ISBN 979-99168-0-1

Proceedings of

Proceedings of

The International Conference on Statistics and Mathematics and Its Applications in the Development of Science and Technology

Bandung, October 4 - 6, 2004

published by
FACULTY OF MATHEMATICS AND NATURAL SCIENCE
BANDUNG ISLAMIC UNIVERSITY
BANDUNG, WEST JAVA - INDONESIA

Table of Contents

PAPERS FOR PARTICIPANTS:

International Conference on Statistics and Mathematics and Its Application in the Development of Science and Technology

SAS MACROS FOR GENERATING DEPENDENT COMPETING RISKS DATA WITH EXPONENTIALLY DISTRIBUTED MARGINAL CAUSE OF FAILURE

Abdul Kudus[1] and Noor Akma Ibrahim[2]

[1]*Department of Statistics, Universitas Islam Bandung, Indonesia*
akudus69@yahoo.com

[2]*Institute for Mathematical Research and Department of Mathematics, University Putra Malaysia, Serdang 43400, Selangor, Malaysia*
nakma@inform.upm.edu.my

Univariate exponentially distribution data may be easily generated using most statistical software packages. For instance, SAS uses an inverse transform method applied to a random variate from the uniform distribution. This is not as simple as generating multivariate data with exponential marginal.

For generating multivariate dependent competing risks data with exponentially marginal distribution of each cause of failure time, we can modify algorithm provided by London and Gennings (London, W. B. and C. Gennings, 1999. *Commun. in Stat. : Simula. and Comput.* 28(2):487-500). This paper proposed to modify that algorithm and implement it by using SAS macros.

Keywords: SAS macros, generating multivariate data, competing risks.

1. Introduction

In the competing risk setting, an individual is exposed to several risks at the same time, but eventual failure of the individual is due to only one of these risks, which is called a cause of failure.

One formulation of the competing risk model is in terms of conceptual or latent failure times for each failure type (Cox, 1959). It assumes that the competing risks are independent of each other. This approach has been critized on the basis of unwarranted assumptions, lack of physical interpretation and identifiability problems.

Alternatively, Prentice *et al.* (1978) proposed cause-specific hazard rates, and showed that they were the basic estimable quantities in the competing risks framework. The competing risk may be dependent on each other.

Under this framework, suppose each failure of individual can be identified as one of p (p>1) possibly dependent causes of failure. In other words, each individual is subject to p distinct risks referred to as competing risks threatening its life. Associated with cause j, there is a nonnegative absolutely continuous random variable T_{ji} representing the lifetime of individual i when no other potential risks are present. So, each individual has a latent vector $\boldsymbol{T}_i = (T_{1i}, \ldots, T_{pi})$. Suppose that vector $\boldsymbol{T}_i$ was distributed as multivariate gamma with marginal exponential. Actually the termination time of an individual is defined as the time to the first failure. Thus, lifetime of an individual i is given by $T_i = \min\{T_{1i}, \ldots, T_{pi}\}$. The available information is usually given by the pair (T_i,Δ_i), where Δ_i indicates the cause(s) of failure. Censored data was specified by zero value of Δ_i.

By assuming multivariate gamma distribution for p latent failure times with marginally exponential distributed, we can set up a dependent competing risk model. For this, London and Gennings (1999) proposed an algorithm to generate data with the desired distribution. Using multivariate normal data to generate Wishart matrix, and then extracting the multivariate gamma vector as dependent latent for competing risk data. Only the correlation structure of the multivariate normal data must be specified in order to generate multivariate gamma vector with the desired mean and variance-covariance structure.

2. The Theory Of Simulation Of Multivariate Gamma Data With Exponential Marginals

2.1. *Notation*

Let the time points be denoted by t; for each time point, denote the censoring indicator by Δ. We will generate n individuals where each individual will contain p elements representing p types of failure time. Therefore, the p elements may be dependent, although the data will be independently distributed across individuals.

