Knudson *et al. Journal of Statistical Distributions and Applications* (2021) 8:6
https://doi.org/10.1186/s40488-021-00119-y

Journal of Statistical Distributions and Applications

RESEARCH Open Access

A flexible multivariate model for high-dimensional correlated count data

Alexander D. Knudson, Tomasz J. Kozubowski, Anna K. Panorska and A. Grant Schissler*

*Correspondence:
aschissler@unr.edu
Department of Mathematics & Statistics, University of Nevada, 89557 Reno, USA

Abstract

We propose a flexible multivariate stochastic model for over-dispersed count data. Our methodology is built upon mixed Poisson random vectors $(Y_1, \ldots, Y_d)$, where the $\{Y_i\}$ are conditionally independent Poisson random variables. The stochastic rates of the $\{Y_i\}$ are multivariate distributions with arbitrary non-negative margins linked by a copula function. We present basic properties of these mixed Poisson multivariate distributions and provide several examples. A particular case with geometric and negative binomial marginal distributions is studied in detail. We illustrate an application of our model by conducting a high-dimensional simulation motivated by RNA-sequencing data.

Keywords: Multivariate count data, Copula, Distribution theory, Big data applications, Gamma-Poisson hierarchy, Mixed Poisson distribution, Negative binomial distribution, High-dimensional multivariate simulation, RNA-sequencing data

AMS Subject Classification: 62E10; 62E15; 62H05; 62H10; 62H30

1 Introduction

As multivariate count data become increasingly common across many scientific disciplines, including economics, finance, geosciences, biology, marketing, and others, there is a growing need for flexible families of multivariate distributions with discrete margins. In particular, flexible models with correlated classical marginal distributions are in high demand in many different applied areas (see, e.g., Barbiero and Ferrari (2017); Madsen and Birkes (2013); Madsen and Dalthorp (2007); Nikoloulopoulos and Karlis (2009); Xiao (2017)). With this in mind, we propose a general method of constructing discrete multivariate distributions with certain common marginal distributions. One important example of this construction is a discrete multivariate model with correlated negative binomial (NB) components and arbitrary parameters. However, our approach is quite general and can produce families with different margins, going beyond the NB case.

One way to generate multivariate distributions with particular margins is an approach through copulas (see, e.g., Nelsen (2006)), and multivariate discrete distributions constructed through this method have been proposed in recent years (see, e.g., Barbiero and Ferrari (2017); Madsen and Birkes (2013); Nikoloulopoulos (2013); Nikoloulopoulos and Karlis (2009); Xiao (2017) and references therein). Recall that a copula is a cumulative

distribution function (CDF) on $[0,1]^d$, describing a random vector with standard uniform margins. Moreover, for any random vector $\mathbf{X} = (X_1, \ldots, X_d)^\top$ with the joint CDF F and marginal CDFs F_i there is a copula function $C(u_1, \ldots, u_d)$ so that

$$F(x_1, \ldots, x_d) = \mathbb{P}(X_1 \leq x_1, \ldots, X_d \leq x_d) = C(F_1(x_1), \ldots, F_d(x_d)), \quad x_i \in \mathbb{R}, i = 1, \ldots, d. \tag{1}$$

Further, for continuous distributions with marginal probability density functions (PDFs) $f_i(x) = F_i'(x)$, the copula function C is unique, and the joint PDF of the $\{X_i\}$ is given by

$$f(x_1, \ldots, x_d) = \left\{ \prod_{i=1}^{d} f_i(x_i) \right\} c(F_1(x_1), \ldots, F_d(x_d)), \quad x_i \in \mathbb{R}, i = 1, \ldots, d, \tag{2}$$

where the function $c(u_1, \ldots, u_d)$ is the PDF corresponding to the copula $C(u_1, \ldots, u_d)$. However, for discrete distributions, the copula is no longer unique and there is no analogue of (2) for calculating the relevant probabilities. Using this concept, one can define a random vector $\mathbf{Y} = (Y_1, \ldots, Y_d)^\top$ in $\mathbb{R}^d$ with arbitrary marginal CDFs F_i viz.

$$(Y_1, \ldots, Y_d) = \left(F_1^{-1}(U_1), \ldots, F_d^{-1}(U_d) \right), \tag{3}$$

where $\mathbf{U} = (U_1, \ldots, U_d)^\top$ is a random vector with standard uniform margins and the CDF given by

$$F_{\mathbf{U}}(u_1, \ldots, u_n) = \mathbb{P}(U_1 \leq u_1, \ldots, U_d \leq u_d) = C(u_1, \ldots, u_d), \quad (u_1, \ldots, u_d)^\top \in [0,1]^d, \tag{4}$$

with a particular copula C. While one can use any of the multitude of different copula functions in this construction, an approach based on Gaussian copula, known as NORTA (**NOR**mal **T**o **A**nything, see, e.g., Chen (2001); Song and Hsiao (1993)), is especially popular due to its flexibility, particularly in the case of discrete distributions (see, e.g., Barbiero and Ferrari (2017); Madsen and Birkes (2013); Nikoloulopoulos (2013)).

While our approach involves copulas as well, the latter connect with *continuous* multivariate distributions rather than discrete, which avoids the issues with non-uniqueness of the copula function. Additionally, compared with the direct approach (3), in our scheme the computation of relevant probabilities is straightforward. Our methodology is based on mixtures of Poisson distributions, which is a common way of obtaining discrete analogs of continuous distributions on nonnegative reals with a particular stochastic interpretation. Indeed, discrete univariate *mixed Poisson* distributions have been proven useful stochastic models in many scientific fields (see, e.g., Karlis and Xekalaki (2005), where one can find a comprehensive review of these distributions with over 30 particular examples). This construction can be described through a randomly stopped Poisson process. More precisely, let $\{N(t), t \in \mathbb{R}_+\}$ be a homogeneous Poisson process with rate $\lambda > 0$, so that the marginal distribution of $N(t)$ is Poisson with parameter (mean) λt. Then, for any random variable T with cumulative distribution function (CDF) F_T, supported on $\mathbb{R}_+$, the quantity $Y = N(T)$ is an integer-valued random variable, with distribution determined viz. standard conditioning argument as follows:

$$\mathbb{P}(Y = n) = \int_{\mathbb{R}_+} \frac{e^{-\lambda t}(\lambda t)^n}{n!} dF_T(t), \quad n \in \mathbb{N}_0 = \{0, 1, \ldots\}. \tag{5}$$

Many standard probability distributions on $\mathbb{N}_0$ arise from this scheme. In particular, if T has a standard gamma distribution with shape parameter $r > 0$, given by the PDF

$$f(x) = \frac{1}{\Gamma(r)} x^{r-1} e^{-x}, \quad x \in \mathbb{R}_+, \tag{6}$$

then $Y = N(T)$ will have a NB distribution $\mathrm{NB}(r, p)$ with the probability mass function (PMF)

$$\mathbb{P}(Y = n) = \frac{\Gamma(n+r)}{\Gamma(r)n!} p^r (1-p)^n, \quad n \in \mathbb{N}_0, \tag{7}$$

where $p = 1/(1+\lambda)$ (see Section 3.2 in the Appendix). As the NB model is quite important across many applications and can be extended to more general stochastic processes (see, e.g., Kozubowski and Podgórski (2009)), it shall serve as a basic example of our approach.

An extension of this scheme to the multivariate case can be accomplished in two different ways, leading to mixed multivariate Poisson distributions of Kind (or Type) I and II in the terminology of Karlis and Xekalaki (2005). The former arises viz.

$$\mathbf{Y} = (Y_1, \ldots, Y_d) = (N_1(T), \ldots, N_d(T)), \tag{8}$$

where the $\{N_i(\cdot)\}$ are Poisson processes with rates λ_i and T is, as before, a random variable on $\mathbb{R}_+$, independent of the $\{N_i\}$. While in general the marginal distributions of $(N_1(t), \ldots, N_d(t))$ can be correlated multivariate Poisson (see, Johnson et al. (1997)), we shall assume that the processes $\{N_i\}$ are mutually independent. In this case, the joint probability generating function (PGF) of the $\{Y_i\}$ in (8) is of the form

$$G(s_1, \ldots, s_d) = \mathbb{E}\left\{\prod_{i=1}^d s_i^{Y_i}\right\} = \phi_T\left(\sum_{i=1}^d \lambda_i - \sum_{i=1}^d \lambda_i s_i\right), \quad (s_1, \ldots, s_d)^\top \in [0,1]^d, \tag{9}$$

where ϕ_T is the Laplace transform (LT) of T, while the relevant probabilities can be conveniently expressed as

$$\mathbb{P}(\mathbf{Y} = \mathbf{y}) = \prod_{i=1}^d g_i(y_i) h(\mathbf{y}), \quad \mathbf{y} = (y_1, \ldots, y_d)^\top \in \mathbb{N}_0^d, \tag{10}$$

where the $\{g_i\}$ in (10) are the *marginal* PMFs of the $\{Y_i\}$. As shown in the Appendix, the function h is of the form

$$h(\mathbf{y}) = \frac{\nu_T\left(\sum_{i=1}^d y_i, \sum_{i=1}^d \lambda_i\right)}{\sum_{i=1}^d \nu_T(y_i, \lambda_i)}, \quad \mathbf{y} = (y_1, \ldots, y_d)^\top \in \mathbb{N}_0^d, \tag{11}$$

where

$$\nu_T(y, \lambda) = \mathbb{E}\left\{e^{-\lambda T} T^y\right\}, \quad \lambda, y \in \mathbb{R}_+. \tag{12}$$

In case of gamma distributed T, with shape parameter $r > 0$ and unit scale, the functions ν and h above can be evaluated explicitly (see the Appendix for details), and the above distribution is know in the literature as the multivariate *negative multinomial* distribution (see Chapter 36 of Johnson et al. (1997) and references therein). Since the marginal distributions in this case are NB, the distribution has also been termed multivariate NB. In cases where the function $\nu(\cdot, \cdot)$ is not available explicitly, it can be easily evaluated numerically, viz. Monte Carlo simulations.

