Journal of Modern Applied Statistical Methods

Volume 3 | Issue 1 Article 8

5-1-2004

On Polynomial Transformations For Simulating Multivariate Non-normal Distributions

Todd C. Headrick
Southern Illinois University at Carbondale, headrick@siu.edu

Journal of Modern Applied Statistical Methods
May, 2004, Vol. 3, No. 1, 65-71

1538 – 9472/04/$95.00

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 non-normality is the method of Monte Carlo. For example, consider the following polynomial transformation in general form

$$Y_1 = c_0 + \sum_{i=1}^{m} c_i Z_1^i \qquad (1)$$

where Z_1 ~ NID(0,1), and $i \in \mathbb{N} = \{1, 2, ..., m\}$. Setting $m = 3$, Fleishman (1978) derived a system of four equations that would solve for the four coefficients $c_0, ..., c_3$ 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_1 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).

Todd C. Headrick is Associate Professor of Statistics. Address: Section on Statistics and Measurement, Department of EPSE, 222-J Wham Building, Mail Code 4618, Southern Illinois University-Carbondale, IL, 62901. His areas of research interest are statistical computing, nonparametric statistics, and optimization. Email address: headrick@siu.edu.

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_1, *in general*, is not exact. Headrick (2004) has derived the probability density function and distribution function for Y_1 when the transformation between Y_1 and Z_1 is monotic.

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

$$Y_1 = c_0 + c_1 Z_1 + c_2 Z_1^2 + c_3 Z_1^3 + c_4 Z_1^4 + c_5 Z_1^5 \quad (2)$$

or

$$Y_1 = c_0 + Z_1\left(c_1 + Z_1\left(c_2 + Z_1\left(c_3 + Z_1\left(c_4 + c_5 Z_1\right)\right)\right)\right). \quad (3)$$

If the algorithm used to generate Y_1 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 $(Y_1 + Y_2)/2$ is by inducing a negative correlation on Y_1 and Y_2. To demonstrate, if Y_1 and Y_2 were identically distributed, then

$$\text{Var}\left[\frac{Y_1 + Y_2}{2}\right] = \frac{1}{2} + \frac{\text{Corr}[Y_1, Y_2]}{2}. \quad (4)$$

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

Assume that a monotone relationship between Z_1 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

$$Y_1 = f_1(c_0, c_1, c_2, c_3, c_4, c_5, Z_1) \quad (5)$$

$$Y_2 = f_2(c_0, -c_1, c_2, -c_3, c_4, -c_5, Z_1). \quad (6)$$

Because the structure between Y_i and Z_1 is standard bivariate normal, the correlation between Y_1 and Y_2 can be defined as

$$\rho_{Y_1 Y_2} = E[Y_1 Y_2]. \quad (7)$$

Expanding (7) and taking expectations using the moments from the standard normal density yields

$$\rho_{Y_1Y_2} = c_0^2 - c_1^2 + 2c_0(c_2 + 3c_4) - 6c_1(c_3 + 5c_5) + 3(c_2^2 + 10c_2c_4 - 5(c_3^2 - 7c_4^2 + 14c_3c_5 + 63c_5^2)). \quad (8)$$

Thus, the correlation between Y_1 and Y_2 can be determined by evaluating (8) using specified values for $c_0,...,c_5$. For example, evaluating (8) using the coefficients that approximate the exponential density (see Headrick, 2002, Table 1) gives $\rho_{Y_1Y_2} \cong -0.647$.

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 θ and θ^2. From the definition of the product moment of correlation exists

$$E[X_1X_2] = \theta^2 \int_0^1 \ln v \ln(1-v)dv = \theta^2(2 - \pi^2/6).$$

As such, the correlation between X_1 and X_2 is

$$\rho_{X_1X_2} = 1 - \pi^2/6 \cong -0.645. \quad (9)$$

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.

Table 1. Confidence Intervals (CI's) on the estimate of $(Y_1 + Y_2)/2$ with and without a negative correlation induced. Y_1 and Y_2 are approximate exponential distributions with population means of $\gamma_1 = 5$. The CI's are based on 50,000 sample estimates.

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 (γ_3) and kurtosis (γ_4) plane given by the general expression

$$\gamma_4 \geq \gamma_3^2 - 2. \quad (10)$$

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.

