An Efficient and Theoretically Consistent Procedure for Generating Correlated, Non-Normal Random Variables in Simulation Models.

By

O. A. Ramirez

W. G. Boggess

C. B. Moss*

February 18, 1990

*W. G. Boggess, O. A. Ramirez, and C. B. Moss are Professor, Graduate Assistant and Assistant Professor, respectively, in the Department of Food and Resource Economics at the University of Florida, Gainesville, 32611.

An Efficient and Theoretically Consistent Procedure for Generating Correlated, Non-Normal Random Variables in Simulation Models

Abstract

In recent years, simulation has become an important methodology for applied decision analysis under uncertainty. A typical simulation effort requires generating a set of possibly correlated and non-normal random variables, using information regarding their underlying joint probability density function contained in a presumably random sample. A few techniques have been suggested to accomplish this task, but most have not met the requirement of being correct and efficient from the statistical point of view. This study proposes a multivariate hyperbolic sine probability density function as a basis to develop an efficient and theoretically consistent approach for generating correlated, non-normal random variables.

Keywords: Simulation, non-normal, density function, random sample, efficient, multivariate, hyperbolic sine.

#8450

An Efficient and Theoretically Consistent Procedure for Generating Correlated, Non-Normal Random Variables in Simulation Models

Introduction

In recent years, simulation has become an important methodology for applied decision analysis under uncertainty. Anderson provides a review of early simulation methods and applications in agricultural economics. He identifies the failure to take into account correlation among random variables as a major shortcoming of stochastic simulation models. A typical simulation effort requires generating a set of possibly correlated and non-normal random variables, using information regarding their underlying joint probability density function contained in a presumably random sample. Furthermore, realizations of these random variables are usually observed through time, so that the trend component of the sample has to be taken into account before attempting to make any inferences regarding the variances, covariances and higher order moments of such random variables. Until recently, few techniques had been suggested to accomplish this task (i.e. Clements, et al.; Richardson and Condra; King), but none have met the requirement of being correct and efficient from the statistical point of view (Fackler and King). In addition, research aimed at developing such techniques had not focused on the basic problem, which as Johnson recently pointed out is the restrictiveness of available multivariate probability density functions. Taylor (1990), proposes two procedures for empirically fitting multivariate nonnormal probability density functions. As he points out, both procedures are small sample alternatives to assuming a particular

theoretical distribution for empirical analysis. This study proposes the multivariate hyperbolic sine probability density function as a basis to develop a simple, efficient and theoretically consistent approach to simulation analysis.

The multivariate hyperbolic sine density function

The inverse hyperbolic sine transformation has received increased attention within the field of econometric modeling. Ramirez (1990), formally defines a "hyperbolic sine" random variable y_i as follows:

$$(1) \qquad y_i = \sinh(\theta_i(g_i))/\theta_i + c_i = (e^{\theta_i g_i} - e^{-\theta_i g_i})/2\theta_i + c_i$$

where g_i is a normal random variable with mean μ_i and variance σ_i^2.

The marginal density function associated with the i^{th} hyperbolic sine (HS) random variable can be easily derived using the transformation technique (see Mood et al.):

$$(2) \qquad (2\pi\sigma_i^2)^{-1/2} \exp^{-.5(w_i-\mu_i)^2\sigma_i^{-2}} (1+(\theta_i(y_i-c_i))^2)^{-1/2}$$

Furthermore, a multivariate hyperbolic sine density function can also be derived by applying the transformation technique to a multivariate normal density function:

$$(3) \qquad f_y=(2\pi)^{-P/2} |\Sigma|^{-1/2}\exp\{-.5(w-\mu)'\Sigma^{-1}(w-\mu)\} \prod_{i=1}^{P} (1+(\theta_i(y_i-c_i))^2)^{-1/2}$$

where y is a P by 1 vector of hyperbolic sine random variables; w is a P by 1 vector with i^{th} element $w_i = \theta_i^{-1} \sinh^{-1}(\theta_i(y_i-c_i)) = \ln(\theta_i(y_i-c_i)+(1+(\theta_i(y_i-c_i))^2)^{1/2})/\theta_i$; μ is a P by 1 vector with i^{th} element μ_i; and Σ is a P by P symmetric positive semi-definite matrix of parameters.