Our main focus will be a more flexible family of *mixed Poisson distributions of Kind II*, where each Poisson process $\{N_i(t),\ t \in \mathbb{R}_+\}$ is randomly stopped at a different random variable T_i, leading to

$$\mathbf{Y} = (Y_1, \ldots, Y_d) = (N_1(T_1), \ldots, N_d(T_d)), \tag{13}$$

where the random vector $\mathbf{T} = (T_1, \ldots, T_d)^\top$ follows a multivariate distribution on $\mathbb{R}_+^d$. A particular special case of this construction with the $\{T_i\}$ having correlated log-normal distributions was recently proposed in Madsen and Dalthorp (2007), where this model was referred to as *lognormal-Poisson hierarchy* (L-P model). While that particular model does not allow explicit forms for marginal PMFs, it proved useful for applications. Our generalization, which we shall refer to as $\mathbf{T}$-Poisson hierarchy, will allow $\mathbf{T}$ in (13) to have any continuous distribution on $\mathbb{R}_+^d$, with margins not necessarily belonging to the same parametric family. As will be seen in the sequel, the joint PMF of this more general model can still be written as in (10), with an appropriate function h. In particular, we shall work with families of distributions of $\mathbf{T}$ described by marginal CDFs $\{F_i\}$ and a copula function $C(u_1, \ldots, u_d)$. In this set-up, the PMF of $\mathbf{Y}$, which is still of the form (10), can be expressed as

$$\mathbb{P}(\mathbf{Y} = \mathbf{y}) = \prod_{i=1}^{d} g_i(y_i) \mathbb{E}\left\{c(F_1(X_1), \ldots, F_d(X_d))\right\}, \quad y_i \in \mathbb{N}_0, \quad i = 1, \ldots, d, \tag{14}$$

where the g_i are the marginal PMFs of $\{Y_i\}$, the function $c(u_1, \ldots, u_d)$ is the PDF corresponding to the copula $C(u_1, \ldots, u_d)$, and the $\{X_i\}$ are *independent* random variables with certain distributions dependent on the $\{y_i\}$. This expression, which is an analogue of (2) for discrete multivariate distributions defined through our scheme, provides a convenient way for computing probabilities of these multivariate distributions. This computational aspect of our construction compares favorably with a cumbersome formula for the PMF (see, e.g., Proposition 1.1 in Nikoloulopoulos and Karlis (2009)) of the competing method defined viz. (3).

In what follows, we explore these ideas to provide a flexible multivariate modeling framework for dependent count data — emphasizing computationally convenient expressions and scaleable algorithms for high-dimensional applications. We begin by showing how multivariate count data can be generated as mixtures of Poisson distributions by developing sequences of *independent* Poisson processes randomly stopped at an underlying continuous real-valued random variable $\mathbf{T}$ (a $\mathbf{T}-$Poisson hierarchy). Then we show how our $\mathbf{T}-$Poisson hierarchy scheme gives rise to computationally convenient joint probability mass functions (PMFs) and how particular choices of parameters/distributions can be used to construct well-known models such as the *multivariate negative binomial*. Next, we describe a scaleable simulation algorithm using our construction and copula theory. Two examples are provided: a basic example to produce a multivariate geometric distribution and an elaborate high-dimensional simulation study, aiming to model and simulate RNA-sequencing data. We note that our modeling framework and computationally-convenient formulas may facilitate novel data analysis strategies, but we do not take up that task in this current study. We conclude with an Appendix containing selected proofs of assertions made throughout.

2 Multivariate mixtures of Poisson distributions

Our goal is to produce a random vector $\mathbf{Y} = (Y_1, \ldots, Y_d)^\top$ with *correlated* mixed Poisson components. To this end, we start with a sequence of *independent* Poisson processes $\{N_i(t),\ t \in \mathbb{R}_+\}$, $i = 1, \ldots, d$, where the rate of the process $N_i(t)$ is λ_i. Next, we let

$\mathbf{T} = (T_1, \dots T_d)^\top$ have a multivariate distribution on $\mathbb{R}_+^d$ with the PDF $f_{\mathbf{T}}(\mathbf{t})$. Then, we define

$$\mathbf{Y} = (Y_1, \dots, Y_d) = (N_1(T_1), \dots, N_d(T_d)). \tag{15}$$

In the terminology of Karlis and Xekalaki (2005), this is a special case of multivariate mixed Poisson distributions of Type II. Assuming that the $\{N_i(t)\}$ are independent of $\mathbf{T}$, by standard conditioning arguments (see Lemma 7 in the Appendix) we obtain

$$\mathbb{P}(\mathbf{Y} = \mathbf{y}) = \prod_{i=1}^{d} \frac{\lambda_i^{y_i}}{y_i!} \int_{\mathbb{R}_+^d} e^{-\sum_{i=1}^d \lambda_i t_i} \prod_{i=1}^{d} t_i^{y_i} f_{\mathbf{T}}(\mathbf{t}) d\mathbf{t}. \tag{16}$$

While in some cases one can obtain explicit expressions for the above joint probabilities, in general these have to be calculated numerically. The calculations can be facilitated by certain representations of these probabilities, discussed in the Appendix (see Lemmas 7 and 8 in the Appendix).

This procedure is quite general, and leads to a multitude of multivariate discrete distributions. Flexible models allowing for marginal distributions of different types can be obtained by a popular approach with copulas. Assume that $\mathbf{T}$ has a continuous distribution on $\mathbb{R}_+^d$ with marginal PDFs f_i and CDFs F_i driven by a particular copula $C(u_1, \dots, u_d)$, so that the joint CDF of the $\{T_i\}$ is given by

$$F_{\mathbf{T}}(\mathbf{t}) = \mathbb{P}(T_1 \le t_1, \dots, T_d \le t_d) = C(F_1(t_1), \dots, F_d(t_d)), \quad \mathbf{t} = (t_1, \dots, t_d)^\top \in \mathbb{R}_+^d.$$

Then according to (2), the joint PDF $f_{\mathbf{T}}$ is of the form

$$f_{\mathbf{T}}(\mathbf{t}) = \left\{ \prod_{i=1}^{d} f_i(t_i) \right\} c(F_1(t_1), \dots, F_d(t_d)), \quad \mathbf{t} = (t_1, \dots, t_d)^\top \in \mathbb{R}_+^d, \tag{17}$$

where the function $c(u_1, \dots, u_d)$ is the PDF corresponding to the copula CDF $C(u_1, \dots, u_d)$. When we substitute (17) into (16), we get

$$\mathbb{P}(\mathbf{Y} = \mathbf{y}) = \prod_{i=1}^{d} \frac{\lambda_i^{y_i}}{y_i!} \int_{\mathbb{R}_+^d} e^{-\sum_{i=1}^d \lambda_i t_i} \prod_{i=1}^{d} \left[t_i^{y_i} f_i(t_i) \right] c(F_1(t_1), \dots, F_d(t_d)) d\mathbf{t}. \tag{18}$$

Using the results presented in the Appendix (see Lemma 7 in the Appendix), one can show that the marginal PMFs of the $\{Y_i\}$ are given by

$$\mathbb{P}(Y_i = y) = \frac{\lambda_i^y}{y!} \mathbb{E}\left[e^{-\lambda_i T_i} T_i^y \right] = \mathbb{E}\left[f_{\lambda_i T_i}(W) \right], \tag{19}$$

where $f_{\lambda_i T_i}(\cdot)$ is the PDF of $\lambda_i T_i$ and W has a standard gamma distribution with shape parameter $y + 1$. With this notation, we can write (18) in the form

$$\mathbb{P}(\mathbf{Y} = \mathbf{y}) = \prod_{i=1}^{d} \mathbb{P}(Y_i = y_i) \int_{\mathbb{R}_+^d} c(F_1(t_1), \dots, F_d(t_d)) g(\mathbf{t}|\mathbf{y}) d\mathbf{t}, \tag{20}$$

where the quantity $g(\mathbf{t}|\mathbf{y})$ in the above integral is the joint PDF of multivariate distribution with *independent* margins,

$$g(\mathbf{t}|\mathbf{y}) = \prod_{i=1}^{d} g_i(t_i|y_i) \tag{21}$$

with

$$g_i(t|y) = \frac{t^y e^{-\lambda_i t} f_i(t)}{\mathbb{E}\left[T_i^y e^{-\lambda_i T_i} \right]}, \quad t \in \mathbb{R}_+. \tag{22}$$

Thus, the integral in (20) can be expressed as

$$\int_{\mathbb{R}_+^d} c(F_1(t_1),\ldots,F_d(t_d))g(\mathbf{t}|\mathbf{y})d\mathbf{t} = \mathbb{E}\left\{c(F_1(X_1),\ldots,F_d(X_d))\right\}, \tag{23}$$

where $\mathbf{X} = (X_1,\ldots,X_d)^\top$ has a multivariate distribution with *independent* components, governed by the PDF specified by (21) - (22). This leads to the following result.