2.1.1 Generate One Individual

This theory is provided by London and Gennings (1999). We reformulate some notation for simplification. Let's proceed to generate the gth individual of multivariate gamma data, $g = 1,\ldots,n$. To do so, we begin with multivariate normal data, convert this to Wishart data, and then extract the multivariate gamma vector, where the marginals will be exponential, a special case of the gamma. We know that the diagonal elements of a Wishart matrix are distributed as multivariate gamma(κ,τ) (Johnson and Kotz, 1972). We begin by generating a Wishart matrix A_g ($p \times p$), $A_g = X_g'X_g \sim W_p(r,\Sigma_g)$, where X_g is ($r \times p$), $vec(X_g) \sim N(0, I_r \otimes \Sigma_g)$, i.e. one row of X_g is $\underline{x}_{gi} = (x_{gi1}, \ldots, x_{gip}) \sim N_p(\underline{0}, \Sigma_g)$ for $i=1,\ldots, r$ rows (Johnson and Wichern 1992) and

$$\Sigma_g = \begin{bmatrix} \sigma_{g11} & \cdots & \sigma_{g1p} \\ \vdots & \ddots & \\ \sigma_{gp1} & & \sigma_{gpp} \end{bmatrix} \tag{1}$$

So the diagonal elements of A_g are $(a_{g11}, \ldots, a_{gpp}) \sim$ multivariate gamma(κ,τ), where $a_{gjj} = \sum_{i=1}^{r} x_{gij}^2$, $j=1,\ldots, p$. In the marginal, if $\kappa=1$ and $\tau=\lambda$, then $(a_{g11}, \ldots, a_{gpp}) \sim$ exponential(λ) (Rothschild and Logothetis 1986). Now we must find the correct approach to use in setting up the multivariate normal data in order to get $\kappa=1$. We know that the chi-square distribution is a special case of the gamma, i.e., if $y \sim \text{gamma}\left(\kappa = \frac{v}{2}, \tau = \frac{1}{2}\right)$, then $y \sim \chi_v^2$ (Evans *et al.*, 1993). Let $r = 2$, then $\sum_{i=1}^{2}\left(\frac{x_{gij}}{\sqrt{\sigma_{gjj}}}\right)^2 \sim \chi_2^2$, $j=1,\ldots,p$.

Therefore, $\sum_{i=1}^{2}\left(\frac{x_{gij}}{\sqrt{\sigma_{gjj}}}\right)^2 \sim \text{gamma}\left(\kappa = \frac{2}{2} = 1, \tau = \frac{1}{2}\right)$, $j=1,\ldots,p$ for $r=2$, i.e., $v=2$. So $r=2$ will give us $\kappa=1$ and therefore $(a_{g11}, \ldots, a_{gpp}) \sim$ multivariate gamma($1,\tau$), and exponential(λ) in the marginal. This vector of multivariate gamma data represents one individual of p elements.

The ($p\times p$) covariance matrix Σ_g^* may be determined for the gth multivariate gamma vector starting from the covariance matrix Σ_g of the original multivariate normal data. We know that for Wishart matrix A_g,

$$Cov(a_{gij}, a_{gkl}) = r(\sigma_{gik}\sigma_{gjl} + \sigma_{gil}\sigma_{gjk}),\ i,j,k,l = 1, \ldots, p$$

(Muirhead, 1982). Since we are only concerned with the diagonal elements of A_g,

$$Cov(a_{gii}, a_{gjj}) = r(\sigma_{gij}\sigma_{gij} + \sigma_{gij}\sigma_{gij}),\ i,j = 1, \ldots, p$$

$$Cov(a_{gii}, a_{gjj}) = 2r\sigma_{gij}^2, \quad i,j = 1,\ldots,p \tag{2}$$

Then

$$\Sigma_g^* = 2r\begin{bmatrix} \sigma_{g11}^2 & \cdots & \sigma_{g1p}^2 \\ \vdots & \ddots & \\ \sigma_{gp1}^2 & & \sigma_{gpp}^2 \end{bmatrix} \tag{3}$$

is the covariance matrix of the gth multivariate gamma random vector, where

$$Var(a_{gii}) = 2r\sigma_{gii}^2 \tag{4}$$

The correlation matrix R_g ($p \times p$) of this multivariate gamma vector can then be determined. Let the (i,j)th elements of R_g be ρ_{ij}, where using (2) and (4),

$$\rho_{gij} = \frac{Cov(a_{gii}, a_{gjj})}{\sqrt{Var(a_{gii})}\sqrt{Var(a_{gjj})}} = \frac{2r\sigma_{gij}^2}{\sqrt{2r\sigma_{gii}^2}\sqrt{2r\sigma_{gjj}^2}} = \frac{\sigma_{gij}^2}{\sigma_{gii}\sigma_{gjj}}, \quad i,j = 1,\ldots,p \tag{5}$$

