Research Article

Applied Stochastic Models in Business and Industry

Received 23 October 2009, Revised 23 November 2010, Accepted 12 March 2011 Published online 3 May 2011 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/asmb.901

On generating multivariate Poisson data in management science applications

Inbal Yahav*, † and Galit Shmueli

Generating multivariate Poisson random variables is essential in many applications, such as multi echelon supply chain systems, multi-item/multi-period pricing models, accident monitoring systems, etc. Current simulation methods suffer from limitations ranging from computational complexity to restrictions on the structure of the correlation matrix, and therefore are rarely used in management science. Instead, multivariate Poisson data are commonly approximated by either univariate Poisson or multivariate Normal data. However, these approximations are often not adequate in practice.

In this paper, we propose a conceptually appealing correction for *NORTA* (NORmal To Anything) for generating multivariate Poisson data with a flexible correlation structure and rates. *NORTA* is based on simulating data from a multivariate Normal distribution and converting it into an arbitrary *continuous* distribution with a specific correlation matrix. We show that our method is both highly accurate and computationally efficient. We also show the managerial advantages of generating multivariate Poisson data over univariate Poisson or multivariate Normal data. Copyright © 2011 John Wiley & Sons, Ltd.

Keywords: multivariate; Poisson; simulation; NORTA

1. Introduction

Stochastic simulation has been an integral part of the Management Science field in the last 40 years [1]. The role of stochastic simulation is best described in [2]: '[Simulation] is descriptive of the performance of a given configuration of the system [...] simulation does more than yield a numerical measure of the performance of the system. It provides a display of the manner in which the system operates.' Stochastic simulation has been used for the evaluation of increasingly complex models, for sensitivity analysis studies, for real-world approximation, and much more (see e.g. [3–6]).

Simulation from multivariate distributions, though relatively uncommon, is of high importance. Multivariate distributions can describe environments in which multiplicity in providers, consumers, products, horizons, etc. takes place. The most common multivariate distribution in the statistical literature is the multivariate Normal (*Gaussian*) distribution. Generating multivariate Normal data is relatively easy and fast. It has therefore been used for many purposes and in a vast number of applications. In many applications, however, the multivariate data that arise in practice are not well approximated by a multivariate Normal distribution.

For example, consider a classic multi-item inventory model, such as a manufacturing system or retail system that faces a single class of demand per item. Demand per item arrives according to a Poisson process with parameter λ. Note that if demand is low for one or more of the items (e.g. $\lambda < 5$), the Normal distribution cannot be used as an approximation of the demand arrival. It is reasonable to assume that demand for different items within a certain set has a correlation structure (see e.g. [7]). The correlation coefficient of each item pair can differ, depending on the nature of these two items; we would expect positive correlation between complementary items. Substitute items may have negative correlation.

Another example is traffic and accident monitoring systems. Here, multiple intersections are being monitored simultaneously and the number of annual/monthly accidents per intersection is recorded. The form of dependency in this application can arise from spatial correlation (i.e. geographical proximity) or temporal correlation (e.g. time period dependency). We describe two additional applications in greater detail in Section 2.

Department of Decisions, Operations and Information Technologies, R. H. Smith School of Business, University of Maryland, College Park, MD, U.S.A.

**Correspondence to: Inbal Yahav, Department of Decisions, Operations and Information Technologies, R. H. Smith School of Business, University of Maryland, College Park, MD, U.S.A.*

†*E-mail: iyahav@rhsmith.umd.edu*

Generating multivariate Poisson random variables has been addressed massively in the statistics literature, with a major focus on the bivariate case. We survey the variety of approaches in Section 3. However, in spite of the multiple approaches and the obvious need for such data in management science, there is hardly any use of them in the management literature. Instead, researches use either multivariate Normal distributions to approximate Poisson data (see e.g. [8–11]) or simply assume that the multivariate data are independent, and use univariate Poisson data (see e.g. [12–14]).

In this paper, we propose a conceptually simple and computationally efficient method for generating multivariate Poisson data for use in simulation studies. Our method is a correction of the NORTA (NORmal To Anything) approach [15], which is used to generate a multivariate distribution with arbitrary *continuous* marginals (described in Section 3.1). We show that our method is powerful enough to allow a flexible correlation structure (with negative and positive values) and a wide range of rates (low and high). We make our code available in Appendix A.

2. Motivating examples

2.1. *Example 1*: *pricing of nondurable goods*

Setting the price of nondurable goods (e.g. cosmetics, fashion, office supplies) on a finite, multi-planning period is a non-trivial problem. Finding an optimal pricing scheme is shown to be computationally difficult even when demand is assumed to be independent across planning periods [16]. When demand tends to be correlated, the problem becomes even more complex. One practical solution to overcome this complexity is to develop heuristic approaches that account for demand correlation, in place of finding an optimal solution. The role of simulation here is crucial in evaluating the performance of the heuristic approach.

One example for such a pricing problem is described by Gupta *et al*. [17], who study the problem of setting prices for clearing retail inventories of fashion goods. Demand for the goods is assumed to be stochastic and *correlated across time periods*. The authors propose a heuristic approach to estimate the optimal pricing scheme over the planning periods. They evaluate and illustrate the approach using a simulation-based numerical study, in which the demand error term is modeled as a multivariate Normal random variable.

Whereas the Normal distribution is a fair approximation of high demand, it may perform poorly when low demand is considered. Low demand counts are common for high-value products or for short time periods. Hence, a simulation based on multivariate Poisson data would be more adequate for such a model.

2.2. *Example 2*: *Biosurveillance*: *disease outbreak detection*

A main aspect of biosurveillance is the early detection of disease outbreaks. In modern biosurveillance daily aggregates of pre-diagnostic and diagnostic data sets are monitored for the purpose of improving the early response to disease outbreaks (see e.g. [18, 19]). A major feature of biosurveillance data is multiplicity in several dimensions, such as multiple data sources (e.g. pharmacy sales, nurse hotline calls, and emergency department visits), multiple locations (e.g. different hospitals within a given geographical area), multiple disease symptoms, etc.

