Multivariate extended Poisson-Tweedie regression model

Prof. Wagner Hugo Bonat

Laboratory of Statistics and Geionformation - LEG

Paraná Federal University - UFPR

Conference on Multivariate Count Analysis
Besançon/France

July/2018

Motivation

- Count data are frequent in many research areas:
  1. Medical research, biology.
  2. Environmental and crop sciences.
  3. Economical, social and political sciences.
  4. Computer science, electronic engineering and etc.
- Statistical challenges:
  1. Overdispersion (mean $>$ variance).
  2. Underdispersion (mean $<$ variance).
  3. Zero-inflation.
  4. Heavy tail.
- Multiple responses each one with its own set of challenges!
- Practical interest:
  1. Estimation of the covariance structure.
  2. Multivariate hypothesis tests (MANOVA-type test).
  3. and much more!!

Review: Univariate approaches

- Models for overdispersion:
    1. Poisson-Tweedie and Hinde-Demétrio (Kokonendji, 2004).
    2. Negative binomial, Poisson-inverse Gaussian, Sichel, Poisson-SGIG, Delaporte, Poisson–Tweedie (Rigby, et.al., 2008).
    3. Normalized tempered stable distribution (Kolossiatis, 2011).
    4. Discrete Linnik distribution (Barabesi, 2016).
- Models for under and overdispersion:
    1. Conway–Maxwell-Poisson (Sellers, 2010).
    2. Gamma-Count (Zeviani, 2014).
    3. Discrete Weibull (Kalktawi, 2015).
- Models for heavy tail or zero-inflation:
    1. Generalised Poisson-inverse Gaussian family (Zhu, 2009).
    2. Hurdle models (Zeileis, 2008).
    3. Zero-inflated and Poisson and negative binomial (Loeys, 2012).
    4. Zero-inflated COM-Poisson (Sellers, 2016).

Review: Multivariate approaches

- Extensions of univariate distributions (David, 2017):
  1. Marginals are Poisson or negative binomial, etc.
  2. Mixture of Poissons or negative binomials, etc.
  3. Conditional Poissons or negative binomials, etc.
- Constructing multivariate distributions:
  1. Multivariate dispersion models (Jørgensen, 2000).
  2. Multivariate exponential dispersion models (Jørgensen, 2012).
  3. Convolution and extended convolution method (Jørgensen, 2013).
- General statistical modelling frameworks:
  1. Multivariate generalized linear mixed models.
  2. Copula based-models.
  3. Multivariate hierarchical generalized linear models.
  4. and . . .

Introduction

- Plethora of distributions/approaches to deal with count data.
- Multivariate probability distribution is not available in closed-form.
- Difficult to fit (problems due to badly behaved likelihood function).
- Difficult to interpret model parameters.
- Difficult to point, with conviction, the best practical choice.
- Demand for a unified model that can automatically adapt to the underlying multivariate count distribution.
- Easy implementation in practice.
- SOLUTION: Multivariate extended Poisson-Tweedie model!

Multivariate extended Poisson-Tweedie model

- Extend the Poisson-Tweedie regression model to deal with multiple response variables.
- Propose an estimating function approach for parameter estimation.
- Propose an extension of the orthodox MANOVA for dealing with multivariate count data.
- Provide R code for fitting the models.
- Illustrative examples to show the flexibility of the multivariate extended Poisson-Tweedie regression model.

Tweedie distribution

- Tweedie distributions (Jørgensen, 1997)

$$f(z;\mu,\phi,p) = a(z,\phi,p)\exp\{(z\psi - k(\psi))/\phi\},$$

  where $\mu = \mathrm{E}(Z) = k'(\psi)$ is the mean.
- $\phi > 0$ and ψ are the dispersion and canonical parameters. $k(\psi)$ is the cumulant function and $a(z,\phi,p)$ is a normalizing constant.
- $\mathrm{Var}(Z) = \phi\mu^p$ where $p \in (-\infty, 0] \cup [1, \infty)$ is the index determining the distribution.
- Special cases: Gaussian ($p = 0$), Poisson ($p = 1$), non-central gamma ($p = 1.5$), gamma ($p = 2$), inverse Gaussian ($p = 3$) and stable distributions ($p > 2$).
- Notation $Z \sim Tw_p(\mu, \phi)$.

