Comparison of Algorithms for Generating Poisson Random Vectors: A Review

Masahiko Gosho* and Kazushi Maruo

Clinical Data Science Dept.,
Kowa Company Ltd, Tokyo, Japan

*Corresponding author: m-gosyo@kowa.co.jp

Abstract

Multivariate correlated Poisson data are commonly analyzed in longitudinal or clustered studies. Monte Carlo simulation studies are often useful to evaluate the properties of the parameter estimator under the finite sample size when the Poisson model is fitted to such data. In this paper, we review two algorithms proposed by Sim, J Stat Comput Sim 47:1–10, (1993) and Krummenauer, Biometrical J 40:823–832, (1998), for generating the multivariate Poisson random numbers with non-identity covariance matrix. In addition, we graphically represent the range constraints for the correlation parameter of the multivariate Poisson distribution with the two algorithms. Consequently, the constraint of the algorithm proposed by Krummenauer (1998) was much stricter than that of the algorithm proposed by Sim (1993).

Key Words: Covariance Matrix, Correlation Matrix, Multivariate Poisson Random Numbers

1 Introduction

Many scientific investigations focus on the correlated measurement of the frequency of a specific behavior or activity. The first example considers data from a randomized clinical trial of 59 epileptics reported by Diggle et al. (2002). The number of incidents of epileptic seizures for each patient was longitudinally measured in four consecutive two-week intervals during a baseline period of eight weeks. Patients were randomized to treatment with an anti-epileptic drug or a placebo in addition to standard chemotherapy. The second example considers data from a study of wave damage to cargo ships (McCullagh and Nelder, 1990). The purpose of this study was to evaluate the association of the damage incidents with the ship type, year of construction, period of operation, and aggregate months service. The final example regards suicide data from the National Center for Health Statistics for each US country reported by Hedeker and Gibbons (2006). The number of suicides was reported as clustered data in the US countries to investigate the effect of antidepressant drug use, age, race, and gender on the suicide rates.

For analyzing the above mentioned correlated count data, Poisson distribution can typically be obtained and a generalized linear mixed model approach and generalized estimating equations method are often applied to such data. Since the validity of most such methods is based on asymptotic theory, the use of Monte Carlo simulation studies has become necessary to assess the finite sample property of these methods.

Although the multivariate Poisson distribution is one of the most well known and important multivariate discrete distributions, it has not found many practical applications apart from the special case of the bivariate Poisson distribution. The main reason for this is the awkward probability function, which causes the inferential procedures to become too complicated for practical purposes. Nevertheless, a few researchers have proposed methods for generating multivariate Poisson random numbers. For example, Sim (1993) has provided an algorithm to generate the random vectors of multivariate Poisson distribution with a given covariance matrix based on a transformation method. Moreover, Krummenauer (1998) has proposed an algorithm for simulation of the multivariate Poisson variables by applying the trivariate reduction method.

In this paper, we review the two procedures proposed by Sim (1993) and Krummenauer (1998) and graphically represent the limitation of the use of the two procedures. In Sect. 2, we provide a brief overview of the notation and definition of the multivariate Poisson distribution. In Sect. 3, we represent the algorithms proposed by Sim (1993) and Krummenauer (1998). In Sect. 4, we illustrate the constraints of the algorithms proposed by Sim (1993) and Krummenauer (1998). Finally, in Sect. 5, we provide some discussions and concluding remarks.

2 The multivariate Poisson distribution

Suppose that random p-dimensional vector $\mathbf{Z} = (Z_1, \ldots, Z_p)^{\mathrm{T}}$ follows multivariate Poisson distribution with some parameter collection $\mathbf{\Lambda}$, where $\mathbf{\Lambda} := \{\lambda_I : \emptyset \neq I \subseteq \{1, \ldots, p\}\}$ and the superscript T denotes the transpose. The probability generating function of the multivariate Poisson distribution is given by

$$g(s_1, \ldots, s_p) = \exp\left\{ \sum_{\emptyset \neq I \subseteq \{1,\ldots,p\}} \lambda_I \cdot \sum_{i \subseteq I} (s_i - 1) \right\}.$$