Multivariate monitoring of biosurveillance data has received attention in the recent literature. One of the challenges addressed in the literature is directionally sensitive multivariate monitoring, where data are monitored for *increases* in the mean of one or more series (rather than traditional monitoring that detects *shifts* in the mean in *any* direction). Solutions range from simple corrections to traditional multivariate monitoring [20, 21], to operation research approaches [22], and to application-specific solutions [23, 24]. The properties of these methods and their performance have been based on and evaluated using the multivariate Normal distribution.

In the context of biosurveillance, Joner *et al*. [21] mention that although the actual distribution is more likely to follow a Poisson distribution, the assumption is that 'each of these Poisson means is sufficiently large to permit the use of normal approximations to the Poisson distributions.' This assumption is essential, as the current directionally sensitive multivariate monitoring methods are not sufficiently robust to support highly skewed distributions [25].

Another main challenge of biosurveillance studies is the lack of available authentic syndromic data to researchers due to privacy and proprietary restrictions. The absence of data limits the ability to develop, evaluate, and compare monitoring methods across different research groups. To tackle this challenge, Lotze *et al*. [26] proposed a method for simulating multivariate syndromic count data, in the form of daily counts from multiple series. The underlying data that the authors generate have a multivariate Normal nature.

The normal distribution assumption, however, is often violated in authentic data, when the actual counts are low (e.g. in daily counts of cough complaints in a small hospital, or daily counts of school absences in a local high school). In low-count situations, a reasonable approximation that has been used in the practice is a Poisson distribution (see e.g. [21, 27]).

3. Existing methods

The p-dimensional Poisson distribution is characterized by a mean (or rate) vector $\vec{\lambda}$ and covariance matrix Σ_{Pois} that has diagonal elements equal to $\vec{\Lambda}$. It is customary to use the term 'multivariate Poisson' for any extension of the univariate Poisson distribution where the resulting marginals are of univariate Poisson form. In other words, the same term is used to describe different multivariate distributions, which have in common the property that their marginals are univariate Poisson.

One of the best known methods for generating bivariate Poisson data is the *Trivariate Reduction*, which was proposed in [28]. In this method three independent Poisson random variables Z_1, Z_2, Z_{12} are first generated with rates λ_1, λ_2 and λ_{12}, respectively. Then, the variables are combined to generate two dependent random variables in the following way:

$$X_1 = Z_1 + Z_{12}$$
$$X_2 = Z_2 + Z_{12}.$$

It is shown that:

$$X_1 \sim \text{Poisson}(\lambda_1 + \lambda_{12}),$$
$$X_2 \sim \text{Poisson}(\lambda_2 + \lambda_{12}),$$
$$\rho_{X_1,X_2} = \frac{\lambda_{12}}{\sqrt{(\lambda_1+\lambda_{12})(\lambda_2+\lambda_{12})}}.$$

The main drawbacks of the Trivariate Reduction method are that it does not support negative correlation values and it does not cover the entire range of feasible correlations. In a recent paper, Shin and Pasupathy [29] presented a computationally fast modification to the Trivariate Reduction method that enables generating a bivariate Poisson with a specified negative correlation. Their method first generates two dependent Poisson variables with rates $\lambda_{X_1}, \lambda_{X_2}$ and some correlation $\widetilde{\rho}_{X_1X_2}$ and then iteratively adjusts the rates to achieve the desired correlation $\rho_{X_1X_2}$.

Krummaenauer [30, 31] proposed a convolution-based method to generate bivariate Poisson data in polynomial time. The algorithm first generates and then convolves independent univariate Poisson variates with appropriate rates. The author presented a recursive formula to carry out the convolution in polynomial time. This method enables the simulation of multivariate Poisson data with *arbitrary* covariance structure. The main limitation of this method is its high complexity (the recursions become very inefficient when the number of series p increases). Also, the method does not support negative correlation.

Minhajuddin *et al.* [32] presented a method for simulating multivariate Poisson data based on the Negative Binomial–Gamma mixture distribution. First, a value k is generated from a Negative Binomial distribution with rate r and success probability $\Pi = \lambda/(\lambda+\theta)$. Then, conditional on k, a set of p independent Gamma variates are generated $(X_1, \ldots, X_p)$. The sum over k of the joint distribution of k and $X_1, \ldots, X_k$ has a Gamma marginal distribution with rates r and λ. The correlation between a pair X_i and X_j $(i \neq j)$ is $\theta/(\lambda+\theta)$. There are two main drawbacks to this approach: First, it requires the correlation between each pair of variates to be identical ($\rho_{ij} = \rho$ for all $i \neq j$). Second, it does not support negative correlations.

Karlis [33] points out that the main obstacle limiting the use of multivariate simulation methods for Poisson data, including the above-mentioned methods, is the complexity of calculating the joint probability function. He mentions that the required summations might be computationally exhausting in some cases, especially when the number of series p is high.

A summary comparison of the methods discussed in this paper is given in Table I.

3.1. *NORTA*: *NORmal TO Anything*

A different approach for generating data from a multivariate distribution with given univariate marginals and a pre-specified correlation structure is known as *NORTA*. The idea is to first generate a p-vector from the multivariate Normal distribution with correlation structure R_N and then to transform it into any desired distribution (say F) using the inverse cumulative distribution function [15, 37]. The resulting distribution is referred to as *normal-copula*.

When the desired distribution F is continuous, the normal-copula has a well-defined correlation structure. However, when F is discrete (as in the Poisson case), the matching between the initial correlation structure R_N and the normal-copula correlation structure R_F is a non-trivial problem [35]. For example, consider the following steps for generating a p-dimensional Poisson vector with arbitrary correlation matrix R_{Pois} and rates $\vec{\Lambda}$:

(1) Generate a p-dimensional Normal vector $\vec{X_N}$ with mean $\vec{\mu} = 0$, variance $\vec{\sigma} = 1$, and a correlation matrix R_N.
(2) For each value X_{N_i}, $i \in 1, 2, \ldots, p$, calculate the Normal CDF:

$$\Phi(X_{N_i}).$$

Table I. Summary of properties of the methods for generating multivariate Poisson data.

