Simulating Controlled Variate and Rank Correlations Based on the Power Method Transformation

Todd C. Headrick, Southern Illinois University Carbondale
Simon Y. Aman
T. Mark Beasley

Available at: https://works.bepress.com/todd_headrick/34/
Simulation Methods

Simulating Controlled Variate and Rank Correlations Based on the Power Method Transformation

TODD C. HEADRICK¹, SIMON Y. AMAN², AND T. MARK BEASLEY³

¹Department of EPSE, Southern Illinois University – Carbondale, Carbondale, Illinois, USA
²Department of Mathematics, Truman College, Chicago, Illinois, USA
³Department of Biostatistics, University of Alabama – Birmingham, Birmingham, Illinois, USA

The power method transformation is a popular algorithm used for simulating correlated non normal continuous variates because of its simplicity and ease of execution. Statistical models may consist of continuous and (or) ranked variates. In view of this, the methodology is derived for simulating controlled correlation structures between non normal (a) variates, (b) ranks, and (c) variates with ranks in the context of the power method. The correlation structure between variate-values and their associated rank-order is also derived for the power method. As such, a measure of the potential loss of information is provided when ranks are used in place of variate-values. The results of a Monte Carlo simulation are provided to confirm and demonstrate the methodology.

Keywords Cumulants; Monte Carlo; Non normal; Rank-order statistics; Simulation.

Mathematics Subject Classification Primary 65C60; Secondary 65C05.

1. Introduction

The power method (Fleishman, 1978; Headrick, 2002; Headrick and Kowalchuk, 2007) is a polynomial transformation that generates continuous non normal variates often used in Monte Carlo or simulation studies. The primary advantage of this transformation is that it provides computationally efficient algorithms for
Simulating Controlled Variate and Rank Correlations

Simulating multivariate non normal distributions with arbitrary correlation matrices (Headrick, 2002; Headrick and Sawilowsky, 1999; Headrick et al., 2007; Vale and Maurelli, 1983).

The power method has been used in studies that have included such topics or techniques as: ANCOVA (Harwell and Serlin, 1988; Headrick and Sawilowsky, 2000; Headrick and Vineyard, 2001; Klockars and Moses, 2002; Olejnik and Algina, 1987), computer adaptive testing (Zhu et al., 2002), hierarchical linear models (Shieh, 2000), item response theory (Stone, 2003), logistic regression (Hess et al., 2001), microarray analysis (Powell et al., 2002), regression (Headrick and Rotou, 2001; Serlin and Harwell, 2004), repeated measures (Beasley and Zumbo, 2003; Kowalchuk et al., 2003; Lix et al., 2003), structural equation modeling (Hipp and Bollen, 2003; Reiniart et al., 2002), and other multivariate (non)parametric or rank-based tests (Beasley, 2002; Habib and Harwell, 1989; Rasch and Guiard, 2004; Steyn, 1993). For example, in the context of ANCOVA, the power method was used to simultaneously control the strength of the correlation between the variate and covariate and their degree of non normality.

Let \( Y_1, \ldots, Y_T \) be power method polynomials where \( Y_i \) and \( Y_j \) are expressed as

\[
Y_i = \sum_{k=1}^{m} c_{ik} Z_k^{k-1}
\]

\[
Y_j = \sum_{k=1}^{m} c_{jk} Z_k^{k-1}.
\]

The variates \( Y_i \) and \( Y_j \) are correlated non normal distributions with zero means, unit variances, and have derivatives such that \( Y'_i > 0 \) and \( Y'_j > 0 \) (i.e., the polynomials are strictly increasing monotonic functions). The variables \( Z_i \) and \( Z_j \sim N(0, 1) \) and have an intermediate correlation denoted as \( \rho_{Z_i, Z_j} \). The constants \( c_{ik} \) and \( c_{jk} \) determine the shapes of \( Y_i \) and \( Y_j \) and can be computed by solving the system of equations given in Headrick and Kowalchuk (2007, Appendix A for \( m = 6 \); or Appendix B for \( m = 4 \)) for a set of prespecified standardized cumulants \( \gamma_{k=1, \ldots, m} \). For a program that solves the constant coefficients for \( m = 6 \), see Headrick et al. (2007).

The probability density function (pdf) and distribution function associated with \( Y_i \) are known and their derivations are given in Headrick (2004) and in Headrick and Kowalchuk (2007, Proposition 3.1). To illustrate, depicted in Fig. 1 is an example of a fifth-order \((m = 6)\) power method pdf’s approximation of the chi-square distribution with \( df = 3 \). As indicated in Fig. 1, the power method’s pdf provides an excellent approximation of the upper 5% of this chi-square pdf. The upper 5% of the power method’s pdf was calculated using the integration techniques described in Headrick and Kowalchuk (2007, Table 2). It should be noted that the class of power method pdfs associated with \( m = 4 \) (Fleishman, 1978) does not include the standardized cumulants associated with the chi-square family of distributions (Headrick and Kowalchuk, 2007, Property 4.4).

