Revista Colombiana de Estadística
Número especial en Bioestadística
Junio 2012, volumen 35, no. 2, pp. 255 a 270

A Statistical Model for Analyzing Interdependent Complex of Plant Pathogens

Un modelo estadístico para analizar complejos interdependientes de patógenos vegetales

Eduardo Dávila[a], Luis Alberto López[b], Luis Guillermo Díaz[c]

Departamento de Estadística, Universidad Nacional de Colombia, Bogotá, Colombia

Abstract

We introduce a new approach for modeling multivariate overdispersed binomial data, from a plant pathogen complex. After recalling some theoretical foundations of generalized linear models (GLMs) and Copula functions, we show how the later can be used to model correlated observations and overdispersed data. We illustrate this approach using fungal incidence in vegetables, which we analyzed using Gaussian copula with Beta-binomial margins. Compared to classical and generalized linear models, the model using Gaussian copula function best controls for overdispersion, being less prone to the underestimation of standard errors, the major cause of wrong inference in the statistical analysis of plant pathogen complex.

***Key words*:** Epidemiological methods, Extra-binomial variation, Multivariate data.

Resumen

Se introduce un nuevo enfoque para modelar datos binomiales multivariados con sobredispersión, obtenidos de complejos de patógenos vegetales. Después de revisar los conceptos básicos de los modelos lineales generalizados (GLMs) y las funciones Cópula, se muestra cómo estas últimas pueden usarse para modelar observaciones correlacionadas y datos con sobredispersión. Se ilustra el método usando la incidencia de hongos en hortalizas, analizando el caso por medio de la función cópula Gaussiana con marginales Beta-binomiales. Comparado con los modelos lineales clásicos y generalizados, el modelo construido con la cópula Gaussiana es el que mejor controla la sobredispersión, siendo menos propenso a la subestimación de los errores

[a]Ph.D. student. E-mail: jedavilas@unal.edu.co
[b]Professor. E-mail: lalopezp@unal.edu.co
[c]Professor. E-mail: lgdiazm@unal.edu.co

estándar, la causa más importante de inferencia inapropiada en el análisis estadístico de complejos de patógenos vegetales.

***Palabras clave*:** métodos epidemiológicos, variación extra-binomial, datos multivariados.

1. Introduction

The use of single-parameter family of distributions can sometimes be problematic for statistical inference (Cox 1983). For example, in the binomial distribution the variance is totally determined by the mean, and when this is satisfied there is nominal dispersion, an assumption that cannot be hold in some data analyses. In fact, vector data may display a lack of independence as is commonly the case in experimental trials in plant pathology; in these data, the presence of a fungus often increases the probability of damage in neighboring leaves, leading to marginal dependence in the data. Moreover, the analysis of plant-pathogen complex can also be complicated by the presence of multivariate dependence, as was shown by Dávila (2005).

To get a correct analysis of multivariate binomial data, an overdispersion diagnostic is necessary in order to compare the nominal dispersion against the actual dispersion. To this end, Smith & Heitjan (1993) provided an appropriate statistical tool to detect extra binomial variation. McCullagh & Nelder (1989) maintain that "overdispersion is a common attribute of data arising in many fields, and statistical practitioners shall assume that overdispersion is present in some extent". Accordingly, there are two main approaches to deal with univariate overdispersion: First, the use of full parametric models like dispersion models (Joe 1997), and second, the choice of families of estimating functions (Heyde 1997). In the case of multivariate data, multivariate dispersion models (Jørgensen & Lauritzen 2000) and copula function based models (Song, Li & Yuan 2009) can be used.

The literature on copula model with count data is not abundant, with some references in financial and actuarial sciences. Nikoloulopoulos & Karlis (2010) present a recent review for the use of this methodology with application to discrete data in marketing exchanges. Some applied works have been done in joint modeling of correlated data using Gaussian copulae (Song et al. 2009). Furthermore, a recent approximation to the Gaussian copula likelihood is given in Madsen & Fang (2011), who found that for finite samples the estimator of generalized estimating equations is more efficient than the maximum likelihood estimator (MLE). However, Song, Li & Yuan (2011) maintain that MLE is more efficient.

With respect to applications in the biological sciences, the next are some useful references. Lambert & Vandenhende (2002) propound a model for non-normal longitudinal data with illustration in a dose titration safety study in human medicine. A work in multivariate logistic regression was presented by Li & Wong (2011) and, because of a lack of constraints in the parameters and the admission of a limited range of dependence in the copula, this paper was criticized and corrected (Nikoloulopoulos 2012). A more basic study was carried out by Trégouët,

Ducimetière, Bocquet, Visvikis, Soubrier & Tiret (1999), with binary data on nuclear families, in this analysis the response was the presence or the absence of a disease in each member of the family.

In the particular situation of plant-disease complex, the presence of two or more pathogenic fungi can be strongly correlated, thereby violating the assumption of independence amongst observations (Dávila 2005, Dávila & López 2010). In such a situation, it is necessary to use a statistical model with multivariate distributions which include both marginal overdispersion and multivariate dependence (Fischer 2011, Joe 1997, Song 2007).

