Journal of Data Science **12**(2014), 217-234

Generating correlated random vector by Johnson system

Qing Xiao
Shanghai University

Abstract:

This paper aims to generate multivariate random vector with prescribed correlation matrix by Johnson system. The probability weighted moment (PWM) is employed to assess the parameters of Johnson system. By equating the first four PWMs of Johnson system with those of the target distribution, a system of equations solved for the parameters is established. With suitable initial values, solutions to the equations are obtained by the Newton iteration procedure. To allow for the generation of random vector with prescribed correlation matrix, approaches to accommodate the dependency are put forward. For the four transformation models of Johnson system, nine cases are addressed. Analytical formulae are derived to determine the equivalent correlation coefficient in the standard normal space for six cases, the rest three ones are handled by an interpolation method. Finally, several numerical examples are given out to check the proposed method.

Key words: correlation coefficient, Johnson system, normal transformation, probability weighted moment.

1. Introduction

A general problem in stochastic simulation modeling entails representing data with the underlying distribution unknown, and generating random vector with desired correlation matrix and marginal distributions. To achieve this goal, several empirical distributions have been proposed, such as the generalized lambda distribution (GLD) (Ramberg & Schmeiser, 1974), the power transformation (Fleishman, 1978), and the Johnson system (Johnson, 1949). The related works have been summarized by Tadikamalla (1980). Among these families of distributions, the Johnson system stands out for its generality to simulate various probability distributions and the potential ability to model the correlated random variables.

The Johnson system comprises four models:

$$\begin{aligned} S_N: &\quad z = \frac{x-\mu_z}{\sigma_z} \\ S_L: &\quad z = \gamma + \delta \cdot ln\,(x-\xi) \quad (\xi < x) \\ S_U: &\quad z = \gamma + \delta \cdot sinh^{-1}\left(\frac{x-\xi}{\lambda}\right) \\ S_B: &\quad z = \gamma + \delta \cdot ln\left(\frac{x-\xi}{\xi+\lambda-x}\right) \quad (\xi < x < \xi+\lambda) \end{aligned} \tag{1}$$

where z is a standard normal variable. γ and δ are parameters determined by the skewness and kurtosis, ξ and λ are parameters determined by the mean and the standard deviation respectively. S_N is the normal distribution with mean μ_z and standard deviation σ_z. S_L represents the Lognormal family. S_U denotes the unbounded family, that is, the variation range of x is infinite. S_B is the bounded family, handling x with a boundary. These four models together cover all types of unimodal probability distributions.

To utilize the Johnson system to fit data or simulate distribution, it is necessary to select an appropriate model and to evaluate the parameters. Based on the skewness and kurtosis, a suitable model for the target random variable can be determined. The central problem is to estimate the parameters. The standardized central moment matching method (Hill et al., 1976) and the percentile method (Slifker & Shapiro, 1980) are examples of various approaches developed for this purpose. This paper focuses on the moment matching method.

Since the statistical information of a random variable is characterized by its statistical moments, the underlying principle for the moment matching method is to relate the moments of the corresponding transformation model to those of the target random variable. The parameters of Johnson system would be obtained by solving the system of nonlinear equations established. In general, the central moment or raw moment is involved, leading to complicated equations. Recently, L-moments have been widely used as good alternatives to conventional moments, which are implemented to characterize the GLD (Karvanen & Nuutinen, 2008) and the power transformation (Headrick, 2011), and much simpler procedure for parameter estimation is achieved. Since L-moments are defined as linear combinations of probability weighted moment (PWM) (Hosking, 1990), it is worthwhile to attempt to perform the moment matching in terms of PWM directly, which may further simplify the problem.

On the issue of simulating correlated random vector, relatively little research has been conducted, which may be due to the tedious and difficult computations for evaluating the correlation coefficient in the standard normal space. When modeling correlated multivariate distributions with samples available, the John-

son system is workable. By normalizing each sample value via the associated model in Eq.(1), the correlation matrix in the standard normal space can be calculated directly. However, random vectors with prescribed correlation matrix cannot be generated in this way. To circumvent the computational complexity, Stanfield et al. (1996) develop an alternative approach. Yet, inadequacies are found both theoretically and practically. This method fails to match the kurtosis for some distributions and is unfeasible for a near singular correlation matrix.

In this paper, PWM is employed to characterize the Johnson system. By setting the first four PWMs of Johnson system equal to those of the simulated random variable, the system of equations solved for the parameters is established. With suitable initial values, the solutions are obtained by Newton iteration method. Furthermore, approaches to accommodate the dependency are also put forward. For cases related to S_N, S_L, S_U and case of $S_N - S_B$, six analytical formulae are derived to evaluate the equivalent correlation coefficient in the standard normal space. The other three cases involving $S_L - S_B$, $S_U - S_B$ and $S_B - S_B$ are handled by an interpolation method.

2. Model selection

With the Lognormal distribution being the boundary, the skewness and kurtosis of a random variable are used for the model selection. Let κ_3, κ_4 denote the skewness and kurtosis respectively. For normal distribution S_N, it follows that $\kappa_3 = 0$ and $\kappa_4 = 3$. In the case of Lognormal distribution, the following relationships are satisfied (Johnson, 1949):

$$\kappa_3^2 = (\omega - 1)(\omega + 2)^2 \tag{2}$$

$$\kappa_4 = \omega^4 + 2\omega^3 + 3\omega^2 - 3 \quad (\omega = e^{\sigma^2}) \tag{3}$$

Suppose x is a non-normal random variable, solving the equation in Eq.(2) for $\omega > 0$, the model is chosen as follows:

$$\begin{aligned} S_L: &\quad \kappa_4 = \omega^4 + 2\omega^3 + 3\omega^2 - 3 \\ S_U: &\quad \kappa_4 > \omega^4 + 2\omega^3 + 3\omega^2 - 3 \\ S_B: &\quad \kappa_4 < \omega^4 + 2\omega^3 + 3\omega^2 - 3 \end{aligned} \tag{4}$$

Since the S_L family is actually the Lognormal distribution, the parameters of S_L

model can be easily obtained.

$$\begin{aligned} \delta &= (ln\omega)^{-\frac{1}{2}} \\ \gamma &= \frac{1}{2}\delta \cdot ln\left(\frac{\omega^2-\omega}{\sigma^2}\right) \\ \xi &= \mu - exp\left(\frac{1-2\delta\gamma}{2\delta^2}\right) \end{aligned} \tag{5}$$

where μ is the mean of x, σ is the standard deviation of x.