Method	Extends to $p>2$	Allows negative corr.	Covers entire corr. range	Allows specific corr.	Complexity limitation
Trivariate reduction [28]				√	√
Modification by Shin and Pasupathy [29]		√	√	√	Iterative corr. search
Convolution-based [30, 31]	√	√	√		Ineffective for large p
NB-Gamma mixture [32]	√		√	Only equal corr.	Based on mixture method (NP complete, [34])
NORTA corr. matching [35]*	√	√	√	√	Based on root finding problem (NP complete, [36])
Our method†	√	√	√	√	

*Method described in Section 3.1.
†Method described in Section 4.1.

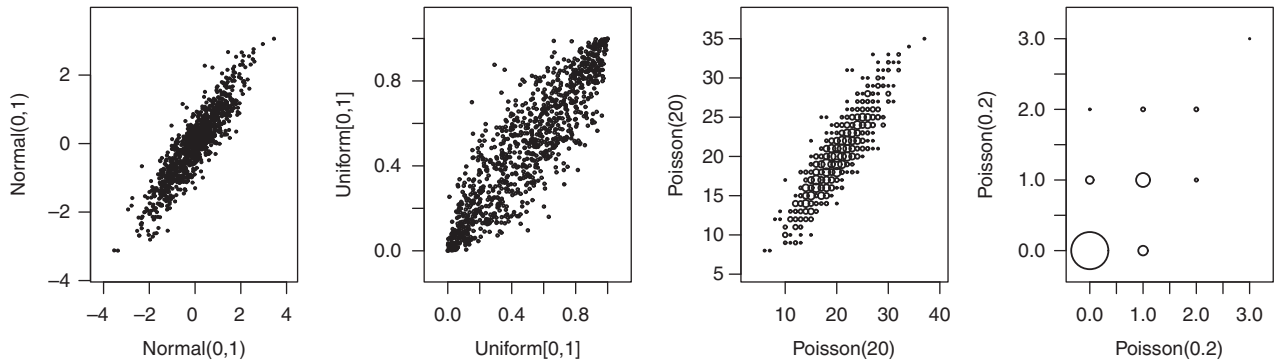

Figure 1. Scatter plots for bivariate simulated variables using NORTA, for Normal, Uniform, Poisson ($\lambda=20$) and Poisson ($\lambda=0.2$).

(3) For each $\Phi(X_{N_i})$, calculate the Poisson inverse CDF (quantile) with rate λ_i

$$X_{\text{Pois}_i} = \Xi^{-1}(\Phi(X_{N_i})),$$

where,

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\sigma^2}} \mathrm{e}^{\frac{-u^2}{2}}\, \mathrm{d}u \tag{1}$$

$$\Xi(x) = \sum_{i=0}^{x} \frac{\mathrm{e}^{-\lambda}\lambda^i}{i!}. \tag{2}$$

The vector $\vec{X}_{\text{Pois}}$ is then a p-dimensional Poisson vector with correlation matrix R_{Pois} and rates $\vec{\Lambda}$. When $\vec{\Lambda}$ is sufficiently high, the Poisson distribution is known to be asymptotically Normal and $R_{\text{Pois}} \approx R_N$. However, when one or more of the rates (λ) is low, the normal-copula correlation deviates from the Normal correlation ($R_{\text{Pois}} \neq R_N$). The reason is that the feasible correlation between two random Poisson variables is no longer in the range $[-1, 1]$, but rather [38]:

$$[\underline{\rho} = \text{corr}(\Xi_{\lambda_1}^{-1}(U), \Xi_{\lambda_2}^{-1}(1-U)), \quad \bar{\rho} = \text{corr}(\Xi_{\lambda_1}^{-1}(U), \Xi_{\lambda_2}^{-1}(U))]. \tag{3}$$

In fact, Shin and Pasupathy [29] show that when $\lambda_1, \lambda_2 \to 0$, the minimum feasible correlation $\underline{\rho} \to 0$. Therefore, the NORTA transformation maps a correlation range of $[-1, 1]$ (multivariate normal) to a much smaller range $[\underline{\rho} \geqslant -1, \bar{\rho} \leqslant 1]$.

To illustrate this reduction in the correlation range, consider Figure 1. The figure depicts a bivariate NORTA transformation process with correlation $\rho = 0.9$, and the resulting Poisson bivariates with high (20) and low (0.2) rates. The 'bubble' size

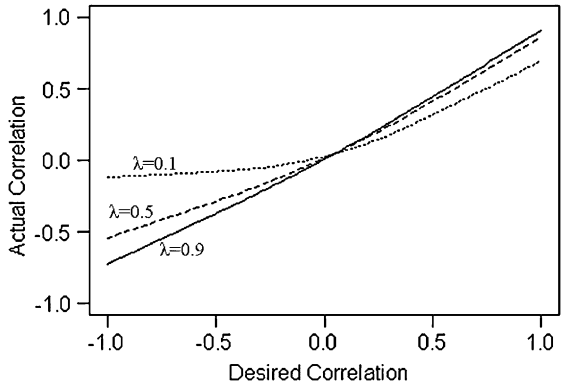

Figure 2. Comparing the desired correlation to the resulting actual correlation for Poisson bivariate data with low rates.

in each panel corresponds to the number of bivariates with the same value. Naturally, the bivariate Poisson with high rates has a fairly similar distribution to that of the Normal distribution. The bivariate Poisson with low rates, however, not only takes very few possible values ($\{(X_1, X_2)|X_1, X_2 \in (0, 1, 2, 3)\}$) but also is a much more skewed distribution (the majority of the bivariate values are the pair $(0, 0)$).

Figure 2 illustrates the relationship between the desired correlation and the resulting actual correlation when generating bivariate Poisson data with low rates ($\lambda < 1$).

