Wiley StatsRef:
Statistics Reference Online

Interpoint Distance Distribution

By *Marco Bonetti*

Keywords: *distance, dissimilarity, U-statistic, point process*

Abstract: The interpoint distance distribution (IDD) is the distribution of the random variable defined as the distance, or dissimilarity, between two i.i.d. random variables. Estimation and testing procedures are available, based on samples of i.i.d. observations for the one-sample and the two-sample problem. The IDD allows for a reduction in dimensionality and can be used to perform shape classification. It is also related to local quantities in spatial point processes.

Consider a distribution $F_{\mathbf{X}}$ taking values in a (possibly highly dimensional) space $\mathcal{X}$. Define the random variable $D = d(\mathbf{X}_1, \mathbf{X}_2)$ as the dissimilarity between the two *i.i.d.* elements $\mathbf{X}_1$ and $\mathbf{X}_2$ extracted from $F_{\mathbf{X}}$. The distribution F_D of D is the interpoint distance distribution (IDD). Note that the dissimilarity function $d(\cdot, \cdot)$ can also be chosen so that it is *not* a proper distance. Bartlett[1] reported on the distribution of the interpoint distances for uniformly distributed points on the unit square and the unit circle (results originally due to Ref. 2), and he suggested measuring the deviation between the observed and the expected frequencies of the observed distances over a grid as a goodness-of-fit measure.

Consider the observed value $(\mathbf{x}_1, \ldots, \mathbf{x}_n)^T$ of an *i.i.d.* sample $(\mathbf{X}_1, \ldots, \mathbf{X}_n)^T$ extracted from $F_{\mathbf{X}}$. Borrowing from the usual terminology, one defines the empirical cumulative distribution function (ECDF) of D from the $\binom{n}{2}$ dependent interpoint distances $d_{i,j} = d(\mathbf{x}_i, \mathbf{x}_j)$, $1 \le i < j \le n$, as

$$F_n(d) = \binom{n}{2}^{-1} \sum_{1\le i<j\le n} 1(d(\mathbf{x}_i, \mathbf{x}_j) \le d), \quad d \ge 0$$

After centering by F_D and scaling by $\sqrt{n}$, the ECDF process F_n converges to a Gaussian process as n tends to infinity. A first proof can be found in Ref. 3, while another proof, which makes use of the theory of U-processes, is in Ref. 4.

If one fixes a grid of values $\mathbf{d} = (d_1, \ldots, d_k)^\mathrm{T}$, then the centered and $\sqrt{n}$-scaled random vector $\mathbf{F}_n(\mathbf{d}) = (F_n(d_1), \ldots, F_n(d_k))^\mathrm{T}$ converges in distribution to a multivariate normal random vector with known variance–covariance matrix. This motivates the construction of the *M statistic*, a Mahalanobis-like quadratic form that can be used to compare the observed vector $\mathbf{F}_n(\mathbf{d})$ to an assumed population vector $\mathbf{F}_\mathrm{D}(\mathbf{d}) = (F_\mathrm{D}(d_1), \ldots, F_\mathrm{D}(d_k))^\mathrm{T}$

$$M_n = (\mathbf{F}_n(\mathbf{d}) - \mathbf{F}_\mathrm{D}(\mathbf{d}))^\mathrm{T} \mathbf{S}^{-} (\mathbf{F}_n(\mathbf{d}) - \mathbf{F}_\mathrm{D}(\mathbf{d}))$$

Carlo F. Dondena Centre for Research on Social Dynamics and Public Policy, Bocconi University, Milan, Italy

Update based on an original article published in the Encyclopedia of Statistical Sciences, Wiley StatsRef: Statistics Reference Online © 2014 John Wiley & Sons, Ltd.

DOI: 10.1002/9781118445112.stat02525.pub2

where $\mathbf{S}^{-}$ is a generalized inverse of the estimated variance–covariance matrix of $\mathbf{F}_n(\mathbf{d})$. M_n can be used as an omnibus test to detect deviations from $\mathbf{F}_{\mathrm{D}}$.

While M_n can be shown to converge in distribution to a Chi-squared random variable, empirical experience shows that the convergence is slow. For this reason, it is often preferable to use empirical testing routines, such as Monte Carlo or permutation testing. The M_n statistic was initially developed to study spatial disease clustering and biosurveillance[4,5]. The choice of the number of bins to use in defining M_n has been studied in Ref. 6.

The method has also been adapted to the two-sample setting and made available in Stata to test for differences between the distributions of distances across two groups[7,8]. The IDD approach has also been used to analyze gene region heterogeneity associated with drug-resistance phenotype, using appropriately defined weighted dissimilarity measures between viral genetic sequences[9].

Given $\mathbf{Y}_1$ and $\mathbf{Y}_2$, two random vectors in $\mathfrak{R}^k$ with distribution function $F_{\mathbf{Y}}$, then the distributions $F_{\mathbf{Y}}$ and $F_{\mathbf{X}}$ above are the same if and only if two interpoint distance variables $d(\mathbf{Y}_1, \mathbf{Y}_2)$ and $d(\mathbf{Y}_1, \mathbf{X}_1)$ (which has the same distribution as $d(\mathbf{Y}_2, \mathbf{X}_1)$) have the same IDD of $d(\mathbf{X}_1, \mathbf{X}_2)$. Consequently, another multidimensional goodness-of-fit test can be constructed as follows. To test that a k-dimensional random sample $\mathbf{Y}_1, \mathbf{Y}_2, \dots, \mathbf{Y}_n$ follows the distribution G, one considers a triangle formed by two random selected data points $\mathbf{Y}_i$ and $\mathbf{Y}_j$ and a vector $\mathbf{X}$ with distribution G and estimates the likelihood that the sides formed by the line from $\mathbf{X}_i$ to $\mathbf{X}_j$ is the smallest, the middle, or the largest side of the triangle. Under H_0, the chances of each event are one-third[10,11].

