M2 Internship : improving inference of Poisson log-normal models with normalizing flows

Contacts:
Nicolas Jouvin - nicolas.jouvin@inrae.fr
Jean-Benoist Leger - jbleger@hds.utc.fr

To apply send CV and motivation letter with subject 'Application for PLN internship'.

1   Context

Multivariate count data arise in many modern application in biology or ecology. Recently, Chiquet et al. (2018) proposed a family of latent variable models based on the Poisson log-normal (PLN) distribution modeling dependencies via a latent Gaussian structure. These models are applicable to various tasks in ecology or biology such as clustering, dimension reduction or network inference (Chiquet et al. 2021).

2   Improving inference of PLN models with Normalizing Flows

Estimation procedure in PLN models relies on variational inference (Blei et al. 2017), where the main bottleneck is the specification of the variational family. Recently, Rezende et al. (2015) proposed to use Normalizing Flows, a family of invertible neural networks, to enrich the variational family while keeping computations efficient.

A key observation in PLN models is that the posterior distribution belongs to the so-called *mean-field* (fully factorized) variational family. Hence, the only approximation made by mean-field variational inference is the parametric choice of the variational family for the posterior. The current Gaussian assumption may be too reductive, and refining the variational family may help reach better estimates of the model parameters and their uncertainty, which can be crucial in applications.

3   Goal of this internship

This internship aims at implementing normalizing flows in the context of PLN variational inference, and comparing with the standard variational inference approach. Precisely, the comparison will leverage on numerical simulations to investigate

- the achieved evidence lower bound,

- estimation error on the model parameters,

- sensitivity and efficiency with respect to the number of samples, the chosen flow, and other hyper-parameters.

- generalization to distributions in the exponential family.

Implementation will be made in **Python**, within the `pytorch` ecosystem leveraging on existing code in the `pyplnmodels` package. Possible applications on single-cell RNA-seq with the PLN-PCA model.

4 Profile & environment

The candidate should be 2nd year master student in Statistics/Machine Learning, with courses on latent variable modeling, EM algorithm, and scientific programming in Python.

- Location : UMR MIA Paris-Saclay, Palaiseau Campus, 22 place de l'agronomie, 91120

- Supervision : Nicolas Jouvin & Jean-Benoist Léger

- Start: Flexible, starting in February or after.

- Duration : 5-6 months

The candidate will have an office, and benefit from the work environment of the Solstis team, with many PhD. students & postdocs working on statistical modeling and machine learning for the life sciences.

References

Chiquet, Julien, Mahendra Mariadassou, and Stéphane Robin (2018). "Variational inference for probabilistic Poisson PCA". In: *The Annals of Applied Statistics* 12.4, pp. 2674–2698.

— (2021). "The Poisson-lognormal model as a versatile framework for the joint analysis of species abundances". In: *Frontiers in Ecology and Evolution* 9, p. 588292.

Blei, David M, Alp Kucukelbir, and Jon D McAuliffe (2017). "Variational inference: A review for statisticians". In: *Journal of the American statistical Association* 112.518, pp. 859–877.

Rezende, Danilo and Shakir Mohamed (2015). "Variational inference with normalizing flows". In: *International conference on machine learning*. PMLR, pp. 1530–1538.