Probability mass function

- Hierarchical specification:

$$
\begin{aligned}
Y|Z &\sim Poisson(Z) \\
Z &\sim Tw_p(\mu, \phi).
\end{aligned}
$$

- Probability mass function ($p > 1$)

$$
f(y; \mu, \phi, p) = \int_0^\infty \frac{z^y \exp^{-z}}{y!} a(z, \phi, p) \exp\{(z\psi - k(\psi))/\phi\} dz.
$$

- No closed-form available apart of special case - negative binomial.
- It can be approximated by Monte Carlo integration (difficult and time consuming).
- Distribution, mass, quantile, random generation functions available through the functions `p-`, `d-`, `q-`, `rptweedie()` in R.

Moments and special cases

- Marginal mean and variance are easily obtained

$$\begin{aligned} \mathrm{E}(Y) &= \mu \\ \mathrm{Var}(Y) &= \mu + \phi\mu^p. \end{aligned}$$

- Special cases: Hermite ($p = 0$), Neyman-Type A ($p = 1$), Pólya-Aeppli ($p = 1.5$), negative binomial ($p = 2$) and Poisson inverse-Gaussian ($p = 3$).
- Careful - Hermite is a limit case!
- p is an index that distinguishes between important distributions.
- Parameter space of p is not trivially defined, i.e. $p \in 0 \cup [1, \infty)$.
- Estimation of p works as an automatically model selection.
- Notation $Y \sim PTw_p(\mu, \phi)$.

Shapes - Dispersion index and power parameter

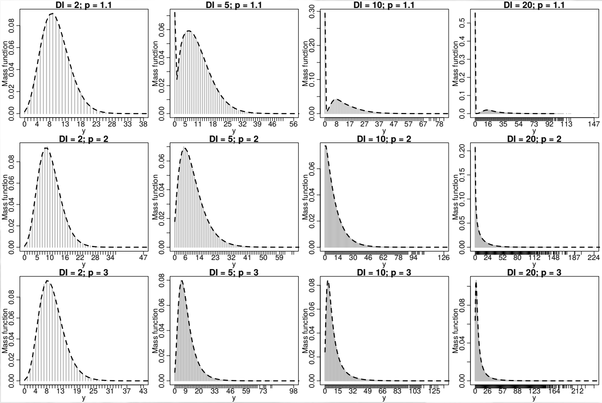


Figure 1. Empirical probability mass function (gray) and approximated probability mass function (black) by dispersion index (DI) and values of the power parameter: Poisson-Tweedie distribution.

Regression models

- Consider $(Y_i, \boldsymbol{x}_i), i = 1, \ldots, n$, where Y_i's are iid rv.
- Full parametric specification:

$$Y_i \sim \mathrm{PTw_p}(\mu_i, \phi).$$

- Second-moment assumptions:

$$\begin{aligned} \mathrm{E}(Y_i) &= \mu_i \\ \mathrm{Var}(Y_i) &= \mu_i + \phi\mu_i^p \end{aligned}$$

  where $g(\mu_i) = \eta_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}$, $\boldsymbol{x}_i$ and $\boldsymbol{\beta}$ are $(Q \times 1)$ vectors of known covariates and unknown regression parameters.
- $\mathrm{Var}(Y_i) > 0$, thus $\phi > -\mu_i^{(1-p)} \Longrightarrow$ under, equi and overdispersion.
- g link function (log link).
- Proposed in Bonat et. al. (2017).

Multivariate regression models

- Consider $(\boldsymbol{Y}_i, \boldsymbol{x}_i), i = 1, \ldots, n$, where $\boldsymbol{Y}_i$'s are i.i.d. random vectors $(R \times 1)$.
- Second-moment assumptions:

$$\begin{aligned} \mathrm{E}(\boldsymbol{Y}_i) &= \boldsymbol{\mu}_i \\ \mathrm{Var}(\boldsymbol{Y}_i) &= \boldsymbol{\Sigma}_i = \mathrm{diag}(\boldsymbol{\mu}_i) + \mathrm{V}(\boldsymbol{\mu}_i; \boldsymbol{p})^{\frac{1}{2}} \boldsymbol{\Omega}(\boldsymbol{\tau}) \mathrm{V}(\boldsymbol{\mu}_i; \boldsymbol{p})^{\frac{1}{2}}, \end{aligned}$$

