An automated MCEM algorithm for hierarchical models with multivariate and multitype response variables.

Vera Georgescu, Nicolas Desassis, Samuel S. Soubeyrand, André Kretzschmar, Rachid Senoussi

Communications in Statistics Theory and Methods (2012), 28 p., DOI: 10.1080/03610926.2012.700372

An automated MCEM algorithm for hierarchical models with multivariate and multitype response variables

Vera Georgescu[1,*], Nicolas Desassis[2], Samuel Soubeyrand[1], André Kretzschmar[1], Rachid Senoussi[1]

May 29, 2012

[1] INRA, UR546 Biostatistique et Processus Spatiaux, F-84914 Avignon, France

[2] École Nationale supérieure des Mines de Paris, Centre de Géosciences, F-77300 Fontainebleau

[*] Corresponding author: vera.georgescu@avignon.inra.fr

Abstract

In this paper, we consider a model allowing the analysis of multivariate data, which can contain data attributes of different types (e.g. continuous, discrete, binary). This model is a two-level hierarchical model which supports a wide range of correlation structures and can accommodate overdispersed data. Maximum likelihood estimation of the model parameters is achieved with an automated Monte Carlo Expectation Maximization (MCEM) algorithm. Our method is tested in a simulation study in the bivariate case and applied to a dataset dealing with beehive activity.

Key words: Continuous data ; Count data ; Mixed mode data ; Monte Carlo EM ; Overdispersion ; Poisson-log normal distribution.

1 Introduction

Data with attributes of different types are encountered in many fields. For instance in ecological studies, abundance data of several species measured at different sites can be counts (discrete), species coverage, weights (continuous), occurrence (binary). Nevertheless, there is a lack of classes of distributions which can take into account these different types of data and are easy to adapt to different situations, while allowing a wide range of correlations between the variables.

The scope of this article is to provide a maximum likelihood estimation method adapted to a model describing multiple response data which can contain variables of different types (discrete, continuous, binary) and which supports a wide range of correlation structures.

The model of interest generalizes the multivariate Poisson log normal (MPLN) model studied by Aitchison and Ho (1989). The MPLN model is a multivariate log normal mixture of independent Poisson distributions. This model provides a parametric class of distributions for the analysis of multivariate count data, that is able to describe a wide range of correlation and overdispersion situations. Unlike other multivariate discrete distributions, such as the multivariate Poisson distribution (first proposed by McKendrick and Wicksell), the MPLN model supports negative correlation between counts. Moreover, it can fit overdispersed data, whereas in the multivariate Poisson model the marginal mean and variance coincide (see Aitchison and Ho, 1989 for a detailed comparison between the MPLN model and the multivariate Poisson model). It seems therefore better suited to model multivariate count data such as species count data in ecological studies, which is generally overdispersed and can be negatively correlated.

The general model that we deal with in this article is a two-layer hierarchical model, in which the hidden layer is a multivariate Gaussian distribution, and the observed layer is a multivariate distribution formed by independent univariate distributions chosen according to the type of variable. We chose the multivariate normal distribution for the hidden layer because it has been extensively studied and it provides a full range of correlations between variables (including negative correlation). The hierarchical structure of the model allows overdispersion in the marginal distributions.

Very recently, Chagneau et al. (2010) proposed a spatial model for random variables of different types with a Bayesian estimation procedure based on MCMC simulations. The principle of our approach is similar in that the dependence between variables is expressed at the hidden level of a hierarchical model and the obtention of different types of variables is achieved by using different conditionally independent univariate distributions and link functions. In this paper, we formalize these multivariate hierarchical models for conditional distributions belonging to the exponential class and present some of their properties in a non spatial framework.

Then, we provide a maximum likelihood estimation procedure for these models easy to adapt to different distributions from the exponential family. This procedure is based on an extension (Wei and Tanner, 1990; Booth and Hobert, 1999) of the Expectation-Maximization (EM) algorithm of Dempster et al. (1977). See McLachlan and Krishnan (2008) for a broad presentation of the EM

algorithm and its extensions.

In the next section we present the general model and its properties for given conditional distributions in the observed layer. The maximum likelihood based estimation procedure is presented in Section 3. In Section 4 we test our method on simulated bivariate data of different types. An application on a real dataset concerning the honey-bee hive activity in the South of France is presented in Section 5. Our results and perspectives are discussed in Section 6.

2 Multivariate hierarchical model

2.1 Definition of the general model

Let $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ denote a random sample of size n of the d-dimensional random vector $\boldsymbol{Y} = (Y_1, \ldots, Y_d)$. In practice $\boldsymbol{Y}_i$ could correspond to the abundances of d species observed at location i. Throughout this article i labels the observation and j the variable.

We define the following hierarchical model for $\boldsymbol{Y}$:

$$\begin{cases} \boldsymbol{Y}_i|\boldsymbol{\theta}_i & \sim & \mathscr{L}(g^{-1}(\boldsymbol{\theta}_i)) \\ \boldsymbol{\theta}_i & \sim & \mathscr{N}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \end{cases}$$

where $\mathscr{N}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the d-dimensional normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, g is a set of link functions and $\mathscr{L}$ is a multivariate distribution with parameters $g^{-1}(\boldsymbol{\theta}_i)$.

In this article only the case where $\mathscr{L}$ is formed by d independent univariate distributions $\mathscr{L} = \mathscr{L}_1 \times \ldots \times \mathscr{L}_d$ is considered. The choice of $\mathscr{L}_j$ and g_j^{-1}, for $j \in \{1, \ldots, d\}$, depends on the type of data (discrete, continuous, ordinal, binary). Note that the variables are not necessarily of the same type, different univariate distributions and different link functions can be used for the d variables. In the remainder of the article, the choice of $\mathscr{L}$ will be restricted to exponential families. This is not a very restrictive choice since the exponential family encompasses a broad set of parametric distributions including the most commonly-used (such as Gaussian, Poisson, Bernoulli, Binomial, Gamma).

The probability density $f_{\boldsymbol{Y}}$ of $\boldsymbol{Y}$ is defined by:

$$f_{\boldsymbol{Y}}(\boldsymbol{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \int_{\mathbb{R}^d} f_{\boldsymbol{Y}|\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{\theta}) f_{\boldsymbol{\theta}}(\boldsymbol{\theta}|\boldsymbol{\mu}, \boldsymbol{\Sigma})\, d\boldsymbol{\theta}, \tag{1}$$

where $f_{\boldsymbol{Y}|\boldsymbol{\theta}}$ denotes the conditional probability density function of $\boldsymbol{Y}$, given the variable $\boldsymbol{\theta}$, $f_{\boldsymbol{\theta}}$ is the multivariate Gaussian density and a realization of a random vector is denoted by the corresponding

lower-case letter. Since the distributions of the d variables are conditionally independent given $\boldsymbol{\theta}$, we have:

$$f_{\boldsymbol{Y}|\boldsymbol{\theta}}(y_1 \dots y_d|\boldsymbol{\theta}) = \prod_{j=1}^{d} f_{Y_j|\boldsymbol{\theta}}(y_j|\boldsymbol{\theta}). \tag{2}$$

Unless conjugate distributions are chosen, there is no simplification of the multiple integral in equation (1) for most choices of $\mathscr{L}$ and g^{-1}. Nevertheless, in some cases, as will be seen in Section 2.3, its first two moments can be obtained in terms of the moments of $\boldsymbol{\theta}$ by using conditional expectation results and properties of the chosen distributions. Conversely, the moments of the hidden variable $\boldsymbol{\theta}$ can then be written in terms of the moments of the data $\boldsymbol{Y}$ and used to initialize the parameters in the estimation procedure presented in Section 3.

Remark : A more general form of the model could be defined by allowing the dimension of $\boldsymbol{\theta}_i$ to differ from the dimension of $\boldsymbol{Y}_i$, but for the sake of simplicity we chose the same dimension d. This could be used to introduce spatial correlations.

2.2 Exponential families

If the density of $\mathscr{L}_j$, the conditional distribution of the jth variable given $\boldsymbol{\theta}$, belongs to an exponential family, it can be written in the form:

$$f_{Y_j|\boldsymbol{\theta}}(y_j|\boldsymbol{\theta}) = c_j(y_j) \exp\Big(\sum_{l=1}^{r_j} \eta_{jl}(\boldsymbol{\theta}) T_{jl}(y_j) - b_j(\boldsymbol{\eta}_j)\Big),$$

where:

- r_j is the number of parameters of $\mathscr{L}_j$,
- $\boldsymbol{\eta}_j = (\eta_{j1}, \dots, \eta_{jr_j})$ is the vector of natural parameters of $\mathscr{L}_j$, which can be expressed in terms of $\boldsymbol{\theta}$ and the link funtion g_j,
- $\boldsymbol{T}_j = (T_{j1}, \dots, T_{jr_j})$ is the vector of minimal sufficient statistics of $\mathscr{L}_j$, which can be written in terms of Y_j,
- and $b_j(\boldsymbol{\eta}_j)$ a normalization factor.

