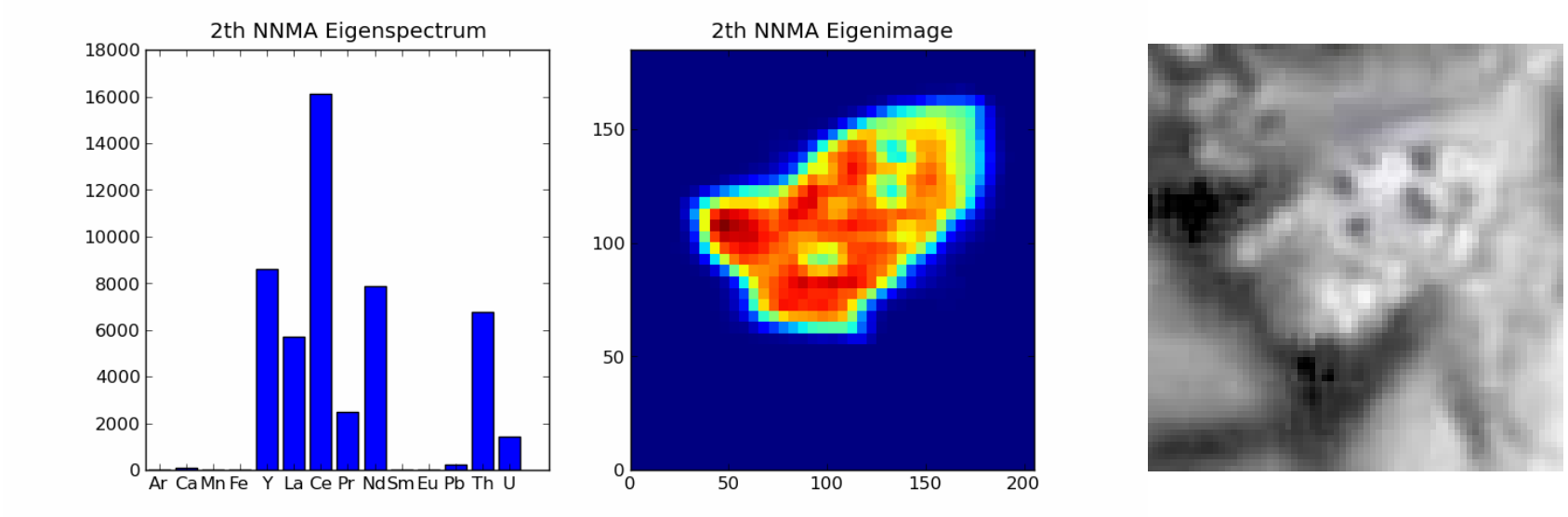
NNMA in fluorescence analysis

Why Non-Negativity is so positive for data analysis,
and what we should consider to get there.



