KU LEUVEN

STADIUS
Center for Dynamical Systems, Signal Processing and Data Analytics

Citation/Reference	Nijs M., Smets T., Waelkens E., De Moor B., (2021) ***A Mathematical Comparison of Non-negative Matrix Factorization-Related Methods with Practical Implications for the Analysis of Mass Spectrometry Imaging Data*** Rapid Communication in Mass Spectrometry,
Archived version	Author manuscript: the content is identical to the content of the published paper, but without the final typesetting by the publisher
Published version	insert link to the published version of your paper
Journal homepage	https://analyticalsciencejournals.onlinelibrary.wiley.com/journal/10970231
Author contact	melanie.nijs@esat.kuleuven.be your phone number *+ 32 (0)491 22 74 10*
Abstract	**RATIONALE:** Non-negative matrix factorization (NMF) has been used extensively for the analysis of mass spectrometry imaging (MSI) data, visualizing simultaneously the spatial and spectral distributions present in a slice of tissue. The statistical framework offers two related NMF methods: probabilistic latent semantic analysis (PLSA) and latent Dirichlet allocation (LDA), which is a generative model. This work offers a mathematical comparison between NMF, PLSA, and LDA, and includes a detailed evaluation of Kullback-Leibler NMF (KL-NMF) for MSI for the first time. We will inspect their results for MSI data analysis as these different mathematical approaches impose different characteristics on the data and the resulting decomposition. **METHODS:** The four methods (NMF, KL-NMF, PLSA, and LDA) are compared on seven different samples: three originated from mice pancreas and four from human-lymph-node tissues, all obtained using MALDI-TOF MS.

RESULTS: Where matrix factorization methods are often used for the analysis of MSI data, we find that each method has different implications on the exactness and interpretability of the results. We have discovered promising results using KL-NMF, which has only rarely been used for MSI so far, improving both NMF and PLSA, and have shown that the hitherto stated equivalent KL-NMF and PLSA algorithms do differ in case of MSI data analysis. LDA, assumed to be the better method in the field of text-mining, is shown to be outperformed by PLSA in the setting of MALDI-MSI. Additionally, the molecular results of the human-lymph-node data have been thoroughly analysed for better assessment of the methods under investigation.

CONCLUSIONS: This paper offers an in-depth comparison of multiple NMF related factorization methods for MSI. We aim to provide fellow researchers in the field of MSI a clear understanding of the mathematical implications using each of these analysis techniques, which might affect the exactness and interpretation of the results.

IR url in Lirias https://lirias.kuleuven.be/handle/123456789/xxxxxx

(article begins on next page)

JOURNAL: Rapid Communications in Mass Spectrometry

Full title:

A Mathematical Comparison of Non-negative Matrix Factorization-Related Methods with Practical Implications for the Analysis of Mass Spectrometry Imaging Data

Short title (max. 70 characters):

A Mathematical Comparison of NMF-Related Methods for MSI Data Analysis

Names and affiliations of all authors:

Melanie Nijs[1], Tina Smets[1], Etienne Waelkens[2], Bart De Moor[1]

[1]STADIUS Center for Dynamical Systems, Signal Processing, and Data Analytics, Department of Electrical Engineering (ESAT), KU Leuven, 3001 Leuven, Belgium

[2]Department of Cellular and Molecular Medicine, KU Leuven Campus Gasthuisberg O&N 2, 3000 Leuven, Belgium

Full address, including email, telephone and fax of the author who is to check the proofs:

Melanie Nijs,
KU Leuven, ESAT/STADIUS, Kasteelpark Arenberg 10, BE-Leuven, Belgium.
melanie.nijs@esat.kuleuven.be
+32 491 22 74 10

Abstract:

RATIONALE: Non-negative matrix factorization (NMF) has been used extensively for the analysis of mass spectrometry imaging (MSI) data, visualizing simultaneously the spatial and spectral distributions present in a slice of tissue. The statistical framework offers two related NMF methods: probabilistic latent semantic analysis (PLSA) and latent Dirichlet allocation (LDA), which is a generative model. This work offers a mathematical comparison between NMF, PLSA, and LDA, and includes a detailed evaluation of Kullback-Leibler NMF (KL-NMF) for MSI for the first time. We will inspect their results for MSI data analysis as these different mathematical approaches impose different characteristics on the data and the resulting decomposition.

METHODS: The four methods (NMF, KL-NMF, PLSA, and LDA) are compared on seven different samples: three originated from mice pancreas and four from human-lymph-node tissues, all obtained using MALDI-TOF MS.

RESULTS: Where matrix factorization methods are often used for the analysis of MSI data, we find that each method has different implications on the exactness and interpretability of the results. We have discovered promising results using KL-NMF, which has only rarely been used for MSI so far, improving both NMF and PLSA, and have shown that the hitherto stated equivalent KL-NMF and PLSA algorithms do differ in case of MSI data analysis. LDA, assumed to be the better method in the field of text-mining, is shown to be outperformed by PLSA in the setting of MALDI-MSI. Additionally, the molecular results of the human-lymph-node data have been thoroughly analysed for better assessment of the methods under investigation.

CONCLUSIONS: This paper offers an in-depth comparison of multiple NMF related factorization methods for MSI. We aim to provide fellow researchers in the field of MSI a clear understanding of the mathematical implications using each of these analysis techniques, which might affect the exactness and interpretation of the results.

Keywords: Non-negative matrix factorization, Mass Spectrometry Imaging, Kullback-Leibler, probabilistic latent semantic analysis, latent Dirichlet allocation

1. INTRODUCTION

Mass Spectrometry Imaging (MSI) is an exploratory research technique aiming to visualize the spatial distributions of biomolecules present in a slice of tissue.[1] Well-known MSI analysis techniques include clustering and dimensionality reduction methods, which entail amongst others factorization methods and non-linear dimensionality reduction (NLDR) techniques. Clustering, such as k-means[2], identifies similar pixels based on the distance between their spectra. NLDR methods, such as t-distributed stochastic neighbour embedding (t-SNE)[3] and uniform manifold approximation (UMAP)[4], embed the dataset on a high-dimensional non-linear manifold. In contrast to these techniques, factorization methods reduce the traditionally very large MSI matrices to a product of two low-rank matrices, containing the spectral and spatial distributions of the data. Each column-row combination of this product can be seen as an individual component describing an important section of the data. Because of the simultaneous extraction of the spectral and spatial elements of MSI data and their straight-forward interpretability, factorization methods have been extensively proven useful over the last decade.[5]

A well-known example of matrix factorization is principal component analysis (PCA).[6] PCA projects the data orthogonally and linearly onto a number K of orthogonal principal components, i.e. the eigenvectors corresponding to the K largest eigenvalues of the matrix. The data is factored in a score matrix and a loadings matrix. The score matrix, which maximizes variation along each principal component, tries to capture as much of the data as possible on the reduced rank-K subspace. The loadings matrix describes the weights of the original data elements in the new space. PCA is an unconstrained factorization method and does not take into account the non-negative nature of MSI data. This can be resolved using non-negative matrix factorization (NMF) instead.

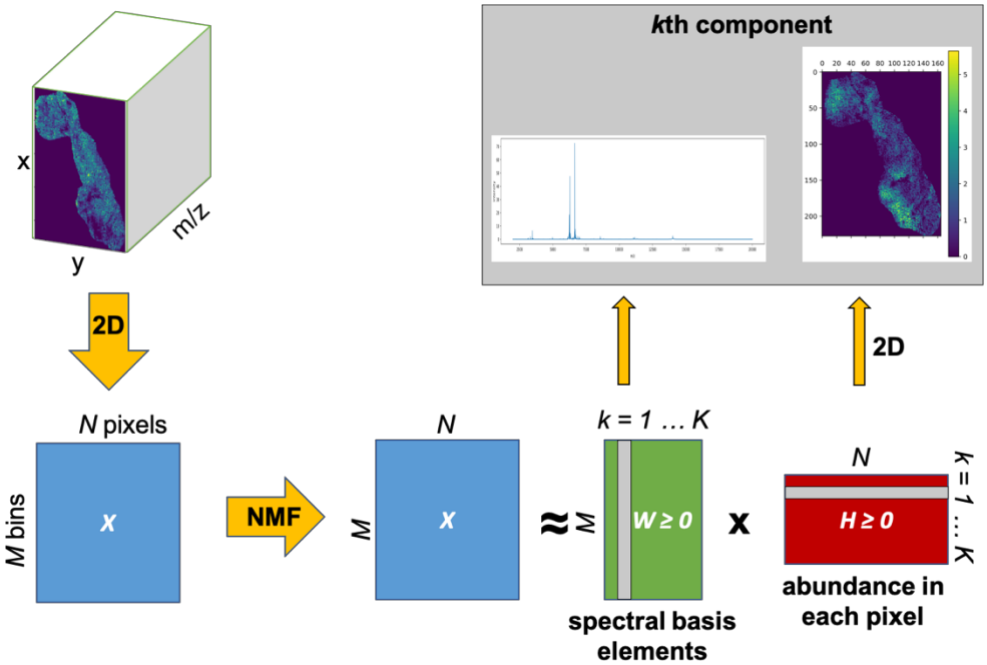

Figure 1: Graphical representation of non-negative matrix factorization.