We have the following conditional moments for $l \in \{1, \dots, r_j\}$:

$$\mathbb{E}(T_{jl}|\boldsymbol{\theta}) = \frac{\partial b(\boldsymbol{\eta}_j)}{\partial \eta_{jl}}, \tag{3}$$

$$\mathbb{V}(T_{jl}|\boldsymbol{\theta}) = \frac{\partial^2 b(\boldsymbol{\eta}_j)}{\partial \eta_{jl}^2}. \tag{4}$$

Suppose for example that $\mathscr{L}_j$ is the Poisson distribution with mean $\lambda_\theta = g_j^{-1}(\theta_j)$ and g_j^{-1} is the exponential function:

$$Y_j|\boldsymbol{\theta} \sim \mathscr{P}(e^{\theta_j}).$$

Then $r_j = 1$, $\eta_j = \log(\lambda_\theta) = \log e^{\theta_j} = \theta_j$, $T_j = Y_j$ and $b_j(\eta_j) = e^{\eta_j}$. We can verify that equation (3) and (4) hold:

$$\mathbb{E}(T_j|\boldsymbol{\theta}) = \mathbb{E}(Y_j|\boldsymbol{\theta}) = \frac{\partial b(\eta_j)}{\partial \eta_j} = e^{\eta_j} = e^{\log(\lambda_\theta)} = \lambda_\theta,$$
$$\mathbb{V}(T_j|\boldsymbol{\theta}) = \mathbb{E}(Y_j|\boldsymbol{\theta}) = \frac{\partial^2 b(\eta_j)}{\partial \eta_j^2} = \lambda_\theta.$$

Suppose now that $\mathscr{L}_j$ is a two-parameter distribution, say the Gamma distribution with the usual shape parameter $k_{\boldsymbol{\theta}}$ and scale parameter $\lambda_{\boldsymbol{\theta}}$, which can be expressed in terms of $\boldsymbol{\theta}$ by using two link functions, g_{j1} and g_{j2}:

$$k_{\boldsymbol{\theta}} = g_{j1}^{-1}(\boldsymbol{\theta}),$$
$$\lambda_{\boldsymbol{\theta}} = g_{j2}^{-1}(\boldsymbol{\theta}).$$

Then $r_j = 2$, $\boldsymbol{\eta}_j = (k_{\boldsymbol{\theta}} - 1, -\frac{1}{\lambda_{\boldsymbol{\theta}}})$, $\boldsymbol{T}_j = (\log(Y_j), Y_j)$ and $b(\boldsymbol{\eta}_j) = -(\eta_1 + 1)\log(-\eta_2) + \log(\eta_1\Gamma(\eta_1))$. We can verify that the first moments of Y obtained using equation (3) and (4) are indeed the mean and variance of the Gamma distribution:

$$\mathbb{E}(T_{j2}|\boldsymbol{\theta}) = \mathbb{E}(Y_j|\boldsymbol{\theta}) = \frac{\partial b(\eta_j)}{\partial \eta_{j2}} = -\frac{\eta_{j1} + 1}{\eta_{j2}} = k_{\boldsymbol{\theta}}\lambda_{\boldsymbol{\theta}},$$
$$\mathbb{V}(T_{j2}|\boldsymbol{\theta}) = \mathbb{E}(Y_j|\boldsymbol{\theta}) = \frac{\partial^2 b(\eta_j)}{\partial \eta_{j2}^2} = k_{\boldsymbol{\theta}}\lambda_{\boldsymbol{\theta}}^2.$$

2.3 Submodel examples

2.3.1 Multivariate Poisson-Log Normal model (MPLN)

The MPLN model (Aitchison and Ho, 1989) is obtained when $\mathscr{L}$ is formed by d independent Poisson distributions and g^{-1} is the exponential function. For all observations $i \in \{1, \ldots, n\}$ and variables $j \in \{1, \ldots, d\}$ we write:

$$\begin{cases} Y_{ij}|\theta_{ij} & \sim & \mathscr{P}(e^{\theta_{ij}}) \\ (\theta_{i1}, \ldots, \theta_{id})^T & \sim & \mathscr{N}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \end{cases}$$

where the superscript T denotes the transpose of the matrix and $\mathscr{P}$ is the univariate Poisson distribution.

Communications in Statistics Theory and Methods (2012), 28 p., DOI: 10.1080/03610926.2012.700372

The unconditional moments of this distribution can be calculated by using properties of the lognormal distribution and of the conditional expectation (Aitchison and Ho, 1989; Tunaru, 2002):

$$\begin{aligned}
\mathbb{E}(Y_j) &= \mathbb{E}[\mathbb{E}(Y_j|\theta_j)] = e^{\mu_j+\frac{1}{2}\sigma_{jj}} \stackrel{def}{=} m_j, \\
\mathbb{V}(Y_j) &= \mathbb{E}[\mathbb{V}(Y_j|\theta_j)] + \mathbb{V}[\mathbb{E}(Y_j|\theta_j)] \\
&= m_j + m_j^2\ (e^{\sigma_{jj}} - 1), \\
\mathrm{cov}(Y_j, Y_{j'}) &= \mathbb{E}[\mathrm{cov}(Y_j, Y_{j'}|\boldsymbol{\theta})] + \mathrm{cov}[\mathbb{E}(Y_j|\theta_j), \mathbb{E}(Y_{j'}|\theta_{j'})] \\
&= m_j\ m_{j'}\ (e^{\sigma_{jj'}} - 1), \\
\mathrm{cor}(Y_j, Y_{j'}) &= \frac{e^{\sigma_{jj'}} - 1}{\sqrt{(e^{\sigma_{jj}} - 1 + m_j^{-1})(e^{\sigma_{j'j'}} - 1 + m_{j'}^{-1})}},
\end{aligned}$$

where $\boldsymbol{\Sigma} = (\sigma_{jj'})$ and $j, j' \in \{1, \ldots, d\}$ for $j \neq j'$. Some interesting features of the model appear:

(i) $\mathbb{V}(Y_j) \geq \mathbb{E}(Y_j)$, so there is overdispersion for the marginal distributions with respect to the Poisson distribution,

(ii) the signs of the correlation between observed variables and the correlation between the hidden normally distributed variables $\boldsymbol{\theta}$ correspond,

(iii) the range of correlation is not as wide as that of the corresponding normal distribution:

$$|\mathrm{cor}(Y_j, Y_{j'})| < |\mathrm{cor}(\theta_j, \theta_{j'})|.$$

Aitchison and Ho (1989) studied the regions of count correlation and overdispersion attainable by the bivariate Poisson-log normal model for different mean counts m.

2.3.2 Bivariate Poisson-Normal model

Data of different types (e.g. continuous and discrete) can be obtained by using different distributions and link functions for the variables. We define the bivariate Poisson-Normal model by:

$$\begin{cases}
Y_{i1}|\theta_{i1} & \sim & \mathscr{P}(e^{\theta_{i1}}) \\
Y_{i2}|\theta_{i2} & = & g_2^{-1}(\theta_{i2}) \\
(\theta_{i1}, \theta_{i2})^T & \sim & \mathscr{N}_2(\boldsymbol{\mu}, \boldsymbol{\Sigma}),
\end{cases}$$

where g_2^{-1} could be for instance the exponential function, if a positive continuous variable is needed. Notice that the likelihood is easier to compute here than for the MPLN model, because the variable $\boldsymbol{\theta}_2$ is observed in this model.

The moments of this distribution can be calculated in a similar way to the previous model.

$$\begin{aligned}
\mathbb{E}(Y_1) &= \mathbb{E}[\mathbb{E}(Y_1|\theta_1)] = e^{\mu_1+\frac{1}{2}\sigma_{11}} \stackrel{def}{=} m_1, \\
\mathbb{V}(Y_1) &= \mathbb{E}[\mathbb{V}(Y_1|\theta_1)] + \mathbb{V}[\mathbb{E}(Y_1|\theta_1)] \\
&= m_1 + m_1^2\ (e^{\sigma_{11}} - 1), \\
\mathrm{cov}(Y_1, Y_2) &= \mathbb{E}[\theta_2\ \mathbb{E}(Y_1|\theta_1)] - \mathbb{E}[\mathbb{E}(Y_1|\theta_1)]\ \mathbb{E}(\theta_2) \\
&= m_1\sigma_{12}, \\
\mathrm{cor}(Y_1, Y_2) &= \frac{\sigma_{12}}{\sqrt{(e^{\sigma_{11}} - 1 + m_1^{-1})\ \sigma_{22}}}.
\end{aligned}$$

The same properties hold: overdispersion for the count variable $\boldsymbol{Y}_1$ and large correlation range between variables, but smaller than the correlation range between θ_{i1} and θ_{i2}.

2.3.3 Bivariate Binomial-Poisson model

We define the bivariate Binomial-Poisson model by:

$$\left\{\begin{array}{rcl}
Y_{i1}|\theta_{i1} & \sim & \mathscr{B}(n_b, \mathrm{logit}^{-1}(\theta_{i1})) \\
Y_{i2}|\theta_{i2} & \sim & \mathscr{P}(e^{\theta_{i2}}) \\
(\theta_{i1}, \theta_{i2})^T & \sim & \mathscr{N}_2(\boldsymbol{\mu}, \boldsymbol{\Sigma}),
\end{array}\right.$$

where $\mathscr{B}$ denotes the univariate binomial distribution, with parameters n_b the number of Bernoulli trials and success probability $\mathrm{logit}^{-1}(\theta_{i1})$, where $\mathrm{logit}^{-1}(x) = \frac{1}{1+e^{-x}}$.

The moments of this distribution cannot be written in a closed form (see Appendix A) but their properties can be studied by simulation or numerical computation. We studied the range of the count correlation coefficient $\mathrm{cor}(Y_1, Y_2)$ for given values of $\mu_1, \mu_2, \sigma_{11}, \sigma_{22}$. The results given in Figure 1 show once more that there is a direct correspondence between the signs of $\mathrm{cor}(Y_1, Y_2)$ and r=$\mathrm{cor}(\theta_1, \theta_2)$, while the range of $\mathrm{cor}(Y_1, Y_2)$ is smaller.

