Statistics and Computing **15**: 255–265, 2005
© 2005 Springer Science + Business Media, Inc. *Manufactured in The Netherlands.*

Multivariate Poisson regression with covariance structure

DIMITRIS KARLIS* and LOUKIA MELIGKOTSIDOU†

**Department of Statistics, Athens University of Economics and Business, 76, Patission Str., 10434, Athens, Greece*
karlis@aueb.gr
†*Department of Mathematics and Statistics, Lancaster University, Lancaster, LA1 4YF, United Kingdom*
l.meligotsidou@lancaster.ac.uk

Received June 2003 and accepted March 2005

In recent years the applications of multivariate Poisson models have increased, mainly because of the gradual increase in computer performance. The multivariate Poisson model used in practice is based on a common covariance term for all the pairs of variables. This is rather restrictive and does not allow for modelling the covariance structure of the data in a flexible way. In this paper we propose inference for a multivariate Poisson model with larger structure, i.e. different covariance for each pair of variables. Maximum likelihood estimation, as well as Bayesian estimation methods are proposed. Both are based on a data augmentation scheme that reflects the multivariate reduction derivation of the joint probability function. In order to enlarge the applicability of the model we allow for covariates in the specification of both the mean and the covariance parameters. Extension to models with complete structure with many multi-way covariance terms is discussed. The method is demonstrated by analyzing a real life data set.

Keywords: data augmentation, EM algorithm, Markov chain Monte Carlo, multivariate reduction, crime data

1. Introduction

Multivariate data analysis has a long history for continuous data based on the multivariate normal and related distributions. However, this is not true for discrete data. When treating multivariate count data, approximations by continuous multivariate normal models can be used, but they can be misleading especially when the observed means are not large and there are several zero counts. The multivariate Poisson distribution, while the most important among discrete multivariate distributions (see, e.g., Johnson, Kotz and Balakrishnan 1997), has several shortcomings for its application. The main drawback of the application of the multivariate Poisson distribution is the complicated form of the joint probability function.

Inferential procedures for a special case of the multivariate Poisson model, with a single common covariance term, are described in the recent papers of Tsionas (1999, 2001) and Karlis (2003). This model is rather restrictive for real applications since it assumes that all the pairs of variables have the same covariance.

In this paper we extend the above model by allowing for larger covariance structure between the variables. Namely we construct a model that allows for a different covariance for each pair of variables. The model is initially presented in its general form, with full (but perhaps unnecessarily large) structure. Then, we focus on a useful reduced model where only two-way covariance terms are used. We also include covariates in the model. Inference from the classical point of view through maximum likelihood (ML) estimation is proposed via an EM algorithm, while Bayesian inference is proposed via an MCMC algorithm. Both approaches make use of a data augmentation scheme based on multivariate reduction.

Potential application of the model can be made in a variety of disciplines where count data occur quite often, like epidemiology (e.g., incidences of different types of illness), marketing (purchases of different products), industrial control (different types of faults) etc. In all these circumstances traditional analysis using a multivariate normal approximation can be misleading due to the nature of the data (small marginal means with a lot of zero counts).

0960-3174 © 2005 Springer Science + Business Media, Inc.

The advantages of our model are the following. Firstly, it generalizes the univariate Poisson model and therefore it is a standard reference model for multivariate count data. Secondly, it allows for realistic covariance structure among the variables. Furthermore, one can introduce covariates in the covariance terms in order to model the covariances in a flexible way (this is not true for other competing models). Finally, the proposed model is less complicated, and therefore less computationally demanding, than several competing models. Limitations of the model are the facts that it does not allow for overdispersion and negative correlation as, for example, the model of Chib and Winkelmann (2001). There are also some other models that allow for negative correlation (see van Ophem 1999, Berkhout and Plug 2004), which, however, are much more complicated and require special efforts for parameter estimation.

The remainder of the paper proceeds as follows. The multivariate Poisson distribution in its general form is described in Section 2. In Section 3 we describe our proposed models. ML estimation for these models is considered in Section 4, while Section 5 describes the Bayesian approach. Section 6 presents an application to real data, while Section 7 contains some concluding remarks.

2. The multivariate Poisson distribution

2.1. *General definition*

The derivation of the multivariate Poisson distribution is based on a general multivariate reduction scheme. Assuming Y_r, $r = 1, \ldots, k$, are independent univariate Poisson random variables, i.e. $Y_r \sim Po(\theta_r)$, $r = 1, \ldots, k$, then the general definition of the multivariate Poisson distribution is made through the vector $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_k)'$ and an $m \times k$ matrix $\boldsymbol{A}$, $m \leq k$, with 0 and 1 elements. Specifically, the vector $\boldsymbol{X} = (X_1, X_2, \ldots, X_m)'$ defined as $\boldsymbol{X} = \boldsymbol{AY}$ follows a multivariate Poisson distribution.

