}{2n^2} \right), \\
\mathbb{E}\hat{K}_{2,n}(r) - K(r) &= r^2 \left( -\frac{8r}{3n} + \frac{r^2}{2n^2} \right) - r^2 \left( 1 + \rho n^2 \right) e^{-\rho n^2} \left( \pi - \frac{8r}{3n} + \frac{r^2}{2n^2} \right).
\end{align*}
\]

**Notes.**

- The assumption \( r/n < 1/2 \) means that at least some balls of radius \( r \) are included in the square \( A_n \).
- The additional term for \( \hat{K}_{2,n} \) corresponds to the probability to draw a sample with zero or one point in the square. This term gives a zero contribution as soon as the mean number of points \( \rho n^2 \) is larger than 20.
- The proof may be adapted for a convex polygon of perimeter \( L_n \) to compute the first order term of the bias; for \( u = 1 \) or 2:
  \[
  \mathbb{E}\hat{K}_{u,n}(r) - K(r) = -\frac{2L_n^2 r}{3} + O\left( \frac{r^2}{n^2} \right).
  \]

### 2.3. Variance

We compute the covariance matrix of \( \hat{K}_{u,n}(r) \) for \( u = 1 \) or 2. We get an exact computation for the variance, that can be used for any value of \( n \).

**Proposition 2.2.** For \( 0 < r < r' \),

\[
\begin{align*}
\text{var}(\hat{K}_{1,n}(r)) &= \frac{2\epsilon_{r,n}}{\rho^2} + \frac{4n^2 \epsilon_{r,n}^2}{\rho} + \frac{4n^2}{\rho} \mathbb{E}h_1^2(U, r), \\
\text{cov}(\hat{K}_{1,n}(r), \hat{K}_{1,n}(r')) &= \frac{2\epsilon_{r,n}}{\rho^2} + \frac{4n^2 \epsilon_{r,n}^2}{\rho} + \frac{4n^2}{\rho} \mathbb{C}ov\left(h_1(U, r'), h_1(U, r)\right).
\end{align*}
\]
2.3. Theorem. Let the parameter we have to use an estimate of $\rho$ vector. Let the covariance matrix $\Sigma$.

2.4. Central Limit Theorem

Notes.

The variances of both estimators are exact and can be computed with any precision, as inverse moments of the Poisson variable correspond to fast converging series.

The covariances are not explicit because the terms $\text{cov}(h_1^2(U, r'), h_1^2(U, r))$ involve parts that have to be numerically integrated.

The leading terms of the variances of $K_{1,n}(r)$ and $K_{2,n}(r)$ as $n$ tends to infinity are $2\pi r^2/n^2 \rho^2 + 4\pi r^4/n^2 \rho$ and $2\pi r^2/n^2 \rho^2$.

2.4. Central Limit Theorem

We show that a normalized vector of Ripley statistics for different $r$ converges in distribution to a normal vector. Let $N(0, \Sigma)$ denote the Gaussian multivariate centred distribution with covariance matrix $\Sigma$.

**Theorem 2.3.** Let $d$ be an integer, $0 < r_1 < \ldots < r_d$ a set of reals and for $u = 1$ or 2, define $K_{u,n} = (\hat{K}_{u,n}(r_1), \ldots, \hat{K}_{u,n}(r_d))$. Then $n\sqrt{\pi}(K_{u,n} - \pi(r_1^2, \ldots, r_d^2))$ converges in distribution to $N(0, \Sigma)$ as $n$ tends to infinity, where for $s$ and $t$ in $\{1, \ldots, d\}$

- if $u = 1$, $\Sigma_{s,t} = \frac{2\pi(r_s^2 \wedge r_t^2)}{\rho} + 4\pi^2 r_s^2 r_t^2$
- if $u = 2$, $\Sigma_{s,t} = \frac{2\pi(r_s^2 \wedge r_t^2)}{\rho}$.

**Note.** The first term of the variance corresponds to a case where the couples of points are independent from each others; this was used as an approximation without proof in [30]; our work proves that the actual variance and limit process are different in the first case and that the approximation holds only in the second case.

2.5. Applications to test statistics

From Theorem 2.3, we deduce that $T_u = \Sigma^{-1/2} K_{u,n}$ is asymptotically $N(0, I_d)$ distributed. For the hypothesis

$H_0$: $X$ is a homogeneous Poisson process of intensity $\rho$

we use $T^2 = \|T_u\|_2^2$ as a test statistic with rejection zone for the level $\alpha$:

$T^2 > \chi^2_\alpha(d)$.

where $\chi^2_\alpha(d)$ is the $(1 - \alpha)$-quantile of the $\chi^2(d)$ distribution.

**Note.** the covariance matrix $\Sigma$ depends on the intensity parameter $\rho$, so that in the case of the unknown parameter we have to use an estimate of $\rho$ in the formula defining $\Sigma$.  

\[
\text{var}(\hat{K}_{2,n}(r)) = 2n^4 \mathbb{E} \left( \frac{I[N > 1]}{N(N - 1)} \right) (e_{r,n} - e_{r,n}^2) = 4n^4 \mathbb{E} \left( \frac{I[N > 1](N - 2)}{N(N - 1)} \right) E h_1^2(U, r)
\]

\[
+ n^4 e^{-\rho n^2} (1 + \rho n^2) \left( 1 - e^{-\rho n^2} - \rho n^2 e^{-\rho n^2} \right) e_{r,n}^2.
\]

\[
\text{cov}(\hat{K}_{2,n}(r), \hat{K}_{2,n}(r')) = 2n^4 \mathbb{E} \left( \frac{I[N > 1]}{N(N - 1)} \right) (e_{r,n} - e_{r',n} e_{r,n})
\]

\[
+ 4n^4 \mathbb{E} \left( \frac{I[N > 1](N - 2)}{N(N - 1)} \right) \text{cov}(h_1(U, r'), h_1(U, r))
\]

\[
+ n^4 e^{-\rho n^2} (1 + \rho n^2) \left( 1 - e^{-\rho n^2} - \rho n^2 e^{-\rho n^2} \right) e_{r',n} e_{r,n},
\]

where $e_{r,n} = \frac{\pi r^2}{n^2} - \frac{8 r^3}{3 n^3} + \frac{r^4}{2 n^4}$ and $E h_1^2(U, r) = \frac{r^5}{n^5} \left( \frac{8}{3} \pi - \frac{256}{45} \right) + \frac{r^6}{n^6} \left( \frac{11}{48} \pi - \frac{56}{9} \right) + \frac{8 r^7}{3 n^7} - \frac{1 r^8}{4 n^8}$.
3. Simulations

We study the empirical variance of the proposed statistics by a Monte-Carlo simulation. Then we apply the test procedure to simulated data sets, observe the number of rejections and compare it to the level of the test.

3.1. Variance

We simulate a sample of 1000 repetitions with \( \rho = 5 \) and compare (after renormalization by \( n\sqrt{\rho} \)) the empirical variance and the exact computed variance with the limit variance for different value of \( n \) (Fig. 1). With 1000 repetitions, the oscillations of the empirical variance are still large; we will use a larger number of repetitions in the following study of the test.

The convergence of the computed variance to the limit value is not so fast and for applications with hundreds of points (corresponding in Fig. 1 to \( n < 15 \)) the distance between the variances is still large. A preliminary study, not presented here, showed that the test procedure is perturbed by any small error in the covariance matrix, as we tried simplified versions of the covariance by ignoring the contribution of points in the corner of the observation window. It is crucial to use an accurate computation of the covariance matrix to have a correct approximation of the square root inverse matrix \( \Sigma^{-1/2} \). Therefore we will use the exact variance formula instead of the asymptotic formula in the test procedure.

3.2. Test level

In the known parameter case, the computation of the test statistic \( T_1 \) is straightforward. In the unknown parameter case, the computation of the test statistic \( T_2 \) is done by replacing the unknown parameter \( \rho \) by the estimator \( N/n^2 \). We also choose to replace the expectation \( \mathbb{E}(\mathbb{I}\{N > 1\}/(N(N - 1))) \) by the observed value \( 1/(N(N - 1)) \) and \( \mathbb{E}(\mathbb{I}\{N > 1\}(N - 2)/(N(N - 1))) \) by \( (N - 2)/(N(N - 1)) \), because the dispersion of a Poisson variable is low with respect to the expectation when its parameter is large. For comparison, a chi-square test \( T_3 \) based on the unbiased Ripley estimator \( \hat{K}_{3,n} \) is given using the asymptotic variance as proposed in [14]. The correction of the bias consists in dividing the indicator function not by the constant area of the
Percentiles of rejection in Table 1 should lie in $[4.57; 5.43]$. Stars indicate values outside this confidence interval. Percentiles of rejection in Table 1 should lie in $[4.57; 5.43]$. Stars indicate values outside this confidence interval.