NMF, as shown in Figure 1, factorizes the matrix into a spatial and spectral component by minimization of the Frobenius norm whilst constraining all values to be greater than or equal to zero. NMF has been used extensively for the analysis of MSI data and extensions and variations of this algorithm is still actively being researched.[7,8] Besides non-negativity

as an important characteristic of MSI data, MSI data is known to follow a Poisson distribution because of the count nature of the mass spectrometer. With MALDI-MSI, the tissue under investigation is ionized by means of a laser beam, after which the ions are accelerated through the mass spectrometer until they reach a sensor that measures the number of hits as well as the time between ionization and reaching the sensor. Due to the heterogeneous distribution of molecules (proteins) within the tissue, the counts (or intensity) measured for each m/z ratio can be seen as a random number of counts, allowing the assumption of MSI data being Poisson distributed, as previously claimed by Harn et al. and Piehowski et al.[9–12] Standard NMF minimizes the squared error assuming a Gaussian distribution of the noise, but Lee and Seung's 1999 paper on NMF already proposed the use of Kullback-Leibler divergence (KL-NMF) instead of least squares as error measure to incorporate Poisson noise for the factorization of face images.[13] Another approach to include the Poisson prior to NMF can be found in the statistical framework: Probabilistic latent semantic analysis (PLSA) or indexing (PLSI) is a topic modelling technique often used for text mining. Here, a word count matrix of documents, a co-occurrence matrix, will be factorized in rank-K topics, each with a document-to-topic distribution and a word-to-topic distribution. It has been shown that maximizing the PLSA likelihood function yields the same objective as minimizing the KL divergence error. Under some conditions the KL-NMF and PLSA methods are equivalent, but they generally converge to different optima.[14,15] In text mining, it is shown that latent Dirichlet allocation (LDA) outperforms PLSA in terms of disambiguation and precision due to the sparse Dirichlet prior.[16] LDA is generally not categorized as a factorization method but as a generative statistical model closely related to PLSA and thus indirectly connected to NMF.[17]

Besides being non-negative and Poisson distributed, we also know that MSI data can be modelled as non-linear. Methods such as t-SNE and UMAP perform a non-linear mapping of the data onto a lower dimensional structure and provide very intuitive visualizations because they take the entire feature space into account. One notable draw-back of these methods is the inability to simultaneously provide a spatial and spectral distribution of the data, making them harder to interpret than linear factorization methods.

In this paper, we will compare the results of the following methods: NMF, KL-NMF, PLSA, and LDA for the application on MSI data analysis. Although there are many more variants of NMF, we limit ourselves to these four because there is a clear resemblance in what they try to obtain, but their mathematics imply fairly different results. All but KL-NMF have already been investigated for the analysis of MSI data.[18–20] KL-NMF has been used in context of MSI in previous work, but has not yet been examined in detail or evaluated against standard NMF.[21] This is likely due to the theoretical resemblance with PLSA, which has also been stated specifically for MSI.[22] However, it has already been proven that both methods differ practically, and the theoretical resemblance only holds under certain circumstances.[15] In general, the selection of a certain algorithm for analysis is rarely substantiated, nor are the implications of a certain framework mentioned. Because MSI data is an exploratory technique, it is important to keep in mind the deviations resulting from the mathematical framework of the chosen algorithm. By comparing the results of different approaches, the correctness of results can be assessed with more certainty. Some frequently used numerical performance measures are the reconstruction error, computation time, and memory usage. Comparison of the non-linear mapping with the spatial distributions obtained with factorization methods could serve as an additional performance measure. As UMAP was found to be excellent for MSI data analysis[23,24], we will examine whether the factorization method is able to capture the same regions as UMAP, for example. Furthermore, the mathematical complexity, interpretability, and scalability will also be used as evaluation criteria.

The next sections provide a mathematical description of the different algorithms, an experimental comparison and a discussion on the different methods. With this paper we hope to offer a comprehensive overview of the NMF-related methods mentioned and evaluate their performance on MSI data analysis by means of multiple measures, including the so far not yet compared KL-NMF method. We aim to provide a clear comparison and understanding of each method, listing their advantages and disadvantages. We will show that KL-NMF offers the best analysis for our data, when trying to match the obtained distributions of UMAP.

2. METHODS

Define $\mathbf{X} \in R_+^{M \times N}$, an MSI dataset consisting of M rows, describing the measured m/z values, and N columns, describing all the (vectorized) pixels in the sample. We want to extract K non-negative components that represent the data as good as possible, according to a certain optimization function. The value K is called the rank of the decomposition. One straightforward way to determine the rank-K of a matrix is to look at the singular values. By taking K equal to the number of singular values of X that are distinctively higher than the other singular values, we get a good estimate for the rank of the non-noise part of the data. Alternatively, there exist several other, more complex methods for the rank estimation of a matrix as the determination of the rank is a research discipline on its own.[25] Extracting the non-negative components can be done using a variety of algorithms. Here, we will restrict ourselves to the examination of non-negative matrix factorization (NMF), Kullback-Leibler NMF (KL-NMF), probabilistic latent semantic analysis (PLSA), and latent Dirichlet allocation (LDA).

2.1. Non-negative Matrix Factorization (NMF)

The standard non-negative matrix factorization decomposes the matrix $\boldsymbol{X} \in R_+^{M \times N}$ in two factor matrices: $\boldsymbol{W} \in R_+^{M \times K}$, of which the columns represent the rank-K spectral basis elements, and $\boldsymbol{H} \in R_+^{K \times N}$, describing the contribution of each basis element to the spectrum at each pixel p_i. This can be written as:

$$\boldsymbol{X} \approx \boldsymbol{W}\boldsymbol{H},$$

and yields the following optimization problem:

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

where the standard Frobenius norm is used, denoted as $||\ \cdot\ ||_F^2$.

The main advantage of NMF is the easy interpretability of the factor matrices. Using the additive linear combination of the spectral basis elements in $\boldsymbol{W} \in R_+^{M \times K}$, weighed by the contribution matrix $\boldsymbol{H} \in R_+^{K \times N}$, the spectral signature of each pixel can be approximated. This way, NMF provides meaningful information, both spectral and spatial, making it a well-suited method for MSI data analysis.
Unfortunately, the NMF problem is NP-hard (non-deterministic polynomial-time hard) and not unique. In most cases, the objective cannot be solved exactly and the optimum

needs to be approximated numerically instead. This typically results in only a local optimum. For the numerical approximation, many algorithms exist such as multiplicative update (MU) rules and alternating least-squares.
NMF decomposition, from the scikit-learn Python package[26], was configured with a coordinate descent (CD) solver. This is the framework used most for NMF; it optimizes in alternation one of the factor matrices $\boldsymbol{W}$ or $\boldsymbol{H}$, while keeping the other one fixed. This will result in a nonnegative least squares subproblem with only one unknown, which is convex and thus easier to solve. Because this optimization problem is symmetric in $\boldsymbol{W}$ and $\boldsymbol{H}$, the same update can be used for both factor matrices. These updates guarantee a decrease of the objective function. The computational complexity of NMF on most real-world problems equals $O(mnk)$ operations per iteration.[27]

2.2. Kullback-Leibler Non-negative Matrix Factorization (KL-NMF)

As mentioned before, there exists a variation on traditional NMF that minimizes the KL divergence.[28] In this case, Poisson distribution is included when factorization is performed, instead of the Gaussian noise implied by standard NMF. For each m/z value i, and for each pixel j, the observed intensity x_{ij} follows a Poisson distribution $P(x_{ij})$. The matrix $\boldsymbol{X} \in R_+^{M\times N}$ is factored in similar matrices $\boldsymbol{W} \in R_+^{M\times K}$ and $\boldsymbol{H} \in R_+^{K\times N}$, but the optimization problem now includes the following maximum likelihood function:

$$L(X|W,H) = \prod_{ij} P(x_{ij}) = \prod_{ij} \exp(-\lambda_{ij}) \frac{\lambda_{ij}^{x_{ij}}}{x_{ij}!},$$

with rate parameter $\lambda_{ij} = (WH)_{ij}$. Or, similarly, it minimizes the negative log-likelihood:

$$\min_{W,H\geq 0} -\ln L\,(X|W,H)$$

$$= \min_{W,H\geq 0} \sum_{i=1}^{N}\sum_{j=1}^{M} \left(\lambda_{ij} - x_{ij}\ln\lambda_{ij} + \ln x_{ij}\,!\right)$$

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

with c a constant depending only on $\boldsymbol{X}$, not on $\boldsymbol{W}$and $\boldsymbol{H}$.

To solve KL-NMF, multiplicative update rules (MU rules) are most commonly used.[29] However, similar to standard NMF, the problem is NP-hard and not unique. The interpretation of the factor matrices is similar but using the KL-NMF now implies Poisson distribution on the result. This corresponds better than standard NMF to the prior information known about MSI data, but has, as known so far, not yet been examined and only rarely been used for MSI data analysis. In practice, many NMF estimators exist for different stochastic distributions of data.
The scikit-learn Kullback-Leibler extension of NMF uses multiplicative update rules (MU) with β-divergence for $\beta = 1$ as factorization algorithm, as this is equivalent to the KL divergence.[26] The MU are the most widely used approach for KL-NMF where matrices W and H are updated with component-wise multiplications, defined as follows:

$$\boldsymbol{H} \leftarrow \boldsymbol{H} \circ \frac{\boldsymbol{W}^T \dfrac{\boldsymbol{X}}{\boldsymbol{WH}}}{\boldsymbol{W}^T \boldsymbol{1}},$$

$$\boldsymbol{W} \leftarrow \boldsymbol{W} \circ \frac{\dfrac{\boldsymbol{X}}{\boldsymbol{WH}} \boldsymbol{H}^T}{\boldsymbol{1} \boldsymbol{H}^T}.$$

These rules are easy to implement and scale well. This approach also has a computational complexity of $O(mnk)$.

2.3. Probabilistic Latent Semantic Analysis (PLSA)

The probabilistic equivalent of KL-NMF is PLSA. In essence, both methods have the same theoretical objective, but they use different approaches of the MU rules to achieve the optimum. Also, PLSA, as a statistical approach, will result in probability distributions, while KL-NMF will consist of values relative to the given dataset. This makes them inherently dissimilar in practice and causes them to converge to different local optima.

