On Polynomial Transformations For Simulating Multivariate Non-normal Distributions

Todd C. Headrick
Southern Illinois University - Carbondale

Procedures are introduced and discussed for increasing the computational and statistical efficiency of polynomial transformations used in Monte Carlo or simulation studies. Comparisons are also made between polynomials of order three and five in terms of (a) computational and statistical efficiency, (b) the skew and kurtosis boundary, and (c) boundaries for Pearson correlations. It is also shown how ranked data can be simulated for specified Spearman correlations and sample sizes. Potential consequences of nonmonotonic transformations on rank correlations are also discussed.

Key words: Correlated data, cumulants, Monte Carlo methods, polynomial transformations, nonnormality

Introduction

A common practice used to investigate the relative Type I error and power properties of competing statistical procedures under nonnormality is the method of Monte Carlo. For example, consider the following polynomial transformation in general form

\[ Y_i = c_0 + \sum_{i=1}^{m} c_i Z_i^i \]  

(1)

where \( Z_i \sim \text{NID}(0,1), \) and \( i \in \mathbb{N} = \{1,2,\ldots,m\} \). Setting \( m = 3 \), Fleishman (1978) derived a system of four equations that would solve for the four coefficients \( c_0,\ldots,c_3 \) in (1) for a specified non-normal distribution. Specifically, these coefficients are determined by simultaneously solving this system of equations for the first four standardized cumulants of a distribution. The coefficients are subsequently entered into (1) to generate \( Y_i \) with the specified cumulants.

Equation (1) was extended to \( m = 5 \) by Headrick (2002) for controlling the first six standardized cumulants from a specified probability density function.

The third-order polynomial (Fleishman, 1978) and the fifth-order polynomial (Headrick, 2002) transformations were also extended for the purpose of generating multivariate non-normal distributions (Headrick, 2002, Equation, 26; Headrick & Sawilowsky, 1999, Equation 9; Vale & Maurelli, 1983, Equation 11). These extensions have been demonstrated to be quite useful when there is a need for correlated non-normal data sets in a Monte Carlo study.

Some examples include analysis of covariance (Harwell & Serlin, 1988; Headrick & Sawilowsky, 1999; Headrick & Vineyard, 2001; Klockers & Moses, 2002), hierarchical linear models (Shieh, 2000), regression (Harwell & Serlin, 1989; Headrick & Rotou, 2001; Whittaker, Fauladi, & Williams, 2002) repeated measures (Beasley & Zumbo, 2003; Harwell & Serlin, 1997), and multivariate nonparametric tests (Beasley, 2002; Habib & Harwell, 1989). The multivariate extension of the fifth-order polynomial has also demonstrated to be useful for simulating continuous with ranked or ordinal data structures (Headrick & Beasley, 2003) and for generating systems of correlated non-normal linear statistical equations (Headrick & Beasley, 2004).
Although the primary advantages of the third and fifth-order polynomials are their ease of execution and computationally efficiency, there are limitations to these transformations. More specifically, the primary limitations are (a) the transformations are limited in terms of the possible combinations of skew and kurtosis, (b) the polynomials are not, in general, monotonic transformations and therefore have the potential to produce biased rank correlation coefficients, and (c) distributions with bivariate non-normal structures may have lower and upper boundary points \((-a, a)\) for Pearson correlations \((r)\) such that \(r \in [-1 < -a, a < +1]\) and where it is possible, for example, that \(|a| < 0.70\). It should be noted that the distribution of \(Y_i\), in general, is not exact. Headrick (2004) has derived the probability density function and distribution function for \(Y_i\) when the transformation between \(Y_i\) and \(Z_i\) is monotonic.