  where $\boldsymbol{\mu}_i = (g^{-1}(\boldsymbol{x}_i^\top \boldsymbol{\beta}_1), \ldots, g^{-1}(\boldsymbol{x}_i^\top \boldsymbol{\beta}_R))^\top$ is the $R \times 1$ vector of expected values and $\boldsymbol{p}$ the $R \times 1$ vector of power parameters.
- $\mathrm{V}(\boldsymbol{\mu}_i; \boldsymbol{p}) = \mathrm{diag}(\boldsymbol{\mu}_i^{\boldsymbol{p}})$.
- $\boldsymbol{\Omega}(\boldsymbol{\tau})$ controls the part of the covariance structure that does not depend on the expected values.
- Easy interpretation!

Example: Trivariate case

- Second-moments specification

$$\mathrm{E}\left[\begin{pmatrix} Y_{1i} \\ Y_{2i} \\ Y_{3i} \end{pmatrix}\right] = \begin{pmatrix} g(\mu_1) = \mathbf{x}_i^\top \boldsymbol{\beta}_1 \\ g(\mu_2) = \mathbf{x}_i^\top \boldsymbol{\beta}_2 \\ g(\mu_3) = \mathbf{x}_i^\top \boldsymbol{\beta}_3 \end{pmatrix}$$

$$\mathrm{Var}\left[\begin{pmatrix} Y_{1i} \\ Y_{2i} \\ Y_{3i} \end{pmatrix}\right] = \begin{pmatrix} \mu_1 + \mu_1^{p_1}\tau_1 & \sqrt{\mu_1^{p_1}\mu_2^{p_2}}\tau_{12} & \sqrt{\mu_1^{p_1}\mu_3^{p_3}}\tau_{13} \\ \sqrt{\mu_1^{p_1}\mu_2^{p_2}}\tau_{12} & \mu_2 + \mu_2^{p_2}\tau_2 & \sqrt{\mu_2^{p_2}\mu_3^{p_3}}\tau_{23} \\ \sqrt{\mu_1^{p_1}\mu_3^{p_3}}\tau_{13} & \sqrt{\mu_2^{p_2}\mu_3^{p_3}}\tau_{23} & \mu_3 + \mu_3^{p_3}\tau_3 \end{pmatrix}.$$

- Note that

$$\boldsymbol{\Omega}(\boldsymbol{\tau}) = \begin{pmatrix} \tau_1 & \tau_{12} & \tau_{13} \\ \tau_{12} & \tau_2 & \tau_{23} \\ \tau_{13} & \tau_{23} & \tau_3 \end{pmatrix}.$$

Example: Trivariate case

- Standardized dispersion matrix

$$\boldsymbol{R}(\boldsymbol{\tau}) = \begin{pmatrix} 1 & \frac{\tau_{12}}{\sqrt{\tau_1\tau_2}} & \frac{\tau_{13}}{\sqrt{\tau_1\tau_3}} \\ \frac{\tau_{12}}{\sqrt{\tau_1\tau_2}} & 1 & \frac{\tau_{23}}{\sqrt{\tau_2\tau_3}} \\ \frac{\tau_{13}}{\sqrt{\tau_1\tau_3}} & \frac{\tau_{23}}{\sqrt{\tau_2\tau_3}} & 1 \end{pmatrix},$$

  gives us a notion of the correlation between response variables.
- These measures do not depend on the expected values.

Parametrization

- Let $\mathcal{Y} = (\boldsymbol{Y}_1^T, \ldots, \boldsymbol{Y}_n^T)^T$ be the stacked vector ($nR \times 1$) of the outcomes.
- Let $\mathcal{M} = (\boldsymbol{\mu}_1^T, \ldots, \boldsymbol{\mu}_n^T)^T$ be the stacked vector ($nR \times 1$) of the expected values.
- Let $\mathbf{C} = \mathrm{diag}(\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_n)$ denotes a block-diagonal matrix.
- Thus, the $\mathbf{C}$ matrix is symmetric and $nR \times nR$.
- Let $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_R^T)^T$ be a vector $P \times 1$ of regression parameters.
- Let $\boldsymbol{\lambda} = (p_1, \ldots, p_R, \boldsymbol{\tau}^T)^T$ be a $Q \times 1$ vector of dispersion parameters.
- Multivariate extended Poisson-Tweedie model is specified by two sets of parameters $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\lambda})$.
- Quasi-score function for regression parameters.
- Pearson estimating function for dispersion parameters.

