Journal of Korean
Data & Information Science Society
2006, Vol. 17, No. 1, pp. 161 ~ 171

On the Multivariate Poisson Distribution with Specific Covariance Matrix

Daehak Kim[1] · Heong Chul Jeong[2] · Byoung Cheol Jung[3]

Abstract

In this paper, we consider the random number generation method for multivariate Poisson distribution with specific covariance matrix. Random number generating method for the multivariate Poisson distribution is considered into two part, by first solving the linear equation to determine the univariate Poisson parameter, then convoluting independent univariate Poisson variates with appropriate expectations. We propose a numerical algorithm to solve the linear equation given the specific covariance matrix.

Keywords : 공분산행렬, 다변량 이항분포, 다변량 포아송분포, 랜덤넘버, 선형방정식의 해, 의존적 구조

1. 서 론

포아송분포는 발생빈도가 드문 희귀한 사건의 횟수 등 다양한 현상에서 발견되며 생물학 등 제 학문분야에 활용성이 높다고 할 수 있다. 특히, 교통사고 사망자 발생 현황이나 전염성 질환 발생 현황 등은 포아송분포를 따른다고 하겠다.

이제, k개의 포아송 확률변수가 각각 포아송분포 $P(\lambda_i), i=1,\cdots,k$를 따르며, 서로 복잡한 상관구조를 형성한다고 하자. 포아송분포와 같은 이산분포가 다변량 구조를 형성하면, 많은 수의 모수 때문에 확률계산, 모수추정 및 난수발생 등 통계적 추론에 상당한 어려움이 따를 수 있다. 다변량 포아송분포의 통계적 추론에 있어서 Tsiamyrtzis와 Karlis(2004)는 다변량 포아송 확률을 계산하는 알고리즘을 제안한 바 있다. 그런데, 무엇보다도 우선되는 것은 다변량 포아송분포와 그의 모수에 대한 정확한 정의가 요구된다고 하겠다. 그러므로 본 논문에서는 먼저 Krummenauer(1998a,b)

1) 제 1 저자 : 교수, 경북 경산시 하양읍 대구가톨릭대학교 정보통계학과
E-mail : dhkim@cu.ac.kr
2) 조교수, 경기도 화성시 봉담읍 수원대학교 통계정보학과
E-mail: jhc@suwon.ac.kr
3) 조교수, 서울시 동대문구 전농동 서울시립대학교 통계학과,
E-mail: bcjung@empal.com

의 k 변량 베르누이 분포로부터 유도된 다변량 포아송분포를 소개하고, 다변량 포아송분포를 표현하는 모수의 특징을 살펴보고자 한다. 이를 토대로 본 논문에서는 Krummenauer(1998a)와 Karlis(2003)에 의해 정의된 일반적 다변량 포아송분포로 부터 난수를 생성하는 방법을 다루고자 한다. 이를 위해 선형방정식의 해를 구하는 수치해석적 알고리즘을 제안하고, Park 등(1996)의 다변량 베르누이 난수 생성에 활용된 알고리즘과의 연관성을 다루었다. 무엇보다도, 본 논문에서는 3 변량 이상의 관계를 설명하는 공통모수가 주어지지 않은 상황에서, 두 변수들 간의 분산-공분산 행렬을 만족하는 다변량 포아송 난수를 생성하는 방법을 중심으로 다루고자 한다.

2장에서는 다변량 포아송분포의 일반적 모수에 대해, 3장에서는 주어진 모수 행렬로부터 난수를 생성하는 방법에 대해 소개하였다. 그리고 4장에서는 특별한 분산-공분산 행렬이 주어졌을 때, 이를 만족하는 난수를 생성하는 알고리즘을 제안하고 5장에서는 4변량 포아송분포의 경우에 난수 발생의 예를 설명하였다. 마지막으로 6장에서는 결론 및 토론을 기술하였다.

