KL-NMF as best suited NMF variant for the analysis of Mass Spectrometry Imaging data

Melanie Nijs, Tina Smets, Etienne Waelkens, Bart De Moor,
KU Leuven, ESAT/STADIUS, Kasteelpark Arenberg 10, 3001 Leuven, Belgium,
melanie.nijs@kuleuven.be

Abstract

Over the last few decades, non-negative matrix factorization (NMF) has proven to be a very effective approach towards the analysis of biomedical data. Within the field of mass spectrometry imaging (MSI) data analysis, it offers intuitive results visualizing simultaneously the spatial and spectral distributions present in a slice of tissue, see Figure 1.

Given the Poisson distributed nature of the mass spectrometer detector, our goal was to find the most suitable NMF related method. To this end, we compared standard NMF with its Kullback-Leibler variant, imposing Poisson distributed on the data. To clarify the difference with related methods from the field of text mining, probabilistic latent semantic analysis (PLSA) and latent Dirichlet allocation (LDA) were added to the comparison. We performed analysis on three mouse pancreas and four human lymph node samples acquired using MALDI-TOF.

Our work[1] shows that each method relies on a different framework which has immediate implications on the results. KL-NMF, which is not typically used for MSI data analysis so far, improves both NMF and PLSA. PLSA, while too overfitting in text mining, improves LDA in the setting of MALDI-MSI. Also, we proved that KL-NMF and PLSA algorithms reach a different local optimum in the case of MSI data analysis.

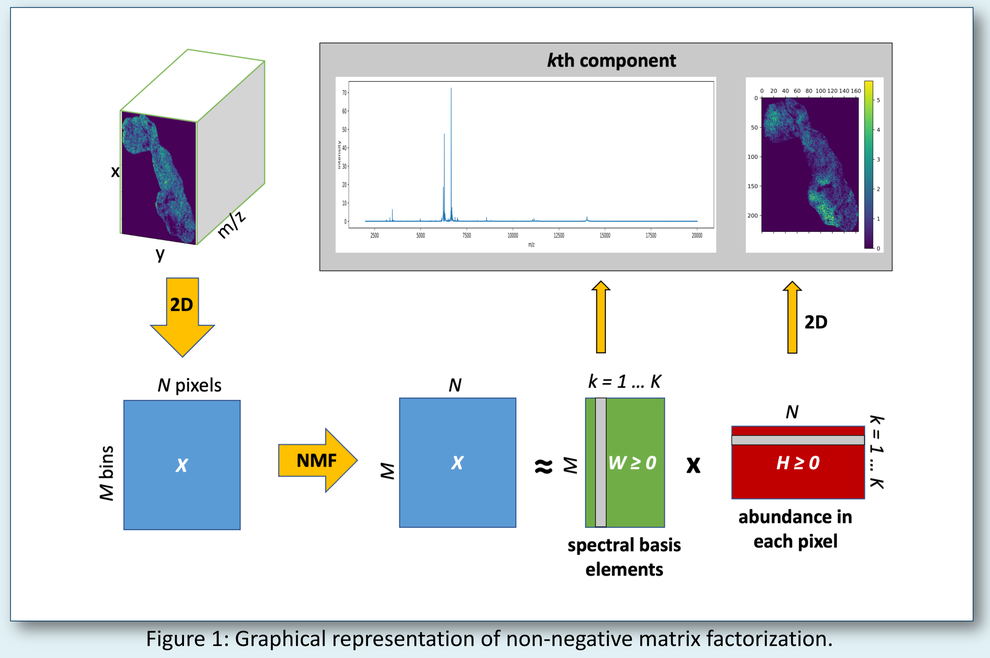

Figure 1: Graphical representation of non-negative matrix factorization.