Regression parameters

- The quasi-score function is defined by,

$$\psi_{\boldsymbol{\beta}}(\boldsymbol{\beta}, \boldsymbol{\lambda}) = \mathbf{D}^\top \mathbf{C}^{-1}(\mathcal{Y} - \mathcal{M})$$

  where $\mathbf{D} = \nabla_{\boldsymbol{\beta}} \mathcal{M}$ is an $NR \times P$ matrix.
- The $P \times P$ sensitivity matrix of $\psi_{\boldsymbol{\beta}}$ is given by

$$\mathrm{S}_{\boldsymbol{\beta}} = \mathrm{E}(\nabla_{\boldsymbol{\beta}} \psi_{\boldsymbol{\beta}}) = -\mathbf{D}^\top \mathbf{C}^{-1} \mathbf{D}. \quad (1)$$

- The $P \times P$ variability matrix of $\psi_{\boldsymbol{\beta}}$ is given by

$$\mathrm{V}_{\boldsymbol{\beta}} = \mathrm{Var}(\psi_{\boldsymbol{\beta}}) = \mathbf{D}^\top \mathbf{C}^{-1} \mathbf{D}. \quad (2)$$

Dispersion parameters

- The Pearson estimating function is defined by,

$$\psi_{\boldsymbol{\lambda}_i}(\boldsymbol{\beta}, \boldsymbol{\lambda}) = \mathrm{tr}(W_{\boldsymbol{\lambda}i}(\mathbf{r}^\top \mathbf{r} - \mathbf{C}))$$

where $W_{\boldsymbol{\lambda}i} = -\frac{\partial \mathbf{C}^{-1}}{\partial \boldsymbol{\lambda}_i}$ and $\mathbf{r} = (\mathcal{Y} - \mathcal{M})$.

- The entry (i, j) of the $Q \times Q$ sensitivity matrix of $\psi_{\boldsymbol{\lambda}}$ is given by,

$$\mathrm{S}_{\boldsymbol{\lambda}_{ij}} = \mathrm{E}\left(\frac{\partial}{\partial \boldsymbol{\lambda}_i}\psi_{\boldsymbol{\lambda}_j}\right) = -\mathrm{tr}\left(W_{\boldsymbol{\lambda}i}\mathbf{C}W_{\boldsymbol{\lambda}j}\mathbf{C}\right). \tag{3}$$

- The entry (i, j) of the $Q \times Q$ variability matrix of $\psi_{\boldsymbol{\lambda}}$ is given by,

$$\mathrm{V}_{\boldsymbol{\lambda}_{ij}} = \mathrm{Cov}(\psi_{\boldsymbol{\lambda}_i}; \psi_{\boldsymbol{\lambda}_j}) = 2\mathrm{tr}(W_{\boldsymbol{\lambda}i}\mathbf{C}W_{\boldsymbol{\lambda}j}\mathbf{C}) + \sum_{l=1}^{NR} k_l^{(4)}(W_{\boldsymbol{\lambda}i})_{ll}(W_{\boldsymbol{\lambda}j})_{ll} \tag{4}$$

where $k^{(4)}$ denotes the fourth cumulant of $\mathcal{Y}$.

Cross sensitivity and variability matrix

- The entry (i, j) of the $Q \times P$ cross sensitivity matrix between $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}$ is given by,

$$\mathrm{S}_{\boldsymbol{\beta}_i \boldsymbol{\lambda}_j} = \mathrm{E}\left(\frac{\partial}{\partial \boldsymbol{\lambda}_j} \psi_{\beta_i}\right) = 0. \tag{5}$$

- The entry (i, j) of the $P \times Q$ cross sensitivity matrix between $\boldsymbol{\lambda}$ and $\boldsymbol{\beta}$ is given by,

$$\mathrm{S}_{\boldsymbol{\lambda}_i \boldsymbol{\beta}_j} = \mathrm{E}\left(\frac{\partial}{\partial \beta_j} \psi_{\boldsymbol{\lambda}_i}\right) = -\mathrm{tr}\left(W_{\boldsymbol{\lambda} i} \mathbf{C} W_{\boldsymbol{\beta} j} \mathbf{C}\right). \tag{6}$$

