Concise Representation Of MS Images By Probabilistic Latent Semantic Analysis

Michael Hanselmann[1], Bernhard Y. Renard[1,*], Marc Kirchner[1,*], Andriy Kharchenko[2], Leendert A. Klerk[2], Ullrich Koethe[1], Ron M. A. Heeren[2], Fred A. Hamprecht[1]

[1] Interdisciplinary Center for Scientific Computing (IWR), University of Heidelberg, Heidelberg, Germany
[2] FOM AMOLF, FOM-Institute for Atomic and Molecular Physics, Amsterdam, The Netherlands
* contributed equally

Introduction

Imaging Mass Spectrometry (IMS) has evolved into a promising technology allowing for a detailed analysis of the spatial distribution of biomolecules. However, the enormous size of data sets acquired with state-of-the-art instrumentation makes a direct manual analysis difficult, and automated (pre-)processing becomes indispensable.

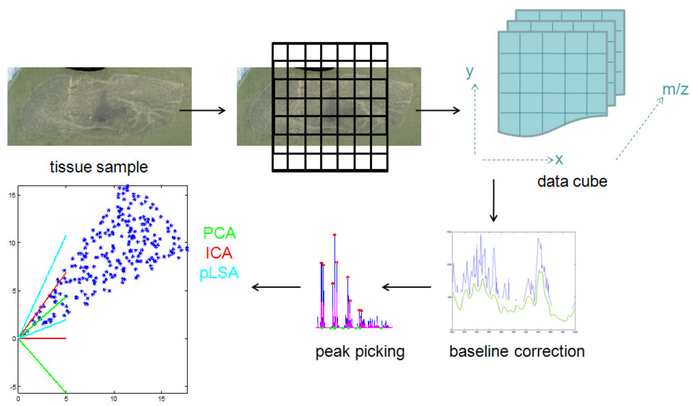
When little or no prior information on the composition of a sample is available, it is useful to decompose the spectral image into a small number of component spectra and abundance maps of these components. Conventional techniques such as Principal Component Analysis (PCA) or Independent Component Analysis (ICA) [3] have successfully been used in this setting; but they suffer from certain drawbacks:

- the component spectra found by PCA are mutually orthogonal and feature negative counts
- ICA components often feature negative counts
- PCA and ICA components are not physically motivated and cannot recover the true mass spectra of the tissue components.

Probabilistic Latent Semantic Analysis (pLSA) [1] directly results in normalized, non-negative components which can be interpreted as ion abundance rates.

Data Processing

- simple baseline correction by channel-wise subtraction of the minimum
- feature extraction by local maximum detection
- beneficial for quality and speed of further analysis



We have successfully applied pLSA in the exploratory analysis of mass spectral images of snap-frozen, cryo-sectioned rat brain samples acquired with a TRIFT II instrument that combines MALDI ionization with a stigmatic imaging TOF mass analyzer. The spatial resolution of the data is re-binned to 1μm and the spectral resolution is re-binned to 0.1Da.

Results

Spatial/spectral component distributions

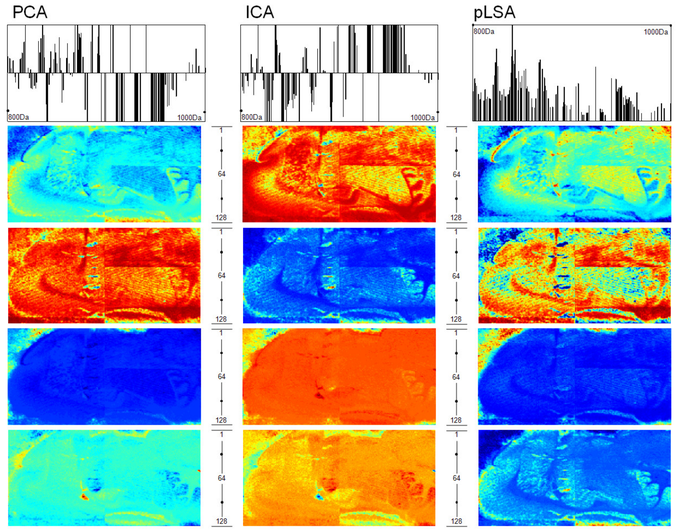


Figure 2: Abundance maps for PCA (left), ICA (middle) and pLSA (right), ordered by explained variance (PCA).