Concerning the choice for the range for distances $r$, there are two situations. From the theoretical point of view, all the scales are of the same interest. One may plot the statistic $K$ and choose the range where the empirical values depart from expected and investigate if the difference is significant. From the practical point of view, when observing real data, practitioners often know in advance the scale they are interested in: range from 2 to 50 meters for tree locations for example, or ten meters to one kilometer for locations of shops in a city; ...

The test output is a Bernoulli random variable with parameter $\alpha$. With a sufficient index of repetition $m$, the mean number of rejection is close to a normal variable with expectation $\alpha$ and variance $\alpha(1 - \alpha)/m$. We consider that the test works correctly when the observed frequency of rejection is in the 95% Gaussian confidence interval $[\alpha - 1.96 \sqrt{\alpha(1 - \alpha)/m}, \alpha + 1.96 \sqrt{\alpha(1 - \alpha)/m}]$. With $m = 10\ 000$ and $\alpha = 0.05$, the interval is $[0.0457; 0.0543]$. Percentiles of rejection in Table 1 should lie in $[4.57; 5.43]$. Stars indicate values outside this confidence interval. The performances of $T_1$ (known parameter $\rho$) are good except when the number of points is less than 100. The test $T_2$ (unknown parameter $\rho$) performs better than $T_1$ for small data sets. The comparison of line 5 and 6 in Table 1 shows that $T_2$ has a bad level if the distances are so small that the corresponding balls have a large probability to be void. The test $T_3$ is systematically affected by edge effects. This is due to the use of an asymptotic formula for the variance that is not sufficiently accurate even for samples with 500 points.

### 3.3. Test power against dependence

We investigate the power of the test $T_2$ against the alternative of dependent point processes. In Table 2, we simulate six Thomas cluster processes [28] and two Hardcore Strauss processes. A Thomas process is a clustered Neyman-Scott process; the germs of the clusters are drawn as a sample of a homogeneous Poisson process of intensity $\kappa$. For each germ, an inhomogeneous Poisson process is drawn with intensity measure $\mu f$, where $f$ is the density of the Gaussian two-dimensional vector centered on the germ and with independent coordinates of standard error $\sigma$. The Thomas process results from the superimposition of these inhomogeneous Poisson processes. The germs are not conserved. Note that this process is homogeneous, with resulting intensity $\rho = \kappa \mu$. Figure 2 presents a sample of these Thomas processes compared to a sample of a Poisson process of the same intensity. It shows that the visual inspection is not sufficient to distinguish between the processes, especially when the number of points is more than 250. The Hardcore Strauss process is a over-dispersed Markov process defined by a density with respect to a homogeneous Poisson process. The density of a point set is equal to zero when two points are at distance less than a constant radius $R$ and constant for other point sets.
Figure 2. Samples of processes of table 2. Expected number of points: first row 250, second row 20, third row 2000. For the processes Thomas 1 to 6, the expected size of the clusters are respectively 10, 5, 4, 2, 10 and 4.

We compare test $T_2$ with Heinrich test $T_3$ and with the Monte Carlo test $Lm$ based on the uncorrected Besag function $L(r)$ estimated in the band $r \in [0, 2.5]$ (see [11]); the test statistic is $Lm = \max |\hat{L}(r) - E(\hat{L}(r))|$ computed on this band. The value of $E(L(r))$ is estimated by a first Monte-Carlo sampling of size 10000 and the distribution of $Lm$ is estimated by a second Monte-Carlo sampling of the same size. The rejection zone corresponds to the largest values corresponding to 5% of the Monte-Carlo sample.

**Edge effects.** The first row of Table 2 shows that test $T_2$ rejects a bit less than Heinrich test $T_3$; this is mainly due to the incorrect level of the test $T_3$ as shown in the third line (28.9% rejection instead of 5% expected for the reference Poisson process). This means that edge effects are too strong for $T_3$ when the ratio $\max(r/n)$ is equal to 0.2. The test $T_2$ has a correct level, detects perfectly the large clusters of model Thomas 1 and quite well (67%) the clusters in model Thomas 2. It performs better than the $Lm$ test. The second row displays tests with lesser edge effects ($\max(r/n)$ is equal to 0.1), for a sample with very few points (20 points expected). The third line shows that the level of $T_2$ and $T_3$ are acceptable even for small samples. The power of the two tests
Table 2. Percentile of rejection over 10,000 repetitions of the test with level $\alpha = 0.05$ for dependent processes compared to Poisson processes.

<table>
<thead>
<tr>
<th>$n$ = 10</th>
<th>$T_2$</th>
<th>$T_3$</th>
<th>$Lm$</th>
</tr>
</thead>
<tbody>
<tr>
<td>250 points</td>
<td>$r = (0.5, 1, 2)$</td>
<td>Thomas 1 $(\kappa, \mu, \sigma) = (0.25, 10, 1)$</td>
<td>94.8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Thomas 2 $(\kappa, \mu, \sigma) = (0.5, 5, 1)$</td>
<td>67.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Poisson 1 $\rho = 2.5$</td>
<td>4.9</td>
</tr>
<tr>
<td>20 points</td>
<td>$r = (0.2, 0.5, 1)$</td>
<td>Thomas 3 $(\kappa, \mu, \sigma) = (0.05, 4, 1)$</td>
<td>60.8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Thomas 4 $(\kappa, \mu, \sigma) = (0.1, 2, 1)$</td>
<td>32.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Poisson 2 $\rho = 0.2$</td>
<td>6.7</td>
</tr>
<tr>
<td>2000 points</td>
<td>$r = (0.2, 0.5, 1)$</td>
<td>Thomas 5 $(\kappa, \mu, \sigma) = (2, 10, 1)$</td>
<td>72.8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Thomas 6 $(\kappa, \mu, \sigma) = (5, 4, 1)$</td>
<td>32.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Poisson 3 $\rho = 20$</td>
<td>5.0</td>
</tr>
<tr>
<td>2000 points</td>
<td>$r = (0.5, 1, 3)$</td>
<td>Thomas 6 $(\kappa, \mu, \sigma) = (5, 4, 1)$</td>
<td>42.6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Poisson 3 $\rho = 20$</td>
<td>5.3</td>
</tr>
<tr>
<td>2000 points</td>
<td>$r = (1, 2, 5)$</td>
<td>Thomas 6 $(\kappa, \mu, \sigma) = (5, 4, 1)$</td>
<td>46.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Poisson 3 $\rho = 20$</td>
<td>5.1</td>
</tr>
<tr>
<td>2000 points</td>
<td>$r = (1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, 5)$</td>
<td>Thomas 6 $(\kappa, \mu, \sigma) = (5, 4, 1)$</td>
<td>25.6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Poisson 3 $\rho = 20$</td>
<td>5.5</td>
</tr>
<tr>
<td>10,000 points</td>
<td>$r = (1, 2, 5)$</td>
<td>Thomas 6 $(\kappa, \mu, \sigma) = (5, 4, 1)$</td>
<td>69.6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Thomas 8 $(\kappa, \mu, \sigma) = (25, 4, 1)$</td>
<td>30.9</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Poisson 4 $\rho = 100$</td>
<td>5.3</td>
</tr>
<tr>
<td>69 points</td>
<td>$r = (0.1, 0.5, 2)$</td>
<td>Strauss 1 $R = 0.4$</td>
<td>44.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Poisson 5 $\rho = 0.69$</td>
<td>6.2</td>
</tr>
<tr>
<td>74 points</td>
<td>$r = (0.1, 0.5, 2)$</td>
<td>Strauss 2 $R = 0.35$</td>
<td>21.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Poisson 6 $\rho = 0.74$</td>
<td>5.4</td>
</tr>
</tbody>
</table>

are similar and quite good for a sample of Thomas 3 (clusters of expected size 4); they still detect around 30% of the samples of Thomas 4, that is much closer to a Poisson process (clusters of expected size 2).

**Comparison with the $Lm$ estimator.** In [11], the authors claim that estimators based on maximum absolute deviation as $Lm$ work better for small samples when edge effects are not corrected by the band correction, that discards the data in the band of width $r$ from the edges of the square of observation. We see in row 1 and 2 that the test $T_2$ performs better than the Monte-Carlo test $Lm$, even for small sample. This is not contradictory with the conclusions of [11], as our edge effects correction does not discard data. The simulation is also different because we do not fix the number of points as it was done in [11]. For small samples the variation of the number of points is significative and may explain the different conclusions. We think that for very small sample, test $Lm$ is advantageous because it is very easy to compute with a perfect level by construction. But for medium size sample of 250 points, $T_2$ is easier to compute, with a perfect level and better power.

