

On modelling multivariate count data

Dimitris Karlis,
Department of Statistics, Athens University of Economics

NIPS Whistler, December 2008

Outline

- Introduction
- Multivariate Poisson models
- Models based on mixing
- Models based on copulas
- Time Series models for discrete data
- Further research - Comments

Motivation

Multivariate data are usually modelled via

- Multivariate Normal models
- Multinomial models (for categorical data)

What about multivariate count data?

- Small counts with a lot of zeros
- Normal approximation may not be adequate at all

Idea: Use multivariate extensions of simple Poisson models

Multivariate Count data

- Incidences of different diseases across time or space
- Different type of crimes in different areas
- Purchases of different products
- Accidents (different types or in different time periods)
- Football data
- Different types of faults in production systems
- Number of faults in parts of a large system etc
- number of spikes in certain neurons in a given time frame

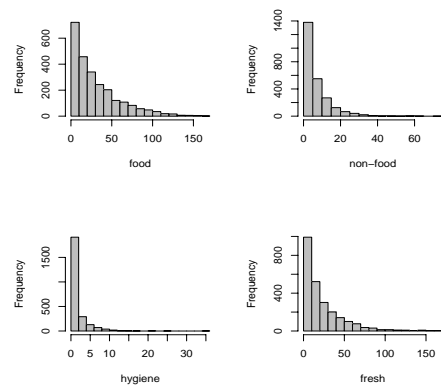
Multivariate Count data

Other kinds of dependent count data

- Time series data with discrete outcomes
- Repeated measurements with discrete data for patients.
- Count Data spatially correlated, like diseases in an area (map)

While such procedures for continuous data are well established, working with count data is less known in the literature.

Example of data: supermarket data



To start with ...

For continuous data the normal and the multivariate normal distributions play the most significant role. Most models can be derived from them or they are based on them.

To start with we need to consider the discrete counterpart of the normal distribution. The Poisson distribution is by far the most prominent distribution for counts.

A lot of research is based on this starting assumption, namely considering Poisson and multivariate Poisson counterparts!

So...

- Define multivariate Poisson distributions
- Make them flexible (allow flexible covariance structure, allow covariate information etc)
- Consider extensions by standard procedures like mixing and define alternative models that in some sense expand the multivariate Poisson model (e.g. multivariate negative binomial model)

Bivariate Poisson model

Let $X_i \sim \text{Poisson}(\theta_i)$, $i = 0, 1, 2$

Consider the random variables

$$X = X_1 + X_0$$

$$Y = X_2 + X_0$$

$(X, Y) \sim \text{BP}(\theta_1, \theta_2, \theta_0)$,

Joint probability function given:

$$P(X = x, Y = y) = e^{-(\theta_1 + \theta_2 + \theta_0)} \frac{\theta_1^x}{x!} \frac{\theta_2^y}{y!} \sum_{i=0}^{\min(x,y)} \binom{x}{i} \binom{y}{i} i! \left(\frac{\theta_0}{\theta_1 \theta_2} \right)$$

Properties of Bivariate Poisson model

- Marginal distributions are Poisson, i.e.

$$X \sim \text{Poisson}(\theta_1 + \theta_0)$$

$$Y \sim \text{Poisson}(\theta_2 + \theta_0)$$

- Conditional Distributions : Convolution of a Poisson with a Binomial
- Covariance: $\text{Cov}(X, Y) = \theta_0 \geq 0$
For a full account see Kocherlakota and Kocherlakota (1992) and Johnson, Kotz and Balakrishnan (1997)

Bivariate Poisson model (more)

Computational problems: evaluation of the probability function implies calculating a sum continuously which can be quite slow!

Recursive relationships:

$$\begin{aligned} xP(x, y) &= \theta_1 P(x-1, y) + \theta_0 P(x-1, y-1) \\ yP(x, y) &= \theta_2 P(x, y-1) + \theta_0 P(x-1, y-1). \end{aligned} \quad (1)$$

with the convention that $P(x, y) = 0$, if $\min(x, y) < 0$.

Need for "clever" use of these relationships (See, e.g. Tsiamirtzis and Karlis, 2002).

Bivariate Poisson model (estimation)

Various techniques:

- Moment method, Maximum likelihood, Even points etc (see, Kocherlakota and Kocherlakota, 1992).
- Recently: Bayesian estimation (Tsiomas, 1999, Karlis and Tsiamirtzis, 2008).

Bivariate Poisson regression model

$$(X_i, Y_i) \sim BP(\theta_{1i}, \theta_{2i}, \theta_{0i})$$

$$\log(\theta_{ji}) = \mathbf{x}_i' \boldsymbol{\beta}_j, \quad j = 0, 1, 2$$

- Allows for covariate-dependent covariance.
- Separate modelling of means and covariance
- Standard estimation methods not easy to apply.
- Computationally demanding.
- Application of an easily programmable EM algorithm

Multivariate Poisson model : A simple derivation

Extend the derivation to more dimensions. Consider the independent random variables $X_i \sim \text{Poisson}(\theta_i)$, $i = 0, 1, \dots, m$.

Then define

$$Y_1 = X_1 + X_0$$

$$Y_2 = X_2 + X_0$$

$$\dots$$

$$Y_m = X_m + X_0$$

Then the vector (Y_1, \dots, Y_m) follows a m -variate Poisson distribution. All the marginals are Poisson and θ_0 is the covariance for all the pairs of the random variables!

The common term X_0 has introduced this kind of dependence between the variables

Properties

- The joint probability function is given by

$$P(\mathbf{X}) = P(X_1 = x_1, X_2 = x_2, \dots, X_m = x_m)$$

$$= \exp\left(-\sum_{i=1}^m \theta_i\right) \prod_{i=1}^m \frac{\theta_i^{x_i}}{x_i!} \sum_{i=0}^s \left[\prod_{j=1}^m \binom{x_j}{i} \right] i! \left(\frac{\theta_0}{\prod_{i=1}^m \theta_i} \right)^i.$$

where $s = \min(x_1, x_2, \dots, x_m)$.

- Marginally each X_i follows a Poisson distribution with parameter $\theta_0 + \theta_i$.
- Parameter θ_0 is the positive covariance between all the pairs of random variables.
- If $\theta_0 = 0$ then the variables are independent.

Problems

- Probability function too complicated for
 - even the calculation of the function (remedy: use of recurrence relationships)
 - for estimation purposes (remedy: use of an EM algorithm)
- Assumes common covariance for all pairs - unrealistic. We need to extend!

Multivariate Poisson model

Let $\mathbf{X} = (X_1, X_2, \dots, X_m)$ and $X_i \sim \text{Poisson}(\theta_i)$, $i = 1, \dots, m$. Then the general definition of multivariate Poisson models is made through the matrix \mathbf{A} of dimensions $k \times m$, where the elements of the matrix are zero and ones and no duplicate columns exist.

Then the vector $\mathbf{Y} = (Y_1, Y_2, \dots, Y_k)$ defined as

$$\mathbf{Y} = \mathbf{A}\mathbf{X}$$

follows a multivariate Poisson distribution.

Complete Specification

$$\mathbf{A} = [A_1 \ A_2 \ \dots \ A_k]$$

where A_i is a matrix of dimensions $k \times \binom{k}{i}$ where each column has exactly i ones and $k - i$ zeroes.

Example $k = 3$

$$A_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad A_2 = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix} \quad A_3 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

and then

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 & 1 \end{bmatrix}$$

Complete Specification (cont.)

This corresponds to

$$\begin{aligned} Y_1 &= X_1 + X_{12} + X_{13} + X_{123} \\ Y_2 &= X_2 + X_{12} + X_{23} + X_{123} \\ Y_3 &= X_3 + X_{13} + X_{23} + X_{123} \end{aligned} \quad (2)$$

where all X_i 's, are independently Poisson distributed random variables with parameter θ_i , $i \in (\{1\}, \{2\}, \{3\}, \{12\}, \{13\}, \{23\}, \{123\})$

Note: Parameters θ_{ij} are in fact covariance parameters between Y_i and Y_j . Similarly θ_{123} is a common 3-way covariance parameter.