Standardized Parameters Uniform Distribution (γ_i)	$\hat{\gamma}_i$	RMSE
Third-Order Polynomial		
$\gamma_1 = 0.0$	0.000	0.142
$\gamma_2 = 1.0$	1.000	0.132
$\gamma_3 = 0.0$	0.002	**0.338**
$\gamma_4 = -6/5$	–1.152[1]	**1.673**
$\gamma_5 = 0.0$	0.095	**15.771**
$\gamma_6 = 48/7$	8.711	**161.61**
Fifth-Order Polynomial		
$\gamma_1 = 0.0$	0.000	0.142
$\gamma_2 = 1.0$	1.000	0.127
$\gamma_3 = 0.0$	0.001	**0.278**
$\gamma_4 = -6/5$	− 1.200	**0.354**
$\gamma_5 = 0.0$	0.006	**0.897**
$\gamma_6 = 48/7$	6.841	**3.301**

[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 γ_i define

$$\gamma_i = \frac{E[X - E[X]]^i}{\sigma_X^i} = E\left(\frac{X - E[X]}{\sigma_X}\right)^i. \quad (11)$$

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 $E[XW]^2 \le E[X^2]E[W^2]$. If the two random variables in the covariance inequality are X and $X^2 - 1$, then

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

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

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

$$\gamma_3^2 \le \gamma_4 - 1, \text{ thus}$$

$$\gamma_4 \ge \gamma_3^2 + 1, \text{ and where}$$

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 (γ_4', γ_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.

Table 3. Lower boundaries of kurtosis for the third (γ_4') and fifth (γ_4^*) order polynomials for a given value of skew (γ_3). The coefficients $c_0,...,c_5$ are associated with the fifth-order polynomial.

γ_3	γ_4'	γ_4^*	c_0	c_1	c_2	c_3	c_4	c_5
0.00	-1.151320	-1.385081	0.000000	-1.643734	0.000000	0.320242	0.000000	-0.011361
0.25	-1.045100	-1.296301	-0.160182	-1.597079	0.195003	0.302208	-0.011607	-0.010437
0.50	-0.741671	-1.038260	-0.298119	1.492904	0.036292	-0.266933	-0.021600	0.008682
0.75	-0.252697	-0.614627	-0.419443	1.357093	0.508113	-0.228251	-0.029554	0.006969
1.00	0.424841	-0.020321	-0.529477	1.190353	0.637194	-0.187141	-0.035906	0.005314
1.25	1.297258	0.753833	-0.632000	0.981640	0.754682	-0.141828	-0.040894	0.003602
1.50	2.370670	1.724592	-0.732543	0.690295	0.866255	-0.087835	-0.044570	0.001719
1.75	3.652341	2.757983	-0.503230	0.829259	0.623359	0.006876	-0.040043	-0.002257
2.00	5.151620	3.983870	-0.524421	0.710491	0.645056	0.048321	-0.040213	-0.004000

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 $r \in [-1 < -a, a < +1]$. 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*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.

Monotinicity and Spearman Correlations

A monotonic relationship between Y_1 and Z_1 in (3) is defined as

$$Z_{1i} > Z_{1j} \Rightarrow Y_{1i} > Y_{1j}, \ \forall_{i \neq j}. \quad (12)$$

Testing for a monotonic relationship can be accomplished by solving $dY_1/dZ_1 = 0$ for Z_1. If only complex solutions of Z_1 exist then the transformation between Y_1 and Z_1 is considered globally monotonic. If real solutions of Z_1 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_1 and Y_2 from equations of the form in (3) with a Pearson correlation $\rho_{Y_1 Y_2}$. Let $R(Y_1)$ and $R(Y_2)$ denote the ranks of Y_1 and Y_2 and $R(Z_1)$ and $R(Z_2)$

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 ρ_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 ρ_s and *n* by making use of the following expression (Moran, 1948)

$$\rho_s = \frac{6}{\pi}\left\{\frac{n-2}{n+1}\sin^{-1}\left(\frac{\rho_{Z_1Z_2}}{2}\right)+\frac{1}{n+1}\sin^{-1}\left(\rho_{Z_1Z_2}\right)\right\}. \quad (13)$$

More specifically, to generate $R(Y_1)$ and $R(Y_2)$ with a specified rank correlation ρ_s and sample size, one need only numerically solve (13) for $\rho_{Z_1Z_2}$ given values of ρ_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 specifed values of ρ_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 $Z : |Z| \geq 3.00$ 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

Beasley, T. M. (2002). Multivariate aligned rank test for interactions in multiple group repeated measures. *Multivariate Behavioral Research, 37*, 197-226.