Ultimately, in relation to the disadvantages of copula-based analysis of count data, two important references shall be mentioned: Genest & Nešlehovà (2007) for details on the danger and limitations of the use of copulae to model discrete data, and Embrechts (2009) who in a personal view gives some review on this theory, recommends some important lectures and analyzes future developments. Additionally, the reader is encouraged to review the controversial article of Mikosch (2006), which is a critical point of view of copula methodology, with discussion and rejoinder. Despite some problems in copula modeling with discrete data, nowadays this model constructions are valid but subject to cautions.

The present paper contains four sections. Section 2 presents the characterization of multivariate vectors, reviews some concepts on overdispersion diagnostics and model selection. Section 3 is dedicated to theoretical details of the proposed model. Section 4 shows an application to empirical data in diseases management on vegetables. Finally, Section 5 presents discussions and conclusion.

2. Material and Methods

In this section we present the characterization of data and parameter vectors, overdispersion diagnostic and a short reminder on copula theory and model selection.

2.1. Structure of data and parameter vectors

In plant pathology studies, data are typically made of binomial observations representing the presence/absence of pathogenic fungi. Data obtained for d fungi are modeled by a d-variate vector:

$$Y = (Y_1, Y_2, \ldots, Y_d)^T \tag{1}$$

where Y_i is a binomial random variable associated to the incidence of the ith fungus, $i = 1, 2, \ldots, d$. A common assumption is that the probabilistic mechanism that generates marginal data is the binomial law, whose density with respect to the counting measure is given by

$$f_{Y_i}(y_i \mid \pi_i, m_i) = \binom{m_i}{y_i} \pi_i^{y_i} (1 - \pi_i)^{m_i - y_i} \tag{2}$$

where $y_i = 0, 1, \ldots, m_i$ and with given probability of success π_i; we write formally that $Y_i \sim bin(m_i, \pi_i)$, with

$$E[Y_i] = m_i \pi_i$$

and

$$Var[Y_i] = m_i \pi_i (1 - \pi_i),\ i = 1, 2, \ldots, d \tag{3}$$

Provided that multivariate data are generated by the same designed experiment, there is an identical design matrix X associated to any margin Y_i; hence, under the GLM framework, the three components are (see McCullagh & Nelder 1989):

1. The class of densities in (2) with π_i varying in the interval $(0, 1)$, which belongs to the exponential family of distributions,
2. The systematic part $X\theta_i$, where X is a $n \times p$ matrix and $\theta_i = (\theta_{i1}, \theta_{i2}, \ldots, \theta_{ip})^T$ is a vector of unknown parameters with $\theta_i = ln\big(\frac{\pi_i}{1-\pi_i}\big)$, and
3. The link function $g_i(\cdot)$.

In GLM modeling, it is supposed that there is independence between any subset of random variables from (1) and that (3) holds.

Because this work is dealing with the lack of independence and overdispersion ($Var[Y_i] \gg m_i \pi_i (1 - \pi_i)$), a natural characteristic of multivariate data arising in plant-disease complex, then a new model shall be considered. Hence a full likelihood inference procedure requires a family of distributions with a great vector of total marginal parameters

$$\Theta = (\theta_1^T, \theta_2^T, \ldots, \theta_d^T)^T$$

and an association matrix

$$\Gamma = \begin{pmatrix} \gamma_{11} & \gamma_{12} & \cdot & \cdot & \cdot & \gamma_{1d} \\ \gamma_{21} & \gamma_{22} & \cdot & \cdot & \cdot & \gamma_{2d} \\ \cdot & \cdot & & \cdot & & \cdot \\ \cdot & \cdot & & & \cdot & \cdot \\ \cdot & \cdot & & & & \cdot \\ \gamma_{d1} & \gamma_{d2} & \cdot & \cdot & \cdot & \gamma_{dd} \end{pmatrix}$$

where γ_{ii^*}, $i \neq i^*$, $i^*, i = 1, 2, \ldots, d$, will be taking in account the bivariate association between each pair of transformed margins; the construction of the desired multivariate distribution is the objective of the Section 3. However, an important prerequisite lies in the detection of extra binomial variation, which we now detail.