Other cases

Independent Poisson variables

Corresponds to the case $A = A_1$.

Example for $k = 3$

$$A = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

i.e. product of independent Poisson probability functions.

Full covariance structure

If we want to specify only up to 2-way covariances we take the form

$$\mathbf{A} = [A_1 \ A_2]$$

Example $k = 3$

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 \end{bmatrix}$$

This model is very interesting as it assumes different covariances between all the pairs and thus it resembles the multivariate normal model.

Full covariance structure(cont.)

This corresponds to

$$\begin{aligned} Y_1 &= X_1 + X_{12} + X_{13} \\ Y_2 &= X_2 + X_{12} + X_{23} \\ Y_3 &= X_3 + X_{13} + X_{23} \end{aligned}$$

where all X_i 's, are independently Poisson distributed random variables with parameter θ_i , $i \in (\{1\}, \{2\}, \{3\}, \{12\}, \{13\}, \{23\})$ the covariance matrix of $\mathbf{Y} = (Y_1, Y_2, Y_3)$ is now

$$\text{Var}(\mathbf{Y}) = \begin{bmatrix} \theta_1 + \theta_{12} + \theta_{13} & \theta_{12} & \theta_{13} \\ \theta_{12} & \theta_2 + \theta_{12} + \theta_{23} & \theta_{23} \\ \theta_{13} & \theta_{23} & \theta_3 + \theta_{13} + \theta_{23} \end{bmatrix}$$

Properties

For the general model we have

$$E(\mathbf{Y}) = \mathbf{A}\mathbf{M}$$

and

$$\text{Var}(\mathbf{Y}) = \mathbf{A}\mathbf{\Sigma}\mathbf{A}^T$$

where \mathbf{M} and $\mathbf{\Sigma}$ are the mean vector and the variance covariance matrix for the variables X_0, X_1, \dots, X_k respectively.

$\mathbf{\Sigma}$ is diagonal because of the independence of X_i 's and has the form

$$\mathbf{\Sigma} = \text{diag}(\theta_1, \theta_2, \dots, \theta_m)$$

Similarly

$$\mathbf{M}^T = (\theta_1, \theta_2, \dots, \theta_m)$$

Other models

While simple Poisson model is widely use we need to create counterparts to account for overdispersion¹ for example. So, we need to generalize in certain extent the univariate models.

A good starting point is to start from them. in the univariate setting such extensions are for example

- mixtures by assuming the parameters to be random variables
- inflated models by increasing the probability for certain values
- truncated models
- Hurdle models
- others

¹variance to the mean ratio larger than 1

Introducing dependence

A standard procedure to introduce dependence is mixing operation:

Assume conditional on a parameter θ we have a series of independent variables, θ introduces the dependence, i.e.

$$Y_i | \theta \sim \text{Poisson}(\alpha_i \theta), \quad i = 1, \dots, k$$

$$\theta \sim g(\theta)$$

or alternatively use different θ 's with some joint distribution:

$$Y_i | \theta_i \sim \text{Poisson}(\alpha_i \theta_i) \quad i = 1, \dots, k$$

$$\theta_1, \dots, \theta_k \sim g_k(\theta_1, \dots, \theta_k)$$

Introducing dependence

- The above models while starting from conditionally independent models they end up to dependent variables. They are simple to derive.
- For example if $g(\theta)$ is a gamma distribution we derive a kind of multivariate negative binomial distribution.
- But, be aware that actually we can derive different models with different properties by considering different kind of mixing!
- Also, in certain models, computations can be very large in large dimensions, so we need to consider some kind of feasibility when extending this way the models.

Mixtures of multivariate Poisson distribution

Several different ways to define such mixtures:

- Assume

$$(Y_1, \dots, Y_k) \sim k - \text{Poisson}(\alpha \theta)$$

$$\alpha \sim G(\alpha)$$

- Assume

$$(Y_1, \dots, Y_k) \sim m - \text{Poisson}(\theta)$$

$$\theta \sim G(\theta)$$

- Part of the vector θ varies, while some of the parameters remain constant. For example (in 2 dimensions)

$$(Y_1, Y_2) \sim \text{Biv. Poisson}(\theta_1, \theta_2, \theta_0)$$

$$\theta_1, \theta_2 \sim G(\theta_1, \theta_2)$$

Dependence Structure

Consider the case

$$(Y_1, \dots, Y_k | \theta) \sim k - \text{Poisson}(\theta)$$

$$\theta \sim G(\theta)$$

The unconditional covariance matrix is given by

$$\text{Var}(\mathbf{Y}) = \mathbf{A} \mathbf{D} \mathbf{A}^T$$

where \mathbf{A} is the matrix used to construct the conditional variates from the original independent Poisson ones and

$$\mathbf{D} = \begin{bmatrix} \text{Var}(\theta_1) + E(\theta_1) & \text{Cov}(\theta_1, \theta_2) & \dots & \text{Cov}(\theta_1, \theta_m) \\ \text{Cov}(\theta_1, \theta_2) & \text{Var}(\theta_2) + E(\theta_2) & \dots & \text{Cov}(\theta_2, \theta_m) \\ \dots & \dots & \dots & \dots \\ \text{Cov}(\theta_1, \theta_m) & \dots & \dots & \text{Var}(\theta_m) + E(\theta_m) \end{bmatrix}$$

Important findings

Remark 1: The above formula imply that if the mixing distribution allows for any kind of covariance between the θ 's then the resulting unconditional variables are correlated. Even in the case that one starts with independent Poisson variables the mixing operation can lead to correlated variables.

Remark 2: More importantly, if the covariance between the pairs (θ_i, θ_j) is negative the unconditional variables may exhibit negative correlation. It is well known that the multivariate Poisson distribution cannot have negative correlations, this is not true for its mixtures.

Remark 3: The covariance matrix of the unconditional random variables are simple expressions of the covariances of the mixing parameters and hence the moments of the mixing distribution. Having fitted a multivariate Poisson mixture model, one is able to estimate consistently the reproduced covariance structure of the data. This may serve as a goodness of fit index.

Random effects interpretation

Let (X_i, Y_i) follow jointly a mixture of 2 bivariate Poisson distributions, i.e.

$$P(x_i, y_i) = \sum_{j=1}^2 p_j \text{BP}(x_i, y_i; \lambda_{1j,i}, \lambda_{2j,i}, \lambda_{3j})$$

and

$$\log \lambda_{kj,i} = \mathbf{z}_{kj,i}' \beta_{kj}$$

$k = 1, 2, j = 1, 2, i = 1, \dots, n$, where $\mathbf{z}_{kj,i}$ are vectors of coefficients associated to parameter λ_k of the j -th component with info for the i -th observation and β 's are the associated regression coefficients. Similarly in a random effects formulation

$$\log \lambda_{kj,i} = \mathbf{z}_{k,i}' \beta_k + u_{ji}$$

where jointly the random effects (u_{1i}, u_{2i}) follow a finite distribution

Example -Multivariate Poisson lognormal

The model assumes that Y_i 's are conditionally independent and

$$Y_i | \theta_i \sim \text{Poisson}(\theta_i), \quad i = 1, \dots, k$$

$$(\theta_1, \dots, \theta_k) \sim \text{Multivariate Lognormal}(\mu, \Sigma)$$

Any correlation comes from the joint density of θ 's. The joint distribution cannot be written in closed form so we need approximations or computer intensive methods to estimate the parameters! The model can be easily extended to allow for covariates

Finite Mixtures of multivariate Poisson distribution

If we assume that θ can take only a finite number of different values finite multivariate Poisson mixture arise. The pf is given as

$$P(\mathbf{Y}) = \sum_{j=1}^g p_j P(\mathbf{Y} | \theta_j)$$

where $P(\mathbf{Y} | \theta_j)$ denotes the pf of a multivariate Poisson distribution, and as usual $0 < p_j < 1$, $j = 1, \dots, g$ are the mixing proportions with $\sum p_j = 1$.

