Modeling Minimal Spanning Trees with Beta Vectors

Haigang Liu and Reza Modarres[*]

Department of Statistics

The George Washington University

Washington, D.C. 20052

Abstract

We model the distribution of normalized interpoint distances (IDs) on the minimal spanning tree (MST) using multivariate beta vectors. We define overlapping sums of the components of a Dirichlet distribution to construct multivariate beta distributions. We also use a multivariate normal copula with beta marginals to define beta vectors. Based on the ordered IDs of the MST, we define a multivariate Gini index to measure their scatter. A simulation study compares the Gini index, the maximum and the range of the IDs with the results of modeling the distances on the MST.

Keywords

Minimal Spanning Tree; Multivariate beta; Dirichlet; Gini index; Lorenz Curve.

[*] Corresponding Author. Email: reza@gwu.edu

1 Introduction

The goal of this article is to model the distribution of normalized IDs on the minimal spanning tree (MST) using beta random vectors. The MST is a multivariate technique that helps one visualize the observations in $\mathbf{R}^p$, identify any clusters, the central vertices and observations at the outskirts of the distribution. We are interested in the sampling distribution of various statistics on the MST. We denote a random data matrix by $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)$ where $\mathbf{X}_i$ is an independent p-dimensional random vector with distribution function (DF) F. We represent the observations of a random sample in $\mathbf{R}^p$ with vertices of a complete weighted graph G. A sample is denoted by $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$, which is also used to represent the set of vertices in the graph. A weighted and connected simple graph G has a set of vertices $v(G) = \mathbf{x}$ and a set of edges $e(G) = \{(\mathbf{x}_i, \mathbf{x}_j) | 1 \leq i < j \leq n\} \subset v \times v$. We also associate a weight $w(G) : e(G) \mapsto \mathbf{R}^+$ with every edge of G. The weight of an edge is the squared Euclidean distance between the two vertices it connects. The squared interpoint distance between $\mathbf{X}_i$ and $\mathbf{X}_j$ is defined by $(\mathbf{X}_i - \mathbf{X}_j)'(\mathbf{X}_i - \mathbf{X}_j)$ for $1 \leq i < j \leq n$.

A subgraph of G with no cycles, the MST has $n-1$ edges and connects all vertices with minimal total weight. The total weight of a tree is sum of the weights of its edges. The MST is unique if the set of weights contain no ties. There is interest in obtaining the sampling distribution of various statistics on the MST such as length of the longest edge for outliers detection (Rohlf, 1975), the diameter of the tree (Addario-Berry, Broutin and Reed, 2006), the depth of the vertices for ranking observations on the tree (Small, 1997) or node degrees in the Euclidean maximum spanning tress (Willemain and Bennett, 2001). The expectation of the desired statistic with respect to all possible MSTs with vertices that are i.i.d. random vectors with distribution function F defines the corresponding population parameter on the MST.

Since the distribution of the vertices on the MST depends on F, computer simulation is a natural method for obtaining the sampling distribution of a statistic of interest. One performs the following steps a large number of times B to obtain the simulated distribution of a statistic of interest.

- Obtain the data matrix $\mathbf{X}$ by generating a random sample of size n from F.
- Compute the $m = n(n-1)/2$ IDs and build the MST.
- Calculate the value of the MST-based statistic such as the Gini index.

The sampling distribution of the MST-based statistic is based on the B values. The major drawback of the simulation approach is the computational cost of the IDs and construction of the MST. While the computational complexity of the MST is $O(n \log n)$ it takes $O(pn^2)$ time to compute all IDs. We also need to store the computed distances before constructing the MST. To side-step the computational and storage requirements we normalize the edge weights of the MST and model them with a multivariate beta distribution.

In comparison, the methods we discuss are based on estimating the parameters of the model by T runs of the simulation algorithm where T is much smaller then B. For example, the copula method uses the probability integral transformation to induce univariate beta marginal distributions on the components of a multivariate normal copula to generate a beta vector of the normalized IDs. The Dirichlet interpoint distance method fits a Dirichlet distribution to the normalized IDs. In both cases, the parameters of the beta vector are estimated by T runs of the simulation algorithm. Of course, one needs to consider the cost finding a good fit and generation of the vectors. There is also the added burden of computing the correlation of beta vectors, which involves the quantile function of the beta distribution. However, once a good model is estimated, we are no longer dependent on the computation of the IDs and construction of the MST.

The above discussion focuses on simulating an MST to measure various characteristics of the tree structure. However, the approach of modeling the edge weights applies equally well to other types of proximity graphs. Proximity graphs include the MST and serve as indispensable tools in disciplines where understanding of shape and structure are vital, including visual perception, computer vision and pattern recognition, geography, and biology. The MST, the Delaunay Triangulation, the relative neighborhood graph (Toussaint, 1980), and the Gabriel Graph (Gabriel and Sokal, 1969) are prominent representatives of proximity graphs.

There are existing multivariate modeling methods for general tree structures. For example, Kirshner (2008) proposed a tree–structured copula of multivariate distributions with uniform marginal. This model can approximate distributions with complex variable dependencies. Ma et al. (2012) proposed a nonparametric estimation algorithm for dependence tree structure learning via copula. Diimann et al. (2013) also discussed vine copula for multivariate modeling using a tree. A tree-structured dependence is sometime too restrictive. A richer dependence structure can be obtained by averaging over all possible or a sample of the tree structures. Parameter estimation for tree-structured and tree-averaged models requires optimization over univariate and bivariate densities. These steps can be costly. In comparison, the beta marginals and normal copula methods enjoy simplicity and ease of implementation. However, the normal copula model has shortcomings, as we discuss in Section 3.

In the next Section, we discuss modeling the MST in order to obtain the sampling distribution of the Gini index, the range of the IDs and the length of the longest edge on the MST. Section 3 investigates methods of obtaining beta vectors that are used to model the MST. These methods are are compared using Monte Carlo simulation in Section 4.

2 Modeling the MST

Since the MST provides a unique path that connects all vertices with minimal total weight, it finds applications in diverse areas such as biological systems (Dussert et al, 1986), two-sample problems (Friedman and Rafsky, 1979; Modarres, 2008), outliers and hotspot detection (Patil and Taillie, 2004), among others. There are well-known connections between the MST and the single-linkage clustering algorithm (Gower and Ross, 1969). The MST is defined and is useful for high dimensional data visualization and robust estimation since it only depends on the IDs. Interpoint distances are the building blocks of the MST and remain invariant after an affine (orthogonal) transformation, depending on whether Mahalanobis or Euclidean distances are used. In fact, the maximal invariant statistic under affine invariance is the set of all pairwise distances.