Presented in Fig. 2 is a schematic of the bivariate correlation structure for the power method based on (1), (2), and their respective rank-orders \( R(Y_i) \) and \( R(Y_j) \). One problem associated with the power method is the inability to simulate controlled correlations between continuous non normal variates and ranks. Specifically, the inability to determine the intermediate correlation \( \rho_{Z_i, Z_j} \) in Fig. 2 such that \( Y_i \) and \( R(Y_j) \) would have a prespecified correlation \( \rho_{Y_i, R(Y_j)} \).
Figure 1. A fifth-order polynomial power method (PM) pdf (dashed lines) approximation to the chi-square distribution with $df = 3$ (with parameters of $\mu = 3$ and $\sigma = \sqrt{6}$) using (1). Equation (1) is expressed as $Y_i = \sum_{k=1}^{m=6} c_{ik}Z_{i}^{-1}$, where the $c_{ik}$ are: $c_{i1} = -0.259037$; $c_{i2} = 0.867102$; $c_{i3} = 0.265362$; $c_{i4} = 0.021276$; $c_{i5} = -0.002108$; and $c_{i6} = 0.000092$.

Figure 2. Schematic of the bivariate correlation structure for the power method based on Eqs. (1), (2), and their rank-orders.
Simulating Controlled Variate and Rank Correlations

Some common applications where this could be useful are simulating correlation structures between (a) percentile (or class) ranks and achievement test scores, (b) variate-values and ranked data in the context of multivariate statistics (e.g., Hotelling’s $T^2$ where one variable is continuous and the other is a rank), or (c) rank transformed covariates and a non normal variate in the context of the GLM.

More specifically, the power method currently allows one to control $\rho_{Y_i, Y_j}$ which, for example, enables a methodologist to simultaneously investigate departures from normality and sphericity. However, a methodologist may also be interested in controlling the correlation between non normal variates when investigating rank-based alternatives to parametric models for repeated measures or ANCOVA designs (Beasley, 2002; Brunner et al., 2002; Stevenson and Jacobson, 1988). Thus, the ability to control $\rho_{R(Y_i), R(Y_j)}$ or $\rho_{Y_i, R(Y_j)}$ would also give a researcher further insight in terms of how converting the non normal variates to ranks affects either the sphericity of a covariance matrix or the strength of the correlation between a variate $Y_i$ and the rank of a covariate.

Another problem associated with the power method is that there is no measure of the amount of potential loss of information when using ranks $R(Y_i)$ in place of the variate-values $Y_i$. One approach that could be used to address this problem would be analogous to Stuart’s (1954) use of the product-moment coefficient of correlation between the variate-values and their rank-order (i.e., $\rho_{Y_i, R(Y_i)}$ in Fig. 2) as the measure of this potential loss in efficiency. This measure is based on the coefficient of mean difference denoted as $\Delta$ (Gini, 1912). More specifically, if two independent continuous variables $X_1$ and $X_2$ are both distributed as $X$, then $\Delta$ can be expressed as (Johnson et al., 1994, p. 3; Kendall and Stuart, 1977, p. 47)

$$\Delta = E[|X_1 - X_2|] = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} |x_1 - x_2| f(x_1) f(x_2) dx_1 dx_2. \tag{3}$$

It follows from Stuart (1954) that the correlation between the variate $X$ and its associated rank-order $R(X)$ is

$$\rho_{X, R(X)} = \sqrt{(n-1)/(n+1)(\Delta \sqrt{3})/(2\sigma)}, \tag{4}$$

where $\sigma$ is the population standard deviation associated with $X$. An alternative derivation of (4) is also given in Gibbons and Chakraborti (1992, pp. 132–135). Some specific evaluations of (4) for various well-known pdfs (e.g., gamma, normal, uniform) are given in Stuart (1954) or in Kendall and Gibbons (1990, pp. 165–166).

2. Purpose of the Study

In view of the above, the present aim is to resolve the aforementioned problems associated with the power method by deriving the general formulae for the variate and rank correlation structure in Fig. 2. In so doing, a methodologist will be able to conduct simulations with controlled correlations between variates ($\rho_{Y_i, Y_j}$), ranks ($\rho_{R(Y_i), R(Y_j)}$), and variates with ranks ($\rho_{Y_i, R(Y_j)}$). Moreover, the derivations also include a measure ($\rho_{Y_i, R(Y_j)}$) of the potential loss of information when using ranks $R(Y_j)$ in place of the variate-values of $Y_i$. A numerical example and the results of a Monte Carlo simulation will be provided to demonstrate and confirm the methodology.
3. Mathematical Development

We assume the polynomials in (1) and (2) are of order five (i.e., $m = 6$) and the variates $Y_i$ and $Y_j$ are strictly increasing monotonic transformations in $Z_i$ and $Z_j$ to ensure that the method produces valid pdfs (Headrick and Kowalchuk, 2007, Proposition 3.1). As such, this implies rank correlations of $\rho_{R(Y_i), R(Z_i)} = \rho_{R(Y_j), R(Z_j)} = 1$ in Fig. 2.