In a recent paper, Avarmidis *et al*. [35] studied the NORTA correlation matching problem when the marginal distributions are discrete. The authors provide several approximations for mapping the desired correlation with the actual correlation. Their approximation involves a bivariate normal integral and the approximation of the derivative of the matching function with respect to the actual correlation. They illustrate the performance of their method on multivariate Binomial and Negative Binomial distribution.

In contrast to the approximation by Avarmidis *et al*. [35], we provide a conceptually simple, empirically based approximation method for mapping R_N to R_{Pois}. We show that our method is highly accurate (with absolute error of less than 6×10^{-2}) and can be computed within milliseconds. We hope that the simplicity of this method, along with the availability of the code will lead to a wider use of simulated multivariate Poisson data, which can be used for studying and evaluating algorithms in the management science field.

4. Generating multivariate Poisson random variables

4.1. Model description

We describe our approximation for the bivariate case ($p = 2$). One can easily extend the method to higher dimensional data by simply applying the correlation mapping to each pair of bivariates.

We define $\vec{\Lambda} = \{\lambda_1, \lambda_2\}$ and,

$$\begin{aligned} \underline{\rho} &= \text{corr}(\Xi_{\lambda_1}^{-1}(U), \Xi_{\lambda_2}^{-1}(1-U)), \\ \bar{\rho} &= \text{corr}(\Xi_{\lambda_1}^{-1}(U), \Xi_{\lambda_2}^{-1}(U)). \end{aligned} \tag{4}$$

To determine an adequate approximation for the relationship between the desired correlation (ρ_{pois}) and the actual correlation (ρ_N), we compose a large simulation study that produces pairs ($\rho_{\text{pois}}, \rho_N$) for different levels of Poisson rates (λ_1, λ_2). We then examine possible parametric relationships (such as exponential, quadratic, and linear) between ρ_{pois} and ρ_N. We fit each approximation by estimating the appropriate generalized linear model, and then evaluate goodness-of-fit (GOF) by examining the residuals. Because we have many different sets of Poisson rates, we summarize the GOF for each parametric approximation using the Root Mean-squared Error (RMSE) for each rate pair (λ_1, λ_2). The average and standard deviation of RMSE taken across all the different rate pairs are provided in Table II. In Figure 3, we illustrate the fit of the different parametric approximations for the particular pair $\lambda_1 = \lambda_2 = 0.4$, by displaying QQ plots of the different model residuals. Similar results were obtained for other values of λ_1, λ_2.

Table II. Goodness-of-fit of various parametric relationships between ρ_{pois} and ρ_N, summarized by the average and standard deviation of RMSE across a range of rate pairs λ_1, λ_2.

Relationship	Mean RMSE	STD of RMSE
Exponential	**0.037**	0.006
Double exponential	**0.066**	0.022
Linear	0.046	0.011
Quadratic	0.096	0.034
Polynomial	0.089	0.040

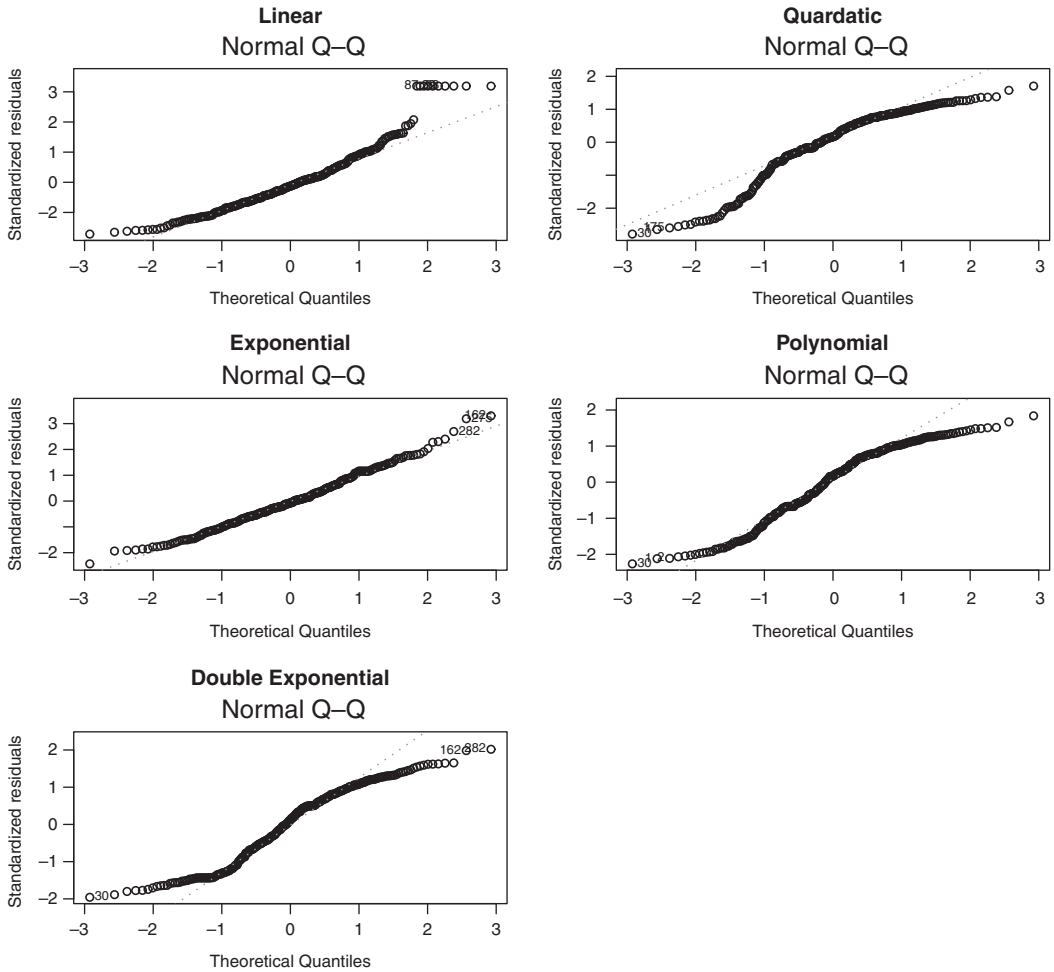

Figure 3. Goodness-of-fit of several parametric approximating models fitted to the pairs $(\rho_{\text{pois}}, \rho_N)$ for the case $\lambda_1 = \lambda_2 = 0.4$.