A hyperbolic sine random variable exhibits several important characteristics. First, normality is a special case of a HS random variable.

As θ_i goes to zero, the i^{th} random variable is accounted for as a normal random variable in the multivariate density function equation (3). In the limit, as all θ_i's go to zero, equation (3) is a multivariate normal density function.

Second, a HS random variable can exhibit any positive or negative expected value and any variance independently of the skewness and kurtosis coefficients. This feature is not shared by the most commonly used random variables (i.e. normal, lognormal, exponential, beta, gamma, chi-square, F). The HS random variable can also exhibit any positive or negative degree of skewness. If the skewness coefficient is zero, the kurtosis coefficient can take on any positive value (fat tails). If the skewness coefficient is not zero, there is a wide range of allowable combinations of skewness and kurtosis.

Third, knowledge of the multivariate HS density function allows joint estimation of the parameters entering the individual marginal density functions as well as the covariances among the random variables under consideration. Since the functional forms of the multivariate density functions associated with the non-normal[1] random variables commonly used in simulation analysis are not known, full information maximum likelihood estimation has not been possible in the past.

The only restriction on the degree of correlation allowed by the HS multivariate density function, results from the requirement that Σ be positive semi-definite. This characteristic is not a limitation of this specific multivariate density function; it merely reflects the theoretical

---

[1]Except for the case of the lognormal multivariate density which is very restrictive in terms of the combination of moments and the degree of negative correlation allowed.

limit to the degree of positive or negative correlation between random variables with non-symmetric and possibly different marginal density functions. The multivariate HS density function allows for correlation coefficients ranging from -1 to 1 if such values are consistent with the nature of the stochastic processes underlying the random variables of interest.

Fourth, the null hypothesis of no skewness for the i^{th} random variable can be specified as $\mu_i = 0$; and the null hypothesis of normality (no skewness and no kurtosis) can be specified as $\theta_i = 0$ (Ramirez, 1990). None of the other density functions commonly used in simulation analysis allow direct testing for normality and skewness.

Finally, a hyperbolic sine random variable can be easily transformed into a normal random variable and vice versa. A normal random variable can be transformed into a HS random variable as follows:

$$(4) \qquad y_i = \frac{\exp^{\theta_i Z_i} - \exp^{-\theta_i Z_i}}{2\theta_i} + c_i$$

where Z_i is a normal random variable with mean μ_i and variance σ_i^2. A HS random variable can be transformed into a normal random variable as follows:

$$(5) \qquad Z_i = \sinh^{-1}(\theta_i(y_i - c_i))/\theta_i$$

This property is of paramount importance within the framework of simulation analysis because it allows correlated HS random variables to be generated by applying a simple transformation to correlated normal random variables.

Another important advantage of this technique is that, if time series data is to be used, a trend and an intercept can be estimated jointly with the rest of the parameters entering the individual marginal density

functions as well as the covariances among the random variables under consideration by setting

(6) $c_i = b_{oi} + t\ b_{1i}$; i=1,...,P;

where t is a trend vector.

This is an appropriate specification since it can be shown that (Ramirez, 1990):

(7) $E[\ y_i\ ] = K_i + t\ b_{1i}$

where K_i is a constant that depends on b_{oi}, θ_i, σ_i^2, and μ_i. The parameters θ_i, σ_i^2, and μ_i control the variance and higher order moments of y_i (Ramirez, 1990), that are assumed to remain constant over time, while b_{oi} and $t\ b_{1i}$ control the mean of y_i which is assumed to change over time according to the process $K_i + t\ b_{1i}$.

Therefore, a genuine full information effort can be attempted using the technique proposed in this study. The costumary approach of first "removing the trend", and then attempting to make some inferences regarding the stochastic properties of the random variables under consideration based on ordinary least squares residuals, is not very attractive, theoretically. It is widely known that the statistical properties of the ordinary least squares residuals are quite different from those of the theoretical errors (deviations from the trend), even under the assumption of normality (Judge et. al.). If these residuals are hypothesized to be nonnormal, simple detrending using ordinary least squares is even more problematic.