Rohlf (1975) proposes a gap test for the detection of multivariate outliers using the MST. Caroni and Prescott (1995, 2002) question the assumption of the independence of the edge lengths on the MST. **It is not difficult to show that two IDs that share a common vertex are dependent (Modarres, 2014)**. Moreover, the MST is based on selecting the minimum distances among m IDs, so that many distances that appear on the MST are dependent.

Prim (1957) and Kruskal (1956) provide two well-known algorithms for building an MST. Prim's algorithm, given a vertex set $\mathbf{x}$ and an edge set E, constructs the MST by starting with an empty vertex set $\mathbf{x}^*$ and edge set E^*. We start with an arbitrary vertex in $\mathbf{x}^*$. Maintaining a tree at all stages, the algorithm proceeds to identify the shortest edge from $\mathbf{x}$ to $\mathbf{x}^*$. This edge is next added to E^*. The corresponding vertex is removed from vertex $\mathbf{x}$ and added to $\mathbf{x}^*$. The algorithm terminates when all vertices are in $\mathbf{x}^*$. Kruskal's algorithm constructs the MST by iteratively adding the next shortest edge that does not creates a cycle to the tree. The algorithm ends when all $n-1$ edges are found. The major difference between the two algorithms is that Prim's maintains a connected tree in all stage of construction while Kruskals' algorithm deals with a forest that eventually merges into a single tree.

In order to model the MST, we will normalize its edge weights. Denote the ordered squared distances on the MST by e_1^2,..., e_{n-1}^2, and define the ordered normalized edge weights with

$$d_i = \frac{e_i^2}{\sum_{i=1}^{n-1} e_i^2}, \quad i = 1, 2, \ldots, n-1.$$

The normalizing process maps the edge weights in the MST to variables on $(0, 1)$

interval to facilitate modeling by a multivariate beta distribution.

To measure the amount of scatter among the distances on the MST we define a multivariate Gini index on the normalized MST. The empirical Lorenz curve was proposed by Lorenz (1905) to measure wealth inequality. Suppose the ordered normalized distances on the MST are denoted with $d_1 < d_2 < \ldots < d_{n-1}$ and $s_i = \sum_{j=1}^{i} d_i$ for $i = 1, \ldots, n-1$. With the initial point at the origin $(0,0)$, the Lorenz curve of the normalized distances is the continuous piece-wise function that linearly interpolates the points $(u_i, L_d(u_i))$ where $u_i = \frac{i}{n-1}$ and $L_d(u_i) = s_i/s_{n-1}$ for $i = 1, \ldots, n-1$. The Lorenz curve ordinate of the normalized distance is the scaled partial integral of the inverse empirical distribution function or $L_d(u) = \frac{1}{\mu_d}\int_0^u H_d^{-1}(t)dt$ where $H_d^{-1}(t)$ is the quantile function and $0 \leq u \leq 1$. The Lorenz curve is nondecreasing with $L_d(0) = 0$ and $L_d(1) = 1$. The Lorenz curve can be used to display the discrepancy between the normalized IDs of the MST. Twice the area between $L_d(u) = u$ and the Lorenz curve equals the Gini index.

We will next study methods of generating beta random vectors in order to model the MST and obtain the distribution of Gini index, the range of the IDs and the length of the longest edge on the MST

3 Multivariate Beta Models

We consider a copula-based method that induces univariate beta marginal distributions on the components of a multivariate normal copula. We also propose two methods based on the Dirichlet distribution and the ordered Dirichlet distribution. The proofs for some of the results appear in the Appendix. The univariate beta distribution naturally arises as the ratio of two independent χ^2 random variables. The strength of the beta distribution lies in its flexibility since its shape can be unimodal, right-skewed, left skewed or symmetric depending on two parameters. The multivariate Dirichlet distribution, defined on the simplex and with negatively correlated univariate beta marginals, is often used to model dependent proportions.

The probability density function of the beta distribution with parameter a and b is $h(z) = \frac{1}{B(a,b)} z^{a-1}(1-z)^{b-1}$ for $0 \leq z \leq 1$ and $a, b > 0$ where $B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$. Kotz, Balakrishnan and Johnson (2000) devote a chapter to the exploration of the beta distribution. We use the following properties to develop beta vectors. If Z has a beta(a,b) distribution, then $1 - Z \sim$ beta(b,a). The Kth moment of the beta distributions, the Mellin transformation of Z, is $E(Z^k) = \frac{\Gamma(a+k)\Gamma(a+b)}{\Gamma(a)\Gamma(a+b+k)}$. Hence, $E(Z) = \frac{a}{a+b}$ and Var$(Z) = \frac{ab}{(a+b)^2(a+b-1)}$.

3.1 Copula Method

The following copula-based method uses the probability integral transformation to induce univariate beta marginal distributions on the components of a multivariate normal copula. Suppose $\mathbf{\Sigma}$ is the correlation matrix of the multivariate normal distribution with distribution function Φ. Let $\mathbf{Y}_i \sim N_p(\mathbf{0}, \mathbf{\Sigma})$ and $\mathbf{U}_i = \Phi(\mathbf{Y}_i)$, where $i = 1, \ldots, n$. It follows that $\mathbf{U}_i$ has a p-variate uniform distribution with correlation matrix $\mathbf{\Sigma}^*$, which is slightly different from $\mathbf{\Sigma}$. Let H^{-1} be the quantile function of beta$(\mathbf{a}, \mathbf{b})$. We obtain a beta vector by $\mathbf{Z}_i = H^{-1}(\mathbf{U}_i)$ for $i = 1, \ldots, d$.

A linear approximation bivariate normal copula density reveals that the off-diagonal elements of the correlation matrix $\mathbf{\Sigma}$ of $\mathbf{X}_i$ and the correlation matrix $\mathbf{\Sigma}^*$ of $\mathbf{U}_i$ do not differ by more than 0.05. The proof of the following result appears in the Appendix.

Theorem 1. Let $\mathbf{Y} = (Y_1, Y_2, ..., Y_p) \sim N(\mathbf{0}, \mathbf{\Sigma})$ be a p-dimensional normal random vector with mean vector $\mathbf{0}$ and correlation matrix $\mathbf{\Sigma}$. If $\mathbf{U} = (\Phi(Y_1), \Phi((Y_2)$ $, ..., \Phi(Y_p))$, then Corr$(U_j, U_k) \approx \frac{3}{\pi}$Corr$(Y_j, Y_k)$.