3. Evaluating the parameters of S_U and S_B models

3.1 Probability weighted moment

The PWM of a random variable x is defined as (Greenwood et al., 1979):

$$M_{p,r,s} = E\{x^p \cdot [F(x)]^r \cdot [1-F(x)]^s\} \tag{6}$$

where $F(\cdot)$ is the cumulative distribution function (CDF) of x. A particular type of PWM, $\beta_r = M_{1,r,0}$, is considered:

$$\beta_r = \int_{-\infty}^{+\infty} x \cdot F^r(x) \cdot f(x)dx \tag{7}$$

where $f(\cdot)$ is the probability density function (PDF) of x.

For the sake of computational convenience, Johnson system is rewritten in an inverse form:

$$\begin{aligned} S_N: &\quad x = \mu_z + \sigma_z z \\ S_L: &\quad x = \xi + exp\left(\frac{z-\gamma}{\delta}\right) \quad (\xi < x) \\ S_U: &\quad x = \xi + \lambda sinh\left(\frac{z-\gamma}{\delta}\right) \\ S_B: &\quad x = \xi + \lambda - \frac{\lambda}{1+exp\left(\frac{z-\gamma}{\delta}\right)} \quad (\xi < x < \xi+\lambda) \end{aligned} \tag{8}$$

where μ_z and σ_z are the mean and the standard deviation of the normal variable respectively. This paper pays attention to evaluate the parameters of S_U and S_B models.

3.2 S_U model

The β_r of the S_U model is:

$$\begin{aligned}\beta_r &= \int_{-\infty}^{+\infty}\left[\xi+\lambda sinh\left(\frac{z-\gamma}{\delta}\right)\right]\cdot\Phi^r(z)\cdot\varphi(z)dz \\ &= \frac{\xi}{1+r}+\lambda\int_{-\infty}^{+\infty} sinh\left(\frac{z-\gamma}{\delta}\right)\cdot\Phi^r(z)\cdot\varphi(z)dz\end{aligned} \tag{9}$$

where $\Phi(\cdot)$ is the CDF of the standard normal random variable, $\varphi(\cdot)$ is the PDF.

Equating the first four PWMs with those of the target random variable x, four equations with four unknowns (ξ, λ, γ, δ) are established. Denote the equations as follows:

$$G_r(\xi,\lambda,\gamma,\delta)-\beta_r=0 \qquad (r=0,1,2,3) \tag{10}$$

where

$$G_r(\xi,\lambda,\gamma,\delta)=\frac{\xi}{1+r}+\lambda\int_{-\infty}^{+\infty} sinh\left(\frac{z-\gamma}{\delta}\right)\cdot\Phi^r(z)\cdot\varphi(z)dz \tag{11}$$

Solving the system of equations yields the parameters of the S_U model. With proper initial values, the solutions can be found by the Newton's iteration method. The Jacobian matrix is:

$$J=\frac{\partial(G_0,G_1,G_2,G_3)}{\partial(\xi,\lambda,\gamma,\delta)}=\begin{pmatrix}1 & \frac{\partial G_0}{\partial\lambda} & \frac{\partial G_0}{\partial\gamma} & \frac{\partial G_0}{\partial\delta}\\ \frac{1}{2} & \frac{\partial G_1}{\partial\lambda} & \frac{\partial G_1}{\partial\gamma} & \frac{\partial G_1}{\partial\delta}\\ \frac{1}{3} & \frac{\partial G_2}{\partial\lambda} & \frac{\partial G_2}{\partial\gamma} & \frac{\partial G_2}{\partial\delta}\\ \frac{1}{4} & \frac{\partial G_3}{\partial\lambda} & \frac{\partial G_3}{\partial\gamma} & \frac{\partial G_3}{\partial\delta}\end{pmatrix} \tag{12}$$

where

$$\begin{aligned}\frac{\partial G_r}{\partial\lambda} &= \int_{-\infty}^{+\infty} sinh\left(\frac{z-\gamma}{\delta}\right)\cdot\Phi^r(z)\cdot\varphi(z)dz \\ \frac{\partial G_r}{\partial\gamma} &= -\frac{\lambda}{\delta}\cdot\int_{-\infty}^{+\infty} cosh\left(\frac{z-\gamma}{\delta}\right)\cdot\Phi^r(z)\cdot\varphi(z)dz \\ \frac{\partial G_r}{\partial\delta} &= -\frac{\lambda}{\delta^2}\cdot\int_{-\infty}^{+\infty}(z-\gamma)cosh\left(\frac{z-\gamma}{\delta}\right)\cdot\Phi^r(z)\cdot\varphi(z)dz\end{aligned} \tag{13}$$

At each iteration, the integrals involved can be evaluated by the Gauss-Hermite quadrature in Appendix.