2.2. Overdispersion Diagnostic

To test the nominal dispersion in the ith margin, it is important to give an extension of (3), i.e.,

$$Var[Y_i] = \lambda_i m_i \pi_i (1 - \pi_i),$$

and the hypothesis testing problem is formulated for all $i = 1, 2, \ldots, d$ as

$$H_{0_i} : \lambda_i = 1 \; versus \; H_{1_i} : \lambda_i > 1 \tag{4}$$

An appropriate procedure to test (4) is the score statistic of Smith and Heitjan (1993), viz.

$$\chi_i^2 = J_i^T A_i^{-1} J_i, \; i = 1, 2, \ldots, d \tag{5}$$

where $J_i = (J_{i1.}, J_{i2.}, \ldots, J_{ip.})$ is a random vector that registers the difference between actual information and nominal information, in the ith margin with respect to every jth parameter, namely

$$J_{ij.} = \frac{1}{2} \sum_{k=1}^{n} \left[\left(\frac{\partial l_{ijk}}{\partial \theta_{ij}} \right)^2 - \left(\frac{\partial^2 l_{ijk}}{\partial \theta_{ij}^2} \right) \right] \; j = 1, \ldots, p, \; i = 1, 2, \ldots, d \tag{6}$$

and A_i is the covariance matrix of J_i corrected for estimation of θ_i, whose explicit expressions are given in the appendix of Smith & Heitjan's (1993) paper.

In equation (6), l_{ijk} is the log-likelihood of the binomial distribution presented in (2). Hence, for each ith margin with respect to the jth parameter and the kth observation, we have

$$l_{ijk} = y_{ijk} \ln \left(\frac{\pi_{ijk}}{1 - \pi_{ijk}} \right) + m_{ijk} \ln(1 - \pi_{ijk})$$

and

$$J_{ij.} = \frac{1}{2} \sum_{k=1}^{n} \left[(y_{ijk} - m_i \pi_i)^2 - m_i \pi_i (1 - \pi_i) \right] x_{ijk}^2, \; j = 1, \ldots, p, \; i = 1, 2, \ldots, d$$

Under the null hypothesis of nominal dispersion (4), the asymptotic distribution of (5) is the central χ^2-distribution with p degrees of freedom. The eventual reject of (4) will be a clear evidence that $Var[Y_i] \gg m_i \pi_i (1 - \pi_i)$; namely, actual variance is statistically greater than the nominal one.

Hitherto, we have been dealing with marginal overdispersion, whereas the statistical problem in plant-pathogen complex data includes both marginal overdispersion and multivariate dependence. In the following, we show how the latter can be addressed using copulae theory.

2.3. Basics on Copula Modeling

An interesting concept for connecting multivariate cumulative distribution functions and their margins is offered by copulae theory (see Joe 1997, Nelsen 2006). A mapping $C : [0, 1]^d \to [0, 1]$ is called a d-dimensional copula, if it is the distribution of a uniform vector $U = (U_1, U_2, \ldots, U_d)$; that is, copulae are joint distribution functions of standard uniform random variates (Cherubini, Luciano

& Vecchiato 2004). Because any marginal distribution function F_i has a uniform distribution, i.e. $F_i(y) \sim U(0,1)$ with $i = 1, 2, \ldots, d$, the use of copulae has become evident in the last few years, to construct dependency models (Härdle & Simar 2007).

The application of copulae to statistical modeling is based on Sklar's theorem (Nelsen 2006); this useful theorem states that given marginal distributions, it is possible to couple these margins into a joint distribution whose arguments are the F_i's; provided that the margins are continuous, this kind of representation is unique. Hence, following Grønneberg (2011), there are four basic problems in parametric modeling through copulae theory, namely:

- How to estimate the dependence parameter?
- How should the parametric form of the copula family be chosen?
- How to select among several candidate models on the basis of actual data?
- Is the final model adequate?

The scientific context of plant pathology gives us preliminary responses for the first two items, whereas the two later are pure statistical modeling steps and will be reviewed in the following.

2.4. Model Selection and Goodness of Fit

A usual tool for model selection is the Akaike Information Criterion (AIC), which is not appropriate when dealing with semi-parametric estimation, a common method used in the construction of copulae. A proper generalization of AIC, given in Grønneberg (2011), is the Copula Information Criterion (CIC), viz.

$$CIC = 2l_{N,\max} - 2(\widehat{p}^* + \widehat{q}^* + \widehat{r}^*) \tag{7}$$

where $l_{N,\max}$ is the maximum multivariate pseudo-likelihood. The second term of (7) has a more elaborate formula than in AIC –where it depends only on the length of parameter vector. If the model is correctly specified, then $\widehat{q}^* = 0$. Details for deriving the estimates of $\widehat{p}^*$, $\widehat{q}^*$ and $\widehat{r}^*$ from empirical information, and least false copula derivatives are given in Grønneberg (2011).

Genest, Rémillard & Beaudoin (2009) provide a useful tool to test the final model adequacy. Let H be a joint cumulative distribution function the copula representation of H is

$$H(y_1, y_2, \ldots, y_d) = C(F_1(y_1), F_2(y_2), \ldots, F_d(y_d)) \tag{8}$$

provided that C is unknown to model $Y = (Y_1, Y_2, \ldots, Y_d)^T$, we suppose that C belongs to a class

$$\mathcal{C} = \{C_\omega : \omega \in \Omega\}, \Omega \subseteq \mathbb{R}^d, d \geq 1 \tag{9}$$

so we must test,

$$H_0 : C \in \mathcal{C} \quad \text{versus} \quad H_1 : C \notin \mathcal{C} \tag{10}$$