For example, if $p = 3$, $g(s_1, s_2, s_3) = \lambda_1(s_1 - 1) + \lambda_2(s_2 - 1) + \lambda_3(s_3 - 3) + \lambda_{12}(s_1 - 1)(s_2 - 2) + \lambda_{13}(s_1 - 1)(s_3 - 3) + \lambda_{23}(s_2 - 2)(s_3 - 3) + \lambda_{123}(s_1 - 1)(s_2 - 2)(s_3 - 3)$. It shall be referred to as the p-variate Poisson distribution $\mathrm{MVP}_p(\mathbf{\Lambda})$ with the parameter collection $\mathbf{\Lambda}$. Characterization of the parameters λ_I becomes feasible by differentiating $g(\cdot)$, e.g., mean vector is shown as $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_p)^{\mathrm{T}}$ and covariance matrix is a $p \times p$ matrix $\mathbf{\Sigma}$ with diagonal elements $\mathrm{Var}(Z_i) = \lambda_i$ and off-diagonal elements $\mathrm{Cov}(Z_i, Z_j) = \lambda_{ij}$ for $1 \leq i < j \leq p$. The parameter collection $\mathbf{\Lambda}$ therefore contains the mean vector $\boldsymbol{\lambda}$ and the covariance matrix $\mathbf{\Sigma}$. Here, $\mathbf{\Sigma}$ should be positive definite matrix. Let $\rho_{ij} = \mathrm{Cov}(Z_i, Z_j)/\sqrt{\mathrm{Var}(Z_i)\mathrm{Var}(Z_j)} = \lambda_{ij}/\sqrt{\lambda_i \lambda_j}$ denote a correlation parameter between Z_i and Z_j. The literature for the multivariate Poisson distribution is large and many references, as well as historical remarks, can be found in Johnson et al. (1997).

3 Algorithms for generating multivariate Poisson numbers

In this section, we review the two algorithms for generating multivariate Poisson numbers proposed by Sim (1993) and Krummenauer (1998) and summarize some limitations for these algorithms. In addition, we attach the SAS programs of these algorithms in the Appendix.

3.1 Sim's algorithm

Sim (1993) has provided a procedure for generating the random vectors following the multivariate Poisson distribution with non-identical marginals and a fixed covariance matrix, based on a

transformation method. According to Sim (1993), it is only necessary to specify covariance matrix $\mathbf{\Sigma}$ for generating the random vectors. Suppose that X_i are independent Poisson random variables with mean μ_i, for $i = 1, \ldots, p$ and the random p-dimensional vector $\mathbf{X} = (X_1, \ldots, X_p)^{\mathrm{T}}$ with mean vector $\boldsymbol{\mu} = (\mu_1, \cdots, \mu_p)^{\mathrm{T}}$ and identity covariance matrix. The required Poisson random vector $\mathbf{Z}$ with the mean vector $\boldsymbol{\lambda}$ and the covariance matrix $\mathbf{\Sigma}$ is provided through the following transformation

$$\begin{aligned} Z_1 &= X_1 \\ Z_2 &= \alpha_{12} * X_1 + X_2 \\ Z_3 &= \alpha_{13} * X_1 + \alpha_{23} * X_2 + X_3 \\ &\vdots \\ Z_p &= \alpha_{1p} * X_1 + \alpha_{2p} * X_2 + \alpha_{3p} * X_3 + \ldots + \alpha_{(p-1)p} * X_{p-1} + X_p \end{aligned}$$

where $\alpha_{ij} \in [0, 1]$, $1 \leq i < j \leq p$, are the required unknown parameters. The random variable $\alpha * X$ is defined as the sum of X independent units where each unit is retained with a probability α, or removed withe a probability $1-\alpha$. The following algorithm summarizes the necessary steps to generate the Poisson random vector $\mathbf{Z}$ with the mean vector $\boldsymbol{\lambda}$ and the covariance matrix $\mathbf{\Sigma}$.

(1) Determine the covariance matrix $\mathbf{\Sigma}$.

(2) Set $\mu_1 = \lambda_1$.

(3) Compute $\alpha_{1j} = \lambda_{1j}/\mu_1$.

