

A New Integral Representation of the Coverage Probability of a Random Convex Hull

Won Son^a, Chi Tim Ng^{1,b}, Johan Lim^a

^aDepartment of Statistics, Seoul National University, Korea

^bDepartment of Statistics, Chonnam National University, Korea

Abstract

In this paper, the probability that a given point is covered by a random convex hull generated by independent and identically-distributed random points in a plane is studied. It is shown that such probability can be expressed in terms of an integral that can be approximated numerically by function-evaluations over the grid-points in a 2-dimensional space. The new integral representation allows such probability be computed efficiently. The computational burdens under the proposed integral representation and those in the existing literature are compared. The proposed method is illustrated through numerical examples where the random points are drawn from (i) uniform distribution over a square and (ii) bivariate normal distribution over the two-dimensional Euclidean space. The applications of the proposed method in statistics are discussed.

Keywords: Coverage probability, integral representation, random convex hull, random points, stochastic geometry.

1. Introduction

Since the work of Renyi and Sulanke (1963a, b), the convex hull generated by independent and identically-distributed random points (random convex hull in short) has attracted a considerable attention in the literatures of stochastic geometry and has been employed in a variety of statistical procedures. For example, Barnett (1976) defines an ordering of the multivariate data based on the notion of convex hull peeling depth. In Cook (1979), the random convex hull generated by the data points is used to identify the influential observations in linear regression. In the data envelopment analysis, the random convex hull is an important tool for estimating the frontier function (Jeong, 2004; Jeong and Park, 2006) and finding the optimal classifiers (Fawcett and Niculescu-Mizil, 2007; Lim and Won, 2012). Recently, Ng *et al.* (2014) develops a test of independence of two random variables based on the area of the random convex hull.

Though there are endeavors of establishing the probabilistic properties of the random convex hull, most existing works focus on the situations of either (i) the sample size goes to infinity or (ii) the random points are generated from the uniform distribution. The classical results in Renyi and Sulanke (1963a, b) of the expected number of vertexes, the perimeter, and the area of the random convex hull are applicable only if the sample size goes to infinity. Subsequent studies in Hueter (1994, 1999) and

This work was financially supported by the 2013 Chonnam National University Research Program grant (No. 2013-2299).

¹ Corresponding author: Department of Statistics, Chonnam National University, 77 Yongbong-ro, Buk-gu, Gwangju 500-757, Korea. E-mail: easterlyng@gmail.com

Hsing (1994) also focus on the asymptotic behaviors of the random convex hull. Exact formulas of the functionals of the random convex hull are studied in for example, Buchta (2005, 2006). However, the results are obtained only under the assumption that the random points are drawn from the uniform distribution over certain bounded regions.

In order to obtain the probabilistic properties of the random convex hull in the finite-sample cases, Efron (1965) considers the probability that a given point $p = (x, y)$ is not covered by the convex hull of n independent points $\{p_i = (x_i, y_i), i = 1, 2, \dots, n\}$ from a general distribution $F(x, y)$ on \mathcal{R}^2 . Let $q(p, n)$ be such a probability. By providing an integral representation of $q(p, n)$, Efron (1965) further obtains a formula of the expected area of the random convex hull. This integral representation is derived under the coordinate system proposed by Santalo (1953) that is rather complicated to use in practice. As we will indicate in Section 2.3, evaluating $q(p, n)$ under this integral representation requires function-evaluations over the grid-points in a 3-dimensional space.

The main contribution of this paper is to establish a new integral representation that allows $q(p, n)$ be evaluated numerically by function-evaluations over the grid-points in a 2-dimensional space.

This short note is organized as follows. In Section 2, a new integral representation of $q(p, n)$ is given and the numerical algorithm of evaluating $q(p, n)$ is proposed. The results are applicable to a general distribution $F(x, y)$ under finite-sample cases. In Section 3, we show that $q(p, n)$ can further be simplified in the special cases when the random points are drawn from the uniform distribution on a square and from the bivariate normal distribution with an arbitrarily given covariance matrix. Numerical examples are also provided. In the concluding section, the applications to certain statistical problems are discussed.

2. Probability

Let $\text{convH}(p_1, p_2, \dots, p_n)$ be the smallest convex hull containing n independent sample points $p_i = (x_i, y_i)_{i=1, \dots, n}$ distributed according to law $F(x, y)$ over a convex but not necessarily bounded region D . Denote the joint density function corresponding to F by $f(x, y)$. In this section, we derive a new integral representation of $q(p, n)$, the probability that a given point $p = (x, y) \in D$ does not belong to $\text{convH}(p_1, p_2, \dots, p_n)$. General results for arbitrarily given distribution $F(\cdot)$ and convex region D are given in this article. Special cases will be discussed in Section 3.