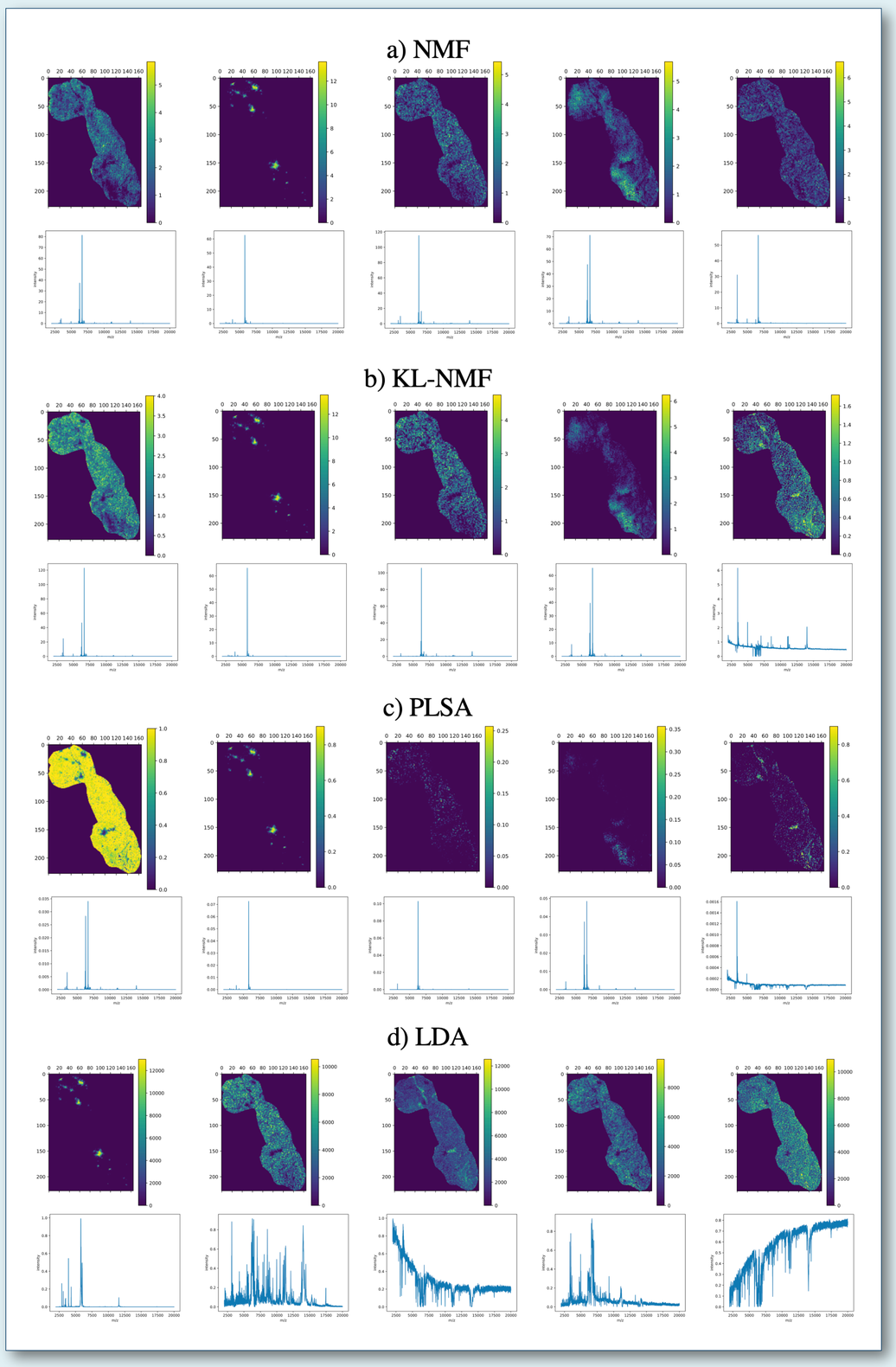

Figure 2: Visual results of the mouse pancreas sample M2, obtained using MALDI-TOF MS, with factorization in 5 components.

Methods

a) Non-negative Matrix Factorization (NMF)

$$\min_{\boldsymbol{W},\boldsymbol{H}\geq 0} \left|\left|\boldsymbol{X}-\boldsymbol{W}\boldsymbol{H}\right|\right|_F^2$$

b) Kullback-Leibler Non-negative Matrix Factorization (KL-NMF)

$$\min_{W,H\geq 0} -\ln L\left(\boldsymbol{X}|\boldsymbol{W},\boldsymbol{H}\right) = \min_{W,H\geq 0} -\ln \prod_{ij} P\left(x_{ij}\right)$$
$$= \min_{W,H\geq 0} \sum_{i=1}^{N}\sum_{j=1}^{M}\left[(\boldsymbol{W}\boldsymbol{H})_{ij} - x_{ij}\ln(\boldsymbol{W}\boldsymbol{H})_{ij}\right] + c$$

