

## **A Method for Simulating Correlated Non-Normal Systems of Linear Statistical Equations**

**Todd C. Headrick<sup>1,\*</sup> and T. Mark Beasley<sup>2</sup>**

<sup>1</sup>Section on Statistics and Measurement, Department of  
EPSE, Southern Illinois University-Carbondale,  
Carbondale, Illinois, USA

<sup>2</sup>Section on Statistical Genetics, Department of Biostatistics,  
University of Alabama-Birmingham,  
Birmingham, Alabama, USA

### **ABSTRACT**

A procedure is derived for simulating correlated non-normal systems of linear statistical equations. The method is based on fifth-order polynomial transformations to generate multivariate non-normal distributions. The procedure allows for the simultaneous control of the correlated non-normal (a) stochastic disturbance distributions, (b) independent variables, and (c) dependent and independent variables for each equation throughout a system. A numerical example is provided to demonstrate the procedure. The results of a Monte

---

\*Correspondence: Todd C. Headrick, Section on Statistics and Measurement, Department of EPSE, 222-J Wham Building, Mail Code 4618, Southern Illinois University-Carbondale, Carbondale, IL 62901, USA; Fax: 618-453-7110; E-mail: headrick@siu.edu.

Carlo simulation are provided to confirm that the proposed method generates the specified standardized cumulants and correlations.

*Key Words:* Monte Carlo; Polynomial transformation; Standardized cumulants.

## 1. INTRODUCTION

The availability of the desktop computer has made simulation and Monte Carlo methods widely applicable in statistical research. For example, Monte Carlo techniques may be used to compare the small sample properties of a test statistic with its competitors or whether these properties are consistent with the statistic's asymptotic approximation (Headrick and Rotou, 2001; Headrick and Sawilowsky, 2000; Holgersson and Shuker, 2001). Further, simulation and Monte Carlo techniques are now applicable to many specific areas of research interest. Some examples include: bootstrap tilting (Hesterberg, 2001); conditional logistic regression (Mehta et al., 2000); likelihood inference with missing data (Gilks et al., 1998); and statistical genetics (Thompson, 2000).

With this plethora of uses for Monte Carlo techniques in mind, there may be occasions when it is desirable to investigate the properties of statistics that involve systems of statistical equations under a variety of conditions. Consider the following system of  $T$  equations

$$\mathbf{y}_p = \mathbf{x}_p \boldsymbol{\beta}_p + \sigma_p \boldsymbol{\epsilon}_p, \quad (1)$$

where  $p = 1, \dots, T$ ,  $\mathbf{y}_p$  and  $\boldsymbol{\epsilon}_p$  have dimension  $(N \times 1)$ ,  $\mathbf{x}_p$  is  $(N \times k_p)$ ,  $\boldsymbol{\beta}_p$  is  $(k_p \times 1)$ , and  $\sigma_p$  is a real positive scalar. These equations can be completely expressed as a linear system as

$$\mathbf{y} = \mathbf{x}\boldsymbol{\beta} + \boldsymbol{\sigma}\boldsymbol{\epsilon}, \quad (2)$$

where  $\mathbf{y}$  and  $\boldsymbol{\epsilon}$  have dimension  $(TN \times 1)$ ,  $\mathbf{x}$  is  $(TN \times K)$ ,  $\boldsymbol{\beta}$  is  $(K \times 1)$ ,  $K = \sum_{p=1}^T k_p$ , and  $\boldsymbol{\sigma} = (\sigma_1, \dots, \sigma_T)$  represents  $T$  scalars associated with each of the  $T$  equations. The stochastic disturbances  $\boldsymbol{\epsilon}$  in (2) are assumed to be centered and have unit variances.

If the disturbance terms in (2) are correlated (e.g.,  $\boldsymbol{\epsilon}_p$  is correlated with  $\boldsymbol{\epsilon}_q$ ) then a gain in efficiency can be realized by using the method of generalized least squares (GLS) as opposed to using ordinary least squares (OLS) (Judge et al., 1985). This approach of joint estimation is perhaps better known as "seemingly unrelated regression equation



estimation" (Zellner, 1962). The stronger the correlations are between the disturbances (or the weaker the correlations are between the independent variables) in (2), then the greater the efficiency GLS has relative to OLS (Dwivedi and Srivastava, 1978). Thus, it may be desirable to study the relative Type I error and power of the GLS and OLS estimators under non-normal conditions. Such an investigation would usually be carried out using Monte Carlo techniques. To determine any advantages of GLS relative to OLS, a variety of non-normal distributions with varying degrees of correlation between the disturbances would usually be included in the study.

There are many methods and applications of linear models that involve a set of statistical equations. Some examples include: confirmatory factor analysis; generalized linear models; hierarchical linear models; models of several time series; structural equation modeling; and other applications of the general linear model (e.g., the analysis of covariance).