2.1. Toy example

Before presenting the new integral representation of $q(p, n)$, let us consider the game described below. The sample space of this game is discrete. Later on, a continuous version of such game is used to obtain $q(p, n)$.

Let us consider the following game of lucky wheel as shown in Figure 1. Assume that the lucky wheel is divided into 12 sectors. These sectors are denoted by $(1, 2), (2, 3), (3, 4), \dots, (12, 1)$ respectively. The first and the second number in the brackets are the starting-points and the end-points of the sector anticlockwise. Each sector can either be occupied or not occupied. Here, the events that the sectors are occupied are not necessarily exclusive. The sample space for this problem is then the set of 12-dimensional ordered tuples with 0 (for non occupied) or 1 (for occupied) as the coordinate values. The probability mass function of this sample space is given. For simplicity, assume that the mass function at $(0, 0, \dots, 0)$ is zero. If there are at least six (i.e., half of 12) consecutive sectors that are not occupied, you win, otherwise, you lose. What is the winning probability?

To answer the above question, rewrite the event E of getting win as the union of the following

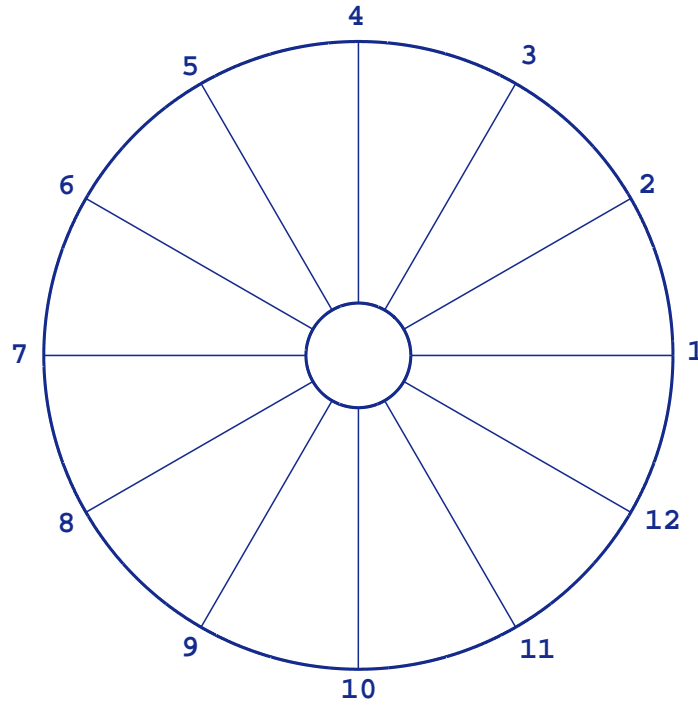


Figure 1: An illustration of lucky wheel

events.

E_1 = The sector (12, 1) is occupied while (1, 7) is not.

E_2 = The sector (1, 2) is occupied while (2, 8) is not.

\vdots

E_{11} = The sector (10, 11) is occupied while (11, 5) is not.

E_{12} = The sector (11, 12) is occupied while (12, 6) is not.

In the above, note that $6 = 12/2$ guarantees that these events E_1, E_2, \dots, E_{12} are mutually exclusive. The event E_1 is equivalent to that (12, 7) is occupied while (1, 7) is not occupied. Therefore,

$$\mathbb{P}(E_1) = \mathbb{P}((12, 7) \text{ is occupied}) - \mathbb{P}((1, 7) \text{ is occupied}).$$

Similarly, the probabilities of events E_2 to E_{12} can be obtained.

2.2. General case

Now we return to our original question, “what is the probability that a fixed point $p = (x, y) \in D$ does not belong to the convex hull”.

The following two statements are equivalent.

1. p does not belong to the convex hull.

2. There is a sector subtended by p and an angle greater than π not occupied by any points p_i , $i = 1, 2, \dots, n$.

Some analogies between the lucky wheel game and the random convex hull are as follows. The lucky wheel in the game has 12 sectors of 30 degrees. In the domain D has infinitely many sectors subtended by the point p and angles of infinitesimal sizes $\Delta\theta$. To win the game, one needs to get at least six consecutive sectors that is not occupied. In order that p is outside the random convex hull, it requires the existence of a sector with subtended angle greater than or equal to π that is not occupied by any random points.