**Effect of the number of points.** In row 3, we study samples with a larger number of points (2000 points expected) in the same space and with the same range of distance. First notice that is not easier to distinguish between Poisson and cluster models when the number of points increases because the clusters are forced to overlap. The performance of $T_2$ is a bit lower than for 250 points but still acceptable. $T_3$ has an incorrect level and $Lm$ has a weak power.

**Effect of the range of distance.** In row 4 and 5, we change the range of distance in $T_2$ for the same process Thomas 6. We do not compute $T_3$ because its level is worse than in row 1. We see that the performance of $T_2$ increases when the range of distance is larger with the best result for the value $r = 5$ corresponding to the maximal ratio $r/n = 1/2$. In row 3, we see that $T_2$ outperforms the $Lm$ estimator even if they are computed on a similar range of distance.
Effect of the number of distances. In row 6, we keep the same process Thomas 6 and the same range of distance but we increase the number of distances. The level of the test is correct, but the power is lower. Notice that increasing the number of distances does not necessarily increase the information (only the jumps in the function \( \hat{K} \) are informative), but it is still surprising that the performances decrease so fast. This could be a consequence of instabilities in the computation of the inverse square root of the covariance matrix as its dimension increases.

Larger set of points. In row 7, we study a small sample (100 repetitions) of processes with a larger number of points (10,000 points expected) with the largest range of distance. Comparing the three processes with cluster size equal to 4, (Thomas 3, 6 and 8), we confirm that the power decreases when the number of clusters increases. Clusters of size 10 of Thomas 7 are still well detected.

Test of over-dispersion. In row 8 and 9, we study a sample (10,000 repetitions) of hardcore Strauss processes with intensity \( \beta = 1 \) and hardcore radius equal to 0.4 and 0.35. The Kolmogorov Smirnov test \( Lm \) is considered as very powerful in this context [15]. We observe that \( Lm \) rejects perfectly the sample where \( T_2 \) and \( T_3 \) have poor performances. Here again, \( T_3 \) seems to have a better power, but its level is not correct as can be seen with Poisson simulation with the corresponding mean number of points.

3.4. Test power against inhomogeneity

In [13], Ho and Chiu propose to test inhomogeneity versus homogeneity with goodness-of-fit tests for the uniform distribution. A general study of those tests is [6]. The advantage is that the distribution is free of edge effects. As our test is also free of edge effects, we investigate here its power against inhomogeneous Poisson processes. Function \( K \) being the same for homogeneous and heterogeneous Poisson process, how could the method work for testing homogeneous Poisson versus heterogeneous? Simply because our definition of \( \hat{K} \) was adapted to the homogeneous case and has a different distribution under inhomogeneous Poisson assumptions.

We derive our models from those of [9]. In this paper, the authors consider five types of intensity functions \( s_i(x) \) for inhomogeneous Poisson processes on the segment \([0, 1]\). We simulate inhomogeneous Poisson processes on the square \([0, 1] \times [0, 1]\) with intensity 100\(s_i(x)s_i(y)\). The functions \( s_i(x) \) are:

\[
\begin{align*}
    s_1(x) &= (1 + \varepsilon) \mathbb{I}\{0 \leq x < 0.125\} + (1 - \varepsilon) \mathbb{I}\{0.125 \leq x < 0.25\} + \mathbb{I}\{0.25 \leq x \leq 1\} \\
    s_2(x) &= \frac{1}{1 + 1.27\varepsilon} \left( 1 + \varepsilon \sum_{j=1}^{11} h_j \mathbb{I}\{x < p_j\} \right) \\
    s_3(x) &= (1 - \varepsilon) \mathbb{I}\{0 \leq x \leq 1\} + \frac{\varepsilon}{0.284} \left( \sum_{j=1}^{11} g_j \left( 1 + \frac{|x - p_j|}{w_j} \right)^{-4} \right) \\
    s_4(x) &= (1 - \varepsilon) \mathbb{I}\{0 \leq x < 0.75\} + (1 + 3\varepsilon) \mathbb{I}\{0.75 \leq x \leq 1\} \\
    s_5(x) &= (1 - \varepsilon) + \varepsilon \beta x^{\beta-1}
\end{align*}
\]

Parameters \( p_j, h_j, g_j \) and \( w_j \) are constant parameters defining the different functions (their values are the same than in [9]). Parameter \( \varepsilon \) corresponds to the strength of heterogeneity within a model. Parameter \( \beta \) modifies the shape of function \( s_5 \). Functions \( s_1, s_2 \) and \( s_4 \) are step functions, function \( s_3 \) shows steep pikes and function \( s_5 \) is smooth. All functions have integral equal to 1, so that the expected number of points in the samples is 100.

The last column corresponds to the level, that is the homogeneous Poisson process with intensity 100.

The powers of the test \( T_2 \) are comparable to those of the tests proposed in [9], for the same expected number of points. They are better for models \( s_2 \) and \( s_5 \), the same for model \( s_4 \) and worse for models \( s_1 \) and \( s_3 \). Notice that the comparison can not be made rigorous, as the Poisson processes in [9] were defined on the line. The range of distance \( r \) has been lowered for Heinrich test \( T_3 \) to keep the level acceptable (for \( r = (0.1, 0.2, 0.5) \) the observed level is 55%). Even then \( T_3 \) performs worse than \( T_2 \). The \( Lm \) test performs better than \( T_2 \) for step functions. This may come from the fact that it uses a finer grid of distances \( r \).
Table 3. Percentile of rejection over 10 000 repetitions of the test with level $\alpha = 0.05$ for 12 inhomogeneous Poisson processes and the reference homogeneous Poisson process.

<table>
<thead>
<tr>
<th>model</th>
<th>s_1</th>
<th>s_1</th>
<th>s_1</th>
<th>s_2</th>
<th>s_2</th>
<th>s_3</th>
<th>s_3</th>
<th>s_4</th>
<th>s_4</th>
<th>s_5</th>
<th>s_5</th>
<th>$\rho$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\varepsilon$</td>
<td>0.5</td>
<td>0.8</td>
<td>1.0</td>
<td>0.5</td>
<td>2.0</td>
<td>0.4</td>
<td>0.6</td>
<td>0.2</td>
<td>0.4</td>
<td>1.0</td>
<td>0.6</td>
<td>2.0</td>
</tr>
<tr>
<td>$T_2$, $r = (0.1, 0.2, 0.5)$</td>
<td>18.3</td>
<td>55.8</td>
<td>84.3</td>
<td>86.5</td>
<td>100</td>
<td>20.9</td>
<td>61.7</td>
<td>90.1</td>
<td>67.6</td>
<td>100</td>
<td>85.9</td>
<td>73.0</td>
</tr>
<tr>
<td>$T_3$, $r = (0.03, 0.05, 0.1)$</td>
<td>13.7</td>
<td>39.3</td>
<td>71.1</td>
<td>76.4</td>
<td>99.5</td>
<td>20.3</td>
<td>67.4</td>
<td>94.1</td>
<td>45.6</td>
<td>99.8</td>
<td>70.7</td>
<td>61</td>
</tr>
<tr>
<td>$Lm$</td>
<td>28.9</td>
<td>72.6</td>
<td>95.8</td>
<td>88.7</td>
<td>100</td>
<td>18.3</td>
<td>67.1</td>
<td>97.2</td>
<td>29.7</td>
<td>99.7</td>
<td>88.0</td>
<td>73.6</td>
</tr>
</tbody>
</table>

4. Conclusion

We provide an efficient test of the null hypothesis of a homogeneous Poisson process for point patterns in a square domain, by proposing a new correction of edge effects. Sample correction (for each point of the data) has rarely been questioned since Ripley’s original paper, except by authors claiming test statistics with no correction as more powerful (see [1, 11]). Instead of correcting on each sample to reduce or cancel the bias, we compute the exact bias, so that we avoid to increase of the variance by discarding some of the observed points. The resulting test is efficient on samples with a few dozens of points as encountered in actual data sets.

This is a theoretical and practical improvement on Monte-Carlo methods as it is quicker and often more powerful. Monte-Carlo simulation of the distribution of $\hat{K}$ is a good method for small samples but becomes tedious when the number of points increases. Marcon and Puech [17] computed $K_1$ for a 36,000-point data set (the largest ever published as far as we know), but had to limit the number of simulations to 20. With a personal computer, calculating the distribution of $Lm$ with 10 000 simulations of a 10 000-points set is 2 days long. But it takes approximatively 3 minutes to compute $T_2$ for three distances with optimized C++ code and 5 minutes with a R routine [19].