Using the multivariate HS density function for simulation

Assume that there is a sample of T observations on P random variables available to the researcher. There is no a priori information regarding the properties of the stochastic process underlying this set of

random variables, so that they could be non-normally distributed and correlated with each other. Furthermore, the mean value of those random variables changes over time according to a process such as $K_i + t\ b_{1i}$; $i = 1,\ldots,P$. The researcher wants to simulate an S by P matrix of realizations of the random variables under consideration, using the information contained in the T by P sample matrix in an efficient and theoretically consistent fashion. The following procedure is suggested:

Step 1. Maximization of the individual likelihood function for each of the random variables under consideration. The likelihood function for the i^{th} HS random variable is proportional to

$$(8) \qquad -T/2\ln(\sigma_i^2)-.5\ln\left(\prod_{t=1}^{T}(1+(\theta_i(y_{it}-c_i)))\right)-.5\sum_{t=1}^{T}(w_{it}-\mu_i)^2\sigma_i^{-2}$$

where $w_{it} = \sinh^{-1}(\theta_i(y_{it}-c_i))/\theta_i$.

In some cases, the nature of the stochastic process underlying the i^{th} random variable can be completely specified without the presence of μ_i. This would imply, as previously mentioned, that the probability density function associated with the i^{th} random variable exhibits no skewness (although if $\theta_i \neq 0$ it has "fat tails"). A null hypothesis of no skewness ($\mu_i = 0$), therefore, can be tested using either a likelihood ratio or an asymptotic t-test.

At this point, a test of the null hypothesis of normality for each of the random variables under consideration is also recommended. A likelihood ratio test can be conducted by comparing twice the difference between the maximum value of (8) and the maximum value of the log-likelihood function for $y_i - c_i$ under normality with a chi-square random variable with two degrees of freedom (Burbidge, 1986). Alternatively, performing t-tests for the null hypotheses $\theta_i = 0$ and $\mu_i = 0$, using the estimates of the

asymptotic standard errors provided by most optimization packages can also be interpreted as an asymptotic test for normality. If the null hypothesis of normality is not rejected, the i^{th} random variable can be treated as a normal random variable throughout the rest of the simulation process. This allows use of all available sample information to jointly estimate all of the covariances and other parameters of interest.

Step 2. After limited information estimates are available for all the parameters entering the marginal density functions, there are two alternative theoretically consistent ways to continue.

Step 2a. Using the limited information parameter estimates of θ_i, and c_i (i.e $b_{oi} + t\ b_{1i}$) apply the following transformation to the data on each of the non-normal random variables under consideration

$$Z_i = \sinh^{-1}(\theta_i(y_i - c_i))/\theta_i \tag{9}$$

where Z_i will be a T by 1 vector consistent with a normal sampling process with mean μ_i and variance σ_i^2. Notice that the raw data on the normal random variables will also have to be detrended before it can be used for computing estimates of the population covariances.

Once the sample has been "transformed to normality" the customary consistent (maximum likelihood) estimator for the covariance parameter when sampling from normal random variables can be applied:

$$\Sigma_{ij} = \sum_{t=1}^{T} ((Z_{it} - \mu_i)(Z_{jt} - \mu_j)) \ / \ T \ ; \ i \neq j. \tag{10}$$

An estimate of the covariance matrix Σ can now be constructed using the σ_i^2's as the diagonal elements and the corresponding Σ_{ij} as the off-diagonal elements. Notice that Σ is not the covariance matrix associated with the original set of random variables but the covariance matrix

associated with the underlying set of normal random variables Z_i (i = 1, 2, ..., P). The matrix Σ, however, contains the parameters that control the variances and covariances associated with the original set of possibly non-normal random variables.

Step 2b. Using the limited information estimates from step 1 as starting values, attempt to maximize the full log-likelihood function which is proportional to

$$(11) \quad -T/2\text{Ln}(\text{abs}(|\Sigma|)) - .5\text{Ln}\left(\prod_{t=1}^{T}\prod_{i=1}^{K}(1+(\theta_i(y_{it}-c_i))^2)\right) - .5\sum_{t=1}^{T}((w_t-\mu)'\Sigma^{-1}(w_t-\mu))$$

where $w_{it} = \sinh^{-1}(\theta_i(y_{it}-c_i))/\theta_i$; i = 1, 2, ... K for the K HS random variables and $w_{it} = y_{it}-c_i$, $\mu_i = 0$; i = K+1, ... P for the P-K normal random variables. Σ is a P by P symmetric positive semi-definite matrix containing the parameters that control the variances and covariances of the random variables under consideration.