Model usage

- Clustering multivariate count data
- We may allow for covariates in the parameters to have finite mixture of multivariate Poisson regressions.
- Modelling multivariate count data with flexible covariance structure (e.g. negative correlations)
- Also the marginal distributions are very flexible as they are finite poisson mixtures that can model a wide range of shapes!
- Inflated models can be considered as special cases!

Properties - Inference

- Identifiability has been proven based on the restriction that the vector of parameters is in lexicographical order
- Estimation through an EM algorithm using the latent structure of the mixture
- Consistency and asymptotic normality of the estimates
- Choice of the number of component via a standard method like AIC, BIC, NEC etc

Interesting things

- Standard model-based clustering procedures can be applied. For example, estimation is feasible via EM algorithm, selection of the number of components can be used in a variety of criteria etc
- Since, mixing operation imposes structure is not a good idea to start with a model with a lot of covariance terms.
- Since we work with counts one may use the frequency table instead of the original observations. This speeds up the process and the computing time is not increased so much even if the sample size increases dramatically

Example

Number of faults in the surface (X) and the interior (Y) in 100 lenses (Aitchinson and Ho, 1989). Negative correlation.

Models to consider:

Poisson-bivariate lognormal

$$X_i \sim \text{Poisson}(\lambda_1)$$

$$Y_i \sim \text{Poisson}(\lambda_2)$$

$$(\lambda_1, \lambda_2) \sim \text{Bivariate Lognormal}(\mu_1, \mu_2, \sigma_1, \sigma_2, \sigma_{12}).$$

and finite mixtures:

$$P_1(x_i, y_i) = \sum_{j=1}^g p_j MP_2(x_i, y_i | \lambda_{1j}, \lambda_{2j}, \lambda_{3j}) \quad (\text{Model 1})$$

$$P_2(x_i, y_i) = \sum_{j=1}^g p_j MP_2(x_i, y_i | \lambda_{1j}, \lambda_{2j}, 0) \quad (\text{Model 2})$$

Model 1	Loglikelihood	d_g	AIC
$g = 1$ (Simple Biv. Poisson)	-450.6038	3	907.208
$g = 2$	-432.6901	7	879.380
$g = 3$	-420.6122	11	863.224
$g = 4$	-419.9137	15	869.827
$g = 5$	-419.2967	19	876.593
$g = 6$ (NPMLE)	-419.004	23	884.008
Model 2			
$g = 1$ (Independent Poisson)	-450.6038	2	905.208
$g = 2$	-433.5881	5	877.176
$g = 3$	-423.6536	8	863.307
$g = 4$	-420.2615	11	862.523
$g = 5$ (NPMLE)	-419.2967	14	866.593
Poisson-Lognormal	-426.4	5	862.800
Independent Poisson lognormal	-428.8	4	865.600

Comparison

Comparing the P-Mult. Lognormal model and the FMMP model:

- The FMMP allows for unconditional correlation, i.e. it has two sources of correlation, intrinsic and due to mixing
- The FMMP is computationally easier, no need to approximate integrals
- The P-Mult. Lognormal is more parsimonious

Note: We may define models based on the negative binomial and/or the Poisson inverse Gaussian distributions. We can also generalize the ideas of inflation etc

Clustering Application

Multivariate Count data: number of 4 different type of crimes in Greece for the year 1997, for 50 prefectures.

Crime type:

rapes, arson, smuggling of antiquities and general smuggling.

The population of each prefecture is used as an offset.

The aim is to cluster the prefectures according to their profiles in those types of crimes.

Results (1)

- An EM type algorithm was used to fit the finite multivariate Poisson mixture model. The number g of components was considered as known for using the EM algorithm, but we fitted the model with increasing value of g in order to decide about the number of components.
- AIC criterion is used to find the optimal number of clusters. For a model with g components there are $11g - 1$ parameters to estimate. We selected a 4 clusters solution.

Results (2)

Mixing proportions $\hat{p} = (0.5915, 0.2266, 0.0638, 0.1181)$.

Parameters in matrix form:

$$\Theta = \begin{bmatrix} \theta_1 & \theta_{12} & \theta_{13} & \theta_{14} \\ & \theta_2 & \theta_{23} & \theta_{24} \\ & & \theta_3 & \theta_{34} \\ & & & \theta_4 \end{bmatrix}, \quad (3)$$

Results (2)

$$\Theta_1 = \begin{bmatrix} 17.339 & 0 & 4.772 & 0 \\ & 2.530 & 0.198 & 1.977 \\ & & 34.112 & 2.398 \\ & & & 6.171 \end{bmatrix}, \Theta_2 = \begin{bmatrix} 0 & 0 & 3.675 & 0 \\ & 9.897 & 1.925 & 1.938 \\ & & 5.320 & 0 \\ & & & 2.172 \end{bmatrix},$$

$$\Theta_3 = \begin{bmatrix} 0 & 20.424 & 0 & 0 \\ & 55.868 & 24.323 & 0 \\ & & 0 & 0 \\ & & & 0 \end{bmatrix}, \Theta_4 = \begin{bmatrix} 14.416 & 12.780 & 0 & 0 \\ & 3.048 & 0 & 0 \\ & & 8.934 & 0 \\ & & & 44.621 \end{bmatrix}$$

p_j	Rapes	Arsons	Manslaughter	Smuggling
0.156	27.708	13.530	16.764	39.945
0.066	19.216	97.243	23.605	0
0.420	23.533	2.217	41.460	4.222
0.357	9.399	11.690	26.497	11.122

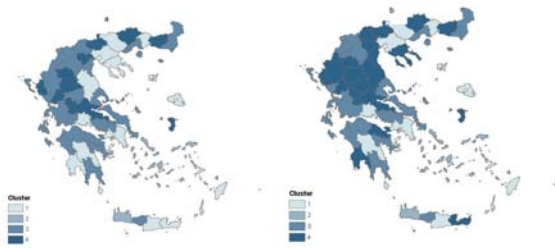
Table: The estimated rates for the four variables for all the components

Motivation

Model	Loglikelihood	d_m	AIC
$\theta_{24} = 0, \theta_{34} = 0$	-344.8708	35	759.742
$\theta_{24} = 0$	-342.3067	39	762.613
$\theta_{13} = 0, \theta_{24} = 0$	-347.6163	35	765.233
$\theta_{13} = 0, \theta_{34} = 0$	-347.9046	35	765.809
$\theta_{13} = 0, \theta_{14} = 0$	-348.8139	35	767.628
$\theta_{34} = 0$	-344.8708	39	767.742
All two way	-341.3377	43	768.675
$\theta_{13} = 0$	-347.6163	39	773.233
$\theta_{12} = 0, \theta_{34} = 0$	-358.2248	35	786.450
$\theta_{14} = 0, \theta_{34} = 0$	-359.5107	35	789.021

Table: The ten best models, selected according to the AIC

Figure: Maps showing the clusters that each prefecture belongs. Map a shows the full two way case, while map b the model with best AIC in Table 4.



Some more results

- Semiparametric models which do not assume a specific parametric mixing distributions but rather they assume some covariance
- Estimation issues: EM type algorithms and Stochastic ML methods. Bayesian estimation is also possible.

Inflated models

- Popular models in the univariate setting. Some specific values have more probability than that predicted by the model, this probability is removed from other points. Very flexible models can be constructed.
- Most common model in the univariate setting the zero-inflated model. i.e. the probability of observing a 0 value is larger than what the model predicts.
- Sparse literature in more dimensions. Inflation only in the (0,0) cell. Inflation in larger dimensions more difficult to handle.
- See paper of Wahlin (2001), Karlis and Ntzoufras (2003) etc.

Zero Inflated model

Inflate only the (0,0) cell.

The model:

$$P_D(x, y) = \begin{cases} (1-p)BP(x, y | \lambda_1, \lambda_2, \lambda_3), & x \text{ or } y \neq 0 \\ (1-p)BP(x, y | \lambda_1, \lambda_2, \lambda_3) + p, & x = y = 0, \end{cases} \quad (4)$$

The model assumes more (0,0) observations than that expected from a simple model.