Proposition 1 *In the above setting, the joint probabilities (18) admit the representation*

$$\mathbb{P}(\mathbf{Y}=\mathbf{y}) = \prod_{i=1}^{d}\mathbb{P}(Y_i=y_i)\mathbb{E}\left\{c(F_1(X_1),\ldots,F_d(X_d))\right\}, \quad \mathbf{y}=(y_1,\ldots,y_d)^\top \in \mathbb{N}_0^d, \tag{24}$$

where the marginal probabilities are given by (19) and the PDF of $\mathbf{X} = (X_1,\ldots,X_d)^\top$ *is given by (21) - (22).*

Let us note that the joint moments of the $Y_1,\ldots Y_d$ exist whenever their counterparts of $T_1,\ldots,T_d$ are finite, in which case they can be evaluated by standard conditioning arguments. In particular, the mean and the covariance matrix of $\mathbf{Y}$ are related to their counterparts connected with $\mathbf{T}$ in a simple way, specified by Lemma 9 in the Appendix. It follows that $\mathbb{E}Y_i = \lambda_i\mathbb{E}T_i$ and $\mathbb{V}arY_i = \lambda_i\mathbb{E}T_i + \lambda_i^2\mathbb{V}arT_i$, so the distributions of the $\{Y_i\}$ are always *over-dispersed*. Moreover, we have

$$\mathbb{C}ov(Y_i,Y_j) = \lambda_i\lambda_j\mathbb{C}ov(T_i,T_j), \quad i\neq j,$$

so that the correlation coefficient of Y_i and Y_j (if it exists) is related to that of T_i and T_j as follows:

$$\rho_{Y_i,Y_j} = c_{i,j}\rho_{T_i,T_j}, \quad i\neq j, \tag{25}$$

where

$$c_{i,j} = \frac{\sqrt{\lambda_i}\sqrt{\lambda_j}}{\sqrt{\lambda_i+\frac{\mathbb{E}(T_i)}{\mathbb{V}ar(T_i)}}\sqrt{\lambda_j+\frac{\mathbb{E}(T_j)}{\mathbb{V}ar(T_j)}}}, \quad i\neq j. \tag{26}$$

Remark 1 *While in general the correlation can be positive as well as negative and admits the same range as its counterpart for* T_i *and* T_j*, the range of possible correlations of* Y_i *and* Y_j *can be further restricted if the margins are fixed. The maximum and minimum correlation can be deduced from (25) - (26) and the range of correlation corresponding to the joint distribution of* T_i *and* T_j*. The later is provided by the minimum and the maximum correlations, corresponding to the lower and the upper Fréchet copulas,*

$$C_L(u_1,u_2) = \max\{u_1+u_2-1,0\}, \quad C_U(u_1,u_2) = \min\{u_1,u_2\}, \quad u_1,u_2\in[0,1]. \tag{27}$$

The upper bound for the correlation is obtained when the distribution of (T_i,T_j) *is driven by the upper Fréchet copula* C_U *in (27), so that* $T_i \overset{d}{=} F_i^{-1}(U)$ *and* $T_j \overset{d}{=} F_j^{-1}(U)$*, where* U *is standard uniform and the* $F_i(\cdot)$*,* $F_j(\cdot)$ *are the CDFs of* T_i*,* T_j*, respectively. Similarly, the lower bound for the correlation is obtained when the distribution of* (T_i,T_j) *is driven by the lower Fréchet copula* C_L *in (27), where we have* $T_i \overset{d}{=} F_i^{-1}(U)$ *and* $T_j \overset{d}{=} F_j^{-1}(1-U)$*. While these correlation bounds are usually not available explicitly, they can be easily obtained by Monte-Carlo approximations viz. simulation from these (degenerate) probability distributions or by other standard approximate methods (see, e.g., Demitras and Hedeker (2011), and references therein).*

Remark 2 *We note that when a bivariate random vector* $\mathbf{Y} = (Y_1, Y_2)^\top$ *is defined viz. (15) and the distribution of the corresponding* $\mathbf{T} = (T_1, T_2)^\top$ *is driven by one of the copulas in (27), then the distribution of* $\mathbf{T}$ *is not absolutely continuous and the above derivations leading to the PDF of* $\mathbf{Y}$ *need a modification. It can be shown that in this case the marginal distributions of the* Y_i *are still given by (19) while the joint PMF of* $(Y_1, Y_2)^\top$ *is also as in (20) with* $d = 2$*, but with the integral term replaced with*

$$\int_0^1 g_1(u|y_1)g_2(u|y_2)du \quad \text{and} \quad \int_0^1 g_1(u|y_1)g_2(1-u|y_2)du \tag{28}$$

under the upper and the lower Fréchet copula cases, respectively, where the $g_i(\cdot|y)$ *in (28) are PDFs on* $(0, 1)$ *given by*

$$g_i(u|y) = \frac{e^{-\lambda_i F_i^{-1}(u)}\left[F_i^{-1}(u)\right]^y}{\mathbb{E}\left[e^{-\lambda_i T_i}T_i^y\right]}, \quad u \in (0,1), y \in \mathbb{N}_0, i = 1, 2. \tag{29}$$

Again, while the integrals in (28) are rarely available explicitly, they can be easily approximated by Monte-Carlo simulations in order to compute the joint PMF of $\mathbf{Y} = (Y_1, Y_2)^\top$*. These two "extreme" distributional cases can also be used to derive the full range of the values for the correlation of* $\mathbf{Y} = (Y_1, Y_2)^\top$ *when the marginal distributions (19) are fixed, if needed.*

2.1 Mixed Poisson distributions with NB margins

We now consider the case where the mixed Poisson marginal distributions of $\mathbf{Y}$ are NB, so that the marginal distributions of $\mathbf{T}$ are gamma (see Lemma 1 in Appendix). Thus, we shall assume that the coordinates of the random vectors $\mathbf{T}$ have univariate standard gamma distributions with shape parameters $r_i \in \mathbb{R}_+$, $i = 1, \ldots, d$. There have been numerous multivariate gamma distributions developed over the years, and we could use any of them here. However, we follow a general approach based on copulas, discussed above. Thus, we assume that the dependence structure of $\mathbf{T}$ is governed by some copula function $C(u_1, \ldots, u_d)$, which admits the PDF $c(u_1, \ldots, u_d)$. In this case, the f_i in (18) are given by (6) where $r = r_i$ and the F_i are the corresponding CDFs. Here, the marginal PMFs of the $\{Y_i\}$ in (19) are given by

$$\mathbb{P}(Y_i = y) = \frac{\Gamma(y + r_i)}{\Gamma(r_i)y!}p_i^{r_i}(1 - p_i)^y, \quad y \in \mathbb{N}_0, \tag{30}$$

where the NB probabilities are given by $p_i = 1/(1+\lambda_i) \in (0, 1)$ (so that $\lambda_i = (1-p_i)/p_i > 0$). Further, the PDF of $\mathbf{X}$ in Proposition 1 is still given by (21), where the marginal PDFs $g_i(\cdot|y_i)$ now admit explicit expressions

$$g_i(t|y_i) = \frac{(1 + \lambda_i)^{y_i+r_i}}{\Gamma(y_i + r_i)}t^{y_i+r_i-1}e^{-(1+\lambda_i)t}, \quad t \in \mathbb{R}_+. \tag{31}$$

We recognize that these are gamma PDFs. Thus, in this special case of multivariate mixed Poisson distributions of Type II with NB marginal distributions, the random vector $\mathbf{X}$ in the representation (14) has multivariate gamma distribution as well, but with independent margins. This fact is summarized in the result below.

Corollary 1 *Let* $\mathbf{Y}$ *have a mixed Poisson distribution defined viz. (15), where the* $\{N_i(\cdot)\}$ *are independent Poisson processes with respective rates* λ_i *and* $\mathbf{T}$ *has multivariate gamma*

distribution with standard gamma margins with shape parameters r_i and CDFs F_i, governed by a copula PDF $c(\mathbf{u})$. Then, the marginal PMFs of $\mathbf{Y}$ are given by (30) with $p_i = 1/(1+\lambda_i) \in (0,1)$ and its joint PMF is given by (14), where $\mathbf{X} = (X_1, \ldots, X_d)^\top$ has multivariate gamma distribution with independent gamma marginal distributions of the $\{X_i\}$ with PDFs given by (31).

Remark 3 *If the expectation in (14) does not admit an explicit form in terms of the $y_1, \ldots, y_d$, one can approximate its value viz. straightforward Monte-Carlo approximation involving random variate generation of independent gamma random variates $\{X_i\}$.*

Let us note that since the $\{T_i\}$ have standard gamma distributions with shape parameters r_i, we have $\mathbb{E}(T_i) = \mathbb{V}ar(T_i) = r_i$, and an application of Lemma 9 leads to the following result.

Proposition 2 *Let $\mathbf{Y}$ have a mixed Poisson distribution defined viz. (15), where the $\{N_i(\cdot)\}$ are independent Poisson processes with respective rates λ_i and $\mathbf{T}$ has multivariate gamma distribution with standard gamma margins with shape parameters r_i and CDFs F_i, governed by a copula PDF $c(\mathbf{u})$. Then, $\mathbb{E}(\mathbf{Y}) = \mathbf{I}(\boldsymbol{\lambda})\mathbf{r}$, where $\mathbf{r} = (r_1, \ldots, r_d)^\top$ and $\mathbf{I}(\boldsymbol{\lambda})$ is a $d \times d$ diagonal matrix with the $\{\lambda_i\}$ on the main diagonal. Moreover, the covariance matrix of $\mathbf{Y}$ is given by*

$$\boldsymbol{\Sigma}_{\mathbf{Y}} = \mathbf{I}(\boldsymbol{\lambda})\mathbf{I}(\mathbf{r}) + \mathbf{I}(\boldsymbol{\lambda})\boldsymbol{\Sigma}_{\mathbf{T}}\mathbf{I}(\boldsymbol{\lambda})^\top,$$

where $\boldsymbol{\Sigma}_{\mathbf{T}}$ is the covariance matrix of $\mathbf{T}$ and $\mathbf{I}(\mathbf{r})$ is a $d \times d$ diagonal matrix with the $\{r_i\}$ on the main diagonal.

Remark 4 *The correlation of Y_i and Y_j is still given by (25), where this time*

$$c_{i,j} = \sqrt{\frac{\lambda_i}{1+\lambda_i}}\sqrt{\frac{\lambda_j}{1+\lambda_j}}, \quad i \neq j,$$

since in (26) we have $\mathbb{E}(T_i) = \mathbb{V}ar(T_i)$. Let us note that while in principle the quantities $c_{i,j}$ can assume any value in $(0,1)$ when we choose appropriate λ_i and λ_j, they are fixed for particular marginal NB distributions, since in this model the NB probabilities are given by $p_i = 1/(1+\lambda_i)$. In the terms of the latter, we have

$$c_{i,j} = \sqrt{1-p_i}\sqrt{1-p_j}, \quad i \neq j.$$

These quantities, along with the full range of correlations for ρ_{T_i,T_j} in (25), can be used to obtain the upper and lower bounds for possible correlations of Y_i and Y_j in this model. We note that the possible range of ρ_{T_i,T_j} depends on the shape parameters r_i and r_j. If the $\{T_i\}$ are exponential (so that $r_i = r_j = 1$), then the upper limit of their correlation can be shown to be 1. However, the full range for the correlation of T_i and T_j is usually a subset of $[-1,1]$, which can be approximated by Monte-Carlo simulations (see Remarks 1-2) or other approximate methods (see, e.g., Demitras and Hedeker (2011)).