The initial values of δ_0 and γ_0 are evaluated as follows:

$$\begin{aligned}
\phi &= \frac{\kappa_3^2}{\kappa_4 - 3} \\
\delta_0 &= \left[\frac{1}{2} ln\left(\sqrt{2\kappa_4 - 2.8\kappa_3^2 - 2} - 1\right)\right]^{-\frac{1}{2}} \\
\gamma_0 &= \begin{cases} \delta_0 \left(0.7739\phi^{0.5473}\right) & \phi \leq 0.1 \\ \delta_0 \left[0.1456 + 0.4436 ln\left(\frac{0.545}{0.545-\phi}\right)\right] & 0.1 < \phi \leq 0.52 \end{cases}
\end{aligned} \tag{14}$$

δ_0 is given by Hill et al. (1976), γ_0 is given byDraper (1952). If $\phi > 0.52$, the S_L model is more appropriate.

Once δ_0 and γ_0 are determined, the initial values ξ_0 and λ_0 are easily calculated by solving the following system of linear equations:

$$\begin{aligned}
&\xi_0 + \lambda_0 \int_{-\infty}^{+\infty} sinh\left(\frac{z-\gamma_0}{\delta_0}\right) \cdot \varphi(z) dz = \beta_0 \\
&\frac{1}{2}\xi_0 + \lambda_0 \int_{-\infty}^{+\infty} sinh\left(\frac{z-\gamma_0}{\delta_0}\right) \cdot \Phi(z) \cdot \varphi(z) dz = \beta_1
\end{aligned} \tag{15}$$

3.3 S_B model

The parameters of the S_B model are obtained in a similar way as the S_U case. β_r is:

$$\beta_r = \frac{\xi + \lambda}{1 + r} - \lambda \int_{-\infty}^{+\infty} \frac{1}{1 + exp\left(\frac{z-\gamma}{\delta}\right)} \cdot \Phi^r(z) \cdot \varphi(z) dz \tag{16}$$

The entries in the Jacobian matrix are:

$$\begin{aligned}
\frac{\partial G_r}{\partial \lambda} &= \frac{1}{1+r} - \int_{-\infty}^{+\infty} \frac{1}{1 + exp\left(\frac{z-\gamma}{\delta}\right)} \cdot \Phi^r(z) \cdot \varphi(z) dz \\
\frac{\partial G_r}{\partial \gamma} &= -\frac{\lambda}{\delta} \int_{-\infty}^{+\infty} \frac{exp\left(\frac{z-\gamma}{\delta}\right)}{\left[1 + exp\left(\frac{z-\gamma}{\delta}\right)\right]^2} \cdot \Phi^r(z) \cdot \varphi(z) dz \\
\frac{\partial G_r}{\partial \delta} &= -\frac{\lambda}{\delta^2} \int_{-\infty}^{+\infty} \frac{(z-\gamma) exp\left(\frac{z-\gamma}{\delta}\right)}{\left[1 + exp\left(\frac{z-\gamma}{\delta}\right)\right]^2} \cdot \Phi^r(z) \cdot \varphi(z) dz
\end{aligned} \tag{17}$$

Notice that the random variable x generated by the S_B model is located in the interval $[\xi,\ \ \xi + \lambda]$, the initial values of ξ and λ can be determined by the variation range of x.

$$\begin{aligned}
\xi_0 &= F^{-1}(\alpha) \\
\lambda_0 &= F^{-1}(1-\alpha) - F^{-1}(\alpha)
\end{aligned} \tag{18}$$

where α represents a small percentage value, $F^{-1}(\cdot)$ is the inverse CDF of x.

As long as ξ_0 and λ_0 are known, ξ_0 and λ_0 are calculated as follows (Johnson & Kitchen, 1971):

$$\begin{aligned}
\mu^* &= \frac{\mu - \xi_0}{\lambda_0} \\
\sigma^* &= \frac{\sigma}{\lambda_0} \\
\delta_0 &= \frac{\mu^*(1-\mu^*)}{\sigma^*} + \frac{\sigma^*}{4}\left[\frac{1}{\mu^*(1-\mu^*)} - 8\right] \\
\gamma_0 &= \delta_0 \cdot ln\left(\frac{1-\mu^*}{\mu^*}\right) + \frac{1}{\delta_0}\left(\frac{1}{2} - \mu^*\right)
\end{aligned} \tag{19}$$

4. Accommodating the dependency

Suppose x_1 and x_2 are two correlated random variables with the correlation coefficient ρ_x. Through the transformation model in Eq.(1), two correlated standard normal variables, z_1 and z_2, are obtained. Let ρ_z denote the correlation coefficient between z_1 and z_2. This section is devoted to evaluate ρ_z for a given value of ρ_x. For the four models of Johnson system, there are ten combinations: $S_N - S_N$, $S_N - S_L$, $S_N - S_U$, $S_N - S_B$, $S_L - S_L$, $S_L - S_U$, $S_L - S_B$, $S_U - S_U$, $S_U - S_B$, $S_B - S_B$. Since $S_N - S_N$ can be easily handled, the other nine cases are discussed in this

4.1 $S_N - S_L$

For two correlated standard normal variables, the following properties hold (Pearson & Young, 1918):

$$\begin{aligned}
E\left(z_1 z_2^{2m}\right) &= 0 \\
E\left(z_1 z_2^{2m+1}\right) &= \rho_z \frac{(2m+1)!}{2^m \cdot m!}
\end{aligned} \tag{20}$$

where m is a nonnegative integer.

Let c denote a constant. Using the formulae in Eq.(20), the derivation below is carried out:

$$E[z_1 exp(cz_2)] = \sum_{i=0}^{\infty} E\left[z_1 \cdot \frac{(cz_2)^i}{i!}\right] = \sum_{j=0}^{\infty} \rho_z \cdot \frac{c^{2j+1}}{2^j \cdot j!} = \rho_z \cdot c \cdot exp\left(\frac{c^2}{2}\right) \tag{21}$$

Let x_1, x_2 denote the random variable obtained by S_N and S_L models in Eq.(8) respectively. The correlation coefficient ρ_x is calculated as:

$$\rho_x = \frac{E[x_1 x_2] - \mu_z \mu_2}{\sigma_z \sigma_2} \tag{22}$$

Substitute the expression of x_1 and x_2 into Eq.(22), using the formula in Eq.(21) yields:

$$\rho_z = \frac{\delta_2 \cdot \sigma_2}{\mu_2 - \xi_2} \cdot \rho_x \tag{23}$$

4.2 $S_N - S_U$

To handle the case of $S_N - S_U$ conveniently, the following derivation is performed.