Based on this simulation study, we find that the relationship between the desired correlation (ρ_{pois}) and the actual correlation (ρ_N) is best approximated by an exponential function:

$$\rho_{\text{Pois}} = a \times e^{b\rho_N} + c. \tag{5}$$

The coefficients a, b, and c can be estimated from the points $(\underline{\rho}, -1)$, $(\bar{\rho}, 1)$ and $(0, 0)$:

$$\begin{aligned} a &= -\frac{\bar{\rho} \times \underline{\rho}}{\bar{\rho} + \underline{\rho}}, \\ b &= \log\left(\frac{\bar{\rho} + a}{a}\right), \\ c &= -a. \end{aligned} \tag{6}$$

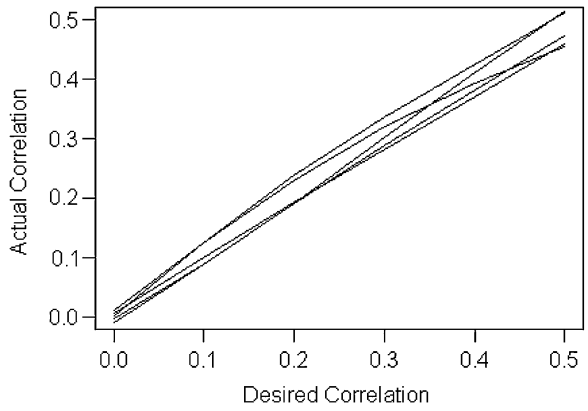

Figure 4. Comparing the desired correlation to the corrected actual correlation.

Following Equations (4)–(6), in order to generate bivariate poisson data with rates λ_1, λ_2, and a desired (*feasible*) correlation ρ_{Pois}, the following steps should be taken:

- Compute $\rho, \overline{\rho}, a, b$, and c from Equations (4) and (6).
- Compute the initial correlation from Equation (5).

$$\rho_N = \frac{\log\left(\frac{\rho_{\text{Pois}} - c}{a}\right)}{b}. \tag{7}$$

- Generate bivariate Normal data with parameters $\mu_1 = \mu_2 = 0$, $\sigma_1 = \sigma_2 = 1$ and correlation ρ_N.
- Follow the NORTA procedure to generate bivariate Poisson data with rates λ_1, λ_2 and correlation ρ_{Pois}.

4.2. Method evaluation

To evaluate the performance of our approximation in terms of accuracy and computation time, we implement the algorithm in R software on a 2.6 GHz Intel dual-core 32 bit-processor running Windows Vista.

Figure 4 illustrates the simulation performance when using the above approximation to correct for the distortion in the resulting correlation. This is illustrated for the bivariate Poisson case with rates that range in $(\lambda_1, \lambda_2) \in (0.1, 0.1), (0.1, 0.5), (0.5, 0.5), (0.5, 0.9), (0.9, 0.9)$. We see that the actual correlation is approximately equal to the desired correlation. Figure 5 shows the mean absolute difference between the actual and desired correlations, for any choice of Λ and ρ. We use white color to represent infeasible correlation values (according to the correlation range in Equation 4). We see that this difference is less than 0.06 (both panels). We also see that the approximation is more accurate for higher rates (left panel), least accurate (though still fairly accurate) for high negative correlation with high rates (right panel), and that the method performs more accurately as the rates increase.

Apart from simplicity, a very important feature of the generator is the short computation time. Figure 6 depicts the computation time (in milliseconds) as a function of data dimension p and series length. The running time is shown to be minor even when generating large data sets.

5. Managerial implications

In this section we exemplify the managerial benefit of generating multivariate Poisson data in management science applications over existing methods such as univariate Poisson and multivariate Normal. For that purpose, we simulate data for the two applications corresponding to the motivating examples presented in Section 2. In both cases, the data are assumed to follow a multivariate Poisson distribution. We discuss the practical implications of using a multivariate Normal approximation of the Poisson distribution, or alternatively ignoring the multivariate structure altogether by assuming independence across series.

5.1. Pricing of nondurable goods

We implement the multi-period clearance pricing problem described in [17]. In this problem, the retailer has N pricing opportunities (we set $N = 2$ for our example). The objective is to maximize the expected revenue by choosing price scheme