PLSA originates from text mining, as it is an extension on latent semantic analysis (LSA). In text mining, PLSA observes the correspondence between words (w) in multiple documents and collects the word counts per document in a co-occurrence matrix.[31] Additionally, in the setting of PLSA, we define as hidden variables the topics (c) across the dataset to which certain words and documents are related.

$$P(w_i, d_i) = \sum_{c=1}^{K} P(c) P\left(d_j \middle| c\right) P(w_i|c).$$

Translating this to MSI, as has been done in previous work, we can see this as the correspondence between the m/z bins (words) and the spectra (documents), located at each pixel entry.[32] The hidden topics, in case of MSI, can be defined as the different cell types within the tissue. When the number of m/z bins ℓ in the dataset is sufficiently large, PLSA results in the following factorization model:

$$\boldsymbol{X} \approx \ell \widehat{\boldsymbol{A}} \mathrm{diag}(\hat{\boldsymbol{s}}) \widehat{\boldsymbol{B}} \; = \; \boldsymbol{WH}.$$

Here, $\hat{\boldsymbol{s}} \in \mathbb{R}_+^{\boldsymbol{K}}$, corresponding to the factor $P(c)$ of the previous equation**,** is a vector describing the probability of a random m/z value being related to the kth topic, matrix $\widehat{\boldsymbol{A}} \in \mathbb{R}_+^{\boldsymbol{M} \times \boldsymbol{K}}$, or $P(w_i|c)$, describes the probability of a certain m/z value contributing to a certain spectrum, and matrix $\widehat{\boldsymbol{B}} \in \mathbb{R}_+^{\boldsymbol{K} \times \boldsymbol{N}}$, or $P\left(d_j \middle| c\right)$, the probability of each spectrum belonging to a certain topic or tissue type. As they entail probabilities, the vector and matrices are exposed to sum-to-one constraints with respect to the spectra. If we define $\boldsymbol{W} = \widehat{\boldsymbol{A}}$ and $\boldsymbol{H} = \ell \mathrm{diag}(\hat{\boldsymbol{s}}) \widehat{\boldsymbol{B}}$, the PLSA model can be seen as an NMF. More specifically, PLSA assumes $\boldsymbol{X}$ to be Poisson distributed in terms of $\widehat{\boldsymbol{A}} \mathrm{diag}(\hat{\boldsymbol{s}}) \widehat{\boldsymbol{B}}$. As the stationary points of KL-NMF preserve the row and column sum of $\boldsymbol{X}$, it can be proved that they correspond to a solution of PLSA.[15] This allows the correspondence of PLSA to KL-NMF in the local optima, despite the normalization difference between two formulations. Nevertheless, in practice both methods do differ. The MU rules of KL-NMF do not take the normalization of PLSA into account explicitly. Moreover, PLSA uses the Expectation-Maximization (EM) algorithm, where it applies a normalization in each step. For a more elaborate mathematical derivation on the

correspondences and differences of these two methods, we refer the reader to a more elaborate work.[30]
Using the EnsTop[33] Python package, the expectation step calculates the probability of each topic given a word-document pair:

$$P(c|w,d) = \frac{P(c|w)P(d|c)}{\sum_{c=1}^{k} P(c|w)P(d|c)}.$$

This is followed by the maximization step that uses the result of the expectation step to calculate the estimates $P(w|c)$ and $P(c|d)$:

$$P(w|c) = \frac{\sum_d X_{w,d}\, P(c|w,d)}{\sum_{d,c} X_{w,d}\, P(c|w,d)},$$

$$P(c|d) = \frac{\sum_w X_{w,d}\, P(c|w,d)}{\sum_{w,d} X_{w,d}\, P(c|w,d)}.$$

The computational complexity of PLSA equals $O(mnk)$.[31]

2.4. Latent Dirichlet Allocation (LDA)

Although LDA is not generally thought of as a factorization method, it is a generalization of PLSA and is often considered as the better of the two in the field of text mining, where they both originate from. The main difference between both methods is that PLSA aims to perform an exact modelling of the dataset, while LDA tries to find unknown groupings within the data. LDA is a generative statistical model that uses a Dirichlet prior to impose sparsity on the solution, making it more robust to modelling the noise in the dataset. For MALDI-TOF MSI, it has been used as a soft clustering technique and with tandem MS, one can use LDA to cluster the fragmentation of the data in substructures, as has been done in previous research.[34–37]
Let us define similarly as for PLSA: N as the number of m/z bins (or words) in the data, K as the number of components (or topics), which can be seen as the different cell-types within the tissue, and D as the number of spectra or pixels in the MSI data set (which corresponds to the number of documents in the setting of text mining). Also, let θ_d be the probability of a spectrum (d) belonging to a cell type (c) according to a Dirichlet prior α, and β_c be the probability of a certain m/z bin (w) belonging to a cell type (c) according to another Dirichlet prior η.

A pseudo algorithm of the generative LDA method can now be defined as[16]:

1. For each component or cell type $c \in K$, draw $\beta_c \sim \text{Dirichlet}(\eta)$.
2. For each spectrum $d \in D$, draw $\theta_d \sim \text{Dirichlet}(\alpha)$.
3. For each m/z bin i in document d:
   a. Draw the cell type assignment $z_{di} \sim \text{Multinomial}(\theta_d)$,
   b. Draw the observed m/z bin $w_{di} \sim \text{Multinomial}(\beta_{z_{di}})$.

This process generates a latent structure that can be used for either prediction of new entries or for data exploration. Using the scikit-learn package for LDA, one can use an efficient batch variational Bayes algorithm to estimate the posterior distribution

$p(z, \theta, \beta | w, \alpha, \eta)$.[26] Using this algorithm, LDA has a computational complexity of $O(mnk)$. For a more detailed description on this algorithm, we refer the reader to the work of Hoffman and Blei.[38]
Note that if we sum over the different topics, LDA can be thought of as a "multinomial PCA" interpretation, i.e., a probabilistic factorization optimizing the squared error.

2.5. Evaluation criteria

To compare the methods described above, multiple evaluation criteria are investigated. First, we compare all methods numerically: we will calculate the reconstruction error, the computation time, and the memory usage for each experiment.

The reconstruction error will be investigated by means of three different measures: the relative ℓ_1-norm, the relative ℓ_2-norm and the KL divergence. We define the relative errors as the norm of the difference between the original and reconstructed data divided by the norm of the original data:

$$\frac{\|\boldsymbol{X}-\boldsymbol{WH}\|_p}{\|\boldsymbol{X}\|_p},$$

with the ℓ_p-norm for a vector x and a real value $p \geq 1$ formulated as:

$$\|x\|_p = (|x_1|^p + |x_2|^p + \cdots + |x_n|^p)^{1/p}.$$

The relative KL divergence between $\boldsymbol{X}$ and $\boldsymbol{WH}$ is calculated as:

$$\sum_i P(i)\, log\left(\frac{P(i)}{Q(i)}\right).$$

Here, P and Q correspond respectively to the vectorized original data $\boldsymbol{X}$ and the vectorized approximation $\boldsymbol{WH}$.
As has been shown in the previous section, NMF and LDA impose a Gaussian distribution of the noise and KL-NMF and PLSA a Poisson distribution. Because of these statistical priors, we can assume NMF to perform better on the squared error (ℓ_2-norm, Frobenius norm) and KL-NMF and PLSA better on the KL divergence error. To this end, we included a third, unrelated error measure: the ℓ_1-norm. This norm calculates the error by means of the absolute value, rather than the squared values used by the ℓ_2-norm, this makes the ℓ_1 reconstruction error more robust towards outliers. Note that for LDA, the comparison using reconstruction errors makes less sense as the model is generative and not aiming to reconstruct the data. We will thus only calculate the KL divergence in the case of LDA, to estimate to what extent the model follows a similar distribution as the given data set.
Additionally, we also measure the average computation time and memory usage for each experiment. For each method and for each dataset, we perform the experiment ten times until convergence, or the maximum number of iterations is reached. Because of the size of real MSI datasets, which is expected to still increase as the technology advances, we monitor time and memory usage as they might reveal practical limitations of the algorithms.
Besides numerical performance measures, there are a number of additional criteria that should be investigated as well. First, we will check how well the spatial distributions resemble the solutions obtained by the non-linear mapping obtained with UMAP. We know these methods are better to capture the non-linearity of MSI data, and thus could serve as an additional quality check. Secondly, since MSI data is ever increasing in size, sparse representations could offer a solution in the future. This way, the large MSI data sets can be represented more efficiently, by using less memory and less computational

power. We want to know how well the examined methods are able to perform with sparse representation, and thus whether they are able to scale up together with the data size. Thirdly, as mentioned before, we want to understand what the meaning is of the results. Depending on the mathematical framework and the used model, interpretation of the components will differ. We will compare each method in terms of their implications on the interpretation. Additionally, the computational complexity and mathematical simplicity of each algorithm will be reviewed. Finally, aside from the mathematical aspects, we will take a closer look at the biological implications for a human-lymph-node sample. Using all these numerical and non-numerical evaluation criteria, we aim to offer a complete comparison of NMF, KL-NMF, LDA, and PLSA for the analysis of MSI data.

2.6. Experimental setup

2.6.1. Data

In order to make proper conclusions on the performance of each method, we executed the experiments on multiple real data sets: mouse pancreas data (M1, M2, and M3) and human-lymph-node data (P1, P2, P3, and P4). The data sets were acquired using MALDI TOF, with a focus on the masses 620-1200 Da of lipids for the lymph data, and masses 2-20 kDa of proteins for the pancreatic data. Total-ion count (TIC) normalization was used for each dataset. Table 1 gives an overview of the specifications of each dataset. For the details on the data acquisition, we refer the reader to the work of T. Smets.[23]

2.6.2. Implementation

We examined the four methods (NMF, KL-NMF, PLSA, and LDA) on the seven tissue samples described above. The experiments were done in Python using the scikit-learn package[26] for NMF, KL-NMF, and LDA, and the EnsTop[33] package for PLSA.

For fair comparison, all experiments were executed using HPC provided by the Flemish Supercomputer Centre (VSC). Each setup used 36 computing nodes on a 2 Xeon Gold 6140 CPUs@2.3 GHz.

All four algorithms require a given number of components for the decomposition, or rank. The meaning of this rank is the same for each of the methods: it is the number of components or cell types we expect to contribute to the data. We determined the rank of the decomposition by means of the singular value decomposition, as explained in section 3.1.

Each experiment was run until the update between two iterations reached a threshold below 1e-3, or until a maximum number of iterations is reached, whichever comes first. We set the maximum number of iterations to 400 for NMF, KL-NMF and PLSA, and to 20 for LDA. The number of iterations for LDA is much lower because LDA has a much longer computation time per iteration.

Finally, we tried to improve the optimization by using a sparsity favouring initialization by using Non-negative double singular value decomposition (NNDSVD) instead of random initialization.[39] This was done for NMF and PLSA. For KL-NMF, solved with multiplicative update rules, we used the

NNDSVD variant that filles the zeros with small random values, such that they can still be updated during the iterations. LDA, being a statistical generative model, uses a prior for the topic word distribution (β) and a prior for the document word distribution (θ). Both were initialized at the default value $\frac{1}{n_comp}$.

3. RESULTS

Each algorithm ran multiple times on each of these samples, while measuring the numerical evaluation criteria. The additional non-numerical criteria are investigated separately.