Genest et al. (2009) advocate the use of "blanket test", based on the empirical copula, viz.,

$$C_N(\mathbf{u}) = \frac{1}{N} \sum_{l=1}^{N} \mathbf{I}(\widehat{\mathbf{U}}_l \leq \mathbf{u}),\ \mathbf{u} \in [0,1]^d \tag{11}$$

where $\widehat{\mathbf{U}}_l$ is a vector of pseudo-observations, whose components are the empirical cumulative distribution functions related to each margin, obtained from actual data, i.e.,

$$\widehat{\mathbf{U}}_l = (\widehat{F}_{l,1}, \ldots, \widehat{F}_{l,d}),\ l = 1, 2, \ldots, N$$

with N being the size of a random sample from (1); it is important to recall that, under probability transformations, it is expected that $\widehat{F}_{l,i} \sim U(0,1)$ for all $l = 1, 2, \ldots, N$ and $i = 1, 2, \ldots, d$. The empirical copula (11) is a consistent estimator of C in (8), and the statistic to test H_0 in (10) is

$$S_N = \sum_{l=1}^{N} \{C_N(\widehat{\mathbf{U}}_l) - C_{\omega_N}(\widehat{\mathbf{U}}_l)\}^2 \tag{12}$$

The asymptotic distribution of (12) cannot be directly tabulated, then approximations of p-values shall be obtained via bootstrap-based procedures. Because of its high computational cost, Kojadinovic, Yan & Holmes (2011) recently proposed a fast large-sample testing procedure based on multiplier central limit theorems.

Now that we have recalled the basics of model selection and goodness of fit tests, we can introduce our alternative model for the statistical analysis of plant-pathogen complex.

3. A Model for Multivariate Overdispersed Binomial Data

Here the objective is to present an alternative statistical model to analyze plant-pathogen complex data. More specifically, we shall focus on the analysis of designed experiments to evaluate substances as possible activators of Systemic Acquired Resistance (SAR) (Durrant & Dong 2004). Because SAR is a mechanism which confers a broad spectrum of protection against plant pathogens, it is expected that all fungi in a complex should be affected and that multivariate data should not present independence; additionally, the natural spreading of pathogen inoculum cannot guarantee marginal independency, then marginal overdispersion can be a natural attribute of such data.

We are going to construct the desired model in two steps, first, fitting margins to an appropriate family of distribution, and second, modeling the given margins in a Gaussian copula family framework.

3.1. Marginal Overdispersion Model

In order to model marginal overdispersion, we make use of Beta-binomial hierarchy, a generalization of binomial distribution (Casella & Berger 2002). In this model, it is supposed that $Y_i \mid P_i \sim bin(m_i, P_i)$, whereas $P_i \sim Beta(\alpha_i, \beta_i)$. Then, from now on, we make the assumption that each margin (Y_i) follows a Beta-binomial law. Therefore, unconditionally the compound density, with respect to the counting measure of Y_i, is given by

$$f_{Y_i}(y_i \mid \alpha_i, \beta_i) = \binom{m_i}{y_i} \frac{B(y_i + \alpha_i, m_i - y_i + \beta_i)}{B(\alpha_i, \beta_i)},\ y_i \in \{0, 1, \ldots, m_i\} \tag{13}$$

furthermore, in (13) $B(.,.)$ is the beta function, $\alpha_i > 0$ and $\beta_i > 0$. Conditional to P_i the expectation is given by

$$E(Y_i \mid P_i) = \mu_i = m_i \pi_i = m_i \frac{\alpha_i}{\alpha_i + \beta_i},\ i = 1, 2, \ldots, d$$

the conditional variance is

$$\begin{aligned} Var(Y_i \mid P_i) &= m_i \pi_i (1 - \pi_i) \frac{\alpha_i + \beta_i + m_i}{\alpha_i + \beta_i + 1} \\ &= m_i \pi_i (1 - \pi_i)\{1 + \phi_i (m_i - 1)\}, \quad i = 1, 2, \ldots, d \end{aligned} \tag{14}$$

from (14) we can see that the marginal dispersion parameter is

$$\phi_i = \frac{1}{\alpha_i + \beta_i + 1}$$

Comparing (3) with (14) it is noted that the later has a greater variance, whose increment is given by a function of ϕ_i and the marginal binomial index m_i. The R package VGAM and its function `vglm` is actually an alternative to fit marginal responses with Beta-binomial distribution.

