




# OnlineNMF.jl: A Julia Package for Out-of-core and Sparse Non-negative Matrix Factorization

Koki Tsuyuzaki <sup>1,2</sup>

<sup>1</sup> Department of Artificial Intelligence Medicine, Graduate School of Medicine, Chiba University, Japan <sup>2</sup> Laboratory for Bioinformatics Research, RIKEN Center for Biosystems Dynamics Research, Japan

DOI: [10.xxxxxx/draft](https://doi.org/10.xxxxxx/draft)

## Software

- [Review](#) 
- [Repository](#) 
- [Archive](#) 

Editor: [Chris Vernon](#) 

## Reviewers:

- [@rahulkhorana](#)
- [@ferchaure](#)

Submitted: 22 October 2025

Published: unpublished

## License

Authors of papers retain copyright<sup>®</sup> and release the work under a Creative Commons Attribution 4.0 International License ([CC BY 4.0](#))<sup>®</sup>.

## Summary

Non-negative Matrix Factorization (NMF) is a widely used dimensionality reduction technique for identifying a small number of non-negative components that minimize the reconstruction error when applied to high-dimensional data (Meng, 2016; Stein-O'Brien, 2018). NMF has been applied across various fields of data science, including face recognition (Lee, 1999), audio signal processing (Kameoka, 2015), recommender system (Sajad, 2025), natural language processing (also known as a “topic model”) (Srivastava & Sahami, 2009), population genetics (also known as “admixture analysis”) (Simanovsky, 2019), and omics studies Rodriques (2019).

Despite its broad applicability, NMF becomes computationally prohibitive for large data matrices, making it difficult to apply in practice. In particular, recent advances in single-cell omics have led to datasets with millions of cells, for which standard NMF implementations often fail to scale. To meet this requirement, I originally developed `OnlineNMF.jl`, which is a Julia package to perform some NMF algorithms (<https://github.com/rikenbit/OnlineNMF.jl>).

## Statement of need

NMF is a workhorse algorithm for most data science tasks. However, as the size of the data matrix increases, it often becomes too large to fit into memory. In such cases, an out-of-core (OOC) implementation — where only subsets of data stored on disk are loaded into memory for computation — is desirable. Additionally, representing the data in a sparse matrix format, where only non-zero values and their coordinates are stored, is computationally advantageous. Therefore, a NMF implementation that supports both OOC computation and sparse data handling is highly desirable.

Similar discussions have been made in the context of Principal Component Analysis (PCA), and we have independently developed a Julia package, `OnlinePCA.jl` (Tsuyuzaki, 2020). `OnlineNMF.jl` is a spin-off version of `OnlinePCA.jl`, implementing NMF.