Sparsity

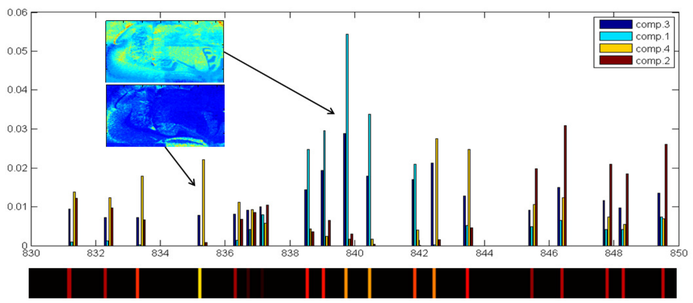


Figure 3: Bar plot of the four component spectra between 830 and 850 Da with the associated sparsity given below. Dark areas correspond to low sparsity, lighter areas to higher sparsity indicating decisive peak positions.

Conclusions

- pLSA components are normalized and non-negative
- pLSA provides superior physical interpretability to PCA and ICA
- pLSA is highly competitive to PCA and ICA in terms of the richness of morphological details revealed by the component abundance maps
- both ICA and pLSA exhibit structures more clearly than PCA

Methods

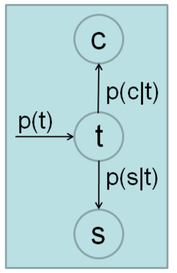
pLSA is equivalent to non-negative matrix factorization with a Kullback-Leibler divergence measure and can be described as a linear model with latent variable t

$$p(s,c) = \sum_{t \in T} p(t)p(s|t)p(c|t) \tag{1}$$

where s is a spectrum, c an m/z-channel and t the hidden variable topic. The decomposition problem is solved by an Expectation Maximization (EM) procedure:

E-step:

$$p(t|s,c) = \frac{p(t)p(s|t)p(c|t)}{\sum_{t' \in T} p(t')p(s|t')p(c|t')} \tag{2}$$

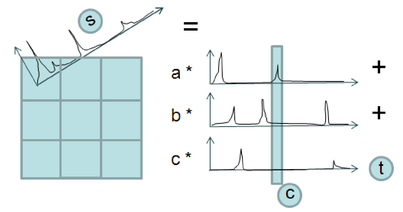


M-step:

$$p(c|t) \propto \sum_{s \in S} n(s,c)p(t|s,c) \tag{3}$$

$$p(s|t) \propto \sum_{c \in C} n(s,c)p(t|s,c) \tag{4}$$

$$p(t) \propto \sum_{s \in S} \sum_{c \in C} n(s,c)p(t|s,c) \tag{5}$$



In the proposed model, each single tissue type is characterized by a distinct distribution over m/z and each acquired spectrum is regarded as a specific mixture of these structures. The decisive peaks can be identified by calculating the sparsity measure [2]

$$sparsity(x) = \frac{\sqrt{|T|} - (\sum |x_i|)/\sqrt{\sum x_i^2}}{\sqrt{|T|} - 1} \tag{6}$$

where x' is a $1 \times |T|$ row vector of the matrix that holds $p(c|t)$.

Acknowledgements

We gratefully acknowledge financial support by the DFG under grant no. HA4364/2-1 (M.H, B.Y.R., F.A.H), the Karl-Steinbuch-Fellowship (B.Y.R.), the Hans L. Merkle foundation (M.K.), 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 A.K.

References

[1] T. Hofmann. Uncertainty in Artificial Intelligence (UIA), Stockholm, 1999.
[2] P. Hoyer. *Journal of Machine Learning Research*, 2004(5):1457–1469, 2004.
[3] A. Hyvärinen and E. Oja. *Neural Networks*, 13(4-5):411–430, 2000.