For any given $\theta_2 > \theta_1$, the probability that (θ_1, θ_2) is occupied is

$$1 - \{\mathbb{P}(p_1 \in (\theta_2, \theta_1 + 2\pi))\}^n.$$

Let us define $G(s, t) = \mathbb{P}(p_1 \in (s, t))$. It is not difficult to see by analogy that the required probability is

$$\begin{aligned} q(p, n) &= \lim_{N \rightarrow \infty} \sum_{k=1}^N \{\mathbb{P}((\theta_k - \Delta, \theta_k + \pi) \text{ is occupied}) - \mathbb{P}((\theta_k, \theta_k + \pi) \text{ is occupied})\} \\ &= \lim_{N \rightarrow \infty} \sum_{k=1}^N \{[G(\theta_k + \pi, \theta_k + 2\pi)]^n - [G(\theta_k + \pi, \theta_k - \Delta + 2\pi)]^n\} \\ &= \int_0^{2\pi} \frac{\partial}{\partial t} G(s + \pi, t)^n \Big|_{t=s+2\pi} ds \\ &= n \int_0^{2\pi} [G(s + \pi, s + 2\pi)]^{n-1} \frac{\partial G(s + \pi, t)}{\partial t} \Big|_{t=s+2\pi} ds. \end{aligned}$$

In the above, $\theta_k = k \cdot 2\pi/N$ for $k = 1, 2, \dots, N$ and $\Delta = 2\pi/N$. In addition,

$$G(s, t) = \int_s^t \int_0^{r(\theta)} f(x + r \cos \theta, y + r \sin \theta) r dr d\theta.$$

If D is bounded, $r(\theta)$ can be defined as shown in Figure 2 and if $D = \mathcal{R}^2$, $r(\theta)$ can be replaced by ∞ .

By differentiating $G(s, t)$ with respect to t , we have

$$\frac{\partial G(s, t)}{\partial t} = \int_0^{r(t)} f(x + r \cos t, y + r \sin t) \cdot r dr.$$

Since the above partial derivative does not involve s , it is convenient to introduce the notation

$$h(t) = \frac{\partial G(s, t)}{\partial t}.$$

Note that $h(t)$ is a function of t only and does not depend on s . Then, the required probability can be written as

$$\begin{aligned} q(p, n) &= \mathbb{P}(p = (x, y) \notin \text{convH}(p_1, \dots, p_n)) \\ &= n \int_0^{2\pi} h(s + \pi) [G(s, s + \pi)]^{n-1} ds. \end{aligned} \tag{2.1}$$

In the above, $G(s, s + \pi)$ is a function of $p = (x, y)$.

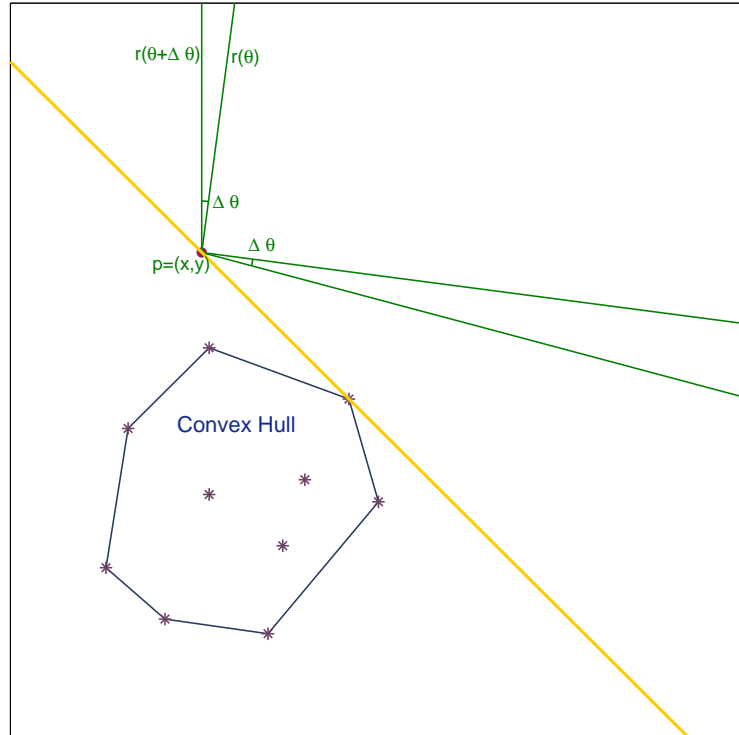


Figure 2: Notations

2.3. Computational issue

In this subsection, we show that the new integral representation given in Subsection 2.2 allows the probability be evaluated numerically by function-evaluations over the grid-points in a 2-dimensional space.