Most statistics textbooks discuss the validity of linear models or test statistics in terms of the various assumptions concerning the stochastic disturbances (Cook and Weisberg, 1999; Neter et al., 1996). For example, the usual OLS regression procedure assumes that the disturbances are independent and normally distributed with conditional expectation of zero and constant variance. Thus, in order to examine the properties of a system of statistical equations using Monte Carlo techniques, it is necessary to have an appropriate data generation procedure that allows for the pre-specification of the distributional shapes and correlation structures of the stochastic disturbances (such as the vector  $\epsilon$  in Eq. 2). It is also desirable that this procedure be both computationally efficient and general enough to allow for the simulation of a variety of conditions such as autocorrelation, heterogeneity of regression coefficients, heteroscedasticity, multicollinearity, non-normality, and other violations of assumptions.

## 2. PURPOSE OF THE STUDY

The purpose of the study is to derive a computationally efficient procedure for simulating systems of correlated non-normal linear statistical equations. The procedure developed in this study is an extension of the methodology described in Headrick (2002) for simulating multivariate non-normal distributions. More specifically, the Headrick (2002) coefficient model is presented in conjunction with the development of the methodology needed to build and model a system of  $T$  equations. The proposed procedure allows for the simultaneous control of the degree



of non-normality and correlations between the (a) stochastic disturbances (b) independent variables, and (c) dependent and independent variables for each equation in a system. A numerical example is provided to demonstrate the procedure. The results of a simulation are provided to confirm that the procedure generates the specified standardized cumulants and correlation structures. *Mathematica* (Wolfram, 1999) notebooks are also available (from the first author) for implementing the procedure.

### 3. MATHEMATICAL DEVELOPMENT

Consider the linear system of  $T$  equations in (2), more explicitly,

$$Y_p = \beta_{p0} + \beta_{p1}X_{p1} + \cdots + \beta_{pi}X_{pi} + \cdots + \beta_{pj}X_{pj} + \cdots + \beta_{pk}X_{pk} + \sigma_p \varepsilon_p, \quad (3)$$

where  $p = 1, \dots, q, \dots, T$ . Note that it is not necessary for each of the  $T$  equations to have the same number of independent variables ( $X$ 's).

Two non-normal independent variables  $X_{pi}$  and  $X_{qj}$  in the system of (3) are generated and correlated according to the fifth-order polynomial transformation derived by Headrick (2002, Eqs. (23) and (24)) as follows

$$X_{pi} = c_{0pi} + c_{1pi}X'_{pi} + c_{2pi}X'^2_{pi} + c_{3pi}X'^3_{pi} + c_{4pi}X'^4_{pi} + c_{5pi}X'^5_{pi}, \quad (4a)$$

$$X_{qj} = c_{0qj} + c_{1qj}X'_{qj} + c_{2qj}X'^2_{qj} + c_{3qj}X'^3_{qj} + c_{4qj}X'^4_{qj} + c_{5qj}X'^5_{qj}, \quad (4b)$$

where  $X'_{pi}$  and  $X'_{qj} \sim N(0,1)$  and have intermediate correlation of  $\rho_{X'_{pi}X'_{qj}}$ . To generate  $X_{pi}$  and  $X_{qj}$  the two sets of power constants,  $c_{0**}, \dots, c_{5**}$  in (4a) and (4b), and an intermediate correlation between  $X'_{pi}$  and  $X'_{qj}$  are first obtained.

The power constants in (4a) are determined by simultaneously solving Headrick's (2002) Eqs. (18), (22), (B.1), (B.2), (B.3), and (B.4). In general, Eqs. (18) and (22) are equated to zero and one while (B.1), (B.2), (B.3), and (B.4) are equated to the third, fourth, fifth, and sixth standardized cumulants ( $\gamma_1, \gamma_2, \gamma_3, \gamma_4$ ) from a specified non-normal distribution. On solving this system, the solutions of the power constants are subsequently entered into (4a) to produce  $X_{pi}$  which has zero (marginal) mean, unit variance, and the specified values of  $\gamma_1, \gamma_2, \gamma_3$ , and  $\gamma_4$ . The power constants in (4b) are determined in the same manner.



The intermediate correlation  $\rho_{X'_{pi}X'_{qj}}$  is determined from the equation derived in the methodology described in Headrick (2002, Eq. 26). This equation is expressed as follows

