Concise Representation of Mass Spectrometry Images by Probabilistic Latent Semantic Analysis –Technical report–

Michael Hanselmann[1], Marc Kirchner[1,‡], Bernhard Y. Renard[1,‡],
Erika R. Amstalden[2], Kristine Glunde[3], Ron M. A. Heeren[2],
Fred A. Hamprecht[1,⋆]

October 8, 2008

[1] Heidelberg Collaboratory for Image Processing (HCI), Interdisciplinary Center for Scientific Computing (IWR), University of Heidelberg, Speyerer Strasse 4, Heidelberg, Germany

[2] FOM-AMOLF, FOM-Institute for Atomic and Molecular Physics, Kruislaan 407, Amsterdam, The Netherlands

[3] Russell H. Morgan Department of Radiology and Radiological Science, Johns Hopkins University School of Medicine, Baltimore, Maryland 21205, USA

[‡] contributed equally

[⋆] corresponding author. Email: fred.hamprecht@iwr.uni-heidelberg.de.

Imaging Mass Spectrometry (IMS) is a promising technology which allows for detailed analysis of spatial distributions of (bio-)molecules in organic samples. In many current applications, IMS relies heavily on (semi-)automated exploratory data analysis procedures to decompose the data into characteristic component spectra and corresponding abundance maps, visualizing spectral and spatial structure. The most commonly used techniques are Principal Component Analysis (PCA) and Independent Component Analysis (ICA). Both methods operate in an unsupervised manner. However, their decomposition estimates usually feature negative counts and are not amenable to direct physical interpretation. We propose Probabilistic Latent Semantic Analysis (pLSA) for non-negative decomposition and the elucidation of interpretable component spectra and abundance maps. We compare this algorithm to PCA, ICA and non-negative PARAFAC and show on simulated and real-world data that pLSA and non-negative PARAFAC are superior to PCA or ICA in terms of complementarity of the resulting components and reconstruction accuracy. We further combine pLSA decomposition with a statistical complexity estimation scheme based on the Akaike Information Criterion (AIC) to automatically estimate the number of components present in a tissue sample dataset and show that this results in sensible complexity estimates.

Imaging Mass Spectrometry (IMS)[1,2] has evolved into a promising technology which allows detailed analysis of spatial distributions of (bio-)molecules, including but not limited to proteins, peptides, lipids or metabolites.[3] Applications include disease studies such as cancer grading[4,5] or Alzheimer's disease studies[6] as well as drug distribution and metabolism analysis.[7,8] However, the enormous size of data sets acquired with state-of-the-art instrumentation renders a direct manual analysis infeasible. Often thousands of spectra each of which comprise thousands of mass channels have to be analyzed and methods of automated preprocessing become indispensable. Many real-world scenarios require exploratory data analysis, where little or no prior information on the composition of a sample is available. This is particularly true when IMS is used as a discovery technique. In such an unsupervised setting it is useful to decompose the spectral image into a small number of characteristic component spectra and corresponding abundance maps to visualize the spectral and spatial structures in the data. These structures are not directly accessible since manual inspection of individual *m/z* channel abundance maps is not only time-consuming, but also tends to neglect spatial and spectral ignoring valuable information regarding interactions of biomolecules. Nonetheless, low-dimensional representations that capture these correlations and make fast and efficient analyses of huge datasets possible are desirable.

Conventional techniques such as Principal Component Analysis (PCA)[9–12] and Independent Component Analysis (ICA)[13] have successfully been applied in such settings, but they suffer from a number of drawbacks. The component spectra found by PCA are mutually orthogonal and hence feature negative counts. This implies that PCA cannot recover the true mass spectra of the tissue components. A rotation of the coordinate system as performed by VARIMAX-enhanced PCA[14] does not solve this problem. ICA suffers from similar non-negativity problems since its objective function tries to minimize the mutual information of the reconstructed components rather than making use of the prior knowledge, that mass spectra are positive. Although physical interpretation of negative ion abundance rates is difficult, PCA is still widely used.[9–12] Broersen[14] considered this problem and applied a non-negative PARAFAC (here abbreviated NN-PARAFAC) to IMS data, Smentkowski[15] applied a multivariate curve resolution method using constrained least squares algorithms. Mathematically, all these techniques perform a bilinear factor analysis, though based on different constraints and objective functions. An inherent problem for all those methods is that the number of components (tissue types) has to be specified – either prior to the decomposition process or in a post-processing step that selects the number of components heuristically, e.g. based on the percentage of variance represented by a given number of components. We therefore propose to combine pLSA with a statistical model selection scheme based on the Akaike information criterion (AIC)[16] which allows for an automated estimation of the number of components in the dataset.

We next revisit PCA, ICA as well as NN-PARAFAC and address their shortcomings that motivate the application of Probabilistic Latent Semantic Analysis (pLSA). The latter two are specifically suitable for IMS data analysis since they do not violate physical properties. We then introduce the AICc-corrected pLSA. Finally, we compare the presented techniques on simulated and real-world data and end with a discussion of our results.

Methods

Principal Component Analysis

Principal Component Analysis (PCA)[17] is a well-known and widely used technique for unsupervised data analysis and dimensionality reduction. PCA essentially performs a linear orthogonal transformation of the data domain. Let $x_l \in \mathbb{R}^{|C|}, l = 1, ..., |S|$ be a set of $|S|$ observed spectra each comprising $|C|$ m/z channels and $X = (x_1, ..., x_{|S|})$ the data matrix where each column holds an observed spectrum x_l. Further define $\tilde{X}$ as its corresponding mean centered version. PCA finds the principal components by diagonalizing the estimated data covariance matrix $\tilde{X}\tilde{X}'$ yielding the principal components (i. e. axes of the new coordinate system) ordered by decreasing non-negative eigenvalues (i. e. the observed variance along the PC axes). PCA projects onto the first k principal components keeping the linear subspace with the largest variance and yielding the best linear k-rank approximation to the data in the least-squares sense. In most applications, one is only interested in the first few components, assuming they hold the most relevant information. Usually the total percentage of variance retained after projection is used as an indicator of how many components should be used. The rationale behind this approach is that in many cases high variance along a direction will allow for good class separation. However, this assumption need not hold.

Mathematically, PCA can also be formulated as a factor analysis model, decomposing the data into two matrices $\tilde{X}' = AB$ where $A'A$ is diagonal and $B'B = I$.[18] Alternatively, PCA also follows from the singular value decomposition (SVD)[17] of $\tilde{X}'$, i.e. $\tilde{X}' = UDV'$ where U and V are orthogonal, D is diagonal and the columns of UD are the principal components.

Independent Component Analysis

Independent Component Analysis (ICA) is a factor analysis model that was introduced in the context of blind-source-separation.[17] The key assumption of ICA is that the source signals, i.e. the characteristic component spectra in our case, are statistically independent with a non-Gaussian distribution. Typical preprocessing steps include centering and whitening which lead to zero mean, unit variance and zero correlation[17] as well as dimensionality reduction. The latter two are often solved with PCA.[19,20] However, unlike PCA which is restricted to second degree cross-moments, ICA uses all cross-moments.

Let $Y = (y_1, ..., y_{|T|})$ be the matrix of independent components which are represented by the set T. The task is to determine the mixing matrix $A = W^{-1}$ and find

$$Y = W\tilde{X} \tag{1}$$

such that the components y_k become maximally independent. This is achieved by either minimizing mutual information which is a measure of the mutual dependence of two random variables[17] or maximizing non-Gaussianity.[19]

Non-negative PARAFAC

PARAFAC (PARAllel FACtors analysis),[21] also known as CANDECOMP (CANonical DECOMPosition),[22] is a multi-way decomposition method. As indicated by Kiers[23] it can also be seen as a constrained two-way PCA-model. In conjunction with non-negativity constraints for the modes this essentially corresponds to a standard non-negative matrix factorization.[24] PARAFAC does not approximate probability densities by marginals and lacks a proper probabilistic foundation. The solution is normally found by alternating least squares,[14,25] minimizing the squared reconstruction error, implicitly assuming Gaussian noise.

Probabilistic Latent Semantic Analysis

Probabilistic Latent Semantic Analysis (pLSA) has been proposed in the realm of automated text analysis[26] and has become a standard method for structure and similarity identification in semantic document analysis. It can be described as a linear model with one latent variable t

$$p(s,c) = \sum_{t \in T} p(t) p(s|t) p(c|t) \tag{2}$$

where s is a document, c a word and t a topic, the hidden variable. In the case of IMS, a spectrum can be considered a document, an m/z channel corresponds to a word and a given tissue type corresponds to a topic. In the proposed model, each single tissue type is characterized by a distinct distribution and each acquired spectrum is regarded as a specific mixture of these structures. The decomposition problem is solved by the following Expectation Maximization (EM) procedure[26] with E-step

$$p(t|s,c) = \frac{p(t)p(s|t)p(c|t)}{\sum_{t' \in T} p(t')p(s|t')p(c|t')} \tag{3}$$

and M-step

$$\begin{aligned} p(c|t) &\propto \sum_{s \in S} X(c,s) p(t|s,c) && (4) \\ p(s|t) &\propto \sum_{c \in C} X(c,s) p(t|s,c) && (5) \\ p(t) &\propto \sum_{s \in S} \sum_{c \in C} X(c,s) p(t|s,c). && (6) \end{aligned}$$

Here, $X(c,s)$ is the abundance of channel c in spectrum s. This can be reformulated as an SVD-like decomposition[26] by

$$P = \widehat{U}\widehat{D}\widehat{V}' \tag{7}$$

where we define $\widehat{U} = p(s|t)$, $\widehat{V} = p(c|t)$ and $\widehat{D} = diag(p(t))$ and where the joint probability P corresponds to the spectra-wise normalized data matrix X.

pLSA provides a probability distribution over the spectral dimension for each tissue type as the resulting components are normalized and non-negative. Unlike PCA, ICA

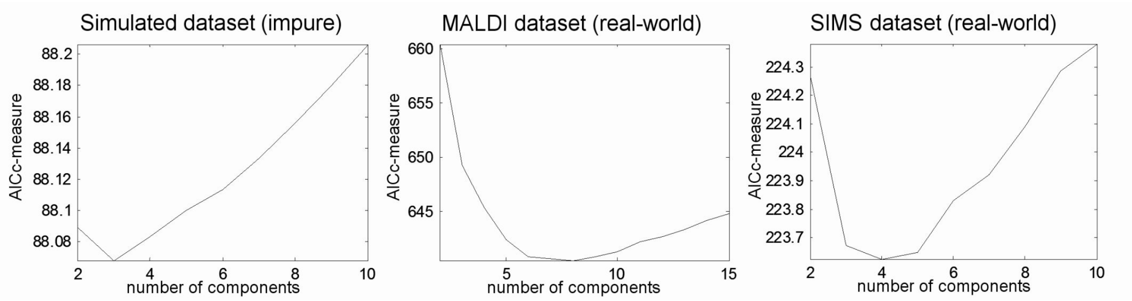

Figure 1: The AICc curves for various datasets used in this study. The simulated dataset is a mixture of three tissue types. The AICc criterion correctly identifies the number of mixture components.

and NN-PARAFAC, it has a sound statistical foundation and defines a proper generative model of the data.[26] This provides physical interpretability and allows to identify the discriminating peaks for a specific tissue type within a spectrum. pLSA is equivalent to non-negative matrix factorization with a Kullback-Leibler (KL) divergence measure,[27] defined by

$$D_{KL}(P||Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)} \tag{8}$$

which quantifies the difference between two probability distributions P and Q. Since peak intensities are ion counts that follow a Poisson distribution, a KL divergence measure is appropriate,[28] further motivating the pLSA approach.