Note that Equation (2.1) can be rewritten as

$$q(p, n) = n \int_0^{2\pi} h(s + \pi) \left[\int_s^{s+\pi} h(\theta) d\theta \right]^{n-1} ds.$$

Algorithm:

Step 1: Choose integers M and N . Set $\theta_k = 2\pi k/2N$ for $k = 0, 1, 2, \dots, 2N - 1$.

Step 2: For $k = 0, 1, 2, \dots, 2N - 1$, evaluate the integral

$$h(\theta) = \int_0^{r(\theta_k)} f(x + r \cos \theta_k, y + r \sin \theta_k) \cdot r dr$$

using any appropriate numerical integration method with M function-evaluations.

Step 3: Compute $I_0 = \pi N^{-1} \sum_{k=0}^{2N-1} h(\theta_k)$.

Step 4: For $k = 1, 2, 3, \dots, 2N - 1$, compute $I_k = I_{k-1} + \pi N^{-1}[h(\theta_k) - h(\theta_{k-N})]$.

Step 5: Obtain $q(p, n) \approx n\pi N^{-1} \sum_{k=0}^{2N-1} h(\theta_{k+N}) I_k^{n-1}$.

This algorithm requires only MN function-evaluations and the computational burden of Step 3–5 is only $O(N)$. If the domain is infinite, Gaussian quadrature method can be chosen in Step 2.

It is interesting to note that the integral representation of $q(p, n)$ in Equation (5.13) of Efron (1965) is

$$q(p, n) = \frac{1}{2}n \int_0^\pi \int_{-\infty}^{\infty} \gamma_{n+1}(\xi(p, \theta), \theta) |\alpha - \beta(p, \theta)| f(x(\xi, \theta, \alpha), y(\xi, \theta, \alpha)),$$

where

$$\begin{aligned} \xi(p, \theta) &= x \cos \theta + y \sin \theta, \\ \beta(p, \theta) &= y \cos \theta - x \sin \theta, \\ x(\xi, \theta, \alpha) &= \xi \cos \theta - \alpha \sin \theta, \\ y(\xi, \theta, \alpha) &= \alpha \cos \theta + \xi \sin \theta, \\ \gamma_{n+1}(\xi, \theta) &= \Gamma^{n-1}(\xi, \theta) - (1 - \Gamma(\xi, \theta))^{n-1}, \\ \Gamma(\xi, \theta) &= \int_{-\infty}^{\infty} \int_{\xi}^{\infty} f(x(\zeta, \theta, \beta), y(\zeta, \theta, \beta)) d\zeta d\beta. \end{aligned}$$

The integral $\Gamma(\xi, \theta)$ involves the three-dimensional function $f(x(\zeta, \theta, \beta), y(\zeta, \theta, \beta))$. As such, the probability $q(p, n)$ has to be approximated numerically with function-evaluations over the grid-points in a three-dimensional space.

3. Two Special Cases

The following two special cases will be discussed in this section. In the first example, $F(\cdot)$ is the uniform distribution function over a bounded convex set D . In the second example, $F(\cdot)$ is the bivariate Gaussian distribution function with mean zero and variance covariance matrix Σ .

3.1. Uniform distribution on $[0, 1]^2$

Consider the case that the random points $p_i, i = 1, 2, \dots, n$ are drawn independently from the uniform distribution over $[0, 1]^2$. Below, we only consider the case $p = (x, y)$ with $0 \leq x, y \leq 1/2$. The probabilities for other values of p can be obtained by symmetry. Under our uniform-distribution assumptions, the function $G(s, t)$ can be written as

$$G(s, t) = \frac{1}{2} \int_s^t r^2(\theta) d\theta$$

and thereby the required probability is

$$q(p, n) = \frac{n}{2} \int_0^{2\pi} r^2(s + \pi) G^{n-1}(s, s + \pi) ds.$$

Here, both $r^2(\theta)$ and $G(s, t)$ depend on $p = (x, y)$.

Let $0 < \theta_1 \leq \pi/2 \leq \theta_2 \leq \pi \leq \theta_3 \leq 3\pi/2 \leq \theta_4$ be the angles defined as follows ,

$$\begin{aligned}\tan(2\pi - \theta_4) &= \frac{y}{1-x}, \\ \tan(\theta_1) &= \frac{1-y}{1-x}, \\ \tan(\pi - \theta_2) &= \frac{1-y}{x}, \\ \tan(\theta_3 - \pi) &= \frac{y}{x}.\end{aligned}$$