Given these assumptions, let $Z_1, \ldots, Z_T$ be standard normal variables where $Z_i$ and $Z_j$ have univariate and bivariate pdfs expressed as:

$$f_i := f_{Z_i}(z_i) = (2\pi)^{-\frac{1}{2}} \exp\left\{-\frac{z_i^2}{2}\right\}$$

$$f_j := f_{Z_j}(z_j) = (2\pi)^{-\frac{1}{2}} \exp\left\{-\frac{z_j^2}{2}\right\}$$

$$f_{ij} := f_{Z_i Z_j}(z_i, z_j, \rho_{z_i z_j}) = (2\pi \sqrt{1 - \rho_{z_i z_j}^2})^{-\frac{1}{2}} \exp\left\{-\left(2(1 - \rho_{z_i z_j}^2)\right)^{-1}\left(z_i^2 - 2\rho_{z_i z_j} z_i z_j + z_j^2\right)\right\}.$$  

(7)

Let the distribution functions associated with (5) and (6) be denoted as

$$\Phi(z_i) = \int_{-\infty}^{z_i} (2\pi)^{-\frac{1}{2}} \exp\left\{-u_i^2/2\right\} du_i$$

$$\Phi(z_j) = \int_{-\infty}^{z_j} (2\pi)^{-\frac{1}{2}} \exp\left\{-u_j^2/2\right\} du_j$$

(8)

(9)

where $\Phi(z_i) \sim U_i[0,1], \Phi(z_j) \sim U_j[0,1]$, and correlation $\rho_{\Phi(z_i), \Phi(z_j)} = (6/\pi) \sin^{-1}(\rho_{z_i z_j}/2)$ (Pearson, 1907).

If $Z_i$ and $Z_j$ are independent, then from (3), (5), and (6) the coefficient of mean difference for the standard normal distribution is:

$$\Delta = 2/\sqrt{\pi} = E[|Z_i - Z_j|] = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} |z_i - z_j| f_i f_j dz_i dz_j.$$  

(10)

Substituting $\Delta$ from (10) into (4) yields for the unit normal distribution

$$\rho_{Z_i, R(Z_i)} = \sqrt{(n-1)/(n+1)}\sqrt{3}/\pi.$$  

(11)

Using (1) and (2) with $c_{jk} = c_{ik}$, and (10), the coefficient of mean difference for the power method is expressed as:

$$\Delta = E\left[ \sum_{k=1}^{m=6} c_{ik} z_i^{k-1} - \sum_{k=1}^{m=6} c_{ik} z_j^{k-1} \right]$$

$$= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \left| \sum_{k=1}^{m=6} c_{ik} z_i^{k-1} \right| \left| \sum_{k=1}^{m=6} c_{ik} z_j^{k-1} \right| f_i f_j dz_i dz_j.$$  

(12)
Changing from rectangular to polar coordinates in (12), to eliminate the absolute value, and integrating yields

\[
\Delta = \int_{0}^{\pi/2} \int_{0}^{\infty} \left( \sum_{k=1}^{m} c_{ik}(r \cos \theta)^{k-1} - \sum_{k=1}^{m} c_{ik}(r \sin \theta)^{k-1} \right) \frac{r}{2\pi} \exp\{-r^2/2\} dr d\theta
\]

\[
- \int_{\pi/2}^{\pi} \int_{0}^{\infty} \left( \sum_{k=1}^{m} c_{ik}(r \cos \theta)^{k-1} - \sum_{k=1}^{m} c_{ik}(r \sin \theta)^{k-1} \right) \frac{r}{2\pi} \exp\{-r^2/2\} dr d\theta
\]

\[
+ \int_{\pi/2}^{\pi} \int_{0}^{\infty} \left( \sum_{k=1}^{m} c_{ik}(r \cos \theta)^{k-1} - \sum_{k=1}^{m} c_{ik}(r \sin \theta)^{k-1} \right) \frac{r}{2\pi} \exp\{-r^2/2\} dr d\theta
\]

\[
\Delta = (2\sqrt{\pi})^{-1}(4c_{i2} + 10c_{i4} + 43c_{i6}). \tag{13}
\]

Substituting \(\Delta\) from (13) into (4) and setting \(\sigma = 1\), because all power method distributions have unit variances, gives the correlations \(\rho_{Y_i,R(Y_j)}\) and \(\rho_{Y_i,R(Z_i)}\) in Fig. 2 as

\[
\rho_{Y_i,R(Y_j)} = \rho_{Y_i,R(Z_i)} = \sqrt{(n-1)/(n+1)} \sqrt{3/\pi}((1/4)(4c_{i2} + 10c_{i4} + 43c_{i6})). \tag{14}
\]

Thus, (14) is the power method’s measure of the potential loss of information when using ranks in place of variate-values. If we were to consider, for example, substituting the even subscripted values of \(c_{ik}\) from the caption below Fig. 1 into (14), then we would obtain as \(n \to \infty\), \(\rho_{Y_i,R(Y_j)} \approx 0.900283\) which is close to the exact correlation for the chi-square distribution with \(df = 3\) based on (4) of \(\rho_{X,R(X)} = 2\sqrt{2/\pi} \approx 0.900316\) where \(\Delta = 8/\pi\) and \(\sigma = \sqrt{6}\). Note also that for the special case where \(Y_i = Z_i\) in (1), i.e., \(c_{i2} = 1\) and \(c_{i(k\neq2)} = 0\), then (14) reduces to (11).