Cross sensitivity and variability matrix

- The entry (i, j) of the $P \times Q$ cross variability matrix between $\boldsymbol{\lambda}$ and $\boldsymbol{\beta}$ is given by,

$$\mathrm{V}_{\boldsymbol{\lambda}_i \boldsymbol{\beta}_j} = \mathrm{E}\left(\sum_{k=1}^{NR} \sum_{l=1}^{NR} \sum_{m=1}^{NR} W_{\boldsymbol{\lambda}_i}^{(lm)} A_m^{(j)} r_k r_l r_m \right), \qquad (7)$$

where $A = \mathbf{D}^T \mathbf{C}^{-1}$ and $A^{(j)}$ denotes the j^{th} collumn of A. In a similar way $W_{\boldsymbol{\lambda}_i}^{(lm)}$ denotes the l^{th} and m^{th} entries of the matrix $W_{\boldsymbol{\lambda}_i}$.

Joint sensitivity and variability matrix

- The joint sensitivity matrix of $\psi_{\boldsymbol{\beta}}$ and $\psi_{\boldsymbol{\lambda}}$ is given by

$$\mathrm{S}_{\boldsymbol{\theta}} = \begin{pmatrix} \mathrm{S}_{\boldsymbol{\beta}} & \mathrm{S}_{\boldsymbol{\beta\lambda}} \\ \mathrm{S}_{\boldsymbol{\lambda\beta}} & \mathrm{S}_{\boldsymbol{\lambda}} \end{pmatrix},$$

  whose entries are defined in equations (1), (3), (6) and (5).
- The joint variability matrix of $\psi_{\boldsymbol{\beta}}$ and $\psi_{\boldsymbol{\lambda}}$ is given by

$$\mathrm{V}_{\boldsymbol{\theta}} = \begin{pmatrix} \mathrm{V}_{\boldsymbol{\beta}} & \mathrm{V}_{\boldsymbol{\beta\lambda}} \\ \mathrm{V}_{\boldsymbol{\lambda\beta}} & \mathrm{V}_{\boldsymbol{\lambda}} \end{pmatrix},$$

  whose entries are defined in equations (2), (4) and (7) above.

Godambe information matrix and asymptotic distribution

- Denote $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\lambda}})$ the estimating function estimator of $\boldsymbol{\theta}$.
- The asymptotic distribution of $\hat{\boldsymbol{\theta}}$ is given by

$$\hat{\boldsymbol{\theta}} \sim \mathrm{N}_{P+Q}(\boldsymbol{\theta}, \mathrm{J}_{\boldsymbol{\theta}}^{-1})$$

where $\mathrm{J}_{\boldsymbol{\theta}}^{-1}$ is the inverse of the Godambe information matrix,

$$\mathrm{J}_{\boldsymbol{\theta}}^{-1} = \mathrm{S}_{\boldsymbol{\theta}}^{-1} \mathrm{V}_{\boldsymbol{\theta}} \mathrm{S}_{\boldsymbol{\theta}}^{-T},$$

where $\mathrm{S}_{\boldsymbol{\theta}}^{-T} = (\mathrm{S}_{\boldsymbol{\theta}}^{-1})^T$.

Algorithms

- The *modified chaser*

$$\begin{aligned}\boldsymbol{\beta}^{(i+1)} &= \boldsymbol{\beta}^{(i)} - \mathrm{S}_{\boldsymbol{\beta}}^{-1}\psi_{\boldsymbol{\beta}}(\boldsymbol{\beta}^{(i)}, \boldsymbol{\lambda}^{(i)}) \\ \boldsymbol{\lambda}^{(i+1)} &= \boldsymbol{\lambda}^{(i)} - \alpha\mathrm{S}_{\boldsymbol{\lambda}}^{-1}\psi_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^{(i+1)}, \boldsymbol{\lambda}^{(i)}).\end{aligned}$$