In this case,

$$h(t) = \frac{1}{2} \int_0^{r(t)} r \, dr = \frac{1}{2} r^2(t),$$

and $r(\theta)$ is defined as:

$$r(\theta) = \begin{cases} (1-x) \frac{1}{\cos \theta}, & \text{if } \theta_4 - 2\pi \leq \theta \leq \theta_1, \\ \frac{1-y}{\sin \theta}, & \text{if } \theta_1 \leq \theta \leq \frac{\pi}{2}, \\ \frac{1-y}{\sin(\pi - \theta)}, & \text{if } \frac{\pi}{2} \leq \theta \leq \theta_2, \\ \frac{x}{\cos(\pi - \theta)}, & \text{if } \theta_2 \leq \theta \leq \pi, \\ \frac{x}{\cos(\theta - \pi)}, & \text{if } \pi \leq \theta \leq \theta_3, \\ \frac{y}{\cos\left(\frac{3\pi}{2} - \theta\right)}, & \text{if } \theta_3 \leq \theta \leq \frac{3\pi}{2}. \end{cases}$$

Finally,

$$G(s, s + \pi) = \int_s^{s+\pi} \frac{1}{2} r(\theta)^2 d\theta$$

and

$$q(p, n) = n \int_0^{2\pi} h(s + \pi) [G(s, s + \pi)]^{n-1} ds. \quad (3.1)$$

Next, numerical examples are presented. Here, we compute the probability $q(p, n)$ for various p and $n = 10, 100$ and 1000 . The points p is chosen as points in $\{(x, y) | x = i * 0.01, y = 0.01 * j, i, j = 1, 2, \dots, 99\}$. Figure 3 shows the probabilities $q(p, n)$ for different p and n .

3.2. Bivariate normal distribution

Consider the case that the random points $p_i, i = 1, 2, \dots, n$ are drawn independently from the bivariate Normal distribution with mean zero and variance covariance matrix Σ .