$$\begin{aligned}
 \rho_{X_{pi}X_{qj}} = & 3c_{4pi}c_{0qj} + 3c_{4pi}c_{2qj} + 9c_{4pi}c_{4qj} + c_{0pi}(c_{0qj} + c_{2qj} + 3c_{4qj}) \\
 & + c_{1pi}c_{1qj}\rho_{X'_{pi}X'_{qj}} + 3c_{3pi}c_{1qj}\rho_{X'_{pi}X'_{qj}} + 15c_{5pi}c_{1qj}\rho_{X'_{pi}X'_{qj}} \\
 & + 3c_{1pi}c_{3qj}\rho_{X'_{pi}X'_{qj}} + 9c_{3pi}c_{3qj}\rho_{X'_{pi}X'_{qj}} + 45c_{5pi}c_{3qj}\rho_{X'_{pi}X'_{qj}} \\
 & + 15c_{1pi}c_{5qj}\rho_{X'_{pi}X'_{qj}} + 45c_{3pi}c_{5qj}\rho_{X'_{pi}X'_{qj}} + 225c_{5pi}c_{5qj}\rho_{X'_{pi}X'_{qj}} \\
 & + 12c_{4pi}c_{2qj}\rho_{X'_{pi}X'_{qj}}^2 + 72c_{4pi}c_{4qj}\rho_{X'_{pi}X'_{qj}}^2 + 6c_{3pi}c_{3qj}\rho_{X'_{pi}X'_{qj}}^3 \\
 & + 60c_{5pi}c_{3qj}\rho_{X'_{pi}X'_{qj}}^3 + 60c_{3pi}c_{5qj}\rho_{X'_{pi}X'_{qj}}^3 + 600c_{5pi}c_{5qj}\rho_{X'_{pi}X'_{qj}}^3 \\
 & + 24c_{4pi}c_{4qj}\rho_{X'_{pi}X'_{qj}}^4 + 120c_{5pi}c_{5qj}\rho_{X'_{pi}X'_{qj}}^5 \\
 & + c_{2pi}\left(c_{0qj} + c_{2qj} + 3c_{4qj} + 2c_{2qj}\rho_{X'_{pi}X'_{qj}}^2 + 12c_{4qj}\rho_{X'_{pi}X'_{qj}}^2\right). \quad (5)
 \end{aligned}$$

The intermediate correlation is computed by solving (5) for  $\rho_{X'_{pi}X'_{qj}}$ , given a specified correlation of  $\rho_{X_{pi}X_{pj}}$  between the independent variables and the specified power constants from (4a) and (4b).

**Remark 3.1.** By inspection of (4a), (4b), and (5), if the structure of  $X_{pi}$  and  $X_{qj}$  is the special case of bivariate normal, then  $c_{1**} = 1$ , and  $c_{0**}, c_{2**}, \dots, c_{5**} = 0$ . Hence,  $X_{pi} = X'_{pi}, X_{qj} = X'_{qj}$ , and  $\rho_{X_{pi}X_{pj}} = \rho_{X'_{pi}X'_{qj}}$ .

Two stochastic disturbance terms  $\varepsilon_p$  and  $\varepsilon_q$  in the system of (3) are generated and correlated in the manner described for the independent variables  $X_{pi}$  and  $X_{qj}$ . Note that  $\varepsilon_p$  and  $\varepsilon_q$  are generated independently of  $X_{pi}$  and  $X_{qj}$ .

The dependent variables  $Y_p$  for the system are generated from the right-hand sides of the  $T$  equations in (3). The method that correlates  $Y_p$  with the  $k_p$  independent variables  $X_{pi}$  is based on the following lemma:

**Lemma 3.2.** If the independent variables  $X_{pi}$  and  $X_{pj}$  have zero means, unit variances, correlation of  $\rho_{X_{pi}X_{pj}}$  based on (4a) and (4b), and  $\text{cov}(\varepsilon_p, X_{pi}) = 0, \forall_{i=1, \dots, k_p}$ , then the correlation between  $Y_p$  and  $X_{pi}$  in (3) is

$$\rho_{Y_p X_{pi}} = \frac{\beta_{pi} + \sum_{pj \neq pi} \beta_{pj} \rho_{X_{pi}X_{pj}}}{\sqrt{\sigma_p^2 + \sum_{pi} \beta_{pi}^2 + 2 \sum_{pj \neq pi} \beta_{pi} \beta_{pj} \rho_{X_{pi}X_{pj}}}}. \quad (6)$$

*Proof.* Equation (6) can be shown by using the definition of the product-moment of correlation and straightforward manipulations using



the algebra of expectations because the structure of  $X'_{pi}$  and  $X'_{pj}$  in (4a) and (4b) is standard bivariate normal.

The coefficients  $\beta_{pi}$  in (3) are determined by simultaneously solving a system of  $k_p$  equations of the form in (6). Specifically, the equations in this system specify pairwise correlations of  $\rho_{Y_p X_{pi}}$  on the left-hand sides. The prespecified constants of  $\rho_{X_{pi} X_{pj}}$  and  $\sigma_p$  are substituted into the right-hand sides. The solutions for  $\beta_{pi}$  are subsequently determined by numerically solving this system. □

**Remark 3.3.** *If the numerical solutions of all  $\beta_{pi}$  are finite real numbers, then the  $(1 + k_p) \times (1 + k_p)$  correlation matrix associated with  $Y_p$  and all  $X_{pi}$  is sufficiently positive definite.*