2.3.4 Bivariate Gamma-Poisson model

This is another example of model combining variables of different types (continuous vs discrete), which uses an exponential family with two parameters, the Gamma distribution. We define the

Communications in Statistics Theory and Methods (2012), 28 p., DOI: 10.1080/03610926.2012.700372

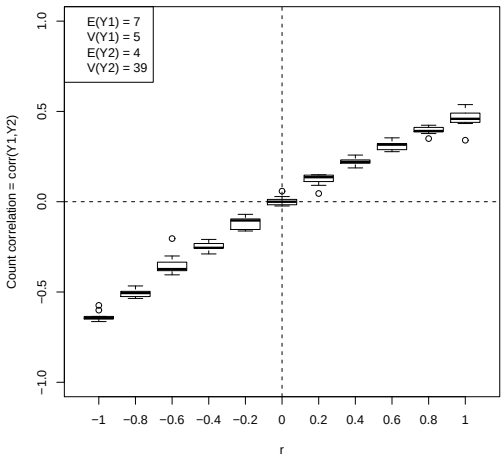

Figure 1: Evolution of the correlation between Y_1, Y_2 with r, (the correlation between θ_1, θ_2) for the Binomial-Poisson model, obtained by simulating 100 samples of size $n = 1000$ with parameters $\mu_1 = \mu_2 = \sigma_{11} = \sigma_{22} = 1$ and $n_b = 10$

bivariate Gamma-Poisson model by:

$$\begin{cases} Y_{i1}|(\theta_{i1}, \theta_{i2}) & \sim & \mathscr{G}(k_{\boldsymbol{\theta}}, \lambda_{\boldsymbol{\theta}}) \\ Y_{i2}|\theta_{i1} & \sim & \mathscr{P}(e^{\theta_{i1}}) \\ (\theta_{i1}, \theta_{i2})^T & \sim & \mathscr{N}_2(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \end{cases}$$

where the shape parameter $k_{\boldsymbol{\theta}}$ and the scale parameter $\lambda_{\boldsymbol{\theta}}$ of the Gamma distribution $\mathscr{G}$ depend on $\boldsymbol{\theta}$, by defining the mean and variance of the Gamma distribution by:

$$\mathbb{E}(Y_{i1}|\theta_{i1}, \theta_{i2}) = k_{\boldsymbol{\theta}}\lambda_{\boldsymbol{\theta}} = e^{\theta_{i1}},$$
$$\mathbb{V}(Y_{i1}|\theta_{i1}, \theta_{i2}) = k_{\boldsymbol{\theta}}\lambda_{\boldsymbol{\theta}}^2 = e^{\theta_{i2}}.$$

This yields:

$$k_{\boldsymbol{\theta}} = g_{11}^{-1}(\boldsymbol{\theta}) = e^{2\theta_{i1} - \theta_{i2}},$$
$$\lambda_{\boldsymbol{\theta}} = g_{12}^{-1}(\boldsymbol{\theta}) = e^{\theta_{i2} - \theta_{i1}}.$$

This model could be interpreted in the following way in an ecological framework: Suppose two species abundances were observed over n locations. Y_1 denotes the weight or surface occupied by species 1, and Y_2 counts of species 2. This model assumes that the expected values of these two variables depend on an unobserved variable θ_1, say resource availability. This unobserved factor θ_1

is linked to another unobserved variable θ_2, which only influences the variance of the abundance of species 1. θ_2 could be for instance a third species which is a competitor of species 1 but has no influence on species 2.

3 Maximum likelihood estimation via the MCEM algorithm

The model that we propose has several interesting properties: it is easy to adapt to different types of data and provides a wide correlation range between variables. The price to pay for these advantages is the increased computational complexity required for parameter estimation. It is therefore important to have a generic estimation procedure that is easy to adapt to different distributions $\mathscr{L}$ and link functions g and thus does not depend on their specific properties. For the MPLN model, Aitchison and Ho (1989) used a maximum likelihood estimation procedure (mix of Newton Raphson and steepest ascent) based on a numerical integration procedure which depends on the specific form of the MPLN likelihood. Tunaru (2002) and Chagneau et al. (2010) use a Bayesian estimation procedure (MCMC algorithm). We built a maximum likelihood estimation procedure based on the EM algorithm.

Let $\mathbf{\Phi}$ denote the unknown parameter vector $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Since $\boldsymbol{\theta}$ is not observed, the Expectation-Maximization (EM) algorithm of Dempster et al. (1977) is well suited for the maximum likelihood estimation of $\mathbf{\Phi}$; see McLachlan and Krishnan (2008) for a complete review of the EM algorithm and its extensions. The idea behind the EM algorithm is to complete the observed data $\boldsymbol{y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$ with the latent variable vector $\boldsymbol{\theta}$, write the complete-data loglikelihood :

$$l_c(\mathbf{\Phi}; \boldsymbol{y}, \boldsymbol{\theta}) = \sum_{i=1}^{n} (\log f_{\boldsymbol{Y}|\boldsymbol{\theta}}(\boldsymbol{y}_i|\boldsymbol{\theta}_i) + \log f_{\boldsymbol{\theta}}(\boldsymbol{\theta}_i; \mathbf{\Phi})) \tag{5}$$

and maximize the conditional likelihood expectation, given the observed data $\boldsymbol{y}$, in terms of $\mathbf{\Phi}$.

The EM algorithm is a two-step iterative algorithm that proceeds as follows: At iteration $t+1$, the current parameter $\mathbf{\Phi}^{(t)}$ is known.

– E-step (Expectation): the conditional expectation of the complete-data log-likelihood given the observed data $\boldsymbol{y}$ and the current parameter estimates is computed:

$$Q(\mathbf{\Phi}, \mathbf{\Phi}^{(t)}) = \mathbb{E}_{\mathbf{\Phi}^{(t)}}[l_c(\mathbf{\Phi}; \boldsymbol{Y}, \boldsymbol{\theta})|\boldsymbol{Y} = \boldsymbol{y}]. \tag{6}$$

– M-step (Maximization): the parameter estimates are updated by:

$$\mathbf{\Phi}^{(t+1)} = \arg\max_{\mathbf{\Phi}} Q(\mathbf{\Phi}, \mathbf{\Phi}^{(t)}).$$

The algorithm is stopped if some convergence criterion is satisfied. In our case, the expression of the Q-function (6) contains an integral over the $\boldsymbol{\theta}$-values in $\mathbb{R}^d$, so the E-step cannot be solved analytically. We use therefore an extension of the EM to approximate Q in the E-step, namely the Monte Carlo EM (Wei and Tanner, 1990), and more specifically the automated MCEM version proposed by Booth and Hobert (1999).

3.1 Expectation step:

Since the first term of equation (5) does not depend on the parameters $\mathbf{\Phi}$, the Q-function can be written:

$$Q(\mathbf{\Phi}, \mathbf{\Phi}^{(t)}) = \mathbb{E}_{\mathbf{\Phi}^{(t)}}[\log f_{\boldsymbol{\theta}}(\boldsymbol{\theta}; \mathbf{\Phi}) | \boldsymbol{Y} = \boldsymbol{y}] + c(\boldsymbol{y}), \tag{7}$$

where c is independent of $\mathbf{\Phi}$.

To calculate the expectation term in equation (7) the density $f_{\boldsymbol{\theta}|\boldsymbol{Y}}(\boldsymbol{\theta}|\boldsymbol{y}; \mathbf{\Phi}^{(t)})$, given by:

$$f_{\boldsymbol{\theta}|\boldsymbol{Y}}(\boldsymbol{\theta}|\boldsymbol{y}; \mathbf{\Phi}) = \frac{f_{\boldsymbol{Y}|\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{\theta}) f_{\boldsymbol{\theta}}(\boldsymbol{\theta}; \mathbf{\Phi})}{f_{\boldsymbol{Y}}(\boldsymbol{y}; \mathbf{\Phi})}, \tag{8}$$

has to be evaluated. The evaluation of $f_{\boldsymbol{Y}}(\boldsymbol{y}; \mathbf{\Phi})$ is difficult because of the integral in equation (1). The solution offered by Monte Carlo EM is to simulate at each EM iteration t and for each observation $\boldsymbol{y}_i$ a random sample $\boldsymbol{\theta}_{i1}^{(t)}, \ldots, \boldsymbol{\theta}_{iN}^{(t)}$ from the distribution $f_{\boldsymbol{\theta}|\boldsymbol{Y}}$ and to replace Q_i, the conditional expectation of the complete-data log-likelihood at observation site i, with a Monte Carlo approximation of the expectation:

$$Q_i(\mathbf{\Phi}, \mathbf{\Phi}^{(t)}) \simeq \frac{1}{N} \sum_{k=1}^{N} \log f_{\boldsymbol{\theta}|\boldsymbol{Y}}(\boldsymbol{\theta}_{ik}^{(t)} | \boldsymbol{y}_i; \mathbf{\Phi}) + c(\boldsymbol{y}_i).$$

Since the observations $\boldsymbol{y}_i$ are independent, we have:

$$Q(\mathbf{\Phi}, \mathbf{\Phi}^{(t)}) = \sum_{i=1}^{n} Q_i(\mathbf{\Phi}, \mathbf{\Phi}^{(t)}).$$

In our case it is difficult to sample from $f_{\boldsymbol{\theta}|\boldsymbol{Y}}$ so we use an alternative of the MCEM algorithm based on importance sampling proposed by Booth and Hobert (1999).