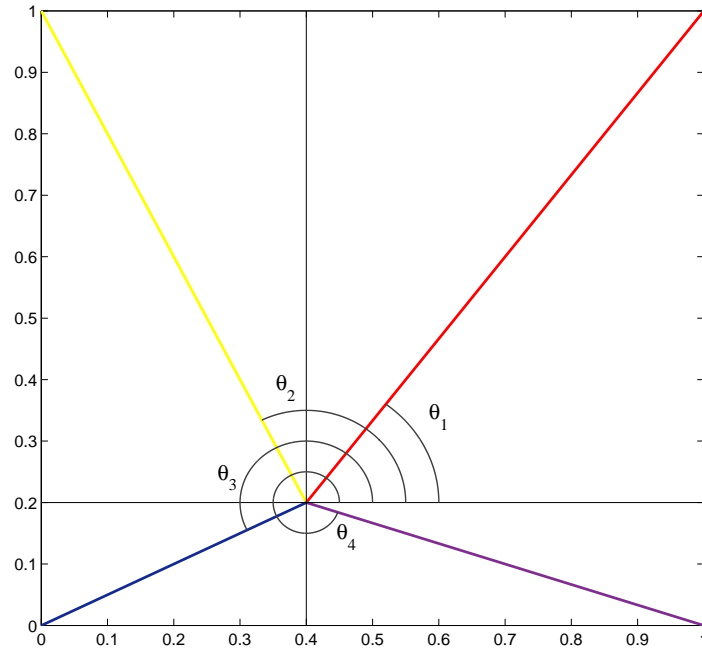


Figure 3: Uniform distribution

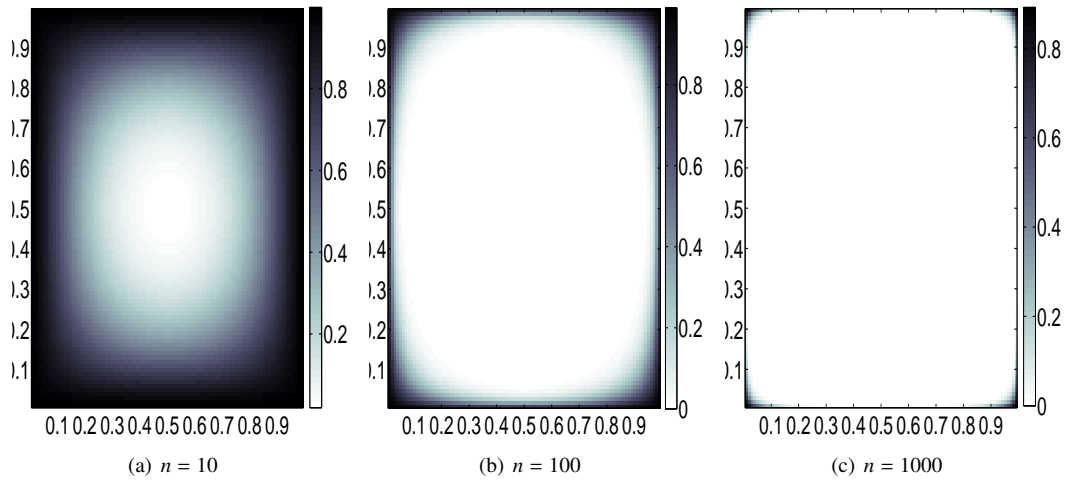
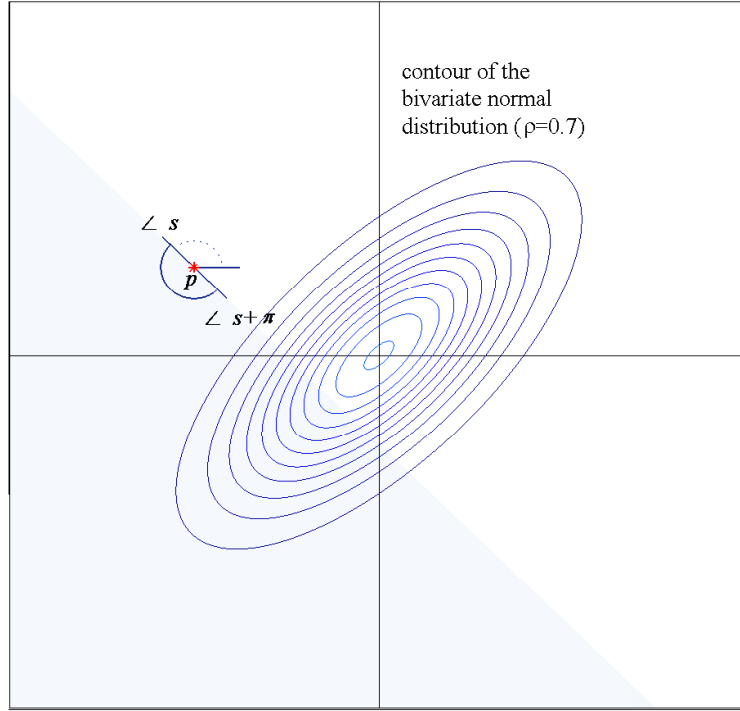


Figure 4: Uniform distribution

To compute the probability $q(p, n)$, we first compute $G(s, s + \pi)$ and $h(s)$. To find $G(s, s + \pi)$, it is beneficial to consider the rectangular coordinates rather than the polar coordinates. In this case, $G(s, s + \pi)$ is the probability measure of the shaded region shown in Figure 5. Let (ζ, η) be the coordinate of a random point. Consider the following rotation and translation,

$$\begin{pmatrix} \zeta' \\ \eta' \end{pmatrix} = U_s^T \begin{pmatrix} \zeta - x \\ \eta - y \end{pmatrix},$$

Figure 5: *Bivariate normal distribution*

where

$$U_s^T = \begin{pmatrix} \cos s & \sin s \\ -\sin s & \cos s \end{pmatrix}.$$

With such transform, the required probability becomes $P(\eta' > 0)$. It can be verified that

$$\begin{pmatrix} \zeta' \\ \eta' \end{pmatrix} \sim N\left(-U_s^T \begin{pmatrix} x \\ y \end{pmatrix}, U_s^T \Sigma U_s\right).$$

Let

$$\mu(p, s) = x \sin s - y \cos s$$

and

$$\sigma^2(p, s) = (\sin s, -\cos s) \Sigma (\sin s, -\cos s)^T.$$

We have

$$G(s, s + \pi) = 1 - \Phi\left(-\frac{\mu(p, s)}{\sigma(p, s)}\right),$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal random variable.

Next, we find $h(t)$. The integral

$$h(p, t) = \int_0^\infty f(x + r \cos t, y + r \sin t) \cdot r dr$$

can be found explicitly. Let

$$\begin{aligned} E(p) &= (x, y)\Sigma^{-1}(x, y)^T, \\ B(p, t) &= (x, y)\Sigma^{-1}(\cos t, \sin t)^T, \\ C(p, t) &= (\cos t, \sin t)\Sigma^{-1}(\cos t, \sin t)^T. \end{aligned}$$

We have

$$h(p, t) = (2\pi)^{-\frac{1}{2}} |\Sigma|^{-\frac{1}{2}} C^{-1}(t) \exp \left\{ -\frac{1}{2} (E - B^2(t)C^{-1}(t)) \right\} \times \left[\phi \left(\frac{B(t)}{C^{\frac{1}{2}}(t)} \right) - \frac{B(t)}{C^{\frac{1}{2}}(t)} \left\{ 1 - \Phi \left(\frac{B(t)}{C^{\frac{1}{2}}(t)} \right) \right\} \right].$$