In view of the above, the purposes of the study are to introduce and discuss methods that minimize the limitations of the polynomial transformations and to develop a procedure for simulating rank correlations. More specifically, the intent is to (a) derive and discuss methods for improving computational and statistical efficiency for a Monte Carlo study, (b) compare and contrast the third and fifth order polynomials in terms of the skew and kurtosis boundary and in terms of boundaries for Pearson correlations, (c) provide a method for simulating Spearman rank correlations with specified samples sizes, and (d) discuss the potential effects of nonmonotonic transformations on rank correlations.

Improving Computational and Statistical Efficiency

Consider (1) with \(m = 5\) as

\[
Y_i = c_0 + c_1 Z_1 + c_2 Z_1^2 + c_3 Z_1^3 + c_4 Z_1^4 + c_5 Z_1^5
\]

or

\[
Y_i = c_0 + Z_1 \left( c_1 + Z_1 \left( c_2 + Z_1 \left( c_3 + Z_1 \left( c_4 + c_5 Z_1 \right) \right) \right) \right) .
\]  

If the algorithm used to generate \(Y_i\) is coded in the manner as in (3) instead of (2) then the run time of a Monte Carlo or simulation study can be substantially reduced. To illustrate (briefly), on a Pentium-based PC it took approximately 25 seconds of computer time to draw 100,000 random samples of size \(n = 550\) from an approximate exponential distribution using (3). On the other hand, using (2), the sample size had to be reduced to \(n = 100\) to obtain the same 100,000 draws within the same 25 second time period. Thus, a considerable gain in computational efficiency can be realized by using (3) in lieu of (2).

Suppose two standardized random variables \(Y_1\) and \(Y_2\) based on (3) are generated. A method that is useful to improve the efficiency of the estimate of \(E[Y_1 Y_2]\) is by inducing a negative correlation on \(Y_1\) and \(Y_2\). To demonstrate, if \(Y_1\) and \(Y_2\) were identically distributed, then

\[
\text{Var} \left[ \frac{Y_1 + Y_2}{2} \right] = \frac{1}{2} + \text{Corr}[Y_1, Y_2] .
\]

By inspection of (4) it would be advantageous if \(Y_1\) and \(Y_2\) were negatively correlated.

Assume that a monotone relationship between \(Z_i\) and \(Y_i\) for \(i = 1, 2\) exists. To induce a negative correlation on \(Y_1\) and \(Y_2\) it is only necessary to simultaneously reverse the signs of the coefficients with odd subscripts in \(Y_2\) as

\[
Y_1 = f_1(c_0, c_1, c_2, c_3, c_4, c_5, Z_i)
\]

\[
Y_2 = f_2(c_0, -c_1, -c_2, c_3, c_4, -c_5, Z_i) .
\]

Because the structure between \(Y_i\) and \(Z_i\) is standard bivariate normal, the correlation between \(Y_1\) and \(Y_2\) can be defined as

\[
\rho_{Y_1 Y_2} = E[Y_1 Y_2] .
\]

Expanding (7) and taking expectations using the moments from the standard normal density yields
\[ \rho_{Y_1 Y_2} = c_0^2 - c_1^2 + 2c_0(c_2 + 3c_4) - 6c_1(c_3 + 5c_5) + 3(c_2^2 + 10c_2c_4 - 5(c_3^2 - 7c_5^2 + 14c_5c_6 + 63c_7^2)). \]

(8)

Thus, the correlation between \( Y_1 \) and \( Y_2 \) can be determined by evaluating (8) using specified values for \( c_0, \ldots, c_7 \). For example, evaluating (8) using the coefficients that approximate the exponential density (see Headrick, 2002, Table 1) gives \( \rho_{Y_1 Y_2} \approx -0.647 \).