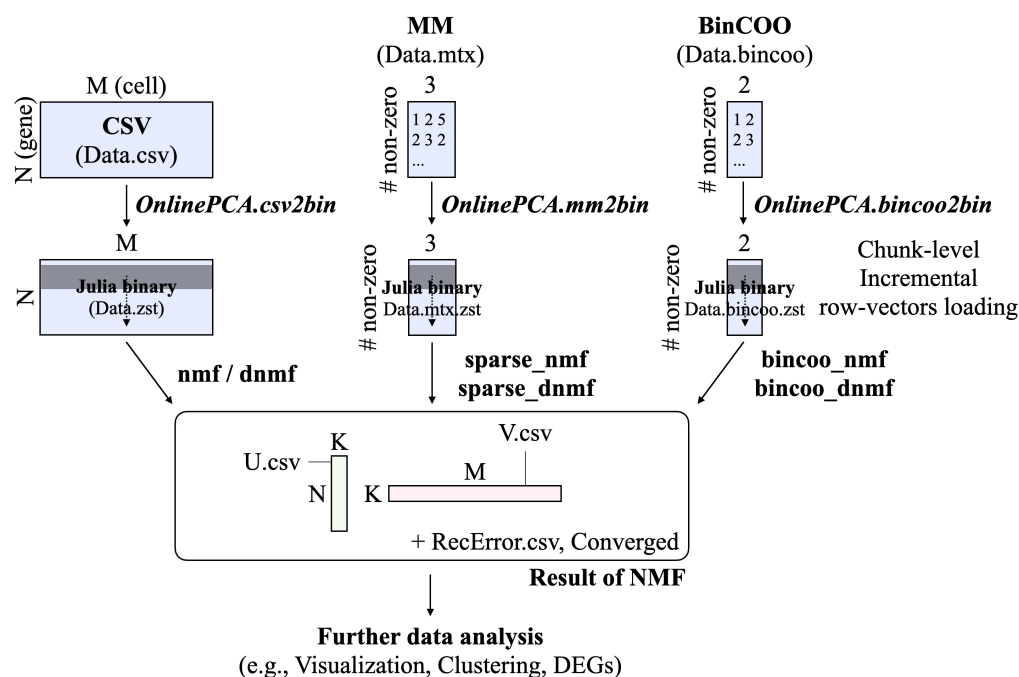


Figure 1: Overview of workflow in OnlineNMF.jl.

## Example

NMF can be easily reproduced on any machine where Julia is pre-installed by using the following commands in the Julia REPL window:

## Installation

First, install OnlinePCA.jl and OnlineNMF.jl from the official Julia package registry or directly from GitHub:

```
# Install OnlinePCA.jl and OnlineNMF.jl from Julia General
julia> Pkg.add("OnlinePCA")
julia> Pkg.add("OnlineNMF")

# or GitHub for the latest version
julia> Pkg.add(url="https://github.com/rikenbit/OnlinePCA.jl.git")
julia> Pkg.add(url="https://github.com/rikenbit/OnlineNMF.jl.git")
```

## Preprocess of CSV

Then, write a synthetic data as a CSV file, convert it to a compressed binary format using Zstandard, and prepare summary statistics for PCA. MM format is also supported for sparse matrices.

```
using OnlinePCA
using OnlinePCA: write_csv
using OnlineNMF
using Distributions
using DelimitedFiles
using SparseArrays
using MatrixMarket
```

```
# CSV
tmp = mktempdir()
data = rand(Binomial(10, 0.05), 300, 99)
data[1:50, 1:33] .= 100*data[1:50, 1:33]
data[51:100, 34:66] .= 100*data[51:100, 34:66]
data[101:150, 67:99] .= 100*data[101:150, 67:99]
write_csv(joinpath(tmp, "Data.csv"), data)

# Matrix Market (MM)
mmwrite(joinpath(tmp, "Data.mtx"), sparse(data))

# Binarization (Zstandard)
csv2bin(csvfile=joinpath(tmp, "Data.csv"),
        binfile=joinpath(tmp, "Data.zst"))

# Sparsification (Zstandard + MM format)
mm2bin(mmfile=joinpath(tmp, "Data.mtx"),
        binfile=joinpath(tmp, "Data.mtx.zst"))
```

#### 40 **Setting for plot**

41 Define a helper function to visualize the results of NMF using the PlotlyJS.jl package. It  
42 generates two subplots: Component-1 vs Component-2 and Component-2 vs Component-3,  
43 with color-coded groups.

```
using DataFrames
using PlotlyJS

function subplots(resnmf, group)
    # data frame
    data_left = DataFrame(pc1=resnmf[:,1], pc2=resnmf[:,2], group=group)
    data_right = DataFrame(pc2=resnmf[:,2], pc3=resnmf[:,3], group=group)
    # plot
    p_left = Plot(data_left, x=:nmf1, y=:nmf2, mode="markers",
                  marker_size=10, group=:group)
    p_right = Plot(data_right, x=:nmf2, y=:nmf3, mode="markers",
                  marker_size=10,
                  group=:group, showlegend=false)
    p_left.data[1]["marker_color"] = "red"
    p_left.data[2]["marker_color"] = "blue"
    p_left.data[3]["marker_color"] = "green"
    p_right.data[1]["marker_color"] = "red"
    p_right.data[2]["marker_color"] = "blue"
    p_right.data[3]["marker_color"] = "green"
    p_left.data[1]["name"] = "group1"
    p_left.data[2]["name"] = "group2"
    p_left.data[3]["name"] = "group3"
    p_left.layout["title"] = "PC1 vs PC2"
    p_right.layout["title"] = "PC2 vs PC3"
    p_left.layout["xaxis_title"] = "pc1"
    p_left.layout["yaxis_title"] = "pc2"
    p_right.layout["xaxis_title"] = "pc2"
    p_right.layout["yaxis_title"] = "pc3"
    plot([p_left p_right])
```

end

```
group=vcat(repeat(["group1"],inner=33), repeat(["group2"],inner=33),
           repeat(["group3"],inner=33))
```