Let ρ_{ij} be an element of $\mathbf{\Sigma}$ and ρ^*_{ij} be an element of $\mathbf{\Sigma}^*$ for $i \neq j = 1, \ldots, d$. It is well-known that $\rho^*_{ij} = \frac{6}{\pi} \arcsin(\frac{\rho_{ij}}{2})$. Thus, if we want to obtain a vector of uniform variables with correlation matrix which is exactly the same with $\mathbf{\Sigma}$, we can first generate correlated normal variates with correlation matrix $\tilde{\mathbf{\Sigma}} = 2 \sin\left(\frac{\pi}{6}\mathbf{\Sigma}\right)$. Note that $\tilde{\mathbf{\Sigma}}$ is not guaranteed to be positive definite even when $\mathbf{\Sigma}$ is positive definite.

Li and Hammond (1975) present a general method for generating a random vector with prescribed marginal probability distributions and correlation matrix. They develop a numerical algorithm to determine the correlation of the induced beta variates. The algorithm in their equation (8) requires the quantile function of the beta variates.

3.2 Dirichlet Normalized Interpoint Distances

Suppose $W_i \sim \Gamma(a_i, 1)$ for $i = 0 \ldots d$ are independent where $0 < W_i < \infty$ and let $W = \sum_{i=0}^{p} W_i$. By the additive property of gamma variables, it follows that $W \sim \Gamma(\sum_{i=0}^{p} a_i, 1)$. For $i = 1 \ldots d$, if we set $Y_i = W_i/W$, then $(Y_1 \ldots Y_p)$ has a Dirichlet distribution with parameters $(a_1, a_2, \ldots, a_p)$. Note that $\sum_{i=0}^{p} Y_i = 1$ and $0 \leq Y_i \leq 1$. If $(Y_1, Y_2, \ldots, Y_p)$ has a Dirichlet distribution, then the marginal distribution of Y_i is beta$(a_i, a - a_i)$, where $a = \sum_{i=1}^{p} a_i$. The expectation, variance and covariance of the marginal distributions are $E(Y_i) = a_i / \sum a_i$, $\text{Var}(Y_i) = \frac{a_i(a-a_i)}{a^2(a+1)}$, and $\text{Cov}(Y_i, Y_j) = \frac{-a_i a_j}{a^2(a+1)}$, respectively. In fact, the sum of any subset of $(Y_1, Y_2, \ldots, Y_p)$ still has a beta distribution. That is, $\sum_{i \in A} Y_i \sim \text{beta}(\sum_{i \in A} a_i, a - \sum_{i \in A} a_i)$ where A can be any subset of $S = \{1, 2, 3 \ldots, d\}$.

Since the sum of the marginal variables of a Dirichlet distribution has a beta distribution, we can construct correlated variables that share common marginal variables. For example, assume Y_1 and Y_2 are marginal variables from a Dirichlet distribution. Define $S_1 = Y_1$ and $S_2 = Y_1 + Y_2$ and note that $S_1 \sim \text{beta}(a_1, a - a_1)$ and $S_2 \sim \text{beta}(a_1 + a_2, a - a_1 - a_2)$. It follows that $\text{Cov}(S_1, S_2) = \frac{a_1(a-a_1-a_2)}{a^2(a+1)}$. The general case is stated below.

Property 1. If Y_i are the marginal variables from the Dirichlet distribution with parameters $(a_1, a_2, \ldots, a_p)$, then $S_i = \sum_{i \in A_p} Y_i$ is distributed as beta$[(\sum_{i \in A_p} a_i), (a - \sum_{i \in A_p} a_i)]$ and $S_j = \sum_{i \in A_q} Y_j$ is distributed as beta$[(\sum_{i \in A_q} a_i), (a - \sum_{i \in A_q} a_i)]$, where $A_p, A_q \subseteq S = \{1, 2, 3, \ldots, d\}$. If $A_p \cap A_q \neq \emptyset$, then S_i and S_j are correlated.

Table 1 provides several examples of S_i and S_j, where Y_i's are are marginal variables of the Dirichlet distribution. The correlation coefficients of S_i and S_j can be positive or negative.

S_1	S_2	$Dir(2,4,7,3)$	$Dir(0.2,0.2,0.5,0.5)$	$Dir(1,4,0.2,0.6)$
Y_1	Y_1+Y_2	0.500	0.650	0.163
Y_1	$Y_1+Y_2+Y_3$	0.245	0.296	0.147
Y_1	$Y_1+Y_2+Y_3+Y_4$	0.049	0.069	0.005
Y_1+Y_2	$Y_1+Y_2+Y_3+Y_4$	-0.013	0.044	0.054
Y_1+Y_2	$Y_2+Y_3+Y_4$	-0.500	-0.650	-0.164
Y_2+Y_3	$Y_2+Y_3+Y_4$	0.537	0.431	0.721

Table 1: Correlations of overlapping sums from the Dirichlet distribution

3.3 Ordered Dirichlet distribution

Wilks (1962) introduced the p-variate ordered Dirichlet distribution with probability density

$$f(z_1, z_2, \ldots, z_p) = \frac{\Gamma\left(\sum_{i=1}^{p+1}\theta_i\right)}{\prod_{i=1}^{p+1}\Gamma(\theta_i)}\prod_{i=1}^{p+1}(z_i - z_{i-1})^{\theta_i - 1},$$

where $z_0 = 0$, $z_{p+1} = 1$, $0 < z_{i-1} < z_i < 1$, $i = 1, \ldots, p$ and $\theta_i > 0$ for $i = 1, \ldots, p+1$. While the Dirichlet distribution is defined on a simplex, the ordered Dirichlet distribution is defined on the upper pyramidal cross section of the unit hyper cube. **The support of a bivariate Dirichlet distribution is $0 \leq z_1 + z_2 \leq 1$ and the support of a bivariate ordered Dirichlet distribution is $0 \leq z_1 \leq z_2 \leq 1$.**

A parameterization of the Dirichlet distribution by van Dorp and Mazzuchi (2003) is more insightful when we consider the Ordered Dirichlet (OD) distribution. Let $\beta = \sum_{i=1}^{p+1}\theta_i$ and $\alpha_i = \frac{\theta_i}{\beta}$ for $i = 1, \ldots, p$. Let $c_1 = \prod_{i=1}^{p}\Gamma(\beta\alpha_i)$ and $c_2 = \Gamma\left(\beta(1 - \sum_{i=1}^{p}\alpha_i)\right)$. The probability density of the ordered Dirichlet distribution can be represented as

$$\frac{\Gamma(\beta)}{c_1 c_2} z_1^{\beta\alpha_1 - 1}\left\{\prod_{i=2}^{p}(z_i - z_{i-1})^{\beta\alpha_i - 1}\right\}(1 - z_p)^{\beta(1 - \sum_{i=1}^{p}\alpha_i) - 1}.$$

Hence, $\mathbf{Z} = (Z_1, Z_2, \ldots, Z_p)$ has an OD$(\boldsymbol{\alpha}, \beta)$ distribution, where $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_p)$. Furthermore, Z_i has a beta$(\sum_{k=1}^{i} \alpha_k, \beta)$ distribution. One can show the correlation coefficient of any two marginal variables is $\mathrm{Corr}(Z_i, Z_j) = \sqrt{\frac{\sum_{k=1}^{i} \alpha_k (1 - \sum_{t=1}^{j} \alpha_t)}{\sum_{k=1}^{j} \alpha_k (1 - \sum_{t=1}^{i} \alpha_t)}} > 0$.