AICc-controlled pLSA

A common feature of decomposition methods like NN-PARAFAC and pLSA is that the number of components $k = |T|$ has to be specified. This property is beneficial if such prior knowledge is available. For scenarios where this is not the case, we propose a method that is capable of automatically estimating k by using a statistical complexity estimation scheme. It is possible to select from a variety of different model selection criteria.[17] Compared to alternatives such as the Bayesian Information Criterion (BIC) or Minimum Description Length (MDL), AIC-type criteria for feature selection tend to select slightly more features if the model specification does not reflect the true generating process of the data;[17,29] it is therefore more "conservative" and hence preferable in this application in which one would rather estimate slightly too many components than loose subtle differences by penalizing model complexity too heavily.

Starting with $k = 2$, we run the pLSA-algorithm repeatedly with increasing k. We then apply a corrected Akaike Information Criterion (AICc)[16,30] (which is based on the KL-divergence) to automatically select the correct model. In our case, the AICc-type criterion is defined as

$$AICc(k) = \underbrace{-\frac{2}{N}\mathcal{L}(k)}_{data\ likelihood} + \underbrace{\frac{2}{N}M\sigma^2}_{penalty\ term} + \underbrace{\frac{1}{N}\frac{2M(M+1)}{N-M-1}}_{correction\ term} \tag{9}$$

where N is the number of observations ($|S| \cdot |C|$) and $\mathcal{L}(k)$ is the data log-likelihood

$$\mathcal{L}(k) = \sum_{s \in S} \sum_{c \in C} X(c,s) \log \sum_{t=1}^{k} p(c|t)p(t|s). \tag{10}$$

The number of free model parameters $M = k \cdot (|S| + |C|)$ is equivalent to the number of elements in the two solution matrices representing $p(c|t)$ and $p(s|t)$. The first term in equation 9 measures how good a model fits to the observed data, the second term penalizes complexity to prevent overfitting, and the last term corrects the AIC for small sample sizes. To robustly estimate the noise variance σ^2 we calculate the median of the squared residuals of the difference between the original spectra with the observed data and their spatially smoothed version (rectangular 3×3 mean filter on each m/z slice), assuming that neighboring tissue consists of similar tissue mixtures. The idea behind AICc-controlled pLSA is to stop the iterations as soon as we can be sure that decompositions with a higher number of components will not yield a lower AICc value than the current minimum. We calculate equation 9 for increasing k until a stopping criterion is met and then take the value of k for which the minimum of the resulting AICc-curve is attained as an estimate for the optimal model complexity (see figure 1). The early stopping criterion is defined as follows.[31] It holds that $\mathcal{L}(k) \leq 0 \; \forall k$ and thus that we can abort the calculations if the (strictly increasing) penalty term for the current k (eq. 9) is higher than any previous value in the AICc-curve. However, this bound is not very tight and therefore not practical. We therefore set a reasonable upper bound $\tilde{k}$ for the number of components, for example 100. Since $-2\mathcal{L}(k)/N$ is monotonously decreasing, we can stop as soon as

$$-\frac{2}{N}\mathcal{L}(\tilde{k}) + \frac{2}{N}M\sigma^2 + \frac{1}{N}\frac{2M(M+1)}{N-M-1} > \min_{2 \leq k \leq k_{curr}} AICc(k) \tag{11}$$

is met for the current number of components k_{curr} (where $M = k_{curr} \cdot (|S| + |C|)$). With this criterion we are guaranteed to find the minimum of the AICc curve within the given bounds $[2; \tilde{k}]$ with low computational overhead.

Sparsity

After decomposing the data it is often of interest to identify decisive peaks that allow for a discrimination of different tissue types. A peak is decisive if it is only present in one or few of the k component spectra. We employ Hoyer's sparsity measure[32]

$$sparsity(u) = \frac{\sqrt{k} - (\sum |u_i|)/\sqrt{\sum u_i^2}}{\sqrt{k} - 1} \tag{12}$$

where u' is a $1 \times k$ row vector of the matrix that holds $p(c|t)$. The measure is based on the ratio between $\sum |u_i|$ (the L_1-norm) and $\sqrt{\sum u_i^2}$ (the L_2-norm), assigning a high sparsity value to a vector u if the intensity distribution over the components u_i is sparse and the respective channel can therefore be used to discriminate between components or tissue types.

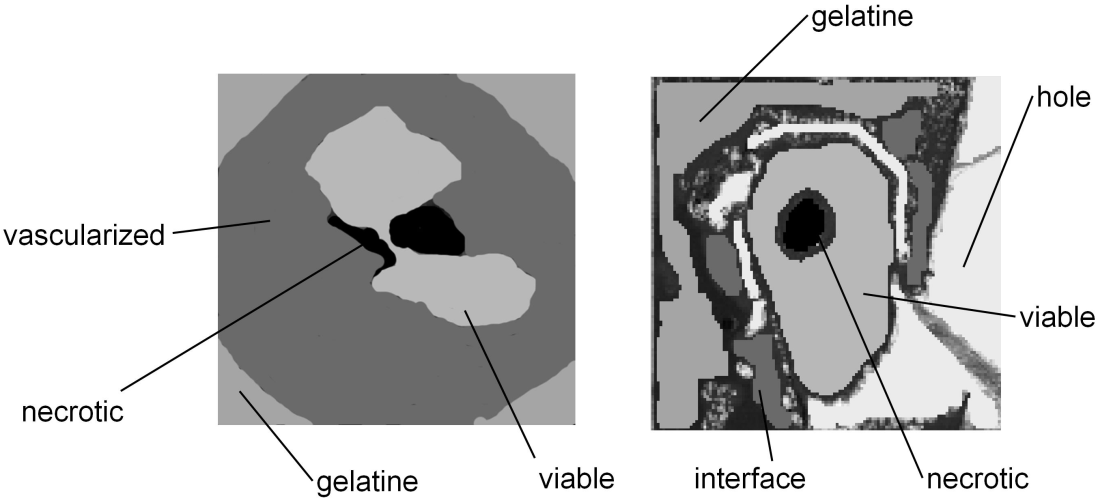

Figure 2: Labeling of the MALDI (left) and SIMS dataset (right) based on total ion count images and histologically stained parallel slices (cf. supporting information A).

Experiments

Simulated data

We first tested the algorithms on two simulated data sets which were generated in the following manner: From a real-world breast cancer dataset (SIMS, see figure 2) we selected three regions of interest that had previously been assigned to different classes by an expert. We treated the average spectra of those regions as characteristic spectra and mixed them according to the two mixture maps shown in figure 3 (top panel) yielding a “pure” and an “impure” ground truth dataset. Two of the characteristic spectra featured considerable spectral overlap which complicated the decomposition process. Before decomposition we added Poisson noise.

Real-world data

In order to evaluate the methods in multiple real-world scenarios, two different real-world datasets were used. Both datasets stem from human breast cancer tissue grown in mice. The first one contains MDA-MB-231, a highly metastatic tumor, and was acquired on a modified MALDI TRIFT II instrument coupled with a time-of-flight (TOF) mass analyzer. For the acquisition of the second set which features MCF-7, a weakly metastatic and estrogen-sensitive tumor, a Physical Electronics TRIFT II TOF SIMS equipped with both 115In+ liquid metal ion gun and Au+ liquid metal ion gun was used.

The data were acquired with the following protocol: the samples were frozen in embedding gelatine, cryo-sectioned and mounted on a cold indium tin oxide-coated glass slide. A 2,5-Dihydroxybenzoic acid (DHB) 30mg/mL in 50% acetonitrile/0,1% trifluoroacetic acid was applied using an air driven thin liquid chromatography spray for MALDI (Matrix Assisted Laser Desorption Ionization) IMS and ME-SIMS (Matrix Enhanced Secondary Ion Mass Spectrometry). After drying, the tissues were additionally sputter coated with 1nm gold. The tissue was not washed prior to SIMS analysis. The spectral resolution

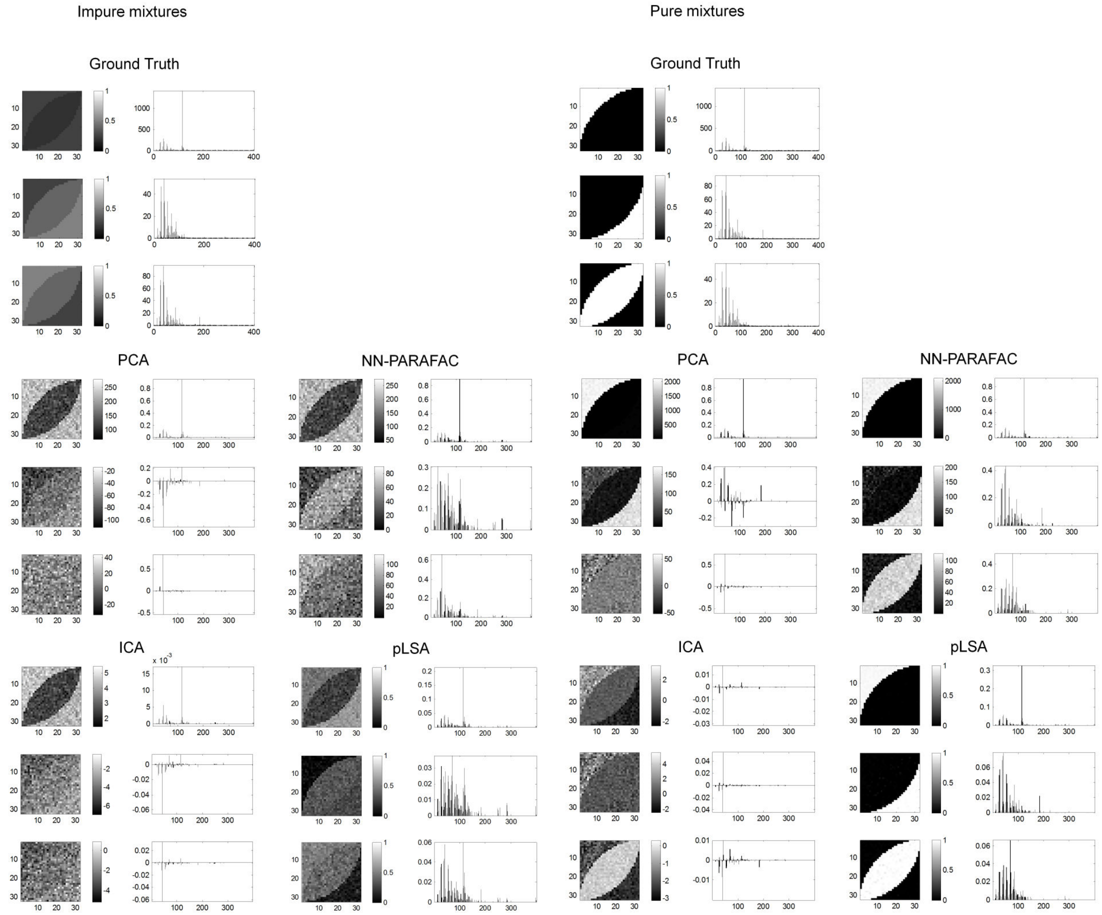

Figure 3: Decomposition results on the simulated dataset with non-pure (left) and pure (right) mixtures. The PCA component abundance maps are ordered from top to bottom according to the eigenvalue associated with the principal components. For ICA, NN-PARAFAC and pLSA the order of components is arbitrary and has been permuted such that the observed abundance maps coincide in their ordering with the ground truth. NN-PARAFAC and, especially, pLSA clearly perform better than PCA and ICA (see also text). For pure and impure mixtures, it was not beneficial to additionally take into account subsequent principal components which mainly contained noise.