c) Probabilistic Latent Semantic Analysis (PLSA)

$$\boldsymbol{X} \approx M\widehat{\boldsymbol{A}}\mathrm{diag}(\hat{\boldsymbol{s}})\widehat{\boldsymbol{B}}$$
$$\boldsymbol{W} = \widehat{\boldsymbol{A}} \text{ and } \boldsymbol{H} = M\,\mathrm{diag}(\hat{\boldsymbol{s}})\widehat{\boldsymbol{B}}$$

d) Latent Dirichlet Allocation (LDA)

$$\text{Estimate } p(z,\theta,\beta|w,\alpha,\eta) \approx p(\boldsymbol{W},\boldsymbol{H}|w,\alpha,\eta):$$

Results and Evaluation

Due to the absence of ground truth information, evaluation of MSI data analysis remains a difficult task. As such, we used a combination of multiple metrics and other criteria to assess each method:

- **Numerically**: In Table 1, we see a numerical evaluation of the methods comparing reconstruction errors ℓ_1 and ℓ_2, as well the statistical correspondence using KL-divergence with the original dataset. We also included runtime and memory usage of each method.
- **Visually**: We compare Figure 2 with spatial regions obtained with the UMAP embedding, see Figure 3.
- **Scalability**: As resolution of MSI increases, how well does each method still perform in terms of runtime and memory usage. We investigated the effect of sparse representation of the algorithms.

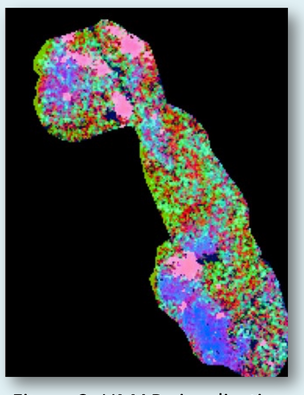

Figure 3: UMAP visualization with cosine distance metric of sample M2.[2]

Table 1: Numerical analysis of the mouse pancreas sample M2. Each experiment ran until convergence in the update was reached, with a threshold of 1e-3.

method	ℓ_1	ℓ_2	KL	time (s)	mem. (GB)
NMF	0.4770	**1.838e-6**	0.2578	**3.7 ± 0.1**	**0.8 ± 1.6**
KL-NMF	**0.4366**	2.290e-6	**0.1922**	88.2 ± 7.2	7.4 ± 1.7
PLSA	0.8283	0.0909	0.2096	63.9 ± 5.1	6.5 ± 2.4
LDA	-	-	1.5376	4 221.0 ± 29.1	1.7 ± 0.002

Conclusion

In terms of runtime, memory usage, and scalability, classical NMF still outperforms the others, which makes it a suitable method for a first impression of the data. It also scores best on the ℓ_2 reconstruction error. For the ℓ_1 error and KL-divergence, we see that KL-NMF outperforms PLSA, and this is also confirmed by the visual comparison with UMAP.

That KL-NMF is best-suited in our analysis, does not mean this is always the case. This work mainly tries to inform peers in the field of MSI data analysis that **each NMF related method relies on different assumptions and thus will have different results.** Additionally, we included a detailed analysis of the performance of KL-NMF on MSI data. It was often assumed PLSA and KL-NMF are identical methods, since they optimize the same objective function. We showed that because these methods use different solvers and other normalization steps, they converge to different local optima.

References

1. Nijs, M, Smets, T, Waelkens, E, De Moor, B. **A mathematical comparison of non-negative matrix factorization related methods with practical implications for the analysis of mass spectrometry imaging data**. *Rapid Commun Mass Spectrom*. 2021; 35(21):e9181. doi: 10.1002/rcm.9181
2. Smets T, Verbeeck N, Claesen M, et al. **Evaluation of Distance Metrics and Spatial Autocorrelation in Uniform Manifold Approximation and Projection Applied to Mass Spectrometry Imaging Data.** Anal Chem. 2019;91(9):5706-5714. doi:10.1021/acs.analchem.8b05827