Property 2. Suppose $Y_i \sim$ beta(a_i, b_i), for $i = 1, 2, \ldots, d$, where $b_i = b_{i-1} + a_{i-1}$. Let $Z_i = 1 - Y_1 \prod_{j=2}^{i} (1 - Y_j)$ where $Z_1 = 1 - Y_1$. The vector $\mathbf{Z} = (Z_1, Z_2, \ldots, Z_p)$ has beta marginals, i.e., $Z_i \sim$ beta$(b_1 + \sum_{j=2}^{i} a_j, a_1)$ where $0 < Z_1 < Z_2 \ldots < Z_p < 1$.

4 Monte Carlo Study

We report on a Monte Carlo study designed to assess the effectiveness of beta generation methods for modeling the edge weights of the MST. We compare normal copula and the Dirichlet generation methods and use the simulated distribution as the gold-standard.

One can use simulation to obtain the sampling distribution of normalized distances. We generate random samples of 10 observations from $N_2(\mathbf{0}, \mathbf{I})$ and use the Kruskal's algorithm to find the MST. **We fit each normalized distance numerically with a beta distribution by minimizing a goodness of fit distance defined by Cramér-von Mises criterion. In order to assess the fit, we performed the Anderson-Darling, the Cramér-von Mises, and the Kolmogorov-Smirnov goodness of fit tests on each fitted beta distribution. These tests confirm that beta distribution provides a good model in all cases. Alternatively, one can compute the maximum likelihood estimates of the distribution parameters with the R base `optim`. Direct optimization of the log-likelihood is performed with the Nelder-Mead method for distributions characterized by more than one parameter. Both methods of fit give similar and results. The folllowing discussion is based on estimates obtained by the goodness of fit. In Figure ??, we show the histograms of the ordered distances d_1 through d_9 on the MST.** The curves of fitted beta distribution is displayed in dotted lines, which demonstrates a good match for the marginal variables. The goodness of overall visual fit is confirmed by Table **??**.

We use the Dirichlet distribution to model the $n-1$ **normalized edge weights,** $d_1, \ldots, d_9$**, on the MST. For the** i**th variable, the mean of the Dirichlet distribution is given by** a_i/a **where** $a = \sum a_i$**. We denote the sample mean of the** i**th edge using** T **runs with** $\bar{d}_i$ **and set** $\bar{d}_i = a_i/a$ **to determine** $\hat{a}_i$**s. The solution is not unique, and we impose a constraint of** $a = 100$ **to force a unique solution. This choice of** $a = 100$ **works well and the Dirichlet method gives good performance under this constraint.**

We repeat the experiment with samples of 10 observations from Log-normal$_2(\mathbf{0}, \mathbf{I})$. In Figure **??**, we plot 9 marginal variables from the shortest edges to the longest edges and fit a beta distribution for each marginal variable. The beta distribution fits the marginal variables well, even though the assumption is changed. In Table **??**, the three statistics and the modeling methods are compared side by side in terms of expectation and variance.

For the copula-based method, we use the normal copula and probability integral transformation to model edges of the MST. We generate 10 observations from $N_2(\mathbf{0}, \mathbf{I})$**, constructed the MST and obtained** $\mathbf{d} = (d_1, d_2, \ldots, d_9)$**. To induce the dependency structure, we formed** $\mathbf{Y} = \mathbf{S}^{1/2}\mathbf{d}$**, where S is the covariance matrix of the 9 nine normalized edges weights obtained by** T **simulation runs (see below). We then apply probability integral transformation to obtain correlated uniform vectors as explained in Section 3.1. Finally, we utilize the inverse beta function to generate a beta vector.**

We obtain estimates of the parameters by T runs of the simulation algorithm described in the introduction where T is much smaller then B. **For example, if one typically uses** $B = 10000$ **runs, we use** $T = 50$ **runs to estimate the parameters.** We judge the performance of the two methods above by a) the MST-based Gini index, b) MST-based range of the IDs and c) the length of the longest edge on the MST. We also examine the performance of beta marginals, which is independent of correlation structure. We replicate the experiment 1000 times and report the mean and variance of the statistics.

The mean and variance of these three statistics of interest appear in Tables **??** and **??**. It is clear that the marginal beta distributions provide good substitutes for the simulated distributions of the IDs on the MST. The resulting distributions of statistics Gini index, the longest edge and the range, using the beta marginals also provide good approximations to simulated distribution. Density plots of the three statistics of interest are provided in the top three panels of Figures **??** and **??**, where the dotted lines denote simulated distribution and the solid lines designate the fitted beta marginals.

The above simulation study shows that the normal copula model with beta marginals provides the best overall model for the distribution of the IDs on the minimal spanning tree when the observations (vertices) follow a normal or log-normal distribution. **It is noteworthy that the copula method suffers from two disadvantages. First, as noted earlier, $\hat{\mathbf{\Sigma}}$ may not be positive definite. While it has a numerical solution, the computation of the correlations of the induced Beta vector is an added burden to modeling the normalized MST. More complex tree-structure modeling techniques such as the methods proposed by Kirshner (2008) and Diimann et al. (2013) are available in cases when a good fit is not available by the method of beta marginals .**

Acknowledgment
The authors would like to thank two anonymous referees and the Editor for comments that led to an improved manuscript.

References

[1] Addario-Berry, L., Broutin, N. and Reed, B. (2006). The diameter of the minimum spanning tree of a complete graph. Proceedings of the Fourth Colloquium on Mathematics and Computer Science, Ed. P. Chassaing, 237-248. Nancy: Discrete Mathematics and Theoretical Computer Science.

[2] Caroni, C. and Prescott, P. (1995). On Rohlf's method for the detection of outliers in multivariate data. Journal of Multivariate Analysis, 52, 295-307.

[3] Caroni, C. and Prescott, P. (2002). Inapplicability of asymptotic results on the minimal spanning tree in statistical testing. Journal of Multivariate Analysis, 83, 487-492.

[4] Diimann J., Brechmann E. C., Czado C., Kurowicka, D. (2013). Selecting and estimating regular vine copulae and application to financial returns. Computational Statistics and Data Analysis, 59, 52-69.