3.2. Multivariate Model

Given the marginal distributions $F_1(Y_1), F_2(Y_2), \ldots, F_d(Y_d)$ from Beta-binomial hierarchies (13) and using the Sklar's theorem, a new family of d-variate distributions can be obtained and represented by

$$C_\Phi(U_1, U_2, \ldots, U_d) = H\left(F_1(Y_1), F_2(Y_2), \ldots, F_d(Y_d) \mid \Gamma\right) \tag{15}$$

where H is the d-variate Gaussian distribution with correlation matrix Γ and, in presence of continuous margin, the density is given by

$$f_Y(y; \mu, \phi, \Gamma) = c_\Phi\{F_1(y_1), F_2(y_2), \ldots, F_d(y_d) \mid \Gamma\} \prod_{i=1}^{d} f(y_i; \pi_i, \phi_i)$$

where $\pi^T = (\pi_1, \pi_2, \ldots, \pi_d) \in [0,1]^d$ is the main vector of marginal parameters and $\phi^T = (\phi_1, \phi_2, \ldots, \phi_d) \in \mathbb{R}^d$ is the ancillary vector of marginal dispersion parameters. Because (13) is a discrete distribution, then we use the more appropriate expression

$$P(Y_1 = y_1, Y_2 = y_2, \ldots, Y_d = y_d) = \sum_{j_1=1}^{2}\sum_{j_2=1}^{2}\cdots\sum_{j_d=1}^{2}(-1)^{j_1+j_2+\ldots+j_d}C_\Phi(u_{1j_1}, u_{2j_2}, \ldots, u_{dj_d} \mid \Gamma) \quad (16)$$

with $u_{i1} = F_i(y_i)$ and $u_{i2} = F_i(y_i - 1)$ $\; i = 1, 2, \ldots, d$, which is the density with respect to the counting measure, namely the Radon-Nikodym derivative of (15).

3.3. Two Step Inference

To make inference on (16) we use the two parts inference procedure, proposed by Joe (1997). In this methodology, in the first step the margins are fitted from (13) and because it is composed of common functions, both numerical methods or maximum likelihood estimation (MLE) are applicable; see Griffiths (1973) for details. The R package VGAM makes use of Fisher scoring for estimation and it operates quite well for overdispersed binomial data. In a particular situation, to model $g(\pi_i) = X\theta_i$, the score equation, for maximum likelihood estimation from (13), is

$$\frac{\partial l_i}{\partial \theta_i} = (\alpha_i+\beta_i)\sum_{k=1}^{n}\left\{ddg(y_i, (\alpha_i+\beta_i)\pi_i) - ddg(m_i - y_i, (\alpha_i+\beta_i)(1-\pi_i))\right\}\frac{1}{g'(\pi_i)}x_{jk}$$

$j = 1, \ldots, p$, where $ddg(a, b) = \log\Gamma(a+b) - \log\Gamma(b)$; additional details can be seen in Hinde & Demetrio (1998).

The second step deals with the selection of an appropriate family of copulae. In the case of Gaussian copula, for the estimation of Γ, can be used some assumptions like the presence of exchangeable Pearson correlation matrix, i.e., $\gamma_{ii^*} = \gamma$, $i \neq i^*$; in any case, from (16) the solution of

$$\frac{\partial C_\Phi(u_{1j_1}, u_{2j_2}, \ldots, u_{dj_d} \mid \Gamma)}{\partial \Gamma} = 0$$

can be obtained using the Gaussian-Hermite quadrature method (see McCulloch, Searly & Neuhaus 2008, pp. 326-331). Finally, consider the vector of marginal and multivariate parameters

$$\eta = (\theta_1, \theta_2, \ldots, \theta_d, \gamma_{12}, \gamma_{13}, \ldots, \gamma_{(d-1)(d)})$$

in order to complete the inference procedure; following Joe (1997), it is necessary to estimate the inverse Godambe information matrix

$$V = D_h^{-1} M_h (D_h^{-1})^T \quad (17)$$

where $D_h = E[\partial h^T(Y,\eta)/\partial\eta]$ and $M_h = E[h^T(Y,\eta)h(Y,\eta)]$, with h being the first derivative of the logarithm of (16) with respect to η. The estimation of $N^{-1}V$, which is the asymptotic covariance matrix of the MLE of η, namely $\widehat{\eta}$, can be done via Jackknife, viz.,

$$\Im = \sum_{l=1}^{N} (\widehat{\eta}^{(l)} - \widehat{\eta})^T (\widehat{\eta}^{(l)} - \widehat{\eta}) \tag{18}$$

In (18), $\widehat{\eta}^{(l)}$ is the estimator of η once the lth observation has been eliminated.

4. Application

Cely (1996) carried out a trial in Colombia in an onion crop, in order to analyze the effect of seven treatments, based on the aspersion of inactive inoculum of the plant pathogen *Peronospora destructor* for cross protection, an approach later included in the SAR methods by Durrant & Dong (2004). The experiment was located under a complete randomized block design, with two blocks (the crop varieties Junca and Monguana). Three responses were captured as binomial data, all of them associated to the incidence of a pathogenic fungus; namely Y_1 represents the downy mildew *Peronospora* sp., Y_2 the leaf blight *Stemphylium* sp. and Y_3 the leaf spot *Cladosporium* sp.; so the dependent response vector to be modeled is

$$Y = (Y_1, Y_2, Y_3)$$

Initially, nominal dispersion was rejected with p-values less than 0.05, for all three margins with respect to the hypothesis testing problem in (4); furthermore, marginal Beta-binomial hierarchy models (13) were fitted; then, given the three CDF's $F_1(y_1), F_2(y_2), F_3(y_3)$ a 3-variate Gaussian copula model was fitted, according to (16). To select the model on the basis of observed data, we use Copula Information Criterion (7), and the goodness of fit was based on Genest et al. (2009); finally, applying Jackknife method (18), the Godambe's asymptotic covariance matrix was estimated. About marginal dispersion parameters, the nominal dispersion (4) was not rejected, under 3-variate framework, for Y_2, i.e., $\phi_2 \simeq 0$ for the random variable associated to *Stemphylium* fungus, an endemic plant pathogen; see Table 1.