Beasley, T. M. & Zumbo, B. D. (2003). Comparison of aligned Friedman rank and parametric methods for testing interactions in split-plot designs. *Computational Statistics and Data Analysis, 42*, 569-593.

Fleishman, A. I. (1978). A method for simulating non-normal distributions. *Psychometrika, 43*, 521-532.

Habib, A. R., & Harwell, M. R. (1989). An empirical study of the Type I error rate and power of some selected normal theory and nonparametric tests of independence of two sets of variables. *Communications in Statistics: Simulation and Computation, 18*, 793-826.

Harwell, M. R. & Serlin, R. C. (1988). An experimental study of a proposed test of nonparametric analysis of covariance. *Psychological Bulletin, 104*, 268-281.

Harwell, M. R. & Serlin, R. C. (1989). A nonparametric test statistic for the general linear model. *Journal of Educational Statistics, 14*, 351-371.

Harwell, M. R. & Serlin, R. C. (1997). An empirical study of finve multivariate tests for the single-factor repeated measures model. *Communications in Statistics: Simulation and Computation, 26*, 605-618.

Headrick, T. C. (2002). Fast fifth-order polynomial transforms for generating univariate and multivariate non-normal distributions. *Computational Statistics and Data Analysis, 40*, 685-711.

Headrick, T. C. & Beasley, T. M. (2004). A method for simulating correlated non-normal systems of linear statistical equations. *Communications in Statistics: Simulation and Computation, 33*, 19-33.

Headrick, T. C. & Beasley, T. M. (April, 2003). A method for simulating correlated structures of continuous and ranked data. American Educational Research Association: Chicago.

Headrick, T. C. & Rotou (2001). An investigation of the rank transformation in multiple regression. *Computational Statistics and Data Analysis, 38*, 203-215.

Headrick, T. C., & Sawilowsky, S. S. (1999). Simulating correlated non-normal distributions: extending the Fleishman power method. *Psychometrika, 64*, 25-35.

Headrick, T. C. & Sawilowsky, S. S. (2000). Properties of the rank transformation in factorial analysis of covariance. *Communications in Statistics: Simulation and Computation, 29*, 1059-1088.

Headrick, T. C., & Sawilowsky, S. S. (2000). Weighted simplex procedures for determining boundary points and constants for the univariate and multivariate power methods. *Journal of Educational and Behavioral Statistics, 25*. 417-436.

Headrick, T. C. & Vineyard, G. (2001). An empirical investigation of four tests for interaction in the context of factorial analysis of covariance. *Multiple Linear Regression Viewpoints, 27*, 3-15.

Headrick, T. C. (2004). Distribution theory for the power method transformation. The Psychometric Society: Monterey.

Klockars, A. J. & Moses, T. P. (2002). Type I error rates for rank-based tests of homogeneity of regression slopes. *Journal of Modern Applied Statistical Methods, 1*, 452-460.

Moran, P.A.P. (1948). Rank correlation and product-moment correlation. *Biometrika, 35*, 203-206

Shieh, Y., (April, 2000). The effects of distributional characteristics on multi-level modeling parameter estimates and Type I error control of parameter tests under conditions of non-normality. American Educational Research Association: New Orleans.

Vale, C. D. & Maurelli, V. A. (1983). Simulating univariate and multivariate non-normal distributions. *Psychometrika, 48*, 465-471.

Whittaker, T., Fouladi, R. T., & Williams, N. (2002). Determining predictor importance in multiple regression under varied correlational and distributional conditions. *Journal of Modern Applied Statistical Methods, 2*, 354-366.

Wolfram, S., (2003). *The Mathematica book*, 4th Edition. Wolfram Media-Cambridge Univeristy Press, Cambridge.