An alternative expression for the multivariate Poisson random vector $\boldsymbol{X}$ arises if we consider each column of $\boldsymbol{A}$ as a vector ϕ_r, $r = 1, \ldots, k$. Then, $\boldsymbol{A} = [\phi_1 \ \phi_2 \ldots \phi_k]$ and hence $\boldsymbol{X} = \boldsymbol{AY} = \sum_{r=1}^{k} \phi_r Y_r$. In this framework, the variability of the random vector $\boldsymbol{X}$, which has the m-variate Poisson distribution, is explained through the variability of k independent univariate Poisson random variables. Note that the elements of $\boldsymbol{X}$ are dependent as indicated by the structure of the matrix $\boldsymbol{A}$.

The most general form of the multivariate Poisson distribution arises if the matrix $\boldsymbol{A}$ has the form $\boldsymbol{A} = [\boldsymbol{A}_1, \boldsymbol{A}_2, \ldots \boldsymbol{A}_m]$, where $\boldsymbol{A}_j$, $j = 1, \ldots, m$ is a sub-matrix of dimensions $m \times C_j^m$, where C_j^m is the number of combinations of picking j from m numbers, each column of $\boldsymbol{A}_j$ has exactly j ones and $(m-j)$ zeroes and no duplicate columns exist. Thus, $\boldsymbol{A}_m$ is the column vector of 1s, while $\boldsymbol{A}_1$ is the identity matrix of size $m \times m$. Then, the vector $\boldsymbol{Y}$ can also be written in the form $\boldsymbol{Y} = (\boldsymbol{Y}_1', \boldsymbol{Y}_2', \ldots, \boldsymbol{Y}_m')'$, where $\boldsymbol{Y}_j$ is a sub-vector of dimension C_j^m, $j = 1, \ldots, m$. Hence, the definition of the vector $\boldsymbol{X}$ becomes $\boldsymbol{X} = \sum_{r=1}^{k} \phi_r Y_r = \sum_{j=1}^{m} \boldsymbol{A}_j \boldsymbol{Y}_j$, which means that $\boldsymbol{X}$ is expressed as a sum of different vector-terms, which explain the variability of its dependent elements. These terms can be interpreted as main effects and two-way up to m-way covariance effects in an ANOVA like fashion.

Under the multivariate Poisson model the mean and the variance-covariance matrix of $\boldsymbol{X}$ are given by

$$E(\boldsymbol{X}) = \boldsymbol{A}\theta$$

and

$$\text{Var}\,(\boldsymbol{X}) = \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T$$

where $\boldsymbol{\Sigma} = \text{diag}\,(\theta_1, \theta_2, \ldots, \theta_k)$ is the variance-covariance matrix of $\boldsymbol{Y}$ (Σ is diagonal because of the independence of Y_i's). Each element of $\boldsymbol{X}$ marginally follows a univariate Poisson distribution. This general model has been theoretically described in Mahamunulu (1967) and Johnson, Kotz and Balakrishnan (1997) among others.

2.2. *The probability distribution*

A main problem, which limits the usage of multivariate distributions in general, is the complexity of calculating the probability distribution function. In the case of the multivariate Poisson distribution, the calculation of the probability mass function can be of great difficulty, as it often demands summations over high-dimensional spaces. If we consider an m-variate Poisson model, the calculation of the probability mass function requires $2^m - m - 1$ summations. For example, the probability function of the 3-variate Poisson distribution can be written with 4 nested sums (see also Mahamunulu 1967 for a symbolic presentation of the problem).

Formally, the definition of the multivariate Poisson distribution was made through a mapping $g : N^k \to N^m$, $k \geq m$, such that $\boldsymbol{X} = g(\boldsymbol{Y}) = \boldsymbol{AY}$. Hence, the joint probability of the m-vector $\boldsymbol{x} = (x_1, x_2, \ldots, x_m)'$ is given by the sum of the joint probabilities of all k-vectors $\boldsymbol{y} = (y_1, y_2, \ldots, y_k)'$ such that $g(\boldsymbol{y}) = \boldsymbol{x}$. If $\boldsymbol{x} \in N^m$, let the set $g^{-1}(\boldsymbol{x}) \subset N^k$ denote the inverse image of $\boldsymbol{x}$ under g. The probability mass function of $\boldsymbol{X}$ is then defined as

$$P(\boldsymbol{X} = \boldsymbol{x}) = \sum_{\boldsymbol{y} \in g^{-1}(\boldsymbol{x})} P(\boldsymbol{Y} = \boldsymbol{y})$$

Since the elements of $\boldsymbol{Y}$ follow independently univariate Poisson distributions, we obtain that

$$P(\boldsymbol{X} = \boldsymbol{x}) = \sum_{\boldsymbol{y} \in g^{-1}(\boldsymbol{x})} \prod_{i=1}^{k} P(y_i; \theta_i). \tag{1}$$

It is clear that the calculation of these probabilities can be computationally expensive, since the summations needed might be exhausting in some cases, especially when the number of dimensions is large. However, computation of the probabilities can be accomplished via recursive schemes. Kano and Kawamura (1991) provided a general scheme for constructing recurrence relations for multivariate Poisson distributions. For computational details on the construction of a recursive scheme for the