was rebinned to 0.1 Da, the spatial resolution was rebinned by a factor of two. Consequently, one pixel equals 150 × 150μm in the MALDI set and 70 × 70μm in the SIMS set. For both, a Hematoxylin-Eosin-(HE)-stained parallel slice is available. Despite some topological differences between the stained and IMS-subjected slices, the stained slices can be used as gold standards enabling further evaluation of the decomposition quality of the four methods. The label information was not used in the decomposition process.

Evaluation criteria

Since two different tumor types have been used for MALDI and SIMS analysis, we refrain from comparing the resulting datasets against each other, but rather concentrate on the performance of the decomposition techniques presented above in each measurement. One of the main motivations for the application of decomposition methods like PCA or pLSA lies in their potential to achieve dimensionality reduction while keeping the relevant

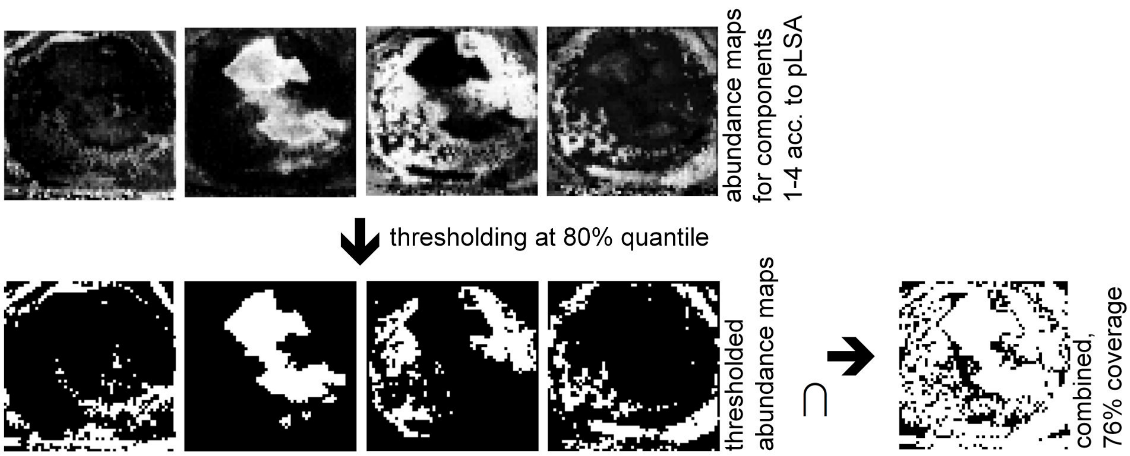
abundance maps for components 1-4 acc. to pLSA
thresholding at 80% quantile
thresholded abundance maps
∩
combined, 76% coverage

Figure 4: Complementarity estimation example for the pLSA decomposition of the MALDI set at the 80% quantile: the first row shows the abundance maps which are thresholded (second row). This means that the top 20 percent of the most intense pixels are set to one and the remaining pixels to zero. The percentage of white pixels in the combined image (right) is an indicator for the complimentarity of the components (see also table 2).

information. When PCA is applied in practical applications, often only a few principal components are considered since the analysis of hundreds of components is simply impractical. Thus, we decided to compare the described methods in a scenario where the number of components is limited to only a few.

We based our evaluation on three criteria: reconstruction quality, complementarity of the resulting components and visual inspection.

The first criterion measures how well the original data can be reconstructed after factorization favoring decompositions that explain the observed data with high accuracy. We calculated the reconstruction error between the observed and reconstructed spectra with respect to three different measures to avoid bias towards one of the methods under consideration: the L_1-norm, the L_2-norm and KL divergence. A detailed description on how these measures have been obtained can be found in supporting information C.

Limiting the number of components (as in many real-world scenarios) means that PCA and ICA can no longer perfectly reconstruct the data. The reconstruction errors are affected by the magnitude of the reduction as well as the stopping criterion for the iterative methods. Since we applied PCA for dimensionality reduction as a preprocessing step for ICA (see ICA section), the estimates for those two methods were the same.

To simplify interpretation, it is often desirable to decompose the data into clearly distinguishable components and the second criterion therefore quantifies the complementarity of the resulting abundance maps. For a given number of components k and a given decomposition method, the complementarity is measured as follows: first, the k estimated abundance maps are thresholded at various quantiles between 95% and 50%. Then, we estimated the complementarity for each quantile by calculating the percentage of area covered after overlaying the k thresholded abundance maps, i. e. the percentage of pixels with an intensity value greater than zero (see figure 4). In the case of PCA and ICA we also calculated the corresponding lower quantiles retaining only the pixels with the most negative contributions to allow for a fair comparison. This was necessary as some structures were better reflected by areas with negative intensities. We then used the best out of the 2^k possible combinations of upper/lower thresholded abundance maps for each method and quantile. The higher the coverage index, the more complementary the analyzed components are.

Data Processing

For ICA and NN-PARAFAC calculations, the freely available MATLAB toolboxes FastICA[20] and N-way[25] were used. We defined a relative change in the fit of 10^{-6} as stopping criterion for both NN-PARAFAC and pLSA. The data sets were baseline-corrected by subtraction of the channel-wise minimum with respect to all spectra in the dataset. We performed feature extraction by an in-house implementation of a threshold-based peak picker to keep the relevant information and simultaneously decrease computation time. The reported results are based on extracted peak-lists.

Results

We first compared the proposed methods on simulated data and confirmed the simulation results on real-world data. In order to minimize random effects, all non-deterministic methods were restarted five times and for each method, the best result was checked.

The simulated data was created from three tissue types and we performed the decompositions with the respective number of components. For the (real-world) MALDI set we expected four different regions in the sample: viable or active tumor, necrotic tissue, vascularized region and embedding gelatine. Based on this prior knowledge (and not in favor of one of the methods), we performed the decompositions with four components for which PCA kept 91% of the total variance. Reconstruction accuracies and complementarity estimates for a varying number of components are presented in supporting information E.

For the (real-world) SIMS set we were interested in the three tissue types and the gelatine region described above, but we furthermore assumed a fifth area since parts of the tissue were torn prior to analysis. As can be seen from the label map in figure 2, in those areas the indium tin oxide-coated glass slide is exposed and thus, we expected an indium peak at 115 Da.

The computation time needed for NN-PARAFAC and pLSA is higher than for PCA and ICA. Whereas for PCA and ICA their calculations were completed in less than one second on the SIMS-set, pLSA required ≈30-60 seconds and NN-PARAFAC up to 240 seconds. The AICc-type-enhanced pLSA needs several passings as we need to calculate decompositions with an increasing number of components. However, often only a few iterations (5-15) are necessary and the calculations finish in reasonable time. We indicate that these measures are highly dependent on the parameter settings like the maximum number of iterations as well as on the random initialization. Furthermore, we used interpreted MATLAB code and computational efficiency was not emphasized.