Suppose x, y are two random variables. Let $y = y_1 + y_2$.

$$\begin{aligned} \rho_{xy} &= \frac{E\left[(x-\mu_x)(y-\mu_y)\right]}{\rho_x \rho_y} \\ &= \frac{E\left[(x-\mu_x)(y_1-\mu_{y_1})\right] + E\left[(x-\mu_x)(y_2-\mu_{y_2})\right]}{\rho_x \rho_y} \\ &= \frac{\sigma_{y_1}}{\sigma_y} \cdot \rho_{xy_1} + \frac{\sigma_{y_2}}{\sigma_y} \cdot \rho_{xy_2} \end{aligned} \tag{24}$$

Let x_2 denote the random variable resulted from S_U model.

$$\begin{aligned} x_2 &= \xi_2 + \lambda_2 sinh\left(\frac{z_2 - \gamma_2}{\delta_2}\right) = x_{21} + x_{22} \\ x_{21} &= \frac{\xi_2}{2} + \frac{\lambda_2}{2} \cdot exp\left(\frac{z_2 - \gamma_2}{\delta_2}\right) \\ x_{22} &= \frac{\xi_2}{2} - \frac{\lambda_2}{2} \cdot exp\left(-\frac{z_2 - \gamma_2}{\delta_2}\right) \end{aligned} \tag{25}$$

Using the formula in Eq.(24), the problem associated with $S_N - S_U$ can be transformed into $S_N - S_L$. ρ_z for ρ_x between $x_1(S_N)$ and $x_2(S_U)$ is evaluated as:

$$\rho_z = \frac{\frac{2\delta_2 \cdot \sigma_2}{\lambda_2} \cdot exp\left(-\frac{1}{2\delta_2^2}\right)}{exp\left(\frac{\gamma_2}{\delta_2}\right) + exp\left(-\frac{\gamma_2}{\delta_2}\right)} \cdot \rho_x \tag{26}$$

4.3 $S_N - S_B$

ρ_x and ρ_z are related by the following equation:

$$\rho_x \sigma_1 \sigma_2 + \mu_1 \mu_2 = E(x_1 x_2) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} x_1 x_2 \varphi(z_1, z_2, \rho_z) dz_1 dz_2 \tag{27}$$

where $\varphi(z_1, z_2, \rho_z)$ is the bivariate standard normal PDF.

Substitute the expression of $x_1(S_N)$ and $x_2(S_B)$ into Eq.(27), the following formula is obtained after some manipulations:

$$\rho_x = \frac{-\lambda_2}{\sqrt{2\pi}\sigma_2} \int_{-\infty}^{+\infty} \frac{t}{1 + exp\left(\frac{t-\gamma_2}{\delta_2}\right)} \cdot exp\left(-\frac{t^2}{2}\right) dt \cdot \rho_z \tag{28}$$

The integral on the right side can be evaluated by the Gauss-Hermite quadrature.

4.4 $S_L - S_L$

As for the case of $S_L - S_L$, x_1 and x_2 both follow the Lognormal distribution. The function relationship between ρ_z and ρ_x is as follows (Mostafa & Mahmoud, 1964):

$$\rho_x = \frac{exp\left(\frac{\rho_z}{\delta_1\delta_2}\right) - 1}{\sqrt{\left[exp\left(\frac{1}{\delta_1^2}\right) - 1\right] \cdot \left[exp\left(\frac{1}{\delta_2^2}\right) - 1\right]}} \tag{29}$$

4.5 $S_L - S_U$

The problem is converted into $S_L - S_L$ using the formula in Eq.(24). ρ_x between $x_1(S_L)$ and $x_2(S_U)$ is estimated by the following equation:

$$\rho_x = \frac{\frac{\lambda_2}{2} exp\left(\frac{1}{2\delta_2^2}\right)}{\sigma_2 \cdot \sqrt{exp\left(\frac{1}{\delta_1^2}\right) - 1}} \left[exp\left(\frac{\rho_z}{\delta_1\delta_2} - \frac{\gamma_2}{\delta_2}\right) - exp\left(-\frac{\rho_z}{\delta_1\delta_2} + \frac{\gamma_2}{\delta_2}\right) + exp\left(\frac{\gamma_2}{\delta_2}\right) - exp\left(-\frac{\gamma_2}{\delta_2}\right)\right] \tag{30}$$

For a given value of ρ_x, solve the equation for ρ_z, the valid solution is restricted by the following conditions:

$$|\rho_z| \leq 1 \quad and \quad \rho_x\rho_z \geq 0 \tag{31}$$

4.6 $S_U - S_U$

Transform the problem into $S_L - S_U$, the equation solved for ρ_z is:

$$\rho_x\sigma_1\sigma_2 + (\mu_1 - \xi_1)(\mu_2 - \xi_2) = \frac{\lambda_1\lambda_2}{4} \cdot exp\left(\frac{1}{2\delta_1^2} + \frac{1}{2\delta_2^2}\right) \left\{ \left[exp\left(\frac{\gamma_1}{\delta_1} + \frac{\gamma_2}{\delta_2}\right) + exp\left(-\frac{\gamma_1}{\delta_1} - \frac{\gamma_2}{\delta_2}\right)\right] exp\left(\frac{\rho_z}{\delta_1\delta_2}\right) - \left[exp\left(\frac{\gamma_1}{\delta_1} - \frac{\gamma_2}{\delta_2}\right) + exp\left(-\frac{\gamma_1}{\delta_1} + \frac{\gamma_2}{\delta_2}\right)\right] exp\left(-\frac{\rho_z}{\delta_1\delta_2}\right) \right\} \tag{32}$$

