A Statistical Model for Analyzing Interdependent Complex of Plant Pathogens

Un modelo estadístico para analizar complejos interdependientes de patógenos vegetales

EDUARDO DÁVILA\textsuperscript{a}, LUIS ALBERTO LÓPEZ\textsuperscript{b}, LUIS GUILLERMO DÍAZ\textsuperscript{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.

\textsuperscript{a}Ph.D. student. E-mail: jedavilas@unal.edu.co
\textsuperscript{b}Professor. E-mail: lalopezp@unal.edu.co
\textsuperscript{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 & Nešlehová (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 cautious.

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:

$$ Y = (Y_1, Y_2, \ldots, Y_d)^T $$

where $Y_i$ is a binomial random variable associated to the incidence of the $i$th 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

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

$$ \text{(2)} $$
where \( y_i = 0, 1, \ldots, m_i \) and with given probability of success \( \pi_i \); we write formally that \( Y_i \sim \text{bin}(m_i, \pi_i) \), with

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

and

\[
\text{Var}[Y_i] = m_i \pi_i (1 - \pi_i), \quad i = 1, 2, \ldots, d
\] (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 \( \pi_i \) 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 \( \theta_i = \ln\left(\frac{\pi_i}{1 - \pi_i}\right) \), 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 \( \text{Var}[Y_i] \gg m_i \pi_i (1 - \pi_i) \), 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

\[
\Gamma = \begin{pmatrix}
\gamma_{11} & \gamma_{12} & \cdots & \gamma_{1d} \\
\gamma_{21} & \gamma_{22} & \cdots & \gamma_{2d} \\
\vdots & \vdots & \ddots & \vdots \\
\gamma_{d1} & \gamma_{d2} & \cdots & \gamma_{dd}
\end{pmatrix}
\]

where \( \gamma_{ij}, \ i \neq i^*, \ i^*, i = 1, 2, \ldots, d \), 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 \( i \)th margin, it is important to give an extension of (3), i.e.,

\[
\text{Var}[Y_i] = \lambda_i m_i \pi_i (1 - \pi_i),
\]

where \( \lambda_i \).
and the hypothesis testing problem is formulated for all $i = 1, 2, \ldots, d$ as

$$H_0_i : \lambda_i = 1 \quad versus \quad H_1_i : \lambda_i > 1$$

(4)

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

$$\chi^2_i = J_i^T A^{-1}_i J_i, \quad i = 1, 2, \ldots, d$$

(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 $i$th margin with respect to every $j$th parameter, namely

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

(6)

and $A_i$ is the covariance matrix of $J_i$ corrected for estimation of $\theta_i$, 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 $i$th margin with respect to the $j$th parameter and the $k$th observation, we have

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

and

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

Under the null hypothesis of nominal dispersion (4), the asymptotic distribution of (5) is the central $\chi^2$-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 $C : [0, 1]^d \rightarrow [0, 1]$ 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
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.

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

where $l_{N, \text{max}}$ 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 $\hat{q}^* = 0$. Details for deriving the estimates of $\hat{p}^*$, $\hat{q}^*$ and $\hat{r}^*$ 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

$$H(y_1, y_2, \ldots, y_d) = C(F_1(y_1), F_2(y_2), \ldots, F_d(y_d))$$

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

$$C = \{C_\omega : \omega \in \Omega, \Omega \subseteq \mathbb{R}^d, d \geq 1\}$$
so we must test,

\[ H_0 : C \in \mathcal{C} \quad \text{versus} \quad H_1 : C \notin \mathcal{C} \quad (10) \]

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

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

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

\[ \tilde{U}_l = (\tilde{F}_{l,1}, \ldots, \tilde{F}_{l,d}), \quad l = 1, 2, \ldots, N \]

with \( N \) being the size of a random sample from (11); it is important to recall that, under probability transformations, it is expected that \( \tilde{F}_{l,i} \sim U(0,1) \) 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

\[ S_N = \sum_{l=1}^{N} \{ C_N(\tilde{U}_l) - C_{\omega_N}(\tilde{U}_l) \}^2 \quad (12) \]