Again,

$$q(p, n) = n \int_0^{2\pi} h(s + \pi) [G(s, s + \pi)]^{n-1} ds.$$

Numerical examples are given in the following. The probability $q(p, n)$ is computed for three bivariate normal distributions with mean $(0, 0)^T$ and covariance matrices

$$\Sigma = \left(\begin{array}{cc} 1 & 0 \\ 0 & 1 \end{array} \right), \left(\begin{array}{cc} 1 & 0.3 \\ 0.3 & 1 \end{array} \right), \quad \text{and} \quad \left(\begin{array}{cc} 1 & 0.7 \\ 0.7 & 1 \end{array} \right) \quad (3.2)$$

respectively. For each Σ , $q(p, n)$ are computed for $n = 10, 100$, and 1000 and p from the grid points over $[-5, 5] \times [-5, 5]$. The results are shown in Figure 6.

4. Conclusion

In this paper, a new integral representation is given for $q(p, n)$, the probability that a random convex hull generated by independent and identically-distributed random points covers p . Under such new representation, $q(p, n)$ can be approximated numerically by function-evaluations over the grid-points in a 2-dimensional space. On the contrary, the classical integral representation obtained in Efron (1965) requires function-evaluations over the grid-points in a 3-dimensional space. The new results allow even more efficient computation of $q(p, n)$ in the finite-sample cases. Moreover, the proofs provided in the present paper is much simpler and intuitive than those in Efron (1965).

One of the future research directions is to apply the new integral representation and to further simplify the formulas of a number of functionals of the random convex hull, including the moments of area and perimeters etc. For example, the expected area of the random convex hull $\text{convH}(p_1, p_2, \dots, p_n)$, denoted by A_n , can be written as

$$A_n = \iint q(p, n) dx dy.$$

The probability that the $n + 1$ th observation $p_{n+1} = (x_{n+1}, y_{n+1})$ does not belong to the random convex hull $\text{convH}((p_1, p_2, \dots, p_n))$ is

