Statistics and Probability Letters 84 (2014) 215–222

Contents lists available at ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

On the interpoint distances of Bernoulli vectors

Reza Modarres *

Department of Statistics, The George Washington University, Washington, DC 20052, United States

ARTICLE INFO

Article history:
Received 17 April 2013
Received in revised form 9 September 2013
Accepted 18 October 2013
Available online 24 October 2013

Keywords:
Bernoulli
Interpoint distances
High dimension
Distribution function

ABSTRACT

We consider the squared Euclidean interpoint distances (IDs) among multivariate Bernoulli observations and determine the mean, covariance and the distribution of the IDs within a single group or across two groups. We discuss testing the equality of two distribution functions when the number of variables is large and exceeds the number of observations.

1. Introduction

The aim of this article is to examine the distribution of IDs within and between samples drawn from multivariate Bernoulli (MB) distributions. Given a sample of n_x observations, one can represent them as nodes of a complete graph with n_x vertices and $m_x = \binom{n_x}{2}$ edges whose weights are the squared IDs. The distribution of IDs constitutes the engine of many graph-based multivariate techniques such as cluster analysis and minimal spanning trees (MST). Also called the Hamming distance in information theory (Hamming, 1950), the squared IDs of Bernoulli data are often used in clustering algorithms to examine the multivariate structure of data in high dimensions (Gasieniec et al., 2004). IDs are used in gene expression (Xu et al., 2001) analysis, hotspot detection (Patil et al., 2006), and disease clustering (Assuncão et al., 2006), among others. There are well-known connections between the MST and the single-linkage clustering algorithm (Gower and Ross, 1969), both of which are functions of the IDs.

One can find numerous uses of IDs when the underlying distribution of the observations is assumed to be absolutely continuous. One reason is to avoid ties among IDs. To obtain a unique MST, one must ensure that the vertices are distinct and represent observations from an absolutely continuous distributions so that ties occur with probability zero. In practice, when the observations are drawn from discrete distributions such as MB or multinomial, ties occur among the IDS. Ties do not cause much difficulty in the testing problem we consider because the dimension is large. The MB distribution is a cornerstone of statistical modeling and there are numerous applications that require modeling binary data. Wilber et al. (2002) consider modeling and variable selection for high-dimensional MB data to evaluate the importance of microbial communities on crop productivity by their DNA profiles obtained from soil samples. These community DNA fingerprints are represented in the form of high-dimensional binary vectors. Modarres (2011) discusses methods of generating high dimension MB vectors.

Suppose $\mathbf{X} = \{\mathbf{X}_i\}$ for $i = 1, \ldots, n_x$ is a sample of independent and identically distributed random vectors in $\{0, 1\}^d$ drawn from an MB distribution with mean vector $\mathbf{P}$, covariance matrix $\mathbf{\Sigma}$ and distribution function (DF) F. We use the notation $\mathbf{X} \sim \text{MB}(\mathbf{P}, \mathbf{\Sigma})$. Similarly, suppose $\mathbf{Y} = \{\mathbf{Y}_j\}$ for $j = 1, \ldots, n_y$ is drawn from an MB($\mathbf{P}^*$, $\mathbf{\Sigma}^*$) and DF G. We also

* Tel.: +1 202 994 9991; fax: +1 202 9946917.
E-mail address: reza@gwu.edu.

0167-7152/$ – see front matter
http://dx.doi.org/10.1016/j.spl.2013.10.012

assume that the **X** and **Y** samples are independent. We are interested to test the hypotheses $H_0 : F = G$ against general alternatives $H_a : F \neq G$ when $d > \max(n_x, n_y)$. The sample covariance of high dimensional Bernoulli observations is singular and a poor estimate of the population covariance matrix in low sample size experiments. This rules out the use of classical multivariate approaches that require estimating the covariance of the observations.

There are not many articles on the exact distribution of the IDs in the literature. When observations are normally distributed, Rohlf (1975) suggests a gap test to detection outliers using the longest edge of the MST. As noted by Caroni and Prescott (1995) the assumption of the independence of the IDs is questionable. Bonetti and Pagano (2005) consider discrete observations and use the asymptotic normality of the empirical DF of the IDs evaluated at a finite number of values to detect spatial disease clusters. Jammalamadaka and Janson (1986) study the asymptotic distribution of small IDs in a sample.

The article covers one sample and two-sample cases and obtains the mean, covariance and the distribution of IDs under an MB distribution. The next section treats the one-sample case, determines the mean and covariance and obtains the probability mass function of the IDs between any two randomly selected observations from an MB distribution. Section 3 establishes the distribution of IDs across two groups and derives their means and covariances. Section 4 considers testing $H_0 : F = G$ when the number of variables exceeds the number of observations and report on a Monte Carlo study designed to examine the test statistics. The last section is devoted to summary.