- The *reciprocal likelihood* algorithm

$$\begin{aligned}\boldsymbol{\beta}^{(i+1)} &= \boldsymbol{\beta}^{(i)} - \mathrm{S}_{\boldsymbol{\beta}}^{-1}\psi_{\boldsymbol{\beta}}(\boldsymbol{\beta}^{(i)}, \boldsymbol{\lambda}^{(i)}) \\ \boldsymbol{\lambda}^{(i+1)} &= \boldsymbol{\lambda}^{(i)} - [\alpha\psi_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^{(i+1)}, \boldsymbol{\lambda}^{(i)})^T\psi_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^{(i+1)}, \boldsymbol{\lambda}^{(i)})\mathrm{V}_{\boldsymbol{\lambda}}^{-1}\mathrm{S}_{\boldsymbol{\lambda}} + \mathrm{S}_{\boldsymbol{\lambda}}]^{-1}\psi_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^{(i+1)}, \boldsymbol{\lambda}^{(i)}\end{aligned}$$

where α is a *tuning constant.*

- Easy implementation through the `mcglm` package (Bonat, 2018).
- Special case of a multivariate covariance generalized linear model (Bonat and Jørgensen, 2016).

Multivariate hypothesis test

- The general linear hypothesis may be stated as

$$\mathrm{H}_0 : \boldsymbol{L}\boldsymbol{\beta} = \boldsymbol{0},$$

where $\boldsymbol{L} = \boldsymbol{G} \otimes \boldsymbol{F}$.

- The $\boldsymbol{G}$ $(R \times R)$ states between responses hypotheses.
- The $\boldsymbol{F}$ $(s \times p)$ states between treatments hypotheses.
- We assume equal linear predictor for all responses, thus $\boldsymbol{L}$ is a $(sR \times P)$ matrix with s denoting the number of linear constraints.
- The alternative hypothesis may be stated in the form,

$$\mathrm{H}_1 : \boldsymbol{L}\boldsymbol{\beta} = \boldsymbol{n},$$

where $\boldsymbol{n}$ is not the null vector.

Multivariate hypothesis test

- Wald statistics given by

$$\mathrm{W}_s = (\boldsymbol{L}\boldsymbol{\beta})^\top (\boldsymbol{L}\mathcal{F}_{\boldsymbol{\beta}}^{-1}\boldsymbol{L}^\top)^{-1}(\boldsymbol{L}\boldsymbol{\beta}),$$

  under the null hypothesis is asymptotically chi-squared distributed with sR degrees of freedom.
- Performing test for all response variables as well as between combination of them.
- All possible contrasts between treatment levels (multivariate multiple hypothesis tests).
- In the Gaussian case the Wald test corresponds to the Hotelling-Lawley statistics.

Data set 1: Australian health survey

- Australian health survey for 1987–1988 (Deb and Trivedi, 1997).
- Five count outcomes (Ndoc, Nndoc, Nmed, Nhosp, Nadm).
- Nine covariates concerning social conditions.
- 5190 respondents and no missing data.
- Goals: assess covariate effects (specifically and overall) and correlation between outcomes.

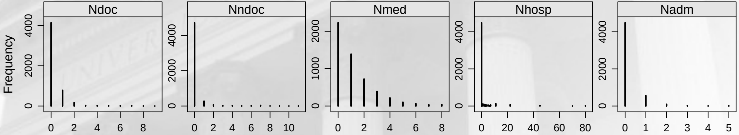


Figure 2. Histograms for each outcome in the Australian health survey data.

Data set 1: Model specification

- Multivariate count regression model

$$\begin{aligned} \mathrm{E}(\boldsymbol{Y}_i) &= \boldsymbol{\mu}_i = \{\exp(\boldsymbol{x}_{i1}^\top \boldsymbol{\beta}_1), \ldots, \exp(\boldsymbol{x}_{i5}^\top \boldsymbol{\beta}_5)\} \\ \mathrm{Var}(\boldsymbol{Y}_i) &= \boldsymbol{\Sigma}_i = \mathrm{diag}(\boldsymbol{\mu}_i) + \mathrm{V}(\boldsymbol{\mu}_i; \boldsymbol{p})^{\frac{1}{2}} \boldsymbol{\Omega}(\boldsymbol{\tau}) \mathrm{V}(\boldsymbol{\mu}_i; \boldsymbol{p})^{\frac{1}{2}} \end{aligned}$$

- Dispersion matrix

$$\boldsymbol{\Omega}(\boldsymbol{\tau}) = \begin{bmatrix} \tau_1 & \tau_{12} & \tau_{13} & \tau_{14} & \tau_{15} \\ \tau_{12} & \tau_2 & \tau_{23} & \tau_{24} & \tau_{25} \\ \tau_{13} & \tau_{23} & \tau_3 & \tau_{34} & \tau_{35} \\ \tau_{14} & \tau_{24} & \tau_{34} & \tau_4 & \tau_{45} \\ \tau_{15} & \tau_{25} & \tau_{35} & \tau_{45} & \tau_5 \end{bmatrix}.$$

- Log link function.
- Poisson-Tweedie dispersion function.