Given all fixed values of the  $\beta$  coefficients, correlations between the independent variables, correlations between the stochastic disturbances, and the scalar terms for the system in (3), other correlations that may be of interest such as  $\rho_{Y_p Y_q}$ ,  $\rho_{Y_q X_{pi}}$ , and  $\rho_{Y_p(\sigma_q \epsilon_q)}$  can be determined by evaluating

$$\rho_{Y_q X_{pi}} = \frac{\sum_{qi} \beta_{qi} \rho_{X_{pi} X_{qi}}}{\sqrt{\sigma_q^2 + \sum_{qi} \beta_{qi}^2 + 2 \sum_{qj \neq qi} \beta_{qi} \beta_{qj} \rho_{X_{qi} X_{qj}}}} \tag{7}$$

$$\begin{aligned} \rho_{Y_p Y_q} &= \frac{\sigma_p \sigma_q \rho_{\epsilon_p \epsilon_q} + \sum_{pi} \sum_{qi} \beta_{pi} \beta_{qi} \rho_{X_{pi} X_{qi}}}{\sqrt{\sigma_p^2 + \sum_{pi} \beta_{pi}^2 + 2 \sum_{pj \neq pi} \beta_{pi} \beta_{pj} \rho_{X_{pi} X_{pj}}} \sqrt{\sigma_q^2 + \sum_{qi} \beta_{qi}^2 + 2 \sum_{qj \neq qi} \beta_{qi} \beta_{qj} \rho_{X_{qi} X_{qj}}}} \end{aligned} \tag{8}$$

$$\rho_{Y_p(\sigma_q \epsilon_q)} = \frac{\sigma_p \rho_{\epsilon_p \epsilon_q}}{\sigma_q \sqrt{\sigma_p^2 + \sum_{pi} \beta_{pi}^2 + 2 \sum_{pj \neq pi} \beta_{pi} \beta_{pj} \rho_{X_{pi} X_{pj}}}}. \tag{9}$$

The derivations of Eqs. (7)–(9) are similar to that of Eq. (6).

#### 4. NUMERICAL EXAMPLE

Suppose we desire the system

$$\mathbf{y} = \mathbf{x}\boldsymbol{\beta} + \boldsymbol{\sigma}\boldsymbol{\epsilon}, \tag{10}$$

where  $p = 1, \dots, T = 4$ ;  $k = 4$  for all  $p = 1, \dots, 4$ ; and  $N = 10$ . Thus,  $\mathbf{y}$  and  $\boldsymbol{\epsilon}$  will have dimension  $(40 \times 1)$ ,  $\mathbf{x}$  is  $(40 \times 16)$ , and  $\boldsymbol{\beta}$  is  $(16 \times 1)$ .



**Table 1.** Specified correlations for the independent variables  $X_{pi}$ . The column and row headings are subscripts identifying the position of the  $pi$ -th independent variable.

$X_{pi}$	11	12	13	14	21	22	23	24	31	32	33	34	41	42	43	44
11	1	.2	.2	.2	.3	.3	.3	.3	.3	.3	.3	.3	.3	.3	.3	.3
12	.2	1	.2	.2	.3	.3	.3	.3	.3	.3	.3	.3	.3	.3	.3	.3
13	.2	.2	1	.2	.3	.3	.3	.3	.3	.3	.3	.3	.3	.3	.3	.3
14	.2	.2	.2	1	.3	.3	.3	.3	.3	.3	.3	.3	.3	.3	.3	.3
21	.3	.3	.3	.3	1	.4	.4	.4	.5	.5	.5	.5	.5	.5	.5	.5
22	.3	.3	.3	.3	.4	1	.4	.4	.5	.5	.5	.5	.5	.5	.5	.5
23	.3	.3	.3	.3	.4	.4	1	.4	.5	.5	.5	.5	.5	.5	.5	.5
24	.3	.3	.3	.3	.4	.4	.4	1	.5	.5	.5	.5	.5	.5	.5	.5
31	.3	.3	.3	.3	.5	.5	.5	.5	1	.6	.6	.6	.7	.7	.7	.7
32	.3	.3	.3	.3	.5	.5	.5	.5	.6	1	.6	.6	.7	.7	.7	.7
33	.3	.3	.3	.3	.5	.5	.5	.5	.6	.6	1	.6	.7	.7	.7	.7
34	.3	.3	.3	.3	.5	.5	.5	.5	.6	.6	.6	1	.7	.7	.7	.7
41	.3	.3	.3	.3	.5	.5	.5	.5	.7	.7	.7	.7	1	.8	.8	.8
42	.3	.3	.3	.3	.5	.5	.5	.5	.7	.7	.7	.7	.8	1	.8	.8
43	.3	.3	.3	.3	.5	.5	.5	.5	.7	.7	.7	.7	.8	.8	1	.8
44	.3	.3	.3	.3	.5	.5	.5	.5	.7	.7	.7	.7	.8	.8	.8	1

Let the non-normal distributions specified for  $\epsilon_p$  and  $x_p$  be (approximate) chi-square with  $p = 1, 2, 3,$  and  $4$  degrees of freedom ( $df$ ). To induce heteroscedasticity in (10), the scalars are set to:  $\sigma = (\sigma_1 = 0.50, \sigma_2 = 1.0, \sigma_3 = 2.0, \sigma_4 = 3.0)$ .

The specified parameters and numerical solutions for this example are summarized in the tables below. Specifically, presented in Tables 1–3 are the specified correlations between the (a) independent variables, (b) stochastic disturbances, and (c) dependent and independent variables for each equation. The standardized cumulants for each chi-square distribution are listed in Table 4.