3.1. Rank selection

MSI data contains a significant amount of noise, it is obvious we do not want to model all this noise exactly. We want to select a number of components of the factorization methods equal to the rank of the non-noise part of the data. Using a singular value decomposition, one can plot the singular values in a semi-log manner and select a rank equal to the knee in this graph.[40–42] This way, we estimate the number of components based on the variance explained by the singular values. As soon as the magnitude of the singular values stops decreasing significantly, one can expect the rest of the singular values are describing the noise present in the data. For a detailed investigation of a MSI data set, one can use the SVD as a first guess, but it is advised to inspect neighboring values as well or use an additional method such as the AIC criterion for rank selection. We refer to Ubaru and Saad for further information on rank estimation.[25]
Figure 2 and Figure 3 display respectively for the M2 pancreas data and the P1 human-lymph-node data, the formation of the knee around the fifth largest singular value. For this, we will decompose the data in five components for both the pancreatic and lymph-node samples. This choice of rank, $k = 5$, can be justified by the visualization done with UMAP where one can find five distinct regions for M2 as for P1, see Figure 6.

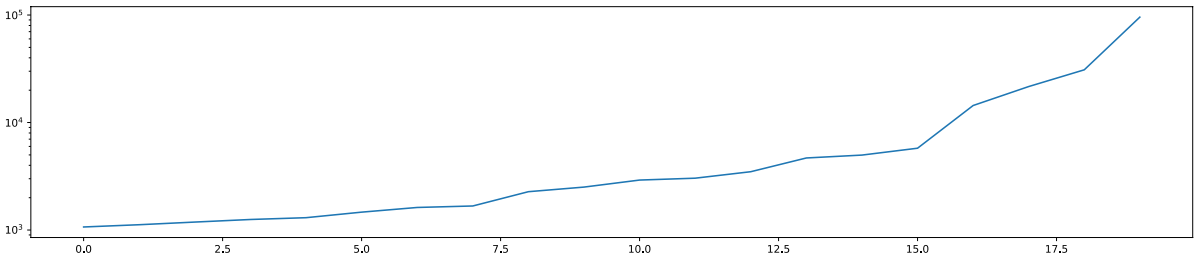

Figure 2: Rank selection of the pancreas sample M2 - 20 largest (right-to-left) singular values are plotted on a semilog scale. There is a kink in the curve around the 5th highest singular value. A good estimate for the rank of this matrix will thus be K=5.

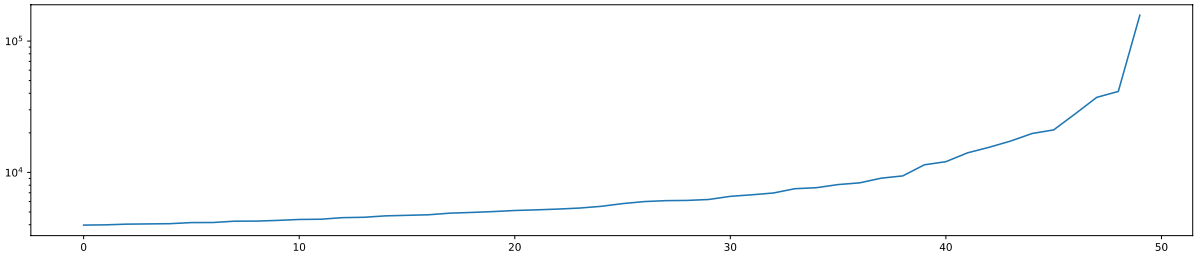

Figure 3: Rank selection of the lymph-node sample P1 - 50 largest (right-to-left) singular values are plotted on a semilog scale. There is a kink in the curve around the 5th highest singular value. A good estimate for the rank of the matrix will thus be K=5.

3.2. Pancreas data

The first data set on which we will compare the factorization methods, is the mouse pancreatic tissue. We have three samples: M1, M2, and M3 of relatively small size for MSI data, see Table 1. Each sample was decomposed in five components, using NMF, KL-NMF, PLSA, and LDA. Table 2 shows the average relative ℓ_1-norm, relative ℓ_2-norm, KL divergence, computation time, and memory usage after 5 runs each. In case of the pancreatic samples, each method reaches convergence before the maximum number of iterations. Note that each method converges to exactly the same value for the multiple runs, making the standard deviation zero for the reconstruction errors. Only the computation time and memory usages fluctuate over multiple runs. We see, as expected, NMF minimizing the relative ℓ_2-norm and KL-NMF the KL divergence error. Additionally, KL-NMF scores also best on the relative ℓ_1-norm. Here, it is interesting to see that PLSA scores second best on the KL divergence error as we would expect. LDA only scores reasonably well on the KL divergence for the M3 sample.

The computation times differ significantly between the investigated methods. NMF clearly executes the fasted, and LDA is by far the slowest, taking into account we limited the number of iterations to twenty. PLSA is a bit faster than KL-NMF, but of the same magnitude. For the memory usage, NMF is clearly the most memory efficient, this can be explained as NMF reaches the update threshold within only one iteration for all samples. LDA, KL-NMF, and PLSA are again of the same magnitude, with LDA the most memory efficient of the three.

The resulting decompositions can be visualized as a spectral and spatial distribution for each component. Figure 4 shows the obtained results for sample M2. A first glance at these results shows that the PLSA decomposition misses a lot of information in its spatial distributions and that the LDA results show poor spectral distributions.

Remark that the spectra have different scales for the intensities across all methods and components. With the pancreas data, most differences between the decompositions are noticeable in in the smaller peaks.

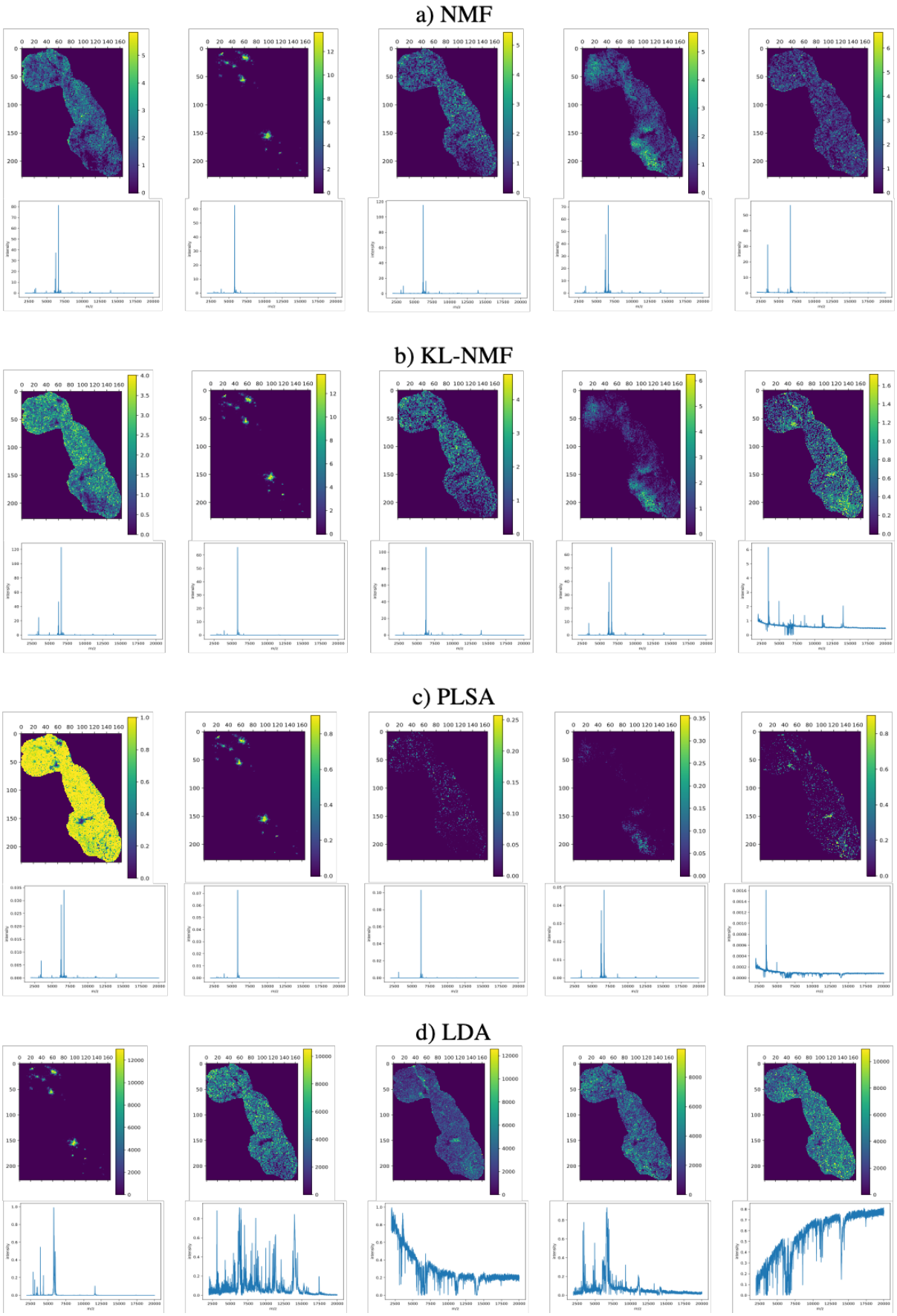

Figure 4: Visual results of the mouse pancreas sample M2. For each method: a) NMF, b) KL-NMF, c) PLSA, and d) LDA we have executed a decomposition in 5 components. The spatial distributions are given in the top rows and the spectral distributions in the bottom rows.