Diagonal Inflated model

Inflate only the diagonal. Such model useful for soccer modelling, in biostatistics to represent patients with no change in before after studies etc

The model:

$$P_D(x, y) = \begin{cases} (1-p)BP(x, y | \lambda_1, \lambda_2, \lambda_3), & x \neq y \\ (1-p)BP(x, y | \lambda_1, \lambda_2, \lambda_3) + pD(x, \theta), & x = y, \end{cases} \quad (5)$$

where $D(y, \theta)$ is discrete distribution with parameter vector θ .

Choices for $D(x, \theta)$ are the Poisson, the Geometric or simple discrete distributions such as the Bernoulli. The Geometric distribution might be of great interest since it has mode at zero and decays quickly. More general than the (0,0) inflation

Zero Inflated model

We inflate the zero counts for one (or both) variables.

The model:

$$P_D(x, y) = \begin{cases} (1-p)BP(x, y | \lambda_1, \lambda_2, \lambda_3), & x \neq 0 \\ (1-p)BP(x, y | \lambda_1, \lambda_2, \lambda_3) + pD(y, \theta), & x = 0, \end{cases} \quad (6)$$

where $D(x, \theta)$ is discrete distribution with parameter vector θ .

Useful Properties

- The marginal distributions of a diagonal inflated model are not Poisson distributions but mixtures of distributions with one Poisson component. They can be zero inflated Poisson as well.
- Secondly, if $\lambda_3 = 0$ the resulting inflated distribution introduces a degree of dependence between the two variables under consideration. For this reason, diagonal inflation may correct both overdispersion and correlation problems.
- Models can be fitted using an EM algorithm.

Bivariate (Multivariate) Negative Binomial

Assume

$$(Y_1, \dots, Y_k) \sim k - \text{Poisson}(\alpha\theta) \\ \alpha \sim G(\alpha)$$

If $G(\cdot)$ is a Gamma density with parameters (γ, β) then a bivariate (multivariate) Negative Binomial distribution can be obtained (Edwards and Gurland, 1961). The bivariate joint probability function is given by

$$P_G(y_1, y_2) = q^\gamma \sum_{i=0}^{\min(y_1, y_2)} \frac{\Gamma(\gamma + y_1 + y_2 - i)}{\Gamma(\gamma) i! (y_1 - i)! (y_2 - i)!} p_1^{y_1 - i} p_2^{y_2 - i} p_3^i,$$

$y_1, y_2 = 0, 1, \dots$, and $p_i = \frac{\lambda_i}{\sum_{j=1}^3 \lambda_j + \beta}$, $q = \frac{\beta}{\sum_{j=1}^3 \lambda_j + \beta}$. This probability function is described in Subrahmaniam (1966).

Bivariate (Multivariate) Negative Binomial

Assume

$$Y_j \sim \text{Poisson}(\alpha\theta_j) \text{ and they are i.i.d.} \\ \alpha \sim G(\alpha)$$

If $G(\cdot)$ is a Gamma density with parameters (γ, γ) (i.e. $E(\alpha) = 1$) then a bivariate (multivariate) Negative Binomial distribution can be obtained (Munkin and Trivedi, 1999). For the bivariate case we have

$$P(y_1, y_2) = \frac{\Gamma(y_1 + y_2 + \gamma)}{y_1! y_2! \Gamma(\gamma)} \left(\frac{\theta_1}{\theta_1 + \theta_2 + 1} \right)^{y_1} \left(\frac{\theta_2}{\theta_1 + \theta_2 + 1} \right)^{y_2} \left(\frac{1}{\theta_1 + \theta_2 + 1} \right)^\gamma$$

The model can allow for covariates by assuming that $\log \theta_j = x_j \beta_j$ for a covariate vector x_j and regression coefficients β_j

Summary

- We defined multivariate Poisson models and based on them and the idea of mixing we derived even more flexible bivariate and multivariate models.
- We have created a package called `bivpois` in R to facilitate the use of some of the models described, especially bivariate Poisson regression allowing also for inflation.
- While the models offer some flexibility they cannot model certain cases, as for example they have some kind of symmetry with respect to the marginal distributions etc.
- There are other derivations based on trivariate reduction or conditional distributions, for example. They have found limited (if any) applications

Part II - copulas

Copulas are currently fashionable in several cases

- Biostatistics
 - Longitudinal data analysis
 - Survival analysis
 - Genetics
- Insurance
- Finance even incorrectly ...
- **Neuroscience**

Pro-definition

- It is well known in nonparametric statistics that if X has a continuous cdf F then $F(X) \sim U(0, 1)$. Moreover by the inversion method $X = F^{-1}(U)$ is a sample from F .
- Then if $U \sim U(0, 1)$ and $V \sim U(0, 1)$, $X = F^{-1}(U)$ and $Y = G^{-1}(V)$ is a independent sample from F and G respectively.
- Dependent bivariate sample \rightarrow Bivariate distribution, whose one dimensional margins are uniform \rightarrow Copula function.

We referred to copulas as distribution functions whose one dimensional margins are uniform.

Definition

A bivariate copula is a function C from \mathbf{I}^2 to \mathbf{I} with the following properties:

- For every u, v in \mathbf{I}

$$C(u, 0) = 0 = C(0, v) \quad \text{and} \quad C(u, 1) = u, C(1, v) = v$$

- For every u_1, u_2, v_1, v_2 in \mathbf{I} such that $u_1 \leq u_2$ and $v_1 \leq v_2$,

$$C(u_2, v_2) - C(u_2, v_1) - C(u_1, v_2) + C(u_1, v_1) \geq 0,$$

(Nelsen, 2006).

If C is be considered to be a distribution function of two random variables U and V , the first condition ensures that U and V have uniform marginal distributions.

The second condition, often referred to as the rectangular inequality, simply requires that C is a valid distribution function, i.e.

$$\text{Prob}\{u_1 \leq U \leq u_2, v_1 \leq V \leq v_2\} \geq 0.$$

Sklar's theorem

Let H be a joint distribution function with margins F and G . Then there exists a copula C such for all x, y in \mathbb{R}

$$H(x, y) = C(F(x), G(y)) \quad (7)$$

If F and G are continuous, then C is unique, otherwise, C is uniquely determined on $\text{Range}F \times \text{Range}G$. Conversely, if C is a copula and F and G are distribution functions, then the function H defined by (7) is a joint distribution function with margins F and G (Nelsen, 2006).

Dependence structure

- The dependence between random variables is completely described by their joint distribution.
- Dependence and marginal behavior can be separated. (not exactly true for the discrete case ...)
- A copula of a multivariate distribution can be considered to be the part describing the dependence structure (Joe, 1997).

Copula based measures of dependence

- Kendall's tau,

$$\tau_C = 4 \int \int_{\mathbf{I}^2} C(u, v) dC(u, v) - 1.$$

- Spearman's rho,

$$\rho_C = 12 \int \int_{\mathbf{I}^2} [C(u, v) - uv] dudv.$$

Copulas, Kendall's tau and Spearman's rho are invariant under strictly increasing transforms (Nelsen, 2006).

Elliptical copulas

• Normal copula.

For $-1 \leq \theta \leq 1$, $C(u, v; \theta) = \Phi_\theta(\Phi^{-1}(u), \Phi^{-1}(v))$, where Φ is the $N(0,1)$ c.d.f., Φ^{-1} is the functional inverse of Φ and Φ_θ is the bivariate standard normal c.d.f. with correlation θ .

• t-copula

For $-1 \leq \theta \leq 1$, $C(u, v; \theta, \nu) = T_{\theta, \nu}(t_\nu^{-1}(u), t_\nu^{-1}(v))$, where t_ν is the t c.d.f. with ν degrees of freedom, t_ν^{-1} is the functional inverse of t_ν and T is the bivariate standard t c.d.f. with ν degrees of freedom and correlation θ .

For both of them we need to evaluate a bivariate(multivariate) integral.

Archimedean copulas

Theorem 1 (Nelsen, 1998) Let φ be a continuous, strictly decreasing function from $[0, 1] \rightarrow [0, \infty]$ such that $\varphi(1) = 0$ and $\varphi(0) = \infty$. Then the function

$$C(u, v) = \varphi^{-1}(\varphi(u) + \varphi(v))$$

from $[0, 1]^2 \rightarrow [0, 1]$ is a copula if and only if φ is convex.