Communications in Statistics Theory and Methods (2012), 28 p., DOI: 10.1080/03610926.2012.700372

Student importance sampling

The random sample $\boldsymbol{\theta}_{i1}^{(t)}, \ldots, \boldsymbol{\theta}_{iN}^{(t)}$ is simulated from the importance density h_t, which has the same support as $f_{\boldsymbol{\theta}|\boldsymbol{Y}}$. The importance sampling Monte Carlo estimate of Q is defined for a given observation i by the following expression:

$$Q_i(\boldsymbol{\Phi}, \boldsymbol{\Phi}^{(t)}) \simeq \frac{1}{N} \sum_{k=1}^{N} w_{ik} \log f_{\boldsymbol{\theta}}(\boldsymbol{\theta}_{ik}^{(t)}|\boldsymbol{\Phi}) + c(\boldsymbol{y}_i),$$

where w_{ik} are the importance weights defined by:

$$w_{ik} = \frac{f_{\boldsymbol{\theta}|\boldsymbol{Y}}(\boldsymbol{\theta}_{ik}^{(t)}|\boldsymbol{y}; \boldsymbol{\Phi}^{(t)})}{h_t(\boldsymbol{\theta}_{ik}^{(t)})} \propto \frac{f_{\boldsymbol{Y}|\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{\theta}) f_{\boldsymbol{\theta}}(\boldsymbol{\theta}_{ik}^{(t)}; \boldsymbol{\Phi}^{(t)})}{h_t(\boldsymbol{\theta}_{ik}^{(t)})}$$

and evaluated up to the normalizing constant $f_{\boldsymbol{Y}}(\boldsymbol{y}; \boldsymbol{\Phi}^{(t)})$ (which does not depend on $\boldsymbol{\Phi}$ and therefore has no effect on the M-step).

The importance density h_t we use is a multivariate Student t-distribution, as Booth and Hobert (1999) suggested. This has proved to be a very efficient choice when the unknown distribution is approximately ellipsoidal and has a mode (Evans and Swartz, 1996). Its expectation m_t and covariance matrix Σ_t are re-evaluated at each step in order to be approximately:

$$\begin{aligned} m_t &= \mathbb{E}_{\boldsymbol{\Phi}^{(t)}}[\boldsymbol{\theta}|\boldsymbol{y}], \\ \Sigma_t &= \mathbb{V}_{\boldsymbol{\Phi}^{(t)}}[\boldsymbol{\theta}|\boldsymbol{y}]. \end{aligned}$$

These quantities are obtained at each MCEM iteration by an iterative algorithm, which corresponds to the procedure used to obtain Penalized Quasi Likelihood (PQL) estimators in GLMM models (Breslow and Clayton, 1993). This algorithm is provided in Appendix B and has to be adapted to the distributions $\mathscr{L}_j$ and link functions g_j used.

Size of the importance sample

The size N of the importance sample is re-evaluated at each step in order to obtain a compromise between the speed of the first iterations and the final precision of the estimation. N is increased with the number of iterations by using the automatic procedure proposed by Booth and Hobert (1999) based on a normal approximation of the Monte Carlo error. The algorithm is initialized with a small value of N, in order to allow a fast evolution at the start when the current estimator of the parameter may be far from the true value, and N is increased if $\|\boldsymbol{\Phi}^{(t+1)} - \boldsymbol{\Phi}^{(t)}\|$ is small compared to the Monte Carlo error, which means that the $(t+1)$th iteration was useless because it was "swamped" by Monte Carlo error.

3.2 Maximization step

In our case, the M-step has an explicit solution. The parameter estimates are obtained by the following expressions:

$$\boldsymbol{\mu}^{(t+1)} = \frac{\sum_{i=1}^{n}\sum_{k=1}^{N} w_{ik}\boldsymbol{\theta}_{ik}^{(t)}}{\sum_{i=1}^{n}\sum_{k=1}^{N} w_{ik}}, \tag{9}$$

$$\boldsymbol{\Sigma}^{(t+1)} = \frac{\sum_{i=1}^{n}\sum_{k=1}^{N} w_{ik}(\boldsymbol{\theta}_{ik}^{(t)} - \boldsymbol{\mu}^{(t+1)})(\boldsymbol{\theta}_{ik}^{(t)} - \boldsymbol{\mu}^{(t+1)})^T}{\sum_{i=1}^{n}\sum_{k=1}^{N} w_{ik}}, \tag{10}$$

which represent the weighted average and the weighted empirical variance of the importance sample simulated at the final iteration. These expressions are obtained by deriving Q_N with respect to $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ and solving the equation $\partial Q_N/\partial\boldsymbol{\mu} = 0$ and $\partial Q_N/\partial\boldsymbol{\Sigma} = 0$ respectively. The reader is referred to Appendix C for a detailed proof of equation (10) in the bivariate case. The proof of equation (9) does not pose any difficulty.

3.3 Stopping rule

The following stopping rule is used:

$$\max_l \frac{|\boldsymbol{\Phi}_l^{(t+1)} - \boldsymbol{\Phi}_l^{(t)}|}{\sqrt{\mathbb{V}(\hat{\boldsymbol{\Phi}}_l)} + \delta_1} < \delta_2, \tag{11}$$

where $\delta_1 = 0.001$, $\delta_2 = 0.01$ and l labels the parameters. The asymptotic variance of the parameter estimates $\mathbb{V}(\hat{\boldsymbol{\Phi}})$ is obtained by using an estimate of the observed Fisher information evaluated at the current parameter estimate (Booth and Hobert, 1999; Tanner, 1991).

Dividing by $\sqrt{\mathbb{V}(\hat{\boldsymbol{\Phi}}_l)}$ instead of $|\boldsymbol{\Phi}_l^{(t)}|$, which is used in standard convergence criterions for deterministic EM algorithms, avoids unnecessary iterations when the estimate is very small compared to its standard error. The algorithm is stopped when rule (11) is satisfied for 3 consecutive iterations, in order to "reduce the risk of stopping the algorithm prematurely because of an unlucky Monte Carlo sample" (Booth and Hobert, 1999).

4 Simulation studies

Results for two of the submodels presented in Section 2, namely the bivariate Poisson-lognormal model (BPLN) and the Binomial-Poisson model, are shown in this section. The Poisson-Normal model is in fact easier to estimate than the bivariate Poisson-lognormal model, since there is only one hidden variable $\boldsymbol{\theta}_1$ instead of two. The results of this submodel are very similar to those of the BPLN submodel and therefore not presented here.

The range of parameters which can be estimated and problems of "practical" identifiability of parameters are studied briefly in the context of the BPLN model. In the Binomial-Poisson case we discuss the precision of the asymptotic standard deviation of the parameter estimates for different sample sizes.

Computer code (in R) is available from the authors upon demand.

4.1 BPLN model

A result of a single run of our estimation procedure is given in Figure 2.

	μ_1	μ_2	σ_1^2	r	σ_2^2
true value φ	1	0	0.5	-0.3	2
mean estimate $\bar{\varphi}$	1.00	0.00	0.49	-0.30	1.98
mASD	0.05	0.10	0.07	0.10	0.27
ESD	0.05	0.09	0.07	0.08	0.23
% IC95	96	99	97	95	97

Table 1: Results for the BPLN model obtained on $n_s = 100$ runs with samples of size $n = 400$.

Our algorithm converged in about 60 iterations. The size N of the importance sample was plotted to illustrate the automatic increase of N with the number of iterations. N increased from an initial value of 10 to 10000 at convergence of the algorithm. The boxplot at the final iteration indicates the results obtained on $n_s = 100$ datasets of size $n = 400$ simulated with the same parameters ($\mu_1 = 1$, $\mu_2 = 0$, $\sigma_1^2 = 0.5$, $\sigma_2^2 = 2$, $r = -0.3$). These results show that our estimation is centered on the true value of the parameters, and the variance of the estimators is reasonably small.

Each run of our algorithm provides an estimation of the asymptotic standard deviation (ASD)

Communications in Statistics Theory and Methods (2012), 28 p., DOI: 10.1080/03610926.2012.700372

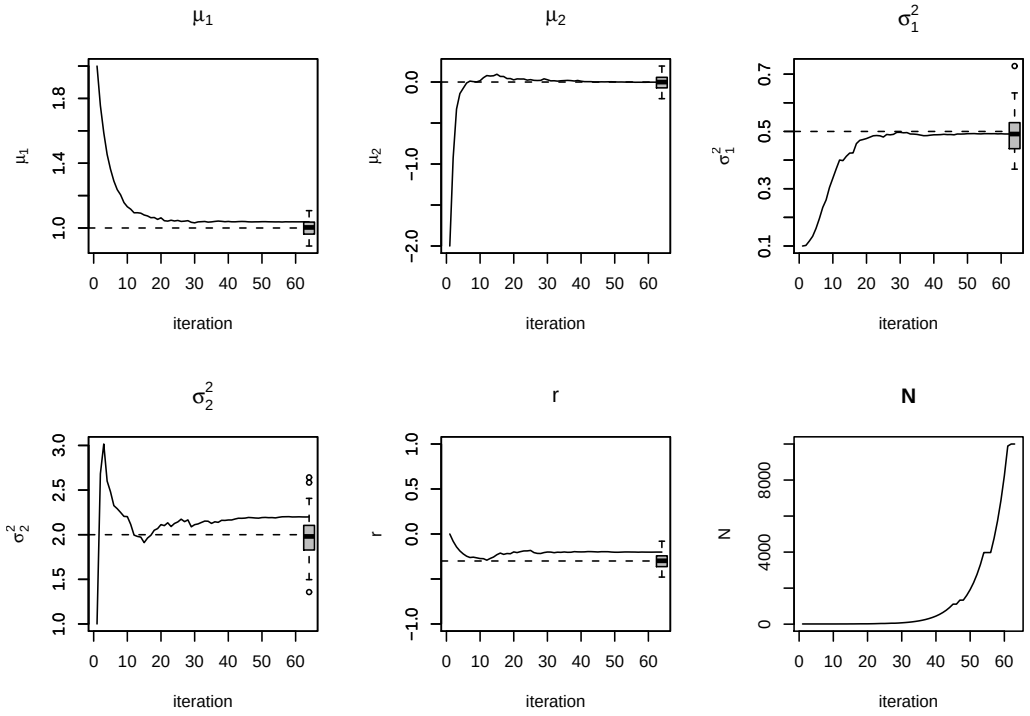

Figure 2: MCEM simulation result for the BPLN model. The true values of the parameters are represented by the horizontal dashed line. The boxplots at the last iteration were obtained on 100 datasets simulated with the same parameters with $n = 400$

of the parameter estimates. Let φ denote an element of the set of parameters $\mathbf{\Phi}$. The final estimates of the parameters ($\hat{\varphi}$), the mean asymptotic standard deviation (mASD), the empirical standard deviation (ESD) and the percentage of ASD leading to a 95% confidence interval ($\hat{\varphi} \pm 1.96$ ASD) containing the true value φ (% IC95) are given in Table 1. The ESD was calculated for each parameter φ by the following formula :

$$\mathrm{ESD}(\hat{\varphi}) = \sqrt{\frac{\sum_{s=1}^{n_s}(\hat{\varphi}_s - \bar{\varphi})^2}{n_s - 1}},$$

where n_s is the number of runs of our algorithm, $\hat{\varphi}_s$ is the estimate of φ obtained at run s and $\bar{\varphi}$ is the mean calculated over the n_s simulations.