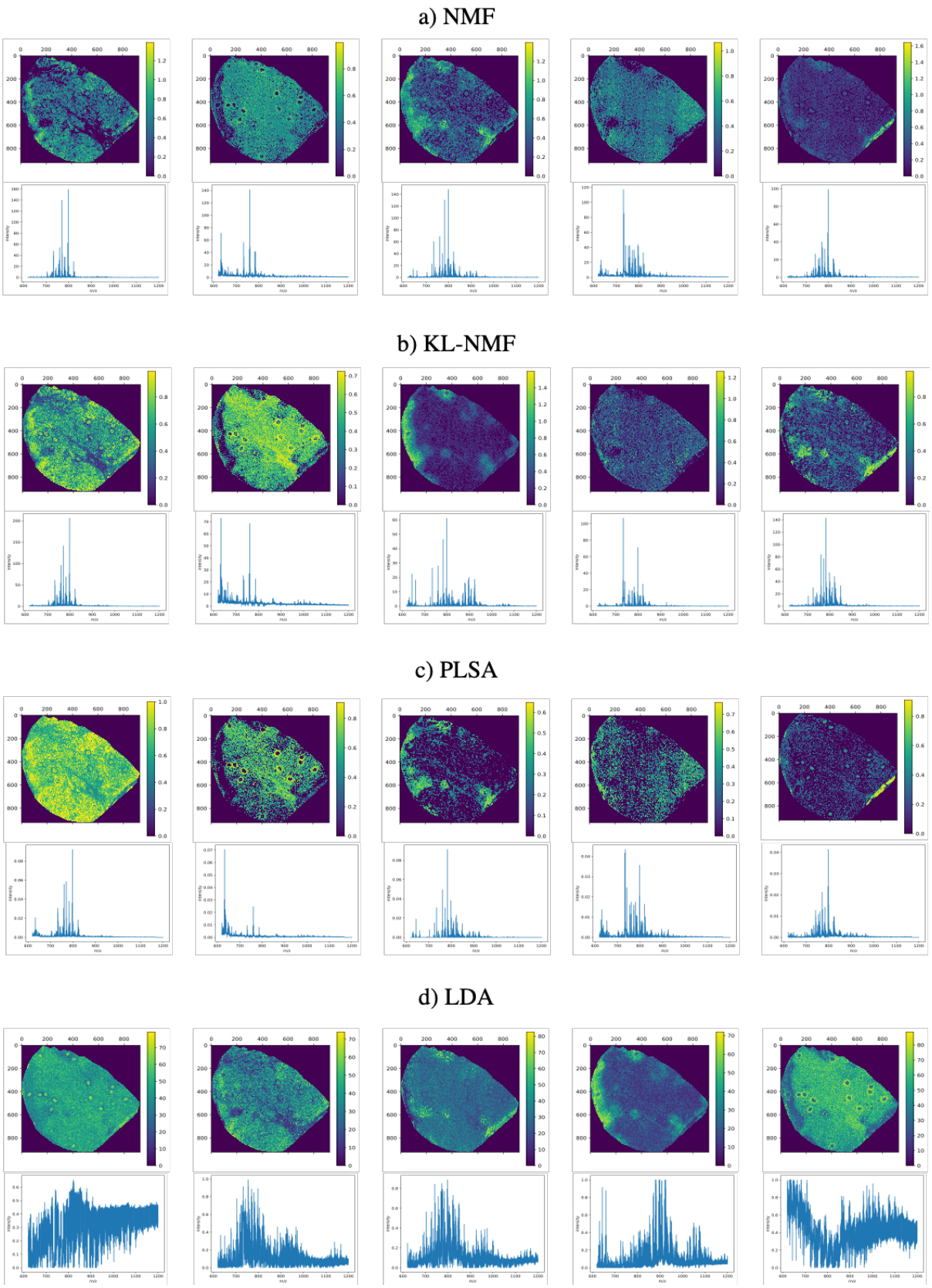

Figure 5: Visual results of the human-lymph-node sample P1. For each method: a) NMF, b) KL-NMF, c) PLSA, and d) LDA we have executed a decomposition in 5 components. The spatial distributions are given in the top rows and the spectral distributions in the bottom rows.

3.3. Lymph-node data

The second data set we will investigate are the human-lymph-node samples. We perform the same experiments on the four lymph-node samples: P1, P2, P3 and P4, with more or less comparable data size, see Table 1. Their sizes are however multitudes greater than the pancreatic samples, affecting as a consequence the peak memory usage significantly. In fact, the lymph-node data resembles much more the current standard of MSI data with higher spatial resolutions. Again, each sample was decomposed in five components, using NMF, KL-NMF, PLSA, and LDA. From Table 3: Numerical analysis of NMF, KL-NMF, PLSA, and LDA on four human lymph samples: P1, P2, P3 and P4. we can deduce similar trends as for the pancreatic sample. Here, as was the case for the pancreas samples, NMF outperforms the other in minimizing the relative ℓ_2-norm and the same is true with KL-NMF for the relative ℓ_1-norm and the KL divergence error. Looking at the standard deviation for each method, we note that they all keep converging to the same solution. Only for LDA convergence of the update between two iterations was not reached, LDA was stopped after its maximum number of iterations (12) as the amount of computation time, roughly twelve hours, was already many times higher those of the other methods.
For the computation times, again NMF is by far the fastest, PLSA and KL-NMF are again of the same order of magnitude and still reasonably fast, with KL-NMF being to slower of the two, but reaching better reconstruction errors. Memory usage is now the lowest for LDA.
Visually, Figure 5 shows the spectral and spatial distributions per method and per component obtained for P1. A gamma-correction of $\gamma = 0.5$ was applied to the 2D visualisations for higher contrast. NMF, KL-NMF and PLSA all show clear decompositions, but differ in their details. Each method extracts different contributions of the relevant peaks. Again, as for the pancreatic sample, LDA results in poor spectral representations.

3.4. Additional evaluation criteria

Aside from the numerical experiments, measuring reconstruction errors, computation time and memory usage, we want to assess the performance of each method more generally as well. Since there no criterion is the most designated one for the evaluation of MSI data analysis, we will include several more evaluation criteria in order to obtain a complete review of every method.

3.4.1. Visual inspection using UMAP

As already described in the works of Smets et al.[23] and some recent computational review papers[22,24], UMAP-based data processing is very well suited for the analysis of MSI data sets. Figure 6a, obtained from Smets et al.[23], shows the visualization of a three-dimensional UMAP embedding with cosine distance of sample M2. From the UMAP embedding, we can distinguish five different regions based on the five distinct colours: pink, blue, black, red and green. Based on previous research[43], we know that the pink region corresponds to the Islets of Langerhans, the blue region is linked to exocrine tissue and the black region to blood vessels. The red and green region are not very clear to interpret and have not yet been related to an anatomical structure. To this end,

we will perform a visual comparison based solely on the pink, blue and black regions of the UMAP embedding.
For all methods, the two most first components give a partition between the islets and the rest of the tissue. This can be explained by the high intensity of insulin (m/z = 5,800) in the dataset. For the exocrine tissue, we see a similar component in NMF and KL-NMF, which is not detected by LDA. PLSA also uncovers the exocrine tissue, but as a less clear image. And lastly the blood vessels are not found in NMF, but are in KL-NMF, PLSA, and LDA. Note that for this anatomical component, we see more noise in the spectra than for the others. LDA shows visually rather good distributions but lacks precision in the spectra. For our MALDI data set, LDA is too general in the decomposition of the spectra. Similarly, we take a look at the correspondence of the decompositions to the UMAP embedding of P1, see Figure 6b, also obtained from Smets et al.[10] The five coloured regions we can distinct here are: yellow, red, blue, green and orange. The spatial distributions of NMF for P1 shows green, blue/green, orange/yellow, and two unidentifiable regions compared to the UMAP decomposition. KL-NMF shows the green region in its first component, next blue/red, yellow, an unidentifiable region and lastly the orange part. For the P1 sample, we see that PLSA performs better than for the M2 sample, we now get a visualisation of similar good correspondence to UMAP as KL-NMF. Both methods however show clearly different decompositions of the sample, but neither are able to extract the exact five regions visible in the UMAP representation. The spatial components of LDA show less detailed visualisations, they are of similar correspondence to UMAP as NMF.

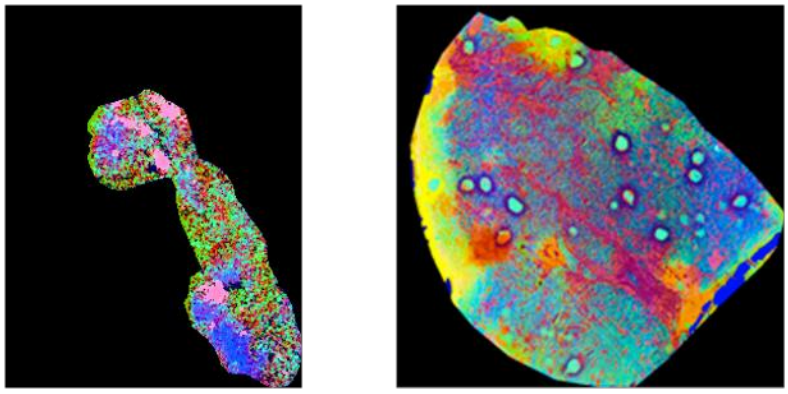

Figure 6: UMAP visualization with cosine distance metric of samples M2 (a) and P1 (b). Obtained from the work of T. Smets et al.[23], see acknowledgements for permission granted.

3.4.2. Scalability

Anticipating the increase in resolution of future MSI datasets, we want to investigate how the factorization methods will be able to cope with even bigger data in the future. The use of sparse representations will be essential, but so far, not all algorithms are able to provide equally good results using sparse operations. With the algorithms we have used for the numerical experiments, we tested their sparse equivalents to gain some insight on their behaviour. NMF improves in speed with little loss in performance. Sparse KL-NMF also

improves computation time, but there is a significant loss in accuracy when doing so. The sparse equivalent of MU is not yet satisfying for MSI data analysis. The EnsTop PLSA algorithm still shows some bugs when there is a sparse input. LDA, however, is faster and more accurate using a sparse input because of the sparse Dirichlet prior.
Note that research towards factorization methods is still ongoing and might lead to new extensions for big data analysis in the future.
Additionally, when looking at Table 1: Specifications of the data sets do not exhibit large sparsity properties. Not many zeros are present, but there are many small values. By adding a cut-off, one might increase sparseness. How to establish a cut-off value and use sparse MSI data analysis, is still something to be explored in the future.

3.4.3. Interpretability

The interpretation of the decomposition depends on the method that was used. For the algebraic methods, the obtained factors can be seen as the spectral basis elements of the data matrix and a weight matrix describing their contribution to each pixel. The components retrieved using NMF and KL-NMF contain information which is deduced directly from the data matrix.
For PLSA, we get similar components except that the weight matrix is now a probability matrix. The results show how likely a certain m/z value is present at a certain pixel. Although the spectra may look similar, they have to be approached differently regarding interpretation.
LDA adds a Dirichlet prior to PLSA, imposing more sparsity in the result and better properties for generation of new "documents". This generative property has little use in the field of MALDI-MSI and that is why the resulting spectral distributions are worse when decomposing with LDA; it generalizes what spectral information might be involved in that component. It could be more useful when an MSI data sets contains fragmentation effects we do not want to model, as is the case for tandem MS.