The numerical solutions obtained from *Mathematica* (Wolfram, 1999) are presented in Tables 5–8. Listed in these tables are the power

**Table 2.** Specified correlations for the stochastic disturbances  $\epsilon_p$ .

Disturbance term	$\epsilon_1$	$\epsilon_2$	$\epsilon_3$	$\epsilon_4$
$\epsilon_1$	1	0.60	0.50	0.40
$\epsilon_2$	0.60	1	0.50	0.40
$\epsilon_3$	0.50	0.50	1	0.40
$\epsilon_4$	0.40	0.40	0.40	1



**Table 3.** Specified correlations for the dependent  $Y_p$  and independent variables  $X_{pi}$ .

Equation ( $p$ )	$X_{p1}$	$X_{p2}$	$X_{p3}$	$X_{p4}$
$Y_{p=1}$	0.20	0.20	0.10	0.10
$Y_{p=2}$	0.40	0.40	0.20	0.20
$Y_{p=3}$	0.60	0.60	0.40	0.40
$Y_{p=4}$	0.80	0.80	0.60	0.60

**Table 4.** The first six standardized cumulants for the specified chi-square distributions.

Distribution	$\mu$	$\sigma^2$	$\gamma_1$	$\gamma_2$	$\gamma_3$	$\gamma_4$
$p = 1(df)$	0	1	$2\sqrt{2}$	12	$48\sqrt{2}$	480
$p = 2(df)$	0	1	2	6	24	120
$p = 3(df)$	0	1	$2\sqrt{2}/\sqrt{3}$	4	$16\sqrt{2}/\sqrt{3}$	$160/3$
$p = 4(df)$	0	1	$\sqrt{2}$	3	$6\sqrt{2}$	30

**Table 5.** Power constants for simulating the chi-square distributions.

Distribution	$c_{0p}$	$c_{1p}$	$c_{2p}$	$c_{3p}$	$c_{4p}$	$c_{5p}$
$p = 1(df)$	-0.397725	0.621071	0.416907	0.068431	-0.006394	0.000044
$p = 2(df)$	-0.307740	0.800560	0.318764	0.033500	-0.003675	0.000159
$p = 3(df)$	-0.259037	0.867102	0.265362	0.021276	-0.002108	0.000092
$p = 4(df)$	-0.227508	0.900716	0.231610	0.015466	-0.001367	0.000055

**Table 6.** The  $\beta$  coefficients for generating the specified correlations between the dependent variable  $Y_p$  and the independent variables  $X_{pi}$  in the  $p$ th equation.

Equation ( $p$ )	$\beta_{p1}$	$\beta_{p2}$	$\beta_{p3}$	$\beta_{p4}$
$p = 1$	0.0809574	0.0809574	0.0161915	0.0161915
$p = 2$	0.3454030	0.3454030	-0.0345403	-0.0345403
$p = 3$	1.1633501	1.1633501	-0.1938917	-0.1938917
$p = 4$	4.4790534	4.4790534	-1.8662722	-1.8662722



**Table 7.** Specified and intermediate correlations for the independent variables.

Position of $X_{p^*}$ and $X_{q^*}$	Specified correlations	Intermediate correlations
$p = 1$	0.20	0.262727
$p = 1, q = 2$	0.30	0.361334
$p = 1, q = 3$	0.30	0.356923
$p = 1, q = 4$	0.30	0.355551
$p = 2$	0.40	0.446285
$p = 2, q = 3$	0.50	0.540676
$p = 2, q = 4$	0.50	0.539103
$p = 3$	0.60	0.630916
$p = 3, q = 4$	0.70	0.724123
$p = 4$	0.80	0.815464

**Table 8.** Specified and intermediate correlations for the stochastic disturbances.

Position of $\varepsilon_p$ and $\varepsilon_q$	Specified correlation	Intermediate correlation
$p = 1, q = 2$	0.60	0.664454
$p = 1, q = 3$	0.50	0.566602
$p = 1, q = 4$	0.40	0.463104
$p = 2, q = 3$	0.50	0.540676
$p = 2, q = 4$	0.40	0.437150
$p = 3, q = 4$	0.40	0.428906

constants, beta coefficients, intermediate correlations for the independent variables, and the intermediate correlations for the stochastic disturbances.

## 5. DATA GENERATION

The data generation procedure for the numerical example begins with creating a correlation matrix for the independent variables of the size in Table 1 with the intermediate correlations listed in Table 7 as its entries. A second correlation matrix is also constructed for the stochastic disturbances using the intermediate correlations in Table 8 as its entries. The two matrices are then separately factored (e.g., a Cholesky factorization). The factored matrices are then used to produce standard



normal random deviates with intercorrelations equal to the intermediate correlations. The random deviates are subsequently transformed by using the power constants from Table 5 and the polynomials of the form in (4a) and (4b) to generate the specified non-normal distributions with the specified correlations. The system is then generated using  $p = 4$  equations of the form in (3) with the  $\beta$  coefficients listed in Table 6 and the specified scalar terms  $\sigma$ .