2. 다변량 포아송분포

2.1 k변량 포아송분포

$B_1, \cdots, B_k$가 각각 베르누이분포 $B_i \sim B(p_i), i = 1, \cdots, k$를 따른다고 하자. 여기서 $p_1, \cdots, p_k$는 베르누이 주변 확률분포의 성공확률을 의미한다. 자연수 $1, 2, \cdots, k$를 원소로 갖는 집합을 $K = \{1, 2, \cdots, k\}$로 나타내자. 이제 집합 K의 부분집합 $I \subseteq K (I \neq \varnothing)$에 대해 $p_I = P(B_i = 1, \forall i \in I)$를 정의하자. K의 임의의 부분집합 $\{i_1, i_2, \cdots, i_l\} (l \leq k)$에 대해 첨자로 표현된 $p_{i_1, i_2, \cdots, i_l}$과 $p_{\{i_1, i_2, \cdots, i_l\}}$그리고 $p_{i_1 i_2 \cdots i_l}$등은 같은 의미라 하자. 또한 $\{i_1, i_2, \cdots, i_l\} (l \leq k)$의 모든 순열 $j_1, j_2, \cdots, j_l$에 대해서도 $p_{i_1, i_2, \cdots, i_l} = p_{j_1, j_2, \cdots, j_l}$이라 놓자. 그러면 k변량 베르누이분포에 대한 모수 집합은 다음과 같이 표현된다.

$$P = \{p_I \mid \forall I \subseteq K, I \neq \varnothing\}.$$

예로서 $k = 3$이면, I는 $\{1\}, \{2\}, \{3\}, \{1, 2\}, \{1, 3\}, \{2, 3\}, \{1, 2, 3\}$등의 7개의 집합중의 하나이며 $p_{12} = p_{21} = p_{\{1,2\}}$등의 관계로부터 모수집합은 $P = \{p_1, p_2, p_3, p_{12}, p_{13}, p_{23}, p_{123}\}$로 표현된다.

이제, $\boldsymbol{X}^{(1)}, \cdots, \boldsymbol{X}^{(n)}$을 서로 독립인 k변량 베르누이분포를 따르는 확률변수라 두면 $\boldsymbol{X}^{(1)} + \cdots + \boldsymbol{X}^{(n)}$은 k변량 이항분포를 따르며, 이를 다변량 이항분포라 부른다. Krummenauer(1998a)는 이를 $MVB_k(n, \boldsymbol{p})$라 표현한 바 있다. 이때 $\boldsymbol{p} = (p_1, p_2, \cdots, p_k)$이다. 이제, 집합 $\Lambda = \{\lambda_I \mid \forall I \subseteq K, I \neq \varnothing\}$에 대해 Krummenauer(1998a)는 $n\boldsymbol{p} \rightarrow \Lambda$를 가정하면 $MVB_k(n, \boldsymbol{p})$는 주변 확률분포가 포아송

분포 $P(\lambda_i), i=1,\cdots,k$를 따르는 다변량 포아송분포 $MVP(\Lambda)$로 확률 약수렴함을 밝힌바 있다. 또한 여기서 $MVP(\Lambda)$의 확률생성함수는

$$f(s_1, s_2, \cdots, s_k) = \exp\{\sum_{I \subseteq K, I \neq \varnothing} \lambda_I \prod_i (s_i - 1)\}$$

로 표현됨을 밝혔다.

이와 같은 표현을 이용하면 $I \subseteq K(I \neq \varnothing)$에 대한 모수집합 Λ의 원소 λ_I중에서 λ_{i_1}은 각 포아송분포의 주변 확률분포의 모수를, $\lambda_{i_1 i_2}$는 두 변수의 공분산 관계를, $\lambda_{i_1 i_2 \cdots i_k}$은 k개 포아송변수의 공통관계를 나타낸다고 하겠다.

2.2 k변량 포아송분포의 의존적 구조

Dwass와 Teicher(1956)는 k변량 포아송분포를 3변량 축소방법(trivariate reduction method)으로 표현한 바 있다. 이제, 이변량 포아송분포에 대한 성질을 기초로 하여 자연스럽게 k변량 포아송분포의 일반적인 구조를 살펴보기로 하자.

다음의 확률변수

$$\begin{aligned} X_1(\lambda_1) &= Y_1(\mu_1) + Y_0(\mu_0) = Y_1(\lambda_1 - \lambda_{12}) + Y_0(\lambda_{12}) \\ X_2(\lambda_2) &= Y_2(\mu_2) + Y_0(\mu_0) = Y_2(\lambda_2 - \lambda_{12}) + Y_0(\lambda_{12}) \end{aligned}$$

를 생각하자. 여기서 각 $Y_i, i=0,1,2$는 포아송 모수 $\mu_i, i=0,1,2$인, 서로 독립인 포아송 확률변수이다. 그러면, X_1, X_2는 각각 모수 λ_1, λ_2와 $Cov(X_1, X_2) = \mu_0 = \lambda_{12}$를 갖는 이변량 포아송분포를 따르게 된다. $E(X_1 X_2) = \lambda_1\lambda_2 + \lambda_{12}$이므로 $\rho_{12} = \lambda_{12}/\sqrt{\lambda_1}\sqrt{\lambda_2}$가 되고 따라서, $\lambda_{12} = \rho_{12}\sqrt{\lambda_1}\sqrt{\lambda_2}$가 된다. 그러나 $\lambda_{12} \leq \lambda_1, \lambda_2$이며, 모든 $\lambda_I \geq 0, I \subseteq \{i,j\}$이므로, 상관계수는

$$0 \leq \rho_{12} \leq \frac{\min(\sqrt{\lambda_1}, \sqrt{\lambda_2})}{\max(\sqrt{\lambda_1}, \sqrt{\lambda_2})}$$

의 범위에서 결정된다.