4. DISCUSSION

For the selection of a MSI data analysis approach there will always be a trade-off; there is no holy grail (yet). This work focusses on NMF-related factorization methods as there are many it is not always clear how they are alike and how they differ. We examined NMF, KL-NMF, PLSA, and LDA both experimentally and theoretically. We can conclude that each method has its own strengths and depending on the goal of the analysis, one might out-perform the other. Interestingly, KL-NMF was not yet thoroughly evaluated in MSI context but proved to be an, at least, equally useful approach.
NMF is a well-known method for data analysis in many fields as well as for MSI. Its popularity comes from the well interpretable decomposition. From the numerical assessment, we find that NMF outperforms all other methods by far in terms of computation time and has often the smallest memory usage as well. Naturally, NMF also scores best on the relative ℓ_2-norm, as its objective function is the minimization of the squared error. Additionally, the results are easy to interpret as basis spectral elements and their spatial weighing across the samples. For MSI data factorization, we advise to always start with

NMF. Even towards future, possibly bigger MSI data sets, NMF is easy to extend to its sparse equivalent with only small loss in information.
A mathematically more suited extension of NMF for the factorization of MSI data is KL-NMF. By imposing Poisson distributed noise, rather than Gaussian, we see that the visual results are even better. Comparison with the UMAP embedding of Figure 6 shows that the KL-NMF components (Figure 4 and Figure 5) extract all regions visible for M2 and almost all regions for P1. This can be seen as a huge advantage, offering a spectral distribution for the spatial regions UMAP identifies. It is in fact the only method that consistently is able to capture the spatial distributions so well, making it even more surprising it has rarely been used for MSI so far. KL-NMF also was able to outperform the other methods in terms of both the relative ℓ_1 reconstruction error and the KL divergence error. Nevertheless, there are some disadvantages of KL-NMF. First, computationally it takes a bit longer than NMF, with a difference of one order of magnitude. Also, the sparse implementation of the algorithm, which could be essential in the future, performs much worse.
At the statistical end, we took a closer look at PLSA. It has already been known for decades that it theoretically resemblances KL-NMF.[44] More recently, it has been shown in many fields that there are in fact differences between both methods.[45] PLSA and KL-NMF optimize the same objective but are only identical methods if KL-NMF adds an additional column normalization and if they would use the same algorithm. MU and EM are different and will result in different solutions when tested on the same data with the same initialization.[14] This statement is supported by looking at the reconstruction errors; every run they each seem to go to their own optimum. The same applies to the visual results, where we can clearly see that they are converging to different local optima. When running until the update between two iterations reaches a threshold of 1e-3, KL-NMF shows a better result than PLSA, which is still missing information. PLSA is a popular method for MSI, where good results are obtained using it, while in text-mining it is becoming less and less popular. Looking at the mathematics, PLSA will work best when there are a lot of samples available.[30] This especially is the case for MSI data, where each spectrum can be seen as a sample. Additionally, PLSA is often critiqued in text-mining for being too sensitive to noise in comparison to LDA; it approximates the given dataset following the objective and does not try to find a generative model for the data. However, in the case of MSI, we specifically want every m/z bin to be equally important. Unfortunately, due to the normalization and the use of the EM algorithm, it does not perform consistently good visually. Although it is a bit faster in runtime, the statistical, normalized counterpart of KL-NMF, shows less accurate spatial distributions when we look at all samples tested, with an exception for P1 where the quality of its spatial distribution matched that of KL-NMF. The spectral basis elements resemble the ones of KL-NMF, albeit in a different scale due to its statistical nature. PLSA does not show absolute contributions of m/z values in its spectral components, but rather the probabilities of each bin contributing.
Finally, LDA is often seen as a generalization of PLSA, as it is able to capture more general trends, emitting the overfitting of noise. For MALDI-MSI, however, generalization is not preferred. In the details of the data, crucial information might still be present. To define what is noise and what is meaningful is very hard for MSI data. Here arises a difference with the field of text mining, where it might be easier to determine what is essential information and what are filler words. MSI is an exploratory technique where little known information is available, and it is harder to check whether something is essential or not. In our experimental set-up, LDA performed worst. Computationally, it took much longer to run until convergence. For the lymph-node data, it took approximately 12 hours to run the 12 iterations. This is the reason the iterations were limited to 12. Even with a sparse input, the computation time stays within the same order of magnitude. Looking at the UMAP

resemblance, the spatial components often performed better than PLSA, as they are smoother and have better contrast, but not as good as KL-NMF. Spectrally, the distributions display little information on the chemical properties making it harder to interpret the spatial results. It has been used on MSI for the clustering of the data, improving memory usage with respect to hierarchical clustering and for tandem MS to cope with fragmentation effects.[34,46] From Table 2 Table 3, we see that the memory usage is lower than for PLSA, but NMF is still performing better.

Aside from the mathematical comparison, we also want to take a closer look at the biochemical interpretation of the resulting spectra. For the pancreatic tissue M2, we refer to the works of T. Smets et al.[23,43] We focus on the human-lymph-node sample.
It has become obvious that MALDI based Mass spectrometry imaging with a focus on the lipid mass range (600-1200 m/z,"Lipid Imaging") is also very well suited to acquire profiles directly from the tissue. These profiles can be applied for facilitating diagnosis, prognosis and follow-up of diseases and therapy. An example of this 'lipid profiling' of selected anatomical regions is shown in Figure 7 and Figure 8. The differences in MS-profiles around the lymph nodules inside the tissue are explored. As shown, the majority of the molecular information (= the number and the height of mass peaks) in positive ion mode, is concentrated in a relative small mass range (700-850 m/z). This range indeed corresponds to masses of (membrane) lipids (especially phosphocholine derivates (PC's) in positive ion mode). The profiles extracted from various regions, demonstrate that an important number of lipid mass peaks are common between the various anatomical structures, and that their relative abundance is of crucial importance. On the other hand, some masses seem to have a higher discriminative value. The masses of 630.7 and 632.7 m/z are found particularly in the mantle zone (see Figure 7). Comparison of normal versus tumoral lymph node tissue in the P1/P3/P4-samples (P2 did not show tumoral tissue) suggest a slightly stronger appearance of m/z 741.6, 761.6, 820.6 and 849.6 in the tumoral transformed regions (see Figure 8 for P1). These observations all underscore the importance of the data processing methods such as the UMAP based methods described earlier or the KL-NMF based method described in this paper, which yield excellent decomposition and high quality visualization with high overall sensitivity.

It should be noted, that a positive identification of the mass peaks is less straightforward as multiple lipids can share the same or nearly the same molecular weight. For instance, lipids with the same number of double bounds but with a difference in the position of these double bonds, all have identical masses. A suggestion of the possible identity of the mass peaks can be retrieved be consulting the Lipid Maps database (www.lipidmaps.org), but for final identification complementary analytical strategies are required, such as ESI-LC-MS with internal lipid standards to support the identity and provide (semi)-quantitative data on the lipid content.

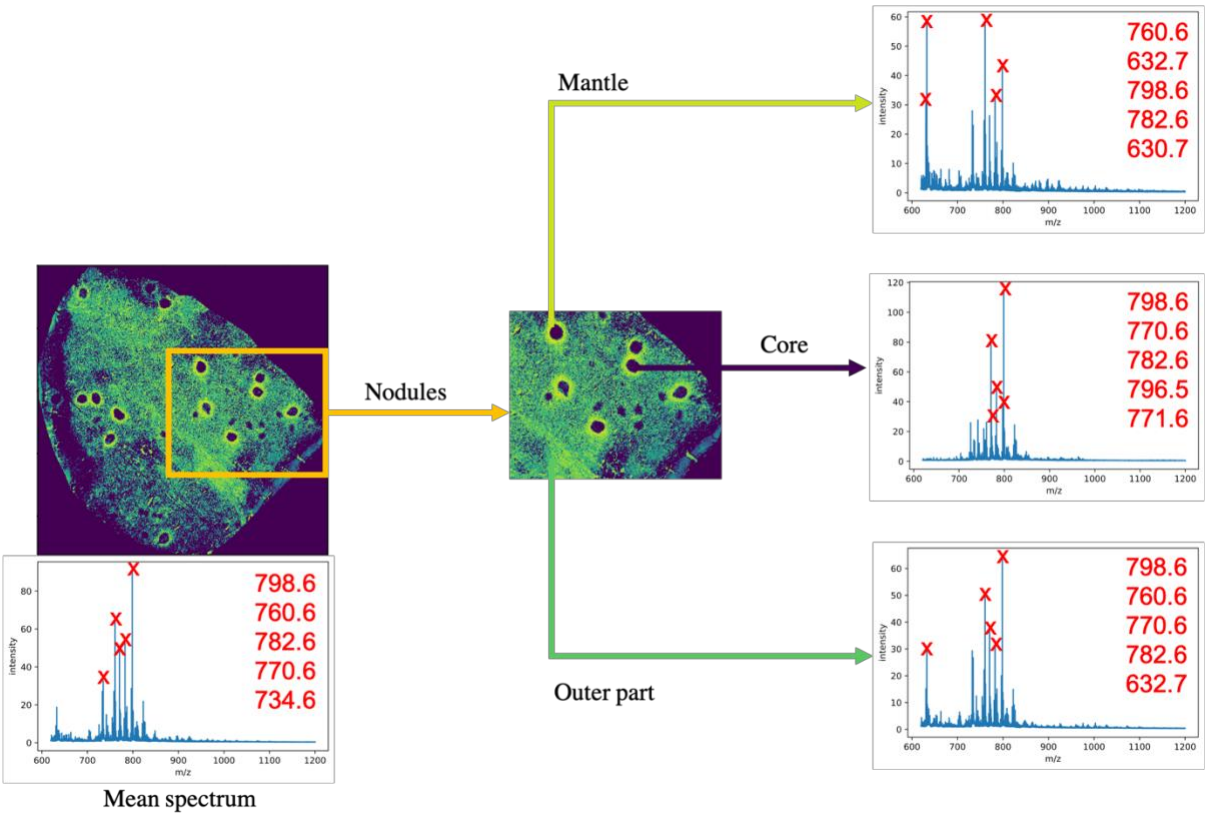

Figure 7: Detailed analysis of the human lymph sample P1. From the entire sample, with its mean spectrum given, we deduce three regions present in the orange box around the nodules: 1. the mantle zone, 2. the core zone, and 3. the outer part. In red, we show the m/z values belonging to the top five intensities present in each region.