Step 3. Given the estimates of the θ_i's, the c_i's and Σ from either step 2a or 2b, simulating the random variables of interest is a straightforward process. First, generate an S by P matrix of standard normal random variables. Second, transform the matrix of standard normal random variables to a matrix of correlated normal random variables with a mean vector $\mu = [\mu_1, \mu_2 \ldots \mu_p]'$ and variance vector $\sigma^2 = [\sigma_1^2, \sigma_2^2, \ldots \sigma_p^2]'$ using the Cholesky decomposition of the estimate of Σ and the estimates of the μ_i's. Finally, transform the columns of the resulting matrix that correspond to the random variables that rejected the null hypothesis of normality using the appropriate θ_i's and c_i's and equation (4).

Furthermore, maximum likelihood estimates of the expected values, variances, third and fourth central moments, as well as the covariance coefficients associated with the original random variables can be easily computed given the maximum likelihood estimates of the parameters entering the joint hyperbolic sine probability density function (Ramirez, 1990).

Application: Simulating U.S. average crop yields of corn, soybeans and wheat

U.S. average crop yields for corn, soybeans and wheat are good candidates for illustrating the proposed technique. The yields are expected to be correlated, have an increasing expected value over time, and there is no reason to believe that the underlying stochastic processes conform to normality.

Define y_1, y_2, and y_3 to be U.S. average yields for corn, soybeans and wheat respectively, from the year 1950 to 1989. Using equation (8), and the maximum likelihood algorithm available within the matrix algebra programming language GAUSS 2.0, limited information parameter estimates for θ_i, μ_i, σ_i^2, b_{oi}, and b_{1i} ($i = 1, 2, 3$) were obtained (see table 1).

The maximum likelihood parameter estimate for θ_1 equals 0.2259, and the estimate of its asymptotic standard error is only 0.1020. This suggests that the probability density function associated with the first random variable exhibits a significant departure from normality. The maximum likelihood parameter estimate for μ_1 equals -4.1177, and the estimate of its asymptotic standard error is only 2.1297. This suggests that the probability density function associated with the first random variable also exhibits a significant degree of skewness to the left side. In order to further explore those issues, restricted models (first setting $\mu_1 = 0$, and then setting $\theta_1 = \mu_1 = 0$) are estimated and sequential likelihood ratio tests, as proposed in

the previous section, are conducted (see table 2). Within this framework, the null hypothesis of no skewness is first rejected. Furthermore, the joint null hypothesis of no skewness and normality is also rejected. The probability density function associated with the first random variable is shown to depart from normality, with a very reasonable degree of statistical certainty. It is also shown that such density function exhibits a considerable degree of left skewness.

On the other hand, notice that the maximum likelihood parameter estimates for θ_2 and μ_2 are both equal to zero. That is, equation (8) asymptotically approaches its maximum as θ_2 and μ_2 go to zero (i.e equation (8) goes to a normal density function with mean $b_{o2} + t\ b_{12}$ and variance σ_2^2). The stochastic process underlying y_2 is obviously normal.

In addition, notice that even though the maximum likelihood parameter estimate for θ_3 equals 0.4027, the estimate of its associated asymptotic standard error is relatively high. The estimate of the parameter μ_3, controlling the degree of skewness associated with the third random variable, exhibits this same characteristic. Restricted models (first setting $\mu_3 = 0$, and then setting $\theta_3 = \mu_3 = 0$) are therefore estimated. Sequential likelihood ratio tests, as proposed in the previous section, first fail to reject the null hypothesis of no skewness, an then that of normality (see table 2). The probability density function associated with the third random variable can not be shown to depart from normality, with a reasonable degree of statistical certainty.

The second step involves estimation of the parameters controlling the covariances between the three random variables under consideration. If the procedure described in 2a is used, the resulting estimators are asymptotically less efficient than the full information maximum likelihood

estimators that can be obtained using the procedure outlined in 2b. For illustrative purposes, estimates resulting from application of both procedures are presented in table 3 and 4.