(4) Compute $\mu_j = \lambda_j - \sum_{k=1}^{j-1} \alpha_{kj}\mu_k$ for $2 \leq j \leq p$ and $\alpha_{ij} = (\lambda_{ij} - \sum_{k=1}^{i-1} \alpha_{ki}\alpha_{kj}\mu_k)/\mu_i$ for $2 \leq i < j \leq p$, iteratively.

(5) Generate realizations $(x_1, \cdots, x_p)$ of Poisson random vector $\mathbf{X}$ with mean vector $\boldsymbol{\mu}$ and identity covariance matrix, which can be generated by using standard software packages providing random call routines for univariate distributions, and let $Z_1 = X_1$, deliver realization z_1 of Z_1.

(6) Generate realizations $\alpha_{ij} * x_i$ of independent random variables $\alpha_{ij} * X_i$ following binomial distribution $\mathrm{Bin}(X_i, \alpha_{ij})$ and obtain realizations z_j from $\alpha_{ij} * x_i$ and x_j.

This algorithm is very simple and elegant but has some limitations. The underlying limitations are $\alpha_{ij} \in [0, 1]$ and $\mu_j > 0$ for $1 \leq i < j \leq p$ to conduct the algorithm. Under these limitations, it is necessary to satisfy $0 < \lambda_{ij} < \min(\lambda_i, \lambda_j)$ which imply all the elements of the random vector $\mathbf{Z}$ must be positively correlated and that the full range of positive correlation are allowed only when $\lambda_i = \lambda_j$. In practice, it is easier to determine whether the given covariance matrix $\mathbf{\Sigma}$ would ensure that $\alpha_{ij} \in [0, 1]$ and $\mu_j > 0$ for $1 \leq i < j \leq p$, by checking whether $\lambda_j > \sum_{k=1}^{j-1} \alpha_{kj}\mu_k$ and $\lambda_{ij} \in [a, \mu_i + a]$ where $a = 0$ for $i = 1$ and $a = \sum_{k=1}^{i-1} \alpha_{ki}\alpha_{kj}\mu_k$ for $i \geq 2$.

3.2 Krummenauer's algorithm

Krummenauer (1998) has proposed another algorithm for generating the random vectors $\mathbf{Z}$ following the multivariate Poisson distribution with the parameter collection $\mathbf{\Lambda}$ based on an extension of "trivariate reduction method" shown by Holgate (1964). Let $\{X_I : \emptyset \neq I \subseteq \{1, \ldots, p\}\}$ denote a family stochastically independent Poisson random variables with X_I following Poisson distribution $\mathrm{Poi}(\mu_I)$ for $\emptyset \neq I \subseteq \{1, \ldots, p\}$. The algorithm is given as follows:

(1) Determine the parameter collection $\mathbf{\Lambda}$ that characterizes the p-variate Poisson distribution to be simulated, that is specify the parameter λ_I for all $I \neq \emptyset$.

(2) Solve the following linear equation system for μ_I for all $I \neq \emptyset$

$$\lambda_I \quad := \sum_{I \subseteq J \subseteq \{1,\ldots,p\}} \mu_J \forall \emptyset \quad \neq I \subseteq \{1,\ldots,p\}$$

by backward substitution.

(3) Generate realizations x_I of independent random variables X_I from Poi(μ_I) for $\emptyset \quad \neq I \subseteq \{1,\ldots,p\}$ using μ_I derived from step (2), which can be generated by using standard software packages providing random call routines for univariate distributions.

(4) Compute marginal realizations

$$z_i := \sum_{I \subseteq \{1,\ldots,p\}: i \in I} x_I \qquad (i = 1,\ldots,p),$$

which provide the realization $(z_1,\ldots,z_p)$ of the random vector $\mathbf{Z}$.