Our work should be extended in two directions: to other domain shapes that are of interest for the practitioners and to 3-dimensional data for high resolution medical imaging. A further study of the asymptotics of the distribution of $\hat{K}(r)$ for dependent point process models such as Markov or Cox processes should also be achieved to inform on the power of the test.

5. Proofs

5.1. Proof of Proposition 2.1

Let $U$ and $V$ be two independent uniform variables on $[0, 1]^2$. The expectations of the Ripley statistics are

$$
\mathbb{E}\hat{K}_{1,n}(r) = \frac{1}{n^2 \rho^2} \mathbb{E}\left( \sum_{X_i \neq X_j \in S} I\{d(X_i, X_j) \leq r\} \right) = \frac{\mathbb{E}(N(N-1))}{n^2 \rho^2} \mathbb{E}( I\{d(U, V) \leq r\}) = n^2 e_{r,n}.
$$

$$
\mathbb{E}\hat{K}_{2,n}(r) = n^2 \mathbb{E}\left( \frac{1}{N(N-1)} \sum_{X_i \neq X_j \in S} I\{d(X_i, X_j) \leq r\} \right) = n^2 \mathbb{P}(N > 1) \mathbb{E}( I\{d(U, V) \leq r\})
$$

$$
= n^2 \left( 1 - (1 + \rho n^2) e^{-\rho n^2} \right) e_{r,n}.
$$

The following lemma allows to conclude:

**Lemma 5.1.**

$$
e_{r,n} = \frac{\pi r^2}{n^2} - \frac{8 r^3}{3 n^3} + \frac{r^4}{2 n^4}.
$$
Proof. We split $A_n$ into four parts to compute $e_{r,n}$:

\[
e_{r,n} = \int_{\xi \in A_n^1} \int_{\eta \in A_n} I\{d(\xi, \eta) \leq r\} \frac{1}{n^2} d\xi d\eta \tag{5.1}
\]

\[
+ \int_{\xi \in A_n^2} \int_{\eta \in A_n} I\{d(\xi, \eta) \leq r\} \frac{1}{n^2} d\xi d\eta \tag{5.2}
\]

\[
+ \int_{\xi \in A_n^3} \int_{\eta \in A_n} I\{d(\xi, \eta) \leq r\} \frac{1}{n^2} d\xi d\eta \tag{5.3}
\]

\[
+ \int_{\xi \in A_n^4} \int_{\eta \in A_n} I\{d(\xi, \eta) \leq r\} \frac{1}{n^2} d\xi d\eta \tag{5.4}
\]

where (see Fig. 2)
- (interior) $A_n^1 = \{\xi, \xi \text{ is at distance larger than } r \text{ from the boundary}\}$
- (one edge) $A_n^2 = \{\xi, \xi \text{ is at distance less than } r \text{ from an edge, larger than } r \text{ from the others}\}$
- (two edges) $A_n^3 = \{\xi, \xi \text{ is at distance less than } r \text{ from two edges and larger than } r \text{ from the corner}\}$
- (corner) $A_n^4 = \{\xi, \xi \text{ is at distance less than } r \text{ from the corner}\}$

Note that $A_n^2$, $A_n^3$ and $A_n^4$ are composed of four parts that contribute identically. We establish formulas only for one of these parts.

Lemma 5.2. Define function $g(x) = \arccos(x) - x\sqrt{1-x^2}$.

If $\xi \in A_n^1$,

\[
\int_{\eta \in A_n} I\{d(\xi, \eta) \leq r\} d\eta = \pi r^2.
\]

If $\xi \in A_n^2$, with $n-r < \xi_1 < n$, $x_1 = \frac{1}{r}(n-\xi_1)$,

\[
\int_{\eta \in A_n} I\{d(\xi, \eta) \leq r\} d\eta = r^2(\pi - g(x_1))
\]

If $\xi \in A_n^3$, with $n-r < \xi_1 < n$, $n-r < \xi_2 < n$ and $(x_1, x_2) = \frac{1}{r}(n-\xi_1, n-\xi_2)$,

\[
\int_{\eta \in A_n} I\{d(\xi, \eta) \leq r\} d\eta = r^2(\pi - g(x_1) - g(x_2)).
\]

\begin{figure}[h]
\centering
\includegraphics[width=0.5\textwidth]{zones_square.png}
\caption{Zones in the square.}
\end{figure}
If $\xi \in A_n^2$, with $n-r < \xi_1 < n$, $n-r < \xi_2 < n$ and $(x_1, x_2) = \frac{1}{n}(n - \xi_1, n - \xi_2)$,
\[
\int_{\eta \in A_n} I\{d(\xi, \eta) \leq r\}d\eta = r^2 \left( \frac{3\pi}{4} + x_1x_2 - \frac{g(x_1) + g(x_2)}{2} \right).
\]

**Note.** For $0 \leq x \leq 1$, function $g(x)$ is the area of the intersection of a ball of radius 1 with a half plane, when the center of the ball lies outside the half plane at a distance $x$ from its boundary.

**Proof.** Let $B(\xi, r)$ denote the ball of center $\xi$ and radius $r$. For the interior points $\xi \in A_n^1$, $B(\xi, r) \subset A_n$.

Let $\xi \in A_n^2$. We compute the area of $B(\xi, r) \cap A_n$.
\[
\int_{\eta \in A_n} I\{d(\xi, \eta) \leq r\}d\eta = \frac{\pi r^2}{2} + 2r^2 \int_0^{x_1} \sqrt{1 - t^2} dt = r^2 \left( \pi - \arccos(x_1) + x_1\sqrt{1 - x_1^2} \right) = r^2 \left( \pi - g(x_1) \right).
\]

Note that $r^2 g(x)$ is the part of the ball that lies out of the square $A_n$ if the center is at distance $xr$ from the edge of the square.

Let $\xi \in A_n^3$. Here the ball intersects two edges of the square and the area of $B(\xi, r) \cap A_n$ is
\[
\int_{\eta \in A_n} I\{d(\xi, \eta) \leq r\}d\eta = r^2 \left( \pi - g(x_1) - g(x_2) \right).
\]

Let $\xi \in A_n^4$. Divide the ball into four quarters along axes parallel to the coordinate axes. One of the quarter is inside the square, two intersect the edges, leaving outside an area equal to $(g(x_1) + g(x_2))/2$. The area of the intersection of the last quarter with the square is $x_1x_2$ so that the area of $B(\xi, r) \cap A_n$ is
\[
\int_{\eta \in A_n} I\{d(\xi, \eta) \leq r\}d\eta = r^2 \left( \frac{3\pi}{4} + x_1x_2 - \frac{g(x_1) + g(x_2)}{2} \right).
\]

**Proof of Lemma 5.1 (continued).** The left-hand side of (5.1) is $m(A_n^1)\pi r^2 = \pi(n - 2r)^2 r^2$. Recall that $A_n^2$ is composed of four parts that contribute identically. Using the integration formula of the arccos function, we get the integral of $g$:
\[
G(x) = \int_0^x g(u)du = x \arccos(x) - \sqrt{1 - x^2} + \frac{1}{3}(1 - x^2)^{3/2} + \frac{2}{3}.
\]

Then the contribution (5.2) is equal to
\[
4r \int_r^{n-r} d\xi_2 \int_0^{1/2} r^2(\pi - g(x))dx = 4r^3(\pi - 2r)(\pi - G(1)) = \left( 4\pi - \frac{8}{3} \right) r^3(n - 2r).
\]

We consider $A_n^3$; the domain of integration is symmetric in $(x_1, x_2)$ so that the contribution (5.3) is equal to
\[
4r^4 \int_0^1 dx_1 \int_0^{1/2} (\pi - 2g(x_1))dx_2 = r^4 \left( 4\pi \left( 1 - \frac{\pi}{4} \right) - 8G(1) + 8 \int_0^1 g(x_1)\sqrt{1 - x_1^2}dx_1 \right).
\]

But $\int_0^1 g(x_1)\sqrt{1 - x_1^2}dx_1 = \frac{\pi^2}{16}$, so that contribution (5.3) is equal to $r^4 \left( 4\pi - \frac{\pi^2}{2} - \frac{16}{3} \right)$.