2. One sample IDs

The random vector $\mathbf{X}_i = (X_{i1}, \ldots, X_{id})$ has a d-variate Bernoulli distribution with the probability mass function $\mathbf{P}^{(\mathbf{d})} = Pr(X_{i1} = k_1, \ldots, X_{id} = k_d)$ where $k_j \in \{0, 1\}$ for $j = 1, \ldots, d$ and $i = 1, \ldots, n_x$. The mean is denoted by **P**, and the covariance with $\mathbf{\Sigma}$. The vector of central moments captures all 2-way or higher level dependences that might exist between the components of **X**. The marginal distribution of any set of the components of $\mathbf{X}_i$ is MB with means, variances and covariances obtained from the corresponding elements of **P** and $\mathbf{\Sigma}$. Since the sum of the 2^d probabilities must equal one, one needs $2^d - 1$ parameters to fully describe an MB distribution. A d-variate MB distribution can be represented as a multinomial distribution with 2^d cells and $2^d - 1$ parameters. The squared ID between $\mathbf{X}_i$ and $\mathbf{X}_j$ is defined as

$$d^2_{(x)ij} = (\mathbf{X}_i - \mathbf{X}_j)'(\mathbf{X}_i - \mathbf{X}_j) = \sum_{r=1}^{d} (X_{ir} - X_{jr})^2 = \sum_{r=1}^{d} T_{(x)ijr}, \tag{1}$$

for $1 \leq i < j \leq n_x$. One can show that $d^2_{(x)ij}$ is a proper distance function since it satisfies (a) $d^2_{(x)ij} = 0$ if and only if $\mathbf{X}_{it} = \mathbf{X}_{jt}$ for $t = 1, \ldots, d$, (b) $d^2_{(x)ij} = d^2_{(x)ji}$, and (c) $d^2_{(x)ij} \leq d^2_{(x)ik} + d^2_{(x)kj}$ for any $k \neq i$ and $k \neq j$. One can establish the triangle inequality by maximizing $d^2_{(x)ij}$ and minimizing $d^2_{(x)ik} + d^2_{(x)kj}$. The maximum of $d^2_{(x)ij}$ is d, which occurs when $\mathbf{X}_{it} \neq \mathbf{X}_{jt}$ for $t = 1, \ldots, d$. The minimum value of $d^2_{(x)ik} + d^2_{(x)kj}$ is $1 + (d - 1) = d$ and occurs when $\mathbf{X}_{kt} = \mathbf{X}_{it}$ for all but one position, $t = 1, \ldots, d$, thus, forcing $\mathbf{X}_{kt}$ and $\mathbf{X}_{jt}$ to be different in $d - 1$ positions since $k \neq i$ and $k \neq j$. In other cases, $d^2_{(x)ij}$ is strictly less than $d^2_{(x)ik} + d^2_{(x)kj}$ for any $k \neq i$ and $k \neq j$.

2.1. Means and covariances

Let $\mathbf{T}_{(x)ij} = (T_{(x)ij1}, \ldots, T_{(x)ijd})$ be a vector in $\{0, 1\}^d$ corresponding to the disagreements between observations $\mathbf{X}_i$ and $\mathbf{X}_j$. One can show that $\mathbf{T}_{(x)ij} \sim \text{MB}(\boldsymbol{\theta}_{(x)ij}, \mathbf{\Gamma}_{(x)ij,ij})$ with the mean being

$$\theta_{(x)ijr} = pr(T_{(x)ijr} = 1) = 2p_r(1 - p_r), \tag{2}$$

and the variances being $\gamma_{(x)rr} = 2p_r(1 - p_r)(1 - 2p_r(1 - p_r))$. The covariances are given by

$$\text{Cov}(T_{(x)ijr}, T_{(x)ijs}) = 2\sigma_{rs}(2\sigma_{rs} + (1 - 2p_r)(1 - 2p_s)), \tag{3}$$

where $\sigma_{rs} = \text{Cov}(X_{ir}, X_{is})$ for $r, s = 1, \ldots, d$. One can obtain the higher level dependences among the elements of $\mathbf{T}_{(x)ij}$ to fully specify the MB distribution. Let $c = m_x \times d$ and let $\mathbf{T} = (\mathbf{T}_{(x)ij})$ for $1 \leq i < j \leq n_x$ be a vector in $[0, 1]^c$ composed of m_x d-dimensional vectors $\mathbf{T}_{(x)ij}$. The vector **T** has a c-dimensional MB distribution with mean composed of m_x copies of $\boldsymbol{\theta}_{(x)ij}$ and covariance matrix $\mathbf{\Gamma}_x = (\mathbf{\Gamma}_{(x)ij,kh}) = \text{Cov}(T_{(x)ijr}, T_{(x)khs})$. The main diagonal matrices of $\mathbf{\Gamma}_x$ are $\mathbf{\Gamma}_{(x)ij,ij}$ and the off-diagonals are $\mathbf{\Gamma}_{(x)ij,kh}$ with entries

$$\gamma_{(x)rs} = \text{Cov}(T_{(x)ijr}, T_{(x)khs}) = \sigma_{rs}(1 - 2p_r)(1 - 2p_s)I_C \tag{4}$$

where $I_C = 1$ if (i, j) and (k, h) share an index, $1 \leq i < j \leq n_x$, $1 \leq k < h \leq n_x$ and zero, otherwise. When $i = k$ and $j = h$ the covariance is given by Eq. (3).

There are $t_x = \binom{m_x}{2}$ pairs of IDs among the $m_x = \binom{n_x}{2}$ distances on n_x data points. One can show that there are $m_x(n_x - 2) = n_x(n_x - 1)(n_x - 2)/2$ pairs of dependent distances among the t_x pairs while the rest of the pairs are independent. Two IDs $d^2_{(x)ij}$ and $d^2_{(x)kh}$ are dependent if they have an index in common. Since $i < j$ and $k < h$, the only remaining dependence patterns are obtained when $(i = k) \vee (j = h) \vee (j = k) \vee (j = h)$. Note that $\text{Cov}(d^2_{(x)ij}, d^2_{(x)kh}) = \sum_{r=1}^{d} \sum_{s=1}^{d} \text{Cov}(T_{(x)ijr}, T_{(x)khs})$.