Data set 1: Dispersion structure

- Dispersion and power parameter estimates.

Table 1. Dispersion parameter estimates and standard errors (SE).

	Ndoc	Nndoc	Nmed	Nhosp	Nadm
	Estimate (SE)	Estimate (SE)	Estimate (SE)	Estimate (SE)	Estimate (SE)
$\hat{p}$	1.9042 (0.1121)	1.7495 (0.1309)	1.3153 (0.2442)	1.4184 (0.3702)	1.7595 (0.3179)
$\hat{\tau}_0$	1.2496 (0.2052)	7.7676 (1.3134)	0.2928 (0.0526)	21.3497 (4.4126)	1.1121 (0.5192)

- Nmed and Nhosp present excess of zeros ($\hat{p}$ near 1).
- Weak over-dispersion Ndoc, Nmed and Nadm.
- High over-dispersion Nhosp and Nndoc.
- Neyman-type A and negative binomial.

Data set 1: Multivariate hypothesis tests

- MANOVA-type test for multivariate count data.

Effects	Df	Hotelling-Lawley	Chi-squared	p-value
Intercept	5	0.3689	1914.80	<0.0001
sex	5	0.0322	167.5387	<0.0001
age	5	0.0410	213.1572	<0.0001
Levyplus	5	0.0046	23.9239	0.0002
freepoor	5	0.0012	6.5851	0.2533
freerepa	5	0.0054	28.2852	<0.0001
illness	5	0.1320	685.1295	<0.0001
actdays	5	0.1025	531.9883	<0.0001
hscore	5	0.0080	41.7531	<0.0001

Data set 1: Standardized dispersion matrix

- Standardized dispersion matrix (lower) and standard errors (upper).

$$\hat{R} = \begin{bmatrix} - & 0.0141 & 0.01420 & 0.0152 & 0.0142 \\ 0.0426 & - & 0.0145 & 0.0159 & 0.0144 \\ 0.1198 & 0.0788 & - & 0.0154 & 0.0140 \\ 0.0609 & 0.0680 & 0.0664 & - & 0.055 \\ 0.0893 & 0.0619 & 0.0690 & 0.5142 & \end{bmatrix}$$

Data set 2: Ant data

- Abundances of 20 epigaeic ant species across 30 sites in south-easthern Australia.
- Covariates: Percent cover of bare ground and shrub.
- Goal: assess the covariates overall effects and describe the relation between species abundances.

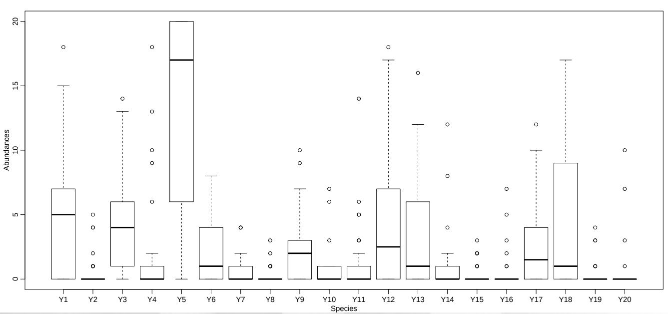


Figure 3. Boxplots of abundances by species.

Data set 2: Model specification

- Linear predictor for each response variable

$$y_i \sim Bare.ground + Shrub.cover.$$

- 60 regression parameters.
- 20 dispersion parameters.
- 190 correlation parameters.
- Total number of parameters: 270.