If e.g., $p = 3$, then step (2) means solving the linear equation system for μ_I

$$\begin{aligned}
\lambda_{123} &= \mu_{123}, \\
\lambda_{12} &= \mu_{12} + \mu_{123}, \quad \lambda_1 = \mu_1 + \mu_{12} + \mu_{13} + \mu_{123}, \\
\lambda_{13} &= \mu_{13} + \mu_{123}, \quad \lambda_2 = \mu_2 + \mu_{12} + \mu_{13} + \mu_{123}, \\
\lambda_{23} &= \mu_{23} + \mu_{123}, \quad \lambda_3 = \mu_3 + \mu_{13} + \mu_{23} + \mu_{123}
\end{aligned}$$

by backward substitution, i.e., first determining μ_{123}, next computing μ_{12}, μ_{13}, μ_{23} and then μ_1, μ_2, μ_3.

This algorithm has also a limitation relevant to the parameter collection $\mathbf{\Lambda}$. All parameters λ_I are necessarily non-negative and therefore the above modeling implies necessarily non-negative correlation between Poisson distributed random variables. For example, it is more concurrently constraint to satisfy $\lambda_1 > \lambda_{12} + \lambda_{13}$, $\lambda_2 > \lambda_{12} + \lambda_{23}$, and $\lambda_3 > \lambda_{13} + \lambda_{23}$ when $p = 3$.

3.3 Check of the algorithms

We conducted a simple simulation study to check the validity of the algorithms by Sim (1993) and Krummenauer (1998) in this section. Henceforth, we refer to the algorithms proposed by Sim (1993) and Krummenauer (1998) as SIM and KRM, respectively.

In the simulation study, we generated the random 4-dimensional vector $\mathbf{z} = (z_1,\ldots,z_4)^{\mathrm{T}}$ with $\boldsymbol{\lambda} = (\lambda_1,\ldots,\lambda_4)^{\mathrm{T}} = (2,\ldots,2)^{\mathrm{T}}$ and AR(1) structure with $\rho_{ij} = \rho^{j-i}$ (when $\rho = 0.4$) by SIM and KRM. We assumed that the parameters with more than 3 subscripts (e.g. λ_{123} and λ_{1234}) were zero. Data generation was repeated 1,000,000 times.

Fig. 1 shows the histograms of the generated random numbers z by SIM and KRM. We illustrated the only result of z_4 in Fig. 1 because the histograms of z_1, z_2, and z_3 were close to the same as that of z_4, regardless of the generation algorithm. According to Fig. 1, the histograms of z_4 by SIM and KRM were quite fitted to the probability function of the univariate Poisson distribution with $\lambda = 2$.

In addition, Tab. 1 shows the marginal mean and variance of z and the correlation coefficient between z_i and z_j by SIM and KRM. As shown in Tab. 1, the mean, variance, and correlation coefficient of the generated random numbers were quite similar to the predefined theoretical values, that is, $\lambda = 2$ and $\rho_{ij} = 0.4^{j-i}$, regardless of the algorithm.

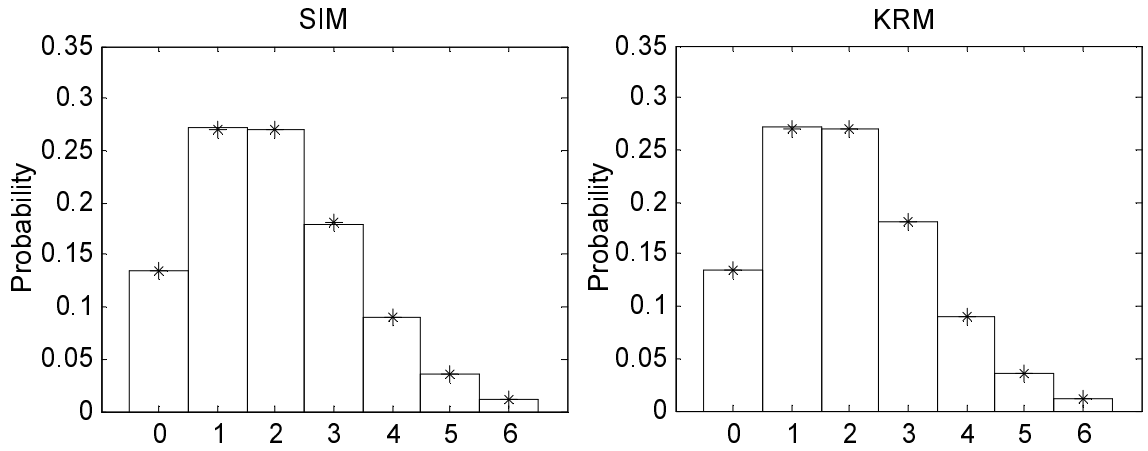

Figure 1: Histograms of the generated random numbers z_4 by SIM and KRM. Here, * refers to the probability function of the univariate Poisson distribution with $\lambda = 2$.