Theorem 2 (Nelsen, 1998) Let U and V be uniform random variables whose copula is Archimedean with generator φ . For $0 < t \leq 1$ set,

$$\lambda(t) = \varphi(t)/\varphi(t)'$$

Then the random variable $C(U, V)$ is distributed as,

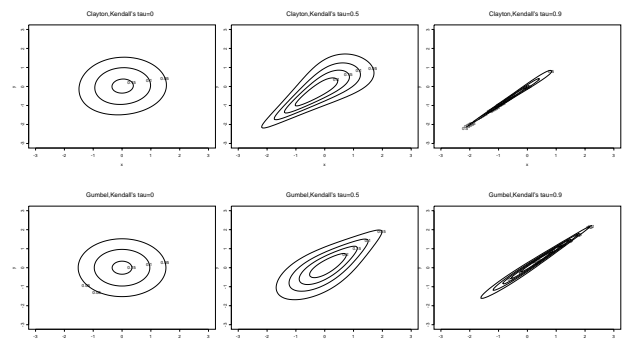
$$K(t) = t - \lambda(t),$$

on unit interval.

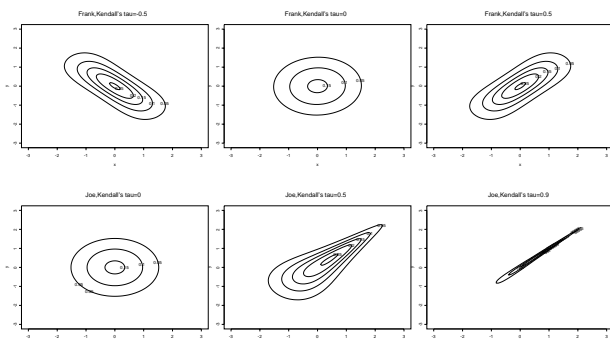
Members of Archimedean class

Family	$\varphi_\theta(t)$	$\theta \in$
Clayton	$\frac{1}{\theta}(t^{-\theta} - 1)$	$(0, \infty)$
Gumbel	$(-\ln t)^\theta$	$[1, \infty)$
Frank	$-\ln \frac{e^{-\theta t} - 1}{e^{-\theta} - 1}$	$(-\infty, \infty) \setminus \{0\}$
Joe	$-\ln[1 - (1 - t)^\theta]$	$[1, \infty)$

Contour plots (1)



Contour plots (2)



Estimation

How?

- Parametric
- Semi-parametric
- Non-parametrical

Methods

- Inference functions of margins (IFM)
- Full maximum likelihood (FML)

Likelihoods

Consider a copula based parametric model for the random variables X, Y , with cumulative distribution function (Joe, 1997),

$$H(x, y; \alpha_1, \alpha_2, \theta) = C(F(x; \alpha_1), F(y; \alpha_2); \theta)$$

We assume that H has a density $h(x, y; \alpha_1, \alpha_2, \theta)$. More on this later!
We can consider the two log-likelihood functions for the univariate margins,

$$L_x(\alpha_1) = \sum_{i=1}^n \log f(x_i; \alpha_1) \quad \text{and} \quad L_y(\alpha_2) = \sum_{i=1}^n \log g(y_i; \alpha_2),$$

and the log-likelihood function for the joint distribution,

$$L(\theta, \alpha_1, \alpha_2) = \sum_{i=1}^n \log h(x, y; \alpha_1, \alpha_2, \theta).$$

Density of bivariate distribution

- Continuous data,

$$h(x, y) = \frac{\partial^2 H(x, y)}{\partial x \partial y} = \frac{\partial^2 C(u, v)}{\partial u \partial v} f(x) g(y)$$

- Mixed data

$$h(x, y) = \frac{\partial H(x, y)}{\partial x} \frac{\partial H(x, y - 1)}{\partial x} = f(x) \left(\frac{\partial C(u, G(y))}{\partial u} - \frac{\partial C(u, G(y - 1))}{\partial u} \right)$$

- Discrete data

$$h(x, y) = C(F(x), G(y)) - C(F(x - 1), G(y)) - C(F(x), G(y - 1)) + C(F(x - 1), G(y - 1))$$

Example: 3-variate case, discrete data

$$\begin{aligned} h(y_1, y_2, y_3) = & C(F_1(y_1), F_2(y_2), F_3(y_3)) - C(F_1(y_1 - 1), F_2(y_2), F_3(y_3)) - \\ & - C(F_1(y_1), F_2(y_2 - 1), F_3(y_3)) - C(F_1(y_1), F_2(y_2), F_3(y_3 - 1)) + \\ & + C(F_1(y_1 - 1), F_2(y_2 - 1), F_3(y_3)) + C(F_1(y_1 - 1), F_2(y_2), F_3(y_3 - 1)) \\ & + C(F_1(y_1), F_2(y_2 - 1), F_3(y_3 - 1)) - \\ & - C(F_1(y_1 - 1), F_2(y_2 - 1), F_3(y_3 - 1)) \end{aligned}$$

It becomes awkward...

Inference functions of margins (IFM)

- Step 1: Maximize

$$L_x(\alpha_1) = \sum_{i=1}^n \log f(x_i; \alpha_1) \quad \text{and} \quad L_y(\alpha_2) = \sum_{i=1}^n \log g(y_i; \alpha_2),$$

- Step 2. Plug-in in the estimates $\hat{\alpha}_1$ and $\hat{\alpha}_2$ and maximize for θ only the log-likelihood function for the joint distribution,

$$L(\theta, \alpha_1, \alpha_2) = \sum_{i=1}^n \log h(x, y; \hat{\alpha}_1, \hat{\alpha}_2, \theta).$$

Full maximum likelihood (FML)

Maximize for all parameters simultaneously the log-likelihood function for the joint distribution,

$$L(\theta, \alpha_1, \alpha_2) = \sum_{i=1}^n \log h(x, y; \alpha_1, \alpha_2, \theta).$$

Methods can be used complimentary, i.e the IFM can be the basis for good initial values for FML. IFM more useful for larger dimensions! Joe (2005) has proved that IFM is asymptotically efficient apart from cases in the boundary.

Copulas and count data

- In fact when the random variables are not continuous we do not have the uniqueness property of the copula. In practice this is not a problem as all the copulas will give us the same description.
- For discrete data the interpretation of copulas is somewhat different.
- Recall that in order to calculate the joint probabilities one need to evaluate the copula a number of times! But a copula is a cdf and for certain copulas this imply the evaluation of a multivariate integral!
- This is an obstacle on the use of certain copulas with no closed form cdf!

Literature of Copulas and count data

- van Ophem (1999) and Lee(2001) exploit the use of bivariate normal copula to model count data.
- Song (2000,2007) defined multivariate dispersion models through multivariate normal copula. Note, severe computational problems.
- Lee (1999) used Frank copula for modelling rugby data.
- Cameron et al (2004) used Frank and other Archimedean copulas for health economic application.
- Mc Hale and Scarf (2007) used Frank copula for soccer data.
- Nikoloulopoulos and Karlis (2008a) used mixture of maxid to allow for greater flexibility with 4-variate data.
- Nikoloulopoulos and Karlis (2008b) used a new flexible copula based on finite normal mixtures to model 6-variate data.
- Berkes (yesterday) and Onken (yesterday) used them in neuroscience application

Copulas and count data

- The latter is even more problematic when estimating parameters since it may result to negative probabilities during maximization and in general produces a lot of problems.
- Kendall's tau is also restricted in certain cases. In general if there are n different values that the data can take we have a limit of the form $1/(n-1)$ for Kendall's tau. So interpreting the correlation must be cautious
- Kendall's tau measures concordance.

$$\tau(X, Y) = P[(X_1 - X_2)(Y_1 - Y_2) > 0] - P[(X_1 - X_2)(Y_1 - Y_2) < 0],$$

For discrete data it is not true that $P(\text{concordance}) + P(\text{discordance}) = 1$, since we have a positive probability of a tie!