Gerd Wellenreuther
HDF5 workshop
Grenoble, 11.-13. January 2010

HELMHOLTZ GEMEINSCHAFT

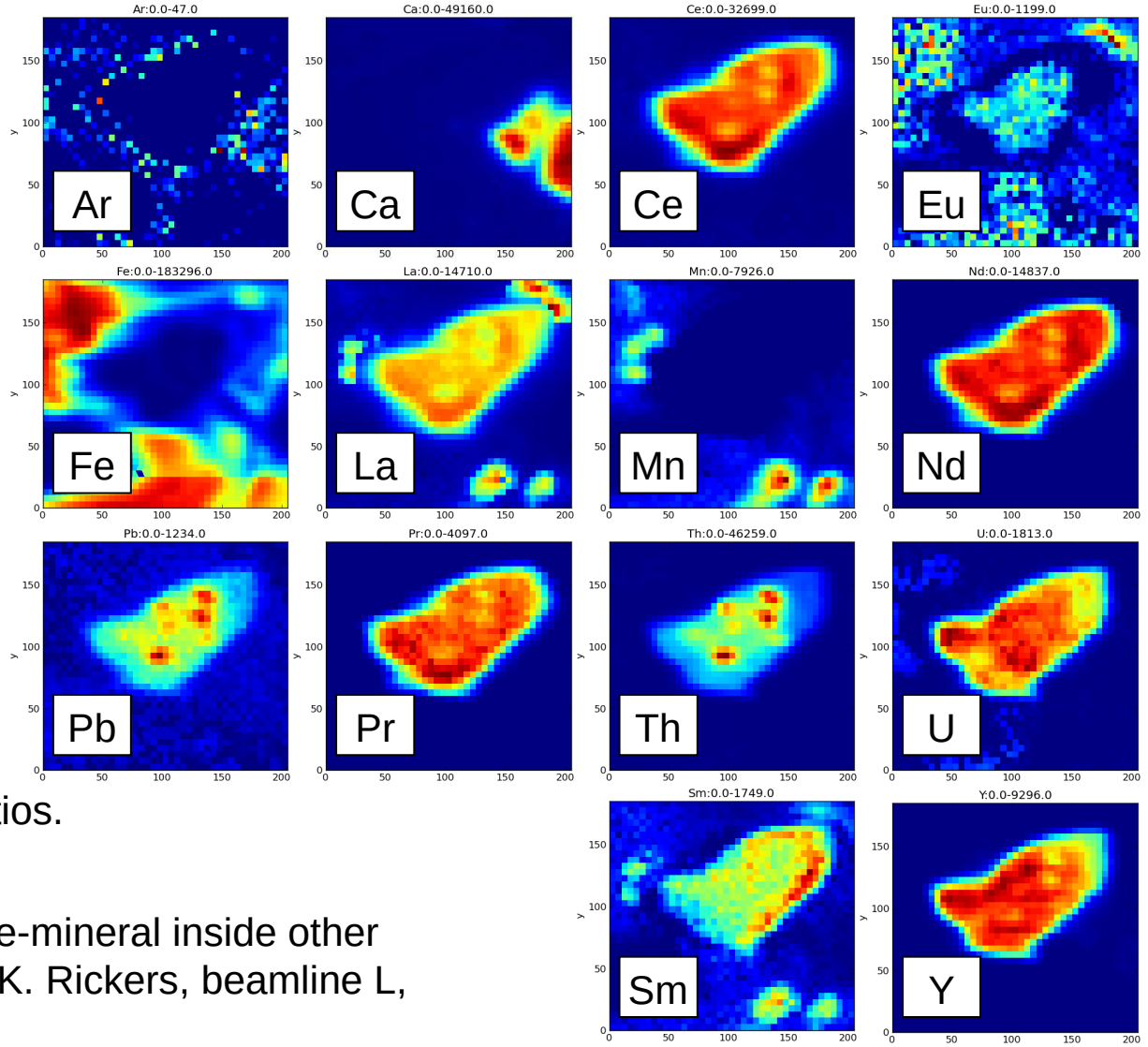
DESY

Outline

- My personal starting point: Elemental maps
- Motivation
- Factor analysis of a monazite-mineral in the presence of other crystalline phases
  - PCA: Introduction + results
  - Non-Negative Matrix Approximation (NNMA)
- Conclusion
- Outlook / Ideas

Picture of monazite-minerals
(stolen from Wikipedia)

Starting point: Elemental maps



Above: microscope image taken with beamline microscope

Scientific question:
Image age of mineral by examination of U : Th : Pb – ratios.

Right: Fitted data from monazite-mineral inside other crystalline phases (courtesy of K. Rickers, beamline L, HASYLAB).