Given either the limited or full information maximum likelihood estimates of all the parameters necessary to describe the underlying stochastic processes; average yields for corn, soybeans and wheat can be easily simulated following step 3. Furthermore, maximum likelihood estimates of the expected values, variances, third and fourth central moments, as well as the covariance coefficients associated with the original random variables can be easily computed given the maximum likelihood estimates of the parameters entering the joint hyperbolic sine probability density function (Ramirez, 1990).

References

Anderson, J. R. "Simulation: Methodology and Application in Agricultural Economics." Review of Marketing and Agricultural Economics. 42(1974):3-55.

Clements, A. M. Jr., H. P. Mapp Jr., and V. R. Eidman. "A Procedure for Correlating Events in Farm Firm Simulation Models." Oklahoma State University Agr. Exp. Sta. Bull. No. T-131, Aug. 1971.

D'Agostino, R.B. "An Omnibus Test of Normality for Moderate and Large Sample Sizes." Biometrika 58(1971):341-8.

Fackler, P. L. and R. P. King. "Generation of Dependent Random Variates with Given Marginal Distributions and Rank Correlation Structure." Mimeograph. North Carolina State University. Raleigh, N.C. 1988.

Johnson, M. E. Multivariate Statistical Simulation. New York. John Wiley and Sons, 1987.

King, Robert P. "Operational Techniques for Applied Decision Analysis Under Uncertainty." Unpublished dissertation. University of Minnesota. St. Paul, MN. 1979.

Ramirez, O. A. and J. S. Shonkwiler. "Statistical Properties of the Inverse Hyperbolic Sine Random Variable and its Multivariate Equivalent." Mimeograph. University of Florida, Gainesville, Florida. 1989.

Richardson, J. W. and G. D. Condra. "A General Procedure for Correlating Events in Simulation Models." Mimeograph. Texas Agri. Exp. Sta., Department of Agricultural Economics. College Station, TX. 1978.

Table 1. Limited information maximum likelihood parameter estimates.

Variable	Parameter θ	μ	σ^2	b_o	b_1	ωmax[a]
Y_1	0.2259[b]	-4.1177	15.4694	35.2728	2.3015	-100.2678
	(0.1020)	(2.1297)	(7.4609)	(2.9333)	(0.0723)	
Y_2	0.0000	0.0000	3.5982	19.8477	0.2962	-46.7491
	(0.4825)	(17.9206)	(0.7948)	(17.9108)	(0.0250)	
Y_3	0.4027	0.2998	3.2542	15.8734	0.5591	-53.9469
	(0.3201)	(0.8087)	(1.6126)	(0.8894)	(0.0347)	

[a] ωmax is the value of the likelihood function at the optimum.
[b] The estimates of the asymptotic standard errors are given in parenthesis.

Table 2. Statistics for the sequential likelihood ratio tests.

Variable	$\alpha^0 max$	$\alpha^1 max$	$x^2_{(1)}$	$\alpha^2 max$	$x^2_{(2)}$
Y_1	-100.2678	-103.8444	7.1532*	-109.2631	17.9906**
Y_3	-53.9469	-54.0237	0.1536	-54.4381	0.9824

*Rejects the null hypothesis of no skewness at the 5% level of statistical confidence.
**Rejects the null hypothesis of normality (no skewness and no kurtosis) at the 5% level of statistical confidence.

Table 3. Full information maximum likelihood parameter estimates.

	Parameter				
Variable	θ	μ	σ^2	b_o	b_1
Y_1	0.1072	-5.4398	30.5154	36.9473	2.2258
	(0.0526)	(4.6112)	(13.8374)	(5.5351)	(0.0885)
Y_2	-	-	3.6384	19.4909	0.3132
	-	-	(0.8128)	(0.5850)	(0.0239)
Y_3	-	-	5.2444	16.6885	0.5381
	-	-	(1.1609)	(0.7299)	(0.0303)

Table 4. Covariance matrices from limited information maximum likelihood (2a) and full information maximum likelihood (2b).

	Σ_{2a}			Σ_{2b}		
Y_1	15.4694	4.8570	2.5047	30.5154	7.6726	3.5659
Y_2	-	3.5983	.7819	-	3.6384	0.8006
Y_3	-	-	5.2358	-	-	5.2444