Table 1: Summary statistics for the generated random numbers by SIM and KRM. Top and bottom rows of diagonal elements are the mean and variance of z, respectively. Off-diagonal elements are the correlation coefficient between z_i and z_j.

	SIM					KRM				
	z_1	z_2	z_3	z_4		z_1	z_2	z_3	z_4
z_1	2.00	0.400	0.160	0.063	z_1	2.00	0.402	0.161	0.064
	2.00					2.00			
z_2		2.00	0.402	0.159	z_2		2.00	0.400	0.160
		2.00					2.00		
z_3			2.00	0.400	z_3			2.00	0.400
			2.00					2.00	
z_4				2.00	z_4				2.00
				2.00					2.00

4 Restrictions of the algorithms

In Sect. 4, we reviewed the properties of the constraints for the algorithms by Sim (1993) and Krummenauer (1998).

We graphically presented the limitations of SIM and KRM in terms of the upper bound within the range constraint for ρ_{ij} of multiple Poisson distribution. In fact, we evaluated the upper bounds within the range constraint of the correlation parameter ρ with SIM and KRM, because we assumed that the true correlation structures were exchangeable structure with $\rho_{ij} = \rho$ and AR(1) structure with $\rho_{ij} = \rho^{j-i}$. In this investigation, we considered $\rho \geq 0$. We set the

two scenarios: (i) the relationship between the upper bound of ρ and the ratio of λ_k to λ_1 (i.e., λ_k/λ_1) for $k = 2, \ldots, p$, where $\lambda_k = \lambda$, and (ii) the relationship between the upper bound of ρ and the slope β_1 derived from $\log(\lambda_t) = \beta_0 + \beta_1 t$ for $t = 1, \ldots, p$. Each algorithm was applied with $p = 3, 5$. We also assumed that the dimension of I is less than 2 when KRM is applied.

The upper bounds within the range constraint of ρ with SIM and KRM for Scenarios 1 and 2 are shown in Figs. 2 and 3, respectively. According to Figs. 2 and 3, the upper bound within the range constraint of ρ with SIM was higher rather than that with KRM in almost all settings. In addition, the constraint of KRM was much stricter than that of SIM and with increasing the dimension of the multivariate Poisson random numbers. In particular, KRM was not available in the case when ρ was more than approximately 0.6 and $p \geq 3$.

In Fig. 2, the upper bounds of ρ with SIM and KRM were significantly dependent on λ_i/λ_1, regardless of the true correlation structure. Moreover, the upper bounds of ρ with SIM and KRM were generally stricter as p increases, except the case of when SIM was applied and the true correlation structure was exchangeable. The upper bounds of ρ with SIM and KRM did not completely depend on the magnitude of λ itself, regardless of the true correlation structure and p (results not shown).

In Fig. 3, the upper bounds of ρ with SIM and KRM were quite dependent on β_1, regardless of the true correlation structure. In addition, the upper bounds of ρ with SIM and KRM were generally stricter as p increased. The range bounds of ρ with SIM and KRM were completely independent of the magnitude of β_0 itself, regardless of the true correlation structure and p (results not shown). Therefore, we conclude that the SIM algorithm is better than the KRM one in many settings.

5 Conclusions

We have reviewed the two algorithms proposed by Sim (1993) and Krummenauer (1998) to generate the random vectors of the multivariate Poisson distribution with non-identity covariance matrix and represented the range constraints of the correlation parameter with SIM and KRM. Based on our review, we conclude that the application settings of SIM should be more convenient than that of KRM, because the range constraint of KRM is much stricter than that of SIM. We therefore recommend the application of SIM for generating the multivariate correlated Poisson random numbers.