Parameter identifiability and estimation limits:

Due to the properties of the Poisson distribution, the correlation coefficient r of the BPLN model is not identifiable for some parameter values: a large mean associated with a small variance for the variable e^{θ} leads to a bad estimation of r (see Table 2), because a Poisson distribution with high

mean will have a high variance, so the variance of the Poisson will erase or "swamp" the correlation between variables.

	$\mathbb{E}(e^\theta)$	$\mathbb{V}(e^\theta)$	$\mu = \mathbb{E}(\theta)$	$\sigma^2 = \mathbb{V}(\theta)$	r
true value φ	2	2	0.49	0.41	-0.9
estimate $\bar{\varphi} \pm 1.96$ ESD			0.475 $\pm$0.2	0.41 $\pm$0.23	$\mathbf{-0.78 \pm 0.35}$
true value φ	2	100	-0.94	3.26	-0.9
estimate $\bar{\varphi} \pm 1.96$ ESD			-0.97 $\pm$0.18	3.35 $\pm$0.2	$\mathbf{-0.88 \pm 0.2}$
true value φ	100	2	4.61	$2 \cdot 10^{-4}$	-0.9
estimate $\bar{\varphi} \pm 1.96$ ESD			4.6 ± 0.02	$10^{-3} \pm 2 \cdot 10^{-3}$	$\mathbf{-0.09 \pm 1.8}$

Table 2: Identifiability of the correlation coefficient r for the BPLN model. Results obtained over 100 datasets of size $n = 400$ simulated with parameters $\mu_1 = \mu_2 = \mu$, $\sigma_1^2 = \sigma_2^2 = \sigma^2$ and $r = -0.9$.

4.2 Binomial-Poisson model

The final estimates obtained on $n_s = 500$ data samples of size $n = 400$ simulated according to the Binomial-Poisson model with parameters $\mu_1 = 0$, $\mu_2 = 0$, $\sigma_1^2 = 2$, $\sigma_2^2 = 1$, $r = -0.8$ and $n_b = 10$ are given in Figure 3 and in Table 3.

One of the advantages of this estimation procedure is the possibility to obtain with a single run of our algorithm the asymptotic standard deviation (ASD) of the parameter estimates. The ASD is very reliable when the data size is large, so our algorithm can be run only once to obtain the parameter estimates with an accurate confidence interval. To show this, we compared the precision of the mean ASD (mASD) with the empirical standard deviation (ESD) calculated over 500 runs. To take into account the variability of the ASD estimates over the n_s simulations, we tested for each run if the true value of the parameter was in the 95% confidence interval obtained using the corresponding ASD estimate (% IC95). The percentage of positive tests are given above each plot and in Table 3.

Influence of sample size on the ASD estimates:

To see if the ASD remains reliable with smaller sample sizes, we performed a similar procedure on 100 runs of our algorithm on data samples simulated according to a Binomial-Poisson model with parameters $\mu_1 = 2$, $\mu_2 = 1$, $\sigma_1^2 = 2$, $\sigma_2^2 = 1$, $r = 0.8$ and different sample sizes.

Communications in Statistics Theory and Methods (2012), 28 p., DOI: 10.1080/03610926.2012.700372

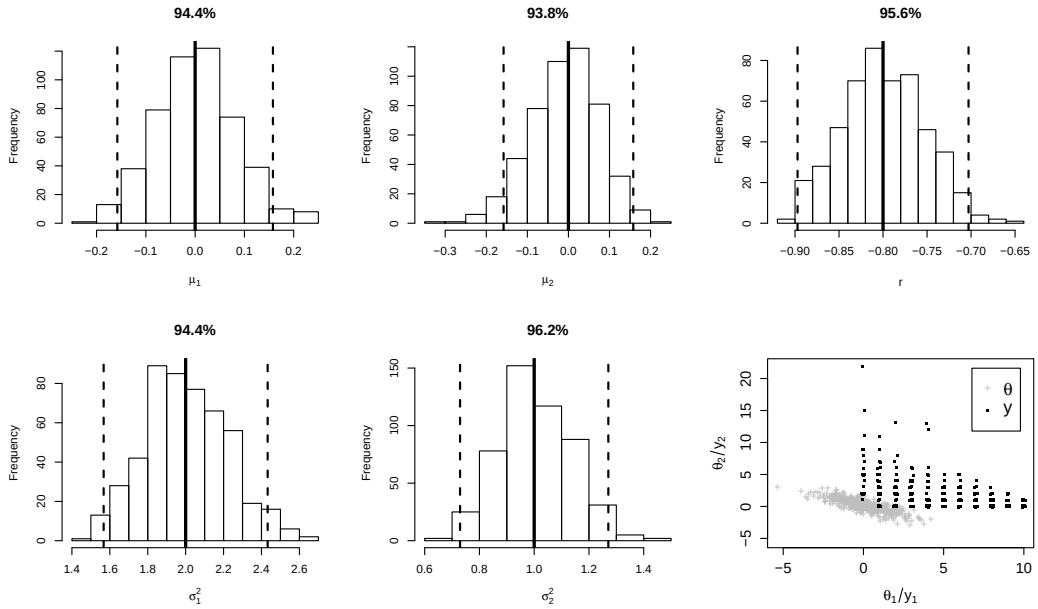

Figure 3: Histograms of the parameter estimates of the Binomial-Poisson model over 500 runs and true values (bold line). The 95% confidence interval was computed using the mean asymptotic standard deviation (dashed lines). The percentage of ASD leading to a 95% confidence interval which contains the true parameter value is given above each plot. Bottomright: example of a dataset simulated with these parameter values (both $\boldsymbol{\theta}$ and $\boldsymbol{Y}$ are represented)

	μ_1	μ_2	σ_1^2	r	σ_2^2
true value	0.0	0.0	2.0	-0.8	1.0
mean estimate	0.00	-0.01	2.00	-0.80	1.00
mASD	0.08	0.08	0.22	0.05	0.14
ESD	0.08	0.08	0.22	0.05	0.14
% IC95	94.4	93.8	94.4	95.6	96.2

Table 3: Results for the Binomial-Poisson model obtained on 500 runs with samples of size $n = 400$.

		μ_1	μ_2	σ_1^2	r	σ_2^2
size n	true value	2.0	1.0	2.0	0.8	1.0
400	mean estimate	2.00	1.00	1.99	0.81	1.00
200		2.00	0.99	2.03	0.81	1.00
100		1.98	0.99	1.92	0.82	1.04
50		2.00	0.96	2.11	0.82	1.02
30		1.98	1.01	1.88	0.82	0.91

Table 4: Mean estimates of the parameters of the Binomial-Poisson model over 100 runs for different sample sizes. The Standard deviation of the estimates are given in Table 5.

The resulting mean parameter estimates are given in Table 4 and the evolution of the ASD with the sample size is provided in Table 5 for $n \in \{30, 50, 100, 200, 400\}$. As expected, the mean estimates and the ASD become worse when the sample size becomes smaller, but they remain meaningful to test for example for a positive or negative correlation.

5 An application to beehive data

In this section we illustrate our model on two bivariate datasets extracted from a survey of the activity of honey-bee colonies over a large observatory in the south of France.

5.1 Beehive dataset

300 hives nested in 20 apiaries were weighed every two days during 24 days in June 2009. The variation with time of the weight of individual hives was modeled by a logistic curve in order to estimate the maximum weight gain over this period for each hive. This weight gain is a continuous

Communications in Statistics Theory and Methods (2012), 28 p., DOI: 10.1080/03610926.2012.700372

size n		μ_1	μ_2	σ_1^2	r	σ_2^2
400	mASD	0.10 (94%)	0.06 (96%)	0.26 (96%)	0.05 (94%)	0.10 (95%)
	ESD	0.10	0.06	0.26	0.04	0.09
200	mASD	0.14 (91%)	0.09 (96%)	0.37 (92%)	0.07 (93%)	0.15 (95%)
	ESD	0.15	0.09	0.40	0.06	0.15
100	mASD	0.19 (97%)	0.13 (92%)	0.50 (90%)	0.09 (90%)	0.22 (93%)
	ESD	0.16	0.14	0.57	0.12	0.25
50	mASD	0.28 (94%)	0.18 (96%)	0.79 (94%)	0.14 (90%)	0.31 (95%)
	ESD	0.29	0.19	0.80	0.12	0.31
30	mASD	0.36 (93%)	0.23 (95%)	0.98 (83%)	0.22 (90%)	0.39 (86%)
	ESD	0.35	0.22	1.05	0.17	0.35

Table 5: Evolution of the mean asymptotic standard deviation (mASD) and the empirical standard deviation (ESD) with the sample size for the Binomial-Poisson model estimated in Table 4. The percentage of ASD leading to a 95% confidence interval which contains the true parameter value are given between brackets. These results were computed over 100 runs of the algorithm.

variable that corresponds mainly to the production of honey during the 24 day period (in kg) and is denoted WG. The number of capped brood cells was measured in each hive at day 0, 12 and 24 (D0, D12, D24) and is used as a proxy for new bee recruitment (C0, C12, C24). Since the development from a capped cell to an emerging bee is 12 days for a working bee, the counts at a 12-day interval are considered non-overlapping, and thus the sum of C0, C12 and C24 is used to estimate the new bee recruitment over the whole period.