#### 44 NMF based on Alpha-Divergence

45 This example demonstrates NMF using the  $\alpha$ -divergence as the loss function. By setting  
46  $\alpha=2$ , the objective corresponds to the Pearson divergence. The input data is assumed to  
47 be a dense matrix compressed with Zstandard (.zst format).

```
out_nmf_alpha = nmf(input=joinpath(tmp, "Data.zst"),
                    dim=3, alpha=2, numepoch=30, algorithm="alpha")
```

```
subplots(out_nmf_alpha, group)
```

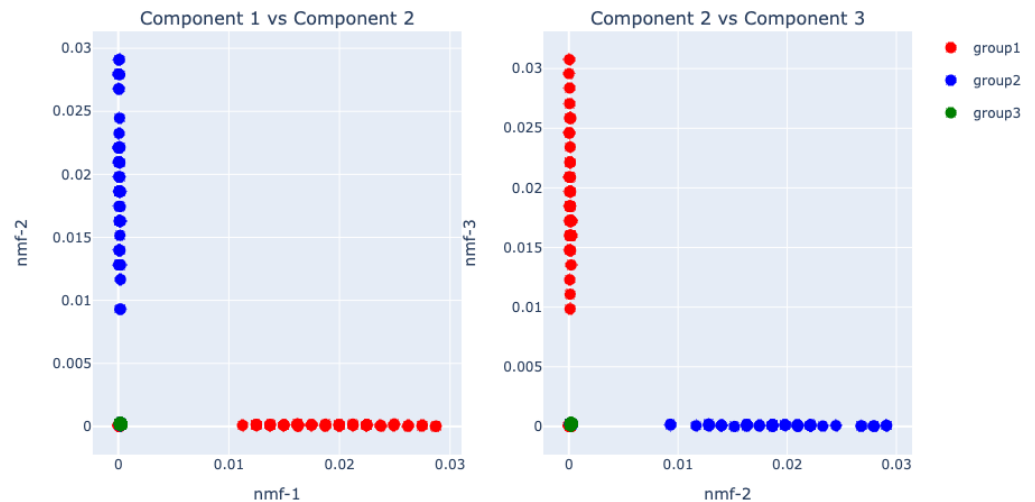


Figure 2: Output of nmf against binarized CSV format.

#### 48 Sparse-NMF based on Beta-Divergence

49 This example performs NMF on a sparse matrix using the  $\beta$ -divergence. The input is a  
50 MM formatted sparse matrix file (.mtx.zst). When  $\beta=1$ , the loss corresponds to the  
51 Kullback-Leibler divergence, and sparse-specific optimization is used internally.

```
out_sparse_nmf_beta = sparse_nmf(input=joinpath(tmp, "Data.mtx.zst"),
                                 dim=3, beta=1, numepoch=30, algorithm="beta")
```

```
subplots(out_sparse_nmf_beta, group)
```

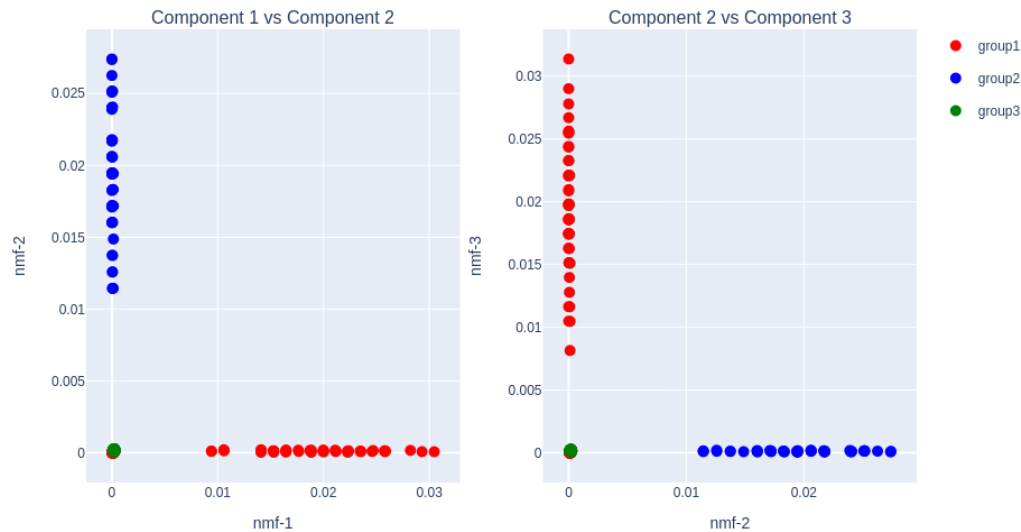


Figure 3: Output of sparse\_nmf against binarized MM format.