2.2 Simulation

One particular way of defining this model, convenient for simulations, is by using the *Gaussian copula* to generate $\mathbf{T}$. This is a very popular methodology due to its flexibility and ease of simulating from a required multivariate normal distribution. The Gaussian

copula is one that corresponds to a multivariate normal distribution with standard normal marginal distributions and covariance matrix $\mathbf{R}$. Since the marginals are standard normal, this $\mathbf{R}$ is also the correlation matrix. If $F_{\mathbf{R}}$ is the CDF of such multivariate normal distribution, then the corresponding Gaussian copula $C_{\mathbf{R}}$ is defined through

$$F_{\mathbf{R}}(x_1, \ldots, x_d) = C_{\mathbf{R}}(\Phi(x_1), \ldots, \Phi(x_d)),$$

where $\Phi(\cdot)$ is the standard normal CDF. Note that the copula $C_{\mathbf{R}}$ is simply the CDF of the random vector $(\Phi(X_1), \ldots, \Phi(X_d))^\top$, where $(X_1, \ldots, X_d)^\top \sim N_d(\mathbf{0}, \mathbf{R})$. If the distribution is continuous (so that $\mathbf{R}$ is non-singular), the copula $C_{\mathbf{R}}$ admits the PDF $c_{\mathbf{R}}$, given by

$$c_{\mathbf{R}}(u_1, \ldots, u_d) = \frac{1}{|\mathbf{R}|^{1/2}} e^{-\frac{1}{2}(\Phi^{-1}(\mathbf{u}))^T(\mathbf{R}^{-1}-\mathbf{I}_d)\Phi^{-1}(\mathbf{u})}, \quad \mathbf{u} = (u_1, \ldots, u_d)^\top \in [0,1]^d, \tag{32}$$

where $\Phi^{-1}(\mathbf{u}) = (\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_d))^\top$ and $\mathbf{I}_d$ is $d \times d$ identity matrix. This $c_{\mathbf{R}}$ will then be used in equations (20), (23), and (14). Simulation of multivariate gamma $\mathbf{T}$ with margins F_i based on this copula is quite simple, and involves the following steps:

(i) Generate $\mathbf{X} = (X_1, \ldots, X_d)^\top \sim N_d(\mathbf{0}, \mathbf{R})$;
(ii) Transform $\mathbf{X}$ to $\mathbf{U} = (U_1, \ldots, U_d)^\top$ viz $U_i = \Phi(X_i)$, $i = 1, \ldots, d$;
(iii) Return $\mathbf{T} = (T_1, \ldots, T_d)^\top$, where $T_i = F_i^{-1}(U_i)$, $i = 1, \ldots, d$;

Remark 5 *This strategy of using Gaussian copula to generate multivariate distributions is quite popular indeed, and it became known in the literature as the NORTA (**NOR**mal **T**o **A**nything) method (see, e.g., Chen (2001); Song and Hsiao (1993)). This methodology has been recently used to generate multivariate discrete distributions, see, e.g., Barbiero and Ferrari (2017), Madsen and Birkes (2013), or Nikoloulopoulos (2013) and references therein. The standard approach discussed in these papers proceeds by simulating the vector* $\mathbf{U}$ *from the Gaussian copula following the steps (i) - (ii) above and then transforming the coordinates of* $\mathbf{U}$ *directly viz. the inverse CDFs of the components of the target random vector* $\mathbf{Y} = (Y_1, \ldots, Y_d)^\top$*, which can be described as*

(iii)' Return $\mathbf{Y} = (Y_1, \ldots, Y_d)^\top$*, where* $Y_i = G_i^{-1}(U_i)$, $i = 1, \ldots, d$;

Here, the G_i *are the CDFs of the* Y_i*. If the distributions of the* Y_i *are discrete (such as NB), the inverse CDF is defined in the standard way as*

$$G^{-1}(u) = \inf\{y : G(y) \geq u\}.$$

The difference of our approach and the one discussed in the literature as described above is in the final step, regardless of the particular copula c that is used. In the standard approach one first simulates random $\mathbf{U}$ *from c and then proceeds viz. (iii)' above to get the target random vector* $\mathbf{Y}$ *(having a multivariate distribution with CDFs* G_i*). On the other hand, our proposal is to first generate* $\mathbf{T}$ *viz. step (iii) above and then obtain the target variable viz. (15). While our methodology involves an extra step compared with this direct method, it offers a simple way of calculating the joint probabilities, which is not available in the other approach. Additionally, our methodology offers a stochastic explanation of the resulting distributions viz. mixing mechanism and its relation to the underlying Poisson processes, which is lacking in the somewhat artificial standard approach. Another advantage of the approach viz. mixed Poisson are possible extensions to more general stochastic*

processes in the spirit of the NB process studied by Kozubowski and Podgórski (2009). On the other hand, its disadvantage is the fact that not all discrete marginal distributions can be obtained, only those that are mixed Poisson to begin with.

Remark 6 *Let us note that the mixed Poisson approach to generate multivariate distributions was used in Madsen and Dalthorp (2007), where* $\mathbf{Y}$ *was obtained viz. (15) with standard Poisson processes and where* $\mathbf{T} = e^{\mathbf{X}}$ *with* $\mathbf{X}$ *being multivariate normal with mean* $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_d)^\top$ *and covariance matrix* $\boldsymbol{\Sigma} = [\sigma_{i,j}]$. *Since in this case the marginals of* $\mathbf{T}$ *have log-normal distributions, the authors referred to this construction as lognormal-Poisson hierarchy. This can be seen as a special case our scheme, where we have* $\lambda_i = e^{\mu_i}$ *and the marginal CDFs of* T_i *of the form* $F_i(t) = \Phi(\log t_i/\sigma_{ii})$. *The copula PDF of the* $\{T_i\}$ *is the Gaussian copula (32) where* $\mathbf{R}$ *is the correlation matrix corresponding to* $\boldsymbol{\Sigma}$.

An important aspect of this problem is how to set the parameters of the underlying copula function so that the distribution of $\mathbf{Y}$ has given characteristics, such as the means and the covariances (and correlations). In the case where a Gaussian copula is used, this has to do with determining the correlation matrix $\mathbf{R}$. This problem arises in the general scheme (i)—(iii) as well — and has been discussed in the literature (see, e.g., Barbiero and Ferrari (2017); Xiao (2017); Xiao and Zhou (2019)). Generally, there is no simple relation between $\mathbf{R}$ and the correlation matrix of $\mathbf{T}$ in (i)—(iii). However, other measure of associations - such as Kendall's τ or Spearman's ρ do transfer directly and may be preferred to use in our set-up. These issues will be the subject of further research.

3 Examples

We provide two examples. The first example describes the $\mathbf{T}$-Poisson hierarchy approach to construct a multivariate geometric distribution. Second, we demonstrate how the $\mathbf{T}$-Poisson hierarchy can be used to conduct a high-dimensional ($d = 1026$) simulation study inspired by RNA-sequencing data — a challenging computational task.

3.1 Multivariate geometric distributions

Suppose that the random vector $\mathbf{T}$ in (15) has marginal standard exponential distributions, so that the marginal CDFs of the $\{T_i\}$ are of the form

$$F_i(t) = 1 - e^{-t}, \quad t \in \mathbb{R}_+. \tag{33}$$

In this case, the $\{Y_i\}$ have geometric distributions with parameters $p_i = 1/(1+\lambda_i)$, so that

$$\mathbb{P}(Y_i = y) = p_i(1-p_i)^y, \quad y \in \mathbb{N}_0. \tag{34}$$

One can then obtain a multitude of multivariate distributions with geometric margins by selecting various copulas for the underlying distributions of $\mathbf{T}$. As an example, consider the case with *Farlie-Gumbel-Morgenstern* (FGM) copula driven by a parameter $\theta \in [-1, 1]$, given by

$$C(\mathbf{u}) = \prod_{i=1}^{d} u_i \left(1 + \theta \prod_{i=1}^{d} (1-u_i)\right), \quad \mathbf{u} = (u_1, \ldots, u_d)^\top \in [0,1]^d. \tag{35}$$

Consider a two dimensional case with $d = 2$, where the PDF corresponding to (35) is of the form

$$c(\mathbf{u}) = 1 + \theta(1-2u_1)(1-2u_2), \quad \mathbf{u} = (u_1, u_2)^\top \in [0,1]^2. \tag{36}$$

In this case, the random vector $\mathbf{X} = (X_1, X_2)^\top$ in Corollary 1 has independent gamma margins (31) with shape parameters $y_i + 1$ and scale parameters $1 + \lambda_i$, $i = 1, 2$. Using this fact, coupled with (33), one can evaluate the expectation in (14), leading to

$$\mathbb{E}\left\{c(F_1(X_1), F_2(X_2))\right\} = 1 + \theta\left[1 - 2\left(\frac{1}{1+p_1}\right)^{y_1+1}\right]\left[1 - 2\left(\frac{1}{1+p_2}\right)^{y_2+1}\right]. \tag{37}$$

In view of Corollary 1, this leads to the following expression for the joint probabilities of bivariate geometric distribution defined by our scheme viz. FGM copula:

$$\mathbb{P}(\mathbf{Y} = \mathbf{y}) = \prod_{i=1}^{2} p_i(1-p_i)^{y_i}\left\{1 + \theta\prod_{i=1}^{2}\left[1 - 2\left(\frac{1}{1+p_i}\right)^{y_i+1}\right]\right\}, \quad \mathbf{y} = (y_1, y_2)^\top \in \mathbb{N}_0^2. \tag{38}$$

We shall denote this distribution by $GEO(p_1, p_2, \theta)$. When $\theta = 0$, the $\{Y_i\}$ are independent geometric variables with parameters $p_i \in (0,1)$, $i = 1, 2$. Otherwise, Y_1, Y_2 are correlated, with

$$\mathbb{C}ov(Y_1, Y_2) = \frac{\theta}{4}\frac{1-p_1}{p_1}\frac{1-p_2}{p_2}, \tag{39}$$

as can be verified by routine, albeit tedious, algebra. In turn, the correlation of Y_1, Y_2 becomes

$$\rho_{Y_1,Y_2} = \frac{\theta}{4}\sqrt{1-p_1}\sqrt{1-p_2}, \tag{40}$$

and can generally take any value in the range $(-1/4, 1/4)$.