The number of ectoparasite mite *Varroa jacobsoni* was measured for each honey-bee colony on a sample of 20 g of adult bees (corresponding to approximately 150 bees) at D0 and D24 (V0, V24). The Varroa mite has an economic impact on the beekeeping industry and may be a contributing factor to colony collapse disorder (CCD). A recent study by Guzmán et al. (2010) shows that it is the main factor for collapsed colonies in Ontario, Canada.

Communications in Statistics Theory and Methods (2012), 28 p., DOI: 10.1080/03610926.2012.700372

We considered 3 variables from this dataset for each hive: the total number of capped brood cells (C=C0+C12+C24 divided by 100), the total number of parasites (V=V0+V24) and the weight gain (WG), which is a positive continuous variable.

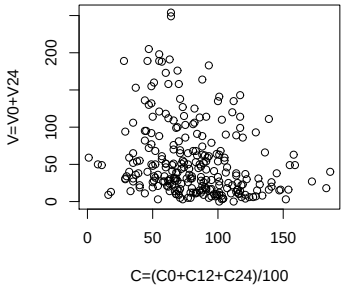

Figure 4: Number of capped cells (C) and number of Varroa mite (V) for 259 beehives (hives with missing data were excluded)

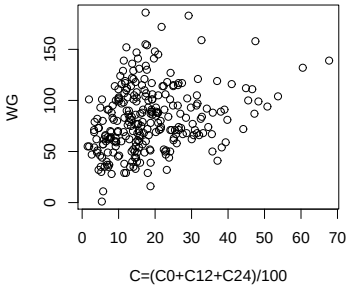

Figure 5: Number of capped cells (C) and weight gain in kg (WG) for 269 beehives (hives with missing data were excluded)

5.2 Results

The BPLN model was used to fit the bivariate distribution of capped brood cell number (C) and bee mite number (V) given in Figure 4. The Poisson-Normal model was used for the count variable C and the continuous variable WG (Figure 5). An exponential inverse link function was used for the continuous variable, in order to satisfy the condition WG > 0, so WG $= Y_2 = e^{\theta_2}$.

The estimation results for the two models are given in Table 6. As expected, the number of capped brood cells C is negatively correlated to the number of parasites V, whereas the hive weight gain is positively correlated to C. The estimates of the parameters μ_1 and σ_1^2, which correspond to the count variable C in both models, are very close, so the estimation procedure is stable. In the Poisson-Normal model the maximum likelihood estimators of μ_2 and σ_2^2 are obtained in 1 iteration and correspond to the empirical marginal mean and variance of the continuous variable log(WG) $= \theta_2$.

	C–V		C–WG	
Parameter	Estimate	ASD	Estimate	ASD
μ_1	4.32	0.03	4.33	0.02
σ_1^2	0.18	0.02	0.16	0.01
r	**-0.26**	0.06	**0.33**	0.06
μ_2	3.59	0.06	2.71	0.04
σ_2^2	0.95	0.09	0.43	0.06

Table 6: Estimation results and asymptotic standard deviation (ASD) of the BPLN model (variables C and V) and of the Poisson-Normal model (variables C and WG). Index 2 for the parameters refers respectively to the number of Varroa mites V (count variable) in the first case and to the weight gain WG (continuous variable) in the second case.

6 Discussion

In this article a general parametric model for multivariate data of various types is presented and a maximum likelihood estimation method, based on a variant of the Monte Carlo EM proposed by Booth and Hobert (1999), is provided. The hierarchical structure of the model and the estimation procedure are easy to adapt to different distributions and link functions which are used to obtain data of different types in our model. It also provides the asymptotic standard deviation of the estimators, which are a useful indication of the precision of the estimation for a small computational effort.

Limits of this method are first due to the model, some parameters cannot be estimated and there may be identifiability issues for some parts of the parameter domain, as shown in Section 4. Estimation issues can arise when the variance of the hidden model is too high.

Possible extensions of this model include spatial studies, GLMM models with multivariate random effects of different types (McCulloch and Searle, 2001), and mixture models. A spatial autocorrelation model could be introduced in the data, in a similar way to the work of Chagneau et al. (2010), but in a maximum likelihood framework without prior distributions on the parameters. This model could also be used in the context of GLMM models, the field in which this estimation procedure was proposed by Booth and Hobert (1999), when multivariate random effects can be defined (see for example Lai and Yau, 2008; Wang et al., 2006). Finally, the class of distributions

we defined in this paper could be used in a multivariate finite mixture model in a clustering context. It would allow clustering of all kinds of variables and even to data containing attributes of different types, which, to our knowledge and according to Fraley and Raftery (2002), has not been achieved yet. Moreover, our MCEM estimation procedure would be a direct extension of the EM procedure which is traditionally used to estimate multivariate mixture models.

Appendices

A Unconditional moments of the bivariate Binomial-Poisson model

Let us recall the bivariate Binomial-Poisson model given in Section 2.3.3:

$$\left\{\begin{array}{rcl} Y_1|\theta_1 & \sim & \mathscr{B}(n_b, \frac{1}{1+e^{-\theta_1}}) \\ Y_2|\theta_2 & \sim & \mathscr{P}(e^{\theta_2}) \\ (\theta_1,\theta_2)^T & \sim & \mathscr{N}_2(\boldsymbol{\mu},\boldsymbol{\Sigma}), \end{array}\right.$$

For sake of simplicity, we omit the index i in this appendix section and write θ_1 instead of θ_{i1}. Let $\boldsymbol{\Phi}$ denote the vector of unknown parameters $(\mu_1, \mu_2, \sigma_{11}, \sigma_{22}, \sigma_{12})$. To initialize our MCEM estimation procedure, an approximation of the parameter vector $\boldsymbol{\Phi}$ can be obtained by the method of moments.

The moment estimators of μ_2, σ_{22} are obtained directly by using the following equations (given in Section 2.3.1):

$$\mathbb{E}(Y_2) = e^{\mu_2+\frac{1}{2}\sigma_{22}} \stackrel{def}{=} m_2, \quad (12)$$

$$\mathbb{V}(Y_2) = m_2 + m_2^2\,(e^{\sigma_{22}}-1) \stackrel{def}{=} v_2. \quad (13)$$

Equations (12) and (13) yield:

$$\begin{aligned} \hat{\mu}_2 &= 2\log(m_2) - \frac{1}{2}\log(v_2 - m_2 + m_2^2), \\ \hat{\sigma}_{22} &= \log(v_2 - m_2 + m_2^2) - 2\log(m_2). \end{aligned}$$

To obtain moment estimators for $\mu_1, \sigma_{11}, \sigma_{12}$, we write the following statistics:

$$\begin{aligned}
\mathbb{E}(Y_1) &= n_b \mathbb{E}\Big(\frac{1}{1+e^{-\theta_1}}\Big) = n_b \int_{\mathbb{R}} \frac{1}{1+e^{-\theta_1}} \frac{1}{\sqrt{2\pi}}\ e^{-\frac{1}{2}\frac{(\theta_1-\mu_1)^2}{\sigma_{11}}} d\theta_1 = n_b F_1(\mathbf{\Phi}),\\
\mathbb{E}(Y_1^2) - \mathbb{E}(Y_1) &= n_b(n_b-1)\mathbb{E}\Big(\frac{1}{(1+e^{-\theta_1})^2}\Big) = n_b(n_b-1)F_2(\mathbf{\Phi}),\\
\mathbb{E}(Y_1Y_2) &= \mathbb{E}(\mathbb{E}(Y_1Y_2|\theta_1,\theta_2)) = \mathbb{E}(\frac{n_b}{1+e^{-\theta_1}}e^{\theta_2}) = n_b F_3(\mathbf{\Phi}),
\end{aligned}$$

where:

$$\begin{aligned}
F_1(\mathbf{\Phi}) &= \int_{\mathbb{R}} \frac{1}{1+e^{-\theta_1}} \frac{1}{\sqrt{2\pi}}\ e^{-\frac{1}{2}\frac{(\theta_1-\mu_1)^2}{\sigma_{11}}} d\theta_1\\
F_2(\mathbf{\Phi}) &= \int_{\mathbb{R}} \frac{1}{(1+e^{-\theta_1})^2} \frac{1}{\sqrt{2\pi}}\ e^{-\frac{1}{2}\frac{(\theta_1-\mu_1)^2}{\sigma_{11}}} d\theta_1\\
F_3(\mathbf{\Phi}) &= \iint_{\mathbb{R}^2} \frac{e^{\theta_2}}{1+e^{-\theta_1}} \frac{1}{2\pi|\Sigma|^{1/2}}\ e^{-\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\mu})\Sigma^{-1}(\boldsymbol{\theta}-\boldsymbol{\mu})} d\theta_1 d\theta_2.
\end{aligned}$$

With the change of variable $Z_1 = \frac{\theta_1-\mu_1}{\sqrt{\sigma_{11}}}$, we have $dZ_1 = \frac{d\theta_1}{\sqrt{\sigma_{11}}}$, $\theta_1 = \sqrt{\sigma_{11}}Z_1 + \mu_1$ and $F_1(\mathbf{\Phi})$ becomes:

$$F_1(\mathbf{\Phi}) = \int_{\mathbb{R}} \frac{1}{1+e^{-(\sqrt{\sigma_{11}}Z_1+\mu_1)}} \frac{1}{\sqrt{2\pi}}\ e^{-\frac{1}{2}Z_1^2} dZ_1$$