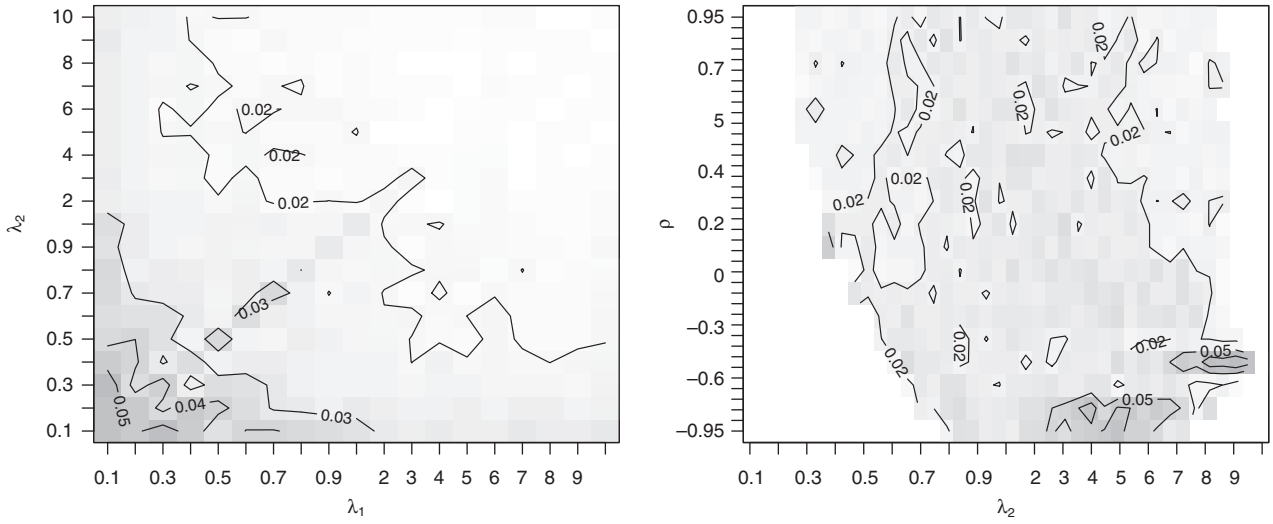

Figure 5. Absolute mean error. Left: error as a function of the Poisson rates. Right: error as a function of the Poisson rate and the desired correlation ($\lambda_1 = 0.4$).

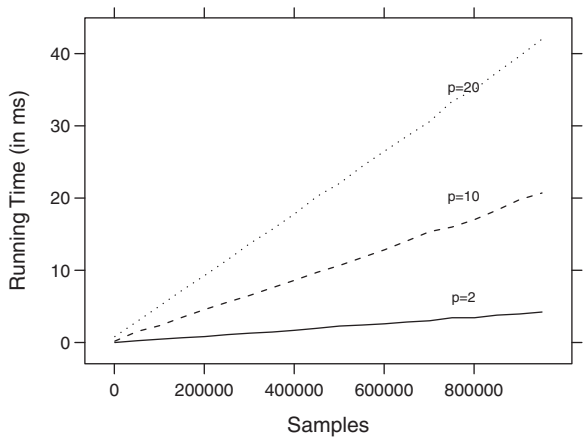

Figure 6. Computation time as a function of the data dimension and length.

$p = (p_1, \ldots, p_N)$, given initial inventory level I, and a salvage value s for any leftover items at the end of period N. Demand is assumed to be stochastic and price dependent of the form:

$$D_i(p_i) = K \mathrm{e}^{-\beta_i p_i} \xi_i, \tag{8}$$

where K represents the market size; β_i is the price sensitivity (for period i); ξ_i is the *correlated* multiplicative random variable that represents a stochastic demand error term, and ρ is the correlation between each consecutive pair of error terms ($\rho = \mathrm{cor}(\xi_i, \xi_{i+1})$).

In their numerical example, Gupta *et al.* [17] assume that $\vec{\xi} = \{\xi_i\}$ follow a multivariate *Normal* distribution. In contrast, we assume that $\vec{\xi}$ follow a multivariate *Poisson* distribution. We use a simulation approach to compute the optimal expected revenue (Equation (22) in [17]):

$$E[\Pi(\mathbf{p})] = p_1 I_1 - \sum_{n=1}^{N} \left(p_n - p_{n+1} E\left[I_1 - \sum_{i=1}^{n} D_i \right]^+ \right). \tag{9}$$

To evaluate the term $E[I_1 - \sum_{i=1}^{n} D_i]^+$ we simulate possible demand scenarios, with the error term ξ_i being taken from the following distributions:

Scenario 1: $\vec{\xi}$ follow a multivariate Poisson distribution with $\Lambda = \{\lambda_1, \ldots, \lambda_N\}$ and correlation ρ.

Scenario 2: $\vec{\xi}$ follow a multivariate Normal distribution with $\mu = \sigma^2 = \{\lambda_1, \ldots, \lambda_N\}$ (an approximation of the actual multivariate Poisson distribution) and correlation ρ.

Table III. Revenue under multivariate Normal, univariate Poisson and multivariate Poisson distributions.

λ_1	λ_2	β_1	β_2	ρ	Revenue under *Scenario 1*: multivariate Poisson	Revenue under *Scenario 2*: multivariate Normal	Revenue under *Scenario 3*: independent Poisson
4	3	1	2	0.5	1381	1453 (*4.96%*)	1395 (*1.00%*)
4	3	1	2	0.9	1347	1432 (*5.94%*)	1417 (*4.94%*)
4	3	1	2	0	1405	1464 (*4.03%*)	1400 (*0.36%*)
4	3	1	2	−0.5	1434	1487 (*3.56%*)	1391 (*−3.09%*)
4	3	1	2	−0.9	1469	1517 (*3.16%*)	1402 (*−4.78%*)
4	3	0.5	1	0.5	1903	1960 (*2.91%*)	1938 (*1.81%*)
4	3	2	3	0.5	731	758 (*3.56%*)	754 (*3.05%*)
3	2	1	2	0.5	1133	1186 (*4.47%*)	1169 (*3.08%*)
2	1	1	2	0.5	838	847 (*1.06%*)	853 (*1.76%*)
Mean running time					*13.2 s*	*11.85 s*	*11.4 s*

Scenario 3: ξ_i follow a set of independent univariate Poisson distributions with $\Lambda = \{\lambda_1, \ldots, \lambda_N\}$ (i.e. $\rho = 0$).

Running several experiments, we find that the pricing scheme under the different distributions remains unchanged. However, the *actual revenue* varies. Under the multivariate Normal distribution the revenue is constantly higher (by 2–5% in our examples), implying that the retailer who faces multivariate Poisson demand, yet uses a Normal approximation to study his revenue opportunities, would constantly overestimate his actual revenue. If the retailer ignores the demand cross-correlation (i.e. use an independent set of univariate Poisson distributions to approximate revenue), the estimated revenue would be slightly higher (1–4%) than the actual revenue if the correlation coefficient is positive, and lower (3–4%) if the correlation is negative.

Table III illustrates this result. The simulated data have $I = 1000$, $s = 0.1$, and $K = 1000$. The values of $\lambda_1, \lambda_2, \beta_1, \beta_2$, and ρ vary.