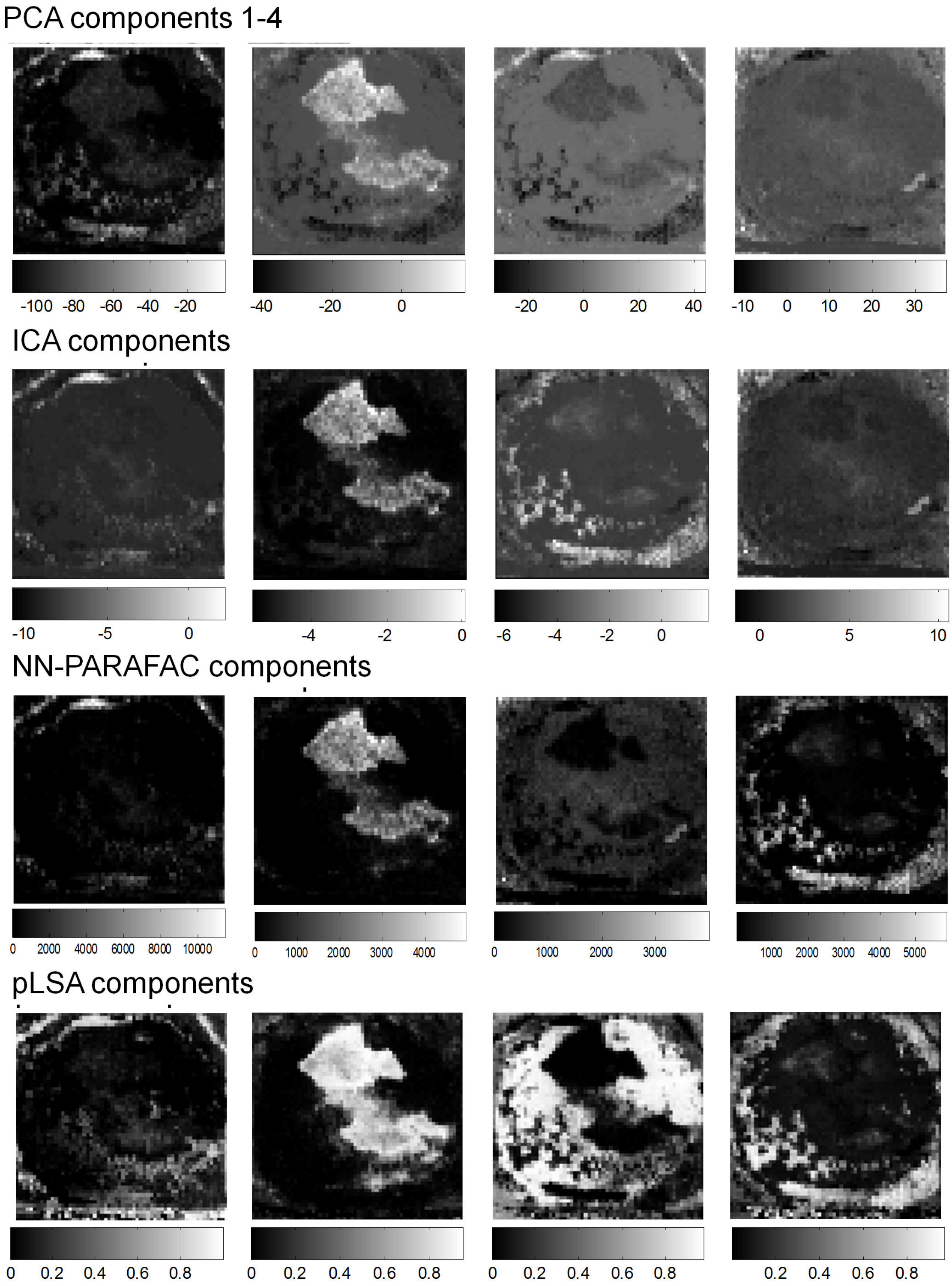

Figure 5: Decomposition of the MALDI set with four components. Again, the ICA, NN-PARAFAC and pLSA components have been reordered to match the PCA components for which the ordering is unique. We further inverted some of the PCA and ICA components for better visual comparison. Only NN-PARAFAC and pLSA have entirely non-negative abundance maps, and only pLSA components are normalized and hence interpretable as tissue probability by definition. The more black-and-white a suite of components, the better their complementarity, cf. table 2. Please refer to the text for interpretations.

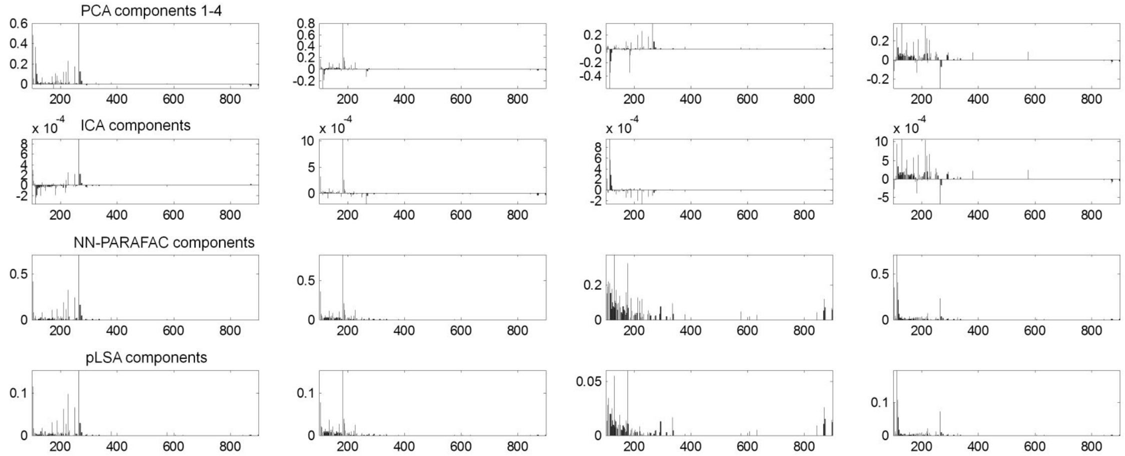

Figure 6: The characteristic spectra of the four tissue types from the MALDI sample (see figures 2 and 5) as reconstructed by PCA, ICA, NN-PARAFAC and pLSA. The latter two result in all positive spectra and therefore allow direct physical interpretability. The most prominent peak in component spectrum two – which seems to correspond to the viable tumor area – lies at 184.5 Da (see also figures 5 and 8). This corresponds to recent findings[33] that show that Phosphocholine which appears at that mass position seems to play an important role in the discrimination of necrotic and viable tumor tissue.

Discussion

Simulated data

The decomposition results on our simulated data sets are illustrated in figures 1 and 3.

Impure mixtures. The abundance map and the characteristic spectrum of the first component are well reconstructed by all methods; the dominant indium peak at 115 Da is well visible. Nevertheless, PCA and ICA completely fail to recover the remaining two components which feature considerable spectral overlap. pLSA is able to recover significantly more structure which can be seen from the abundance maps in figure 3. NN-PARAFAC performs nearly as well as pLSA, but does worse for the third component.

Pure mixtures. As expected, the decomposition task for pure mixtures is simpler and yields better results. PCA exhibits difficulty in extracting the second and third component, but is able to reconstruct the first component very well. ICA shows problems with representing all three components and even fails to extract the indium peak. Possible explanations are given below. In contrast, NN-PARAFAC and pLSA are able to deliver a convincing result and clearly outperform PCA and ICA. In spite of noise being present in the data, the reconstruction of the abundance maps is highly accurate for all three components.

The spectral components estimated by NN-PARAFAC and pLSA are much closer to the ground truth spectra than their PCA and ICA counterparts and outperform PCA and ICA with respect to reconstruction of major peaks (see supporting information F). The AICc-type criterion correctly estimated the number of components in the data set to be three.

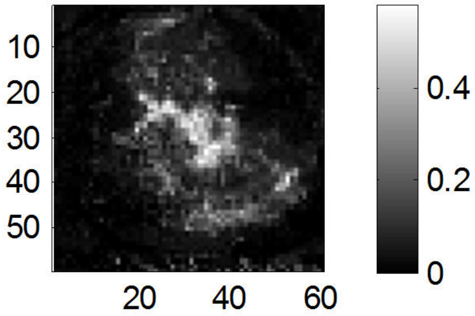

Figure 7: This component of an eight-component decomposition with pLSA is convincingly correlated with the necrotic area in the label map. See also supporting information D for complete eight-component decompositions for all methods.

Real-world data

MALDI dataset. Results are shown in figures 5 and 6. The reconstruction accuracy of both NN-PARAFAC and pLSA is high (see table 1). pLSA delivers the best result with respect to KL-divergence, but it does also well regarding the other norms. NN-PARAFAC ranks highest with respect to the L_2-norm that is the only metric for which PCA and ICA also give a good result. We observe, that in a setting where only a few components are analyzed, NN-PARAFAC and pLSA outperform PCA and ICA with respect to reconstruction accuracy. For PCA and ICA, the complementarity of the estimated components is low (see table 2) and the assignments of reconstructed components to tissue types is not obvious, especially for components three and four. In contrast, the complementarity of the abundance maps calculated by pLSA is very high and we get a clear spatial separation of the four types which simplifies interpretation. Component two seems to represent the viable part of the tumor, component three the vascularized region and components one and four seem to stand for gelatine. NN-PARAFAC performs better than PCA and ICA but does not pick up the vascularized part very well. The contrast of the abundance maps is lower than that of the respective pLSA components which mirrors in the lower complementarity estimation values (see table 2). All methods have problems to separate the necrotic from the viable part in the four component decomposition. The AICc criterion opted for a total of eight components. Manual inspection showed that this basically leads to splitting of the four components of interest described above for which we are currently evaluating biological reason. Indeed, the necrotic part of the tumor is much better represented in the eight component solution as shown in figure 7.

In figure 8, an excerpt of the m/z range of the dataset between 170 and 190 Da is displayed for the pLSA decomposition and we show how the sparsity criterion described earlier can be used to automatically identify discriminating peaks. The spectral components estimated by PCA and ICA (cf. figure 6) are, as expected, partly negative. The Phosphocholine-peak at 184.5 Da is detected as the most dominant peak by all methods, but physical interpretation of the negative parts of the PCA and ICA component spectra is difficult. This does not render the distribution of major peaks in the PCA and ICA components uninformative, but the characteristic spectra of the underlying tissue types

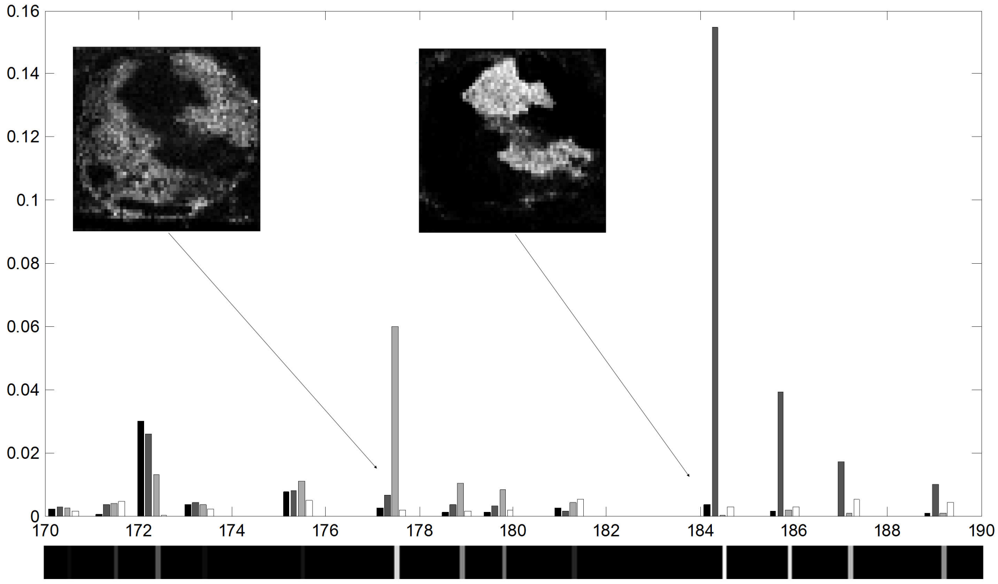

Figure 8: An excerpt of the *m/z* range of the MALDI dataset between 170 and 190 Da. The bar plot shows the intensities of these channels for the four components obtained with pLSA (from left to right). The bar at the bottom color-codes the level of sparsity for the intensity distribution for each *m/z* channel where light color indicates high sparsity (discriminating peak) and dark color indicates low sparsity. We also give abundance maps corresponding to those *m/z* channels with the highest sparsity value. At 177.5 Da, we see a highly intense peak in the characteristic spectrum of tissue type three and low intensities for the characteristic spectra of the other components. Apparently, this peak is typical for tissue type three and can be used to distinguish this component from the others. In contrast, if a peak appears with equal intensity in all four component spectra, it has no discriminatory power.

are not directly revealed.