## 6. SIMULATION STUDY

To evaluate the proposed procedure, the specified parameters  $\mu$ ,  $\sigma^2$ ,  $\gamma_1$ ,  $\gamma_2$ ,  $\gamma_3$ ,  $\gamma_4$  listed in Table 4 and all specified correlations from the numerical example were simulated using an algorithm coded in Fortran 77. The algorithm employed the use of subroutines NORMB1 and UNI1 from RANGEN (Blair, 1987) to generate pseudo-random normal and uniform deviates. Independent sample sizes of  $N = 10, 100, 1000$ , and  $10,000$  were generated for simulating the specified parameters and correlations. Values of all standardized cumulants for the independent variables and the stochastic disturbances were calculated for each repetition and then averaged across 50,000 repetitions. Thus, the average values of the specified parameters and correlations were based on  $N \times 50,000$  random deviates.

## 7. RESULTS

The overall average values of correlation and specified parameters computed in the simulation are reported below in Tables 9 and 10. The results reported are a subset of the entire set of averages computed for the system in (10). This subset consists of a representative sample of all the different kinds of averages computed for the full set. The purposes of reporting this subset are to confirm that the underlying derivations of the procedure are correct and to omit redundant reporting.

Inspection of Table 9 indicates that the procedure generated overall average correlations that were in close agreement with all specified correlations even for sample sizes as small as  $N = 10$ . These results also confirm the derivations underlying (7)–(9). Specifically, the specified correlations of:  $\text{Corr}(Y_2, X_{11}) = 0.164$ ;  $\text{Corr}(Y_1, Y_2) = 0.569$ ; and  $\text{Corr}(Y_1, \sigma_2 \varepsilon_2) = 0.579$  were obtained by evaluating (7)–(9). Inspection of Table 9 indicates excellent agreement between the specified and simulated correlations.

The proposed method also generated overall averages of standardized cumulants that were in close agreement with the parameters



**Simulation of Correlated Non-Normal Systems**

**Table 9.** Values of average correlation. Entries are based on  $N \times 50,000$  random deviates.

Variables	Specified correlation	Average correlation ( $N = 10$ )	Average correlation ( $N = 100$ )	Average correlation ( $N = 1000$ )	Average correlation ( $N = 10,000$ )
$X_{12}, X_{13}$	0.200	0.199	0.199	0.200	0.200
$X_{22}, X_{23}$	0.400	0.397	0.399	0.400	0.400
$X_{32}, X_{33}$	0.600	0.603	0.600	0.600	0.600
$X_{42}, X_{43}$	0.800	0.799	0.801	0.800	0.800
$X_{11}, X_{21}$	0.300	0.299	0.301	0.300	0.300
$X_{22}, X_{31}$	0.500	0.499	0.501	0.500	0.500
$X_{33}, X_{41}$	0.700	0.702	0.701	0.700	0.700
$X_{44}, X_{12}$	0.300	0.299	0.300	0.300	0.300
$X_{12}, X_{14}$	0.200	0.196	0.199	0.200	0.200
$X_{14}, X_{24}$	0.300	0.300	0.299	0.300	0.300
$X_{24}, X_{34}$	0.500	0.502	0.500	0.500	0.500
$X_{34}, X_{44}$	0.700	0.702	0.701	0.700	0.700
$Y_1, X_{11}$	0.200	0.199	0.200	0.200	0.200
$Y_1, X_{14}$	0.100	0.099	0.100	0.100	0.100
$Y_2, X_{21}$	0.400	0.402	0.399	0.400	0.400
$Y_2, X_{24}$	0.200	0.200	0.200	0.200	0.200
$Y_3, X_{31}$	0.600	0.601	0.601	0.600	0.600
$Y_3, X_{34}$	0.400	0.400	0.400	0.400	0.400
$Y_4, X_{41}$	0.800	0.801	0.799	0.800	0.800
$Y_4, X_{44}$	0.600	0.600	0.600	0.600	0.600
$\sigma_1 \varepsilon_1, \sigma_2 \varepsilon_2$	0.600	0.601	0.600	0.600	0.600
$\sigma_1 \varepsilon_1, \sigma_3 \varepsilon_3$	0.500	0.497	0.499	0.500	0.500
$\sigma_1 \varepsilon_1, \sigma_4 \varepsilon_4$	0.400	0.402	0.400	0.400	0.400
$Y_2, X_{11}$	0.164	0.165	0.164	0.164	0.164
$Y_1, Y_2$	0.569	0.572	0.570	0.569	0.569
$Y_1, \sigma_2 \varepsilon_2$	0.579	0.580	0.579	0.579	0.579

listed in Table 4. In terms of the independent variables, inspection of Table 10 indicates that the larger the sample sizes drawn, the closer the approximation was to the specified parameter. Similar results were obtained with respect to the disturbance terms.

**8. DISCUSSION**

The proposed method is useful for generating other systems of statistical equations based on the general linear model (GLM).



**Table 10.** Values of the average standardized cumulants. The entries are based on  $N \times 50,000$  random deviates. The specified parameters are listed in Table 4.