4.7 $S_L/S_U/S_B - S_B$

For the cases of $S_L - S_B$, $S_U - S_B$ and $S_B - S_B$, analytical expressions are unobtainable, other approaches should be tried. Suppose x_2 represents the random variable generated by S_B model. Substituting the expression of x_1 into Eq.(27) yields:

$S_L - S_B$

$$\rho_x = -\frac{(\mu_1-\xi_1)(\mu_2-\xi_2-\lambda_2)}{\sigma_1\sigma_2} - \frac{\lambda_2(\mu_1-\xi_1)}{\sqrt{2\pi}\sigma_1\sigma_2}\cdot \int_{-\infty}^{+\infty} \frac{1}{1+exp\left(\frac{t}{\delta_2}+\frac{\rho_z}{\delta_1\delta_2}-\frac{\gamma_2}{\delta_2}\right)} exp\left(-\frac{t^2}{2}\right) dt \tag{33}$$

$S_U - S_B$

$$\rho_x = -\frac{(\mu_1-\xi_1)(\mu_2-\xi_2-\lambda_2)}{\sigma_1\sigma_2} - \frac{\lambda_1\lambda_2 exp\left(\frac{1}{2\delta_1^2}\right)}{2\sqrt{2\pi}\sigma_1\sigma_2}\cdot \int_{-\infty}^{+\infty} \left[\frac{exp\left(-\frac{\gamma_1}{\delta_1}\right)}{1+exp\left(\frac{t}{\delta_2}+\frac{\rho_z}{\delta_1\delta_2}-\frac{\gamma_2}{\delta_2}\right)} - \frac{exp\left(\frac{\gamma_1}{\delta_1}\right)}{1+exp\left(\frac{t}{\delta_2}-\frac{\rho_z}{\delta_1\delta_2}-\frac{\gamma_2}{\delta_2}\right)}\right] exp\left(-\frac{t^2}{2}\right) dt \tag{34}$$

$S_B - S_B$

$$\rho_x = -\frac{(\mu_1-\xi_1-\lambda_1)(\mu_2-\xi_2-\lambda_2)}{\sigma_1\sigma_2} + \frac{\lambda_1\lambda_2}{2\pi\sigma_1\sigma_2}\cdot \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \frac{1}{1+exp\left(\frac{v-\gamma_2}{\delta_2}\right)}\cdot \frac{1}{1+exp\left(\frac{\sqrt{1-\rho_z^2}u}{\delta_1}+\frac{\rho_z v-\gamma_1}{\delta_1}\right)} exp\left(-\frac{v^2}{2}\right) exp\left(-\frac{u^2}{2}\right) dudv \tag{35}$$

For the three equations above, it is obvious that ρ_x is a continuous function with respect to ρ_z, which is located in the interval $[-1,\ 1]$. According to Weierstrass approximation theorem (Saxe, 2001), ρ_x can be approximated to any required accuracy by a polynomial of ρ_z.

Choose several values of ρ_z over the interval $[-1, 1]$, calculate the integral by Gauss-Hermite quadrature, evaluate the value of ρ_x. Then, a polynomial is established by an interpolation method.

$$\rho_x = a_n\rho_z^n + \cdots + a_1\rho_z + a_0 \tag{36}$$

For a specified ρ_x, ρ_z can be determined by solving the polynomial equation above, the valid solution is also restricted by conditions in Eq.(31). In general, a satisfactory result would be obtained by a polynomial of degree less than 10.

From numerical investigations where $|\kappa_3| \leq 20$ and $|\kappa_4| \leq 120$, it is found that the integral with $|\delta| > 0.5$ can be accurately evaluated by a 9-point Gauss-Hermite quadrature. If $|\delta| < 0.5$, more quadrature nodes are required.

4.8 Generating multivariate random vector

Suppose $\boldsymbol{X} = (x_1, \ldots, x_i, \ldots, x_m)^T$ is an m-dimensional random vector with correlation matrix $\boldsymbol{R_x}$, the steps for generating $\boldsymbol{X}$ are as follows:

1. Choose a suitable transformation model for each variable x_i of $\boldsymbol{X}$ and determine the parameters.

2. Calculate $\rho_z(i,j)$ $(i \neq j)$ for each entry $\rho_x(i,j)$ in the matrix $\boldsymbol{R_x}$, obtain the equivalent correlation matrix $\boldsymbol{R_z}$ in the standard normal space.

3. Generate m-dimensional vector of independent standard normal variables, $\boldsymbol{U} = (u_1, \ldots, u_i, \ldots, u_m)^T$, transform $\boldsymbol{U}$ into correlated standard normal vector $\boldsymbol{Z} = (z_1, \ldots, z_i, \ldots, z_m)^T$ by performing $\boldsymbol{Z} = \boldsymbol{LU}$. $\boldsymbol{L}$ is the lower triangular matrix resulted from Cholesky decomposition of $\boldsymbol{R_z}$, that is, $\boldsymbol{R_z} = \boldsymbol{LL^T}$.

4. Convert each element z_i into x_i by the corresponding model. The random vector $\boldsymbol{X}$ with the prescribed marginal distributions and correlation matrix $\boldsymbol{R_x}$ could be obtained.

As long as the correlation matrix $\boldsymbol{R_x}$ is a positive definite symmetric matrix (Cario & Nelson, 1997), and the marginal distribution of x_i can be well simulated by Johnson system, the correlated random vector $\boldsymbol{X}$ can be generated by the proposed method.

5. Numerical example

Four numerical examples are performed in Matlab to demonstrate the proposed approach. The integrals involved are all evaluated by a 21-point Gauss-Hermite quadrature. The Newton's iteration method is required to get five digits of accuracy.