SIMS dataset. Decomposition results are given in figure 9. The reconstruction error for PCA and ICA is significantly higher than for NN-PARAFAC and pLSA and the complementarity of the estimated components is not well expressed (cf. tables 1 and 2). The contrast in the PCA and ICA abundance maps is very low and assigning these components to regions in the label map is difficult. In comparison, the NN-PARAFAC and pLSA solutions show higher contrast and complementarity. In the NN-PARAFAC and pLSA decompositions, the most prominent peak in the spectra corresponding to the first and third component (cf. supporting information B) lies at 115 Da (indium). Thus, components one and three seem to correspond to background/holes, component two and four to tumor and interface region and component five seems to represent gelatine. Necrotic and viable part can be distinguished as well. The NN-PARAFAC result is similar to the pLSA solution; the reconstruction error is slightly reduced whereas the complementarity estimation shows that the components estimated with pLSA are slightly more sparse.

The AICc-type-controlled pLSA was capable of automatically estimating the number of components and only slightly preferred the decomposition with four components over the five component solution (see figure 1).

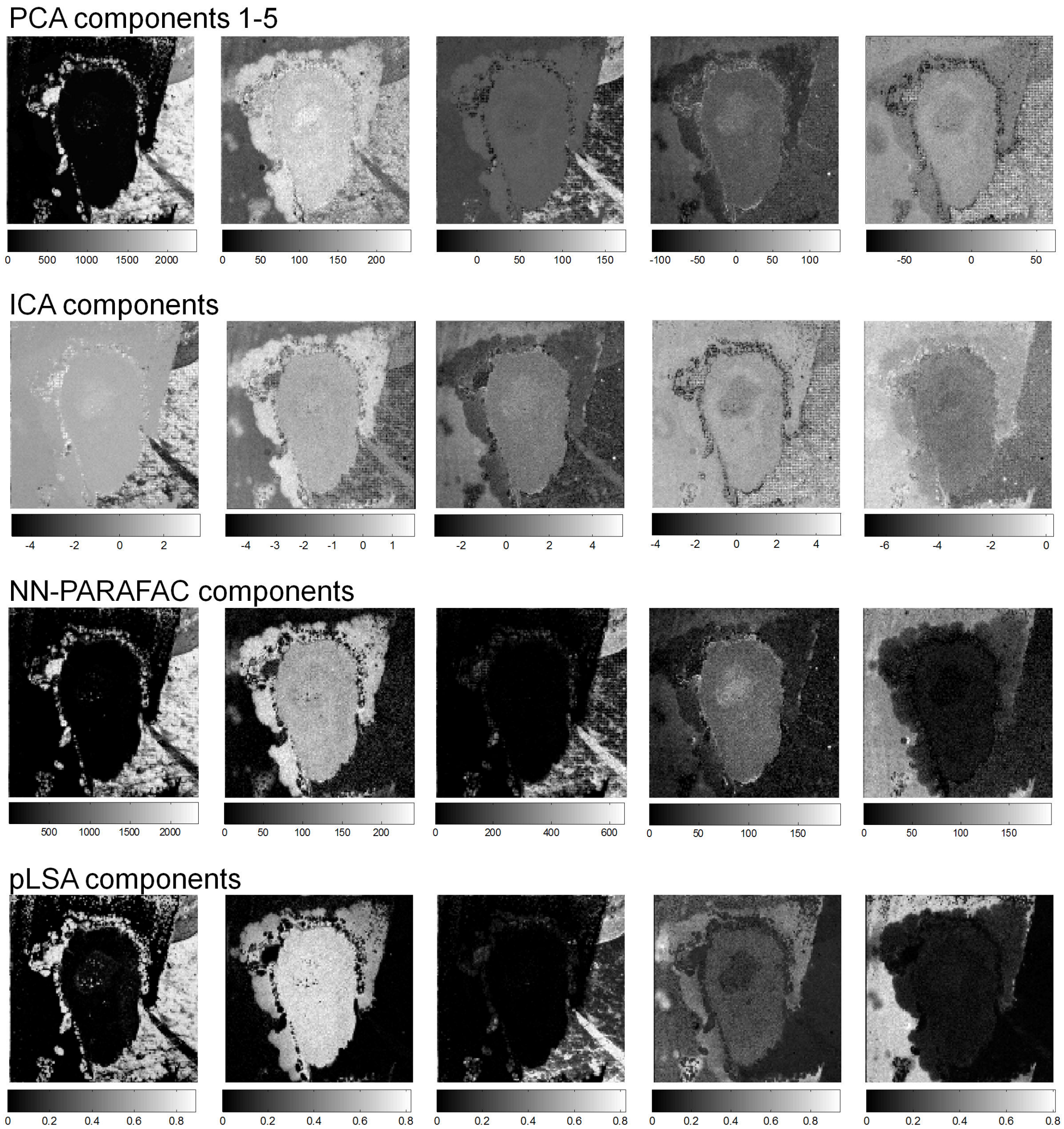

Figure 9: Decomposition of the SIMS set with five components; see comments on figure 5. The corresponding spectral components are shown in supporting information B.

	Reconstruction Error					
	MALDI			SIMS		
Norm	PCA/ICA	NN-P.	pLSA	PCA/ICA	NN-P.	pLSA
L_1	$3.5 \cdot 10^1$	$2.4 \cdot 10^1$	$\mathbf{2.3 \cdot 10^1}$	$9.4 \cdot 10^0$	$\mathbf{3.7 \cdot 10^0}$	$3.9 \cdot 10^0$
L_2	$1.0 \cdot 10^{-1}$	$\mathbf{8.0 \cdot 10^{-2}}$	$9.5 \cdot 10^{-2}$	$1.4 \cdot 10^{-2}$	$\mathbf{5.5 \cdot 10^{-3}}$	$6.6 \cdot 10^{-3}$
KL	$2.2 \cdot 10^0$	$1.7 \cdot 10^{-1}$	$\mathbf{1.3 \cdot 10^{-1}}$	$1.2 \cdot 10^0$	$2.9 \cdot 10^{-2}$	$\mathbf{2.5 \cdot 10^{-2}}$

Table 1: Reconstruction error for the two real-world datasets. The NN-PARAFAC and pLSA reconstructions outperform the other methods by a large margin. PCA/ICA does only well with respect to the L_2-norm. pLSA does best for KL-divergence, but it also performs well with respect to the other norms. Furthermore, NN-PARAFAC and pLSA grant increased interpretability.

Interpretations and method's properties

Constraints and their effects. Possible explanations for the inferior performance of PCA and ICA in our experiments lie in the constraints used by these methods. ICA relies on the assumption that the reconstructed sources have minimal mutual information, i. e. their statistical independence is maximal. The effect of this assumption is data-dependent and may be beneficial in some scenarios and harmful in others. In situations where we want to unmix data containing tissue types that differ only slightly and thus feature similar spectral signatures, ICA may not be the right choice as a decomposition method. It is unlikely to yield two similar characteristic spectra since these would feature high mutual information. However, if this assumption holds we can hope for a good performance of ICA.

PCA is handicapped in the detection of differences in tissue composition that manifest themselves merely in small spectral changes, because such subtleties contribute little to the overall variance of the data and are hence relegated to higher principal components which are easily overlooked in the routine visual inspection of the first few components. In our experiments, NN-PARAFAC and pLSA were better able to detect minor differences in unmixing both, the pure and impure mixtures (cf. figure 3). In contrast to the constraints used by PCA and ICA, the non-negativity constraint in NN-PARAFAC and pLSA is well motivated by physical properties and valid for all count data sets.

Number of components and reconstruction accuracy. In contrast to PCA, NN-PARAFAC and pLSA require the number of components k to be specified in advance. This means that all observed spectral intensity and variability is distributed among the k component spectra. This behavior is desirable if k equals the correct number of components. If k underestimates the correct number of components, such an approach does not fully exploit information on further tissue types (see for example figure 5 where the four component decomposition of the MALDI set does not reveal the necrotic part). For PCA, the number of components is not specified in advance and the full decomposition is always computed, even in cases where prior knowledge is available. Especially in such situations, PCA results in way too many components since the number of tissue types that one is interested in is normally much smaller than the number of mass channels, i. e. the number of principal components estimated. In typical situations, one can only examine the first few of possibly hundreds of principal components. This leads to a loss of information and

	Complementarity									
	MALDI (4 comp.)					SIMS (5 comp.)				
Quant.	PCA	ICA	NN-P.	pLSA	th.max.	PCA	ICA	NN-P.	pLSA	th.max.
95	0.20	0.19	0.19	**0.20**	0.20	**0.25**	0.24	**0.25**	**0.25**	0.25
90	0.37	0.35	0.35	**0.40**	0.40	0.47	0.46	0.49	**0.50**	0.50
85	0.54	0.50	0.50	**0.59**	0.60	0.63	0.64	0.68	**0.73**	0.75
80	0.67	0.63	0.61	**0.76**	0.80	0.75	0.77	0.82	**0.86**	1.00
75	0.76	0.74	0.69	**0.89**	1.00	0.84	0.86	0.91	**0.95**	1.00
70	0.85	0.84	0.75	**0.98**	1.00	0.92	0.93	0.97	**0.99**	1.00
65	0.94	0.91	0.80	**1.00**	1.00	0.97	0.96	0.99	**1.00**	1.00
60	0.98	0.95	0.84	**1.00**	1.00	0.99	0.98	**1.00**	**1.00**	1.00
55	0.99	0.98	0.88	**1.00**	1.00	**1.00**	0.99	**1.00**	**1.00**	1.00
50	0.99	0.99	0.91	**1.00**	1.00	**1.00**	**1.00**	**1.00**	**1.00**	1.00

Table 2: Complementarity estimation for the two real-world datasets (MALDI/SIMS). The numbers reported correspond to the percentage of the region of interest that is covered after combining the four/five thresholded component abundance maps at various quantiles. The theoretical maximum which is only reached for perfectly complementary abundance maps is also given. In both scenarios, the pLSA solution is significantly more complementary than the PCA and ICA counterparts, and more complementary than the NN-PARAFAC solution.

reconstruction accuracy. We have demonstrated that in scenarios where the number of components is limited, NN-PARAFAC and pLSA are superior to PCA and ICA.

Complementarity. In our experiments on both simulated and real-world data, pLSA clearly outperformed PCA, ICA and NN-PARAFAC with respect to the complementarity of the estimated components (cf. table 2). It is not surprising that PCA results in low complementary estimates since it always performs the full decomposition. For pLSA, very often the theoretical maximum of the complementarity measure is reached or almost reached. The numbers presented are backed by visual inspection that also suggests that the NN-PARAFAC and pLSA partitionings are more sparse than the PCA and ICA solutions, even though sparsity is not explicitly enforced. Sparsity simplifies interpretation and should be recovered if present in the data, that is if most of the tissue predominantly belongs to one (but not necessarily the same) class. The results obtained on the simulated datasets show that NN-PARAFAC and pLSA not only perform well in the pure mixture case, but also for heterogeneous tissue (cf. figure 3). The different mixture areas were better reconstructed than with PCA or ICA. Furthermore, the NN-PARAFAC and pLSA decomposition maps corresponding to the real-world data were convincingly correlated with the structures that are visible in the label maps (cf. figures 2, 5 and 9).