이제, 이변량 포아송분포의 의존적 구조를 자연스럽게 k변량의 경우로 확장하자. k개의 확률변수

$$\begin{aligned} X_1(\lambda_1) &= Y_1(\mu_1) + Y_0(\mu_0) = Y_1(\lambda_1 - \lambda_{ab}) + Y_0(\lambda_{ab}) \\ X_2(\lambda_2) &= Y_2(\mu_2) + Y_0(\mu_0) = Y_2(\lambda_2 - \lambda_{ab}) + Y_0(\lambda_{ab}) \\ &\cdots, \\ X_k(\lambda_k) &= Y_k(\mu_k) + Y_0(\mu_0) = Y_k(\lambda_k - \lambda_{ab}) + Y_0(\lambda_{ab}) \end{aligned} \tag{2.1}$$

를 고려하여 보자. $X_1, X_2, \cdots, X_k$는 $k+1$개의 모수가 있는 다변량 포아송분포라 할 수 있다. 여기서도 두 변수 X_i 와 X_j간의 상관계수의 범위는

$$0 \leq \rho_{ij} \leq \frac{\min(\sqrt{\lambda_i}, \sqrt{\lambda_j})}{\max(\sqrt{\lambda_i}, \sqrt{\lambda_j})}$$

로 결정되며, 모든 변수들 간에 비음(non-negative)인 상관구조를 지니게 된다. 그런데, 위의 $k+1$개 모수를 사용한 표현에 의하면 다변량 포아송분포의 모든 모수를 표현 할 수 없으므로 일반적인 표현이라고 할 수 없다. 일반적인 다변량 포아송분포의 구조를 표현 하는 방법에 대해 Loukas와 Kemp(1983), Karlis(2003)등을 참고할 수 있다.

예를 들어 3변량 포아송분포인 경우

$$\begin{aligned} X_1 &= Y_1 + Y_{12} + Y_{13} + Y_{123} \\ X_2 &= Y_2 + Y_{12} + Y_{23} + Y_{123} \\ X_3 &= Y_3 + Y_{13} + Y_{23} + Y_{123} \end{aligned} \tag{2.2}$$

로 표현하는 것이 합리적이라 할 수 있다. 여기서, Y_i는 서로 독립인 포아송 확률변수이며 i는 $\{1\},\{2\},\{3\},\{1,2\},\{1,3\},\{2,3\},\{1,2,3\}$이다. (2.2)식을 참고하면, 분산분석의 모형을 표현하는 방법처럼 k변량 포아송분포가 표현된다고 하겠다.

(2.2)식의 표현을 보다 일반화 하여 $\boldsymbol{A} = [A_1, A_2, \cdots, A_k]$의 행렬로 표현하여 보자. 여기서, A_i는

$$k \times \binom{k}{i}$$

의 차원을 지니며, 각 열에 정확히 i개의 1과 $k-i$개의 0이 존재하는 구조의 행렬이다. 이제 서로 독립인 2^k-1 개의 포아송 확률변수 $\boldsymbol{Y} = \{Y_I \mid \forall I \subseteq K, I \neq \varnothing\}$와 $\boldsymbol{X} = (X_1, X_2, \cdots, X_k)'$에 대해 다변량 포아송분포는 Karlis(2003)에 의하면

$$\boldsymbol{X}_{k\times 1} = \boldsymbol{A}_{k\times(2^k-1)}\boldsymbol{Y}_{(2^k-1)\times 1}$$

로 표현할 수 있다. 이를 이용하면 (2.2)식에서

$$\boldsymbol{A} = \begin{pmatrix} 1 & 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 & 1 \end{pmatrix}$$

로 나타난다. 또한 하나의 포아송 난수 X_i를 생성하기 위해 요구되는 독립된 포아송 확률변수 Y_i는 $\sum_{i=1}^{k}\binom{k-1}{i-1}$개 이다.