3.2 Simulating RNA-seq data

This section describes how to simulate data using a **T**-Poisson hierarchy, aiming to replicate the structure of high-dimensional dependent count data. In fact, simulating RNA-sequencing (RNA-seq) data is a one of the primary motivating applications of the proposed methodology, seeking scaleable Monte Carlo methods for realistic multivariate simulation (for example, see Schissler et al. (2018)).

The RNA-seq data generating process involves counting how often a particular messenger RNA (mRNA) is expressed in a biological sample. Since this is a counting process with no upper bound, many modeling approaches use discrete random variables with infinite support. Often the counts exhibit over-dispersion and so the negative binomial arises as a sensible model for the expression levels (*gene counts*). Moreover, the counts are correlated (co-expressed) and cannot be assumed to behave independently. RNA-seq platforms quantify the entire transcriptome in one experimental run, resulting in high-dimensional data. In humans, this results in count data corresponding to over 20,000 genes (coding genomic regions) or even over 77,000 isoforms when alternating spliced mRNA are counted. This suggests simulating high-dimensional multivariate NB with heterogeneous marginals would be useful tool in the development and evaluation of RNA-seq analytics.

In an illustration of our proposed methodology applied to real data, we seek to simulate RNA-sequencing data by producing simulated random vectors generated from the Type II

T-Poisson framework (as in Eq. (13)). Our goal is to replicate the structure of a breast cancer data set (BRCA: breast cancer invasive carcinoma data set from The Cancer Genome Atlas). For simplicity, we begin by filtering to retain the top 5% highest expressing genes of the 20,501 gene measurements from $N = 1212$ patients' tumor samples, resulting in $d = 1026$ genes. All these genes exhibit over-dispersion and, so, we proceed to estimate the NB parameters $(r_i, p_i), i = 1, \ldots, d$, to determine the target marginal PMFs $g_i(y_i)$ (via method of moments). Notably, the $\hat{p}_i's$ are small — ranging in $[\,3.934 \times 10^{-6}, 1.217 \times 10^{-2}]$. To complete the simulation algorithm inputs, we estimate the Pearson correlation matrix $\mathbf{R_Y}$ and set that as the target correlation.

With the simulation targets specified, we proceed to simulate $B = 10,000$ random vectors $\mathbf{Y} = (Y_1, \ldots, Y_d)^\top$ with target Pearson correlation $\mathbf{R_Y}$ and marginal PMFs $g_i(y_i)$ using a **T**-Poisson hierarchy of Kind II. Specifically, we first employ the *direct Gaussian copula* approach to generate B random vectors following a standard multivariate Gamma distribution **T** with shape parameters r_i equal to the target NB sizes and Pearson correlation matrix $\mathbf{R_T}$. Care must be taken when setting the specifying **R** (refer to Eq. (32)) — we employ Eq. (25) to compute the scaling factors $c_{i,j}$ and adjust the underlying correlations to ultimately match the target $\mathbf{R_Y}$. Notably, of the 525,825 pairwise correlations from the 1026 genes, no scale factor was less than 0.9907, indicating the model can produce essentially the entire range of possible correlations. Here we are satisfied with approximate matching of the specified Gamma correlation and set $\mathbf{R} = \mathbf{R_T}$ in our Gaussian copula scheme (**R** indicating the specified multivariate Gaussian correlation matrix). Finally, we generate the desired random vector $Y_i = N_i(T_i)$ by simulating Poisson counts with expected value $\mu_i = \lambda_i \times T_i$, for $i = 1, \ldots, d$, (with $\lambda_i = \frac{(1-p_i)}{p_i}$) and repeat $B = 10,000$ times.

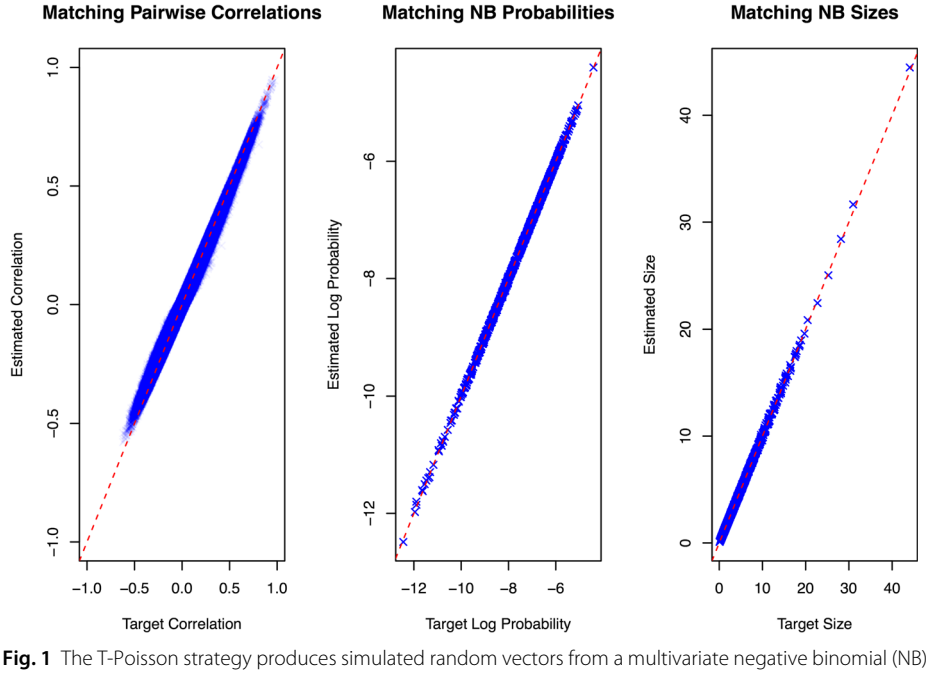

Fig. 1 The T-Poisson strategy produces simulated random vectors from a multivariate negative binomial (NB) that replicate the estimated structure from an RNA-seq data set. The dashed red lines indicated equality between estimated parameters (vertical axes; derived from the simulated data) and the specified target parameters (horizontal axes)

Figure 1 shows the results of our simulation by comparing the specified target parameter (horizontal axes) with the corresponding quantities estimated from the simulated data (vertical axes). The evaluation shows that the simulated counts approximately match the target parameters and exhibit the full range of estimated correlation from the data. Utilizing 15 CPU threads in a MacBook Pro carrying a 2.4 GHz 8-Core Intel Core i9 processor, the simulation completed in less than 30 seconds.

Appendix

Gamma-Poisson mixtures

For the convenience of the reader, we include a short proof of the well-known fact stating that Poisson distribution with gamma-distributed parameter is NB (see, e.g., Solomon (1983)).

Lemma 1 *If* $\{N(t), t \in \mathbb{R}_+\}$ *is a homogeneous Poisson process with rate* $\lambda = (1-p)/p > 0$ *and* T *is an independent standard gamma variable with shape parameter* r*, then the randomly stopped process,* $Y = N(T)$*, has a NB distribution* $NB(r,p)$ *with the PMF (7).*

Proof Suppose that T has a standard gamma distribution with the PDF (6) and the corresponding CDF F_T. When we substitute the latter into (5), we obtain

$$\mathbb{P}(Y = n) = \int_{\mathbb{R}_+} \frac{e^{-\lambda t}(\lambda t)^n}{n!} \frac{1}{\Gamma(r)} t^{r-1} e^{-t} dt.$$

After some algebra, this produces

$$\mathbb{P}(Y = n) = \frac{\Gamma(n+r)}{\Gamma(r)n!} \frac{\lambda^n}{(1+\lambda)^{n+r}} \int_{\mathbb{R}_+} \frac{(1+\lambda)^{n+r}}{\Gamma(n+r)} t^{n+r-1} e^{-t(1+\lambda)} dt.$$

Since the integrand above is the PDF of gamma distribution with shape $n+r$ and scale $1+\lambda$, the integral becomes 1 and we have

$$\mathbb{P}(Y = n) = \frac{\Gamma(n+r)}{\Gamma(r)n!} \left(\frac{1}{1+\lambda}\right)^r \left(\frac{\lambda}{1+\lambda}\right)^n,$$

which we recognize as the NB probability from (7) with $p = (1+\lambda)^{-1}$. The result follows when we set $\lambda = (1-p)/p$ in the above analysis. □

Mixed multivariate Poisson distributions of type I

Here we provide basic distributional facts about mixed multivariate Poisson distributions of Type I, which are the distributions of $\mathbf{Y} = (Y_1, \ldots, Y_d)^\top = (N_1(T), \ldots, N_d(T))^\top$, where the $\{N_i(\cdot)\}$ are independent Poisson processes with rates λ_i and T is a random variable on $\mathbb{R}_+$, independent of the $\{N_i\}$.