If we decided that we'd like R_g to have an *exchangeable correlation* structure, then for $i \neq j$, $\rho = \alpha$, i.e.,

$$R_g(\alpha) = \begin{bmatrix} 1 & \alpha & \cdots & \alpha \\ \alpha & 1 & \cdots & \alpha \\ \vdots & \vdots & \ddots & \vdots \\ \alpha & \alpha & \cdots & 1 \end{bmatrix} \tag{6}$$

So α is the value of every off-diagonal element of R_g in the case of an exchangeable correlation structure.

We may now think of the a_{gii} as the t_{gii} time points in a survival analysis, i=1,...,p, some of which are censored and the balance of which are failures. The mean, variance, and α for these gth individual survival times are related to σ_{gii} and the σ_{gij} as elements of Σ_g in equation (1). This means, we can generate t_{gii} with the desired mean and variance by controlling the value σ_{gii}, σ_{gij} and α as will be shown below.

If marginal distribution of t_{gii} is exponential(λ), so its mean and variance are $1/\lambda$ and $(1/\lambda)^2$ respectively. This means, we only need to input "mean" to calculate "variance" and use it to solve σ_{gii} in (4), where r=2.

$$\sigma_{gii} = \sqrt{\frac{Var(t_{gii})}{4}} \qquad (7)$$

To obtain σ_{gij} from (5), with $\rho_{gij} = \alpha$ (off-diagonal elements), and $\sigma_{gii} = \sigma_{gjj}$ (diagonal elements), we use the following equation

$$\sigma_{gij} = \sigma_{gii}\sqrt{\alpha} \qquad (8)$$

Algorithm *Generating one individual dependent competing risk data.*
Input mean, correlation (α) and number of failure type (p) for desired generated marginal exponential data.
Calculate σ_{gii} and σ_{gij} using equation (7) and (8) respectively.
Construct Σ_g of equation (1)
Generate 1 observation from the 2p-normal multivariate distribution with mean 0 and covariance matrix $I_2 \otimes \Sigma_g$. Take the first p elements as first row of matrix X and the others as second row of X.
Construct $A_g = X_g' X_g$, where $X_g = \begin{pmatrix} x_{g11} & \cdots & x_{g1p} \\ x_{g21} & \cdots & x_{g2p} \end{pmatrix}$
Extract diagonal elements of A_g. Let it's $T_g = (t_{g11}, \ldots, t_{gpp})$ as sample from latent multivariate gamma with marginal exponential.
Pick the minimum of T_g as the observed failure time. If the position of its minimum is at jth element, then its failure type is j, where j=1,...,p.
Generate binary number from Bernoulli distribution with q probability of failure, where 100q is percentage of censoring. If resulted generated binary number is zero, then the failure time is censored, otherwise the individual is failure with jth failure type (keep result from step 7).

Notice that we employ the censoring independent from failure time data generation. After generating failure time with its failure type in step 7, we proceed to decide whether it was censored or not through generating binary number in step 8.

2.1.2 Generate n Individual

For generating n individual dependent competing risk data, we can repeat step 4 to 8 n times by using input from step 1. The multivariate normal data generation in step 4 is employed just like the macro MVN in SAS™, so that this macro can be developed in an effective way.

3. Implementation of The Theory and Application

The above algorithm is implemented in SAS Macros using the name `depcrEx` which stands for "dependent competing risk with marginal Exponential" with six input parameters and one output parameters. The six input parameters are `corr` (correlation between failure time, α in equation (8)), `means` (mean of exponential distribution, λ), `p` (number of variate in multivariate gamma and also number of type of failure), `nos` (number of observations to be generated), `seed` (starting seed value for the random number generator), `percent` (censoring percentage), and the only output is `scmprsk` , the "sample competing risk".

For example the syntax :

```
%depcrEx(corr=0.7,means=1,p=3,seed=210369,nos=10000,percent=50,scmprsk=expo);
```

is for generating dependent competing risk data with marginal exponential for correlation between failure time α=0.7, mean of each exponential variate equal to 1, number of failure type was 3, 10000 number of observations, using starting seed 210369, 50% of censoring and output data is assigned to `expo.sd2`.