Consider the T-distribution with 4 degrees of freedom ($T(4)$), the first four standardized central moments are: $\mu = 0$, $\sigma = 1.4142$, $\kappa_3 = 0$, $\kappa_4 = +\infty$. Solving the equation in Eq.(2) yields $\omega = 1$. According to Eq.(4), the S_U model is chosen. To enable the calculation of the initial values, κ_4 is set to be 10^{16}. The initial

values are: $\xi_0 = 0$, $\lambda_0 = 0.0806$, $\delta_0 = 0.4617$, $\gamma_0 = 0$. The solutions of the system of equations are: $\xi = 0$, $\lambda = 1.4399$, $\delta = 1.3760$, $\gamma = 0$. The graphs of PDF are depicted in Fig.1.

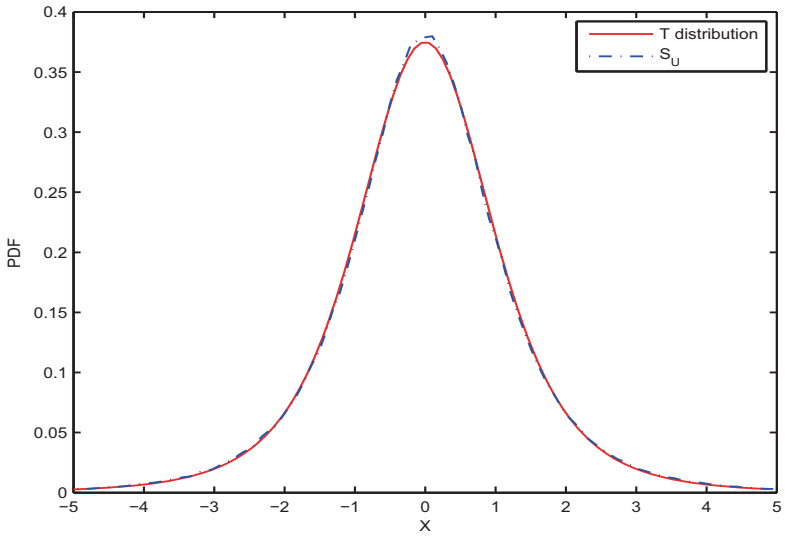

Figure 1: The PDF of T-distribution with 4 degrees of freedom $T(4)$

For a given percentage p, evaluate the associated percentile t_p from $T(4)$ and the percentile x_p given by the S_U model in Eq.(8), the relative error between t_p and x_p is calculated:

$$\varepsilon = \left| \frac{x_p - t_p}{t_p} \right| \times 100[\%] \qquad t_p = T^{-1}(p) \qquad x_p = \xi + \lambda sinh\left(\frac{\Phi^{-1}(p) - \gamma}{\delta} \right) \tag{37}$$

where $T^{-1}(\cdot)$ is the inverse CDF of $T(4)$, $\Phi^{-1}(\cdot)$ is the inverse CDF of the standard normal variable.

10000 values of p are chosen evenly within the interval $[0.01, 0.99]$. Evaluating ε for each percentage p as stated in Eq.(38), the average value is 1.55%, the maximum value is 4.63%, the minimum value is 0.0002%. The Johnson system offers satisfactory accuracy for simulating $T(4)$.

Another example is given related to the Gamma distribution $\Gamma(10,\ 1)$. The first four standardized central moments are: $\mu = 10.0000$, $\sigma = 3.1623$, $\kappa_3 = 0.6325$, $\kappa_4 = 3.6000$. The S_B model is preferable. Setting $\alpha = 0.01$, the initial values are obtained: $\xi_0 = 4.1302$, $\lambda_0 = 14.6529$, $\delta_0 = 0.9057$, $\gamma_0 = 0.4748$. The results are: $\xi = -1.8372$, $\lambda = 72.2970$, $\delta = 3.1212$, $\gamma = 5.1961$. The graphs of PDF are shown in Fig.2.

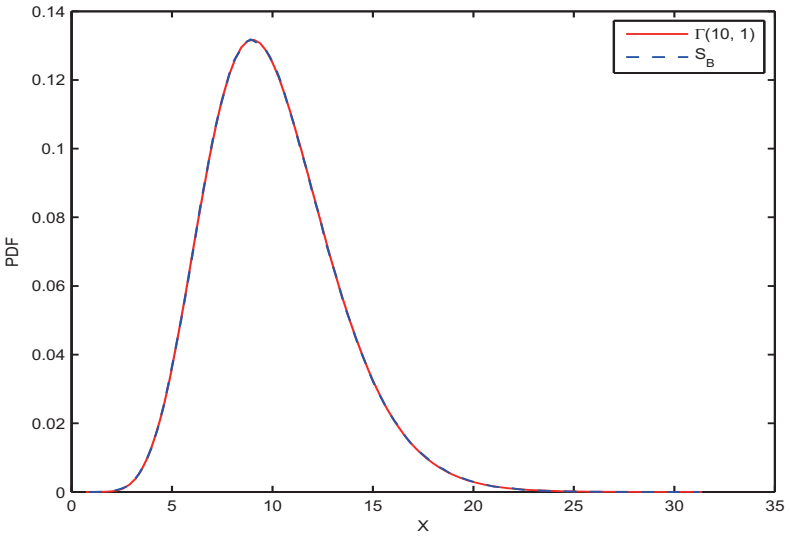

Figure 2: The PDF of Gamma distribution $\Gamma(10,\ 1)$

10000 values of p are chosen evenly within the interval $[0.01, 0.99]$. Evaluating ε for each percentage p as the former case:

$$\varepsilon = \left| \frac{x_p - t_p}{t_p} \right| \times 100[\%] \qquad t_p = G^{-1}(p) \qquad x_p = \xi + \lambda - \frac{\lambda}{1 + exp\left(\frac{\Phi^{-1}(p) - \gamma}{\delta}\right)} \tag{38}$$

where $G^{-1}(\cdot)$ is the inverse CDF of $\Gamma(10,\ 1)$.