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 $Y_i | P_i \sim bin(m_i, P_i)$, 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

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

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

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

and the conditional variance is

$$Var(Y_i | P_i) = m_i \pi_i (1 - \pi_i) \frac{\alpha_i + \beta_i + m_i}{\alpha_i + \beta_i + 1}$$

$$= m_i \pi_i (1 - \pi_i) \{1 + \phi_i (m_i - 1)\}, \quad i = 1, 2, \ldots, d \quad (14)$$

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 $\phi_i$ 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

$$C_\Phi(U_1, U_2, \ldots, U_d) = H(F_1(Y_1), F_2(Y_2), \ldots, F_d(Y_d) | \Gamma) \quad (15)$$

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

$$f_Y(y; \mu, \phi, \Gamma) = c_\Phi \{F_1(y_1), F_2(y_2), \ldots, F_d(y_d) | \Gamma\} \prod_{i=1}^{d} f(y_i; \tau_i, \phi_i)$$
Statistical Model for Plant Pathogens

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

\[
P(Y_1 = y_1, Y_2 = y_2, \ldots, Y_d = y_d) = \sum_{j_1=1}^{2} \sum_{j_2=1}^{2} \cdots \sum_{j_d=1}^{2} (-1)^{j_1+j_2+\cdots+j_d} C_d(u_{1j_1}, u_{2j_2}, \ldots, u_{dj_d} | \Gamma) \tag{16}
\]

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 \( \Gamma \), 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

\[
\frac{\partial C_d(u_{1j_1}, u_{2j_2}, \ldots, u_{dj_d} | \Gamma)}{\partial \Gamma} = 0
\]

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

\[
V = D_n^{-1} M_h(D_n^{-1})^T \tag{17}
\]
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 \( \eta \). The estimation of \( N^{-1}V \), which is the asymptotic covariance matrix of the MLE of \( \hat{\eta} \), can be done via Jackknife, viz.,

\[
\Omega = \sum_{l=1}^{N} (\hat{\eta}^{(l)} - \hat{\eta})^T (\hat{\eta}^{(l)} - \hat{\eta})
\]

(18)

In (18), \( \hat{\eta}^{(l)} \) is the estimator of \( \eta \) once the \( l \)th 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 \textit{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 \textit{Peronospora} sp., \( Y_2 \) the leaf blight \textit{Stemphylium} sp. and \( Y_3 \) the leaf spot \textit{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 (1); 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 (15), the Godambe’s asymptotic covariance matrix was estimated. About marginal dispersion parameters, the nominal dispersion (14) was not rejected, under 3-variate framework, for \( Y_2 \), i.e., \( \phi_2 \simeq 0 \) for the random variable associated to \textit{Stemphylium} fungus, an endemic plant pathogen; see Table 1.

<table>
<thead>
<tr>
<th>Parameter estimations</th>
<th>Standard error</th>
<th>Lower Limit</th>
<th>Upper Limit</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \phi_1 )</td>
<td>0.01983</td>
<td>0.0037</td>
<td>0.0125</td>
</tr>
<tr>
<td>( \phi_2 )</td>
<td>0.00377</td>
<td>0.1979</td>
<td>-0.3840</td>
</tr>
<tr>
<td>( \phi_3 )</td>
<td>0.01735</td>
<td>0.0035</td>
<td>0.0104</td>
</tr>
</tbody>
</table>

\( (*) \quad \alpha = 0.05. \)

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.,

\[ \hat{\Gamma} = \begin{pmatrix} 1.000 & 0.484** & 0.475** \\ 0.484** & 1.000 & 0.688** \\ 0.475** & 0.688** & 1.000 \end{pmatrix} \]

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.