Table 1. Summary of Generated Data

	Failure Type		
	1	2	3
Observed (freq)	1589	1633	1688
Censored (freq)	1709	1712	1669
Total Censored (freq)	5090		
$\hat{\lambda}$	0.9463	1.0878	1.1493
Mean	1.0567	0.9193	0.8701

Generated data summary in Table 1 showed that the censoring percentage which is closed to 50%, and all maximum likelihood estimated sample mean are closed to 1. We can further explore the generated data by employing cumulative incidence function (CIF) estimation. We use SAS macros `comprisk` from Bergstralh (2000) for CIF estimation.

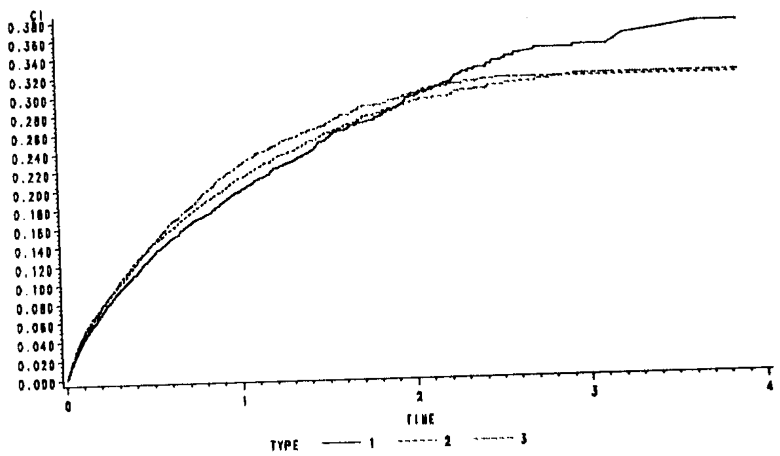

Fig. 1. Cumulative Incidence Function for Three Failure Types

CIF for all three failure types have almost the same pattern. This figure confirmed the equality of marginal exponential for all failure type as supposed to be.

4. Conclusion and Discussion

Based on maximum likelihood estimation result and CIF estimation results, we conclude that the simulated dependent competing risk data have the means and distribution that were intended.

The exponential distribution was chosen because exponential models have been repeatedly used as parametric models for failure time data. The example presented here made use of an exchangeable correlation structure, but other structures could also be possible.

5. Reference

Bergstralh, E. 2000. *%COMPRISK macro: Estimation of cumulative incidence (CI) in presence of competing risks.* URL `www.mayo.edu/hsr/sasmac.html`.

Cox, D.R. 1959. The analysis of exponentially distributed lifetimes with two types of failure. *J. Roy. Statist. Soc. B.* 21:411-421.

Evans, M., Hastings, N. and Peacock, B. 1993. *Statistical Distributions.* New York:Wiley.

Johnson, N.L. and Kotz, S. 1972. *Distributions in Statistics: Continuous Multivariate Distributions*, New York: Wiley.

Johnson, R.A. and Wichern, D.W. 1992. *Applied Multivariate Statistical Analysis.* Englewood Cliffs, NJ: Prentice-Hall.

London, W.B and Gennings, C. 1999. Simulation of multivariate gamma data with exponential marginals for independent clusters. *Commun. in Stat. : Simul. and Comput.* 28(2):487-500.

Muirhead, R.J. 1982. *Aspect of Multivariate Statistical Theory.* New York:Wiley

Prentice, R.L., Kalbfleisch, J.D., Peterson, A.V., Fluornoy, N., Farewell, V.T., and Breslow, N.E. 1978. The analysis of failure times in the presence of competing risks. *Biometrics.* 34:541-554.

Rothschild, V. and Logothetis, N. 1986. *Probability Distributions.* New York: Wiley.

SAS Institute Inc. 1996. *%MVN macro: Generating multivariate normal data.* URL `http://ftp.sas.com/techsup/download/stat/mvn.html`.

SAS Institute Inc. 1997. SAS® *Macro Language: Reference, First Editions.* Cary, NC: SAS Institute Inc.

SAS Institute Inc. 1989. SAS/IML® *Software: Usage and Reference, Version 6, First Edition.* Cary, NC: SAS Institute Inc.