Table 1: Estimations, standard errors, and confidence intervals for dispersion parameters(*).

Parameter estimations	Standard error	Lower Limit	Upper Limit
$\widehat{\phi}_1 = 0.01983$	0.0037	0.0125	0.0270
$\widehat{\phi}_2 = 0.00377$	0.1979	−0.3840	0.3915
$\widehat{\phi}_3 = 0.01735$	0.0035	0.0104	0.0242

(*) $\alpha = 0.05$.

The standard errors of the parameter estimators appear on Table 2 for normal linear models (MVN) with Box and Cox transformations –the original model used

by Cely (1996)–, generalized linear model (GLM), marginal overdispersion model (ODM), and multivariate overdispersion model with Gaussian copula and Beta-binomial margins (CGB). As it can be seen, ODM and CGB are the models with less significant effects. In fact, both ODM and CGB show a total of six standard errors associated with significant estimations; nevertheless, without differences in relation to the number of significant effects, CGB offers higher values of standard errors.

With respect to the estimation of the association parameters, the correlation matrix, i.e.,

$$\widehat{\Gamma} = \begin{pmatrix} 1.000 & 0.484** & 0.475** \\ 0.484** & 1.000 & 0.688** \\ 0.475** & 0.688** & 1.000 \end{pmatrix}$$

shows a positive dependence between normal scores; all three estimations were highly significant (p-value < 0.0001), leading to the consideration that the appropriate copula, for the analyzed data, is not the independent one.

Table 2: Standard errors for parameter estimators.

Factor(variable)	MVN(1)	GLM(1)	ODM(1)	CGB(1)
T0(y_1)	0.0266*	0.103*	0.163*	0.157*
T1(y_1)	0.0266*	0.106*	0.166*	0.141*
T2(y_1)	0.0266	0.112	0.179	0.221
T3(y_1)	0.0266	0.110*	0.174	0.176
T4(y_1)	0.0266*	0.107*	0.168*	0.159*
T5(y_1)	0.0266	0.112	0.177	0.166
T6(y_1)	0.0266	0.113	0.179	0.197
JUNCA(y_1)	0.0133*	0.052*	0.082*	0.086*
T0(y_2)	0.0186	0.093	0.106	0.113
T1(y_2)	0.0186*	0.091*	0.103*	0.136*
T2(y_2)	0.0186*	0.094	0.107	0.135
T3(y_2)	0.0186*	0.091*	0.104*	0.122*
T4(y_2)	0.0186	0.096	0.109	0.121
T5(y_2)	0.0186	0.095	0.108	0.142
T6(y_2)	0.0186	0.098	0.111	0.107
JUNCA(y_2)	0.0092	0.046	0.053	0.061
T0(y_3)	0.0265	0.108	0.164	0.169
T1(y_3)	0.0265	0.104	0.160	0.196
T2(y_3)	0.0265	0.109	0.167	0.202
T3(y_3)	0.0265	0.105	0.161	0.210
T4(y_3)	0.0265	0.107	0.163	0.176
T5(y_3)	0.0265	0.105	0.162	0.208
T6((y_3)	0.0265	0.113*	0.170	0.161
JUNCA(y_3)	0.0130	0.053	0.081	0.089

(1)*= significative effect (α=0.05).

In relation to the early work of Cely (1996), the author made use of the assumption of independence between the three count variables; hence, let's see that a wrong assumption can lead to an incorrect inference. In the original report of Cely, the SAR-treatment (T2), with respect to the random variable Y_2, was considered a significant one, i.e., it was statistically different from chemical and mixed

treatments and its use was not taken account: why shall a small difference lead to significant effect? The answer is an underestimation of standard error, given by lack of independence and marginal overdispersion, that were not considered in the assumed probability model.

In this new data analysis, based on dependence concepts, the treatment T2 does not have differences with respect to chemical and mixed ones, according to response Y_2; therefore, the new position will be that T2 is a good solution to implement an integrated pathogen handing in that crop, because it controls the three pathogens together with statistical significance. In Table 3 we present two inferential situations; first, the analysis under MVN, whose significant effects are represented by "$*$"; second, the analysis via CGB, whose significant effects are indicated by "$\diamond$". Because in CGB the treatment T2 is statistically similar to the chemical ones, this new analysis is in favour of T2, the natural SAR-fungicide.