The average value of ε is $2.52 \times 10^{-5}\%$, the maximum value is 1.81%, the minimum value is 0.18%.

Suppose x_1 and x_2 both follow the Gamma distribution $\Gamma(10,\ 1)$. Select four values of ρ_z that are evenly spaced over the interval $[-1,\ 1]$, evaluate the double integral in Eq.(35), obtain the associated ρ_x, the polynomial is established by Newton's interpolation method.

$$\rho_x = 4.1721 \times 10^{-5}\rho_z^3 + 0.0215\rho_z^2 + 0.9780\rho_z - 0.0025 \tag{39}$$

201 values of ρ_z are chosen from -1 to 1 in steps of 0.01. For each value of ρ_z, 10^6 bivariate Gamma random vectors are generated by the procedures in Section 4.8, and ρ_x is calculated by evaluating the correlation coefficient of bivariate Gamma random samples. Along with the ρ_x evaluated by the polynomial in Eq.(39), the function curves are plotted in Fig. 3.

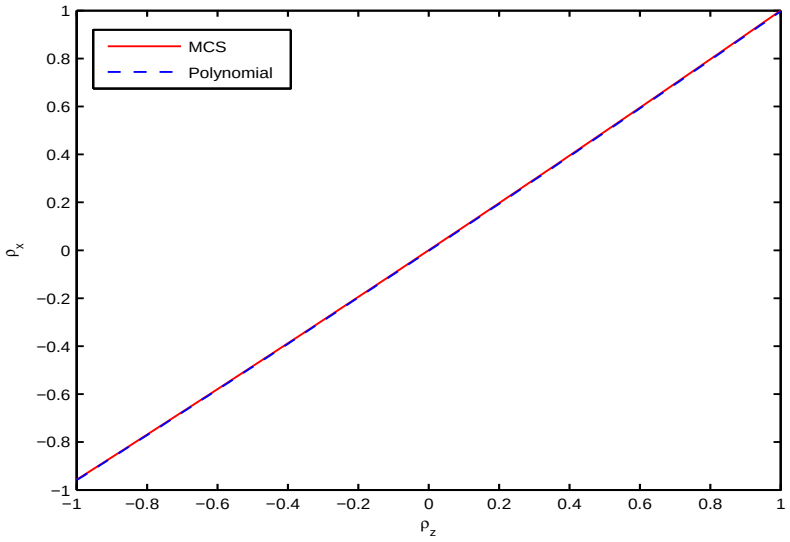

Figure 3: The function curve between ρ_z and ρ_x

Several pairs of (ρ_z, ρ_x) values are presented in Table 1.

Table 1: The values of $(\rho_z,\ \rho_x)$

ρ_z	ρ_x(Sample)	ρ_x(Polynomial)	ρ_z	ρ_x(Sample)	ρ_x(Polynomial)
-0.9	-0.863	-0.865	0.1	0.098	0.096
-0.7	-0.674	-0.677	0.3	0.296	0.293
-0.5	-0.484	-0.486	0.5	0.494	0.492
-0.3	-0.292	-0.294	0.7	0.695	0.693
-0.1	-0.097	-0.100	0.9	0.898	0.895

Inspection of Fig.3 and Table 1 indicates that the polynomial in Eq.(39) suffices for a good approximation of the functional relationship between ρ_z and ρ_x.

Finally, an example for generating multivariate random vector with specified marginal distributions and correlation matrix is worked. Consider the random vector $\boldsymbol{X} = (x_1,\ x_2,\ x_3,\ x_4)^T$, x_1 follows the standard normal distribution, x_2, Lognormal distribution $lnN(0\ 1)$, x_3, T-distribution with 4 degrees of freedom,

x_4, Gamma distribution $\Gamma(10,\ 1)$. The desired correlation matrix $\boldsymbol{R_x}$ is:

$$\boldsymbol{R_x} = \begin{pmatrix} 1 & 0.7 & 0.4 & 0.5 \\ 0.7 & 1 & 0.3 & 0.6 \\ 0.4 & 0.3 & 1 & 0.2 \\ 0.5 & 0.6 & 0.2 & 1 \end{pmatrix} \tag{40}$$

The equivalent correlation matrix $\boldsymbol{R_z}$ evaluated by the proposed method is:

$$\boldsymbol{R_z} = \begin{pmatrix} 1 & 0.918 & 0.406 & 0.507 \\ 0.918 & 1 & 0.402 & 0.737 \\ 0.406 & 0.402 & 1 & 0.209 \\ 0.507 & 0.737 & 0.209 & 1 \end{pmatrix} \tag{41}$$

5×10^6 random vectors $\boldsymbol{X} = (x_1,\ x_2,\ x_3,\ x_4)^T$ are generated, the correlation matrix of the samples is:

$$\boldsymbol{R_x^*} = \begin{pmatrix} 1 & 0.698 & 0.397 & 0.502 \\ 0.698 & 1 & 0.303 & 0.598 \\ 0.397 & 0.303 & 1 & 0.202 \\ 0.502 & 0.598 & 0.202 & 1 \end{pmatrix} \tag{42}$$

Comparing $\boldsymbol{R_x^*}$ with $\boldsymbol{R_x}$ indicates that the correlation matrix of samples is in close agreement to the desired one.

6. Conclusion

Through the transformation models of Johnson system, a variety of distributions can be simulated by the standard normal distribution. Random numbers with a specified distribution can be efficiently generated by an elementary transformation of standard normal deviates. In the context of probability weighted moments, this paper develops a simpler approach to evaluate the parameters of Johnson system. To allow for the generation of multivariate non-normal distributions with desired correlation matrix, formulae are derived to calculate the equivalent correlation coefficient in the standard normal space. Extension of Johnson system to the multivariate data generation makes it more applicable to stochastic modeling and simulation study.