$$u_{n+1} = 1 - \iint q(p, n) dF(x, y).$$

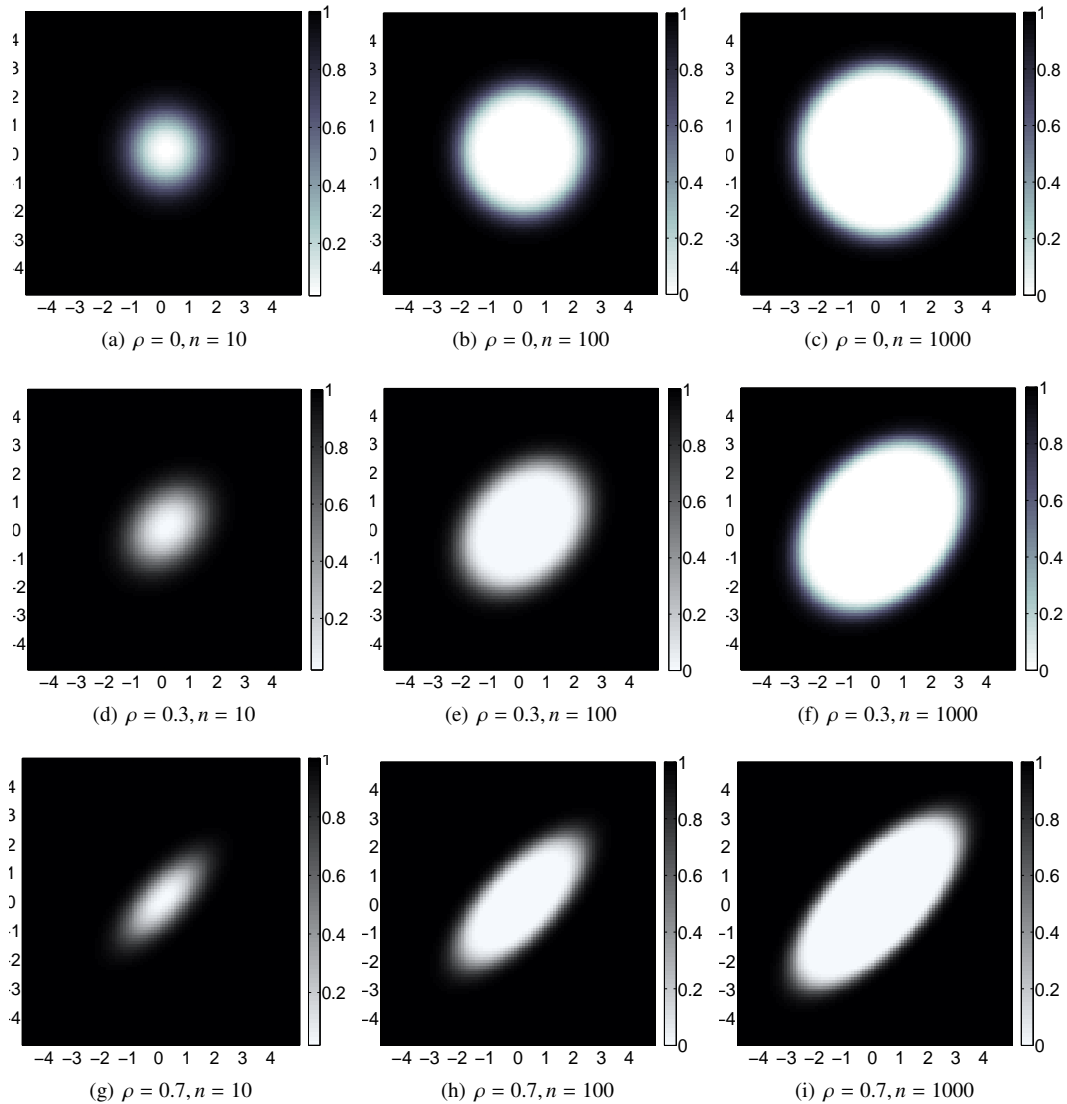


Figure 6: Normal Distribution

Indeed, this is the probability that a new random point forms a new vertex of the convex hull. Therefore, u_{n+1} can be useful in the simulation of the random convex hull.

In the data envelopment analysis, the probability $q(p, n)$ can be used to obtain the required number of sample size n so that the given point p is within the random convex hull. Such sample size is an estimator of the frontier function (or the domain) of the bivariate distribution on a plane (Jeong, 2004; Jeong and Park, 2006).

It is also an interesting research direction of generalizing the results to the higher dimensional situations. The improved efficiency in the computation would facilitates many applications of random convex hull in the multivariate data analysis in the future.

References

- Barnett, V. (1976). The ordering of multivariate data. (with discussion), *Journal of the Royal Statistical Society, Series A*, **139**, 318–339.
- Buchta, C. (2005). An identity relating moments of functionals of convex hulls, *Discrete Computational Geometry*, **33**, 125–142.
- Buchta, C. (2006). The exact distribution of the number of vertices of a random convex chain, *Mathematika*, **53**, 247–254.
- Cook, R. D. (1979). Influential observations in linear regression, *Journal of the American Statistical Association*, **74**, 169–174.
- Efron, B. (1965). The convex hull of random set of points, *Biometrika*, **52**, 331–343.
- Fawcett, T. and Niculescu-Mizil, A. (2007). PAV and the ROC convex hull, *Machine Learning*, **68**, 97–106.
- Hsing, T. (1994). On the asymptotic distribution of the area outside a random convex hull in a disk, *Annals of Applied Probability*, **4**, 478–493.
- Hueter, I. (1994). The convex hull of normal samples, *Journal of Applied Probability*, **26**, 855–875.
- Hueter, I. (1999). Limit theorems for the convex hull of random points in higher dimensions, *Transactions of the American Mathematical Society*, **351**, 4337–4363.
- Jeong, S. (2004). Asymptotic distribution of DEA efficiency scores, *Journal of Korean Statistical Society*, **33**, 449–458.
- Jeong, S. and Park, B. U. (2006). Large sample approximation of the distribution for convex hull estimators of boundaries, *Scandinavian Journal of Statistics*, **33**, 139–151.
- Lim, J. and Won, J. (2012). ROC convex hull and nonparametric maximum likelihood estimation, *Machine Learning*, **88**, 433–444.
- Ng, C. T., Lim, J., Lee, K. E., Yu, D. and Choi, S. (2014). A fast algorithm to sample the number of vertexes and the area of the random convex hull on the unit square, *Computational Statistics*, **29**, 1187–1205.
- Renyi, A. and Sulanke, R. (1963a). Über die konvexe Hülle von n zufällig gewählten Punkten, *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, **2**, 75–84.
- Renyi, A. and Sulanke, R. (1963b). Über die konvexe Hülle von n zufällig gewählten Punkten, *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, **3**, 138–147.
- Santalo, L. A. (1953). Introduction to integral geometry, *Actualities Scientifiques et Industrielles*, **1198**, Hermann, Paris.