Data set 2: R code

- R code

```r
# Loading extra packages
require(mcglm)
require(mvabund)
require(Matrix)
require(corrplot)

# Loading data set
data(antTraits)
y <- antTraits$abund[,11:30] # Selecting response variables
names(y) <- paste0("y",1:20)
X <- antTraits$env[,c(3,5)] # Selecting covariates
data <- data.frame(y, X)

# Linear predictor
lp <- paste(names(y), "~", "Shrub.cover + Feral.mammal.dung")
form <- lapply(lp, formula)

# Matrix linear predictor
Z0 <- mc_id(data)
```

- Fitting the model to data.

```r
# Fitting multivariate model
fit1 <- mcglm(linear_pred = c(form), matrix_pred = rep(list(Z0), 20),
              link = rep("log", 20), variance = rep("poisson_tweedie", 20),
              control_algorithm = list(max_iter = 100),
              data = data)
```

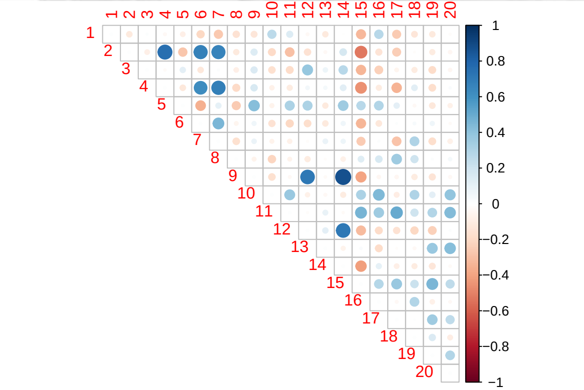
Data set 2: Results

- Multivariate hypothesis test.

```
manova.mcglm(fit1)
```

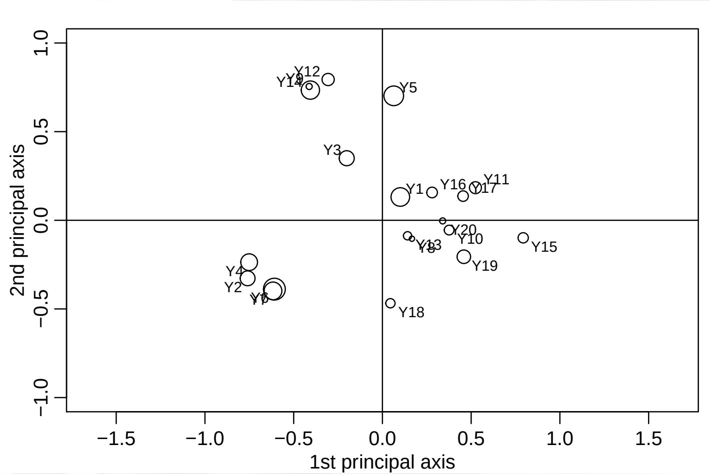
```
##       Effects Df Hotelling.Lawley Qui.square p_value
## 1   Intercept 20        60.876617 1826.29851   0e+00
## 2 Bare.ground 20         2.110956   63.32867   2e-06
## 3 Shrub.cover 20         2.636528   79.09584   0e+00
```

- Standardized dispersion matrix.



Data set 2: Biplot

Biplot: Points proportional to logarithm of body length.



Main results

- Flexible multivariate statistical model to deal with count data.
- Second-moment assumptions.
- General framework for estimation based on estimating function.
- Efficient algorithms for estimation.
- Asymptotic theory (Godambe Information).
- General software implementation in `R` (`mcglm` package).
- Extension of orthodox MANOVA to deal with count data.

Topics for research

- MANOVA-type tests
  1. Hotelling-Lawley statistics $\rightarrow$ Wald test.
  2. Pillai statistics $\rightarrow$ Score test.
  3. Wilks statistics $\rightarrow$ likelihood ratio test.
- Can we have analogous to
  1. principal components,
  2. factor analysis,
  3. correspondence analysis,
  4. canonical correlation analysis and
  5. Redundancy analysis

  in the context of count data?
- Perhaps, all of them based on the standardized dispersion matrix. Does it make sense?
- How to deal with high-dimensional data in both outcomes and covariates?
- How to deal with missing and/or censored data?

Contact

- Work in progress ...
- Name: Wagner Hugo Bonat
- e-mail: wbonat@ufpr.br
- Webpages
  1. https://cran.r-project.org/web/packages/mcglm
  2. https://github.com/wbonat/mcglm
  3. www.leg.ufpr.br/papercompanions
  4. http://www.leg.ufpr.br/∼wagner
- Thank you !