ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

D-BSSE
Department of Biosystems
Science and Engineering

Statistical Models in Computational Biology

Jack Kuipers
David Dreifuss
Xiang Ge Luo
Rudolf Schill

Due 25th of May 2023
Please submit your project with the filename `Lastname(s)_Project11.pdf`.

Problem 31: Uniqueness of NMF solutions (2 points)

Given a non-negative matrix factorization

$$V \approx WH \quad \text{with} \quad V \in \mathbb{R}_{\geq 0}^{M\times N}, W \in \mathbb{R}_{\geq 0}^{M\times K}, H \in \mathbb{R}_{\geq 0}^{K\times N},$$

are the factor matrices unique? If they are not, provide alternative factor matrices $\tilde{W} \neq W$ and $\tilde{H} \neq H$ such that

$$\tilde{W}\tilde{H} = WH \quad \text{and} \quad \tilde{W} \in \mathbb{R}_{\geq 0}^{M\times K}, \tilde{H} \in \mathbb{R}_{\geq 0}^{K\times N}. \tag{1}$$

If they are unique, provide a proof that eq(1) necessarily implies $\tilde{W} = W$ and $\tilde{H} = H$.

Problem 32: NMF of spatial gene expression patterns (3 points)

The file `DrosophilaExpressions.rda` contains a data matrix V derived from early stage embryos of the fruit fly *Drosophila melanogaster* and a function `imageBatchDisplay` for its visualization[1]. Each column of V is an observation of a different gene in terms of its spatial expression pattern. Each row of V is a feature that represents a different location (pixel) in the embryo.

Install and load the package `NMF` from CRAN to perform the following analyses.

1. Display the first 16 observations using `imageBatchDisplay(V[,1:16])` as elliptical images that resemble the embryo. Compute a factorization $V \approx \hat{V} = WH$ with `rank=15`, `seed=123` and the default `method="brunet"`. Report the generalized KL divergence[2] of the approximated data matrix $\hat{V}$ and display its first 16 columns as images. Then display all computed basis patterns, i.e. columns of W. Repeat these steps for a rank 10 factorization.

   (1.5 points)

2. For the rank 10 factorization, display the expression pattern of the gene `Mkp3` approximately as a linear combination of the computed basis patterns. To this end, pass the corresponding coefficients as the argument `imgNames` to `imageBatchDisplay` in order to print the coefficients below the basis patterns. Then do the same for the gene `CG31909`.

   (1.5 points)

[1] the data and code are taken from the Berkeley Drosophila Genome Project with slight modifications
[2] The `NMF` package reports the KL divergence misleadingly as “residuals”

Problem 33: Implementing NMF from scratch **(5 points)**

Implement your own NMF algorithm in R and test it by computing a rank 10 factorization WH of the matrix V from `DrosophilaExpressions.rda`. To this end, set a seed for reproducibility and initialize the entries of W and H with random numbers drawn uniformly between 0 and 1. Then iteratively optimize the generalized KL divergence using the multiplicative update rules

$$H_{kj} \leftarrow H_{kj} \frac{\sum_i W_{ik} V_{ij}/(WH)_{ij}}{\sum_i W_{ik}} \quad \text{and} \quad W_{ik} \leftarrow W_{ik} \frac{\sum_j H_{kj} V_{ij}/(WH)_{ij}}{\sum_j H_{kj}}.$$

Report the generalized KL divergence[3] after 1000 iterations, where one iteration updates both matrices.

Hint: Rather than manipulating individual matrix entries according to the update formulae, it is clearer and more efficient to use higher level operations in R. These include the matrix product `A%*%B`, transpose `t(A)`, `rowSums(A)`, `colSums(A)` and the elementwise product `A*B`. The operation `A*v` multiplies each column of `A` elementwise with a vector `v`. Elementwise division is analogous.

If your implementation turns out too slow, you may use fewer iterations or a smaller rank.

[3] when some entries of V are exactly zero, use the convention $x \log(x/0) = 0$ to compute the generalized KL divergence