With the change of variable $\boldsymbol{Z}^T = (\boldsymbol{\theta}-\boldsymbol{\mu})^T|\mathbf{\Sigma}^{-1/2}|$, we have $\boldsymbol{\theta} = |\mathbf{\Sigma}^{1/2}|\boldsymbol{Z} + \boldsymbol{\mu}$. We have to find:

$$\mathbf{\Sigma}^{1/2} = \begin{pmatrix} a & b \\ b & c \end{pmatrix}.$$

such that $a > 0$, $c > 0$ and $ac - b^2 \geq 0$ ($\Sigma^{1/2}$ has to be positive definite). The solution is:

$$\begin{aligned}
a &= \frac{\sigma_{11}+\Delta}{H}\\
c &= \frac{\sigma_{22}+\Delta}{H}\\
b &= \frac{\sigma_{12}}{H}.
\end{aligned}$$

where $\Delta = \sqrt{\sigma_{11}\sigma_{22} - \sigma_{12}^2}$ and $H = \sqrt{\sigma_{11}+\sigma_{22}+2\Delta}$. Thus we have:

$$\theta_1 = aZ_1 + bZ_2 + \mu_1$$

$$\theta_2 = bZ_1 + cZ_2 + \mu_2$$

and $F_3(\mathbf{\Phi})$ becomes:

$$F_3(\mathbf{\Phi}) = \iint_{\mathbb{R}^2} \frac{e^{bZ_1+cZ_2+\mu_2}}{1+e^{-(aZ_1+bZ_2+\mu_1)}} \frac{1}{2\pi}\ e^{-\frac{1}{2}\boldsymbol{Z}^T\boldsymbol{Z}} d\theta_1 d\theta_2.$$

The Newton-Raphson iterative procedure was used to obtain moment estimators for $\mu_1, \sigma_{11}, \sigma_{12}$.

We need to compute the derivatives:

$$\begin{aligned}
\frac{\partial F_1}{\partial \mu_1} &= \int_{\mathbb{R}} \frac{e^{-(\sqrt{\sigma_{11}}Z_1+\mu_1)}}{(1+e^{-(\sqrt{\sigma_{11}}Z_1+\mu_1)})^2}\,\Phi_0(dZ_1)\\
\frac{\partial F_1}{\partial \sigma_{12}} &= 0\\
\frac{\partial F_1}{\partial \sigma_{11}} &= \int_{\mathbb{R}} \frac{Z_1}{2\sqrt{\sigma_{11}}}\,\frac{e^{-(\sqrt{\sigma_{11}}Z_1+\mu_1)}}{(1+e^{-(\sqrt{\sigma_{11}}Z_1+\mu_1)})^2}\,\Phi_0(dZ_1)\\
\frac{\partial F_2}{\partial \mu_1} &= \int_{\mathbb{R}} \frac{2e^{-(\sqrt{\sigma_{11}}Z_1+\mu_1)}}{(1+e^{-(\sqrt{\sigma_{11}}Z_1+\mu_1)})^3}\,\Phi_0(dZ_1)\\
\frac{\partial F_2}{\partial \sigma_{12}} &= 0\\
\frac{\partial F_2}{\partial \sigma_{11}} &= \int_{\mathbb{R}} -\frac{Z_1}{\sqrt{\sigma_{11}}}\,\frac{e^{-(\sqrt{\sigma_{11}}Z_1+\mu_1)}}{(1+e^{-(\sqrt{\sigma_{11}}Z_1+\mu_1)})^3}\,\Phi_0(dZ_1)\\
\frac{\partial F_3}{\partial \mu_1} &= \iint_{\mathbb{R}^2} \frac{R_1R_2}{(1+R_1)^2}\,\Phi_0(dZ_1,dZ_2)\\
\frac{\partial F_3}{\partial \sigma_{12}} &= \iint_{\mathbb{R}^2} \frac{(b_{12}Z_1+c_{12}Z_2)R_1(1+R_2)+R_1(a_{12}Z_1+b_{12}Z_2)R_2}{(1+R_2)^2}\\
\frac{\partial F_3}{\partial \sigma_{11}} &= \iint_{\mathbb{R}^2} \frac{(b_{11}Z_1+c_{11}Z_2)R_1(1+R_2)+R_1(a_{11}Z_1+b_{11}Z_2)R_2}{(1+R_2)^2},
\end{aligned}$$

with the notations:

$$\begin{aligned}
R_1 &= e^{-(aZ_1+bZ_2+\mu_1)}\\
R_2 &= e^{(bZ_1+cZ_2+\mu_2)}\\
a_{11} &= \frac{\partial a}{\partial \sigma_{11}} = \frac{(\sigma_{22}+2\Delta)H^2-(\sigma_{11}+\Delta)(\sigma_{22}+\Delta)}{2\Delta H^3}\\
a_{12} &= \frac{\partial a}{\partial \sigma_{12}} = \frac{\sigma_{12}(\sigma_{11}-H^2+\Delta)}{\Delta H^3}\\
b_{11} &= \frac{\partial b}{\partial \sigma_{11}} = -\frac{\sigma_{12}(\sigma_{22}+\Delta)}{2\Delta H^3}\\
b_{12} &= \frac{\partial b}{\partial \sigma_{12}} = \frac{\sigma_{12}^2+\Delta H^2}{\Delta H^3}\\
c_{11} &= \frac{\partial c}{\partial \sigma_{11}} = \frac{\sigma_{22}H^2-(\sigma_{22}+\Delta)^2}{2\Delta H^3}\\
c_{12} &= \frac{\partial c}{\partial \sigma_{12}} = \frac{\sigma_{12}(\sigma_{22}-H^2+\Delta)}{\Delta H^3}.
\end{aligned}$$

Distribution	$p = g^{-1}(\theta)$	$\lambda = \mathbb{E}(Y_j\|\theta)$	$g(\lambda)$	$\partial g(\lambda)/\partial\lambda$	$v = \mathbb{V}(Y_j\|\theta)$
Poisson(p)	e^θ	$p = e^\theta$	$\log(\lambda)$	$e^{-\theta}$	e^θ
Binomial(n_b,p)	$\dfrac{1}{1+e^{-\theta}}$	$n_b p = \dfrac{n_b}{1+e^{-\theta}}$	$\log\dfrac{\lambda}{1-\lambda}$	$\dfrac{(1+e^{-\theta})^2}{n_b e^{-\theta}}$	$n_b p(1-p) = \dfrac{n_b e^{-\theta}}{(1+e^{-\theta})^2}$

Table 7: Expressions of λ, v and $g_\lambda(\lambda)$ for the Poisson distribution and the Binomial distribution with fixed number of trials n_b.

B PQL estimators of the conditional moments for different distributions and link functions

The parameters $\boldsymbol{m}_t$ and $\boldsymbol{\Sigma}_t$ of the multivariate Student t distribution used in the E-step of the MCEM to obtain an importance sample, are re-evaluated at each step in order to be approximately:

$$\boldsymbol{m}_t = \mathbb{E}_{\boldsymbol{\Phi}^{(t)}}[\boldsymbol{\theta}|\boldsymbol{y}],$$
$$\boldsymbol{\Sigma}_t = \mathbb{V}_{\boldsymbol{\Phi}^{(t)}}[\boldsymbol{\theta}|\boldsymbol{y}].$$

The Penalized Quasi Likelihood (PQL) estimators of these conditional moments are obtained by using a Laplace approximation of the likelihood and a Fisher scoring maximization procedure.

At iteration $(t+1)$ of the MCEM algorithm, the Fisher scoring iterative algorithm is used, given by:

$$\boldsymbol{m}_t^{(k+1)} = \boldsymbol{m}_t^{(k)} + (\boldsymbol{W}(\boldsymbol{m}_t^{(k)}) + \boldsymbol{\Sigma}^{-1})^{-1}\Big(\boldsymbol{W}(\boldsymbol{m}_t^{(k)})\boldsymbol{\Delta}(\boldsymbol{m}_t^{(k)})(\boldsymbol{y} - \boldsymbol{\lambda}(\boldsymbol{m}_t^{(k)})) - \boldsymbol{\Sigma}^{-1}(\boldsymbol{m}_t^{(k)} - \boldsymbol{\mu})\Big),$$
$$\boldsymbol{\Sigma}_t^{(k+1)} = \Big(\boldsymbol{W}(\boldsymbol{m}_t^{(k+1)}) + \boldsymbol{\Sigma}^{-1}\Big)^{-1},$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are the current estimators of the mean and variance of $\boldsymbol{\theta}$ (from the MCEM iteration t), and the matrices $\boldsymbol{\lambda}(\boldsymbol{m}_t^{(k)})$, $\boldsymbol{W}(\boldsymbol{m}_t^{(k)})$ and $\boldsymbol{\Delta}(\boldsymbol{m}_t^{(k)})$ are defined (for a single-parameter exponential distribution) by:

$$\lambda_i = \frac{\partial b(\eta)}{\partial \eta} = \mathbb{E}(Y_j|\theta),$$
$$v_i = \frac{\partial^2 b(\eta)}{\partial \eta^2} = \mathbb{V}(Y_j|\theta),$$
$$g_\lambda(\lambda_i) = \frac{\partial g(\lambda_i)}{\partial \lambda_i},$$
$$\boldsymbol{W}(\boldsymbol{\lambda}) = \mathrm{diag}\Big(\frac{1}{v_i g_\lambda^2(\lambda_i)}\Big),$$
$$\boldsymbol{\Delta}(\boldsymbol{\lambda}) = \mathrm{diag}(g_\lambda(\lambda_i)),$$

where $\mathrm{diag}(x_i)$ denotes a diagonal matrix with diagonal element x_i (at row and column i) and with the notations of section 2.2. The PQL estimators of $\boldsymbol{m}_t$ and $\boldsymbol{\Sigma}_t$ should thus be adapted to the model by calculating these expressions for the chosen distributions $\mathscr{L}$ and link functions g (cf table 7).