We consider $A_n^4$; the contribution (5.4) is equal to
\[
4r^4 \int_0^1 dx_1 \int_0^{1/2} \left( \frac{3\pi}{4} + x_1x_2 - g(x_1) \right)dx_2 = r^4 \left( \frac{3\pi^2}{4} + \frac{1}{2} - 4 \int_0^1 g(x_1)\sqrt{1 - x_1^2}dx_1 \right) = r^4 \left( \frac{\pi^2}{2} + \frac{1}{2} \right).
\]

Gathering the four contributions, we get
\[
\epsilon_{r,n} = \frac{r^2}{n^2} \left( \pi \left( 1 - \frac{2r}{n} \right)^2 + \left( 4\pi - \frac{8}{3} \right) \frac{r}{n} \left( 1 - \frac{2r}{n} \right) + \left( 4\pi - \frac{29}{6} \right) \frac{r^2}{n^2} \right) = \frac{r^2}{n^2} \left( \pi - \frac{8}{3} \frac{r}{n} + \frac{1}{2} \frac{r^2}{n^2} \right).\]
5.2. Proof of Proposition 2.2

For \( u = 1 \) or \( 2 \), we decompose the variance of \( K_{u,A_n}(r) \) by conditioning the variable with respect to the number \( N \) of points in the sample. Conditionally to \( N, K_{u,A_n}(r) \) has the form of a \( U \)-statistic. Then we apply the Hoeffding decomposition to this \( U \)-statistic. We use the relation

\[
\text{var}(\hat{K}_{u,A_n}(r)) = \text{var}(\hat{K}_{u,A_n}(r)|N) + \mathbb{E}\text{var}(\hat{K}_{u,A_n}(r)|N).
\]

We first consider the conditional expectation of \( \hat{K}_{u,A_n}(r) \).

\[
\mathbb{E}(\hat{K}_{1,n}(r)|N) = \frac{1}{n^2 \rho^2} \left( \sum_{i \neq j=1}^{N} \mathbb{E} \{ d(X_i, X_j) \leq r \} \right) = \frac{N(N-1)c_{r,n}}{n^2 \rho^2},
\]

\[
\mathbb{E}(\hat{K}_{2,n}(r)|N) = \frac{n^2}{N(N-1)} \sum_{i \neq j=1}^{N} \mathbb{E} \{ d(U_i, U_j) \leq r \} = n^2 c_{r,n} \mathbb{I}\{N > 1\}.
\]

Because \( N \) is a Poisson variable with intensity \( \rho n^2 \)

\[
\mathbb{E}N^2(N-1)^2 = \rho^4 n^8 + 4 \rho^3 n^6 + 2 \rho^2 n^4.
\]

\[
\text{var}(N(N-1)) = 4 \rho^3 n^6 + 2 \rho^2 n^4.
\]

Then

\[
\text{var}(\hat{K}_{1,n}(r)|N) = \frac{(4 \rho n^2 + 2)c_{r,n}}{\rho^2}, \quad (5.6)
\]

\[
\text{var}(\hat{K}_{2,n}(r)|N) = 4 \mathbb{P}\{N > 1\}(1 - \mathbb{P}\{N > 1\})c_{r,n}^2 = n^4 e^{-\rho n} \left( 1 + \rho n^2 \right) \left( 1 - e^{-\rho n^2} \left( 1 + \rho n^2 \right) \right) c_{r,n}^2. \quad (5.7)
\]

We compute the conditional variances.

\[
\text{var}(\hat{K}_{1,n}(r)|N) = \frac{1}{n^4 \rho^4} \text{var} \left( \sum_{i \neq j=1}^{N} h(X_i, X_j, r) \right),
\]

\[
\text{var}(\hat{K}_{2,n}(r)|N) = \frac{n^4}{N^2(N-1)^2} \text{var} \left( \sum_{i \neq j=1}^{N} h(X_i, X_j, r) \right).
\]

Conditionally to \( N \), the locations of the points are independent and uniformly distributed variables \( U_i \) over \( A_n \).

We introduce the Hoeffding decomposition of the \( U \)-statistic kernel \( h \):

\[
h(x, y, r) = h_1(x, r) + h_1(y, r) + h_2(x, y, r),
\]

where \( h_1(x) = \mathbb{E}(h(U, V, r)|V = x) \), \( (U, V) \) being two independent uniform random variables on \( A_n \).

Then \( \mathbb{E}h_1(U, r) = 0 \) and \( \mathbb{E}(h_2(U, V, r)|U) = \mathbb{E}(h_2(U, V, r)|V) = 0 \), so that

\[
\text{var} h(U, V, r) = \text{var} h_1(U, r) + \text{var} h_1(V, r) + \text{var} h_2(U, V, r) = 2 \mathbb{E}h_1^2(U, r) + \text{var} h_2(U, V, r).
\]

From

\[
\sum_{i \neq j=1}^{N} h(U_i, U_j, r) = 2(N-1) \sum_{i=1}^{N} h_1(U_i, r) + \sum_{i \neq j=1}^{N} h_2(U_i, U_j, r).
\]
Lemma 5.3.

We now apply the same decomposition to \( \text{cov} (\hat{K}_{1,n}(r)|N) \) where each variance term appears four times in the expansion of the variance of the sum over \( i \neq j \).

\[
\text{var} (\hat{K}_{1,n}(r)|N) = \frac{4(N-1)^2}{n^4 \rho^4} \text{var} \left( \sum_{i=1}^{N} h_1(U_i, r) \right) + \frac{1}{n^4 \rho^4} \text{var} \left( \sum_{i \neq j=1}^{N} h_2(U_i, U_j, r) \right)
\]

\[
= \frac{4N(N-1)^2}{n^4 \rho^4} \mathbb{E} h_1^2(U, r) + \frac{2N(N-1)}{n^4 \rho^4} \sum_{i \neq j=1}^{N} \text{var} h_2(U_i, U_j, r)
\]

\[
= \frac{4N(N-1)^2}{n^4 \rho^4} \mathbb{E} h_1^2(U, r) + \frac{2N(N-1)}{n^4 \rho^4} (\text{var} h(U, V, r) - 2\mathbb{E} h_1^2(U, r))
\]

\[
= \frac{4N(N-1)(N-2)}{n^4 \rho^4} \mathbb{E} h_1^2(U, r) + \frac{2N(N-1)}{n^4 \rho^4} \text{var} h(U, V, r),
\]

Note that the factor 2 in the second line may be surprising in the variance of a sum of independent variables, but each variance term appears four times in the expansion of the variance of the sum over \( i \neq j \). Now \( \text{var} h(U, V, r) = \epsilon_{r,n} - \epsilon_{r,n}^2 \) and using factorial moments of the Poisson distribution

\[
\mathbb{E} \text{var} (\hat{K}_{1,n}(r)|N) = \frac{4N^2}{\rho} \mathbb{E} h_1^2(U, r) + \frac{2}{\rho^2} (\epsilon_{r,n} - \epsilon_{r,n}^2). \quad (5.8)
\]

Lemma 5.3 gives the exact value of \( \mathbb{E} h_1^2(U, r) \). Its proof is postponed at the end of the paper.

Lemma 5.3.

\[
\mathbb{E} h_1^2(U, r) = \frac{r^5}{n^2} \left( \frac{8}{3} \pi - \frac{256}{45} \right) + \frac{r^6}{n^2} \left( \frac{11}{48} \pi - \frac{56}{9} \right) + \frac{8}{3 n^2} - \frac{1}{4 n^2}.
\]

With relations (5.6) and (5.8), we get

\[
\text{var} (\hat{K}_{1,n}(r)) = \frac{2\epsilon_{r,n}}{\rho^2} + \frac{4n^2\epsilon_{r,n}}{\rho} + \frac{4n^2}{\rho} \mathbb{E} h_1^2(U, r)
\]

\[
= \frac{1}{n^2} \left( \frac{2\pi^2}{\rho^2} + \frac{4\pi^2 r^4}{\rho} \right) - \frac{1}{n^2} \left( \frac{16}{3} \pi^4 + \left( \frac{32}{3} + \frac{1024}{45} \right) r^5 \right) + \frac{1}{n^2} \left( \frac{r^4}{\rho^2} + \left( \frac{59}{12} + \frac{32}{9} \right) \frac{r^6}{\rho} \right).
\]

Similarly

\[
\text{var} (\hat{K}_{2,n}(r)) = \frac{4n^4 \mathbb{I} \{ N > 1 \} (N-2)}{N(N-1)} \mathbb{E} h_1^2(U, r) + \frac{2n^4 \mathbb{I} \{ N > 1 \}}{N(N-1)} \text{var} h(U, V, r),
\]

\[
\mathbb{E} \text{var} (\hat{K}_{2,n}(r)) = 4n^4 \mathbb{E} \left( \frac{\mathbb{I} \{ N > 1 \} (N-2)}{N(N-1)} \right) \mathbb{E} h_1^2(U, r) + 2n^4 \mathbb{E} \left( \frac{\mathbb{I} \{ N > 1 \}}{N(N-1)} \right) (\epsilon_{r,n} - \epsilon_{r,n}^2).
\]