[5] Dussert, C., Rasigni, M., Palmari, J., Rasigni, G., Llebaria, A. and Marty, F. (1986). Minimal spanning tree analysis of biological structures. Journal of Theoretical Biology, 125, 317-323.

[6] Friedman, J. H. and Rafsky, L. C. (1979). Multivariate generalization of the Wald-Wolfowitz and Smirnov two-sample tests. The Annals of Statistics, 7, 697-717.

[7] Gabriel, R. K. and Sokal, R. R. (1969). A new statistical approach to geographic variation analysis. Systematic Biology, 18, 259-278.

[8] Gower, J. and Ross, G. (1969). Minimum spanning trees and single linkage cluster analysis. Applied Statistics, 18, 54-64.

[9] Kibble, W. F. (1945). An extension of theorem of Mehler on Hermite polynomials. Proceedings of the Cambridge Philosophical Society, 41,12-15.

[10] Kirshner, S. (2008). Learning with Tree-Averaged Densities and Distributions. Advances in Neural Information Processing Systems. Editors: J.C. Platt, D. Koller, Y. Singer and S.T. Roweis, 20, 761-768.

[11] Kotz, S., Balakrishnan, N. and Johnson, N.L. (2000). Continuous Multivariate Distributions. 2nd Edition, Wiley Press, New York.

[12] Kruskal, J. B. (1956). On the shortest spanning subtree of a graph and the traveling salesman problem. Proc. Amer. Math. Soc. 7, 48-50.

[13] Li, S. D. and Hammond, J. L. (1975). Generation of pseudorandom numbers with specified univariate distributions and correlation coefficients. IEEE transactions on Systems, man, and Cybernetics, 5, 557-561.

[14] Lorenz, M. O. (1905). Methods of measuring the concentration of wealth. Journal of the American Statistical Association, ,9, 208-219.

[15] Ma, J., Sun, Z. Q, Chen, S. and Liu, H. H. (2012). Dependence Tree Structure Estimation via Copula. International Journal of Automation and Computing, 9, 2, 113-121.

[16] Modarres, R. (2008). Tests of bivariate Exchangeability. International Statistical Review, Vol. 76, 203-213.

[17] Modarres, R. (2014). On the interpoint distances of Bernoulli vectors. Statistics and Probability Letters, 84, 215-222.

[18] Patil, G. P. and Taillie, C. (2004). Upper level set statistic for detecting arbitrarily shaped hotspots. Environmental and Ecological Statistics 11, 183-197.

[19] Prim, R. C. (1957). Shortest connection networks and some generalizations. Bell System Tech. J., 36, 1389-1401.

[20] Rohlf, F. J. (1975). Generalization of the gap test for the detection of multivariate outliers. Biometrics, 31, 93-101.

[21] Small, C. (1997). Multidimensional medians arising from geodesics on graphs. The Annals of Statistics, 25, 478-494.

[22] Stuart, A. and Ord, K. (1994). The Advanced Theory of Statistics. Volume 1, sixth edition. Wiley. New York. NY.

[23] Toussaint, G. T. (1980). The relative neighborhood graph of a finite planar set. Pattern Recognition, 12, 261-268.

[24] Van Dorp, J. and Mazzuchi, T. A. (2003). Parameter specification of the beta distribution and its Dirichlet extensions utilizing quantiles. Beta Distribution and its Application, 29(1), 1-37.

[25] Wilks, S.S. (1962). Mathematical statistics. Wiley. New York, NY.

[26] Willemain, T. R. and Bennett, M. V. (2002). The Distribution of Node Degree in Maximum Spanning Trees. Journal of Statistical Computation and Simulation, 72, 2, 101-106

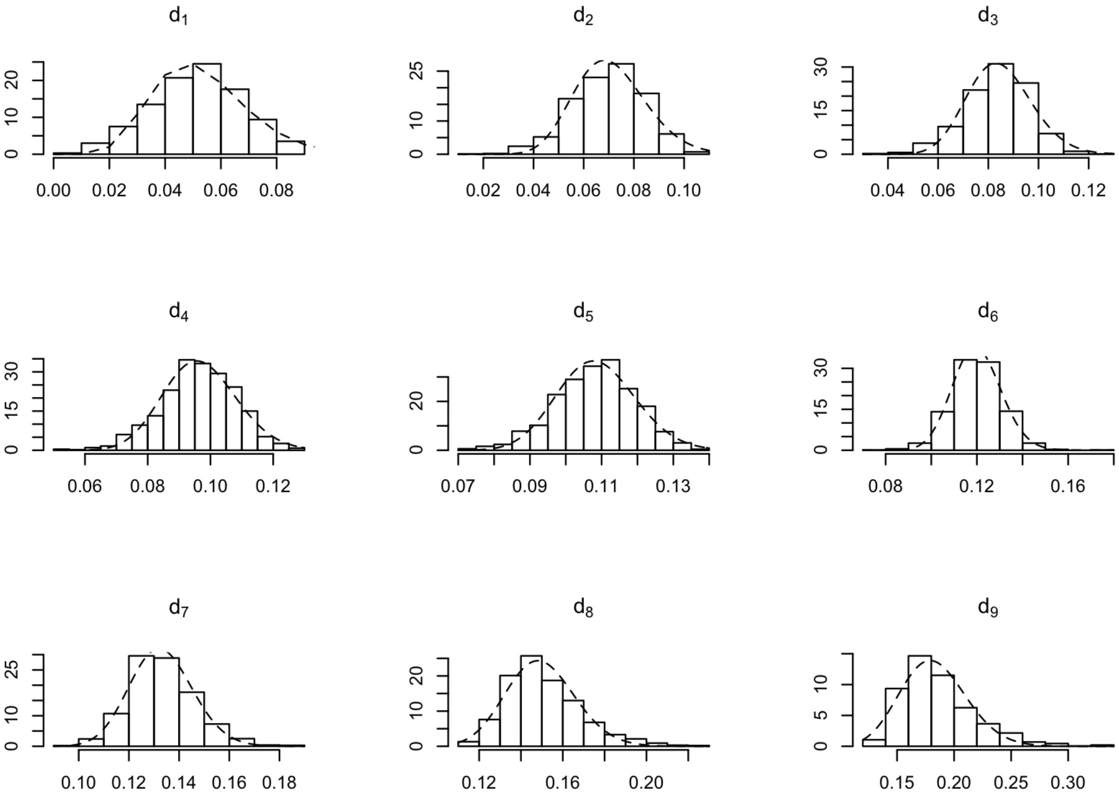

Figure 1: Histograms of the ordered distances $d_1, \ldots, d_9$ from $N_2(\mathbf{0}, \mathbf{I})$.