C Maximum likelihood estimators of the multivariate normal parameters in the M-step of the MCEM

The maximum likelihood estimator of $\boldsymbol{\Sigma}$ is obtained by deriving Q_N with respect to $\boldsymbol{\Sigma}$ and solving the equation $\partial Q_N/\partial\boldsymbol{\Sigma} = 0$. This is equivalent to solving $\partial Q_N/\partial\boldsymbol{\Gamma} = 0$ where $\boldsymbol{\Gamma} = \boldsymbol{\Sigma}^{-1}$. The result, provided in equation (10), is recalled here:

$$\hat{\boldsymbol{\Sigma}} = \frac{\sum_{i=1}^{n}\sum_{k=1}^{N} w_{ik}(\boldsymbol{\theta}_{ik} - \hat{\boldsymbol{\mu}})(\boldsymbol{\theta}_{ik} - \hat{\boldsymbol{\mu}})^T}{\sum_{i=1}^{n}\sum_{k=1}^{N} w_{ik}}$$

Proof of (10) in the bivariate case:

$$\begin{aligned}
Q_N(\boldsymbol{\Phi}, \boldsymbol{\Phi}^{(t)}) &= \sum_{i=1}^{n} \mathbb{E}_{\boldsymbol{\Phi}^{(t)}}[\log(f_{\boldsymbol{\theta}}(\boldsymbol{\theta}_i; \boldsymbol{\mu}, \boldsymbol{\Sigma}))|\boldsymbol{Y}_i = \boldsymbol{y}_i] \\
&= \frac{1}{N}\sum_{i=1}^{n}\sum_{k=1}^{N} w_{ik}\Big(\log(f_{\boldsymbol{\theta}}(\boldsymbol{\theta}_{ik}|\boldsymbol{\mu}, \boldsymbol{\Sigma}))\Big) \\
&= c + \frac{1}{2N}\sum_{i=1}^{n}\sum_{k=1}^{N} w_{ik}\Big(-\log|\boldsymbol{\Sigma}| - (\boldsymbol{\theta}_i - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}_i - \boldsymbol{\mu})\Big) \\
&= c + u(\boldsymbol{\Sigma})
\end{aligned}$$

where $c = -\frac{1}{2N}\sum_{i=1}^{n}\sum_{k=1}^{N} w_{ik}\log(2\pi)$ is a constant.

Let $\boldsymbol{\Gamma} = \boldsymbol{\Sigma}^{-1}$. Then $|\boldsymbol{\Sigma}| = |\boldsymbol{\Gamma}|^{-1}$ and $u(\boldsymbol{\Sigma})$ is replaced by:

$$v(\boldsymbol{\Gamma}) = \frac{1}{2N}\sum_{i=1}^{n}\sum_{k=1}^{N} w_{ik}\Big(\log|\boldsymbol{\Gamma}| - (\boldsymbol{\theta}_i - \boldsymbol{\mu})^T\boldsymbol{\Gamma}(\boldsymbol{\theta}_i - \boldsymbol{\mu})\Big).$$

Solving $\partial u(\boldsymbol{\Sigma})/\partial\boldsymbol{\Sigma} = 0$ is equivalent to solving $\partial v(\boldsymbol{\Gamma})/\partial\boldsymbol{\Gamma} = 0$. We denote $\boldsymbol{\Gamma} = (\gamma_{ll'})$ where

$l, l' \in \{1, 2\}$, $\boldsymbol{\mu} = (\mu_1, \mu_2)$ and $\boldsymbol{\theta} = (\theta_1, \theta_2)$. We have:

$$\frac{\partial v(\boldsymbol{\Gamma})}{\partial \gamma_{11}} = \frac{1}{2N} \sum_{i=1}^{n} \sum_{k=1}^{N} w_{ik} \Big(\frac{\gamma_{22}}{|\boldsymbol{\Gamma}|} - (\theta_1 - \mu_1)^2 \Big)$$
$$\frac{\partial v(\boldsymbol{\Gamma})}{\partial \gamma_{22}} = \frac{1}{2N} \sum_{i=1}^{n} \sum_{k=1}^{N} w_{ik} \Big(\frac{\gamma_{11}}{|\boldsymbol{\Gamma}|} - (\theta_2 - \mu_2)^2 \Big)$$
$$\frac{\partial v(\boldsymbol{\Gamma})}{\partial \gamma_{12}} = \frac{1}{2N} \sum_{i=1}^{n} \sum_{k=1}^{N} w_{ik} \Big(- \frac{\gamma_{12}}{|\boldsymbol{\Gamma}|} - 2(\theta_1 - \mu_1)(\theta_2 - \mu_2) \Big)$$

and $\gamma_{12} = \gamma_{21}$. $\partial v(\boldsymbol{\Gamma}) / \partial \boldsymbol{\Gamma} = 0$ is equivalent to:

$$\frac{\gamma_{11}}{|\boldsymbol{\Gamma}|} = \frac{\sum_{i=1}^{n} \sum_{k=1}^{N} w_{ik} (\theta_2 - \mu_2)^2}{\sum_{i=1}^{n} \sum_{k=1}^{N} w_{ik}}$$
$$\frac{\gamma_{22}}{|\boldsymbol{\Gamma}|} = \frac{\sum_{i=1}^{n} \sum_{k=1}^{N} w_{ik} (\theta_1 - \mu_1)^2}{\sum_{i=1}^{n} \sum_{k=1}^{N} w_{ik}}$$
$$\frac{\gamma_{12}}{|\boldsymbol{\Gamma}|} = \frac{-2 \sum_{i=1}^{n} \sum_{k=1}^{N} w_{ik} (\theta_1 - \mu_1)(\theta_2 - \mu_2)}{\sum_{i=1}^{n} \sum_{k=1}^{N} w_{ik}}.$$

Since:

$$\boldsymbol{\Sigma} = \boldsymbol{\Gamma}^{-1} = \frac{1}{|\boldsymbol{\Gamma}|} \begin{pmatrix} \gamma_{22} & -\gamma_{12} \\ -\gamma_{12} & \gamma_{11} \end{pmatrix}$$

we have:

$$\sigma_{11} = \frac{\gamma_{22}}{|\boldsymbol{\Gamma}|}$$
$$\sigma_{22} = \frac{\gamma_{11}}{|\boldsymbol{\Gamma}|}$$
$$\sigma_{12} = -\frac{\gamma_{12}}{|\boldsymbol{\Gamma}|}.$$

It follows that:

$$\hat{\boldsymbol{\Sigma}} = \frac{\sum_{i=1}^{n} \sum_{k=1}^{N} w_{ik} (\boldsymbol{\theta} - \hat{\boldsymbol{\mu}})(\boldsymbol{\theta} - \hat{\boldsymbol{\mu}})^T}{\sum_{i=1}^{n} \sum_{k=1}^{N} w_{ik}}.$$

Acknowledgement

This work was supported by the "Institut National de la Recherche Agronomique" (INRA) and the French region Provence Alpes Côte d'Azur. The beehive dataset was provided by an INRA - ADAPI ("Association pour le développement de l'apiculture provençale") project, funded by the "Fonds européen agricole de garantie" (FEAGA).

References

Aitchison, J., Ho, C.H.: The multivariate Poisson log-normal distribution. Biometrika **76**, 643–653 (1989)

Booth, J.G., Hobert, J.P.: Standard errors of prediction in generalized linear mixed models. J. Am. Stat. Assoc. **93**, 262–272 (1998)

Booth, J.G.,Hobert, J.P.: Maximizing Generalized Linear Mixed Model likelihoods with an automated Monte Carlo EM algorithm. J. Roy. Stat. Soc. B **61**, 265–285 (1999)

Breslow, N.E., Clayton, D.G.: Approximate inference in generalized linear mixed models. J. Am. Stat. Assoc. **88**, 9-25 (1993)

Chagneau, P., Mortier, F., Picard, N., Bacro, J.-N.: A hierarchical bayesian model for spatial prediction of multivariate non-Gaussian random fields. Biometrics (2010) doi:10.1111/j.1541-0420.2010.01415.x

Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood for incomplete data via the EM algorithm. J. Roy. Stat. Soc. B **39**, 1–38 (1977)

Evans, M., Swartz, T.B.: Bayesian integration using multivariate Student importance sampling. Comp. Sci. Stat. **27**, 456-461 (1996)

Fraley, C., Raftery, A.E.: Model-Based Clustering, Discriminant Analysis, and Density Estimation. J. Am. Stat. Assoc. **97**, 611–631 (2002)

Guzmán-Novoa, E., Eccles, L., Calvete, Y., Mcgowan, J., Kelly, P.G., Correa-Benítez A.: Varroa destructor is the main culprit for the death and reduced populations of overwintered honey bee (Apis mellifera) colonies in Ontario, Canada. Apidologie (2010).

Lai, X., Yau, K.K.W.: Long-term survivor model with bivariate random effects: Applications to bone marrow transplant and carcinoma study data. Stat. Med. **27**, 5692–5708 (2008)

McCulloch, C.E., Searle, S.R.: General, Linear and Mixed Models. Wiley, New York (2001)

McLachlan, G.J., and Krishnan, T.: The EM Algorithm and Extensions. Second Edition. Wiley, Hoboken (2008)

Tanner, A.M.: Tools for Statistical Inference: Observed Data and Data Augmentation Methods. Springer, New York (1991)

Tunaru, R.: Hierarchical bayesian models for multiple count data. Austrian Journal of Statistics. **31**, 221–229 (2002)

Wang, K.,Yau, K.K.W., Lee, A.H., McLachlan, G.J.: Two-component Poisson mixture regression modelling of count data with bivariate random effects. Math. Comput. Model. **46**, 1468–1476 (2007)

Wei, G.C.G.,Tanner, M.A.: A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. J. Am. Stat. Assoc. **85**, 699–704 (1990)