**Remark 3.1.** The strictly increasing monotonicity assumption in (1), i.e., \(Y_i^* > 0\), implies that \(\rho_{Y_i,Z_i} \in (0, 1]\). Thus, for (13) and (14) to both be positive, the restriction on the constants \(c_{i2}\), \(c_{i4}\), and \(c_{i6}\) is that they must satisfy the inequality \(0 < c_{i2} + 3c_{i4} + 15c_{i6} \leq 1\).

**Proof.** The correlation \(\rho_{Y_i,Z_i}\) (see Fig. 2) can be determined by equating \(\rho_{Y_i,Z_i} = E[Y_iZ_i]\) because both \(Y_i\) and \(Z_i\) have zero means and unit variances. Hence,

\[
\rho_{Y_i,Z_i} = E[Y_iZ_i] = E\left[ \sum_{k=1}^{m} c_{ik}Z_i^k \right] = \sum_{k=1}^{m} c_{ik}E[Z_i^k] = c_{i2} + 3c_{i4} + 15c_{i6} \tag{15}
\]

because all odd central moments of the standard normal distribution are zero and the even central moments are \(\mu_2 = 1\), \(\mu_4 = 3\), and \(\mu_6 = 15\).

The correlations \(\rho_{Y_i,Y_j}\) and \(\rho_{Y_i,R(Y_j)}\) in Fig. 2 can be obtained by making use of (1), (2), (7), and the standardized form of (9). More specifically, \(\rho_{Y_i,Y_j}\) and \(\rho_{Y_i,R(Y_j)}\) can be numerically calculated from

\[
\rho_{Y_i,Y_j} = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \left( \sum_{k=1}^{m} c_{ik}z_i^{k-1} \right)\left( \sum_{k=1}^{m} c_{jk}z_j^{k-1} \right) f_{ij} dz_i dz_j \tag{16}
\]
To evaluate the proposed procedure, four (non)normal distributions (\(Y_i\), \(Z_i\)) are simulated with the specified standardized cumulants (\(g_i\)) and correlations (\(\rho_{Y_i,Z_i}\), \(\rho_{Y_i,Y_j}\)), listed in Table 2. The graphs of the distributions’ pdfs depicted in Table 2 are described as: \(Y_1\) is standard normal; \(Y_2\) is asymmetric with moderate tail-weight; \(Y_3\) is symmetric with heavy tails; and \(Y_4\) is symmetric with light tails. The constants \(c_{ik}\) listed in Table 2 were calculated by simultaneously solving the system of equations given in Headrick and Kowalchuk (2007, Appendix A) using the Mathematica 5.2 (Wolfram, 2003) command FindRoot for the specified values of \(g_i\). The correlations \(\rho_{Y_i,Z_i}\) and \(\rho_{Y_i,Y_j}\) were determined from Eqs. (b) and (e) in Table 1.

### 4. Monte Carlo Simulation

To evaluate the proposed procedure, four (non)normal distributions (\(Y_i\) = 1, 2, 3, 4) were simulated with the specified standardized cumulants (\(g_i\)) and correlations (\(\rho_{Y_i,Z_i}\), \(\rho_{Y_i,Y_j}\)), listed in Table 2. The graphs of the distributions’ pdfs depicted in Table 2 are described as: \(Y_1\) is standard normal; \(Y_2\) is asymmetric with moderate tail-weight; \(Y_3\) is symmetric with heavy tails; and \(Y_4\) is symmetric with light tails. The constants \(c_{ik}\) listed in Table 2 were calculated by simultaneously solving the system of equations given in Headrick and Kowalchuk (2007, Appendix A) using the Mathematica 5.2 (Wolfram, 2003) command FindRoot for the specified values of \(g_i\). The correlations \(\rho_{Y_i,Z_i}\) and \(\rho_{Y_i,Y_j}\) were determined from Eqs. (b) and (e) in Table 1.
These parameters represent examples of correlations between variates, associated with the specified correlations of intermediate correlations for simulating the specified correlations in Table 3.