Additional recent work further exploits the idea of studying IDDs to perform shape classification on images on the plane. The IDD is used as a shape summary of a set, and it is constructed as the distribution of the Euclidean distance between two points selected at random and uniformly within the set. Results on the continuity of the transformation from the shape-equivalence class of a set C to its IDD, the explicit relationship between IDD and covariogram, and steps toward identifiability of sets through IDDs have been developed[12]. In that reference, connections to functional data analysis and an application to classification in marine biology are also described.

The estimation of parametric models for the IDDs is addressed in Ref. 13.

Lastly, one may also view the (bounded) sampling region as itself a sample of some larger space on which a stationary point process acts. In particular, in the setting of stationary isotropic processes, one defines the K-function

$$K(t) = \lambda^{-1}E \text{ [number of further events within distance } t \text{ of an arbitrary event]}$$

where λ is the intensity of the process or the (assumed constant) expected number of events per unit of area[14]. The K-function shares some of the properties of the IDD, even though it is not a distribution function; indeed, $K(t) \to \infty$ as $t \to \infty$. For further results and applications in the point processes direction of we refer to Refs 15–19.

Related Articles

Permutation Based Inference; **Point Processes**; **Point Processes, Spatial**; **Size and Shape Analysis**; **Spatial Data Analysis**; **U-Statistics and V-Statistics (Update)**; **U-Statistics**.

References

[1] Bartlett, M.S. (1964) The spectral analysis of two-dimensional point processes. *Biometrika*, **51**, 299–311.
[2] Borel, E. (1925) *Traite du Calcul des Probabilites et de ses Applications*, vol. I, Gauthier-Villars, Paris.
[3] Silverman, B.W. (1976) Limit theorems for dissociated random variables. *Adv. Appl. Probab.*, **8**, 806–819.

DOI: 10.1002/9781118445112.stat02525.pub2

[4] Bonetti, M. and Pagano, M. (2005) The interpoint distance distribution as a descriptor of point patterns, with an application to cluster detection. *Stat. Med.*, **24** (5), 753–773.

[5] Forsberg, L., Bonetti, M., Jeffery, C., Ozonoff, A., and Pagano, M. (2005) Distance-based methods for spatial and spatio-temporal surveillance, in *Spatial and Syndromic Surveillance for Public Health*, John Wiley & Sons, Ltd, Chichester.

[6] White, L.F., Bonetti, M., and Pagano, M. (2009) The choice of the number of bins for the M statistic. *Comput. Stat. Data Anal.*, **53** (10), 3640–3649.

[7] Manjourides, J. (2009) Distance based methods for space-time modelling of the health of populations. PhD diss. Harvard School of Public Health, Department of Biostatistics.

[8] Tebaldi, P., Bonetti, M., and Pagano, M. (2011) M statistic commands: interpoint distance distribution analysis. *Stata J.*, **11** (2), 271–289.

[9] Kowalski, J., Pagano, M., and DeGruttola, V.A. (2002) A nonparametric test of gene region heterogeneity associated with phenotype. *J. Am. Stat. Assoc.*, **97**, 398–408.

[10] Maa, J.-F., Pearl, D.K., and Bartoszyński, R. (1996) Reducing multidimensional two-sample data to one-dimensional interpoint distances. *Ann. Stat.*, **24**, 1069–1074.

[11] Bartoszyński, R., Pearl, D.K., and Lawrence, J. (1997) A multidimensional goodness-of-fit test based on interpoint distances. *J. Am. Stat. Assoc.*, **92**, 577–586.

[12] Berrenderoa, J.R., Cuevasa, A., and Pateiro-Lopez, B. (2016) Shape classification based on interpoint distance distributions. *J. Multivariate Anal.*, **146**:237–247.

[13] Bonetti, M., Olson, K.L., Mandl, K.D., and Pagano, M. (2008) Parametric modelling of interpoint distance distributions, with an application to a mixture model for biosurveillance data. *Biomed. Stat. Clin. Epidemiol.*, **2** (3), 255–266.

[14] Ripley, B.D. (1976) The second-order analysis of stationary point processes. *J. Appl. Probab.*, **13**, 255–266.

[15] Silverman, B.W. and Brown, T.C. (1978) Short distances, flat triangles and Poisson limits. *J. Appl. Probab.*, **15**, 815–825.

[16] Brown, T.C. and Silverman, B.W. (1979) Rates of Poisson convergence for U-statistics. *J. Appl. Probab.*, **16**, 428–432.

[17] Ripley, B.D. (1988) *Statistical Inference for Spatial Processes*, Cambridge University Press, Cambridge.

[18] Diggle, P.J. and Chetwynd, A.G. (1991) Second-order analysis of spatial clustering for inhomogeneous populations. *Biometrics*, **47**, 1155–1163.

[19] Baddeley, A.J. and Gill, R.D. (1997) Kaplan-Meier estimators of interpoint distance distributions for spatial point processes. *Ann. Stat.*, **25**, 263–292.

DOI: 10.1002/9781118445112.stat02525.pub2