Table 3: Two inferential situations

Treatments	$Y_1(\%)$	$Y_2(\%)$	$Y_3(\%)$
T0= control	18.38$*\diamond$	18.37$*\diamond$	11.45
T1= SAR low dosage	16.78$*\diamond$	20.38$*\diamond$	15.10
T2= SAR medium dosage	10.26	16.10$*$	10.10
T3= SAR high dosage	12.62	20.17$*\diamond$	12.60
T4= Mancozeb	14.88$*\diamond$	15.50	11.85
T5= Mancozeb + Cimoxanyl	11.66	15.40	11.80
T6= T2+T4	10.35	14.97	10.3
T7= T2+T5	9.93	15.90	11.15

$*$ = significant effect in relation to MVN modeling, $\alpha = 0.05$.
$\diamond$ = significant effect in relation to CGB modeling, $\alpha = 0.05$.

5. Discussion and Conclusion

Gaussian copulae theory is suitable to construct models with given non-normal margins, which is the particular situation in plant diseases control. A very important issue in model selection is the context, i.e., all modeling shall have scientific foundations and clear proposals (Claeskens & Hjort 2008). Because the application of some therapies associates to natural resistance activation (SAR methodologies) on plants against fungi has a broad spectrum, the lack of independence between the incidence of pathogens is evident, and then the use of independent marginal models is out of scientific context.

Also, it is important to stress the difference of the present methodology with respect to the works of Song et al. (2009) and Song (2000), which is the use of margins not belonging to the class of dispersion models (Jørgensen 1997) in our proposal. Here we are using a Beta-binomial hierarchy to deal with marginal overdispersion, a new application to copulae theory in the broad field of plant pathology, a methodology appropriate to modeling SAR-based experiments, the ones that require modern statistical tools.

It is worth to recall some limitations of the proposed model, according to the work of Genest & Nešlehovà (2007). The first limitation is the lack of uniqueness

of the copula, once the random variables put their mass on few atoms: it is a crucial aspect in binary data and less important if the binomial index tends to infinity in binomial variables. Accordingly, practitioners may be cautious in the use of the present methodology with sparse data, that is, when the binomial index is small ($m_i < 6$), our model is not appropriate.

A second aspect of copula-based regression for discrete data is that dependence is not only a function of the copula; additionally, Kendall's tau an Spearman's rho may not span the entire interval $[-1, 1]$. About this weakness, the use of Gaussian copula guarantees that the association parameter, i.e., the Pearson correlation coefficient, can reach the Fréchet-Hoeffding bounds (Song 2007). Nevertheless, these dependence parameters are governing the association but they do not have direct interpretation. That is, the correlation between normal scores is not the same that the one between the actual variables; hence, we may interpret Γ as a dependence parameter matrix, all but as a correlation matrix of the original binomial variables. Furthermore, because the margins also characterize the dependence in the copula, when dealing with discrete data we may consider a conditional copula model, where the association parameters are varying with the covariates (see Acar, Craiu & Yao 2011).

Even if the conditions under which the dependence parameters are estimable are not elucidated, hitherto the maximum likelihood estimation is a valid methodology for inference. Hence, no further discussion on this topic are exposed here (Genest & Nešlehovà 2007).

In conclusion, we have that in our example, the model based on Gaussian copula (CGB) displayed the highest standard errors associated to parameter estimators, suggesting that this approach controlled the overdispersion in the data. Additionally, it considers both marginal overdispersion and multivariate dependence, whereas the marginal overdispersion model, based on independent Beta-binomial hierarchy (ODM), assigns multivariate dependence to a marginal overdispersion. Provided that multivariate dependence is present, application shows that normal linear models (MVN) does not differ from modeling via GLM without overdispersion fit, leading to a wrong multivariate inference. The model constructed via Gaussian copula with Beta-binomial margins (CGB) is probably preferable for analyzing overdispersed and non-sparse multivariate binomial data, whereas the classical multivariate normal linear model is not appropriate in such situations.

Acknowledgments

We would like to thank the editors and the two anonymous referees for their constructive comments to improve the quality and presentation of this paper.

[Recibido: septiembre de 2011 — Aceptado: febrero de 2012]

References

Acar, E., Craiu, R. & Yao, F. (2011), 'Dependence calibration in conditional copulas: A nonparametric approach', *Biometrics* **67**, 445–453.

Casella, G. & Berger, R. (2002), *Statistical Inference*, 2 edn, Duxbury Press, Florida, United States.

Cely, B. (1996), Control de mildeo velloso (*Peronospora destructor*) en el cultivo de cebolla de rama mediante protección cruzada, Tesis de grado, Universidad Pedagógica y Tecnológica de Colombia, Tunja, Colombia.

Cherubini, U., Luciano, E. & Vecchiato, W. (2004), *Copula Methods in Finance*, John Wiley & Sons, England.

Claeskens, G. & Hjort, N. (2008), *Model Selection and Model Averaging*, Cambridge University Press, Cambridge.