3. Λ로부터 다변량 포아송 난수를 생성하는 방법

$\boldsymbol{X} = \boldsymbol{A}\boldsymbol{Y}$의 구조 하에서 다변량 포아송 난수를 생성하는 방법과 그의 한계점을 살펴보자. Y_I의 모든 모수 집합을 $M = \{\mu_I \in [0, \infty), I \subseteq K, I \neq \varnothing\}$이라 놓자. 이제,

$$\lambda_I = \sum_{J \subseteq K, J \supseteq I} \mu_J \,, \forall\, I \subseteq K,\ I \neq \varnothing \tag{3.1}$$

로부터 모수집합 Λ가 주어지면

$$\Lambda_{(2^k-1)\times 1} = \boldsymbol{B}_{(2^k-1)\times(2^k-1)} \boldsymbol{M}_{(2^k-1)\times 1}$$

으로 표현되므로 $M = \boldsymbol{B}^{-1}\Lambda$로 Y_I의 모든 모수 μ_I를 계산할 수 있다. 여기서 $\boldsymbol{B}$는 (3.1)식의 계획행렬이며 크기는 $(2^k-1)\times(2^k-1)$이다. 이제, θ_I를 포아송분포 $P(\mu_I)$에서 생성한 난수라 가정하면

$$X_i = \sum_{I \subseteq K, i \in I} \theta_I \,,\ i = 1, \cdots, k \tag{3.2}$$

로 되고 모수 Λ를 따르는 다변량 포아송 난수를 생성할 수 있다. 즉, 다변량 포아송분포의 모든 모수 $\Lambda = \{\lambda_I \mid I \subseteq K,\ I \neq \varnothing\}$가 주어지면, 이에 따라 확률변수 Y_I의 모수 μ_I를 결정할 수 있어 다변량 포아송 난수를 생성할 수 있게 된다. Krummenaure(1998a)는 (3.1)식에서 계획행렬 $\boldsymbol{B}$를 사용하지 않고 후향대치 과정을 통해 모수 μ_I의 계산을 언급하였다. 후향대치를 사용하면, $k = 3$일 때, (3.1)식에서

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

의 순서로 모수 μ_I를 결정한 후 (3.2)식의 방법으로 난수를 생성할 수 있다. 그런데, 이 방법의 가장 큰 문제는 연구자가 모든 모수 $\Lambda = \{\lambda_I \mid \forall\, I \subseteq K, I \neq \varnothing\}$를 미리 정의하여야 한다는 점이다. Karlis(2003)는 다변량 포아송분포에서 모든 모수의 추정은 매우 소모적이며, 상당한 계산적 어려움이 존재한다고 하였다. 또한 연구자가 다변량 포아송분포를 따르는 난수를 생성하고자 할 때, 상관구조를 넘어서는 3변수 이상의 공통 모수값을 정의한다는 점은 모든 제약식을 확인해야 한다는 점에서 상당한 어려움이 따르며, 불필요한 일이라 여겨진다. 일반적인 경우, 연구자는 다변량 정규분포의 난수를 생성하는 방법과 비슷하게, 두 변수의 상관구조와 각 주변 확률분포의 평균 등을 정의했을 때 이에 만족하는 난수를 발생시키는 경우가 많이 요구된다 하겠다.

4. 공분산행렬 구조를 만족하는 다변량 포아송 난수 생성

본 절에서는 각 포아송분포의 평균(분산) $\lambda_i, i=1,\cdots,k$들과 $Cov(X_i, X_j)=\lambda_{ij}$만 주어진 상황에서 이 조건들을 만족시키는 난수 생성 방법을 제안하고자 한다.

주어진 문제의 상황은 k개 변수간의 분산-공분산행렬 $\Sigma=\{\sigma_{ij}\}$가 주어졌으며, 3변량 이상의 관계를 표현하는 모수(예를 들어 λ_{ijk})는 주어져 있지 않은 경우이다. 즉 주어진 모수의 개수는 $k(k+1)/2$이다. 그런데, 우리는 $\Lambda = \boldsymbol{B}M$의 행렬 표현을 이용하여 알려진 모수 Λ_1과 알려지지 않은 모수 Λ_2를 다음과 같이 표현할 수 있다.

$$\begin{pmatrix}\Lambda_1\\ \Lambda_2\end{pmatrix}=\begin{pmatrix}B_{11} & B_{12}\\ 0 & B_{22}\end{pmatrix}\begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix}.$$