The specified correlations of concern for this simulation are listed in Table 3. These parameters represent examples of correlations between variates \((\rho_{Y_i,Y_j})\), variates with ranks \((\rho_{Y_i,R(Y_j)})\), and ranks \((\rho_{R(Y_i),R(Y_j)})\). Tables 4–6 give the required intermediate correlations for simulating the specified correlations in Table 3 for sample sizes of \(n = 10\), \(n = 30\), and \(n = 10^6\). The intermediate correlations associated with the specified correlations of \(\rho_{Y_i,Y_j}\), \(\rho_{Y_i,R(Y_j)}\), and \(\rho_{R(Y_i),R(Y_j)}\) were determined from Eqs. (d), (f), and (g) in Table 1. Presented in Table 7 is an example

### Table 2
Fifth-order power method distributions and specifications for the simulation

<table>
<thead>
<tr>
<th>Cumulants</th>
<th>Constants</th>
<th>Correlations</th>
<th>Probability density function for (Y_{i=1,2,3,4})</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\gamma_1 = 0)</td>
<td>(c_{11} = 0)</td>
<td>(\rho_{Y_1,Z_1} = 1.0)</td>
<td><a href="https://example.com/graph1">Graph</a></td>
</tr>
<tr>
<td>(\gamma_2 = 1)</td>
<td>(c_{12} = 1)</td>
<td>(\rho_{Y_1,R(Y_1),n=10} = .883915)</td>
<td></td>
</tr>
<tr>
<td>(\gamma_3 = 0)</td>
<td>(c_{13} = 0)</td>
<td>(\rho_{Y_1,R(Y_1),n=30} = .945156)</td>
<td></td>
</tr>
<tr>
<td>(\gamma_4 = 0)</td>
<td>(c_{14} = 0)</td>
<td>(\rho_{Y_1,R(Y_1),n=10^6} = .977204)</td>
<td></td>
</tr>
<tr>
<td>(\gamma_5 = 0)</td>
<td>(c_{15} = 0)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(\gamma_6 = 0)</td>
<td>(c_{16} = 0)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(\gamma_1 = 0)</td>
<td>(c_{21} = -0.307740)</td>
<td>(\rho_{Y_2,Z_2} = .903441)</td>
<td><a href="https://example.com/graph2">Graph</a></td>
</tr>
<tr>
<td>(\gamma_2 = 1)</td>
<td>(c_{22} = 0.800560)</td>
<td>(\rho_{Y_2,R(Y_2),n=10} = .783164)</td>
<td></td>
</tr>
<tr>
<td>(\gamma_3 = 2)</td>
<td>(c_{23} = 0.318764)</td>
<td>(\rho_{Y_2,R(Y_2),n=30} = .837425)</td>
<td></td>
</tr>
<tr>
<td>(\gamma_4 = 6)</td>
<td>(c_{24} = 0.033500)</td>
<td>(\rho_{Y_2,R(Y_2),n=10^6} = .865819)</td>
<td></td>
</tr>
<tr>
<td>(\gamma_5 = 24)</td>
<td>(c_{25} = -0.003675)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(\gamma_6 = 120)</td>
<td>(c_{26} = 0.000159)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(\gamma_1 = 0)</td>
<td>(c_{31} = 0)</td>
<td>(\rho_{Y_3,Z_3} = .890560)</td>
<td><a href="https://example.com/graph3">Graph</a></td>
</tr>
<tr>
<td>(\gamma_2 = 1)</td>
<td>(c_{32} = 0.374011)</td>
<td>(\rho_{Y_1,R(Y_1),n=10} = .707015)</td>
<td></td>
</tr>
<tr>
<td>(\gamma_3 = 0)</td>
<td>(c_{33} = 0)</td>
<td>(\rho_{Y_1,R(Y_1),n=30} = .756001)</td>
<td></td>
</tr>
<tr>
<td>(\gamma_4 = 25)</td>
<td>(c_{34} = 0.159040)</td>
<td>(\rho_{Y_1,R(Y_1),n=10^6} = .781634)</td>
<td></td>
</tr>
<tr>
<td>(\gamma_5 = 0)</td>
<td>(c_{35} = 0)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(\gamma_6 = 5000)</td>
<td>(c_{36} = 0.002629)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(\gamma_1 = 0)</td>
<td>(c_{41} = 0)</td>
<td>(\rho_{Y_4,Z_4} = .986561)</td>
<td><a href="https://example.com/graph4">Graph</a></td>
</tr>
<tr>
<td>(\gamma_2 = 1)</td>
<td>(c_{42} = 1.248343)</td>
<td>(\rho_{Y_1,R(Y_1),n=10} = .903126)</td>
<td></td>
</tr>
<tr>
<td>(\gamma_3 = 0)</td>
<td>(c_{43} = 0)</td>
<td>(\rho_{Y_1,R(Y_1),n=30} = .965699)</td>
<td></td>
</tr>
<tr>
<td>(\gamma_4 = -1)</td>
<td>(c_{44} = -0.111426)</td>
<td>(\rho_{Y_1,R(Y_1),n=10^6} = .998442)</td>
<td></td>
</tr>
<tr>
<td>(\gamma_5 = 0)</td>
<td>(c_{45} = 0)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(\gamma_6 = 48/7)</td>
<td>(c_{46} = 0.004833)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Table 3
Specified correlations for the distributions in Table 2

| \(\rho_{Y_1,Y_2}\) | \(0.70\) |
| \(\rho_{Y_1,Y_3}\) | \(0.60\) |
| \(\rho_{Y_1,Y_4}\) | \(0.40\) |
| \(\rho_{Y_2,R(Y_1)}\) | \(0.60\) |
| \(\rho_{Y_2,R(Y_2)}\) | \(0.50\) |
| \(\rho_{R(Y_1),R(Y_3)}\) | \(0.80\) |
of solving for the intermediate correlation associated with $\rho_{Y, R(Y)}$ in Table 3 and Eq. (f) in Table 1 using the Mathematica 5.2 (Wolfram, 2003) command NIntegrate.