<table>
<thead>
<tr>
<th>Factor(variable)</th>
<th>MVN(1)</th>
<th>GLM(1)</th>
<th>ODM(1)</th>
<th>CGB(1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>T0(y1)</td>
<td>0.0266*</td>
<td>0.103*</td>
<td>0.163*</td>
<td>0.157*</td>
</tr>
<tr>
<td>T1(y1)</td>
<td>0.0266*</td>
<td>0.106*</td>
<td>0.166*</td>
<td>0.141*</td>
</tr>
<tr>
<td>T2(y1)</td>
<td>0.0266</td>
<td>0.112</td>
<td>0.179</td>
<td>0.221</td>
</tr>
<tr>
<td>T3(y1)</td>
<td>0.0266</td>
<td>0.110*</td>
<td>0.174</td>
<td>0.176</td>
</tr>
<tr>
<td>T4(y1)</td>
<td>0.0266*</td>
<td>0.107*</td>
<td>0.168*</td>
<td>0.159*</td>
</tr>
<tr>
<td>T5(y1)</td>
<td>0.0266</td>
<td>0.112</td>
<td>0.177</td>
<td>0.166</td>
</tr>
<tr>
<td>T6(y1)</td>
<td>0.0266</td>
<td>0.113</td>
<td>0.179</td>
<td>0.197</td>
</tr>
<tr>
<td>JUNCA(y1)</td>
<td>0.0133*</td>
<td>0.052*</td>
<td>0.082*</td>
<td>0.086*</td>
</tr>
<tr>
<td>T0(y2)</td>
<td>0.0186</td>
<td>0.093</td>
<td>0.106</td>
<td>0.113</td>
</tr>
<tr>
<td>T1(y2)</td>
<td>0.0186*</td>
<td>0.091*</td>
<td>0.103*</td>
<td>0.136*</td>
</tr>
<tr>
<td>T2(y2)</td>
<td>0.0186*</td>
<td>0.094</td>
<td>0.107</td>
<td>0.135</td>
</tr>
<tr>
<td>T3(y2)</td>
<td>0.0186*</td>
<td>0.091*</td>
<td>0.104*</td>
<td>0.122*</td>
</tr>
<tr>
<td>T4(y2)</td>
<td>0.0186</td>
<td>0.096</td>
<td>0.109</td>
<td>0.121</td>
</tr>
<tr>
<td>T5(y2)</td>
<td>0.0186</td>
<td>0.095</td>
<td>0.108</td>
<td>0.142</td>
</tr>
<tr>
<td>T6(y2)</td>
<td>0.0186</td>
<td>0.098</td>
<td>0.111</td>
<td>0.107</td>
</tr>
<tr>
<td>JUNCA(y2)</td>
<td>0.0092</td>
<td>0.046</td>
<td>0.053</td>
<td>0.061</td>
</tr>
<tr>
<td>T0(y3)</td>
<td>0.0265</td>
<td>0.108</td>
<td>0.164</td>
<td>0.169</td>
</tr>
<tr>
<td>T1(y3)</td>
<td>0.0265</td>
<td>0.104</td>
<td>0.160</td>
<td>0.196</td>
</tr>
<tr>
<td>T2(y3)</td>
<td>0.0265</td>
<td>0.109</td>
<td>0.167</td>
<td>0.202</td>
</tr>
<tr>
<td>T3(y3)</td>
<td>0.0265</td>
<td>0.105</td>
<td>0.161</td>
<td>0.210</td>
</tr>
<tr>
<td>T4(y3)</td>
<td>0.0265</td>
<td>0.107</td>
<td>0.163</td>
<td>0.176</td>
</tr>
<tr>
<td>T5(y3)</td>
<td>0.0265</td>
<td>0.105</td>
<td>0.162</td>
<td>0.208</td>
</tr>
<tr>
<td>T6(y3)</td>
<td>0.0265</td>
<td>0.113*</td>
<td>0.170</td>
<td>0.161</td>
</tr>
<tr>
<td>JUNCA(y3)</td>
<td>0.0130</td>
<td>0.053</td>
<td>0.081</td>
<td>0.089</td>
</tr>
</tbody>
</table>

(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 Y2, 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 "⋄". 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

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

∗ = significant effect in relation to MVN modeling, $\alpha = 0.05$.  
⋄ = 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 & Nešlehová (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 co-
efficient, 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 $\Gamma$ as a depen-
dence 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 method-
ology for inference. Hence, no further discussion on this topic are exposed here
(Genest & Nešlehová 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. Addi-
tionally, 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 overdis-
persion 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.
References


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


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


