Dealing with overdispersion in multivariate count data

Noemi Corsini & Cinzia Viroli

Statistical methods and models for complex data — Poster Session: 21 September 2022

Framework and motivation

The **overdispersion** or extra-variation is a recurring phenomenon when dealing with **counts and categorical data**.

↪ When fitting a binomial, a multinomial or a Poisson model if the data exhibit a larger variability than that the model is able to explain, the sampling variation will be greater than the estimated variation accounted by the model.

Overdispersion has specific causes and consequences

- **Causes** may be the result of data aggregation such as clumped sampling, the correlation between individual responses or the additional experimental variability.
- **Inferential consequences** are imprecise estimates and biased standard errors that make model selection, interpretability and prediction unreliable.

Our focus is on **multivariate count data**, whose natural probabilistic model is the multinomial distribution. However, in presence of overdispersion, this model may lead to a nominal variance well below the empirical one.

How to cope with overdispersion?

- **Quasi-likelihood approach** where second order hypotheses allow for relaxing constraints on the variance structure;
- **Likelihood-based approach** exploiting distributions that differ in the variance and correlation structure, that can be
  - **Negative** Multinomial (MN), Dirichlet-Multinomial (DM), Random Clumped Multinomial (RCM);
  - **Positive** Negative Multinomial (NM);
  - **General** Generalized Dirichlet Multinomial (GDM).

Main objectives

1. **Comprehensive comparison** of probabilistic models that capture extra-variation in multivariate count data.
2. Introduction of a **new model** that extends the Dirichlet-Multinomial in a deep fashion.
3. Analyses of **empirical performances and properties** through a broad simulation study.

Deep Dirichlet Multinomial

In order to deal with overdispersion, we propose a new model called **Deep Dirichlet-Multinomial (DDM)**. Derived from [2], it is a mixture of Dirichlet-Multinomial distributions with restrictions on the parameters.

Let $\mathrm{DM}(\boldsymbol{\theta}, m)$ be the probability mass function of a Dirichlet Multinomial distribution with parameters $\boldsymbol{\theta}$ and size m, then the probability distribution of the DDM model is defined as

$$\mathbb{P}(\boldsymbol{Y} = \boldsymbol{y}) = \sum_{k=1}^{K} \omega_k \, \mathrm{DM}(\boldsymbol{\beta}(1 + \boldsymbol{\alpha_k}), m)$$

where ω_k, for $k = 1, \ldots, K$, are the weights of the mixture defined such that they satisfy $0 < \omega_k < 1$ and $\sum_{k=1}^{K} \omega_k = 1$ while $\boldsymbol{\beta} > 0$ is a vector of length p.

Each $\boldsymbol{\alpha}_k \in (-1, 1)$ is a p-dimensional vector that can be interpreted as **perturbation parameter** used to adjust β and **get a more flexible model** that behaves better in case of overdispersion.

- Positive values lead to a larger effect of $\theta_j = \beta_j(1 + \alpha_j)$;
- Negative values show which categories have more zeros.

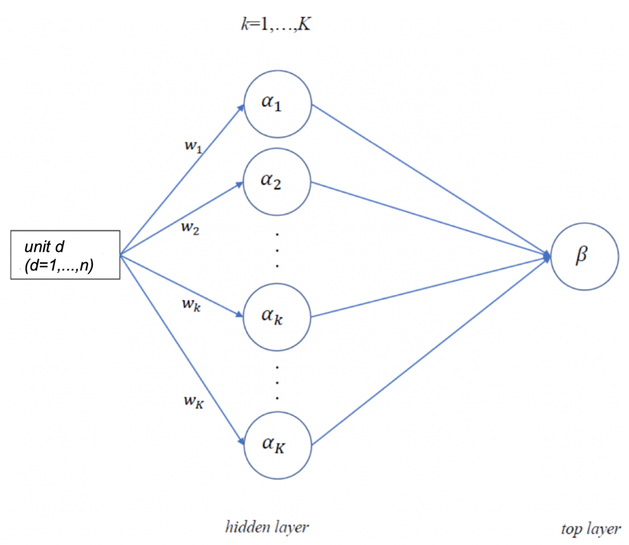

Figure 1. Structure of DDM model. A hidden layer of nodes is introduced to better capture the overdispersion.

Let $\boldsymbol{\theta}_k = \boldsymbol{\beta}(1 + \boldsymbol{\alpha_k})$, $\theta_{0k} = \sum_{j=1}^{p} \beta_j(1 + \alpha_{jk})$, $\boldsymbol{\pi}_k = \frac{\boldsymbol{\theta}_k}{\theta_{0k}}$ and $\rho_k^2 = 1/(1 + \theta_{k0})$, then we can define expectation and variance as:

$$\mathbb{E}[\boldsymbol{Y}] = \sum_{k=1}^{K} \omega_k m \boldsymbol{\pi}_k,$$

$$\mathbb{V}[\boldsymbol{Y}] = \sum_{k=1}^{K} \omega_k m \{diag(\boldsymbol{\pi}_k) - \boldsymbol{\pi}_k \boldsymbol{\pi}_k'\}(1 + \rho_k^2(m-1)) + \sum_{k=1}^{K} \omega_k m^2 \boldsymbol{\pi}_k \boldsymbol{\pi}_k' - m^2 \left(\sum_{k=1}^{K} \omega_k \boldsymbol{\pi}_k\right)\left(\sum_{k=1}^{K} \omega_k \boldsymbol{\pi}_k\right)^{\top}.$$

The variance can be split into two components:

1. A weighted sum of **within variances**;
2. A **between variance** part that captures both over and underdispersion.

This model has a **flexible correlation structure among variables** that comes at the price of an high number of parameters to be estimated which largely increase with K.