An algorithm coded in Fortran 77 was used to simulate the specified distributions and correlations listed in Tables 2 and 3. The algorithm employed the use of subroutines UNI1 and NORMB1 from RANGEN (Blair, 1987) to generate pseudo-random uniform and standard normal deviates. The data generation procedure began by conducting Cholesky factorizations on each of the intermediate correlation matrices in Tables 4–6. The entries from the factored matrices were then used to produce standard normal deviates with intercorrelations equal to the intermediate correlations. These deviates were subsequently transformed by using the constants $c_{ik}$ in Table 2 and the polynomials of the form in (1) and (2) to produce the specified (non)normal distributions with the specified correlations in Table 3 for each of the three different sample sizes.

The algorithm computed 100,000 values of the first four standardized cumulants and correlations listed in Tables 2 and 3. Overall average estimates of the cumulants ($\gamma_1 =$ mean, $\gamma_2 =$ variance, $\gamma_3 =$ skew, and $\gamma_4 =$ kurtosis) and correlations ($\rho_{Y, R(Y)}$, $\rho_{Y, Y'}$, $\rho_{X, R(Y)}$, and $\rho_{R(X), R(Y)}$) were subsequently obtained and were based on $(n = 10) \times 100,000$ and $(n = 30) \times 100,000$ pseudo-random deviates. The algorithm also computed estimates of the standardized cumulants and correlations based on single draws of size $n = 10^6$.

The empirical estimates of $\hat{\gamma}_1$ and $\hat{\rho}_{Y, R(Y)}$ are reported in Tables 8–10. The estimates of $\hat{\rho}_{Y, Y'}$, $\hat{\rho}_{Y, R(Y)}$, and $\hat{\rho}_{R(Y), R(Y)}$ are listed in Tables 11–13. Inspection of the results presented in Tables 8–13 indicates that all estimates were in close proximity with their specified parameter even for sample sizes as small as $n = 10$.

### Table 4
Intermediate correlation matrix for the correlations in Table 3 with $n = 10$

<table>
<thead>
<tr>
<th></th>
<th>$Z_1$</th>
<th>$Z_2$</th>
<th>$Z_3$</th>
<th>$Z_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Z_1$</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$Z_2$</td>
<td>0.774815</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$Z_3$</td>
<td>0.664127</td>
<td>0.759832</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>$Z_4$</td>
<td>0.442751</td>
<td>0.630988</td>
<td>0.862571</td>
<td>1</td>
</tr>
</tbody>
</table>

### Table 5
Intermediate correlation matrix for the correlations in Table 3 with $n = 30$

<table>
<thead>
<tr>
<th></th>
<th>$Z_1$</th>
<th>$Z_2$</th>
<th>$Z_3$</th>
<th>$Z_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Z_1$</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$Z_2$</td>
<td>0.774815</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$Z_3$</td>
<td>0.664127</td>
<td>0.709578</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>$Z_4$</td>
<td>0.442751</td>
<td>0.589522</td>
<td>0.831382</td>
<td>1</td>
</tr>
</tbody>
</table>
Table 6
Intermediate correlation matrix for the correlations in Table 3 with \(n = 10^6\)

<table>
<thead>
<tr>
<th></th>
<th>(Z_1)</th>
<th>(Z_2)</th>
<th>(Z_3)</th>
<th>(Z_4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Z_1)</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Z_2)</td>
<td>0.774815</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Z_3)</td>
<td>0.664127</td>
<td>0.685865</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>(Z_4)</td>
<td>0.442751</td>
<td>0.569937</td>
<td>0.813474</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 7
Mathematica code for estimating the intermediate correlation \(\hat{\rho}_{z_2z_4}\) in Table 4

(*Specified Correlation in Table 3: \(\rho_{Y_2,Y_4} = 0.50\)

Estimated Intermediate Correlation listed in Table 4 (see below): \(\hat{\rho}_{z_2z_4} = 0.630988^\ast\)

\(c_21 = -0.3077396;\)
\(c_22 = 0.8005604;\)
\(c_23 = 0.3187640;\)
\(c_24 = 0.0335001;\)
\(c_25 = -0.0036748;\)
\(c_26 = 0.0001587;\)
\(n = 10;\)
\(\hat{\rho}_{z_2z_4} = 0.630988;\)