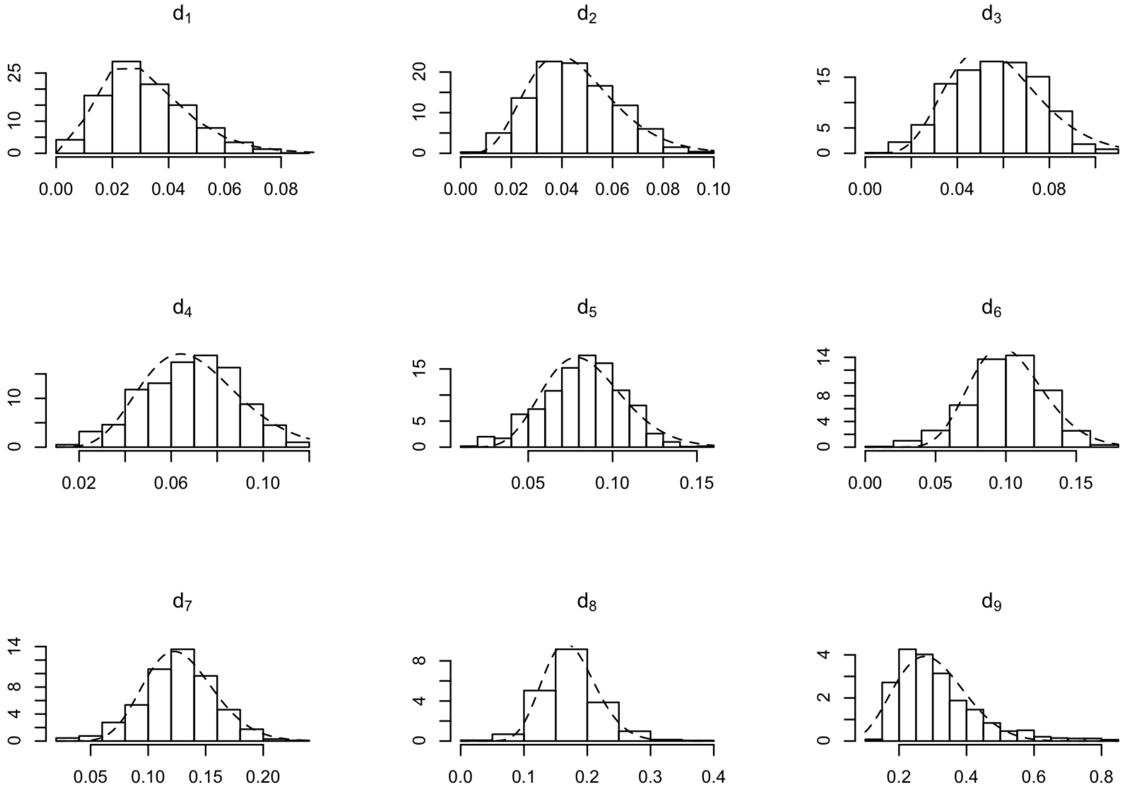

Figure 2: Histograms of the ordered distances $d_1, \ldots, d_9$ from Log-Normal$_2(\mathbf{0}, \mathbf{I})$.

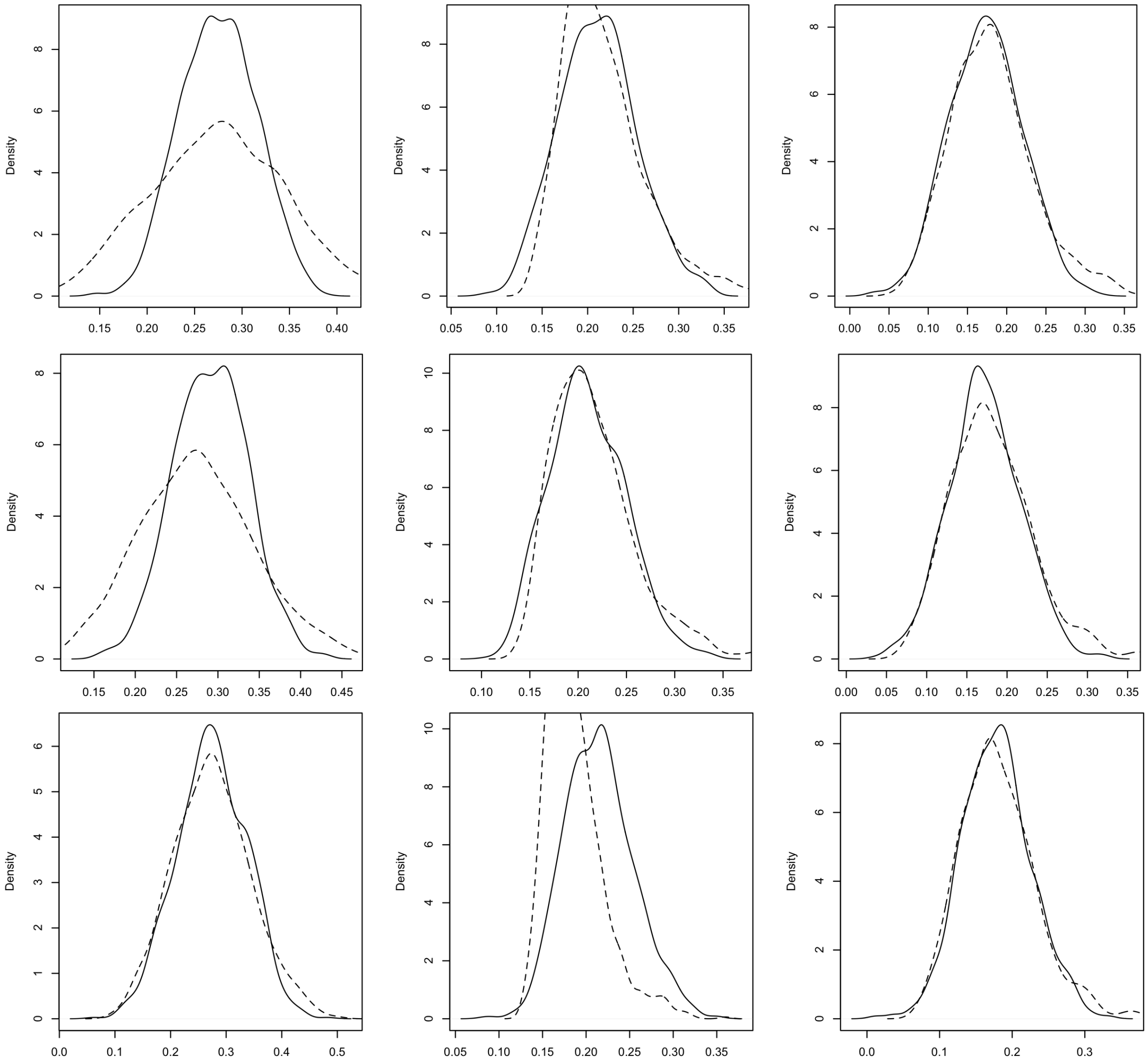

Figure 3: Normal observations: Density plots for fitted beta (top panels), Dirichlet (middle panels) and the normal copula (bottom panels) methods. From left to right: the density plots of Gini index, the longest edges and the range. Dotted lines show the simulated model.