여기서 $\Lambda_1=(\lambda_1,\cdots\lambda_k,\lambda_{12},\cdots,\lambda_{(k-1)k})'$이며 $\Lambda_2=\{\lambda_J \mid J\subseteq K, J\supseteq\{i_1,i_2\}\}$, 단 $\{i_1,i_2\}\subseteq K$이다. 이제, 우리는 알려진 Λ_1의 정보만으로 해당 선형방정식을 만족하는 적절한 $\mu_I\geq 0$을 찾을 수 있으면 주어진 분산-공분산 관계를 유지하는 다변량 포아송 난수를 발생 시킬 수 있다. 물론, 유도해야 할 μ_I는 총 2^k-1 개로 $\mu_I\geq 0$을 만족하는 해가 무수히 많이 존재하는 선형방정식임을 쉽게 알 수 있다. 이에 대해, (3.1)의 선형방정식에서 μ_I를 계산하는 다음의 알고리즘을 제안한다.

[선형방정식의 해를 구하는 알고리즘]

[Step 0] 평균(분산) $\lambda_i, i=1,\cdots,k$과 상관계수 ρ_{ij}로부터 $\lambda_{ij}=\rho_{ij}\sqrt{\lambda_i}\sqrt{\lambda_j}$에 의해 분산-공분산 행렬 $\{\sigma_{ij}^{(0)}\}=\Sigma^{(0)}$를 계산한다. $\{\sigma_{ij}^{(0)}\}$를 기초로 초기값 $\lambda_I^{(0)}$를 결정한다. 초기값은 다음의 방법으로 결정한다.

0-1) $\beta_l=\{\sigma_{rs} \mid \min\{\sigma_{ij}\}, 1\leqq i\leqq j\leqq k\}$로 놓는다. 이제, $T_l=\{r,s\}$로 놓으면, T_l은 $\min\{\sigma_{ij}\}$ 가 있는 원소의 위치를 지칭하는 지시집합(index set)을 의미한다. 여기서, 초기 $l=1$이다.

0-2) 초기 선택된 T_l을 포함하는 모수 집합을 $\Lambda_l=\{\lambda_J \mid J\subseteq K, J\supseteq T_l\}$로 놓고, $\Lambda_l=\Lambda_l-\bigcup_{i=1}^{l-1}\Lambda_i$의 집합을 구한 후, $\lambda_I\in\Lambda_l$에 대해 $\lambda_I=\beta_l$로 할당한다.

0-3) $k\leftarrow l+1$로 놓고, 앞의 단계에서 결정된 σ_{rs}를 행렬 $\{\sigma_{ij}\}$에서 제외한 후, 0-1)과 0-2)의 단계를 반복한다.

0-4) 위의 방법을 $l=1,\cdots,k(k+1)/2$까지 반복하여 모든 초기값 $\lambda_I^{(0)}$를 결정한다.

[Step 1] 초기 $\mu_I^{(0)}=\boldsymbol{B}^{-1}\lambda_I^{(0)}$를 결정한다. 여기서 $\mu_I^{(0)}\geq 0$이면, 모든 조건을 만족으로 알고리즘을 빠져나온다.

[Step 2] $n = k-1$로 둔다. 집합 $C(n) = \{i_1, i_2, \cdots, i_n\} \subseteq K$에 대하여

2-1) $\mu_{C(n)} = \lambda_{C(n)} - \sum_{J \subseteq K, J \supseteq C(n)} \mu_J$ 를 계산하여, $\mu_{C(n)} \geqq 0$을 확인한다.

2-2) 특정 $C(n)$에서 $\mu_{C(n)} < 0$ 면 모든 $C(n+1) \supset C(n)$중에서 $\{\max(\mu_i) \mid i \in C(n+1)\}$를 포함하는 집합 $C(n+1)$을 선택한다. 그리고 해당 $C(n+1)$의 $\lambda_{C(n+1)}$에 대해 $\lambda_{C(n+1)} \leftarrow \lambda_{C(n+1)} + \mu_{C(n)}$으로 조정한 후 모든 $C(n+1) \subseteq K$에 대하여 $\mu_{C(n+1)} = \lambda_{C(n+1)} - \sum_{J \subseteq K, J \supseteq C(n+1)} \mu_J$을 다시 계산하고, 2-1)로 돌아간다.

2-3) 모든 $C(n)$에서, $\mu_{C(n)} \geqq 0$면 [Step 3]으로 간다.

[Step 3] $n \leftarrow n-1$ 로 하여 $n = 1$이 될 때까지 [Step 2]를 반복한다.