Chi-square distribution	$N$	$\hat{\mu}$	$\hat{\sigma}$	$\hat{\gamma}_1$	$\hat{\gamma}_2$	$\hat{\gamma}_3$	$\hat{\gamma}_4$
$X_{11}, 1(df)$	10	-0.00032	1.00440	2.87895	12.61254	75.67786	522.27230
$X_{21}, 2(df)$		0.00461	1.01152	2.05508	6.26645	25.13730	120.50190
$X_{31}, 3(df)$		-0.00059	1.00071	1.62701	3.93257	12.34966	46.06458
$X_{41}, 4(df)$		-0.00068	1.00090	1.40937	2.95257	8.00648	26.51598
$X_{11}, 1(df)$	100	-0.00027	1.00000	2.82779	11.98363	67.50917	473.47560
$X_{21}, 2(df)$		0.00050	1.00151	2.00750	6.03671	24.11410	118.81970
$X_{31}, 3(df)$		0.00018	0.99993	1.63635	4.02221	13.19285	53.83429
$X_{41}, 4(df)$		0.00020	0.99986	1.41692	3.01708	8.582005	30.40368
$X_{11}, 1(df)$	1000	0.00003	1.00046	2.83014	12.00356	67.78110	476.43590
$X_{21}, 2(df)$		0.00020	1.00048	2.00233	6.00923	24.01563	119.50420
$X_{31}, 3(df)$		0.00016	1.00000	1.63367	4.00102	13.04928	52.99338
$X_{41}, 4(df)$		0.00008	0.99999	1.41485	3.00103	8.47810	29.81137
$X_{11}, 1(df)$	10,000	0.00002	1.00000	2.82867	11.99921	67.77151	478.77361
$X_{21}, 2(df)$		0.00000	1.00001	1.99956	5.99476	23.99768	119.84500
$X_{31}, 3(df)$		0.00008	1.00000	1.63278	3.99875	13.03505	53.23197
$X_{41}, 4(df)$		-0.00002	1.00000	1.41404	2.99837	8.46670	29.86888

For example, the method could be used to generate  $T$  independent equations to investigate the statistical properties of competing nonparametric tests in the context of analysis of covariance (ANCOVA) or repeated measures.

With respect to ANCOVA, one attractive feature of the proposed method is that it has an advantage over other competing algorithms to the extent that it allows for the creation of distributions with unequal regression slopes while maintaining equal variances. This can be demonstrated by inspecting Eq. (3) where the slope coefficient(s) could change (i.e., made unequal) while the error terms remain unchanged. Subsequent to any changes made to the slope coefficients, the variate and covariate correlations can be determined from Eq. (6). This feature of the proposed method would be a remedy to the problem with the algorithm used in the Monte Carlo ANCOVA study by Hamilton (1976). Specifically, Hamilton (1976) used the Knapp and Sowyer (1967) algorithm to generate correlated data. This algorithm is restrictive in



the context of ANCOVA because it cannot simulate an unequal slope condition without simultaneously violating the between-group equal variance assumption. See Rogosa (1980) for a discussion on the validity of the Hamilton (1976) study.

Many other applications of the proposed procedure to the GLM are possible. From the GLM perspective, the dependent variables  $Y_p$  could represent the same variable collected under  $T$  different conditions or at time points  $1, \dots, T$ . In either case, the independent variables  $X_{pi}$  could represent static covariates (e.g., pre-existing ability measures often used in ANCOVA models). On the other hand, the independent variables  $X_{pi}$  could also be different for each of the  $T$  equations and may be used to represent time-varying covariates (i.e., some variables measured over the  $T$  periods along with  $Y_p$ ).

In terms of repeated measures, the method could also be used to allow the generation of repeated measures data of nonspherical structures with non-normal disturbances and non-normal covariates. It should also be noted that with non-normal stochastic distributions one could specify all of these distributions to be the same. This assumption is implicit in parametric analyses of repeated measures data because normal disturbance populations are implied.

The method could also be applied in the context of time series analysis. Specifically, the procedure could be used to model instrumental variables which address one of the problems that certain autoregressive models (e.g., the adaptive expectations model) have where the dependent measure from a preceding time period ( $Y_p$ ) is included as an independent variable in the subsequent ( $q$ th) period. As such, the vectors  $Y_p$  and  $\varepsilon_q$  are usually correlated. Using the proposed method, a Monte Carlo study could be arranged to simulate non-normal “proxies” correlated at various levels between  $Y_p$  and  $\varepsilon_q$ .

The application of the proposed method to the GLM is flexible and has the potential to simulate other types of models where the stochastic disturbance distributions may change over time. For example, data sets with repeated measures often have mistimed measures or missing data. Thus, the proposed method could be used to compare and contrast the GLM with generalized estimating equations or hierarchical linear models – procedures that are often considered preferable to standard univariate or multivariate OLS procedures.

It may also be reasonable to consider a model where the correlation structure is a function of time between the observations. Thus, data could also be generated using the proposed procedure for Monte Carlo studies involving dynamic regression models that have distributed lags or moving averages.



