Deconvolution approaches for count data in the context of comparative genomics

Sophie Abby
TIMC, CNRS

Nelle Varoquaux
TIMC, CNRS

November 9, 2022

Internship subject

This internship tackles the problem of decomposing count data using matrix factorization. Count data are encountered in a diversity of research fields such as ecology, biology, psychology. A common question is how to decompose such count data in a lower dimension that encapsulate scientific knowledge. For example, can we decompose a matrix X where each row corresponds to an organism, each column to a protein, and each entry to the number of time each protein is found into each genome, into two matrices H and W, one that summarizes which protein belongs to a protein complex (*e.g.*, proteins involved in the same biological function) and the other summarizing the presence/absence of these protein complexes in each organism.

A popular approach for decomposing such matrices are matrix factorization approaches. These methods consists in decomposing X as the product of two lower dimensional matrices H and W, under some constraints (such as non-negativity). Many distributions can be used to account for specific properties of the data at hand. For example, Poisson generative models are well suited to model count data. Yet, there are still many technological hurdles to be taken, and many methodological aspects to be addressed for pratical applications (*e.g.*, hyperparameter tuning, inclusion of prior knowledge through penalization approaches, ...)

On the application side, we have access to a large database of bacterial and archael genomes, where each protein has been annotated as potentially being part of specific protein complexes (through sequence similarity). We also have groundtruth information on the presence / absence of protein complexes, thus making this an ideal dataset to develop novel methods for real-world problems of major importance.

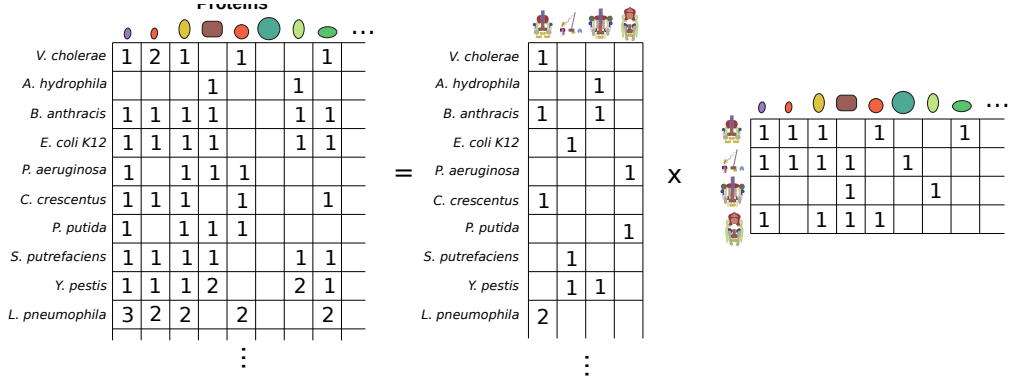

Figure 1: **A schema representing an NMF approach to tackle the problem at hand.** The goal is to decompose a matrix X (where each row corresponds to an organism, each column a protein known to be involved in a secretion system, and each entry the number of times this protein is found in this genome) into two matrix V and U to automatically detect the number of each time of secretion systems in each genome.

Goals include:

1. perform a literature review on this topic, specifically for NMF approaches on count data [1, 2];
2. test some algorithms on our dataset to understand the limits of NMF approaches;
3. investigate how to tune the hyperparameters of the model using prior knowledge and/or stability analysis;
4. investigate better feature engineering to consider each protein's genomic context (*e.g.*, the gene organization along the genome): Taking inspiration from nature language processing, we will extract n-grams of proteins and investigate how to add **structured penalization** to the deconvolution.

Candidate profile

The candidate should be a master (1st or 2nd year) student or equivalent (engineering school) in applied mathematics, computer science, or bioinformatics, with a interest in biological applications. The candidate should have strong communication skills.

The application file should contain:

- Cover letter
- CV

Applications

Applications are to be addressed to Nelle Varoquaux: `nelle.varoquaux@univ-grenoble-alpes.fr`

References

[1] Daniel Lee and H. Sebastian Seung. Algorithms for non-negative matrix factorization. In T. Leen, T. Dietterich, and V. Tresp, editors, *Adv. Neural Inf. Process. Syst.*, volume 13. MIT Press, 2001.

[2] Cédric Févotte and A. Taylan Cemgil. Nonnegative matrix factorizations as probabilistic inference in composite models. In *In In Proc. 17th European Signal Processing Conference (EUSIPCO-09)*, 2009.