From this and relation (5.7), we get

\[
\text{var} (\hat{K}_{2,n}(r)) = 2n^4 \mathbb{E} \left( \frac{\mathbb{I} \{ N > 1 \}}{N(N-1)} \right) (\epsilon_{r,n} - \epsilon_{r,n}^2) + 4n^4 \mathbb{E} \left( \frac{\mathbb{I} \{ N > 1 \} (N-2)}{N(N-1)} \right) \mathbb{E} h_1^2(U, r)
\]

\[
+ n^4 e^{-\rho n^2} (1 + \rho n^2) \left( 1 - e^{-\rho n^2} - \rho n^2 e^{-\rho n^2} \right) \epsilon_{r,n}.
\]

We now apply the same decomposition to \( \text{cov} (\hat{K}_{1,n}(r), \hat{K}_{1,n}(r')) \),

\[
\text{cov} (\mathbb{E} (\hat{K}_{1,n}(r')|N), \mathbb{E} (\hat{K}_{1,n}(r)|N)) = \frac{(4\rho n^2 + 2) \epsilon_{r,n} \epsilon_{r,n}}{\rho^2}. \quad (5.9)
\]
Denoting $r$ evaluated; as the integrand is bounded and very smooth this can be achieved without difficulties. To compute $h_1(U, r')$ cannot be achieved in a close form in the edge bands and in the corner. We consider the following zones:

- (interior) $A_n^{1,1} = \{ \xi, \xi \text{ is at distance larger than } r' \text{ from the boundary}\}$,
- (interior-edge) $A_n^{1,2} = \{ \xi, \xi \text{ is at distance between } r \text{ and } r' \text{ from an edge, larger than } r' \text{ from the others}\}$,
- (edge) $A_n^{2,2} = \{ \xi, \xi \text{ is at distance less than } r \text{ from an edge, larger than } r' \text{ from the others}\}$,
- (corner) $A_n^{3,3} = \{ \xi, \xi \text{ is at distance less than } r' \text{ from two edges}\}$.

Denoting $x_1 = \frac{1}{r}(n - \xi_1)$ and $x_1' = \frac{1}{r'}(n - \xi_1)$ we get

$$h_1(X_j, r')h_1(X_j, r) = \begin{cases} 
\left( \frac{\pi r^2}{n^2} - e_{r', n} \right) \left( \frac{\pi r^2}{n^2} - e_{r, n} \right) & \text{on } A_n^{1,1}, \\
\frac{\pi r^2}{n^2} - e_{r', n} - \frac{r^2}{n^2}g(x_1') \left( \frac{\pi r^2}{n^2} - e_{r, n} \right) & \text{on } A_n^{1,2}, \\
\left( \frac{\pi r^2}{n^2} - e_{r', n} - \frac{r^2}{n^2}g(x_1') \right) \left( \frac{\pi r^2}{n^2} - e_{r, n} - \frac{r^2}{n^2}g(x_1) \right) & \text{on } A_n^{2,2}.
\end{cases}$$

Denote $b_{r, n} = \left( \pi - \frac{n^2 - 2r^2}{2n^2} e_{r, n} \right) = \frac{8r}{3n} - \frac{r^2}{2n^2}$.

$$\text{cov}(h_1(X_j, r'), h_1(X_j, r)) = C(A_n^{1,1}) + C(A_n^{1,2}) + C(A_n^{2,2}) + C(A_n^{3,3})$$

$$C(A_n^{1,1}) = \frac{r'^2 r^2}{n^4} \left( 1 - \frac{2r'}{n} \right)^2 b_{r', n} b_{r, n}$$

$$C(A_n^{1,2}) = 4 \left( 1 - \frac{2r'}{n} \right) \frac{r'^3 r^2}{n^5} b_{r, n} \int_{r/r'}^1 (b_{r', n} - g(x_1'))dx_1'$$

$$C(A_n^{2,2}) = 4 \left( 1 - \frac{2r'}{n} \right) \frac{r'^3 r^2}{n^5} \int_0^1 (b_{r', n} - g(rx_1/r'))(b_{r, n} - g(x_1))dx_1.$$

The first integral may be expressed in terms of function $G$, the second integral is elliptic and has to be numerically evaluated; as the integrand is bounded and very smooth this can be achieved without difficulties. To compute
the term $C(A_{n}^{3,3})$, we rewrite the different values of function $h_{1}$ with the help of indicator functions:

\[
\begin{align*}
    h_{A1}(x, r) &= b_{r,n} \mathbb{I}\{x_1 \geq 1; x_2 \geq 1\} \\
    h_{A2}(x, r) &= (b_{r,n} - g(x_1)) \mathbb{I}\{x_1 \geq 1; x_2 < 1\} + (b_{r,n} - g(x_1)) \mathbb{I}\{x_2 \geq 1; x_1 < 1\} \\
    h_{A3}(x, r) &= (b_{r,n} - g(x_1) - g(x_2)) \mathbb{I}\{x_1 < 1; x_2 < 1; x_1^2 + x_2^2 \geq 1\} \\
    h_{A4}(x, r) &= (b_{r,n} - \pi/4 + x_1 x_2 - (g(x_1) + g(x_2))/2) \mathbb{I}\{x_1^2 + x_2^2 < 1\}
\end{align*}
\]

For $x' = \frac{1}{r}(n - \xi_1, n - \xi_2)$, $C(A_{n}^{3,3}) = 4 \frac{r^2 \rho^4}{n^4} \int_0^1 \int_0^1 \sum_{i=1}^4 h_{A1}(r'x'/r, r) \times \sum_{i=3}^4 h_{A1}(x', r) dx_1 dx_2$

and this integral also can be numerically evaluated.

**Note.** The whole computation of this term of the covariance could be numerically achieved, but we gain some useful precision with an exact computation whenever it is possible.

The case of the covariance of $K_{2,n}(r)$ is analogous:

\[
\begin{align*}
    \text{cov}(E(\hat{K}_{2,n}(r')|N), E(\hat{K}_{2,n}(r)|N)) &= n^4 e^{-\rho n^2} (1 + \rho n^2)(1 - e^{-\rho n^2}(1 + \rho n^2)) e_{r,n} e_{r,n}.
\end{align*}
\]

\[
\begin{align*}
    E \text{ cov}(\hat{K}_{2,n}(r'), \hat{K}_{2,n}(r)|N) &= 4n^4 E \left( \frac{\mathbb{I}\{N > 1\}(N - 2)}{N(N - 1)} \right) \text{cov}(h_1(U, r'), h_1(U, r)) \\
    &+ 2n^4 E \left( \frac{\mathbb{I}\{N > 1\}(N - 2)}{N(N - 1)} \right) (e_{r,n} - e_{r',n} e_{r,n}).
\end{align*}
\]

### 5.3. Proof of Theorem 2.3

We show that any linear combination of the $K_{1,n}(r_i)$ is asymptotically normal. Let $\Lambda = (\lambda_1, \ldots, \lambda_d)$ be a vector of real coefficients. Define $Z_1 = \sum_{i=1}^d \lambda_i K_{1,n}(r_i)$. We use the Bernstein blocks technique [3]: we divide the square $A_n$ into squares of side $p$ with $p = o(n)$. These squares are separated by gaps of width $2r_d$ so that the sums over couples of points in each square are independent. The couples of points with at least one point in the gaps give a negligible contribution, so that the statistic $Z_1$ is equivalent to a sum of independent variables and asymptotically normal.