The method of inducing a negative correlation between \( Y_1 \) and \( Y_2 \) is analogous to the method used on distributions generated by the inverse transform method. More specifically, consider generating \( X_1 \) and \( X_2 \) from the single parameter exponential family with distribution function \( G \) and with an inverse distribution function denoted as \( G^{-1} \). Let \( X_1 = G^{-1}(V) \) and \( X_2 = G^{-1}(1-V) \) where \( V \sim U(0,1) \). Define the parameters for the first and second moments as \( \theta \) and \( \theta^2 \). From the definition of the product moment of correlation exists

\[ E[X_1 X_2] = \theta^2 \int_0^1 \ln v \ln(1-v)dv = \theta^2 (2 - \pi^2/6). \]

As such, the correlation between \( X_1 \) and \( X_2 \) is

\[ \rho_{X_1 X_2} = 1 - \pi^2/6 \approx -0.645. \]

(9)

Thus, the approximation given by (8) for the exponential distribution is very close to the exact result given in (9).

Presented in Table 1 below are confidence intervals from a Monte Carlo simulation study that demonstrate the advantage of inducing a negative correlation on \( Y_1 \) and \( Y_2 \). By inspection of Table 1 when \( Y_1 \) and \( Y_2 \) are uncorrelated it takes over 2.5 times the sample size to obtain a confidence interval that has approximately the same width as the data with an induced negative correlation. Thus, whenever possible it is advantageous to induce a negative correlation to improve the computational and statistical efficiency of a Monte Carlo study.

| Corr[\( Y_1, Y_2 \)] Sample Size 95% C.I. |
|-----------------------------|--------|------------------|
| 0.000 \( n = 10 \)          |        | [4.552, 5.448]  |
| -0.647                      |        | [4.715, 5.252]  |
| 0.000 \( n = 26 \)          |        | [4.726, 5.273]  |
| -0.647                      |        | [4.841, 5.158]  |

Statistical efficiency can also be improved when using the fifth-order polynomial in lieu of the third-order polynomial. For example, consider approximating the uniform distribution. The kurtosis for this distribution is theoretically \(-1.20\). However, the lower-boundary of kurtosis for the third-order polynomial is \(-1.15132\) (Headrick & Sawilowsky, 2000) whereas the fifth-order polynomial can generate this distribution with the required kurtosis (Headrick, 2002, Table 1). Presented in Table 2 is a comparison between the two polynomials' approximations to the uniform distribution. By inspection of the values of RMSE in Table 2, it is evident that the fifth-order polynomial is superior in its approximation to the standardized cumulants of this distribution.

Lower Boundary Points of Kurtosis

The lower boundary points of kurtosis is another topic of concern because neither the third nor the fifth-order polynomial transformations span the entire skew (\( \gamma_3 \)) and kurtosis (\( \gamma_4 \)) plane given by the general expression

\[ \gamma_4 \geq \gamma_3^2 - 2. \]

(10)
Table 2. Estimates of the first six standardized cumulants of the uniform density and the Root Mean Square Errors for the third and fifth-order polynomials. Estimates (\(\hat{\gamma}_i\)) are based on a sample size of \(n = 50\) and averaged across 50,000 repetitions. The same random numbers were used in both polynomials.

<table>
<thead>
<tr>
<th>Standardized Parameters</th>
<th>Uniform Distribution ((\gamma_i))</th>
<th>(\hat{\gamma}_i)</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Third-Order Polynomial</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(\gamma_1 = 0.0)</td>
<td>0.000</td>
<td>0.142</td>
<td></td>
</tr>
<tr>
<td>(\gamma_2 = 1.0)</td>
<td>1.000</td>
<td>0.132</td>
<td></td>
</tr>
<tr>
<td>(\gamma_3 = 0.0)</td>
<td>0.002</td>
<td><strong>0.338</strong></td>
<td></td>
</tr>
<tr>
<td>(\gamma_4 = -6/5)</td>
<td>(-1.152^1)</td>
<td><strong>1.673</strong></td>
<td></td>
</tr>
<tr>
<td>(\gamma_5 = 0.0)</td>
<td>0.095</td>
<td><strong>15.771</strong></td>
<td></td>
</tr>
<tr>
<td>(\gamma_6 = 48/7)</td>
<td>8.711</td>
<td><strong>161.61</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Fifth-Order Polynomial</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(\gamma_1 = 0.0)</td>
<td>0.000</td>
<td>0.142</td>
<td></td>
</tr>
<tr>
<td>(\gamma_2 = 1.0)</td>
<td>1.000</td>
<td>0.127</td>
<td></td>
</tr>
<tr>
<td>(\gamma_3 = 0.0)</td>
<td>0.001</td>
<td><strong>0.278</strong></td>
<td></td>
</tr>
<tr>
<td>(\gamma_4 = -6/5)</td>
<td>(-1.200)</td>
<td><strong>0.354</strong></td>
<td></td>
</tr>
<tr>
<td>(\gamma_5 = 0.0)</td>
<td>0.006</td>
<td><strong>0.897</strong></td>
<td></td>
</tr>
<tr>
<td>(\gamma_6 = 48/7)</td>
<td>6.841</td>
<td><strong>3.301</strong></td>
<td></td>
</tr>
</tbody>
</table>