Related literature Genest and Neshlehova, (2007), Nikoloulopoulos, (2007), Joe (1997), Nikoloulopoulos and Karlis (2008a,b)

Kendall's tau for count data

Theorem(Nikoloulopoulos, 2007) Let X and Y be integer-valued discrete random variables whose joint distribution is H , with marginal cdfs F, G , pmfs f, g and copula C . Then the population version of Kendall's tau for X and Y is given by

$$\tau(X, Y) = \sum_{x=0}^{\infty} \sum_{y=0}^{\infty} h(x, y) \{4C(F(x-1), G(y-1)) - h(x, y)\} + \sum_{x=0}^{\infty} \{f^2(x) + g^2(x)\} - 1, \quad (8)$$

where,

$$h(x, y) = C(F(x), G(y)) - C(F(x-1), G(y)) - C(F(x), G(y-1)) + C(F(x-1), G(y-1))$$

is the pmf of X and Y .

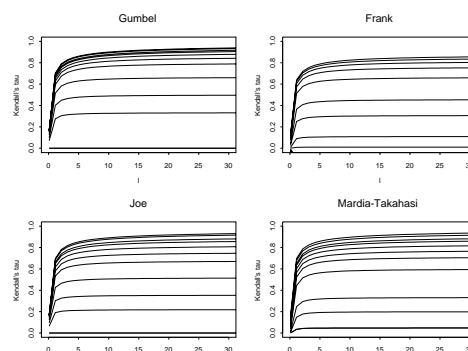


Figure: Kendall'tau values computed using different copulas for a grid of parameter value for each copula (different lines) and Poisson marginal distributions

Application of copulas to discrete data in brief

- It is useful to consider copulas with closed form cdf.
- Unfortunately this may exclude some copulas with flexible structure.
- There are flexible copulas in the literature to play this role. For example the family of mixture of max-id copulas of Joe (1997) or the family proposed in Nikoloulopoulos and Karlis, (2008)
- There is a trade off between very flexible copulas without simple form to carry out the computations.
- Note also the number of parameters we need. For fully describe correlation for k variables we need $k(k-1)/2$ parameters.
- Marginal distributions can now selected based on all the known univariate models, this allows great flexibility.
- Based on simulation, IFM works well.

Proposed approaches for multivariate count data - Summary

- Multivariate normal copula: One needs to calculate the cumulative function in large dimensions of the multivariate normal distribution. This is not easy and only approximate methods exist, which may cause numerical problems during maximization-estimation
- Multivariate version of Archimedean copulas: They offer limited dependence structure which is not realistic in most practical applications
- Mixtures of max id copulas. This idea used mixture formulation to add in two steps dependence structure. The dependence is necessarily positive. It is not so easy to interpret the dependence.

Application

Y_1 : the number of consultations with a doctor or a specialist and
 Y_2 : the total number of prescribed and non-prescribed medications used in past 2 days.
 X_i : age, sex, income, health score, and an indicator of chronic health conditions ($n = 5190$).

number of consultations	number of prescribed and nonprescribed medications								
	0	1	2	3	4	5	6	7	8
0	2009	1144	492	262	124	47	28	18	17
1	164	193	178	98	65	41	21	8	14
2	39	30	40	17	21	10	9	4	4
3	5	12	4	4	1	1	2	1	0
4	6	7	3	4	3	1	0	0	0
5	2	2	4	0	1	0	0	0	0
6	2	0	0	4	1	2	1	1	1
7	1	0	2	3	1	1	1	0	3
8	1	1	0	1	1	0	1	0	0
9	0	0	0	0	0	0	0	0	1

Table: Data from the Australian Health Survey (Cameron and Trivedi, 1986)

Specify marginal distributions

- For our data negative binomial provides a good univariate model.
- It allows for the large over-dispersion in the doctor visit and medication data (the sample variances are roughly two times the sample means).
- For each observation $i = 1, \dots, 5190$ and $j = 1, 2$,

$$F_j(y_{ji} | \mathbf{X}_{ji}, \beta_j) = \sum_{k=0}^{y_{ji}} \frac{\Gamma(\vartheta_j + k)}{\Gamma(\vartheta_j) \Gamma(k+1)} \frac{\mu_{ji}^k \vartheta_j^{\vartheta_j}}{(\mu_{ji} + \vartheta_j)^{\vartheta_j + k}},$$

where $E(y_{ji}) = \mu_{ji} = \exp(\mathbf{X}_{ji} \beta_j)$ and $\text{var}(y_{ji}) = \mu_{ji} + \mu_{ji}^2 / \vartheta_j$.

Computation of estimates and standard errors

Full Maximum Likelihood:

The maximum likelihood estimates and standard errors of $\varrho = (\vartheta, \beta, \theta)$ for each copula model were obtained by numerical maximization of the joint likelihood, using the `nlm` function in the statistical software R (a quasi Newton optimization).

Inference Function of Margins:

The univariate parameters were estimated fitting separate negative binomial models and the copula parameter was estimated from bivariate likelihood using a quasi Newton routine, with univariate parameters fixed as estimated from the separate negative binomial models. Standard errors of estimates $\hat{\varrho}$ were obtained by the jackknife method proposed by Joe (1997) with 52 blocks of 100.

Application results

Covariate	Normal		Clayton		Gumbel		Frank		Joe	
	coeff.	St. Err	coeff.	St. Err	coeff.	St. Err	coeff.	St. Err	coeff.	St. Err
Number of consultations with a doctor or a specialist										
ϑ	0.52	0.05	0.52	0.04	0.47	0.04	0.52	0.04	0.47	0.04
Intercept	-2.08	0.14	-2.07	0.13	-2.03	0.12	-2.07	0.12	-2.04	0.12
Sex	0.24	0.09	0.22	0.07	0.24	0.07	0.23	0.07	0.24	0.07
Age	1.19	0.21	1.22	0.20	1.09	0.17	1.20	0.17	1.08	0.17
Income	-0.15	0.11	-0.14	0.09	-0.16	0.10	-0.15	0.10	-0.16	0.10
H. score	0.16	0.01	0.16	0.01	0.16	0.01	0.16	0.01	0.16	0.01
Chronic	0.07	0.07	0.04	0.07	0.12	0.07	0.06	0.07	0.12	0.07
Number of prescribed and non-prescribed medications used in past 2 days										
ϑ	2.75	0.18	2.69	0.19	2.61	0.18	2.71	0.19	2.61	0.18
Intercept	-1.23	0.06	-1.20	0.06	-1.20	0.06	-1.19	0.06	-1.21	0.06
Sex	0.47	0.04	0.47	0.03	0.46	0.03	0.46	0.03	0.47	0.03
Age	1.76	0.08	1.73	0.08	1.74	0.08	1.73	0.08	1.74	0.08
Income	0.03	0.05	0.03	0.05	0.04	0.05	0.02	0.05	0.04	0.05
H. score	0.10	0.01	0.10	0.01	0.10	0.01	0.10	0.01	0.10	0.01
Chronic	0.34	0.03	0.33	0.03	0.34	0.03	0.33	0.03	0.34	0.03
θ (FML)	0.34	0.03	0.82	0.07	1.15	0.02	2.14	0.15	1.17	0.02
θ (IFM)	0.31	0.02	0.81	0.08	1.14	0.01	2.14	0.15	1.16	0.02
LogL. FML			-10594.10		-10619.63		-10583.86		-10638.40	
LogL. IFM	-10588.39		-10594.61		-10621.27		-10584.19		-10640.17	

Multivariate normal copula

- The copula of the n -variate normal distribution with linear correlation matrix R ,

$$C_R^n(\mathbf{u}) = \Phi_R^n(\Phi^{-1}(u_1), \dots, \Phi^{-1}(u_n)),$$

where Φ is the $N(0,1)$ c.d.f., Φ^{-1} is the functional inverse of Φ and Φ_R^n is the n -variate standard normal c.d.f. with linear correlation matrix R .