Computation time. The computation time needed for NN-PARAFAC and pLSA is higher than for PCA and ICA. However, in relation to the time required for data acquisition, it seems safe to say that all methods are sufficiently fast to be applied in practice.

Conclusions

We have shown on simulated and real-world data that pLSA is a suitable approach for unsupervised analysis of imaging mass spectrometry data. pLSA and NN-PARAFAC outperform PCA and ICA in terms of quality of the decomposition maps as they use an additive model which correctly mirrors the physical properties of the data. In addition, they offer superior physical interpretability as they produce normalized and non-negative components which can directly be interpreted as peak intensity lists. They also lead to more complementary components and retain high reconstruction accuracy. In contrast to non-negative PARAFAC, pLSA is based on a sound probabilistic model. We have further introduced the AICc-controlled pLSA, providing the methodology necessary to automatically estimate the number of tissue types, thus significantly decreasing the dependency on sample-specific prior knowledge. We further have described how a sparsity measure can be used to automatically identify those m/z channels that are relevant for the discrimination of tissue types. This can give further valuable insights in exploratory data analysis.

MATLAB code for the AICc-enhanced pLSA is freely available on request from the authors.

Acknowledgments

We gratefully acknowledge financial support by the DFG under grant no. HA4364/2-1 (M.H, B.Y.R., F.A.H), the Hans L. Merkle foundation (M.K.), the Karl-Steinbuch-Fellowship (B.Y.R.), and the Robert Bosch GmbH (F.A.H.). The financial support of the Netherlands BSIK program Virtual Laboratory for e-science is gratefully acknowledged by R.M.A.H. and E.R.A.

We furthermore like to thank Ivo Klinkert for providing data conversion and visualization tools, Ullrich Koethe for valuable scientific discussions and Tiffany R. Greenwood for cutting the tissue sections.

References

1. Caprioli, R. M.; Farmer, T. B.; Gile, J. *Anal. Chem.* **1997**, *69*, 4751–4760.

2. McDonnell, L. A.; Heeren, R. M. A. *Mass Spectrometry Reviews* **2006**, *26*, 606–643.

3. Chaurand, P.; Schwartz, S. A.; Caprioli, R. M. *Current Opinion in Chemical Biology* **2002**, *6*, 676–681.

4. Simpkins, F.; Czechowicz, J. A.; Liotta, L.; Kohn, E. C. *Pharmacogenomics* **2005**, *6*, 647–653.

5. Yanagisawa, K.; Shyr, Y.; Xu, B.; Massion, P.; Larsen, P.; White, B.; Roberts, J.; Edgerton, M.; Gonzalez, A.; Nadaf, S. *The Lancet* **2003**, *362*, 433–439.

6. Rohner, T. C.; Staab, D.; Stoeckli, M. *Mechanisms of Ageing and Development* **2005**, *126*, 177–185.

7. Belu, A. M.; Davies, M. C.; Newton, J. M.; Patel, N. *Anal. Chem.* **2000**, *72*, 5625–5638.

8. Cornett, D. S.; Frappier, S. L.; Caprioli, R. M. *Anal. Chem.* **2008**, *80*, 5648–5653.

9. Wu, L.; Lu, X.; Kulp, K. S.; Knize, M. G.; Berman, E. S. F.; Nelson, E. J.; Felton, J. S.; Wu, K. J. J. *Intern. Journal of Mass Spectrom.* **2007**, *2-3*, 137–145.

10. Trim, P. J.; Atkinson, S. J.; Princivalle, A. P.; Marshall, P. S.; West, A.; Clench, M. R. *Rapid Comm. in Mass Spectrom.* **2008**, *22*, 1503–1509.

11. van de Plas, R.; Ojeda, F.; Dewil, M.; van den Bosch, L.; de Moor, B.; Waelkens, E. *Proc. of the Pacific Symposium of Biocomputing* **2007**, *12*, 458–469.

12. Ivosev, G.; Burton, L.; Bonner, R. *Anal. Chem.* **2008**, *80*, 4933–4944.

13. Mantini, D.; Petrucci, F.; del Boccio, P.; Pieragostino, D.; di Nicola, M.; Lugaresi, A.; Federici, G.; Sacchetta, P.; di Ilio, C.; Urbani, A. *Bioinformatics* **2008**, *24*, 63–70.

14. Broersen, A.; van Liere, R.; Heeren, R. M. A. *Proc. of the 5th IASTED Intern. Conf. on Visualization, Imaging, and Image Processing* **2005**, 540–545.

15. Smentkowski, V. S.; Ostrowski, S. G.; Kollmer, F.; Schnieders, A.; Keenan, M. R.; Ohlhausen, J. A.; Kotula, P. G. *Surface and Interface Analysis* **2008**, *40*, 1176–1182.

16. Akaike, H. *IEEE Trans. on Automatic Control* **1974**, *19*, 716–723.

17. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning*; Springer, 2001.

18. Bro, R. *Multi-way Analysis in the Food Industry - Models, Algorithms, and Applications*, 1998, PhD-thesis, University of Amsterdam (NL).

19. Hyvärinen, A.; Oja, E. *Neural Networks* **2000**, *13*, 411–430.

20. Gävert, H.; Hurri, J.; Särelä, J.; Hyvärinen, A. http://www.cis.hut.fi/projects/ica/fastica/, 2007.

21. Harshman, R. A. *UCLA Working Papers in Phonetics* **1970**, *16*, 1–84.

22. Carroll, J. D.; Chang, J. J. *Psychometrika* **1970**, *35*, 283–819.

23. Kiers, H. A. *Psychometrika* **1991**, *56*, 97–212.

24. Cichocki, A.; Zdune, R. *Lecture Notes in Computer Science* **2007**, *2007*, 793–802.

25. Bro, R.; Andersson, C. A. http://www.models.kvl.dk/source/nwaytoolbox/index.asp, 2007.

26. Hofmann, T. *Proc. of the 15th Conf. on Uncertainty in Artificial Intelligence* **1999**.

27. Gaussier, E.; Goutte, C. *Proc. of the 28th annual int. ACM SIGIR conf. on Research and Development in Information Retrieval* **2005**, 601–602.

28. Benvenuto, F.; La Camera, A.; Theys, C.; Ferrari, A.; Lantri, H.; Bertero, M. *Inverse Problems* **2008**, *24*, 1–20.

29. Stine, R. A. *Sociological Methods and Research* **2004**, *33*, 230–260.

30. Burnham, K. P.; Anderson, D. R. *Model Selection and Multimodel Inference: A Practical-Theoretic Approach*, 2nd ed.; Springer, 2002.

31. Renard, B. Y.; Kirchner, M.; Steen, H.; Steen, J. A.; Hamprecht, F. A. *BMC Bioinformatics* **2008**, *9*, 355–385.

32. Hoyer, P. *Journal of Machine Learning Research* **2004**, *5*, 1457–1469.

33. submitted (by one of the co-authors), anonymized.

Supporting Information

Concise Representation of Mass Spectrometry Images by Probabilistic Latent Semantic Analysis

Michael Hanselmann[1], *Marc Kirchner*[1,‡], *Bernhard Y. Renard*[1,‡],
Erika R. Amstalden[2], *Kristine Glunde*[3], *Ron M. A. Heeren*[2], *Fred A. Hamprecht*[1,⋆]

[1] Heidelberg Collaboratory for Image Processing (HCI), Interdisciplinary Center for Scientific Computing (IWR), University of Heidelberg, Speyerer Strasse 4, Heidelberg, Germany, [2] FOM-AMOLF, FOM-Institute for Atomic and Molecular Physics, Kruislaan 407, Amsterdam, The Netherlands, [3] Russell H. Morgan Department of Radiology and Radiological Science, Johns Hopkins University School of Medicine, Baltimore, Maryland 21205, USA

[‡] contributed equally

[⋆] corresponding author. Email: fred.hamprecht@iwr.uni-heidelberg.de.

A: Total ion count images and histologically stained parallel slices

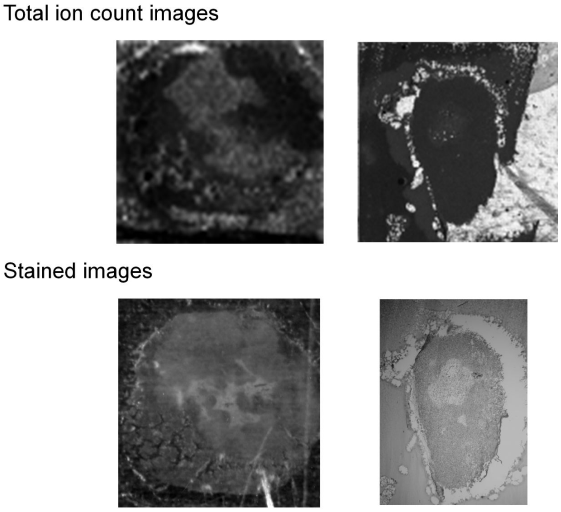

Figure 10: Total ion count images and histologically stained parallel slices for the MALDI (left) and SIMS (right) dataset.

B: Spectral components for the SIMS set

Figure 11 shows the estimated spectral components for the SIMS dataset and a five component decomposition. Please refer to the Experiments section of the paper for the corresponding abundance maps as well as more details on the dataset.

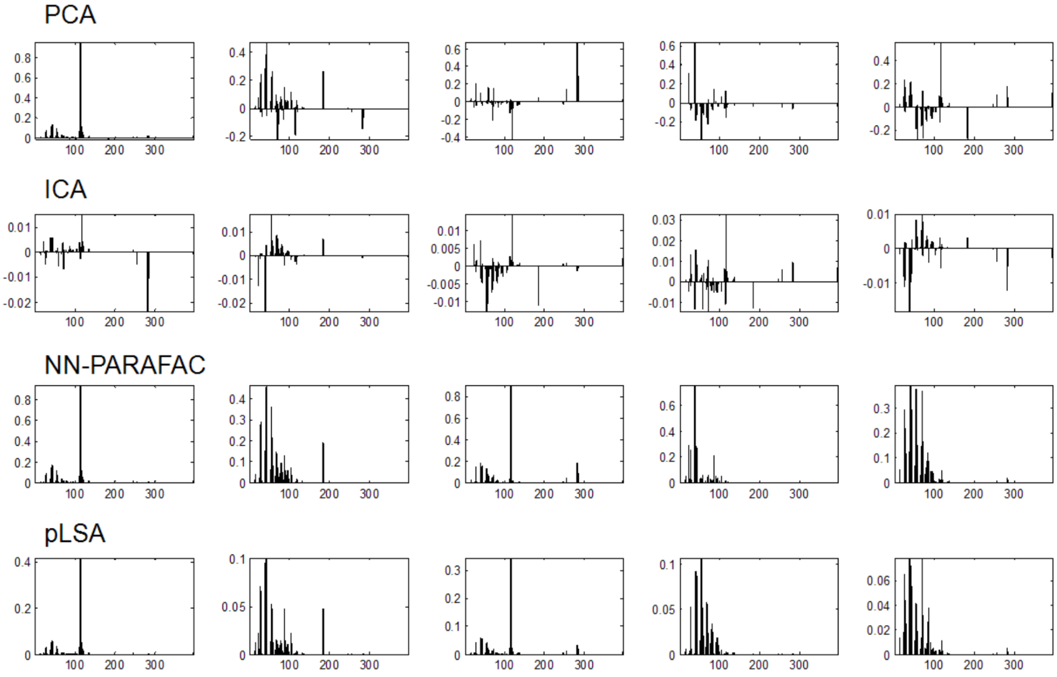

Figure 11: Decomposition of the SIMS set with five components. The components have been arranged according to the ordering of the abundance maps in the paper. The most prominent peak in component one corresponds to indium (115 Da). Indeed, the tissue is torn in the respective areas and the indium tin oxide-coated glass slide is exposed.