Lemma 2 *In the above setting, the PGF of* $\mathbf{Y}$ *is given by*

$$G(\mathbf{s}) = \mathbb{E}\left\{\prod_{i=1}^d s_i^{Y_i}\right\} = \phi_T\left(\sum_{i=1}^d \lambda_i - \sum_{i=1}^d \lambda_i s_i\right), \quad \mathbf{s} = (s_1, \ldots, s_d)^\top \in [0,1]^d,$$

where ϕ_T *is the LT of* T*.*

Proof By using standard conditioning argument, we have

$$G(\mathbf{s}) = \mathbb{E}\left\{\prod_{i=1}^{d} s_i^{Y_i}\right\} = \int_{\mathbb{R}_+} \mathbb{E}\left\{\prod_{i=1}^{d} s_i^{Y_i} \,\middle|\, T = t\right\} dF_T(t). \tag{41}$$

Since given $T = t$ the variables $\{Y_i\}$ are independent and Poisson distributed with means $\{\lambda_i t\}$, respectively, we have

$$\mathbb{E}\left\{\prod_{i=1}^{d} s_i^{Y_i} \,\middle|\, T = t\right\} = \prod_{i=1}^{d} \mathbb{E}\left\{s^{Y_i} \,\middle|\, T = t\right\} = \prod_{i=1}^{d} e^{-\lambda_i t(1-s_i)} = e^{-t\left(\sum_{i=1}^{d}\lambda_i - \sum_{i=1}^{d}\lambda_i s_i\right)}.$$

When we substitute the above into (41) we conclude that the PGF of $\mathbf{Y}$ is indeed of the form stated above. □

Remark 7 *Note that in the dimensional case $d = 1$, we recover the well-known formula for the PGF of $Y = N(T)$,*

$$G(s) = \phi_T(\lambda(1-s)), \quad s \in [0,1], \tag{42}$$

where $\lambda > 0$ is the rate of the Poisson process $\{N(t),\ t \in \mathbb{R}_+\}$. If we further assume that T is standard gamma distributed with shape parameter $r > 0$, so that

$$\phi_T(t) = \left(\frac{1}{1+t}\right)^r, \quad t \in \mathbb{R}_+,$$

and we take $\lambda = (1-p)/p$, we obtain

$$G(s) = \left(\frac{p}{1-(1-p)s}\right)^r, \quad s \in [0,1]. \tag{43}$$

We recognize this as the PGF of the NB distribution $NB(r,p)$, as it should be according to Lemma 1. Similarly, the PGF of a d-dimensional mixed Poisson distribution with such a gamma distributed T takes on the form

$$G(\mathbf{s}) = \left(\frac{1}{Q - \sum_{i=1}^{d} P_i s_i}\right)^r, \quad \mathbf{s} = (s_1, \ldots, s_d)^\top \in [0,1]^d,$$

where $P_i = \lambda_i$ and $Q = 1 + \sum_{i=1}^{d} P_i$. This is a general form of multivariate negative multinomial distribution (see Chapter 36 of Johnson et al. (1997)). Since the PGF of the marginal distributions of Y_i in this setting is of the form (43) with $p = (1+\lambda_i)^{-1}$, all marginal distributions are NB. Due to this property, discrete multivariate distributions with the above PGFs have been termed multivariate NB distributions (for more details, see Johnson et al. (1997)).

Remark 8 *Let us note that changing a scaling factor of the variable T in this model has the same effect as adjusting the rate parameters connected with the Poisson processes $\{N_i(\cdot)\}$. Namely, it follows from Lemma 2 that if we let $\tilde{T} = cT$ in the above setting, then we have the following equality in distribution:*

$$\left(N_1\left(\tilde{T}\right), \ldots, N_d\left(\tilde{T}\right)\right)^\top \overset{d}{=} \left(\tilde{N}_1(T), \ldots, \tilde{N}_d(T)\right)^\top, \tag{44}$$

where the $\{\tilde{N}_i(\cdot)\}$ are independent Poisson processes with rates $c\lambda_i$, respectively. Thus, without loss of generality, we may assume that the scale parameter of the variable T in this model is set to unity.

Lemma 3 *In the above setting, the PMF of* $\mathbf{Y}$ *is given by*

$$\mathbb{P}(\mathbf{Y}=\mathbf{y})=\prod_{i=1}^{d} g_i(y_i)h(\mathbf{y}),\quad \mathbf{y}=(y_1,\ldots,y_d)^\top\in\mathbb{N}_0^d,$$

where

$$g_i(y)=\frac{\lambda_i^y}{y!}\nu_T(y,\lambda_i),\quad y\in\mathbb{N}_0,$$

are the marginal PMFs of the $\{Y_i\}$,

$$\nu_T(y,\lambda)=\mathbb{E}\left\{T^y e^{-\lambda T}\right\},\quad \lambda,y\in\mathbb{R}_+,$$

and the function h is given by

$$h(\mathbf{y})=\frac{\nu_T\left(\sum_{i=1}^d y_i,\sum_{i=1}^d\lambda_i\right)}{\prod_{i=1}^d\nu_T(y_i,\lambda_i)},\quad \mathbf{y}=(y_1,\ldots,y_d)^\top\in\mathbb{N}_0^d.$$

Proof Since given $T=t$ the variables $\{Y_i\}$ are independent and Poisson distributed with means $\{\lambda_i t\}$, respectively, by using standard conditioning argument, followed by some algebra, we have

$$\mathbb{P}(\mathbf{Y}=\mathbf{y})=\prod_{i=1}^d\frac{\lambda_i^{y_i}}{y_i!}\int_{\mathbb{R}_+}e^{-t\sum_{i=1}^d\lambda_i}t^{\sum_{i=1}^d y_i}dF_T(t)=\left[\prod_{i=1}^d\frac{\lambda_i^{y_i}}{y_i!}\right]\left[\nu_T\left(\sum_{i=1}^d y_i,\sum_{i=1}^d\lambda_i\right)\right]. \tag{45}$$

Similarly, the marginal PMFs are given by

$$\mathbb{P}(Y_i=y)=\frac{\lambda_i^y}{y!}\int_{\mathbb{R}_+}e^{-t\lambda_i}t^y dF_T(t)=\frac{\lambda_i^y}{y!}\nu_T\left(y,\lambda_i\right). \tag{46}$$

By combining (45) and (46), we obtain the result. □

Remark 9 *Note that the joint PMF of* $\mathbf{Y}$ *can be also written as*

$$\mathbb{P}(\mathbf{Y}=\mathbf{y})=\nu_T\left(\sum_{i=1}^d y_i,\sum_{i=1}^d\lambda_i\right)\prod_{i=1}^d\frac{\lambda_i^{y_i}}{y_i!}, \tag{47}$$

which is a convenient expression for approximating these probabilities by Monte Carlo simulations if the function $\nu_T(\cdot,\cdot)$ *is not available explicitly and the random variable T is straightforward to simulate. We also note that whenever the marginal PMFs of* Y_i *are explicit, then so is the function* $\nu_T(\cdot,\cdot)$*, which is clear from Lemma 3. For example, if T is standard gamma with shape parameter r, then we have*

$$\nu_T(y,\lambda)=\frac{\Gamma(r+y)}{\Gamma(r)}\left(\frac{1}{1+\lambda}\right)^{r+y}=\frac{y!}{\lambda^y}\mathbb{P}(Y=y),\quad \lambda,y\in\mathbb{R}_+,$$

where Y has a NB distribution with parameters r and $p=1/(1+\lambda)$.

Next, we present an alternative expression for the joint probabilities $P(\mathbf{Y}=\mathbf{y})$, which provides a convenient formula for their computation whenever the variable T is difficult to simulate but its PDF is easy to compute. This representation involves a multinomial random vector $\mathbf{N}=(N_1,\ldots,N_d)^\top$ with parameters n and $\mathbf{p}=(p_1,\ldots,p_d)^\top$, denoted

by $MUL(n, \mathbf{p})$, where $n \in \mathbb{N}$ represents the number of trials, the $\{p_i\}$ represent event probabilities that sum up to one, and

$$\mathbb{P}(\mathbf{N} = \mathbf{y}) = \frac{n!}{y_1! \cdots y_d!} p_1^{y_1} \cdots p_d^{y_d}, \quad \mathbf{y} \in \left\{ \mathbf{k} = (k_1, \ldots, k_d)^d \in \mathbb{N}_0^d : \sum_{i=1}^{d} k_i = n \right\}. \quad (48)$$

Lemma 4 *In the above setting, the PMF of* $\mathbf{Y}$ *is given by*

$$\mathbb{P}(\mathbf{Y} = \mathbf{y}) = \mathbb{P}(\mathbf{N} = \mathbf{y}) \mathbb{E}\left(f_{\lambda T}(W)\right), \quad \mathbf{y} = (y_1, \ldots, y_d)^\top \in N_0^d, \quad (49)$$

where $\lambda = \sum_{i=1}^{d} \lambda_i$, $\mathbf{N} \sim MUL(n, \mathbf{p})$ *with* $n = \sum_{i=1}^{d} y_i$ *and* $p_i = \lambda_i/\lambda$, *the quantity* $f_{\lambda T}$ *is the PDF of* λT, *and* W *has standard gamma distribution with shape parameter* $n + 1$.

Proof Proceeding as in the proof of Lemma 3, we obtain

$$\mathbb{P}(\mathbf{Y} = \mathbf{y}) = \prod_{i=1}^{d} \frac{\lambda_i^{y_i}}{y_i!} \frac{n!}{\lambda^n} \frac{1}{\lambda} \int_{\mathbb{R}_+} \frac{\lambda^{n+1}}{n!} t^{(n+1)-1} e^{-\lambda t} f_T(t) dt. \quad (50)$$

Since the integrand is the product of $f_T(t)$ and the density of gamma random variable X with shape parameter $n + 1$ and scale λ, we have

$$\frac{1}{\lambda} \int_{\mathbb{R}_+} \frac{\lambda^{n+1}}{n!} t^{(n+1)-1} e^{-\lambda t} f_T(t) dt = \mathbb{E}\left[\frac{1}{\lambda} f_T(X)\right] = \mathbb{E}\left[\frac{1}{\lambda} f_T\left(\frac{W}{\lambda}\right)\right],$$

where $W = \lambda X$ has standard gamma distribution with shape parameter $n + 1$ (and scale 1). To conclude the result, observe that the expression

$$\prod_{i=1}^{d} \frac{\lambda_i^{y_i}}{y_i!} \frac{n!}{\lambda^n}$$

in (50) coincides with the multinomial probablity (48) with $p_i = \lambda_i/\lambda$ while

$$\frac{1}{\lambda} f_T\left(\frac{w}{\lambda}\right) = f_{\lambda T}(w).$$

□

Remark 10 *Note that in one dimensional case where* $d = 1$ *the multinomial probability in (49) reduces to 1, and we obtain*

$$\mathbb{P}(Y = y) = \mathbb{E}\left(f_{\lambda T}(W)\right), \quad y \in N_0, \quad (51)$$

where $Y \overset{d}{=} N(T)$, $\{N(t),\ t \in \mathbb{R}_+\}$ *is a Poisson process with rate* λ, *the quantity* $f_{\lambda T}$ *is the PDF of* λT, *the variable* W *has standard gamma distribution with shape parameter* $y + 1$, *and* T *is independent of the Poisson process.*

Finally, we present well-known results concerning the mean and the covariance structure of mixed multivariate Poisson distributions of Type I, which are easily derived through standard conditioning arguments. Generally, whenever the mean of T exists then so does the mean of each Y_i, and we have $\mathbb{E}(Y_i) = \lambda_i \mathbb{E}(T)$. Moreover, the variance of each Y_i is finite whenever T has a finite second moment, in which case we have $\mathbb{V}ar(Y_i) = \lambda_i \mathbb{E}(T) + \lambda_i^2 \mathbb{V}ar(T)$. Thus, the variance of Y_i is greater than the mean, and the distribution of Y_i is *over-dispersed*. Finally, under the latter assumption, the covariance of Y_i and Y_j exists and equals $\mathbb{C}ov(Y_i, Y_j) = \lambda_i \lambda_j \mathbb{V}ar(T)$. The result below summarizes these facts.