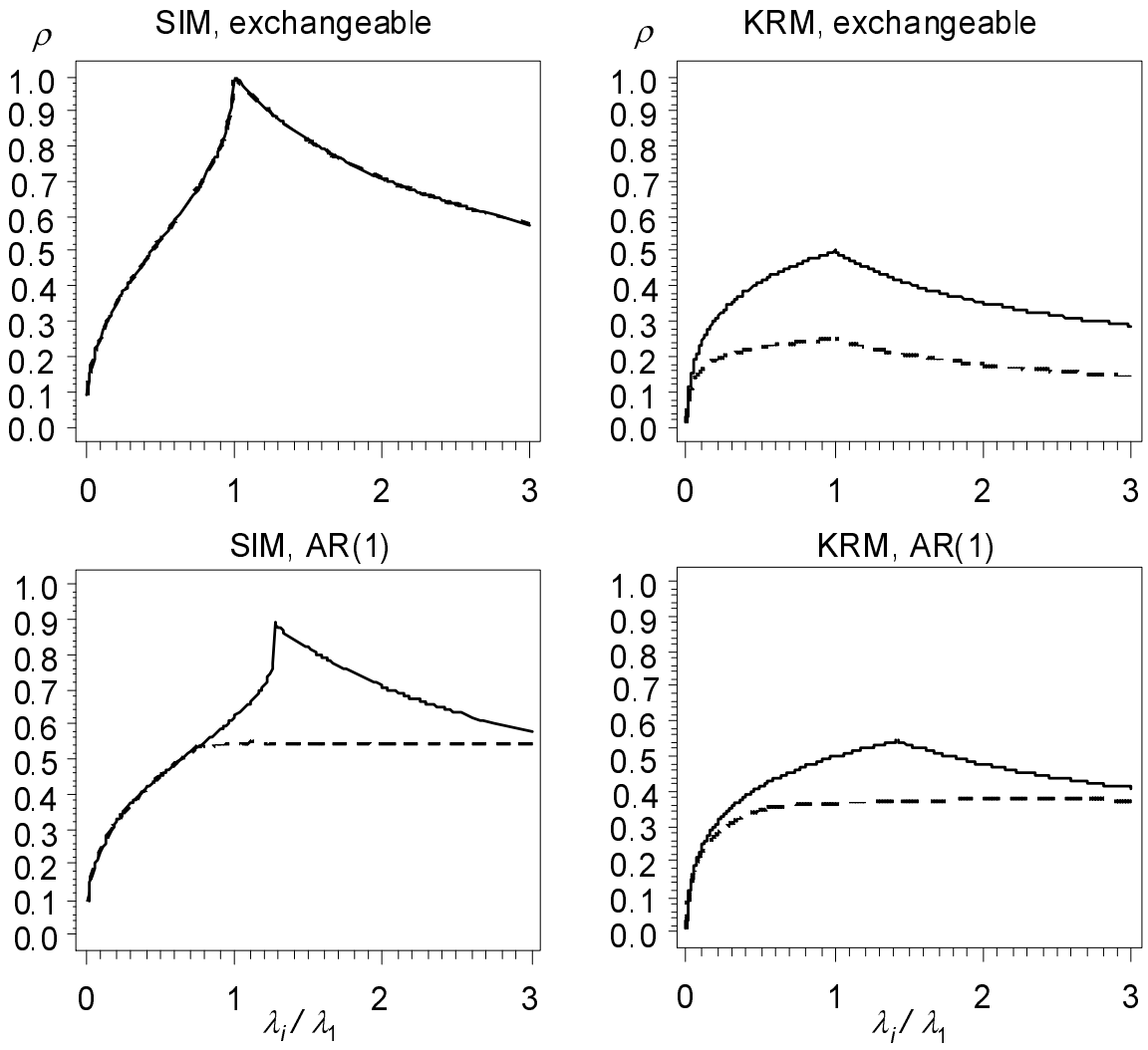

Figure 2: Upper bounds within the range constraint of ρ in Scenario 1. Here, solid and dashed lines refer to the upper bounds within the range constraints of ρ with $p = 3$ and $p = 5$, respectively.