\(^1\)The lower boundary of kurtosis for the third-order polynomial is \(-1.15132\).

Proof (Eq. 10). For any random variable with finite values of \(\gamma_i\) define

\[
\gamma_i = \frac{E[X - E[X]]^4}{\sigma_X^4} = \left(\frac{X - E[X]}{\sigma_X}\right)^4. \tag{11}
\]

Without loss of generality, it can be assumed that the random variable \(X\) is standardized such that \(E[X] = 0\) and \(\sigma_X = E[X^2] = 1\) in (11). From the covariance (or Schwarz) inequality there is \(E[XW]^2 \leq E[X^2]E[W^2]\). If the two random variables in the covariance inequality are \(X\) and \(X^2 - 1\), then

\[
\left(E[X(X^2 - 1)]\right)^2 \leq E[X^2] \times E[(X^2 - 1)^2]
\]

\[
\left(E[X^3 - X]\right)^2 \leq E[X^4 - 2X + 1]
\]

\[
\left(E[X^3]\right)^2 \leq E[X^4] - 1
\]

\(\gamma_3 \leq \gamma_4 - 1\), thus

\(\gamma_4 \geq \gamma_3^2 + 1\), and where

subtracting a constant of 3, such that kurtosis for the normal distribution is zero, gives (10) (It can also be shown that the equality condition in (10) is not possible. However, in the context of this paper, the matter is trivial).

Presented in Table 3 are the lower boundary points of kurtosis for both polynomials. The values of minimum kurtosis \((\gamma_4', \gamma_4'^*)\) were obtained by minimizing Equation 14 (Headrick & Sawilowsky, 2000) and Equation 36 (Headrick, 2002) using the command ‘NMinimize’ (Mathematica, Wolfram, 2003, version 5.0). By inspection of Table 3, it is evident that the fifth-order polynomial spans a much larger space in the plane defined by (10) than the third-order polynomial.

Pearson Correlations

As mentioned, the third and fifth-order polynomial transformations are computationally efficient algorithms for generating multivariate non-normal distributions. In general, and in terms of the fifth-order polynomial, the approach taken is to solve the equation given in Headrick (2002, Equation, 26) for pairwise intermediate correlations between \(k\) variables.

The intermediate correlations are subsequently assembled into a correlation matrix and factored (e.g., a Cholesky factorization). The components from the factorization are used to generate multivariate standard normal random deviates correlated at an intermediate level. These deviates are then transformed by the polynomials to produce the specified non-normal distributions with the desired intercorrelations.
There are limitations in simulating multivariate distributions using the polynomial transformations. Specifically, the third and fifth-order polynomials may have lower and upper boundary points \((a^-, a^+)\) for correlations \((r)\) such that 
\[
1, 1 \quad r \in -a^- < a^+ < 1.
\]
In the context of the bivariate case, this problem is most pronounced when one distribution is symmetric and the other skewed.