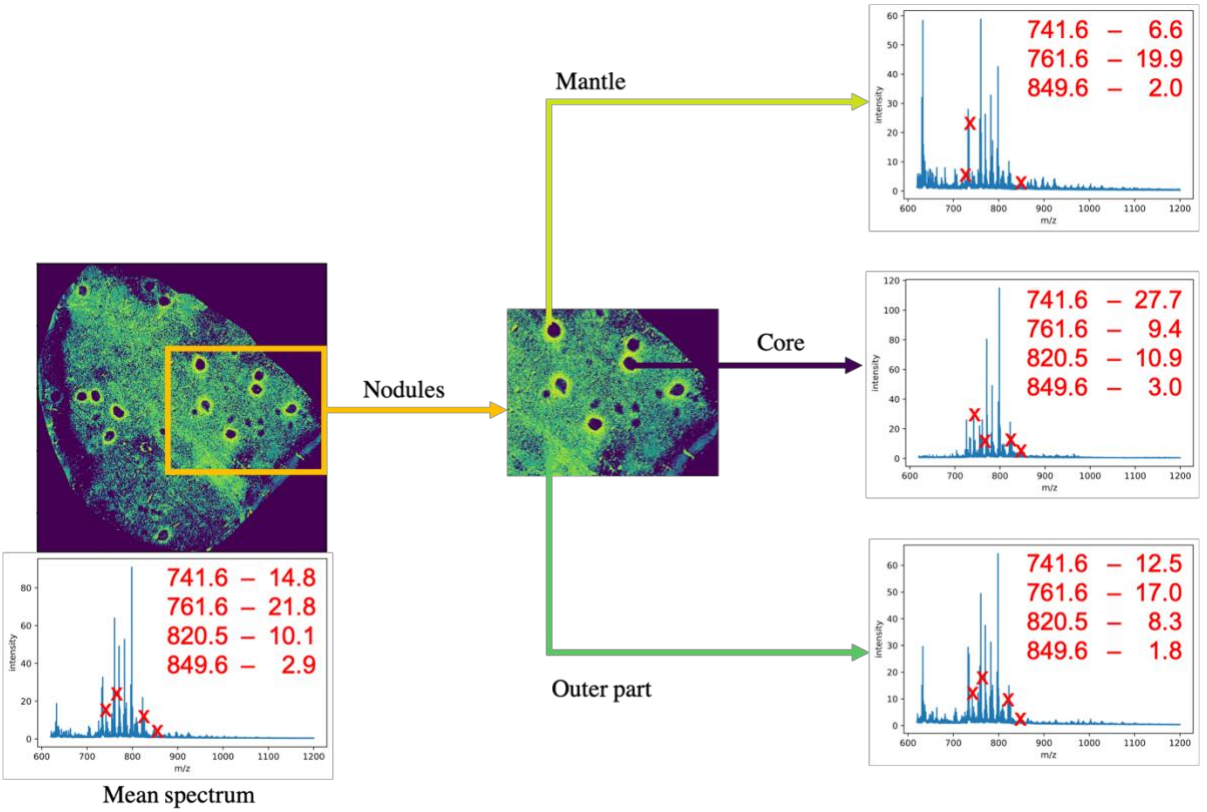

Figure 8: Detailed analysis of the human lymph sample P1. From the entire sample, with its mean spectrum given, we deduce three regions present in the orange box around the nodules: 1. the mantle zone, 2. the core zone, and 3. the outer part. In red, we show respectively the m/z values and intensities of the four m/z values of our interest: 741.6, 761.6, 820.6, and 849.6.

5. CONCLUSION

Within the field of MSI, factorization methods are still a popular tool to interpret the chemical properties present in a tissue sample. We focussed on four NMF-related methods (NMF, KL-NMF, PLSA, and LDA) and evaluated their performance on multiple criteria and on multiple MALDI-MSI data sets. Each method has its significant contribution to the analysis of MSI data, but rarely are they compared to one another. We have shown that NMF is a fast, easy and accurate factorization method. However, it implies Gaussian noise whereas MSI data follows a Poisson distribution. KL-NMF and PLSA both take this characteristic into account, but unlike often assumed, do not result in the same decomposition. LDA can be seen as a generalization on PLSA, and is often considered the better of the two, but does not offer an advantage in the context of MALDI-MSI data factorization. In our experiments, KL-NMF was able to represent the data most similar to the UMAP visualisation, while surprisingly, it has rarely been used for MSI data analysis so far. Were UMAP exhibits state-of-the-art MSI visualisations, factorization methods such as KL-NMF improve data interpretation. They simultaneous decompose MSI data in both spatial and spectral distributions, offering more possibilities for further analysis of the biochemical structures present in tissues. In this work, we restricted ourselves to the four methods, but extrapolating this investigation to the implications and interpretations of other methods will offer better prospects for the interpretation of MSI data.

Acknowledgements

This work was supported by KU Leuven: Research Fund (projects C16/15/059, C3/19/053, C32/16/013, C24/18/022), Industrial Research Fund (Fellowship 13-0260) and several Leuven Research and Development bilateral industrial projects, Flemish Government Agencies: FWO (EOS Project no 30468160 (SeLMA), SBO project S005319N, Infrastructure project I013218N, TBM Project T001919N; PhD Grants (SB/1SA1319N, SB/1S93918, SB/151622)), This research received funding from the Flemish Government (AI Research Program). Bart De Moor and Melanie Nijs are affiliated to Leuven.AI - KU Leuven institute for AI, B-3000, Leuven, Belgium. VLAIO (City of Things (COT.2018.018), PhD grants: Baekeland (HBC.20192204) and Innovation mandate (HBC.2019.2209), Industrial Projects (HBC.2018.0405)), European Commission: This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 885682); (EU H2020-SC1-2016-2017 Grant Agreement No.727721: MIDAS), KOTK fundation.
The resources and services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by the Research Foundation - Flanders (FWO) and the Flemish Government.
Figure 6 was obtained from https://pubs.acs.org/doi/10.1021/acs.analchem.8b05827 and further permission related to the material excerpted should be directed to the ACS.

References

1. Caprioli RM, Farmer TB, Gile J. Molecular Imaging of Biological Samples: Localization of Peptides and Proteins Using MALDI-TOF MS. *Anal Chem*. Published online 1997. doi:10.1021/ac970888i
2. Lloyd SP. Least Squares Quantization in PCM. *IEEE Trans Inf Theory*.

1982;28(2):129-137. doi:10.1109/TIT.1982.1056489
3. Van Der Maaten L, Hinton G. *Visualizing Data Using T-SNE*. Vol 9.; 2008.
4. McInnes L, Healy J, Melville J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. Published online 2018. http://arxiv.org/abs/1802.03426
5. Siy PW, Moffitt RA, Parry RM, et al. Matrix factorization techniques for analysis of imaging mass spectrometry data. *8th IEEE Int Conf Bioinforma Bioeng BIBE 2008*. Published online 2008:1-6. doi:10.1109/BIBE.2008.4696797
6. Ma S, Dai Y. Principal component analysis based Methods in bioinformatics studies. *Brief Bioinform*. Published online 2011. doi:10.1093/bib/bbq090
7. Xiong XC, Fang X, Ouyang Z, Jiang Y, Huang ZJ, Zhang YK. Feature extraction approach for mass spectrometry imaging data using non-negative matrix factorization. *Fenxi Huaxue/ Chinese J Anal Chem*. 2012;40(5):663-669. doi:10.1016/S1872-2040(11)60544-6
8. Leuschner J, Schmidt M, Fernsel P, Lachmund D, Boskamp T, Maass P. Supervised non-negative matrix factorization methods for MALDI imaging applications. *Bioinformatics*. 2019;35(11):1940-1947. doi:10.1093/bioinformatics/bty909
9. Harn YC, Powers MJ, Shank EA, Jojic V. Deconvolving molecular signatures of interactions between microbial colonies. In: *Bioinformatics*. Vol 31. Oxford University Press; 2015:i142-i150. doi:10.1093/bioinformatics/btv251
10. Keenan MR, Kotula PG. Accounting for Poisson noise in the multivariate analysis of ToF-SIMS spectrum images. *Surf Interface Anal*. 2004;36(3):203-212. doi:10.1002/sia.1657
11. Deepaisarn S, Tar PD, Thacker NA, Seepujak A, McMahon AW. Quantifying biological samples using Linear Poisson Independent Component Analysis for MALDI-ToF mass spectra. *Bioinformatics*. 2018;34(6):1001-1008. doi:10.1093/bioinformatics/btx630
12. Piehowski PD, Davey AM, Kurczy ME, et al. Time-of-Flight Secondary Ion Mass Spectrometry Imaging of Subcellular Lipid Heterogeneity: Poisson Counting and Spatial Resolution. *Anal Chem*. 2009;81(14):5593-5602. doi:10.1021/AC901065S
13. Lee DD, Seung HS. Learning the parts of objects by non-negative matrix factorization. *Nature*. 1999;401(6755):788-791. doi:10.1038/44565
14. Zhang Z-Y, Ding C, Tang J. Note on Algorithm Differences Between Nonnegative Matrix Factorization And Probabilistic Latent Semantic Indexing. doi:10.4156/jcit.vol6.issue9.25
15. Ding C, Li T, Peng W. On the equivalence between Non-negative Matrix Factorization and Probabilistic Latent Semantic Indexing. *Comput Stat Data Anal*. 2008;52:3913-3927. doi:10.1016/j.csda.2008.01.011
16. Blei DM, Ng AY, Edu JB. *Latent Dirichlet Allocation Michael I. Jordan*. Vol 3.; 2003.
17. Faleiros T, De A, Lopes A, De T, Faleiros P. *On the Equivalence between Algorithms for Non-Negative Matrix Factorization and Latent Dirichlet Allocation Induction of Topic-Based Bayesian Networks from Text for the Prediction of Sugar Cane Yields, View Project Applications of Bayesian Theory View Project On the Equivalence between Algorithms for Non-Negative Matrix Factorization and Latent Dirichlet Allocation*.; 2016. Accessed January 15, 2021. https://www.researchgate.net/publication/302488923
18. Liu JX, Wang D, Gao YL, Zheng CH, Xu Y, Yu J. Regularized Non-Negative Matrix Factorization for Identifying Differentially Expressed Genes and Clustering Samples: A Survey. *IEEE/ACM Trans Comput Biol Bioinforma*. 2018;15(3):974-987. doi:10.1109/TCBB.2017.2665557