5.2. *Biosurveillance*: *disease outbreak detection*

Consider a disease anomaly detection system that monitors work absences in search of a disease outbreak in a certain neighborhood. We assume that work absences within each workplace follow a Poisson distribution, with λ being proportional to the workplace size[‡]. Owing to geographic proximity, it is reasonable to assume that work absences across workplaces in nearby geographical areas are correlated. We use p to represent the number of workplaces in the neighborhood. For simplicity we assume that the correlation coefficient of each pair of absences from workplaces is equal to ρ.

We use the algorithm in [20] to monitor the series for anomalies. Follmann [20] presents a simple directionally sensitive multivariate Hotelling control chart to detect increases in the mean of one or more series. The monitoring statistic is given as

$$\chi_t^2 = (\underline{X}_t - \underline{\mu})' \Sigma^{-1} (\underline{X}_t - \underline{\mu}), \tag{10}$$

where $\underline{X}_t$ is the daily count vector (work absences, in our example) at time t; $\underline{\mu}$ is the sample mean vector, and Σ is the covariance matrix. An alert is triggered when $\{\chi_t^2 > \chi_{2\alpha}^2(p)$ and $\sum_{j=1}^{p}(X_t^j - \mu^j) > 0\}$.

Follmann proves that the procedure has false alert rate equal to 2α, and uses simulations to illustrate its true alert rate and to compare it with more complicated likelihood ratio tests.

We simulate work absence samples with a varying number of workplaces p, absence mean vector Λ, and correlation ρ, for a *no-disease period* of one year (365 days). We use Follmann's method with the theoretical threshold of $2\alpha = 5\%$. This means that we allow 5% false alerts (on average 1–2 false alerts every month). This threshold is set to meet the system's capability to investigate alerts. Under this set up, the algorithm should ideally produce not more than 5% alerts (which would all be false, due to the lack of outbreaks in this period). Table IV summarizes the actual resulting false alert rate of our experiments, when the underlying data are generated from a multivariate Poisson distribution.

Our experiments show that Follmann's Hotelling method is very sensitive to the underlying distribution. The false alert rate increases significantly when the data follow a multivariate Poisson distribution. In some cases, the false alert rate reaches more than 50% of the desired rate, implying that the system has to investigate twice more alerts than its capability. In practice, high false alert rates often lead many users to ignore alerts altogether. Link no longer available.

[‡] *We assume that the data have been adjusted for seasonal and day-of-week effects.*

Table IV. False alert rates when workplace absences follow a multivariate Poisson distribution.

Number of workplaces (p)	Correlation (ρ)	Absence rate ($\Lambda=(\lambda_1 \ldots \lambda_p)$)	False alert rate (compared to 5%)
2	0	(1, 1)	0.09 (44.44%)
2	0	(4, 4)	0.07 (28.57%)
2	0	(10, 10)	0.06 (16.67%)
2	0.5	(0.1,0.1)	0.15 (66.67%)
2	0.5	(1, 1)	0.08 (37.50%)
2	0.5	(4, 4)	0.07 (28.57%)
2	0.5	(10, 10)	0.05 (00.00%)
2	0.5	(1, 10)	0.09 (41.18%)
2	0.9	(1, 1)	0.20 (75.00%)
2	0.9	(4, 4)	0.10 (50.00%)
2	0.9	(10, 10)	0.07 (28.57%)
2	0.9	(1, 10)	0.10 (50.00%)
5	0.5	(1, 1, 1, 1, 1)	0.14 (64.29%)
5	0.5	(10, 10, 10, 10, 10)	0.08 (33.33%)

Note that the alerting performance degrades steeply as Λ decreases and a multivariate Normal distribution can no longer be used as a proxy for the actual work absences. A similar decrease in the performance is observed when the number of workplaces (p) increases.

6. Conclusions

Simulating multivariate Poisson data is essential in many real-world applications in a wide range of fields such as healthcare, marketing, management science, and many others where multivariate count data arise. Current simulation methods suffer from computational limitations and restrictions on the correlation structure, and therefore are rarely used.

In this paper, we propose an elegant modification of the NORTA method to generate multivariate Poisson data based on a multivariate Normal distribution with a pre-specified correlation matrix and Poisson rate vector. Because multivariate Normal and univariate Poisson simulators are implemented in many standard statistical software packages, implementing our method requires only a few lines of code.

We show that our method works well for different correlation structures (both negative and positive; and varying values) and for high and low Poisson rates. We show that the method is highly accurate in terms of producing Poisson marginal distributions and the pre-specified correlation matrix.

Finally, we show the practical advantages of generating multivariate Poisson data over univariate Poisson or Multivariate Normal data: In pricing of nondurable goods, inadequate simulation can lead to under- or over-estimation of revenue. In biosurveillance, inadequate simulations can lead to excessive false alerts.

Appendix A: generating multivariate Poisson data in R

```
  # Generate a p-dimensional Poisson
  # p     = the dimension of the distribution
  # samples  = the number of observations
  # R     = correlation matrix p X p
  # lambda   = rate vector p X 1
GenerateMultivariatePoisson<-function(p, samples, R, lambda){
    normal_mu=rep(0, p)
    normal = mvrnorm(samples, normal_mu, R)
    unif=pnorm(normal)
    pois=t(qpois(t(unif), lambda))
    return(pois)
}
  # Correct initial correlation between a
  # certain pair of series
  # lambda1  = rate of first series
```

```
  # lambda2  = rate of second series
  # r     = desired correlation
CorrectInitialCorrel<-function(lambda1, lambda2, r){
  samples=500
  u = runif(samples, 0, 1)
  lambda=c(lambda1,lambda2)
  maxcor=cor(qpois(u, lambda1), qpois(u, lambda2))
  mincor=cor(qpois(u, lambda1), qpois(1-u, lambda2))
  a=-maxcor*mincor/(maxcor+mincor)
  b=log((maxcor+a)/a, exp(1))
  c=-a
  corrected=log((r+a)/a, exp(1))/b
  corrected=ifelse ((corrected>1 | corrected<(-1)),
         NA, corrected)
  return(corrected)
}
```