For example, suppose the distributions are approximate chi-square \((1\text{df})\) and normal. The boundaries of correlation for the third-order polynomial are \(a = \pm .67481\) whereas the boundaries for the fifth-order polynomial are \(a = \pm .82024\). As another example, if the normal distribution is replaced by the coefficients for the uniform distribution, then the boundaries for bivariate correlation are \(a = \pm .623033\) and \(a = \pm .738553\) for the third and fifth-order polynomials, respectively. Thus, the fifth-order polynomial can be a remedy for cases where it is needed to simulate the often used correlation of \(r = .70\) when the distributional conditions make it impossible for the third-order polynomial.

### Table 3. Lower boundaries of kurtosis for the third \((\gamma_4')\) and fifth \((\gamma_5')\) order polynomials for a given value of skew \((\gamma_3)\). The coefficients \(c_0, \ldots, c_5\) are associated with the fifth-order polynomial.

<table>
<thead>
<tr>
<th>(\gamma_3)</th>
<th>(\gamma_4')</th>
<th>(\gamma_5')</th>
<th>(c_0)</th>
<th>(c_1)</th>
<th>(c_2)</th>
<th>(c_3)</th>
<th>(c_4)</th>
<th>(c_5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00</td>
<td>-1.151320</td>
<td>-1.385081</td>
<td>0.000000</td>
<td>-1.643734</td>
<td>0.000000</td>
<td>0.000000</td>
<td>0.320242</td>
<td>-0.011361</td>
</tr>
<tr>
<td>0.25</td>
<td>-1.045100</td>
<td>-1.296301</td>
<td>-0.160182</td>
<td>-1.597079</td>
<td>0.195003</td>
<td>0.302208</td>
<td>-0.011607</td>
<td>-0.010437</td>
</tr>
<tr>
<td>0.50</td>
<td>-0.741671</td>
<td>-1.038260</td>
<td>-0.298119</td>
<td>1.492904</td>
<td>0.036292</td>
<td>-0.266933</td>
<td>-0.021600</td>
<td>0.008682</td>
</tr>
<tr>
<td>0.75</td>
<td>-0.252697</td>
<td>-0.614627</td>
<td>-0.419443</td>
<td>1.357093</td>
<td>0.508113</td>
<td>-0.228251</td>
<td>-0.029554</td>
<td>0.006969</td>
</tr>
<tr>
<td>1.00</td>
<td>0.424841</td>
<td>-0.020321</td>
<td>-0.529477</td>
<td>1.190353</td>
<td>0.637194</td>
<td>-0.187141</td>
<td>-0.035906</td>
<td>0.005314</td>
</tr>
<tr>
<td>1.25</td>
<td>1.297258</td>
<td>0.753833</td>
<td>-0.632000</td>
<td>0.981640</td>
<td>0.754682</td>
<td>-0.141828</td>
<td>-0.040894</td>
<td>0.003602</td>
</tr>
<tr>
<td>1.50</td>
<td>2.370670</td>
<td>1.724592</td>
<td>-0.732543</td>
<td>0.690295</td>
<td>0.866255</td>
<td>-0.087835</td>
<td>-0.044570</td>
<td>0.001719</td>
</tr>
<tr>
<td>1.75</td>
<td>3.652341</td>
<td>2.757983</td>
<td>-0.503230</td>
<td>0.829259</td>
<td>0.623359</td>
<td>0.006876</td>
<td>-0.040043</td>
<td>-0.002257</td>
</tr>
<tr>
<td>2.00</td>
<td>5.151620</td>
<td>3.983870</td>
<td>-0.524421</td>
<td>0.710491</td>
<td>0.645056</td>
<td>0.048321</td>
<td>-0.040213</td>
<td>-0.004000</td>
</tr>
</tbody>
</table>