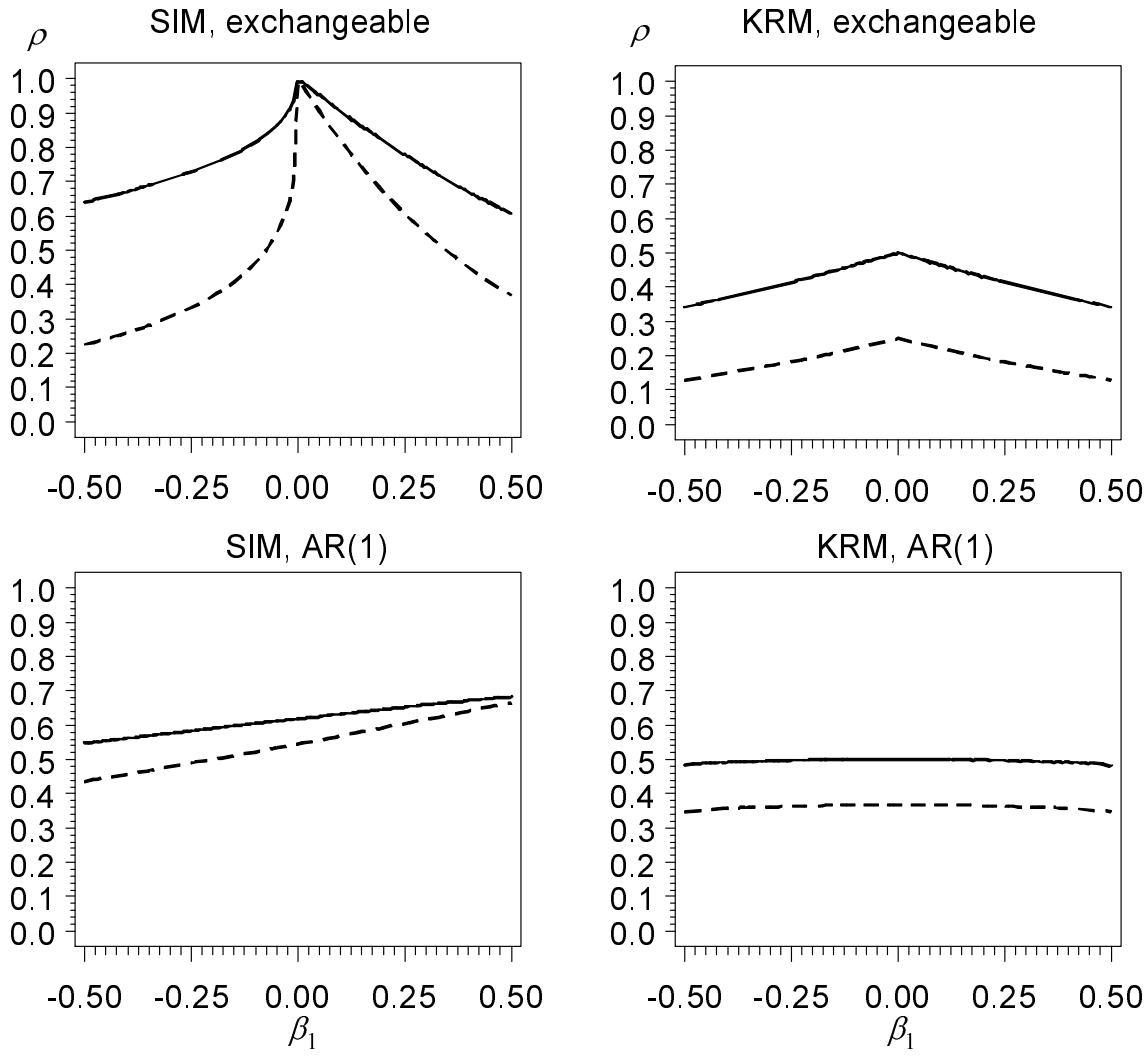

Figure 3: Upper bounds within the range constraint of ρ in Scenario 2. Here, solid and dashed lines refer to the upper bounds within the range constraints of ρ with $p = 3$ and $p = 5$, respectively.