Appendix

The univariate integrals of the form $\int_{-\infty}^{+\infty} e^{-t^2} f(t)dt$ may be approximated with a Gaussian-type formula (Naylor & Smith, 1982).

$$\int_{-\infty}^{+\infty} e^{-t^2} g(t)dt = \sum_{k=1}^{n} \omega_k g(t_k) \tag{43}$$

where n is the number of the quadrature nodes. t_k is the kth zero of the Hermite polynomial $H_n(t)$.

$$H_n(t) = (-1)^n e^{t^2} \frac{d^n}{dt^n} e^{-t^2} \tag{44}$$

The associated weights ω_k are given by:

$$\omega_k = \frac{2^{n-1} n! \sqrt{\pi}}{n^2 [H_{n-1}(t_k)]^2} \tag{45}$$

Using Eq.(43), the following formula can be easily derived:

$$\int_{-\infty}^{+\infty} g(t) \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt = \sum_{k=1}^{n} \frac{\omega_k}{\sqrt{\pi}} g(\sqrt{2} t_k) \tag{46}$$

Extension of Gauss-Hermite quadrature to double integral is as follows:

$$\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} g(t_1, t_2) \cdot \frac{1}{2\pi} e^{-\frac{t_1^2}{2}} \cdot e^{-\frac{t_2^2}{2}} dt_1 dt_2 = \frac{1}{\pi} \sum_{k=1}^{n} \sum_{l=1}^{n} \omega_k \omega_l g(\sqrt{2} t_k, \sqrt{2} t_l) \tag{47}$$

The nodes t and weights ω of the 21-point Gauss-Hermite are presented in Table 2.

Table 2: Nodes t and weights ω for the 21-point Gauss−Hermite formula

t	ω
0	0.6774 4181 7650 738
±0.6780 4569 2440 645	0.6792 5764 5819 217
±1.3597 6582 3211 230	0.6848 3436 1406 712
±2.0491 0246 8257 180	0.6945 8757 5425 530
±2.7505 9298 1052 280	0.7093 1584 3106 891
±3.4698 4669 0475 750	0.7304 1501 1188 135
±4.2143 4398 1687 360	0.7603 4305 2415 214
±4.9949 6394 4784 140	0.8037 1612 0812 564
±5.8293 8200 7301 600	0.8703 3646 0017 752
±6.7514 4471 8719 760	0.9861 0234 0998 038
±7.8493 8289 5113 030	1.2593 6799 7734 310

References

Cario, M. C. and Nelson, B. L. (1997). Modeling and generating random vectors with arbitrary marginal distributions and correlation matrix. Department of Industrial Engineering and Management Sciences, Northwestern University.

Draper, J. (1952). Properties of distributions resulting from certain simple transformations of the normal distribution. *Biometrika*, **39**, 290-301.

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

Greenwood, J. A., Landwehr, J. M., Matalas, N. C. and Wallis, J. R. (1979). Probability weighted moments: definition and relation to parameters of several distributions expressable in inverse form. *Water resources research*, **15**, 1049-1054.

Headrick, T. C. (2011). A characterization of power method transformations through L-moments. *Journal of probability and statistics*, **2011**, 1-22.

Hill, I. D., Hill, R. and Holder, R. L. (1976). Fitting Johnson curves by moments. Journal of the royal statistical society. *Series C (Applied statistics)*, **25**, 180-189.

Hosking, J. R. M. (1990). L-moments: analysis and estimation of distribution using linear combinations of order statistics. Journal of the royal statistical society. *Series B (Methodological)*, **52**, 105-124.

Johnson, N. L. (1949). Systems of frequency curves generated by methods of translation. *Biometrika*, **36**, 149-176.

Johnson, N. L. and Kitchen, J. O. (1971). Tables to facilitate fitting SB curves II: both terminals known. *Biometrika*, **58**, 657-668.

Karvanen, J. and Nuutinen, A. (2008). Characterizing the generalized lambda distribution by L-moments. *Computational Statistics and data Analysis*, **52**, 1971-1983.

Mostafa, M. D. and Mahmoud, M. W. (1964). On the problem of estimation for the bivariate lognormal distribution. *Biometrika*, **51**, 522-527.

Naylor, J. C. and Smith, A. F. M. (1982). Applications of a method for the efficient computation of posterior distributions. Journal of the royal statistical society. *Series C (Applied statistics)*, **31**, 214-225.

Pearson, K. and Young, A. W. (1918). On the product-moments of various orders of the normal correlation surface of two variates. *Biometrika*, **12**, 86-92.

Ramberg, J. S. and Schmeiser, B. W. (1974). An approximate method for generating asymmetric random variables. *Communications of the ACM*, **17**, 78-82.

Saxe, K. (2001). Beginning Functional Analysis. New York: Springer-Verlag.

Slifker, J. F. and Shapiro, S. S. (1980). The Johnson system: selection and parameter estimation. *Technometrics*, **22**, 239-246.

Stanfield, P. M., Wilson, J. R., Mirka, G. A., Glasscock, N. F., Psihogios, J. P. and Davis, J. R. (1996). Multivariate input modeling with Johnson distributions. In *Proceeding WSC '96 Proceedings of the 28th conference on Winter simulation* (pp. 1457-1464). Coranado.

Tadikamalla, P. R. (1980). On simulating non-normal distributions. *Psychometrika*, **45**, 273-279.

Received March 17, 2013; accepted August 16, 2013.

Qing Xiao
School of Mechatronic Engineering and Automation
Shanghai University
Shanghai, 200072, China
xaoshaoying@shu.edu.cn