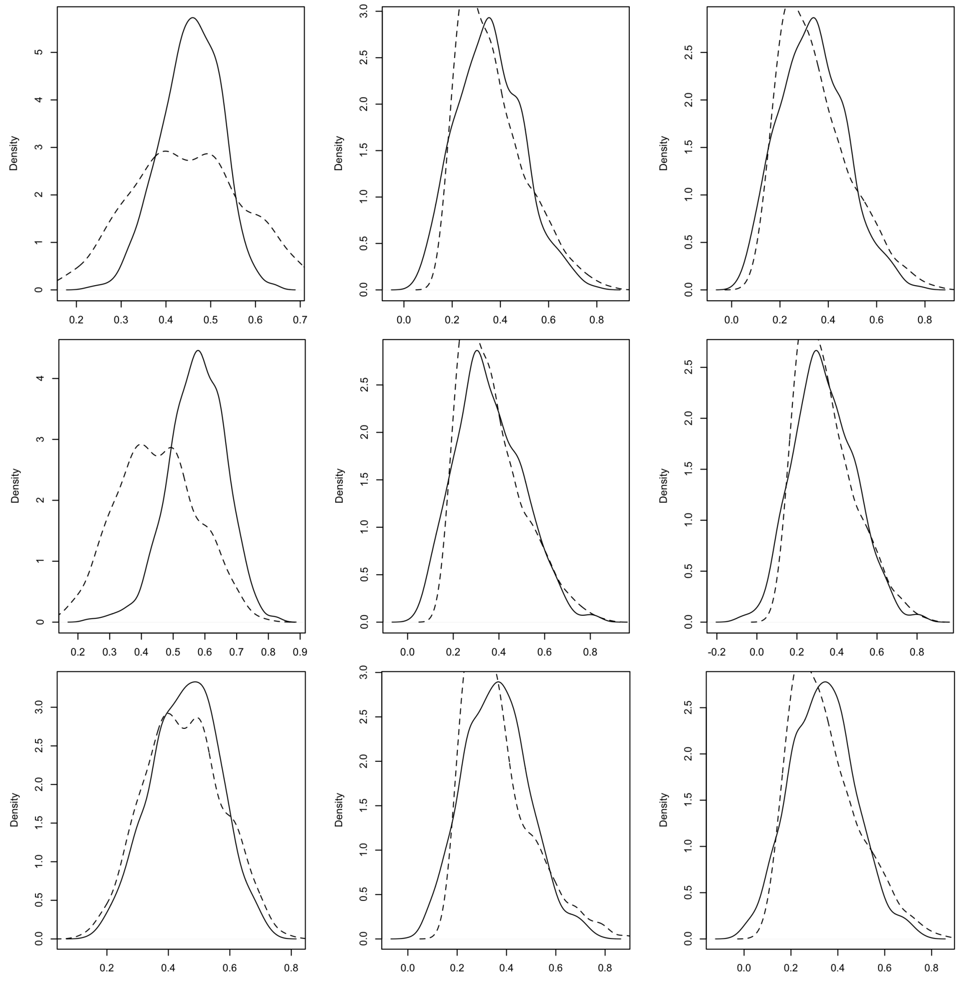

Figure 4: Log-Normal observations: Density plots for fitted beta (top panels), Dirichlet (middle panels) and the normal copula (bottom panels) methods. From left to right: the density plots of Gini index, the longest edges and the range. Dotted lines show the simulated model.

Distance	SM Mean	Beta Mean	SM Variance	Beta Variance
d_1	0.0366	0.0355	0.00035	0.00029
d_2	0.0570	0.0556	0.00035	0.00029
d_3	0.0742	0.0728	0.00027	0.00027
d_4	0.0892	0.0881	0.00026	0.00025
d_5	0.1046	0.1037	0.00022	0.00023
d_6	0.1210	0.1205	0.00022	0.00024
d_7	0.1405	0.1406	0.00030	0.00033
d_8	0.1662	0.1676	0.00051	0.00057
d_9	0.2111	0.2153	0.00164	0.00221

Table 2: Normal observations: Simulated model (SM) and the fitted beta distribution (Beta) for the MST distances.

Distance	SM Mean	Beta Mean	SM Variance	Beta Variance
d_1	0.0213	0.0229	0.00019	0.00021
d_2	0.0342	0.0362	0.00026	0.00031
d_3	0.0455	0.0490	0.00036	0.00038
d_4	0.0580	0.0611	0.00046	0.00048
d_5	0.0731	0.0774	0.00062	0.00081
d_6	0.0931	0.0969	0.00093	0.00095
d_7	0.1237	0.1283	0.00140	0.00151
d_8	0.1826	0.1790	0.00301	0.00262
d_9	0.3685	0.3527	0.01872	0.01554

Table 3: Log-Normal observations: Simulated model (SM) and the fitted beta distribution (Beta) for the MST distances.

	Gini index		Longest edge		Range	
	Mean	Variance	Mean	Variance	Mean	Variance
Simulated Model	0.272	0.0053	0.215	0.0022	0.180	0.0030
Beta Marginals	0.270	0.0015	0.210	0.0017	0.173	0.0020
Dirichlet method	0.312	0.0011	0.276	0.0017	0.238	0.0018
Normal copula	0.270	0.0045	0.211	0.0017	0.176	0.0024

Table 4: Normal observations: Simulated Model, fitted beta distributions, Dirichlet and the normal copula methods for the Gini index, the longest edge, and the range of the MST interpoint distances.

	Gini index		Longest edge		Range	
	Mean	Variance	Mean	Variance	Mean	Variance
Simulated Model	0.450	0.0151	0.368	0.0187	0.347	0.0206
Beta Marginals	0.442	0.0045	0.357	0.0167	0.334	0.0167
Dirichlet method	0.560	0.0077	0.346	0.0200	0.322	0.0237
Normal copula	0.440	0.0120	0.352	0.0154	0.329	0.0170

Table 5: Log-Normal observations: simulated model, fitted beta distributions, Dirichlet and the normal copula methods for the Gini index, the longest edge, and the range of the MST interpoint distances.