The model parameters are estimated through a generalized EM algorithm with a quasi-Newton optimization step.

Simulation study

The capabilities of the proposed model are compared to the likelihood-based models in **two simulation studies** that differ in the way the overdispersion is introduced.

1. Randomly add zeros into the data;
2. Gradually add zeros by replacing the smallest counts starting from cells with frequency one.

Different scenarios are simulated from a $\mathrm{MN}(\pi, m = 100)$, starting from the case of lack of extra-variation (0% of zeros added), to the case of 90% of zeros added.

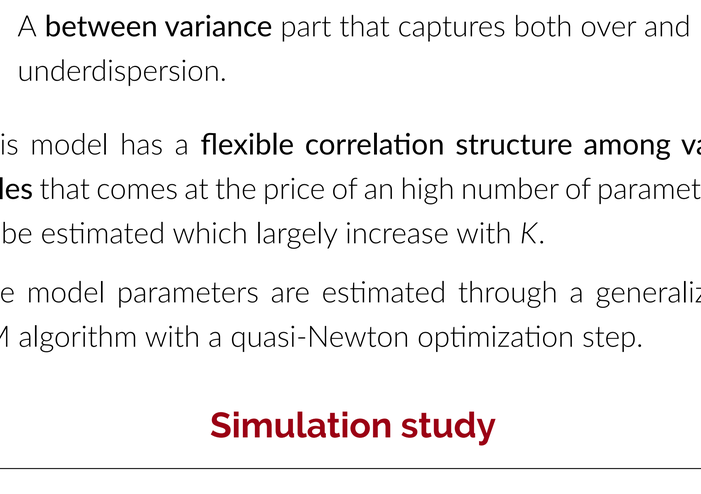

Figure 2. Average euclidean distances between the empirical and estimated variances in the two simulations.

As the mixture components increase from $K = 2$ to $K = 20$, the **DDM model** has a smaller and smaller euclidean distance.

In both scenarios it is evident that the Deep Dirichlet-Multinomial distribution better and better approximates the empirical true variance as K increases.

The price of flexibility is a greater computational burden

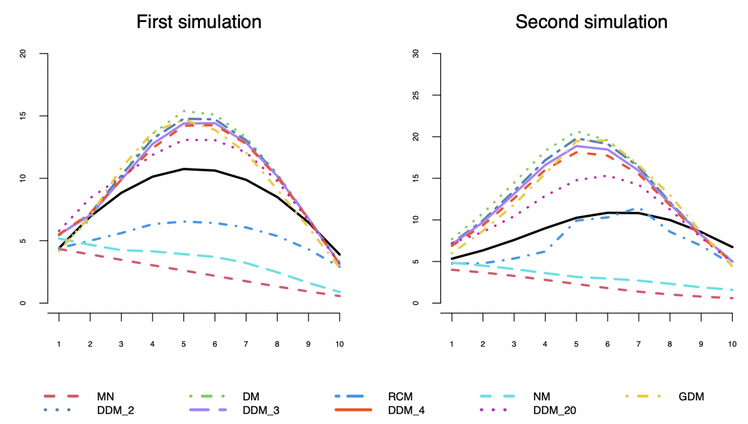

Figure 3. Evolution of the empirical true variance - solid black line - with respect to the estimated variances of the different models as the number of zeros in the dataset increases.

DDM asymptotic behavior

Exploiting the same data generating process, we analyze the **dynamic behavior** of the DDM variance as function of K in the case of maximum overdispersion.

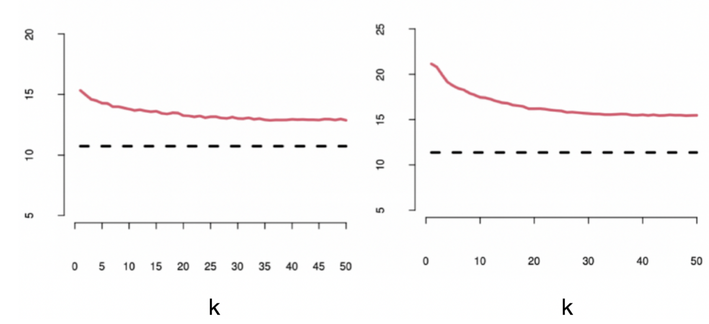

Figure 4. The red line describes the evolution of the fitted DDM variance when K increases to $+\infty$. The dashed black line is the sample variance.

The empirical analysis suggests that the **estimated variance tends to the computed variance** when the number of elements K of the mixture goes to $+\infty$.

Conclusion

- A new approach, the Deep Dirichlet Multinomial, is proposed. Its variance shows desirable properties
  - ⟶ It can ideally be split in within and between variance;
  - ⟶ It tends to the computed one when $K \to \infty$.
- Simulations show that the new model deals better with overdispersed data, compare to other likelihood based solutions.

Further readings

[1] Noemi Corsini and Cinzia Viroli. "Dealing with overdispersion in multivariate count data". In: *Computational Statistics & Data Analysis* 170 (2022), p. 107447.

[2] Cinzia Viroli and Laura Anderlucci. "Deep mixtures of unigrams for uncovering topics in textual data". In: *Statistics and Computing* 31.3 (2021), pp. 1–10.

Contact information

Noemi Corsini, Ph.D. student
University of Padua
noemi.corsini@phd.unipd.it
https://noecors.github.io

SCAN ME