## 9. CONCLUSION

Systems of linear statistical equations with correlated non-normal variables are widely applicable in many experimental or non-experimental settings. Some examples include confirmatory factor analysis, hierarchical linear models, time series analysis, and other applications of the general linear model (e.g., analysis of covariance, repeated measures). The present study develops a method for simulating systems of statistical equations with non-normal variables and with specified correlations between the (a) stochastic disturbance distributions, (b) independent variables, and (c) dependent and independent variables for each equation in the system. *Mathematica* (Wolfram, 1999) notebooks are made available for implementing the procedure. The notebooks solve for (a) power constants, (b) intermediate correlations, and (c) coefficients for each equation in the system. The results of a Monte Carlo simulation confirm that the procedure generates the specified parameters and correlation structures.

## REFERENCES

- Blair, R. C. (1987). *RANGEN*. Boca Raton, FL: IBM.
- Cook, R. D., Weisberg, S. (1999). *Applied Regression Including Computing and Graphics*. New York: John Wiley.
- Dwivedi, T., Srivastava, K. (1978). Optimality of least squares in the seemingly unrelated regression model. *J. Economics* 7:391–395.
- Gilks, W. R., Richardson, S., Spiegelhalter, D. J. (1998). *Markov Chain Monte Carlo in Practice*. Boca Raton, FL: Chapman & Hall/CRC.
- Hamilton, B. L. (1976). A Monte Carlo test of the robustness of parametric and non-parametric analysis of covariance against unequal regression slopes. *J. Amer. Statist. Assoc.* 71:864–869.
- Headrick, T. C. (2002). Fast fifth-order polynomial transforms for generating univariate and multivariate non-normal distributions. *Comput. Statist. Data Anal.* 40:685–711.
- Headrick, T. C., Rotou, O. (2001). An investigation of the rank transformation in multiple regression. *Comput. Statist. Data Anal.* 38:203–215.
- Headrick, T. C., Sawilowsky, S. S. (2000). Properties of the rank transformation in factorial analysis of covariance. *Comm. Statist. Simulation Comput.* 29:1059–1087.
- Hesterberg, T. (2001). Bootstrap tilting. In: The University of Florida Statistics Symposium on Monte Carlo in the New Millennium. Gainesville, FL.



- Holgerson, H. E. T., Shuker, G. (2001). Some aspects of non-normality tests in systems of regression equations. *Comm. Statist. Simulation Comput.* 30:291–310.
- Judge, G. J., Hill, R. C., Griffiths, W. E., Lutkepohl, H., Lee, T. C. (1985). *Introduction to the Theory and Practice of Econometrics*. 2nd ed. New York: John Wiley.
- Knapp, T. R., Swoyer, V. H. (1967). Some empirical results concerning the power of Bartlett's test of the significance of a correlation matrix. *Am. Educational Res. J.* 4:13–17.
- Mehta, C. R., Patel, N. R., Senchaudhuri, P. (2000). Efficient Monte Carlo methods for conditional logistic regression. *J. Am. Statist. Assoc.* 95:99–108.
- Neter, J., Kutner, M. H., Nachtsheim, C. J., Wasserman, W. (1996). *Applied Linear Statistical Models*. 4th ed. Boston, MA: WCB/McGraw-Hill.
- Rogosa, D. (1980). Comparing nonparallel regression lines. *Psychol. Bull.* 88:307–321.
- Thompson, E. A. (2000). MCMC estimation of multi-locus genome sharing and multipoint gene location scores. *Int. Statist. Rev.* 68:53–73.
- Wolfram, S. (1999). *The Mathematica Book*. 4th ed. Wolfram Media/Cambridge University Press.
- Zellner, A. (1962). An efficient method of estimating seemingly unrelated regressions and tests of aggregation bias. *J. Am. Statist. Assoc.* 57:348–368.



## **Request Permission or Order Reprints Instantly!**

Interested in copying and sharing this article? In most cases, U.S. Copyright Law requires that you get permission from the article's rightsholder before using copyrighted content.

All information and materials found in this article, including but not limited to text, trademarks, patents, logos, graphics and images (the "Materials"), are the copyrighted works and other forms of intellectual property of Marcel Dekker, Inc., or its licensors. All rights not expressly granted are reserved.

Get permission to lawfully reproduce and distribute the Materials or order reprints quickly and painlessly. Simply click on the "Request Permission/Order Reprints" link below and follow the instructions. Visit the [U.S. Copyright Office](#) for information on Fair Use limitations of U.S. copyright law. Please refer to The Association of American Publishers' (AAP) website for guidelines on [Fair Use in the Classroom](#).

The Materials are for your personal use only and cannot be reformatted, reposted, resold or distributed by electronic means or otherwise without permission from Marcel Dekker, Inc. Marcel Dekker, Inc. grants you the limited right to display the Materials only on your personal computer or personal wireless device, and to copy and download single copies of such Materials provided that any copyright, trademark or other notice appearing on such Materials is also retained by, displayed, copied or downloaded as part of the Materials and is not removed or obscured, and provided you do not edit, modify, alter or enhance the Materials. Please refer to our [Website User Agreement](#) for more details.

### **[Request Permission/Order Reprints](#)**

Reprints of this article can also be ordered at

<http://www.dekker.com/servlet/product/DOI/101081SAC120028431>