\(\Phi_4 = \int_{-\infty}^{\infty} \left(\sqrt{2\pi}\right)^{-1} e^{-\frac{z_4^2}{2}} dz_4;\)
\(Y_2 = \sum_{k=1}^{6} c_{2k} z_4^{k-1};\)
\(f_{24} = e^{\frac{1}{2(\hat{\rho}_{z_2z_4})^2} (\frac{1}{2(\hat{\rho}_{z_2z_4})^2})};\)
\(\text{int} = \sqrt{\frac{n-1}{n}} \text{NIntegrate} \left((Y_2(\sqrt{3}(2\Phi_4 - 1)) f_{24}, \{z_2, -6, 6\}, \{z_4, -6, 6\}, \text{Method -> Trapezoidal}\right)\)
Solution: int = 0.49999987969495174

Table 8
Empirical estimates of the first four standardized cumulants \((\hat{\gamma}_i)\) and the variate-value and rank correlation \(\hat{\rho}_{Y_i,R(Y_i)}\) listed in Table 2. Sample size is \(n = 10\)

<table>
<thead>
<tr>
<th>Distribution</th>
<th>Mean ((\hat{\gamma}_1))</th>
<th>Variance ((\hat{\gamma}_2))</th>
<th>Skew ((\hat{\gamma}_3))</th>
<th>Kurtosis ((\hat{\gamma}_4))</th>
<th>(\hat{\rho}_{Y_i,R(Y_i)})</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Y_1)</td>
<td>0.000</td>
<td>1.005</td>
<td>0.000</td>
<td>0.000</td>
<td>.884</td>
</tr>
<tr>
<td>(Y_2)</td>
<td>0.001</td>
<td>1.002</td>
<td>2.007</td>
<td>6.021</td>
<td>.783</td>
</tr>
<tr>
<td>(Y_3)</td>
<td>0.000</td>
<td>1.001</td>
<td>0.000</td>
<td>25.181</td>
<td>.707</td>
</tr>
<tr>
<td>(Y_4)</td>
<td>0.000</td>
<td>1.000</td>
<td>0.001</td>
<td>-1.003</td>
<td>.903</td>
</tr>
</tbody>
</table>
### Table 9
Empirical estimates of the first four standardized cumulants ($\hat{\gamma}_i$) and the variate-value and rank correlation $\hat{\rho}_{Y_i,R(Y_i)}$ listed in Table 2. Sample size is $n = 30$

<table>
<thead>
<tr>
<th>Distribution</th>
<th>Mean ($\hat{\gamma}_1$)</th>
<th>Variance ($\hat{\gamma}_2$)</th>
<th>Skew ($\hat{\gamma}_3$)</th>
<th>Kurtosis ($\hat{\gamma}_4$)</th>
<th>$\hat{\rho}_{Y_i,R(Y_i)}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Y_1$</td>
<td>0.000</td>
<td>1.000</td>
<td>0.000</td>
<td>0.000</td>
<td>.945</td>
</tr>
<tr>
<td>$Y_2$</td>
<td>0.000</td>
<td>1.000</td>
<td>2.000</td>
<td>5.99</td>
<td>.837</td>
</tr>
<tr>
<td>$Y_3$</td>
<td>0.000</td>
<td>0.999</td>
<td>0.008</td>
<td>25.067</td>
<td>.756</td>
</tr>
<tr>
<td>$Y_4$</td>
<td>0.000</td>
<td>1.000</td>
<td>0.000</td>
<td>-1.003</td>
<td>.965</td>
</tr>
</tbody>
</table>

### Table 10
Empirical estimates of the first four standardized cumulants ($\hat{\gamma}_i$) and the variate-value and rank correlation $\hat{\rho}_{Y_i,R(Y_i)}$ listed in Table 2. Sample size is $n = 10^6$

<table>
<thead>
<tr>
<th>Distribution</th>
<th>Mean ($\hat{\gamma}_1$)</th>
<th>Variance ($\hat{\gamma}_2$)</th>
<th>Skew ($\hat{\gamma}_3$)</th>
<th>Kurtosis ($\hat{\gamma}_4$)</th>
<th>$\hat{\rho}_{Y_i,R(Y_i)}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Y_1$</td>
<td>0.000</td>
<td>0.997</td>
<td>0.000</td>
<td>-0.014</td>
<td>.977</td>
</tr>
<tr>
<td>$Y_2$</td>
<td>0.001</td>
<td>1.000</td>
<td>2.012</td>
<td>5.98</td>
<td>.866</td>
</tr>
<tr>
<td>$Y_3$</td>
<td>0.000</td>
<td>0.993</td>
<td>0.080</td>
<td>24.976</td>
<td>.782</td>
</tr>
<tr>
<td>$Y_4$</td>
<td>0.000</td>
<td>0.997</td>
<td>0.000</td>
<td>-1.015</td>
<td>.998</td>
</tr>
</tbody>
</table>

### Table 11
Empirical estimates of the correlations in Table 3 for $n = 10$

\[\hat{\rho}_{Y_1,Y_2} = 0.700\]
\[\hat{\rho}_{Y_1,Y_3} = 0.602\]
\[\hat{\rho}_{Y_1,Y_4} = 0.401\]
\[\hat{\rho}_{Y_2,Y_3} = 0.601\]
\[\hat{\rho}_{Y_2,Y_4} = 0.501\]
\[\rho_{R(Y_3),R(Y_4)} = 0.800\]

