Zero-inflation in the Poisson lognormal family

Analyzing high-dimensional count data is a challenge and statistical model-based approaches provide an adequate and efficient framework that preserves explainability. The (multivariate) Poisson-Log-Normal (PLN) model is one such model: it assumes count data are driven by an underlying structured latent Gaussian variable, so that the dependencies between counts solely stems from the latent dependencies. However PLN doesn't account for zero-inflation, a feature frequently observed in real-world datasets. Here we introduce the Zero-Inflated PLN (ZIPLN) model, adding a multivariate zero-inflated component to the model, as an additional Bernoulli latent variable. The Zero-Inflation can be fixed, site-specific, feature-specific or depends on covariates. We estimate model parameters using variational inference that scales up to datasets with a few thousands variables and compare two approximations: (i) independent Gaussian and Bernoulli variational distributions or (ii) Gaussian variational distribution conditioned on the Bernoulli one. The method is assessed on synthetic data and the efficiency of ZIPLN is established even when zero-inflation concerns up to 90% of the observed counts. We then apply both ZIPLN and PLN to a cow microbiome dataset, containing 90.6% of zeroes. Accounting for zero-inflation significantly increases log-likelihood and reduces dispersion in the latent space, thus leading to improved group discrimination.

QQs liens (preprint, code R/C++ et Python)
https://arxiv.org/abs/2405.14711
https://github.com/PLN-team/PLNmodels
https://bbatardiere.pages.mia.inra.fr/pyplnmodels