19. Hanselmann M, Kirchner M, Renard BY, et al. Concise representation of mass spectrometry images by probabilistic latent semantic analysis. *Anal Chem*. 2008;80(24):9649-9658. doi:10.1021/ac801303x
20. Johan J, Van Der Hooft J, Wandy J, Barrett MP, Burgess KE V, Rogers S. Topic modeling for untargeted substructure exploration in metabolomics. doi:10.1073/pnas.1608041113
21. T B, D L, J O, et al. A new classification method for MALDI imaging mass spectrometry data acquired on formalin-fixed paraffin-embedded tissue samples. *Biochim Biophys acta Proteins proteomics*. 2017;1865(7):916-926. doi:10.1016/J.BBAPAP.2016.11.003
22. Verbeeck N, Caprioli RM, Van de Plas R. Unsupervised machine learning for exploratory data analysis in imaging mass spectrometry. *Mass Spectrom Rev*. 2020;39(3):245-291. doi:10.1002/mas.21602
23. 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
24. Alexandrov T. Spatial Metabolomics and Imaging Mass Spectrometry in the Age of Artificial Intelligence. *Annu Rev Biomed Data Sci*. 2020;3(1):61-87. doi:10.1146/annurev-biodatasci-011420-031537
25. Ubaru S, Saad Y. *Fast Methods for Estimating the Numerical Rank of Large Matrices*.; 2016.
26. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine Learning in Python. *J Mach Learn Res*. 2011;12(85):2825-2830. http://jmlr.org/papers/v12/pedregosa11a.html
27. Gillis N. The Why and How of Nonnegative Matrix Factorization. 2014;(June). http://arxiv.org/abs/1401.5226
28. Févotte C, Idier J. Algorithms for nonnegative matrix factorization with the β-divergence. *Neural Comput*. 2011;23(9):2421-2456. doi:10.1162/NECO_a_00168
29. Thi L, Hien K. Algorithms for Nonnegative Matrix Factorization with the Kullback-Leibler Divergence. (arXiv:2010.01935v1 [math.OC]). (679515):1-30. http://arxiv.org/abs/2010.01935
30. Gillis N. *Nonnegative Matrix Factorization*. Society for Industrial and Applied Mathematics; 2020. doi:10.1137/1.9781611976410
31. Hofmann T. *Probabilistic Latent Semantic Analysis*.
32. Hanselmann M, Kirchner M, Renard BY, et al. Concise Representation of Mass Spectrometry Images by Probabilistic Latent Semantic Analysis. *Anal Chem*. 2008;80(24):9649-9658. doi:10.1021/AC801303X
33. GitHub - lmcinnes/enstop: Ensemble topic modelling with pLSA. Accessed December 23, 2020. https://github.com/lmcinnes/enstop
34. Chernyavsky I, Alexandrov T, Maass P, Nikolenko SI. Life and Medical Sciences, J.2 Physical Sciences and Engineering, H.2.8 Database Applications Nikolenko; licensed under Creative Commons License ND German Conference on Bioinformatics. 1998;12:39-48. doi:10.4230/OASIcs.GCB.2012.39
35. Van Der Hooft JJJ, Wandy J, Barrett MP, Burgess KEV, Rogers S. Topic modeling for untargeted substructure exploration in metabolomics. *Proc Natl Acad Sci U S A*. 2016;113(48):13738-13743. doi:10.1073/pnas.1608041113
36. Hooft JJJ van der, Wandy J, Young F, et al. Unsupervised Discovery and Comparison of Structural Families Across Multiple Samples in Untargeted Metabolomics. *Anal Chem*. 2017;89(14):7569-7577. doi:10.1021/ACS.ANALCHEM.7B01391

37. Reder GK, Young A, Altosaar J, et al. Supervised topic modeling for predicting molecular substructure from mass spectrometry. *F1000Research 2021 10403*. 2021;10:403. doi:10.12688/f1000research.52549.1
38. Hoffman MD, Blei DM, Bach F. *Online Learning for Latent Dirichlet Allocation*.
39. Boutsidis C, Gallopoulos E. SVD based initialization: A head start for nonnegative matrix factorization. *Pattern Recognit*. 2008;41(4):1350-1362. doi:10.1016/j.patcog.2007.09.010
40. Yadav SK, Sinha R, Bora PK. An Efficient SVD Shrinkage for Rank Estimation. *IEEE Signal Process Lett*. 2015;22(12):2406-2410. doi:10.1109/LSP.2015.2487600
41. Wall ME, Rechtsteiner A, Rocha LM. Singular value decomposition and principal component analysis. Published online 2003:91-109.
42. Gavish M, Donoho DL. The Optimal Hard Threshold for Singular Values is 4/ √ 3.
43. Smets T, Waelkens E, De Moor B. Prioritization of m/ z-Values in Mass Spectrometry Imaging Profiles Obtained Using Uniform Manifold Approximation and Projection for Dimensionality Reduction. *Anal Chem*. 2020;92(7):5240-5248. doi:10.1021/acs.analchem.9b05764
44. Gaussier E, Goutte C. Relation between PLSA and NMF and implications. *SIGIR 2005 - Proc 28th Annu Int ACM SIGIR Conf Res Dev Inf Retr*. 2005;(August):601-602. doi:10.1145/1076034.1076148
45. Zhang Z-Y. *NMF-Based Models for Tumor Clustering: A Systematic Comparison*.; 2009.
46. Hooft JJJ van der, Wandy J, Barrett MP, Burgess KE V., Rogers S. Topic modeling for untargeted substructure exploration in metabolomics. *Proc Natl Acad Sci*. 2016;113(48):13738-13743. doi:10.1073/PNAS.1608041113

Tables

Table 1: Specifications of the mouse pancreas data (M1, M2, and M3) and human-lymph-node data (P1, P2, P3, and P4). We define the sparsity of the datasets as $1 - \frac{\#_{nonzero}(A)}{\#_{total}(A)}$, which shows the percentage of zero entries in the matrices.

sample	pixels	m/z bins	sparsity	Size (GB)
M1	10 606	14 000	0.030	1.19
M2	14 791	14 000	0.090	1.66
M3	6 937	14 000	0.095	0.78
P1	572 173	8 000	0.005	36.62
P2	552 701	8 000	0.015	33.33
P3	354 444	12 000	0.041	34.03
P4	275 133	12 000	0.020	26.41

Table 2: Numerical analysis of NMF, KL-NMF, PLSA, and LDA on three mouse pancreas samples: M1, M2, and M3. For the reconstruction error, we use three measures: the relative ℓ_1 and ℓ_2-norm, and the Kullback-Leibler divergence. Each experiment ran until convergence in the update was reached, with a threshold of 1e-3.

sample	method	ℓ_1	ℓ_2	KL	time (s)	mem. usage (GB)
M1	NMF	0.4516	**2.486e-6**	0.2108	**2.6 ± 0.0**	**1.1 ± 1.0**
	KL-NMF	**0.3955**	3.233e-6	**0.1480**	60.9 ± 3.1	6.3 ± 1.4
	PLSA	0.8700	0.1350	0.1658	44.4 ± 0.9	6.6 ± 2.3
	LDA	-	-	1.1770	3 970.4 ± 27.7	1.2 ± 0.0
M2	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
M3	NMF	0.5719	**1.266e-5**	0.2872	**1.6 ± 0.0**	**0.2 ± 0.4**
	KL-NMF	**0.5585**	1.442e-5	**0.2782**	35.8 ± 4.7	3.6 ± 1.0
	PLSA	0.7465	0.3067	0.2919	27.1 ± 0.3	2.1 ± 1.6
	LDA	-	-	0.5658	3 761.2 ± 51.6	0.8 ± 0.002

Table 3: Numerical analysis of NMF, KL-NMF, PLSA, and LDA on four human lymph samples: P1, P2, P3 and P4. For the reconstruction error, we use three measures: the relative ℓ_1 and ℓ_2-norm, and the Kullback-Leibler divergence. Each experiment for NMF, KL-NMF, and PLSA ran until convergence in the update was reached, with a threshold of 1e-3. For LDA convergence was not reached within the maximum number of 12 iterations.

sample	method	ℓ_1	ℓ_2	KL	time (s)	mem. usage (GB)
P1	NMF	0.8042	**2.740e-6**	0.5559	**120 ± 3**	104 ± 2
	KL-NMF	**0.7865**	2.869e-6	**0.5365**	2 919 ± 133	218 ± 1
	PLSA	4.9300	0.1475	1.1545	2 159 ± 46	422 ± 8
	LDA	-	-	1.0710	46 297 ± 811	**37 ± 1**
P2	NMF	0.6213	**1.080e-7**	0.4185	**112 ± 8**	92 ± 3
	KL-NMF	**0.6065**	1.147e-7	**0.3308**	2 687 ± 105	196 ± 2
	PLSA	3.0266	0.0778	0.4077	1 754 ± 69	307 ± 15
	LDA	-	-	1.2304	42 981 ± 2 415	**33 ± 1**

P3	NMF	0.6210	**3.233e-7**	0.3342	**119 ± 12**	93 ± 3
	KL-NMF	0.5958	3.406e-7	**0.3133**	3 761 ± 52	197 ± 1
	PLSA	4.2952	0.1926	0.3418	1 856 ± 151	303 ± 15
	LDA	-	-	0.9790	42 455 ± 264	**35 ± 2**
P4	NMF	0.6823	**2.557e-7**	0.3899	**102 ± 11**	73 ± 2
	KL-NMF	**0.6402**	2.706e-7	**0.3531**	2 096 ± 60	155 ± 1
	PLSA	3.4761	0.1882	0.3899	1 379 ± 120	242 ± 14
	LDA	-	-	0.9769	33 897 ± 261	**28 ± 2**