Related work

There are various implementations of NMF (Boureima, 2024; Pedregosa, 2011; Tsuyuzaki, 2023) and some of them are OOC-type or sparse-type (Lab, 2023; Pedregosa, 2011) but OnlineNMF.jl is the only tool that supports both OOC computation and sparse data formats (e.g., MM, BinCOO).

Function Name	Language	OOC	Sparse Format
nnTensor::NMF	R	No	-
sklearn.decomposition.NMF	Python	No	-
pyDNMFk	Python	No	-
NMF.MultUpdate	Julia	No	-
sklearn.decomposition.MinibatchNMF	Python	Yes	-
RcppPlanc/PLANC	R/C++	Yes	dgCMatrix

References

Boureima, I. et al. (2024). Distributed out-of-memory NMF on CPU/GPU architectures. *J Supercomput*, 80, 3970–3999. <https://doi.org/10.1007/s11227-023-05587-4>

Kameoka, H. (2015). Non-negative matrix factorization and its variants with applications to audio signal processing. *Journal of the Japan Statistical Society, Japanese Issue*, 44(2), 383–407. <https://doi.org/10.11329/jssj.44.383>

Lab, W. (2023). *RcppPlanc: R wrapper for the PLANC nonnegative matrix factorization library*. <https://github.com/welch-lab/RcppPlanc>.

Lee, D. D. et al. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755), 788–791. <https://doi.org/10.1038/44565>

Meng, C. et al. (2016). Dimension reduction techniques for the integrative analysis of multi-omics data. *Briefings in Bioinformatics*, 17(4), 628–641. <https://doi.org/10.1093/bib/bbv108>

- 70 Pedregosa, F. et al. (2011). Scikit-learn: Machine learning in python. *Journal of Machine*  
71 *Learning Research*, 12(85), 2825–2830.
- 72 Rodriques, S. G. et al. (2019). Slide-seq: A scalable technology for measuring genome-wide  
73 expression at high spatial resolution. *Science*, 363, 1463–1467. [https://doi.org/10.1126/](https://doi.org/10.1126/science.aaw1219)  
74 [science.aaw1219](https://doi.org/10.1126/science.aaw1219)
- 75 Sajad, A. et al. (2025). Recommender systems based on non-negative matrix factorization: A  
76 survey. *IEEE Transactions on Artificial Intelligence*, 1–21. [https://doi.org/10.1109/TAI.](https://doi.org/10.1109/TAI.2025.3559053)  
77 [2025.3559053](https://doi.org/10.1109/TAI.2025.3559053)
- 78 Simanovsky, A. L. et al. (2019). Single haplotype admixture models using large scale HLA  
79 genotype frequencies to reproduce human admixture. *Immunogenetics*, 71, 589–604.  
80 <https://doi.org/10.1007/s00251-019-01144-7>
- 81 Srivastava, A. N., & Sahami, M. (Eds.). (2009). *Text mining: Classification, clustering, and*  
82 *applications*. Chapman; Hall/CRC. <https://doi.org/10.1201/9781420059458>
- 83 Stein-O'Brien, G. L. et al. (2018). Enter the matrix: Factorization uncovers knowledge from  
84 omics. *Trends in Genetics*, 34(10), 790–805. <https://doi.org/10.1016/j.tig.2018.07.003>
- 85 Tsuyuzaki, K. et al. (2020). Benchmarking principal component analysis for large-scale single-  
86 cell RNA-sequencing. *Genome Biology*, 21(1). <https://doi.org/10.1186/s13059-019-1900-3>
- 87 Tsuyuzaki, K. et al. (2023). nnTensor: An r package for non-negative matrix/tensor decomposi-  
88 tion. *Journal of Open Source Software*, 8(84), 5015. <https://doi.org/10.21105/joss.05015>

DRAFT