Lemma 5 *In the above setting, the mean vector of* $\mathbf{Y}$ *exists whenever the mean of* T *is finite, in which case we have* $\mathbb{E}(\mathbf{Y}) = \boldsymbol{\lambda}\mathbb{E}(T)$, *where* $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_d)^\top$. *Moreover, if* T *has a finite second moment, then the covariance matrix of* $\mathbf{Y}$ *is well defined and is given by*

$$\boldsymbol{\Sigma} = \mathbb{E}(T)\mathbf{I}(\boldsymbol{\lambda}) + \mathbb{V}ar(T)\boldsymbol{\lambda}\boldsymbol{\lambda}^\top,$$

where $\mathbf{I}(\boldsymbol{\lambda})$ *is a* $d \times d$ *diagonal matrix with the* $\{\lambda_i\}$ *on the main diagonal.*

Remark 11 *By the above result, if it exists, the correlation coefficient of* Y_i *and* Y_j *is given by*

$$\rho_{i,j} = \frac{\sqrt{\lambda_i}\sqrt{\lambda_j}}{\sqrt{\lambda_i + \frac{\mathbb{E}(T)}{\mathbb{V}ar(T)}}\sqrt{\lambda_j + \frac{\mathbb{E}(T)}{\mathbb{V}ar(T)}}}.$$

The correlation is always positive, and can generally fall anywhere within the boundaries of zero and one.

Mixed multivariate Poisson distributions of type II

Here we provide basic distributional facts about mixed multivariate Poisson distributions of Type II, which are the distributions of $\mathbf{Y} = (Y_1, \ldots, Y_d)^\top = (N_1(T_1), \ldots, N_d(T_d))^\top$, where the $\{N_i(\cdot)\}$ are independent Poisson processes with rates λ_i and $\mathbf{T} = (T_1, \ldots T_d)^\top$ is a random vector in $\mathbb{R}_+^d$ with the PDF $f_{\mathbf{T}}$, independent of the $\{N_i\}$.

Lemma 6 *In the above setting, the PGF of* $\mathbf{Y}$ *is given by*

$$G(\mathbf{s}) = \mathbb{E}\left\{\prod_{i=1}^d s_i^{Y_i}\right\} = \phi_{\mathbf{T}}(\mathbf{I}(\boldsymbol{\lambda})(\mathbf{1} - \boldsymbol{s})), \quad \mathbf{s} = (s_1, \ldots, s_d)^\top \in [0,1]^d, \tag{52}$$

where $\phi_{\mathbf{T}}$ *is the LT of* $\mathbf{T}$, $\mathbf{I}(\boldsymbol{\lambda})$ *is a* $d \times d$ *diagonal matrix with the* $\{\lambda_i\}$ *on the main diagonal, and* $\mathbf{1}$ *is a d-dimensional column vector of 1s.*

Proof By using standard conditioning argument, we have

$$G(\mathbf{s}) = \mathbb{E}\left\{\prod_{i=1}^d s_i^{Y_i}\right\} = \int_{\mathbb{R}_+^d} \mathbb{E}\left\{\prod_{i=1}^d s_i^{Y_i} \middle| \mathbf{T} = \mathbf{t}\right\} dF_{\mathbf{T}}(\mathbf{t}). \tag{53}$$

Since given $\mathbf{T} = \mathbf{t}$ the variables $\{Y_i\}$ are independent and Poisson distributed with means $\{\lambda_i t_i\}$, respectively, we have

$$\mathbb{E}\left\{\prod_{i=1}^d s_i^{Y_i} \middle| \mathbf{T} = \mathbf{t}\right\} = \prod_{i=1}^d \mathbb{E}\left\{s^{Y_i} \middle| \mathbf{T} = \mathbf{t}\right\} = \prod_{j=1}^d e^{-\lambda_i t_j (1 - s_j)} = e^{-\mathbf{t}^\top \mathbf{I}(\boldsymbol{\lambda})(\mathbf{1} - \boldsymbol{s})}.$$

When we substitute the above into (53) we conclude that the PGF of $\mathbf{Y}$ is as stated in the lemma. □

Remark 12 *Note that the expression (52) is a generalization of (42) to the multivariate case of mixed Poisson. Additionally, observe that if the components of* $\mathbf{T}$ *coincide, that is* $T_i = T$ *for* $i = 1, \ldots d$, *we have*

$$\phi_{\mathbf{T}}(\mathbf{t}) = \mathbb{E}\left(e^{-\mathbf{t}^\top \mathbf{T}}\right) = \mathbb{E}\left(e^{-(t_1 + \cdots + t_d)T}\right) = \phi_T(t_1 + \cdots + t_d),$$

and the PGF in (52) reduces to its counterpart provided in Lemma 2, as it should.

Remark 13 *Let us note that changing scaling factors of the variables T_i in this model has the same effect as adjusting the rate parameters connected with the Poisson processes $\{N_i(\cdot)\}$. Namely, it follows from Lemma 6 that if we let $\tilde{T}_i = c_i T_i$ in the above setting, then we have the following equality in distribution:*

$$\left(N_1(\tilde{T}_1), \ldots, N_d(\tilde{T}_d)\right)^\top \stackrel{d}{=} \left(\tilde{N}_1(T_1), \ldots, \tilde{N}_d(T_d)\right)^\top, \tag{54}$$

where the $\{\tilde{N}_i(\cdot)\}$ are independent Poisson processes with rates $c_i\lambda_i$, respectively. Thus, without loss of generality, we may assume that the scale parameters of the variables T_i in this model are set to unity.

Next, we provide a convenient formula for the PMF of multivariate mixed Poisson distributions of Type II, which is an extension of that given in Lemma 3. To state the result, we extend the definition of the function ν_T described by (12) to vector-valued arguments and random vectors $\mathbf{T}$ in $\mathbb{R}_+^d$. Namely, for $\mathbf{a} = (a_1, \ldots, a_d)^\top, \mathbf{b} = (b_1, \ldots, b_d)^\top \in \mathbb{R}_+^d$ we set

$$\mathbf{a}^{\mathbf{b}} = \prod_{i=1}^{d} a_i^{b_i} \tag{55}$$

and define

$$\nu_{\mathbf{T}}(\mathbf{y}, \boldsymbol{\lambda}) = \mathbb{E}\left\{\mathbf{T}^{\mathbf{y}} e^{-\boldsymbol{\lambda}^\top \mathbf{T}}\right\}, \quad \boldsymbol{\lambda}, \mathbf{y} \in \mathbb{R}_+^d. \tag{56}$$

Lemma 7 *In the above setting, the PMF of $\mathbf{Y}$ is given by*

$$\mathbb{P}(\mathbf{Y} = \mathbf{y}) = \prod_{i=1}^{d} g_i(y_i) h(\mathbf{y}), \quad \mathbf{y} = (y_1, \ldots, y_d)^\top \in \mathbb{N}_0^d,$$

where

$$g_i(y) = \frac{\lambda_i^y}{y!} \nu_{T_i}(y, \lambda_i), \quad y \in \mathbb{N}_0,$$

are the marginal PMFs of the $\{Y_i\}$ and the function h is given by

$$h(\mathbf{y}) = \frac{\nu_{\mathbf{T}}(\mathbf{y}, \boldsymbol{\lambda})}{\prod_{i=1}^{d} \nu_{T_i}(y_i, \lambda_i)}, \quad \mathbf{y} = (y_1, \ldots, y_d)^\top \in \mathbb{N}_0^d.$$

Proof By using standard conditioning argument, we have

$$\mathbb{P}(\mathbf{Y} = \mathbf{y}) = \int_{\mathbb{R}_+^d} \mathbb{P}(N_1(T_1) = y_1, \ldots, N_d(T_d) = y_d | \mathbf{T} = \mathbf{t}) f_{\mathbf{T}}(\mathbf{t}) d\mathbf{t}, \tag{57}$$

where $\mathbf{y} = (y_1, \ldots, y_d)^\top$ and $\mathbf{t} = (t_1, \ldots, t_d)^\top$. Further, by independence, we have

$$\mathbb{P}(N_1(T_1) = y_1, \ldots, N_d(T_d) = y_d | \mathbf{T} = \mathbf{t}) = \prod_{i=1}^{d} \mathbb{P}(N_i(t_i) = y_i). \tag{58}$$

Since the $N_i(t_i)$ are Poisson with parameters $\lambda_i t_i$, we have

$$\mathbb{P}(N_i(t_i) = y_i) = \frac{e^{-\lambda_i t_i} (\lambda_i t_i)^{y_i}}{y_i!}, \quad i = 1, \ldots, d. \tag{59}$$

When we now substitute (58) - (59) into (57), then after some algebra we get

$$\mathbb{P}(\mathbf{Y} = \mathbf{y}) = \prod_{i=1}^{d} \frac{\lambda_i^{y_i}}{y_i!} \int_{\mathbb{R}_+^d} e^{-\sum_{i=1}^{d} \lambda_i t_i} \prod_{i=1}^{d} t_i^{y_i} f_{\mathbf{T}}(\mathbf{t}) d\mathbf{t} = \left[\prod_{i=1}^{d} \frac{\lambda_i^{y_i}}{y_i!}\right] \left[\nu_{\mathbf{T}}(\mathbf{y}, \boldsymbol{\lambda})\right]. \tag{60}$$

Similarly, the marginal PMFs are given by

$$\mathbb{P}(Y_i = y) = \frac{\lambda_i^y}{y!} \int_{\mathbb{R}_+} e^{-t\lambda_i} t^y dF_{T_i}(t) = \frac{\lambda_i^y}{y!} \nu_{T_i}(y, \lambda_i). \tag{61}$$

By combining (60) and (61), we obtain the result. □

We now present an alternative expression for the joint probabilities $P(\mathbf{Y} = \mathbf{y})$, which facilitates their computation if the random vector $\mathbf{T}$ is difficult to simulate but its PDF is readily available.