Motivation

Deliver more than “just” elemental maps to the users:

- Samples consist of parts/clearly distinct contributions:
  - ➔ Spectroscopic data should nicely factorize (I think)
- Go beyond elemental distributions:
  - ➔ Chemical fingerprints of parts of the sample
- 2d-mapping projects the 3d-sample:
  - ➔ Previously well separated parts can overlap

➔ **Use statistical methods, e.g. factor analysis!**

Factor analysis / multivariate analysis

- Given: Spectroscopic data **D** (n_pixels,n_spectrum)
- Find factorisation into **maps M** (n_pixels,n_factors) and **spectra S** (n_factors,n_spectrum)

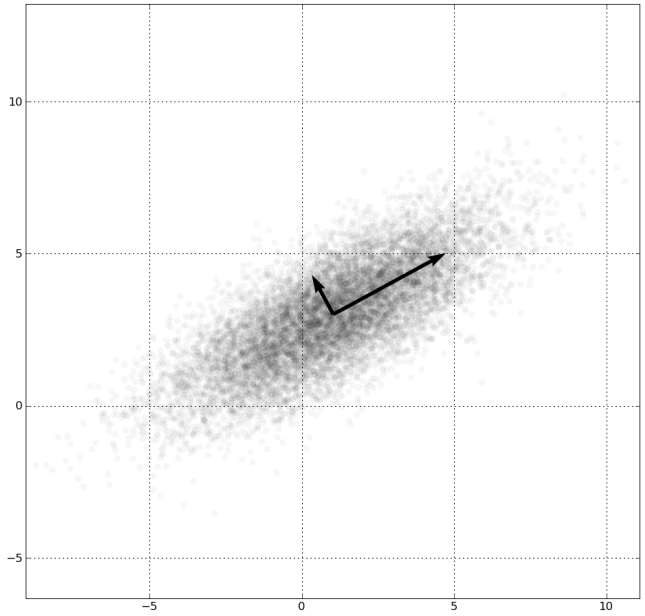
$$D \approx M * S$$

- Goals:
  - **Dimensional reduction** (n_factors << n_spectrum)
  - **Identify underlying factors**
    (good methods should be meaningful/interpretable)
- It does not …
  - **… know anything about chemistry (bad) or your sample (not biased ➔ good)**
  - **… take into account knowledge about adjacent pixels, detector physics etc.**

Why PCA is unfit for the photon counting community

PCA (Principal Component Analysis):

- Most used factor analysis, **quick + unique**
- Does a **principal axis transformation** of the **covariance** matrix
- Assumption: **Variance = Information**
  (not true in general, definitely broken e.g. in trace and ultra-trace analysis)



➔ Resulting spectra are **orthogonal**
(➔ in the presence of pure A a compound like AB will be demixed)