Set $p = n^{1/4}$. Assume that the Euclidean division of $n$ by $(p + 2r_d)$ gives a quotient $a$ and a remainder $q$. For $l = 0, \ldots, a$, we define the segment $I_l = [(p + 2r_d)l, (p + 2r_d)l + p - 1]$. We order the set $\{0, \ldots, a\}$ by the lexicographic order. To any integer $i$ such that $1 \leq i \leq k = (a + 1)^2$, corresponds an element $(j_1, j_2)$ of this set; we define the block $P_{i,n} = I_{j_1} \times I_{j_2}$ and $Q = A_n \setminus \cup_{i,n} P_{i,n}$ the set of points that are in none of the $P_{i,n}$’s. For each block $P_{i,n}$ and $Q$, we define the partial sums:

\[
\begin{align*}
    u_{i,n} &= \frac{1}{n \rho^{5/2}} \sum_{X_i \notin X_m \in P_{i,n}}^{d} \lambda_i \mathbb{I}\{d(X_i, X_m) \leq r_i\}, \\
    v_{i,n} &= \frac{1}{n \rho^{5/2}} \sum_{X_i \notin X_m \in P_{i,n}, X_m \in Q}^{d} \lambda_i \mathbb{I}\{d(X_i, X_m) \leq r_i\} \\
    w_n &= \frac{1}{n \rho^{5/2}} \sum_{X_i \notin X_m \in Q}^{d} \sum_{t=1}^{d} \lambda_t \mathbb{I}\{d(X_i, X_m) \leq r_i\}.
\end{align*}
\]

then

\[
n \sqrt{\rho}(Z_1 - \mathbb{E}Z_1) = \sum_{i=1}^k (u_{i,n} - \mathbb{E}u_{i,n}) + \sum_{i=1}^k (v_{i,n} - \mathbb{E}v_{i,n}) + w_n - \mathbb{E}w_n.
\]

We show that the sum of the $u_{i,n}$ converges in distribution to a Gaussian variable and that the other term are negligible in $L^2$. We check the conditions of the following CLT adapted from [2].
Theorem 5.4. Let \((z_{i,n})_{0 \leq i \leq k(n)}\) be an array of random variables satisfying

1. There exists \(\delta > 0\) such that \(\sum_{i=0}^{k(n)} \mathbb{E}|z_{i,n}|^{2+\delta}\) tends to 0 as \(n\) tends to infinity,
2. \(\sum_{i=0}^{k(n)} \text{var } z_{i,n}\) tends to \(\sigma^2\) as \(n\) tends to infinity,

then \(\sum_{i=0}^{k(n)} z_{i,n}\) tends in distribution to \(\mathcal{N}(0, \sigma^2)\) as \(n\) tends to infinity.

To check Condition 1, we compute the fourth order moment of \(u_{i,n} - \mathbb{E}u_{i,n}\). Let \(N_i\) be the number of points of \(S\) that fall in \(P_{i,n}\). Denote \(f(x,y) = \sum_{l=1}^{d} \lambda_l \left( \mathbb{I}\{d(x,y) \leq r_i\} - c_{r,p}\right) = \sum_{l=1}^{d} \lambda_l h(x,y,r_l)\), then

\[
\mathbb{E}(u_{i,n} - \mathbb{E}u_{i,n})^4|N_i) = \frac{1}{n^4 \rho^6} \mathbb{E}\left( \sum_{l \neq m = 1}^{N_i} f(U_l, U_m)\right)^4
\]

Denote \(f_1\) and \(f_2\) the decomposing functions of \(f\): \(\mathbb{E}(f_1(U_i)) = 0\), \(\mathbb{E}(f_1(U_i)f_2(U_i, U_m)) = \mathbb{E}(f_1(U_m)f_2(U_i, U_m)) = 0\), for \(U_i\) and \(U_m\) two independent uniform variables on \(P_{i,n}\).

\[
\sum_{l \neq m = 1}^{N_i} f(U_l, U_m) = 2(N_i - 1) \sum_{l = 1}^{N_i} f_1(U_l) + \sum_{l \neq m = 1}^{N_i} f_2(U_l, U_m).
\]

Lemma 5.5. \(f_1\) is bounded by \(Cp^{-2}\), \(f_2\) is bounded by a constant and \(f_2(x, y) \leq Cp^{-2}\) as soon as \(\|x - y\| > r\).

Proof. All the quantities computed in Lemma 5.2 for the four different cases are bounded by a constant so that \(\mathbb{E}(\{d(U_1, U_2) \leq r\}|U_1) = O(p^{-2})\). As \(c_{r,p} = O(p^{-2})\), this is also true for \(h_1(x, r)\) for any \(r\) and then for \(f_1\). Because \(f_2(x, y) = f(x, y) - f_1(x) - f_1(y)\), it is bounded by a constant and \(f_2(x, y) = O(p^{-2})\) as soon as the indicator function vanishes.

Define \(M_1 = \mathbb{E}\left(\sum_{l = 1}^{N_i} f_1(U_l)\right)^4\). Then \(M_1 = N_i \mathbb{E}(f_1^4(U)) + 6N_i(N_i - 1) \mathbb{E}(f_1^2(U))^2\) and \(\mathbb{E}(N_i - 1)^4 M_1 = O(p^2)\).

Define \(M_2 = \mathbb{E}\left(\sum_{l \neq m = 1}^{N_i} f_2(U_l, U_m)\right)^4\). Because \(f_2\) is zero mean with respect to one coordinate, only the products where variables appear at least two times contribute.

\[
M_2 = 8 \sum_{l \neq m = 1}^{N_i} \mathbb{E}f_2^4(U_l, U_m) + 48 \sum_{l \neq m \neq u = 1}^{N_i} \mathbb{E}f_2^2(U_l, U_u)f_2^2(U_u, U_m)
\]

\[
+ 96 \sum_{l \neq m \neq u = 1}^{N_i} \mathbb{E}f_2^2(U_l, U_m)f_2(U_m, U_u)f_2(U_u, U_l)
\]

\[
+ 12 \sum_{l \neq m \neq u \neq v = 1}^{N_i} \mathbb{E}f_2^2(U_l, U_m)f_2(U_m, U_u)f_2(U_u, U_v)
\]

\[
+ 48 \sum_{l \neq m \neq u \neq v = 1}^{N_i} \mathbb{E}f_2(U_l, U_m)f_2(U_m, U_u)f_2(U_u, U_v)f_2(U_v, U_l).
\]

Consider the first sum

\[
\sum_{l \neq m = 1}^{N_i} \mathbb{E}f_2^4(U_l, U_m) \leq \sum_{l \neq m = 1}^{N_i} \mathbb{P}\{d(U_l, U_m) \leq r\} + C \mathbb{P}\{d(U_l, U_m) > r\} p^{-8} \leq CN_i(N_i - 1)p^{-2}.
\]
In all the sums, the main term comes from sets of points with all interdistance less than \( r \) and the resulting magnitude of the expectation is \( O(p^2) \), so that

\[
\sum_{i=0}^{k} E(u_{i,n} - E u_{i,n})^4 = O(n^{-2}).
\]

Thus condition 1 is realised. To check condition 2, note that the vector \((K_{1,P_1}(r_1), \ldots, K_{1,P_1}(r_d))\) has a covariance matrix \( \Sigma \) defined by Proposition 2.2 by substituting \( p \) to \( n \) in the expressions. The \( u_{i,n} = \frac{p^2}{n} \sum_{l=1}^{d} \lambda_l \), where \( \Sigma \) is the covariance matrix and \( \text{var} \), magnitude of the expectation is \( \frac{\lambda^2}{\nu^2} A^\dagger \Sigma A \). But \( p^2 \rho \Sigma_p \) tends to \( \Sigma \) as \( p \) tends to infinity and

\[
\sum_{i=0}^{k} \text{var} u_{i,n} = \frac{kp^4 \rho}{n^2} A^\dagger \Sigma A \rightarrow A^\dagger \Sigma A
\]

so that \( \sum_{i=1}^{k} u_{i,n} \) tends in distribution to \( N(0, A^\dagger \Sigma A) \).