- Density of n -variate distribution via normal copula for continuous random variables Y_1, \dots, Y_n (Lambert and Vandenhende, *Statist. Med.*, 2002),

$$h(\mathbf{y}) = \frac{\partial H(\mathbf{y})}{\partial y_1 \dots \partial y_n} = \frac{\partial C_R^n(\mathbf{u})}{\partial u_1 \dots \partial u_n} \prod_{i=1}^n f_i(y_i) = c_R^n(\mathbf{u}) \prod_{i=1}^n f_i(y_i),$$

where

$$c_R^n(\mathbf{u}) = \frac{\phi_R^n(\Phi^{-1}(u_1), \dots, \Phi^{-1}(u_n))}{\prod_{i=1}^n \phi(\Phi^{-1}(u_i))},$$

with ϕ and ϕ_R^n denoting, respectively, the density of a standardized univariate and n -variate normal.

Density of n -variate distribution via normal copula for discrete random variables

Definition Let $\mathbf{c} = (c_1, c_2, \dots, c_n)$ be vertices where each c_k is equal to either y_k or $y_k - 1$. Then the probability mass function h of discrete random variables Y_1, Y_2, \dots, Y_n is given by,

$$h(y_1, y_2, \dots, y_n) = \sum \text{sgn}(\mathbf{c}) H(\mathbf{c}) = \sum \text{sgn}(\mathbf{c}) C_R^n(F_1(c_1), \dots, F_n(c_n)),$$

where the sum is taken over all vertices \mathbf{c} , and $\text{sgn}(\mathbf{c})$ is given by,

$$\text{sgn}(\mathbf{c}) = \begin{cases} 1, & \text{if } c_k = y_k - 1 \text{ for an even number of } k\text{'s.} \\ -1, & \text{if } c_k = y_k - 1 \text{ for an odd number of } k\text{'s.} \end{cases}$$

Multivariate Archimedean copulas

Definition (Widder, 1941) A function $g(t)$ is completely monotonic on an interval J if it is continuous there and has derivatives of all orders which alternate in sign, i.e., if it satisfies $(-1)^k \frac{d^k}{dt^k} g(t) \geq 0$ for all t in the interior of J and $k = 0, 1, 2, \dots$.

Theorem (Nelsen, 1998) If C^n is the function from \mathbf{I}^n to \mathbf{I} given by,

$$C^n(\mathbf{u}) = \varphi^{-1}(\varphi(u_1) + \varphi(u_2) + \dots + \varphi(u_n)),$$

then C^n is a n -copula for all $n \geq 2$ if and only if φ^{-1} is completely monotonic (Laplace transform) on $[0, \infty)$.

Very limited positive dependence structure since all k -margins are identical.

Partially Symmetric copulas (1)

Since the general multivariate result is notationally complex, we indicate the 4-variate copula (Joe, 1997),

$$C(\mathbf{u}; \theta) = \phi_1(\phi_1^{-1} \circ \phi_2(\phi_2^{-1} \circ \phi_3(\phi_3^{-1}(u_1) + \phi_3^{-1}(u_2)) + \phi_2^{-1}(u_3)) + \phi_1^{-1}(u_4))$$

where ϕ_1, ϕ_2, ϕ_3 are Laplace transforms and $\phi_1^{-1} \circ \phi_2, \phi_1^{-1} \circ \phi_3$ are in $\mathbf{L}_\infty^* = \{\omega : [0, \infty) \rightarrow [0, \infty) | \omega(0) = 0, \omega(\infty) = \infty, (-1)^{j-1} \omega^{(j)} \geq 0, j = 1, \dots, \infty\}$.

Properties:

- Bivariate margins, associated with Laplace transforms that are more nested, are larger in concordance than those are less nested.
- For the n -variate copula, $n - 1$ dependence parameters for $n(n - 1)/2$ margins.

Partially Symmetric copulas (2)

For each of the following families of Laplace transforms $\phi(\cdot)$:

- 1 $\exp(-t^{1/\theta}), \theta \geq 1,$
- 2 $(1+t)^{-1/\theta}, \theta \geq 0,$
- 3 $1 - (1 - e^{-t})^{1/\theta}, \theta \geq 1,$
- 4 $-\theta^{-1} \log(1 - (1 - e^{-\theta})e^{-t}), \theta \geq 0,$

the condition $\phi_{\theta_1}^{-1} \circ \phi_{\theta_2} \in \mathbf{L}_\infty^*, \theta_1 < \theta_2$ is satisfied, so the copula function

$$C(\mathbf{u}; \theta) = \phi_{\theta_1}(\phi_{\theta_1}^{-1} \circ \phi_{\theta_2}(\phi_{\theta_2}^{-1} \circ \phi_{\theta_3}(\phi_{\theta_3}^{-1}(u_1) + \phi_{\theta_3}^{-1}(u_2)) + \phi_{\theta_2}^{-1}(u_3)) + \phi_{\theta_1}^{-1}(u_4))$$

is a proper distribution function as described in Joe (1997). Marginally (X_1, X_3) has the same dependence with (X_2, X_3) , similarly $(X_1, X_4), (X_2, X_4)$ and (X_3, X_4) have the same dependence. Moreover the pair (X_1, X_2) has the larger dependence.

These constructions produce positively dependent random vectors.

Multivariate copulas with general dependence

As proposed by Joe (1997) the function of the form

$$C(\mathbf{u}) = \phi\left(-\sum_{i < j} \log C'_{ij}(e^{-p_i \phi^{-1}(u_i)}, e^{-p_j \phi^{-1}(u_j)}) + \sum_{i=1}^m \nu_i p_i \phi^{-1}(u_i)\right),$$

where

- $C'_{ij}, 1 \leq i < j \leq m,$ are bivariate copulas that are max infinitely divisible distributions,
- typically ν_i are non-negative, despite if some of the copulas C'_{ij} corresponds to independence,
- $p_i = (m - 1 + \nu_i)^{-1}, i = 1, \dots, m$

is a multivariate copula with flexible positive dependence.

Properties of multivariate copulas with general dependence

- The Laplace transform ϕ introduce the smallest dependence between random variables, while the copulas C'_{ij} add some pairwise dependence.
- Partial closure under the margins, which is succeeded by the inclusion of parameters ν_j .
- Closed form cumulative distribution functions if parametric families for ϕ and C'_{ij} are chosen appropriately, leading to faster computation of probabilities of the form $Pr(\mathbf{Y} = y|x)$.

Some final Comments n copulas and count data

- For multivariate data there are methods that overcome the curse of dimensionality problem, like the composite likelihood method of Zhao and Joe(2005)
- Goodness of fit: for count data chi-square type statistics are easy to calculate
- Parameters to the copula parameters: this allows to model more flexibly the data and allows for time varying copula effects! (Dependence changes over time)
- Model selection issues. Unfortunately quite often we firstly fit several copulas and then select one. This in general leads to wrong standard errors (and inferences) as the uncertainty is underestimated! Use model averaging.

Summarizing

- While copulas are relatively easy to use in the bivariate case generalization to more dimensions is not easy.
- Multivariate archimedean copulas do not allow for flexible correlation, the normal copula has computational problems, while other construction are complicated and not easy to interpret.
- So, we need to define appropriate copulas or at least to find cheap ways to estimate parameters.
- For example, the covariances can be reproduced by considering joint bivariate pairs.

Part III- Time series models

Why not standard time series models?

Standard time series models are based on an assumption of normal and related distributions which, while reasonable for continuous data, fail substantially for count data.

They fail for one or more of the following reasons:

- Low counts, small mean
- A lot of zeros
- Symmetry not present
- Probabilities hard to compute and interpret

Normal approximations exist and work well for large values (so, usually in aggregated data).