위의 알고리즘은 단순 선형방정식을 재귀적으로 반복하여 $\mu_i \geq 0$을 만족하는 모든 μ_I를 찾는 비교적 단순한 방법이라 할 수 있다. 이 알고리즘은 k가 높을수록 계산시간이 많이 요구되나, 초기값을 잘 선정하면, 반복 연산을 상당 수준 단축 할 수 있는 장점이 있다. 단 $k = 2$에서는 3개의 모든 모수가 주어져 있기에, 단순한 행렬의 연산에 불과하며, $k = 3$에서도, $\lambda_{123} = \mu_{123} = \min\{\sigma_{ij}\}$로 놓으면, 단순한 행렬 연산에 의해 간단히 μ_I를 유도할 수 있음을 알 수 있다.

5. 예제

$k = 4$인 경우에서 살펴보자. 발생하고자 하는 확률변수 (X_1, X_2, X_3, X_4)를 일반적인 다변량 포아송분포로 표현하면

$$
\begin{aligned}
X_1 &= Y_1 + Y_{12} + Y_{13} + Y_{14} + Y_{123} + Y_{134} + Y_{124} + Y_{1234} \\
X_2 &= Y_2 + Y_{12} + Y_{23} + Y_{24} + Y_{123} + Y_{234} + Y_{124} + Y_{1234} \\
X_3 &= Y_3 + Y_{13} + Y_{23} + Y_{34} + Y_{123} + Y_{134} + Y_{234} + Y_{1234} \\
X_4 &= Y_4 + Y_{14} + Y_{24} + Y_{34} + Y_{134} + Y_{234} + Y_{124} + Y_{1234}
\end{aligned}
$$

로 된다. 이제 $\forall I \subseteq \{1,2,3,4\}$에 대해 $\lambda_I = \sum_{J \subseteq \{1,2,3,4\}, I \subseteq J} \mu_J$ 는 다음과 같다.

$\lambda_{1234} = \mu_{1234}$
$\lambda_{123} = \mu_{123} + \mu_{1234}, \lambda_{134} = \mu_{134} + \mu_{1234}, \lambda_{124} = \mu_{124} + \mu_{1234}, \lambda_{234} = \mu_{234} + \mu_{1234}$
$\lambda_{12} = \mu_{12} + \mu_{123} + \mu_{124} + \mu_{1234},\ \lambda_{13} = \mu_{13} + \mu_{123} + \mu_{134} + \mu_{1234},$
$\lambda_{23} = \mu_{23} + \mu_{123} + \mu_{234} + \mu_{1234},\ \lambda_{14} = \mu_{14} + \mu_{124} + \mu_{234} + \mu_{1234},$
$\lambda_{24} = \mu_{24} + \mu_{124} + \mu_{234} + \mu_{1234},\ \lambda_{34} = \mu_{34} + \mu_{134} + \mu_{234} + \mu_{1234},$
$\lambda_1 = \mu_1 + \mu_{12} + \mu_{13} + \mu_{14} + \mu_{123} + \mu_{134} + \mu_{124} + \mu_{1234},$
$\lambda_2 = \mu_2 + \mu_{12} + \mu_{23} + \mu_{24} + \mu_{123} + \mu_{234} + \mu_{124} + \mu_{1234},$
$\lambda_3 = \mu_3 + \mu_{13} + \mu_{23} + \mu_{34} + \mu_{123} + \mu_{134} + \mu_{234} + \mu_{1234}$
$\lambda_4 = \mu_4 + \mu_{14} + \mu_{24} + \mu_{34} + \mu_{134} + \mu_{234} + \mu_{124} + \mu_{1234}$

이제 $\lambda_i = 1, i = 1, \cdots, 4$이며, $\rho_{12} = 0.4$, $\rho_{13} = 0.3$, $\rho_{14} = 0.2$,$\rho_{2,3} = 0.6$, $\rho_{2,4} = 0.4$ 그리고 $\rho_{3,4} = 0.7$의 구조를 지닌 4변량 포아송분포의 난수를 생성하여보자. 우리는 주어진 상관구조가 $0 \leqq \rho_{ij} < \dfrac{\min(\sqrt{\lambda_i}\sqrt{\lambda_j})}{\max(\sqrt{\lambda_i}\sqrt{\lambda_j})}$의 조건을 만족시키는가를 먼저 확인할 필요가 있다. 위의 조건을 만족하지 못하면, 독립적인 포아송분포의 합으로 표현된 다변량 포아송분포에 대한 난수는 생성할 수 없게 된다. 우리는 여기서 모수의 제약조건에 의해 자연스럽게 $\lambda_{i_1} \geqq \lambda_{i_1 i_2} \geqq \cdots \geqq \lambda_{i_1 i_2 \cdots i_k}$임을 확인할 수 있다. 이제, $\lambda_{ij} = \rho_{ij}\sqrt{\lambda_i}\sqrt{\lambda_j}$에 의해, 주어진 상관구조에 대한 분산-공분산 행렬은 다음과 같다.