C: Reconstruction accuracy metrics

To estimate the reconstruction accuracy of PCA, ICA, NN-PARFAC and pLSA, we first created two vectors by concatenating all original (observed) spectra of a dataset and all reconstructed spectra. We then calculated the L_1- and L_2-norm of the difference with equation 13 and divided the result by the number of spectra in the dataset.

Equation 14 for KL divergence requires that the two concatenated spectra are probability distributions and are therefore all-positive and sum up to one. Non-negativity is not guaranteed in the case of PCA and ICA. Especially if only a few components are used in the reconstruction, it is possible that some channels of the reconstructed spectra exhibit negative values. We therefore set all channels in the reconstructed spectra that had negative intensities to zero (which was beneficial for the PCA/ICA error estimates). Irrespective of the method used for the decomposition, we further added a small constant to all values in the two concatenated vectors to avoid division by zero and finally normalized the two vectors before applying equation 14.

L_p-norm of a vector a:

$$\|a\|_p = (\sum_{i=1}^{n} |a_i|^p)^{1/p} \tag{13}$$

KL-divergence for two discrete probability distributions P and Q:

$$D_{KL}(P||Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)} \tag{14}$$

D: Decomposition of the MALDI set with eight components

Figures 12 and 13 as well as tables 3 and 4 illustrate the decomposition results for the MALDI set with eight component. The abundance maps estimated by NN-PARAFAC and pLSA are much clearer than their PCA/ICA counterparts. For many, a direct assignment to one of the expected tissue types like viable or necrotic part is possible. Please refer to the figure captions for further details.

	Reconstruction Error MALDI		
Norm	PCA/ICA	NN-PARAFAC	pLSA
L_1	$1.8 \cdot 10^1$	$1.5 \cdot 10^1$	$\mathbf{1.3 \cdot 10^1}$
L_2	$4.9 \cdot 10^{-2}$	$\mathbf{4.8 \cdot 10^{-2}}$	$5.4 \cdot 10^{-2}$
KL	$2.7 \cdot 10^{-1}$	$7.5 \cdot 10^{-2}$	$\mathbf{4.7 \cdot 10^{-2}}$

Table 3: Reconstruction error for the MALDI set and an eight component decomposition. The NN-PARAFAC and pLSA reconstructions perform best.

	Complementarity MALDI (8 comp.)				
Quant.	PCA	ICA	NN-PARAFAC	pLSA	th.max.
95	0.35	0.36	0.32	**0.39**	0.40
90	0.61	0.61	0.57	**0.72**	0.80
85	0.79	0.79	0.74	**0.92**	1.00
80	0.90	0.89	0.84	**0.99**	1.00
75	0.96	0.96	0.91	**1.00**	1.00
70	0.99	0.99	0.94	**1.00**	1.00
65	**1.00**	0.99	0.96	**1.00**	1.00
60	**1.00**	**1.00**	0.96	**1.00**	1.00
55	**1.00**	**1.00**	0.97	**1.00**	1.00
50	**1.00**	**1.00**	0.98	**1.00**	1.00

Table 4: Complementarity estimation for the MALDI set and an eight component decomposition. The numbers reported correspond to the percentage of the region of interest that is covered after combining the thresholded component abundance maps at various quantiles. The theoretical maximum which is only reached for perfectly complementary abundance maps is also given. Again, the pLSA solution is significantly more complementary than the PCA, ICA and NN-PARAFAC results.

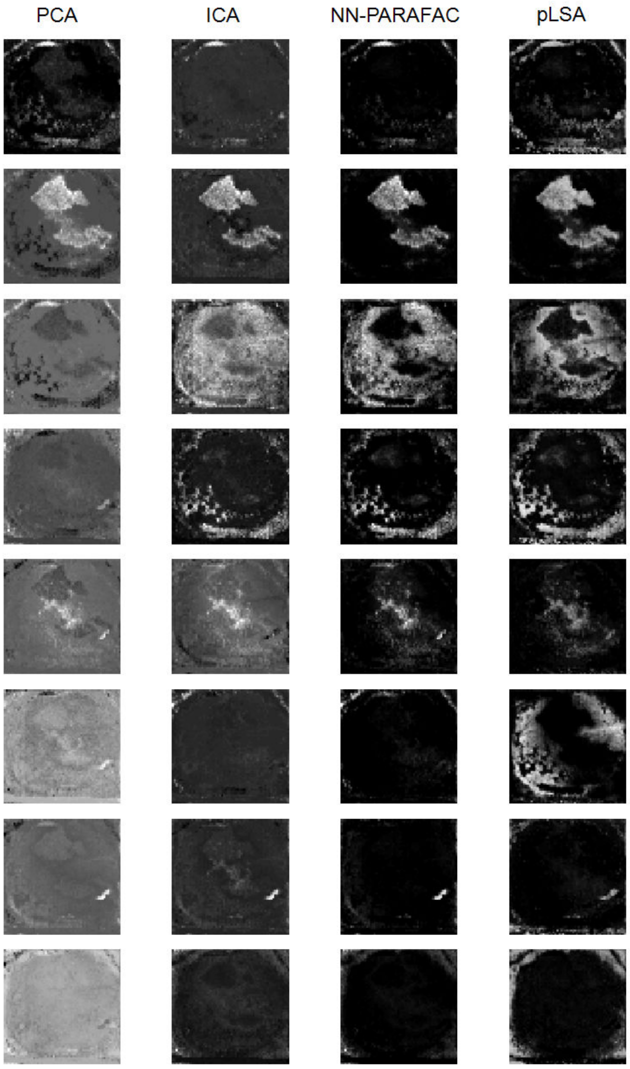

Figure 12: Decomposition of the MALDI set with eight components. Here we have omitted the colorbars for a clearly laid out visualization, but the scaling is chosen separately for each component such that the coloring is over the whole range of zero to one. In PCA and ICA, the sign of the components is arbitrary and has been changed to facilitate comparison with the non-negative methods. For all methods except PCA, the ordering of the components is arbitrary and has been permuted so as to facilitate comparison with the PCA decomposition. The components obtained by NN-PARAFAC and pLSA are much clearer than their PCA/ICA counterparts - especially with respect to component three (vascularized). Components two (viable) and five (necrotic) are also nicely expressed and have a clear localization. The third component is split up into two parts by pLSA yielding components three and six. The latter shows some resemblance with the stained image shown in figure 10. Furthermore, all components estimated by pLSA show spatial coherence which goes well with the assumption that tissue normally has a spatial extent. ICA performs better than PCA (see for example component four). The corresponding spectral components are shown in figure 13.

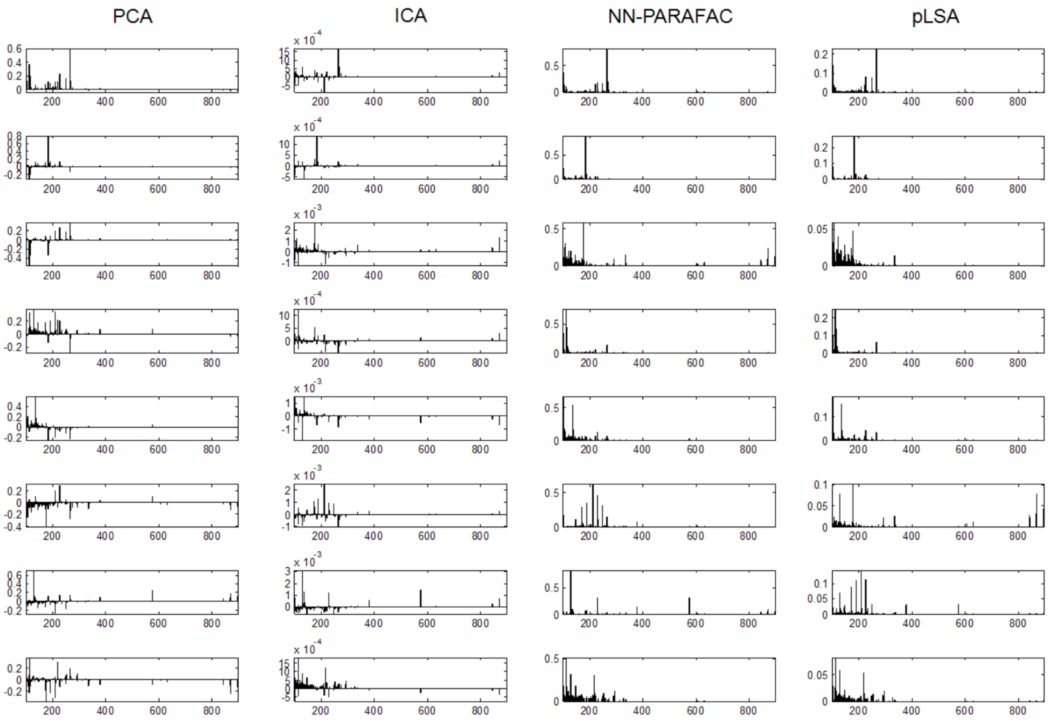

Figure 13: Decomposition of the MALDI set with eight components with all reconstructed spectral components which have been arranged according to figure 12.

E: Reconstruction accuracies and complementarity estimates for a varying number of components

The results obtained for the four and eight component decompositions of the MALDI set suggested that NN-PARAFAC and pLSA are superior to PCA and ICA in terms of reconstruction accuracy and complementarity. In the following experiment we quantified the reconstruction error as well as the complementarity for a varying number of components. Since we always select the best combination of upper/lower quantile abundance maps for PCA/ICA (see Experiments section of the paper) we only calculated decompositions where the combinatorial complexity was still manageable.

The more components we take into account, the better the reconstructions get and if we consider all PCA components, we end up with a perfect reconstruction. However, in real-world applications we are often interested in reducing the dimensionality of the data and examine only few components. In this scenario NN-PARAFAC and pLSA clearly outperform PCA and ICA (cf. table 5). Concerning the complementarity estimates we see that taking more components into account is not necessarily beneficial for PCA with respect to its relative performance (see table 6).