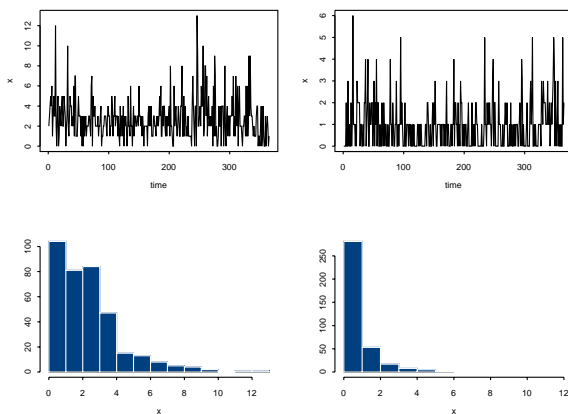


Figure: Example of discrete-valued time series

Time Series Models

- Parameter driven models

$$Y_t | \epsilon_t, X_t \sim \text{Poisson}(\epsilon_t \exp(x_t' \beta))$$

ϵ_t is a latent process usually of a standard form as in classical time series, e.g

$$\epsilon_t = \rho \epsilon_{t-1} + w_t, \quad w_t \text{ iid } N(0, \sigma^2)$$

- Observation driven models

$$Y_t | Y_{t-1}, X_t \sim \text{Poisson}(\exp(x_t' \beta + f(Y_{t-1})))$$

where $f(y)$ can be any function

- Hidden Markov models : there is a hidden process that change the state of the series and add autocorrelation
- Integer Autoregressive Models: They mimic standard autoregressive models suitably defined for integers

Integer Autoregressive

We mimic the classical AR model for normally distributed data. The process is defined as

$$Y_t = \alpha \circ Y_{t-1} + R_t$$

where R_t is a sequence of uncorrelated non-negative integer-valued random variables having mean μ and finite variance σ^2 and X_0 represents an initial value of the process while the operator " \circ " denotes the binomial thinning operator.

The operator " \circ " is defined by

$$\alpha \circ X = \sum_{t=1}^X Y_t,$$

where Y_t are Bernoulli random variables with $P(Y_t = 1) = \alpha = 1 - P(Y_t = 0)$, $\alpha \in [0, 1]$ and is called the binomial thinning operator.

Properties

The mean and variance of a stationary INAR(1) process (i.e. $0 < \alpha < 1$) are constants given by the formulae

$$\mu_X = E(X_t) = \frac{\mu_R}{1 - \alpha} \quad \text{and} \quad \sigma_X^2 = \text{Var}(X_t) = \frac{\alpha \mu_R + \sigma_R^2}{1 - \alpha^2}, \quad (9)$$

where μ_R and σ_R^2 are respectively the (assumed finite) mean and variance of the i.i.d. innovations. The auto-covariance function of a stationary INAR(1) process $\{X_t\}_{t \in \mathbb{Z}}$ is given by the formula

$$\gamma_X(k) = \text{Cov}(X_t, X_{t-k}) = \alpha^{|k|} \sigma_X^2, \quad k \in \mathbb{Z}. \quad (10)$$

From the covariance function, it is easy to obtain the autocorrelation function $\rho(k)$ as follows:

$$\rho(k) = \frac{\gamma(k)}{\gamma(0)} = \alpha^{|k|}.$$

Thus, the autocorrelation function $\rho(k)$ decays exponentially with lag k

Poisson models

$$Y_t = a \circ Y_{t-1} + R_t$$

$$R_t \sim \text{Poisson}(\lambda_t)$$

$$\log \lambda_t = x_t' \beta$$

where x_t' are covariates associated with the t observation and β as usually the coefficients to be estimated. The autocorrelation in lag 1 is α .

A serious limitation is that marginally the data follow a Poisson distribution which is not so realistic

We may add covariates in α but this might cause problems in the interpretability of the model!

Overdispersion

Poisson INAR model assumes that marginally the densities are Poisson which is too restrictive. We need to generalize so as to cover overdispersion. The situation becomes more complicated. It can be seen that the overdispersion of the series relate to the overdispersion of the innovation by

$$ID(Y_t) = \frac{a + ID(R_t)}{1 + a}$$

Hence, we need an overdispersed distribution for the innovations!

A simple idea is to use the negative binomial distribution but this leads to very complicated likelihood. Another idea is to use a finite mixture of Poisson distributions. Estimation is feasible. There is a trade of between the descriptive and the predictive power of such a model.

Extensions

- Not constant α . Flexible but it allows for limited overdispersion
- Define INAR(p) models of the form

$$Y_t = \sum_{i=1}^p a_i \circ Y_{t-i} + R_t$$

- Define INARMA(p,q) models of the form

$$Y_t = \sum_{i=1}^p a_i \circ Y_{t-i} + \sum_{j=1}^q \beta_j \circ R_{t-j} + R_t$$

- Other distributional choices for R_t so as to provide tractable likelihoods!
- Dependence between Y_{t-1} and R_t . This may create more complicated models with more interesting dependence structure but more computationally demanding procedures.

Extensions

- For the INAR(p) model one may relate the innovation terms with covariates. Also the thinning parameters may depend on some parameters. Estimation is easy via EM type schemes (see Brijis et al, 2008).
- Extensions of this may imply negative binomial innovations to account for the overdispersion associated with some covariates.
- Also Bayesian estimation is simple.
- We are now working on multivariate extension! a multivariate extension implies observations that are correlated between them at the same time, but in different times there is also some correlation! this is a very general model!
- Example: Purchases of different products. At the same time point it is correlated since this reflect customer choices but also across time this is correlated as it represents a time series.

Zeger's model

We assume

$$Y_t \sim \text{Poisson}(\mu_t \exp(\epsilon_t))$$

$$\log(\mu_t) = x_t' \beta$$

$$\epsilon_t = \phi \epsilon_{t-1} + R_t$$

$$R_t \sim N(0, \sigma^2)$$

Pros: Overdispersion and correlation are imposed together by ϵ_t .

Cons: Hard to be estimated

Properties

$$\rho(k) = \frac{\rho_\epsilon(k)}{[(1 + (\sigma^2 \mu_t)^{-1})(1 + (\sigma^2 \mu_{t+k})^{-1})]^{1/2}}$$

where $\rho_\epsilon(k)$ is the autocorrelation of the process assumed for ϵ . Hence, autocorrelation from the innovation process. Stationarity is not guaranteed.

Important things

- Good news: Since the correlation properties are coming from the innovation one can use any of the models for continuous time series to create a similar model for count data.
- Bad news: Estimation quite complicated: Use a GEE approach. Extensions are very difficult.

Hidden Markov Models

- Consider that there are two states in which my series is.
- Conditionally on these states my observations are independent.
- But since I cannot know the states the series is at each time point and because the states are correlated I observe data that are correlated because there is the latent process that changes the states at each time point!
- Usually the states are connected through a Markov Process with some transition probabilities
- we may allow for higher order Markovian properties etc.

Properties

Pros

- Easy interpretation
- Markovian and well understood properties

Cons

- Estimation is difficult
- The number of states is another thing that we have to take into account and estimate.
- Being Markovian we cannot have a large range of correlation structure

Other Models

The list is not at all complete

- Ordered Probit models
- Models based on discretizing a continuous model
- Variants based on other thinning operators
- GLM type of models
- Copula based models: the transition probabilities are defined through the conditional distribution of a copula model.
- DARMA models of the form $X_t = V_t X_{t-1} + (1 - V_t) Z_t$ where V_i are binary variables and Z_i 's are iid discrete variables.

Comments

- Most of the approaches are relatively new in the literature
- For the computational burden usually there are available function in R.
- For sure there are some more models that I did not refer to due to time restrictions.

Repeated measurements

Standard repeated measurements approaches assume a correlation matrix and then applying usually a GEE approach estimate the parameters. Usually correlation parameters are treated as nuisance parameters.

The proposed models can be used to describe such data in a more concrete and model based approach by appropriately defining the correlation structure.

Example: Copula based models for repeated measurements

Concluding

- A wealth of models appropriate for correlated count data have been discussed.
- Most of them extend existing ideas from the well studied multivariate continuous literature.
- Note however that in general not everything can generalize to the discrete case and that correlation issues work with different dynamic with small counts!
- The latter is also a warning for misuse of continuous models to discrete data as "approximations".

Thanks

Special thanks to my colleagues that work together in the above topics:

- Ioannis Ntzoufras (AUER)
- Tom Brijs (Hasselt University, Belgium Universitat Centrum)
- Loukia Meligkotsidou (University of Athens)
- Aris Nikoloulopoulos (AUER)
- Jonas Andersson (NHH, Norway)

and the organizers for inviting me in Whistler, giving me the opportunity to discover a whole new area...

THE END