### Monotonicity and Spearman Correlations

A monotonic relationship between \(Y_i\) and \(Z_i\) in (3) is defined as

\[
Z_{ui} > Z_{ij} \Rightarrow Y_{ui} > Y_{ij}, \quad \forall u \neq j.
\]

(12)

Testing for a monotonic relationship can be accomplished by solving \(dY_i/dZ_i = 0\) for \(Z_i\). If only complex solutions of \(Z_i\) exist then the transformation between \(Y_i\) and \(Z_i\) is considered globally monotonic. If real solutions of \(Z_i\) exist, then the transformation is considered non-monotonic. For example, all chi-square distributions \((df > 1)\) approximated by fifth-order polynomials are globally monotonic transformations. The third-order polynomials, however, are not monotone transformations for any approximation of the chi-square family (see Headrick, 2004). The concern for monotonic relationships becomes important when there is a need to simulate ranked data with specified Spearman correlations.

Consider generating \(Y_i\) and \(Y_j\) from equations of the form in (3) with a Pearson correlation \(\rho_{Y_iY_j}\). Let \(R(Y_i)\) and \(R(Y_j)\) denote the ranks of \(Y_i\) and \(Y_j\) and \(R(Z_i)\) and \(R(Z_j)\)
denote the ranks of $Z_1$ and $Z_2$. If monotonic relationships hold for both transformations as defined in (12), then $\rho_{R(Y_1)R(Y_2)} = \rho_{R(Z_1)R(Z_2)} = \rho_s$ and where $\rho_s$ denotes the Spearman rank coefficient of correlation.

Because the structure of $Z_1$ and $Z_2$ is standard bivariate normal, ranked data can be simulated for specified values of $\rho_s$ and $n$ by making use of the following expression (Moran, 1948)

$$
\rho_s = \frac{6}{\pi} \frac{n-2}{n+1} \sin^{-1} \left( \frac{\rho_{Z_1Z_2}}{2} \right) + \frac{1}{n+1} \sin^{-1} \left( \rho_{Z_1Z_2} \right),
$$

(13)

More specifically, to generate $R(Y_1)$ and $R(Y_2)$ with a specified rank correlation $\rho_s$ and sample size, one need only numerically solve (13) for $\rho_{Z_1Z_2}$ given values of $\rho_s$ and $n$. For example, suppose it is desired to generate $R(Y_1)$ and $R(Y_2)$ with a Spearman rank correlation of $\rho_s = .70$, $n = 5$, and where the distributions $Y_1$ and $Y_2$ are approximate exponential. For this example, it is appropriate to use fifth-order polynomial transformations because (12) holds for this case. Thus, solving (13) for the specified values of $\rho_s$ and $n$ gives an intermediate correlation of $\rho_{Z_1Z_2} = .811202$.

Conclusion

In terms of the procedure for simulating ranked data with specified Spearman correlations, it should be pointed out that equation (12) is a sufficient condition for monotonicity. However, the procedure will provide adequate simulations of rank data with specified correlations if the polynomial transformations are locally monotonic. More specifically, the simulated rank correlations may be robust to violations of (12) even though real solutions of $Z_1$ (or $Z_2$) exist for $dY_1/dZ_1 = 0$ (or $dY_2/dZ_2 = 0$). For example, assume more generally, for two symmetric distributions of the same shape that $Z \pm 3.00$ are real solutions for $dY/dZ = 0$.

These distributions could be considered locally monotonic because the probability associated with drawing such values of $Z$ is $|Z| \geq 3.00$ is only .0027. Because the probability of obtaining such values of $Z$ is very low, the amount of bias introduced into a Monte Carlo or simulation study would be negligible.

To provide an empirical definition of local monotonicity, this author conducted simulations using fifth-order transformations with many different non-normal distributions with nonmonotonic relationships. The simulation results indicated that Spearman correlations were close to what (13) would compute ($\rho_s \pm .025$) if the real solutions of $Z$ for $dY/dZ = 0$ were $|Z| \geq 1.75$.

References