### Table 12
Empirical estimates of the correlations in Table 3 for $n = 30$

\[\hat{\rho}_{Y_1,Y_2} = 0.700\]
\[\hat{\rho}_{Y_1,Y_3} = 0.600\]
\[\hat{\rho}_{Y_1,Y_4} = 0.400\]
\[\hat{\rho}_{Y_2,Y_3} = 0.600\]
\[\hat{\rho}_{Y_2,Y_4} = 0.501\]
\[\hat{\rho}_{R(Y_3),R(Y_4)} = 0.800\]

### Table 13
Estimates of the correlations in Table 3 using a single draw of $n = 10^6$

\[\hat{\rho}_{Y_1,Y_2} = 0.700\]
\[\hat{\rho}_{Y_1,Y_3} = 0.602\]
\[\hat{\rho}_{Y_1,Y_4} = 0.401\]
\[\hat{\rho}_{Y_2,Y_3} = 0.600\]
\[\hat{\rho}_{Y_2,Y_4} = 0.500\]
\[\hat{\rho}_{R(Y_3),R(Y_4)} = 0.800\]
5. Discussion

Many methodologists have used the power method to simulate non normal data for the purpose of examining Type I error and power properties among competing (non)parametric and other rank-based statistical procedures. As such, of particular importance is the correlation $\rho_{Y_1, R(Y_1)}$ in Table 1 which can be used as a measure for the potential loss of information when using the ranks in place of the variate-values. For example, distribution $Y_3$ in Table 2 has a correlation between its variate-values and assigned ranks that is rather low ($\rho_{Y_3, R(Y_3)} \approx 0.756$) relative to other pdfs (e.g., $Y_1$ and $Y_4$). Thus, conclusions from inferential procedures based on the ranks of $Y_3$ could differ markedly from analyses performed on the actual variate-values. In contrast, distribution $Y_4$ has a variate-rank correlation of $\rho_{Y_4, R(Y_4)} \approx 0.966$. In this case, with such a high correlation, conclusions from statistical analyses based on the ranks would usually not differ much from the same analyses performed on the variate-values (Gibbons and Chakraborti, 1992, p. 132; Kendall and Gibbons, 1990, p. 166).

To demonstrate, a small Monte Carlo study was conducted to compare the relative Type I error and power properties of the variate-values and their rank-orders for a two-group independent samples design with $n_A = n_B = 30$. We used the two symmetric distributions $Y_3$ and $Y_4$ (see Table 2) which have the disparate variate-rank correlations indicated above.

The data for groups $A$ and $B$ were randomly sampled from either $Y_3$ or $Y_4$ where both groups were sampled from the same (symmetric) distribution. In each condition, the usual OLS parametric $t$-test was applied to both the variate-values and their assigned ranks. It is noted that an OLS procedure applied to ranked data in this context is referred to as a rank-transform (RT) procedure suggested by Conover and Iman (1981) and is functionally equivalent to the Mann–Whitney–Wilcoxon test. Thus, with symmetric error distributions, the null hypothesis for the rank-based analysis $H_0: P(Y_A < Y_B) = P(Y_A > Y_B)$ is equivalent to the null hypothesis for the parametric $t$-test $H_0: \mu_A = \mu_B$; that is, $E[Y_A] = E[Y_B]$ is equivalent to $E[R(Y_A)] = E[R(Y_B)]$ (see Vargha and Delaney, 1998a,b).

Empirical Type I error and power rates were based on 50,000 replications within each condition in the Monte Carlo study. In terms of Type I error (nominal $\alpha = 0.05$), when both groups were sampled from $Y_3$, the empirical Type I error rate for the for the parametric $t$-test performed on the variate-values was conservative $\hat{\alpha} = 0.040$, whereas, the rejection rate for the $t$-test performed on the ranks was close to nominal alpha $\hat{\alpha} = 0.051$. In contrast, when both groups were sampled from $Y_4$, the empirical Type I errors rates were essentially the same (to three decimals places) for both tests $\hat{\alpha} = \hat{\alpha} = 0.051$.

In terms of power analysis, the rejection rates for the tests differed markedly when sampling was from $Y_3$. For example, with a standardized effect size of one-half a standard deviation between the variate means, the rejection rates were 0.55 for the parametric $t$-test performed on the variate-values and 0.89 for the $t$-test performed on the ranks. In contradistinction, when sampling was from $Y_4$, and with the same standardized effect size of 0.50, the rejection rates were similar for the parametric $t$-test and the RT, 0.46 and 0.44, respectively.

Some other applications of the proposed methodology that are worth mentioning are in terms of the use of ranks in principal components analysis (PCA) or exploratory factor analysis (EFA) (Bartholomew, 1983; Besse and Ramsay, 1986). More specifically, this method not only controls the correlation structure
between non normal variates but also allows the researcher insight into how ranking will affect the subsequent correlation matrix submitted to a PCA or EFA. Thus, knowing the correlation between non normal variates and their ranks is useful for investigating estimation bias when using rank-based statistics.

References