$$\sigma_{ij}^{(0)} = \Sigma^{(0)} = \begin{pmatrix} 1 & 0.566 & 0.520 & 0.400 \\ & 2 & 1.470 & 1.131 \\ & & 3 & 2.424 \\ & & & 4 \end{pmatrix}$$

제안된 알고리즘을 적용하기 위해서는 분산-공분산 행렬을 분산이 작은 변수가 위로 올라오도록 중심축화(pivoting)하는 것이 좋다.

[알고리즘에 의한 계산]

1) 초기값 $\lambda_I^{(0)}$를 먼저 결정한다.

1) $\beta_1 = 0.4,\ T_1 = \{1, 4\},\ \Lambda_1 = \{\lambda_{1234}, \lambda_{124}, \lambda_{134}, \lambda_{14}\} = 0.4$
2) $\beta_2 = 0.520,\ T_2 = \{1, 3\},\ \Lambda_2 = \{\lambda_{123}\} = 0.52$
3) $\beta_3 = 0.566,\ T_3 = \{1, 2\},\ \Lambda_3 = \varnothing$
4) $\beta_4 = 1.131,\ T_4 = \{2, 4\},\ \Lambda_4 = \{\lambda_{234}\} = 1.131$

$\cdots$

10) $\beta_{10} = 4.0,\ T_{10} = \{4\},\ \Lambda_{10} = \{\lambda_4\} = 4.0$

2) 초기 모수 집합 $\lambda_I^{(0)}$ 에 대해 조건식을 검사한다. 먼저 $\mu_{1234} = \lambda_{1234}$로 놓고, $\mu_{1234} \geq 0$을 확인한 후 원소의 개수가 3인 모든 $C(3)$ 집합 $\{1,2,3\},\{1,2,4\},\{1,3,4\},\{2,3,4\}$에 대해

$$\mu_{123} = \lambda_{123} - \mu_{1234} = 0.12,\ \mu_{124} = \lambda_{124} - \mu_{1234} = 0,$$
$$\mu_{134} = \lambda_{134} - \mu_{1234} = 0,\ \mu_{234} = \lambda_{234} - \mu_{1234} = 0.731$$

를 계산한 후, 비음을 확인한다.

3) 위의 단계를 통과하면, 원소의 개수가 2인 모든 $C(2)$ 집합 $\{1,2\},\{1,3\},\cdots,\{3,4\}$에 대해

$$\mu_{12} = \lambda_{12} - \mu_{123} - \mu_{124} - \mu_{1234}$$
$$\cdots$$
$$\mu_{34} = \lambda_{34} - \mu_{134} - \mu_{234} - \mu_{1234}$$

를 계산하여 모든 $\mu_{C(2)}$가 비음 인지 확인하고, 0보다 작은 값이 나타나면, 해당 $C(2)$를 포함하는 $C(3)$ 집합 중 최대값 $\mu_i, i \in C(3)$을 선정하여 $\lambda_{C(3)}$을 조정한다.

4) $C(n)$의 원소의 수 $n(C(n))$가 1일 때까지 반복한다.

본 예에서는 초기값이 매우 적절하게 선정되어 반복 연산이 없이도 모든 μ_I가 [Step 1]에서 결정되었다. 즉 초기값에 의해 $\boldsymbol{B}^{-1}\lambda_I^{(0)}$로 계산된 μ_I가 비음임을 알 수 있다. 또한 (3.1)의 방정식에 의한 후향대치 방법으로 모수 λ_I 를 계산한 결과는 다음과 같다.