Note that the \( v_{i,n} \) are \( k \) independent variables. Denote \( N_{i,r_d} \) the number of points \( X_l \) in the boundary region \( P_{i,r_d} \) of \( P_{i,n} \) such that the ball \( B(X_l, r_d) \) intersects \( Q \) and let \( D(X_l) \) denote this intersection. Note that

\[
E N_{i,r_d} = \rho m(P_{i,r_d}) \leq C p r_d.
\]

\[
\text{var} v_{i,n} \leq \frac{C}{n^2} \text{E} \left( \sum_{l=1}^{N_{i,r_d}} \sum_{m=1}^{N_Q} I\{X_m \in D(X_l)\} \right)^2 \leq \frac{C}{n^2} (T_1 + T_2),
\]

where

\[
T_1 = \text{E} \sum_{l=1}^{N_{i,r_d}} \sum_{m=1}^{N_Q} \sum_{u=1}^{N_{i,r_d}} I\{X_m \in D(X_l)\} \{X_u \in D(X_l)\}
\]

\[
T_2 = \text{E} \sum_{m=1}^{N_{i,r_d}} \sum_{u=1}^{N_{i,r_d}} \sum_{v=1}^{N_Q} I\{X_m \in B(X_l, 2r_d)\} \{X_u \in D(X_l) \cap D(X_m)\}\}
\]

\[
T_1 \leq \text{E} N_{i,r_d} \text{E} N_{i,r_d}^2 \text{P}^2\{X_m \in D(X_l) | X_m \in Q\} \leq \rho^3 m(P_{i,r_d})(m^2(Q) + m(Q)) \left( \frac{\pi r_d^2}{2m(Q)} \right)^2 = O(p).
\]

\[
T_2 = \text{E} \sum_{m=1}^{N_{i,r_d}} \sum_{u=1}^{N_{i,r_d}} \sum_{v=1}^{N_Q} I\{X_m \in B(X_l, r_d)\} \{X_u \in D(X_l) \cap D(X_m)\} \text{E} N_{i,r_d} \text{P}\{X_u \in D(X_l) | X_u \in Q\}
\]

\[
\leq \text{E} N_{i,r_d}^2 \text{P}\{X_m \in B(X_l, r_d)\} \text{E} N_{i,r_d} \text{P}\{X_u \in D(X_l) | X_u \in Q\}
\]

\[
\leq \rho^3 (m^2(P_{i,r_d}) + m(P_{i,r_d})) \left( \frac{\pi r_d^2}{m(P_{i,r_d})} \right) m(Q) \left( \frac{\pi r_d^2}{2m(Q)} \right) = O(p)
\]

and \( \text{var} \left( \sum_{i=1}^{k} v_{i,n} \right) = O\left( kp/n^2 \right) = O\left( p^{-1} \right) \), so that this sum is negligible in \( L^2 \). Similarly

\[
\text{var} (w_n) \leq \frac{C}{n^2} \text{E} \left( \sum_{l=1}^{N_Q} \sum_{m=1}^{N_i} I\{X_m \in B(X_l, r_d)\} \right) \leq \frac{C}{n^2} (T_1 + T_2),
\]
where

\[ T_1 = \mathbb{E} \sum_{l=1}^{N_L} \sum_{m=1}^{N_Q} \mathbb{I}\{X_m \in B(X_l, r_d)\} \]

\[ \leq \mathbb{E} N_Q (N_Q - 1) \mathbb{P}\{X_m \in B(X_l, r_d) | X_m \in Q\} \leq \frac{m^2(Q)}{m(Q)} \cdot \frac{\pi r_d^2}{m(Q)}. \]

\[ T_2 = \mathbb{E} \sum_{l=1}^{N_L} \sum_{m=1}^{N_Q} \mathbb{I}\{X_m \in B(X_l, r_d)\} \mathbb{I}\{X_n \in B(X_l, r_d)\} \]

\[ \leq \mathbb{E} N_Q^2 (N_Q - 1) \mathbb{P}^2\{X_m \in B(X_l, r_d) | X_m \in Q\} \leq \left( m^3(Q) + 2m^2(Q) \right) \left( \frac{\pi r_d^2}{m(Q)} \right)^2. \]

Note that \( m(Q) = O(\sqrt{n}) \). Then \( \text{var} (w_n) = O \left( \frac{m(Q)}{n^2} \right) = O \left( \frac{p}{n} \right) \) and \( w_n \) is negligible in \( L^2 \).

Consider now \( K_{2,n}(r) \). Define \( Z_2 = \sum_{i=1}^{d} \lambda_i K_{2,n}(r_i) = A_{N,n} Z_1 \) where \( A_{N,n} = \frac{\sqrt{n} \rho}{N(N-1)} \). We have \( \mathbb{E}(A_{N,n}^{-1}) = 1 \) and from (5.5), \( \text{var} (A_{N,n}^{-1}) = \frac{4}{n^2 \rho} + \frac{2}{n^2 \rho^2} \). For \( \delta > 0 \), the Markov inequality gives

\[ \mathbb{P}(|A_{N,n}^{-1} - 1| > \delta) \leq \frac{\text{var} (A_{N,n}^{-1})}{\delta^2}. \]

Then, with \( \delta = n^{-1/4} \sum_{n=1}^{\infty} \mathbb{P}(|A_{N,n}^{-1} - 1| > n^{-1/4}) < \sum_{n=1}^{\infty} \frac{4}{n^{3/2} \rho} + \frac{2}{n^{1/2} \rho^2} < \infty \). From the Borel–Cantelli lemma, we get that \( A_{N,n}^{-1} \) converges a.s. to 1. By the Slutsky lemma, \( A_{N,n} Z_1 \) converges in distribution to \( \mathcal{N}(0, A^t \Sigma A) \).

**5.4. Proof of Lemma 5.3**

This lemma is equivalent to Result 1 of [27], substituting \( r/n \) to the parameter \( h \) and substracting \( e^2_{r,n} \). From the computation of the bias, denoting \( x_i = \frac{1}{r}(n - \xi_i) \), we get

\[ h_1(\xi, r) = \frac{\pi r^2}{n^2} - e_{r,n} \text{ on } A_n^1 \]

\[ = \frac{r^2}{n^2} (\pi - g(x_1)) - e_{r,n} \text{ on } A_n^2 \]

\[ = \frac{r^2}{n^2} (\pi - g(x_1) - g(x_2)) - e_{r,n} \text{ on } A_n^3 \]

\[ = \frac{r^2}{n^2} \left( \frac{3\pi}{4} + x_1 x_2 - \frac{g(x_1) + g(x_2)}{2} \right) - e_{r,n} \text{ on } A_n^4 \]

Integrating on the four zones, we get

\[ \mathbb{E}(h_1(X_j, r))^2 = \pi^2 \left( 1 - \frac{2r}{n} \right)^2 \frac{r^4}{n^4} - e_{r,n}^2 + T_1 + T_2 + T_3 \]

\[ T_1 = 4 \left( 1 - \frac{2r}{n} \right)^3 \frac{r^3}{n^3} \int_0^1 (\pi - g(x_1))^2 dx_1 \]

\[ T_2 = 4 \frac{r^3}{n^3} \int_0^1 dx_1 \int_{\sqrt{1-x_1^2}}^1 (\pi - g(x_1) - g(x_2))^2 dx_2 \]

\[ T_3 = 4 \frac{r^6}{n^6} \int_0^1 dx_1 \int_0^{\frac{1-x_1^2}{2}} \left( \frac{3\pi}{4} + x_1 x_2 - \frac{g(x_1) + g(x_2)}{2} \right)^2 dx_2. \]
To rewrite these three terms with the notations of [27], we denote \( \theta = \arccos(x_1) \) and \( \phi = \arccos(x_2) \).

\[
\begin{align*}
T_1 &= 4 \left( 1 - \frac{2r}{n} \right) \frac{r^5}{n^5} \int_0^{\pi/2} (\pi - \theta + \cos(\theta) \sin(\theta))^2 \sin(\theta) d\theta \\
T_2 &= \frac{r^6}{n^6} \int_0^{\pi/2} \sin(\theta) d\theta \int_0^{\pi/2-\theta} (\pi - \theta + \cos(\theta) \sin(\theta) - \phi + \cos(\phi) \sin(\phi))^2 \sin(\phi) d\phi \\
T_3 &= \frac{r^6}{n^6} \int_0^{\pi/2} \sin(\theta') d\theta' \int_0^{\pi/2} (3\pi/2 + 2 \cos(\theta) \cos(\phi) - \theta - \phi + \cos(\theta) \sin(\theta) + \cos(\phi) \sin(\phi))^2 \sin(\phi) d\phi
\end{align*}
\]

changing variables by \( \theta' = \pi/2 - \theta \) and \( \phi' = \pi/2 - \phi \). Then formulas (5), (6) and (7) in [27] give respectively

\[
\begin{align*}
T_1 &= \left( 1 - \frac{2r}{n} \right) \frac{r^5}{n^5} \left( 4\pi^2 - \frac{256}{45} - \frac{8\pi}{3} \right) \quad (5.10) \\
T_2 &= \frac{r^6}{n^6} \left( -\pi^3 + 4\pi^2 - \frac{9\pi}{2} - \frac{512}{45} \right) \quad (5.11) \\
T_3 &= \frac{r^6}{n^6} \left( \frac{\pi^3}{4} + \frac{19\pi}{48} + \frac{8}{9} \right) \quad (5.12)
\end{align*}
\]

Note that the upper bound of the second integral in formula (7) of [27] is a mistyping. Gathering the expression of \( e_{r,n} \), (5.10)–(5.12) gives the result.

Acknowledgements. We are thankful to Michel Koskas for his help in accelerating the computation of \( K \) with \( C^{++} \). We also wish to thank the anonymous referees for their suggestions to improve the section concerning the power of the test.

References