Appendix: SAS programs

Sim's algorithm

```
proc iml ;
lambda={4,2,2}; /*Poisson parameter vector*/
rho=0.3;        /*correlation parameter*/
ex=1;           /*0:AR(1), 1:Exchangeable*/
n=100;          /*sample size*/
seed0=12345;    /*seed*/
p=nrow(lambda) ;
if ex=1 then R=(1-rho)*I(p)+rho*J(p) ;
else R=toeplitz(rho##(0:p-1));
Sigma=R#(lambda##0.5*t(lambda##0.5));
mu=j(p,1,0); alpha=j(p,p,0); mu[1]=Sigma[1,1];
do j=2 to p;
  alpha[j,1]=Sigma[1,j]/mu[1];
  if j>2 then do;
  do i=2 to j-1;
    alpha[j,i]=(Sigma[i,j]-sum(alpha[i,1:i-1]#alpha[j,1:i-1]#t(mu[1:i-1])))/mu[i];
  end;
  end;
  mu[j]=Sigma[j,j]-sum(alpha[j,1:j-1]#t(mu[1:j-1]));
end;
X=j(n,p,0); Z=j(n,p,0); X[,1]=ranpoi(J(n,1,seed0),mu[1]); Z[,1]=X[,1];
do j=2 to p;
  sum=J(n,1,0);
  do i=1 to j-1;
    sumi=J(n,1,0);
    sumi[loc(X[,i]>0),]=ranbin(0,X[loc(X[,i]>0),i],alpha[j,i]*j(nrow(loc(X[,i]>0)),1,1));
    sum=sum+sumi;
  end;
  X[,j]=ranpoi(J(n,1,seed0),mu[j]); Z[,j]=sum+X[,j];
end;
create data from Z; append from Z;
quit;
```

Krummenauer's algorithm

```
proc iml;
lambda={4,2,2}; /*Poisson parameter vector*/
rho=0.3;        /*correlation parameter*/
ex=1;           /*0:AR(1), 1:Exchangeable*/
n=100;          /*sample size*/
seed0=12345;    /*seed*/
p=nrow(lambda) ;
if ex=1 then R=(1-rho)*I(p)+rho*J(p) ;
else R=toeplitz(rho##(0:p-1));
Sigma=R#(lambda##0.5*t(lambda##0.5));
Z=j(n,p,0); ZZ=j(n,p**2,0); mu=Sigma;
do i=1 to p;
  do j=1 to p;
    if i^=j then mu[i,i]=mu[i,i]-Sigma[i,j];
  end;
  do j=1 to p;
    if i<=j then ZZ[,(i-1)*p+j]=ranpoi(J(n,1,seed0),mu[i,j]);
    else ZZ[,(i-1)*p+j]=ZZ[,(j-1)*p+i];
  end;
  do j=1 to p;
    Z[,i]=Z[,i]+ZZ[,(i-1)*p+j];
  end;
end;
create data from Z; append from Z;
quit;
```

References

[1] Diggle PJ, Heagerthy P, Liang KY, Zeger SL (2002) Analysis of longitudinal data, 2nd edn. Oxford, Oxford University Press

[2] Farrell PJ, Stewart KR (2008) Methods for generating longitudinally correlated binary data. Int Stat Rev 76:28–38

[3] Hardin J, Hilbe J (2003) Generalized estimating equations. London, Chapman and Hall

[4] Hedeker D, Gibbons RD (2006) Longitudinal data analysis. New Jersey, Wiley

[5] Holgate P (1964) Estimation for the bivariate Poisson distribution Biometrika 51 241–245

[6] Johnson NL, Kotz S, Balakrishnan N (1997) Discrete multivariate distributions, New York, John Wiley

[7] Krummenauer F (1998) Efficient simulation of multivariate binomial and Poisson distributions. Biometrical J 40:823–832

[8] McCulleagh P, Nelder JA (1990) Generalized linear models, 2nd edn. London, Chapman and Hall

[9] Sim CH (1993) Generation of Poisson and gamma random vectors with given marginals and covariance matrix. J Stat Comput Sim 47:1–10