Lemma 8 *In the above setting, the PMF of* $\mathbf{Y}$ *is given by*

$$\mathbb{P}(\mathbf{Y} = \mathbf{y}) = \mathbb{E}\left(f_{\mathbf{I}(\boldsymbol{\lambda})\mathbf{T}}(\mathbf{W})\right), \quad \mathbf{y} = (y_1, \ldots, y_d)^\top \in \mathbb{N}_0^d, \tag{62}$$

where the quantity $f_{\mathbf{I}(\boldsymbol{\lambda})\mathbf{T}}$ *is the PDF of* $\mathbf{I}(\boldsymbol{\lambda})\mathbf{T} = (\lambda_1 T_1, \ldots, \lambda_d T_d)^\top$ *and* $\mathbf{W} = (W_1, \ldots, W_d)^\top$ *with mutually independent* W_i *having standard gamma distributions with shape parameters* $y_i + 1$.

Proof Proceeding as in the proof of Lemma 4, we obtain

$$\mathbb{P}(\mathbf{Y} = \mathbf{y}) = \prod_{i=1}^d \frac{1}{\lambda_i} \int_{\mathbb{R}_+^d} \prod_{i=1}^d \left\{ \frac{\lambda_i^{y_i+1}}{y_i!} t_i^{(y_i+1)-1} e^{-\lambda_i t_i} \right\} f_{\mathbf{T}}(\mathbf{t}) d\mathbf{t}. \tag{63}$$

Note that the product under the integral above is the PDF of $\mathbf{X} = (X_1, \ldots, X_d)^\top$, where the X_i are mutually independent gamma random variables with shape parameters $y_i + 1$ and scale parameters λ_i. This allows us to conclude that

$$\mathbb{P}(\mathbf{Y} = \mathbf{y}) = \mathbb{E}\left[\prod_{i=1}^d \frac{1}{\lambda_i} f_{\mathbf{T}}(\mathbf{X})\right] = \mathbb{E}\left[\prod_{i=1}^d \frac{1}{\lambda_i} f_{\mathbf{T}}\left(\frac{W_1}{\lambda_1}, \ldots, \frac{W_d}{\lambda_d}\right)\right],$$

where $\mathbf{W} = (W_1, \ldots, W_d)^\top = \mathbf{I}(\boldsymbol{\lambda})\mathbf{X}$ has independent standard gamma components with shape parameters $y_i + 1$. To conclude the result, observe that

$$\prod_{i=1}^d \frac{1}{\lambda_i} f_{\mathbf{T}}\left(\frac{W_1}{\lambda_1}, \ldots, \frac{W_d}{\lambda_d}\right) = f_{\mathbf{I}(\boldsymbol{\lambda})\mathbf{T}}(\mathbf{W}).$$

□

Finally, let us summarize standard results concerning the mean and the covariance structure of mixed multivariate Poisson distributions of Type II, which parallel the results for Type I, and are easily derived through standard conditioning arguments. Generally, whenever the means of $\{T_i\}$ exist then so do the means of the $\{Y_i\}$, and we have $\mathbb{E}(Y_i) = \lambda_i \mathbb{E}(T_i)$. Similarly, the variance of each Y_i is finite whenever T_i has a finite second moment, in which case we have $\mathbb{V}ar(Y_i) = \lambda_i \mathbb{E}(T_i) + \lambda_i^2 \mathbb{V}ar(T_i)$. Again, the distribution of Y_i is always over-dispersed. Finally, for any $i \neq j$, the covariance of Y_i and Y_j exists and equals $\mathbb{C}ov(Y_i, Y_j) = \lambda_i \lambda_j \mathbb{C}ov(T_i, T_j)$ whenever the covariance of T_i and T_j exists. These facts are summarized in the result below.

Lemma 9 *In the above setting, the mean vector of* $\mathbf{Y}$ *exists whenever the mean of* $\mathbf{T}$ *is finite, in which case we have* $\mathbb{E}(\mathbf{Y}) = \mathbf{I}(\boldsymbol{\lambda})\mathbb{E}(\mathbf{T})$, *where* $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_d)^\top$ *and* $\mathbf{I}(\boldsymbol{\lambda})$ *is a* $d \times d$

diagonal matrix with the $\{\lambda_i\}$ *on the main diagonal. Moreover, if* $\mathbf{T}$ *has a finite covariance matrix* $\mathbf{\Sigma_T}$ *then the covariance matrix of* $\mathbf{Y}$ *is well defined as well and is given by*

$$\mathbf{\Sigma_Y} = \mathbf{I}(\boldsymbol{\lambda})\mathbf{I}(\mathbb{E}(\mathbf{T})) + \mathbf{I}(\boldsymbol{\lambda})\mathbf{\Sigma_T}\mathbf{I}(\boldsymbol{\lambda})^\top,$$

where $\mathbf{I}(\mathbb{E}(\mathbf{T}))$ *is a* $d \times d$ *diagonal matrix with the diagonal entries* $\{\mathbb{E}(T_i)\}$.

Abbreviations

BRCA: Breast invasive carcinoma; CDF: Cumulative distribution function; FGM: Farlie-Gumbel-Morgenstern; L-P model: lognormal-Poisson model; mRNA: messenger ribonucleic acid; NB: Negative binomial; NORTA: NORmal To Anything; PDF: Probability density functions; PGF: Probability generating function; PMF: Probability mass function; RNA-seq: RNA-sequencing

Acknowledgements

The authors thank the two reviewers for their comments that help improve the paper. We also thank Professors Walter W. Piegorsch and Edward J. Bedrick (University of Arizona) for their helpful discussions.

Authors' contributions

AGS and TJK conceived the study. TJK, AKP, and AGS developed the approach. ADK and AGS conducted the computational analyses. TJK, AKP, and AGS wrote the manuscript. TJK, AKP, AGS, ADK revised the manuscript. All authors read and approved with the final document.

Funding

Research reported in this publication was supported by MW-CTR-IN of the National Institutes of Health under award number 1U54GM104944.

Availability of data and materials

The results shown here are in whole or part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga.

Code reproducing the BRCA data set and computational analyses is available from the corresponding author on reasonable request.

Declarations

Competing interests

The authors declare that they have no competing interests.

Received: 28 October 2020 Accepted: 2 March 2021
Published online: 16 March 2021

References

Barbiero, A., Ferrari, P. A.: An R package for the simulation of correlated discrete variables. Comm. Statist. Simul. Comput. **46**(7), 5123–5140 (2017)

Chen, H.: Initialization for NORTA: Generation of random vectors with specified marginals and correlations. INFORMS J. Comput. **13**(4), 257–360 (2001)

Clemen, R. T., Reilly, T.: Correlations and copulas for decision and risk analysis. Manag. Sci. **45**, 208–224 (1999)

Demitras, H., Hedeker, D.: A practical way for computing approximate lower and upper correlation bounds. Amer. Statist. **65**(2), 104–109 (2011)

Johnson, N., Kotz, S., Balakrishnan, N.: Discrete Multivariate Distributions. Wiley, New York (1997)

Karlis, D., Xekalaki, E.: Mixed Poisson distributions. Intern. Statist. Rev. **73**(1), 35–58 (2005)

Kozubowski, T. J., Podgórski, P.: Distribution properties of the negative binomial Lévy process. Probab. Math. Statist. **29**, 43–71 (2009)

Madsen, L., Birkes, D.: Simulating dependent discrete data. J. Stat. Comput. Simul. **83**(4), 677–691 (2013)

Madsen, L., Dalthorp, D.: Simulating correlated count data. Environ. Ecol. Stat. **14**(2), 129–148 (2007)

Nelsen, R. B.: An Introduction to Copulas (2006)

Nikoloulopoulos, A. K.: Copula-based models for multivariate discrete response data. In: Copulae in Mathematical and Quantitative Finance, 231–249, Lect. Notes Stat., 213. Springer, Heidelberg, (2013)

Nikoloulopoulos, A. K., Karlis, D.: Modeling multivariate count data using copulas. Comm. Statist. Sim. Comput. **39**(1), 172–187 (2009)

Schissler, A. G., Piegorsch, W. W., Lussier, Y. A.: Testing for differentially expressed genetic pathways with single-subject N-of-1 data in the presence of inter-gene correlation. Stat. Methods Med. Res. **27**(12), 3797–3813 (2018)

Solomon, D. L.: The spatial distribution of cabbage butterfly eggs. In: Roberts, H., Thompson, M. (eds.) Life Science Models Vol. 4, pp. 350–366. Springer-Verlag, New York, (1983)

Song, W. T., Hsiao, L.-C.: Generation of autocorrelated random variables with a specified marginal distribution. In: Proceedings of 1993 Winter Simulation Conference - (WSC '93), pp. 374–377, Los Angeles, (1993). https://doi.org/10.1109/WSC.1993.718074

Xiao, Q.: Generating correlated random vector involving discrete variables. Comm. Statist. Theory Methods. **46**(4), 1594–1605 (2017)

Xiao, Q., Zhou, S.: Matching a correlation coefficient by a Gaussian copula. Comm. Statist. Theory Methods. **48**(7), 1728–1747 (2019)

Publisher's Note