Cox, D. R. (1983), 'Some remarks on overdispersion', *Biometrika* **7**(1), 269–274.

Durrant, W. & Dong, X. (2004), 'Systemic acquired resistance', *Annual Review of Phytopathology* **42**, 185–209.

Dávila, E. (2005), Modelación multivariada de la sobredispersión en datos binarios, aplicación en epidemiología vegetal, Tesis de maestría, Universidad Nacional de Colombia, Bogotá, Colombia.

Dávila, E. & López, L. (2010), Modeling multivariate overdispersed binomial data, *in* 'International Biometrics Conference', XXV International Biometric Conference, Florianópolis, Brazil.

Embrechts, P. (2009), 'Copulas: A personal view', *The Journal of Risk and Insurance* **76**(3), 639–650.

Fischer, M. (2011), Multivariate copulas, *in* D. Kurowicka & H. Joe, eds, 'Dependence Modeling Vine Copula Handbook', World Scientific, pp. 19–36,.

Genest, C. & Nešlehovà, J. (2007), 'A primer on copulas for count data', *ASTIN Bulletin* **37**(2), 475–515.

Genest, C., Rémillard, B. & Beaudoin, D. (2009), 'Goodness-of-fit tests for copulas: A review and a power study', *Insurance: Mathematics and Economics* **44**, 199–213.

Griffiths, D. A. (1973), 'Maximum likelihood estimation for the beta-binomial distribution and an application to the household distribution of the total number of cases of a disease', *Biometrics* **29**, 637–648.

Grønneberg, S. (2011), The copula information criterion and its implications for the maximum pseudo-likelihood estimator, *in* Kurovicka & Joe, eds, 'Dependence Modeling Vine Copula Handbook', World Scientific, pp. 113–138.

Härdle, W. & Simar, L. (2007), *Applied Multivariate Statistical Analysis*, Springer-Verlag, Berlin.

Heyde, C. (1997), *Quasi-likelihood And Its Applications: A General Approach To Optimal Methods of Estimation*, Springer, New York.

Hinde, J. & Demetrio, C. (1998), *Overdispersion: Models and estimation*, XIII Sinape, Caxambu, Brazil.

Joe, H. (1997), *Multivariate Models and Dependence Concepts*, Chapman and Hall, London.

Jørgensen, B. (1997), *Dispersion Models*, Chapman and Hall, London.

Jørgensen, B. & Lauritzen, S. (2000), 'Multivariate dispersion models', *Journal of Multivariate Analysis* **74**, 267–281.

Kojadinovic, I., Yan, J. & Holmes, M. (2011), 'Fast large-sample goodness-of-fit for copulas', *Statistica Sinica* **21**, 841–871.

Lambert, P. & Vandenhende, F. (2002), 'A copula-based model for multivariate non-normal longitudinal data: Analysis of a dose titration safety study on a new antidepressant', *Statistics in Medicine* **21**, 3197–3217.

Li, J. & Wong, W. (2011), 'Two-dimensional toxic dose and multivariate logistic regression, with application to decompression sickness', *Biostatistics* **12**, 143–155.

Madsen, L. & Fang, Y. (2011), 'Joint regression analysis for discrete longitudinal data', *Biometrics* **67**(3), 1171–1175.

McCullagh, P. & Nelder, J. (1989), *Generalized Linear Models*, Chapman and Hall/CRC, London.

McCulloch, C., Searly, S. & Neuhaus, J. (2008), *Generalized Linear and Mixed Models*, Wiley, New York.

Mikosch, T. (2006), 'Copulas: Tales and facts (with discussion and rejoinder)', *Extremes* **9**, 3–63.

Nelsen, R. (2006), *An Introduction to Copulas*, 2 edn, Springer, New York.

Nikoloulopoulos, A. (2012), 'Letter to the editor', *Biostatistics* **13**(1), 1–3.

Nikoloulopoulos, A. & Karlis, D. (2010), 'Modeling multivariate count data using copulas', *Statistics in Medicine* **27**, 6393–6406.

Smith, P. & Heitjan, F. (1993), 'Testing and adjusting for departures from nominal dispersion in generalized linear models', *Applied Statistics* **42**(1), 31–34.

Song, P. X. (2000), 'Multivariate dispersion models generated from gaussian copula', *Scandinavian Journal of Statistics* **27**, 305–320.

Song, P. X. (2007), *Correlated Data Analysis: Modeling, Analytics, and Applications*, Springer, New York.

Song, P. X., Li, M. & Yuan, Y. (2009), 'Joint regression analysis of correlated data using gaussian copulas', *Biometrics* **65**, 60–68.

Song, P. X., Li, M. & Yuan, Y. (2011), 'Joint regression analysis for discrete longitudinal data - rejoinder', *Biometrics* **67**(3), 1175–1176.

Trégouët, D., Ducimetière, P., Bocquet, V., Visvikis, S., Soubrier, F. & Tiret, L. (1999), 'A parametric copula model for analysis of familial binary data', *American Journal of Human Genetics* **64**(3), 886–893.