References

1. Nelson BL. Stochastic simulation research in Management Science. *Management Science* 2004; 855–868.
2. Conway RW, Johnson BM, Maxwell WL. Some problems of digital systems simulation. *Management Science* 1959; 92–110.
3. Weeks JK. A simulation study of predictable due-dates. *Management Science* 1979; 363–373.
4. Fu MC. Optimization for simulation: theory vs. practice. *INFORMS Journal on Computing* 2002; **14**(3):192–215.
5. Bhuiyan N, Gerwin D, Thomson V. Simulation of the new product development process for performance improvement. *Management Science* 2004; 1690–1703.
6. Avramidis AN, Deslauriers A, L'Ecuyer P. Modeling daily arrivals to a telephone call center. *Management Science* 2004; 896–908.
7. Song JS. Order-based backorders and their implications in multi-item inventory systems. *Management Science* 2002; 499–516.
8. Eppen GD. Effects of centralization on expected costs in a multi-location newsboy problem. *Management Science* 1979; 498–501.
9. Yang WN, Nelson BL. Multivariate batch means and control variates. *Management Science* 1992; 1415–1431.
10. Caron F, Marchet G. The impact of inventory centralization/decentralization on safety stock for two-echelon systems. *Journal of Business Logistics* 1996; **17**:233–258.
11. Yan XS, Robb DJ, Silver EA. Inventory performance under pack size constraints and spatially-correlated demand. *International Journal of Production Economics* 2009; **117**(2):330–337.
12. Muckstadt JA, Thomas LJ. Are multi-echelon inventory methods worth implementing in systems with low-demand-rate items? *Management Science* 1980; 483–494.
13. Jarvis JP. Approximating the equilibrium behavior of multi-server loss systems. *Management Science* 1985; 235–239.
14. Whitt W. A light-traffic approximation for single-class departure processes from multi-class queues. *Management Science* 1988; 1333–1346.
15. Nelsen RB. *An Introduction to Copulas*. Springer: Berlin, 2006.
16. Heyman DP, Sobel MJ. *Stochastic Models in Operations Research*: *Stochastic Optimization*. McGraw-Hill, 1984; 2.
17. Gupta D, Hill AV, Bouzdine-Chameeva T. A pricing model for clearing end-of-season retail inventory. *European Journal of Operational Research* 2006; **170**(2):518–540.
18. Shmueli G, Fienberg SE. Current and potential statistical methods for monitoring multiple data streams for biosurveillance. *Statistical Methods in Counter-terrorism* 2004; 109–140.
19. Shmueli G, Burkom H. Statistical challenges facing early outbreak detection in biosurveillance. *Technometrics* 2010; **52**(1):39–51.
20. Follmann D. A simple multivariate test for one-sided alternatives. *Journal of the American Statistical Association* 1996; **91**(434):854–861.
21. Joner MD, Woodall WH, Reynolds Jr, MR, Fricker Jr, RD. A one-sided MEWMA chart for health surveillance. *Quality and Reliability Engineering International* 2008; **24**(5):503–518.
22. Testik MC, Runger GC. Multivariate one-sided control charts. *IIE Technometrics* 2006; **38**:635–645.
23. Fricker RD. Directionally sensitive multivariate statistical process control procedures with application to syndromic surveillance. *Advances in Disease Surveillance* 2007; **3**:1–22.
24. Fricker RD, Knitt MC, Hu CX. Comparing directionally sensitive MCUSUM and MEWMA procedures with application to biosurveillance. *Quality Engineering* 2008; **20**(4):478–494.
25. Stoumbos ZG, Sullivan JH. Robustness to non-normality of the multivariate EWMA control chart. *Journal of Quality Technology* 2002; **34**(3):260–276.
26. Lotze T, Shmueli G, Yahav I. *Biosurveillance*: *A Health Protection Priority*. Simulating and evaluating biosurveillance datasets. Chapman & Hall: New York, 2009.
27. Kleinman K, Lazarus R, Platt R. A generalized linear mixed models approach for detecting incident clusters of disease in small areas, with an application to biological terrorism. *American Journal of Epidemiology* 2004; **159**(3):217–224.
28. Mardia KV. *Families of Bivariate Distributions*. Griffin: London, 1970.
29. Shin K, Pasupathy R. A method for fast generation of bivariate Poisson random vectors. *Proceedings of the 39th Conference on Winter Simulation*: *40 Years! The Best is Yet to Come*. IEEE Press: Piscataway, NJ, U.S.A., 2007; 472–479.
30. Krummenauer F. Efficient simulation of multivariate binomial and poisson distributions. *Biometrical Journal* 1998; **40**(7):823–832.
31. Krummenauer F. Limit theorems for multivariate discrete distributions. *Metrika* 1998; **47**(1):47–69.

32. Minhajuddin ATM, Harris IR, Schucany WR. Simulating multivariate distributions with specific correlations. *Journal of Statistical Computation and Simulation* 2004; **74**(8):599–607.
33. Karlis D. An EM algorithm for multivariate Poisson distribution and related models. *Journal of Applied Statistics* 2003; **30**(1):63–77.
34. James LF, Priebe CE, Marchette DJ. Consistent estimation of mixture complexity. *Annals of Statistics* 2001; **29**(5):1281–1296.
35. Avramidis AN, Channouf N, L'Ecuyer P. Efficient correlation matching for fitting discrete multivariate distributions with arbitrary marginals and normal-copula dependence. *INFORMS Journal on Computing* 2009; **21**(1):88–106.
36. Dom M, Guo J, Niedermeier R. Bounded degree closest k-Tree power is NP-complete. *Computing and Combinatorics* 2005; 757–766.
37. Chen H. Initialization for NORTA: generation of random vectors with specified marginals and correlations. *INFORMS Journal on Computing* 2001; **13**(4):312–331.
38. Whitt W. Bivariate distributions with given marginals. *Annals of Statistics* 1976; **4**(6):1280–1289.