		Reconstruction Error MALDI		
Comp.	Norm	PCA/ICA	NN-PARAFAC	pLSA
2	L_1	$5.1 \cdot 10^1$	$3.9 \cdot 10^1$	$\mathbf{3.7 \cdot 10^1}$
	L_2	$1.5 \cdot 10^{-1}$	$\mathbf{1.4 \cdot 10^{-1}}$	$1.8 \cdot 10^{-1}$
	KL	$3.8 \cdot 10^1$	$5.0 \cdot 10^{-1}$	$\mathbf{2.8 \cdot 10^{-1}}$
3	L_1	$4.4 \cdot 10^1$	$3.3 \cdot 10^1$	$\mathbf{2.8 \cdot 10^1}$
	L_2	$1.2 \cdot 10^{-1}$	$\mathbf{1.0 \cdot 1.0^{-1}}$	$1.3 \cdot 10^{-1}$
	KL	$3.6 \cdot 10^0$	$3.8 \cdot 10^{-1}$	$\mathbf{1.7 \cdot 10^{-1}}$
4	L_1	$3.5 \cdot 10^1$	$\mathbf{2.4 \cdot 10^1}$	$2.5 \cdot 10^1$
	L_2	$1.0 \cdot 10^{-1}$	$\mathbf{8.0 \cdot 10^{-2}}$	$1.2 \cdot 10^{-1}$
	KL	$2.2 \cdot 10^0$	$1.7 \cdot 10^{-1}$	$\mathbf{1.4 \cdot 10^{-1}}$
5	L_1	$3.0 \cdot 10^1$	$2.2 \cdot 10^1$	$\mathbf{2.0 \cdot 10^1}$
	L_2	$8.6 \cdot 10^{-2}$	$\mathbf{7.0 \cdot 10^{-2}}$	$8.3 \cdot 10^{-2}$
	KL	$1.6 \cdot 10^0$	$1.5 \cdot 10^{-1}$	$\mathbf{9.1 \cdot 10^{-2}}$
6	L_1	$2.8 \cdot 10^1$	$1.8 \cdot 10^1$	$\mathbf{1.6 \cdot 10^1}$
	L_2	$6.6 \cdot 10^{-2}$	$\mathbf{5.8 \cdot 10^{-2}}$	$7.1 \cdot 10^{-2}$
	KL	$4.9 \cdot 10^{-1}$	$9.7 \cdot 10^{-2}$	$\mathbf{6.8 \cdot 10^{-2}}$
7	L_1	$2.1 \cdot 10^1$	$1.6 \cdot 10^1$	$\mathbf{1.5 \cdot 10^1}$
	L_2	$5.9 \cdot 10^{-2}$	$\mathbf{5.1 \cdot 10^{-2}}$	$6.2 \cdot 10^{-2}$
	KL	$3.8 \cdot 10^{-1}$	$8.0 \cdot 10^{-2}$	$\mathbf{5.7 \cdot 10^{-2}}$
8	L_1	$1.8 \cdot 10^1$	$1.5 \cdot 10^1$	$\mathbf{1.3 \cdot 10^1}$
	L_2	$4.9 \cdot 10^{-2}$	$\mathbf{4.5 \cdot 10^{-2}}$	$5.4 \cdot 10^{-2}$
	KL	$2.7 \cdot 10^{-1}$	$7.1 \cdot 10^{-2}$	$\mathbf{4.7 \cdot 10^{-2}}$
9	L_1	$1.5 \cdot 10^1$	$\mathbf{1.3 \cdot 10^1}$	$\mathbf{1.3 \cdot 10^1}$
	L_2	$4.1 \cdot 10^{-2}$	$\mathbf{4.0 \cdot 10^{-2}}$	$5.2 \cdot 10^{-2}$
	KL	$1.4 \cdot 10^{-1}$	$5.6 \cdot 10^{-2}$	$\mathbf{4.3 \cdot 10^{-2}}$
10	L_1	$1.5 \cdot 10^1$	$\mathbf{1.2 \cdot 10^1}$	$\mathbf{1.2 \cdot 10^1}$
	L_2	$3.8 \cdot 10^{-2}$	$\mathbf{3.6 \cdot 10^{-2}}$	$4.7 \cdot 10^{-2}$
	KL	$1.3 \cdot 10^{-1}$	$5.2 \cdot 10^{-2}$	$\mathbf{3.9 \cdot 10^{-2}}$

Table 5: Reconstruction error for the MALDI dataset: a varying number of components (first column) was used to confirm the results obtained for the four and eight component decompositions. Naturally, the more components we use, the better the reconstruction performance of PCA gets. However, in a real-world scenario we are often interested in reducing the dimensionality of the data and normally consider only a few components. For the MALDI-set, the AICc-type criterion suggests to use eight components for which the reconstructions of NN-PARAFAC and pLSA are more precise.

		Complementarity				
		MALDI				
Comp.	Quantile	PCA	ICA	NN-PARAFAC	pLSA	th.max.
2	95	**0.10**	**0.10**	**0.10**	**0.10**	0.10
	75	**0.50**	**0.50**	0.46	**0.50**	0.50
	55	0.77	0.77	0.69	**0.90**	0.90
3	95	**0.15**	0.14	0.14	**0.15**	0.15
	75	0.65	0.66	0.55	**0.75**	0.75
	55	0.97	0.96	0.75	**1.00**	1.00
4	95	**0.20**	0.19	0.19	**0.20**	0.20
	75	0.76	0.74	0.69	**0.89**	1.00
	55	0.99	0.98	0.88	**1.00**	1.00
5	95	0.24	0.23	0.22	**0.25**	0.25
	75	0.88	0.84	0.77	**0.98**	1.00
	55	**1.00**	0.99	0.94	**1.00**	1.00
6	95	0.28	0.28	0.26	**0.30**	0.30
	75	0.92	0.90	0.83	**1.00**	1.00
	55	**1.00**	**1.00**	0.95	**1.00**	1.00
7	95	0.31	0.32	0.29	**0.34**	0.35
	75	0.95	0.95	0.86	**1.00**	1.00
	55	**1.00**	**1.00**	0.96	**1.00**	1.00
8	95	0.35	0.36	0.32	**0.39**	0.40
	75	0.96	0.96	0.91	**1.00**	1.00
	55	**1.00**	1.00	0.97	**1.00**	1.00
9	95	0.37	0.39	0.36	**0.43**	0.45
	75	0.97	0.97	0.91	**1.00**	1.00
	55	**1.00**	1.00	0.97	**1.00**	1.00
10	95	0.39	0.42	0.36	**0.45**	0.50
	75	0.98	0.98	0.92	**1.00**	1.00
	55	**1.00**	**1.00**	0.97	**1.00**	1.00

Table 6: Complementarity estimation for the MALDI dataset: a varying number of components (first column) was used to check if the results obtained from the four and eight component decompositions also hold here. Again, pLSA outperforms the other methods.

F: Peak reconstruction

In order to verify if the four decomposition methods are capable of reconstructing the major peaks of the three ground truth spectra in the simulated datasets, we compared the set of major peaks in the reconstructed (spectral) components with the set of most intense peaks in the respective ground truth spectra. This was done by the following procedure: We first calculated various quantile spectra of the three ground truth spectra only containing the most intense peaks. For example, the 95% quantile spectrum only features the top five percent of the major peaks in the spectrum. For each decomposition method and each reconstructed component, we also calculated the respective quantile spectra. For each quantile, tissue type and method we then compared the resulting reduced spectrum to the ground truth spectrum corresponding to the same tissue type and quantile. The overlap of the peak positions contained in the two spectra is a measure of how well the major peaks were reconstructed. A value of 1.00 is only reached if the positions of the most intense peaks in the ground truth spectrum and the reconstructed spectrum are completely identical.

The sign of the principal and independent components is arbitrary. In the case of PCA and ICA, we therefore calculated the upper quantiles as well as the lower quantiles to allow for a fair comparison. The latter were calculated by first inverting the sign of the components and then extracting the major peaks as described above. We then used the quantile spectrum (upper/lower) that showed more overlap with the respective ground truth spectrum. Table 7 holds the results obtained for the two simulated datasets used in the paper. The numbers suggest that NN-PARAFAC and pLSA are better suited to extract the positions of the most relevant peaks than PCA or ICA.

		Peak reconstruction							
		Simulated set (impure mixtures)				Simulated set (pure mixtures)			
Quant.	Comp.	PCA	ICA	NN-P.	pLSA	PCA	ICA	NN-P.	pLSA
95	1	0.43	0.57	0.86	0.86	0.63	0.63	1.00	1.00
	2	0.71	0.71	0.71	0.71	0.63	0.63	0.88	0.88
	3	0.71	0.71	0.71	0.58	0.75	0.50	0.75	0.88
90	1	0.46	0.69	0.85	0.85	0.50	0.63	0.81	0.81
	2	0.77	0.85	0.77	0.77	0.38	0.50	0.94	0.94
	3	0.54	0.54	0.54	0.62	0.63	0.38	0.75	0.75
85	1	0.47	0.63	0.79	0.58	0.50	0.63	0.88	0.88
	2	0.79	0.79	0.84	0.84	0.42	0.46	0.92	0.92
	3	0.42	0.42	0.68	0.68	0.46	0.33	0.83	0.83
80	1	0.50	0.58	0.69	0.65	0.44	0.59	0.88	0.88
	2	0.69	0.73	0.88	0.88	0.38	0.44	0.88	0.88
	3	0.38	0.42	0.69	0.73	0.50	0.44	0.88	0.88
⊘ 80-95	1-3	0.57	0.64	0.73	**0.75**	0.50	0.51	**0.88**	0.86

Table 7: Simulated dataset: the table quantifies which fraction of the most intense peaks in the (known) characteristic spectra of the three tissue types is also among the major peaks in the corresponding spectral components estimated by the four methods. For each method and each estimated spectral component we calculated various quantile spectra containing only the most intense peaks. The resulting peak lists were then compared with the corresponding quantile spectrum of the respective characteristic spectrum. The overlap of the two peak lists is expressed as a number in $[0; 1]$. A value of 1.00 is only reached if the positions of the major peaks in the ground truth spectrum and the reconstructed spectrum are identical. In the case of PCA and ICA, the signum of the reconstructed spectral components is arbitrary. Therefore, besides the upper quantile, we also calculated the lower quantile spectrum by first inverting the signum of the component and then extracting the major peaks as described above. We then used the quantile spectrum (upper/lower) that resulted in a higher percentage of overlap. The last row gives the average values of overlap obtained with the four methods. For pure and impure mixtures, the percentage of overlap is highest for NN-PARAFAC and pLSA indicating that those methods are able to better reconstruct the major peaks of the characteristic spectra than PCA or ICA. Furthermore, ICA does better than PCA, especially in the impure setting.