$$\lambda_{1234} = \mu_{1234} = 0.4$$
$$\lambda_{123} = \mu_{123} + \mu_{1234} = 0.52,\ \lambda_{134} = \mu_{134} + \mu_{1234} = 0.4$$
$$\lambda_{124} = \mu_{124} + \mu_{1234} = 0.4,\ \lambda_{234} = \mu_{234} + \mu_{1234} = 1.131$$
$$\lambda_{12} = \mu_{12} + \mu_{123} + \mu_{124} + \mu_{1234} = 0.566,\ \lambda_{13} = \mu_{13} + \mu_{123} + \mu_{134} + \mu_{1234} = 0.520$$
$$\lambda_{14} = \mu_{14} + \mu_{124} + \mu_{234} + \mu_{1234} = 0.4,\ \lambda_{23} = \mu_{23} + \mu_{123} + \mu_{234} + \mu_{1234} = 1.470$$
$$\lambda_{24} = \mu_{24} + \mu_{124} + \mu_{234} + \mu_{1234} = 1.131,\ \lambda_{34} = \mu_{34} + \mu_{134} + \mu_{234} + \mu_{1234} = 2.424$$
$$\lambda_1 = \mu_1 + \mu_{12} + \mu_{13} + \mu_{14} + \mu_{123} + \mu_{134} + \mu_{124} + \mu_{1234} = 1$$
$$\lambda_2 = \mu_2 + \mu_{12} + \mu_{23} + \mu_{24} + \mu_{123} + \mu_{234} + \mu_{124} + \mu_{1234} = 2$$
$$\lambda_3 = \mu_3 + \mu_{13} + \mu_{23} + \mu_{34} + \mu_{123} + \mu_{134} + \mu_{234} + \mu_{1234} = 3$$
$$\lambda_4 = \mu_4 + \mu_{14} + \mu_{24} + \mu_{34} + \mu_{134} + \mu_{234} + \mu_{124} + \mu_{1234} = 4$$

위의 계산에 의해 우리는 $\lambda_{ij} = Cov(X_i, X_j)$의 관계를 유지하고 있음을 확인할 수 있으며, 최종 다변량 포아송 난수는

$$X_i = \sum_{I \subseteq \{1,\cdots,4\},\ i \in I} \theta_I\ , i = 1, \cdots, 4$$

로 발생할 수 있다.

6. 결론 및 토의

본 연구에서는 Krummenauer(1998a)와 Karlis(2003)에서 언급된 일반적인 다변량 포아송분포로부터 난수를 생성하는 방법을 다루었다. 이를 위해 (3.1)식의 선형방정식의 해를 구하는 알고리즘을 제안하였다. 발생된 난수는 주어진 분산-공분산 행렬에 기초하지만 포아송분포의 모수의 성격상 분산-공분산 행렬은 모두 0보다 크므로 각 변수들 간에 양적인 상관관계가 있음을 알 수 있다. 그러므로 제안된 알고리즘에 의하면 음의 상관을 지니는 다변량 포아송 난수는 생성할 수 없음을 알 수 있다.

Park 등(1996)은 다변량 베르누이 난수를 생성하기 위해 포아송분포의 가역성을 사용하였다. Park 등(1996)은 (3.2)식에서 구해진 X_i에 대해 $Z_i = I_{\{0\}}(X_i), i = 1, \cdots, k$로 변환하여 다변량 베르누이 난수를 생성하였다. 여기서 $I_{\{0\}}$은 발생된 포아송 난수가 0이면 1, 그렇지 않으면 0 인 지시함수이다. 본 연구는 일반적인 포아송 모형에서 결정해야 할 $2^k - 1$ 개의 모수 관점에서 접근하였으며, 또한 선형방정식을 계속 풀어가는 일반적인 해와 비교하여 무수히 많은 해 중 분산-공분산 행렬에서 주어진 모수 개수와 같은 개수의 모수를 선택하는 알고리즘 측면을 논의한 점에서 Park 등(1996)과 차이를 둘 수 있다.

참고문헌

1. Dwass, M. and Teicher, H. (1956). On infinitely divisible random vectors, *Annals of Mathematical Statistics*, 27, 461-470.
2. Karlis, D. (2003). An EM algorithm for multivariate Poisson distribution and related models, *Journal of Applied Statistics*, 30, 1, 63-77.
3. Krummenauer, F. (1998a). Efficient simulation of multivariate binomial and Poisson distributions, *Biometrical Journal*, 40, 7, 823-832.
4. Krummenauer, F. (1998b). Limit theorems for multivariate discrete distributions, *Metrika* 47, 47-69.
5. Loukas, S., and Kemp, C.D. (1983). On computer sampling from trivariate and multivariate discrete distribution, *Journal of Statistical Computations and Simulation*, 17, 113-123.
6. Park, C.G, Park, T. and Shin, D.W. (1996). A simple method for

generating correlated binary variates, *The American Statistician*, 50, 4, 306-310.
7. Tsiamyrtzis, P. and Karlis, D. (2004). Strategies for efficient computation of multivariate Poisson probabilities, *Communications in Statistics-Simulation and Computations*, 33, 2, 271-292.

[2006년 1월 접수, 2006년 2월 채택]