Appendix

Proof of Theorem 1: We show that the Theorem holds for bivariate normals and the proof for general p-dimensional multivariate normals follows. Let $\rho = Corr(Y_1, Y_2)$. Consider the Mehler's series (Kibble, 1945) $\phi(y_1, y_2) = \phi(y_1)\phi(y_2)\big\{1 + \sum_{t=1}^{\infty} \frac{1}{t!}\rho^t H_t(y_1)H_t(y_2)\big\}$, where $\phi(y_1, y_2)$ is the p.d.f. of standard bivariate normals, $\phi(y_1)$ is the p.d.f. of univariate standard normal and H_t is the t-th Hermite polynomial. It follows that

$$E\big[\Phi^m(y_1)\Phi^n(y_2)\big] = \frac{1}{m+1}\frac{1}{n+1} + \sum_{t=1}^{\infty} \frac{1}{t!}\rho_{ij}^t E\big[\Phi^m(y_1)H_t(y_1)\big]E\big[\Phi^n(y_2)H_t(y_2)\big]. \quad (1)$$

Substituting Equation (**??**) when $n = m = 1$ into $\text{Cov}(U_1, U_2) = E[U_1U_2] - \frac{1}{4}$, one obtain $\text{Cov}(U_1, U_2) = \sum_{t=1}^{\infty} \frac{1}{t!}\rho^t E^2\big[\Phi(y_1)H_t(y_1)\big]$.

There is no closed form solution for $E\big[\Phi(y_1)H_t(y_1)\big]$. Let $T(t) = E^2\big[\Phi(y_1)H_t(y_1)\big]$. We use a 20-degrees Hermite-Gauss quadrature to obtain the numerical values $T(1) = 1$, $T(2) = -5.2\text{E-}9$, $T(3) = -0.14104$, and $T(4) = -2.4\text{E-}7$. The terms $T(2)$ and $T(4)$ are practically zero while $|\frac{1}{3!}\rho^3 T^2(3)| \leq 0.003$. Hence, it is suitable to use a first order approximation for $E^2\big[\Phi(y_1)H_t(y_1)\big]$. That is, $\text{Cov}(U_1, U_2) \approx \rho E^2\big[\Phi(y_1)H_1(y_1)\big]$ where the first order Hermite polynomial is $H_1(y) = y$. One observes that if $Y \sim N(0, 1)$, then $E[Y\Phi(Y)] = \frac{1}{2\sqrt{\pi}}$ and $\mathrm{Cov}(U_1, U_2) \approx \rho E^2\big[\Phi(y_1)y_1\big] = \frac{1}{4\pi}\rho$. Consequently,

$$\mathrm{Corr}(U_1, U_2) = \frac{\mathrm{Cov}(U_1, U_2)}{\sqrt{\mathrm{Var}(U_1)\mathrm{Var}(U_2)}} \approx 12\frac{1}{4\pi}\rho = \frac{3}{\pi}\rho \approx 0.955\rho.$$

Since $|\rho| \leq 1$, the error of the approximation is below 0.05 under the first order approximation. Furthermore, if $\rho = 0$, then $\mathrm{Corr}(U_1, U_2) = 0$. Finally, one observes that if $\mathbf{Y} = (Y_1, Y_2, ..., Y_p)$, then the marginal distribution of (Y_i, Y_j) is bivariate normal with correlation matrix $\left(\begin{smallmatrix} 1 & \rho_{ij} \\ \rho_{ij} & 1 \end{smallmatrix}\right)$ and the proof for the general p-dimensional case follows. □

Proof of Property 1

We decompose A_p and A_q as $A_p = M \cup C$ and $A_q = N \cup C$, where C is the overlapping set. Note that $M \cap C = \emptyset$ and $N \cap C = \emptyset$. It follows that $S_i = T_M + T_C \sim$ beta$[(\sum_{i\in M\cup C} a_i), (a - \sum_{i\in M\cup C} a_i)]$ and $S_j = T_N + T_C \sim$ beta$[(\sum_{i\in N\cup C} a_i), (a - \sum_{i\in N\cup C} a_i)]$ are correlated. Each component in the decomposition has a beta distribution, namely, $T_M, T_N,$ and T_C have beta$[\theta_M, a - \theta_M]$, beta$[\theta_N, a - \theta_N]$ and beta$[\theta_C, a - \theta_C]$ distributions, respectively, where $T_M = \sum_{i\in M} Y_i$, $T_N = \sum_{i\in N} Y_i$, $T_C = \sum_{i\in C} Y_i$, $\theta_M = \sum_{i\in M} a_i$, $\theta_N = \sum_{i\in N} a_i$ and $\theta_C = \sum_{i\in C} a_i$.
The covariance of S_i and S_j is expressed as

$$\begin{aligned}\mathrm{Cov}(S_i, S_j) &= E(S_iS_j) - E(S_i)E(S_j)\\ &= \frac{\sum_{k\in C} a_k(a - \sum_{k\in C} a_k)}{a^2(a+1)} + \left(\frac{\sum_{k\in C} a_k}{a}\right)^2\\ &\quad - \frac{1}{a(a+1)}\left(\sum_{k\in C}\sum_{j\in N} a_k a_j + \sum_{i\in M}\sum_{k\in C} a_k a_i + \sum_{i\in M}\sum_{j\in N} a_i a_j\right)\\ &\quad - \frac{\sum_{k\in C} a_k + \sum_{i\in M} a_i}{a}\frac{\sum_{k\in C} a_k + \sum_{j\in N} a_j}{a}.\end{aligned}$$

Proof of Property 2

The following property and its generalization appear in Stuart and Ord (1994). If $W_1 \sim$ beta(a, b) and $W_2 \sim$ beta$(a + b, c)$ are independent, then

$$W_1W_2 \sim \text{beta}(a, b + c). \tag{2}$$

Consider Z_1 and $Z_1(1-Z_2)$. Since $1-Z_2 \sim \text{beta}(a_1+b_1, a_2)$, one can show that $Z_1(1-Z_2) \sim \text{beta}(a_1, b_1+a_2)$ by applying EQ **??**. For the remaining terms we have $Z_1 \prod_{j=2}^{i}(1-Z_j) \sim \text{beta}(a_1, b_1) \prod_{j=2}^{i} \text{beta}(b_j, a_j)$. Applying EQ **??** again, terms cancel and we obtain $Z_1 \prod_{j=2}^{i}(1-Z_j) \sim \text{beta}(a_1, b_1 + \sum_{j=2}^{i} a_j)$. Switching the parameters and we obtain $1 - Z_1 \prod_{j=2}^{i}(1-Z_j) \sim \text{beta}(b_1 + \sum_{j=2}^{i} a_j, a_1)$. Hence, the Z_is have beta marginals.