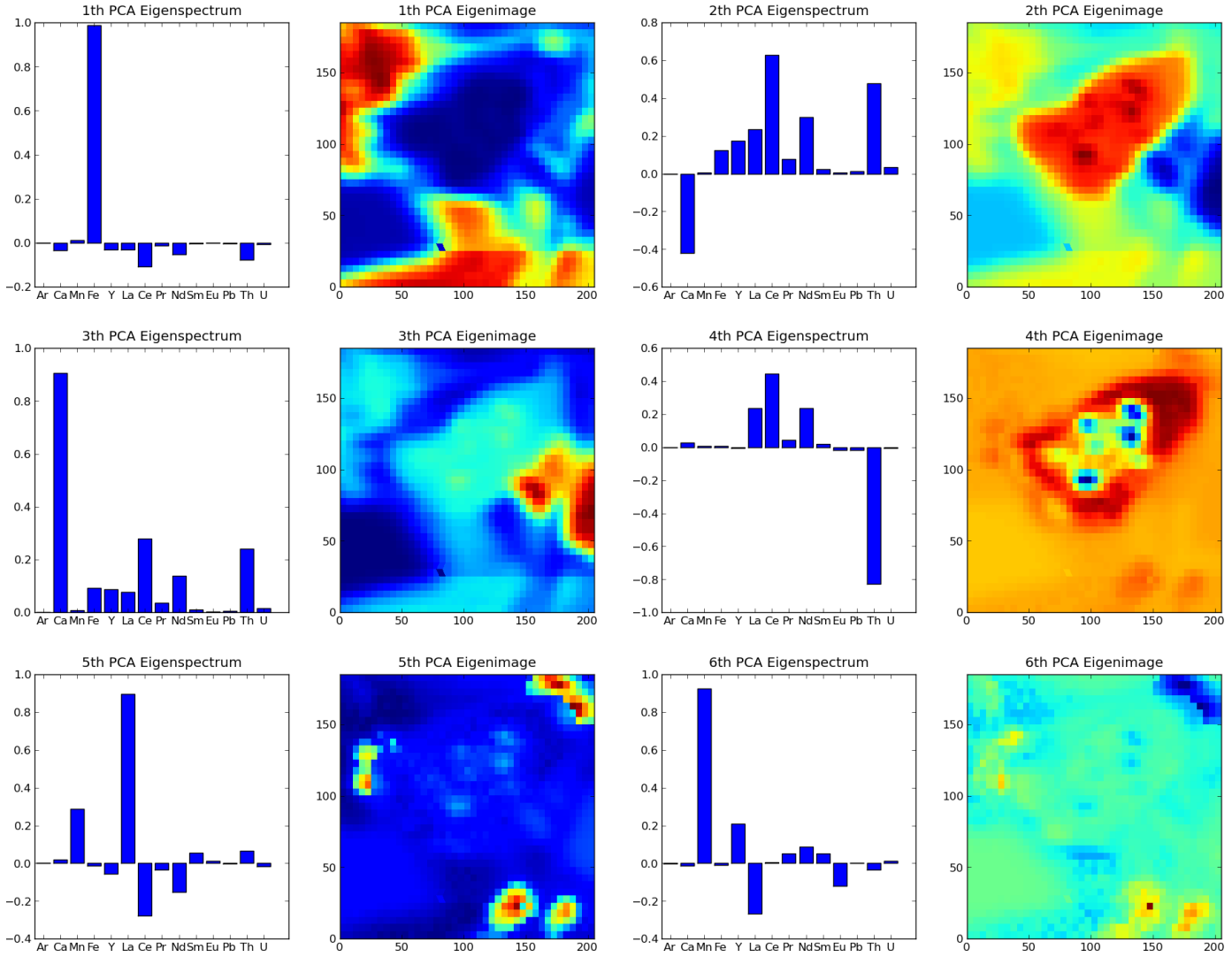
➔ Spectra and maps contain **negative values**
(➔ what is to be gained from negative concentrations or negative photons is unclear at best)

Example: Monazite PCAed

Eigenspectra can be **partly negative** ➔ anti-correlation

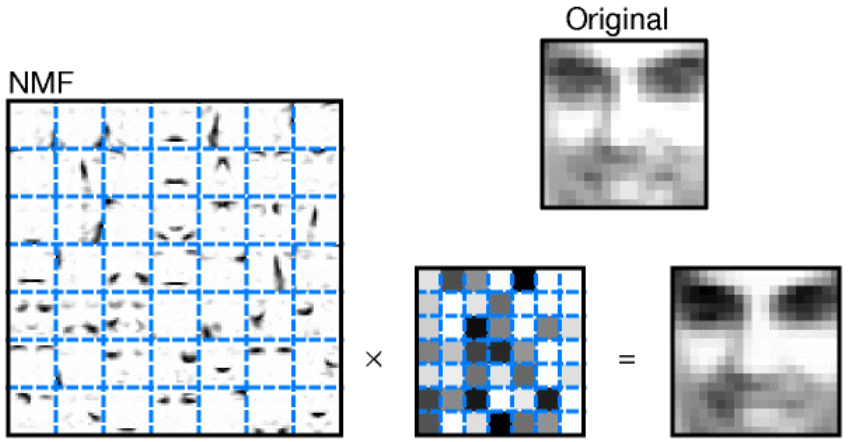
➔ PCA uses cancellation effects.

We need a factor analysis which is **not enforcing orthogonality** of spectra, but **non-negativity**!



Non-negative matrix approximation

- "Here we demonstrate an algorithm for **non-negative matrix factorization** that is able to learn **parts** of faces and semantic features of text.
- This is in contrast to other methods, such as **principal components analysis** and vector quantization, that **learn holistic**, not parts-based, representations.
- Non-negative matrix factorization is distinguished from the other methods by its use of **non-negativity constraints**.
- These constraints lead to a **parts-based representation because they allow only additive, not subtractive, combinations**."

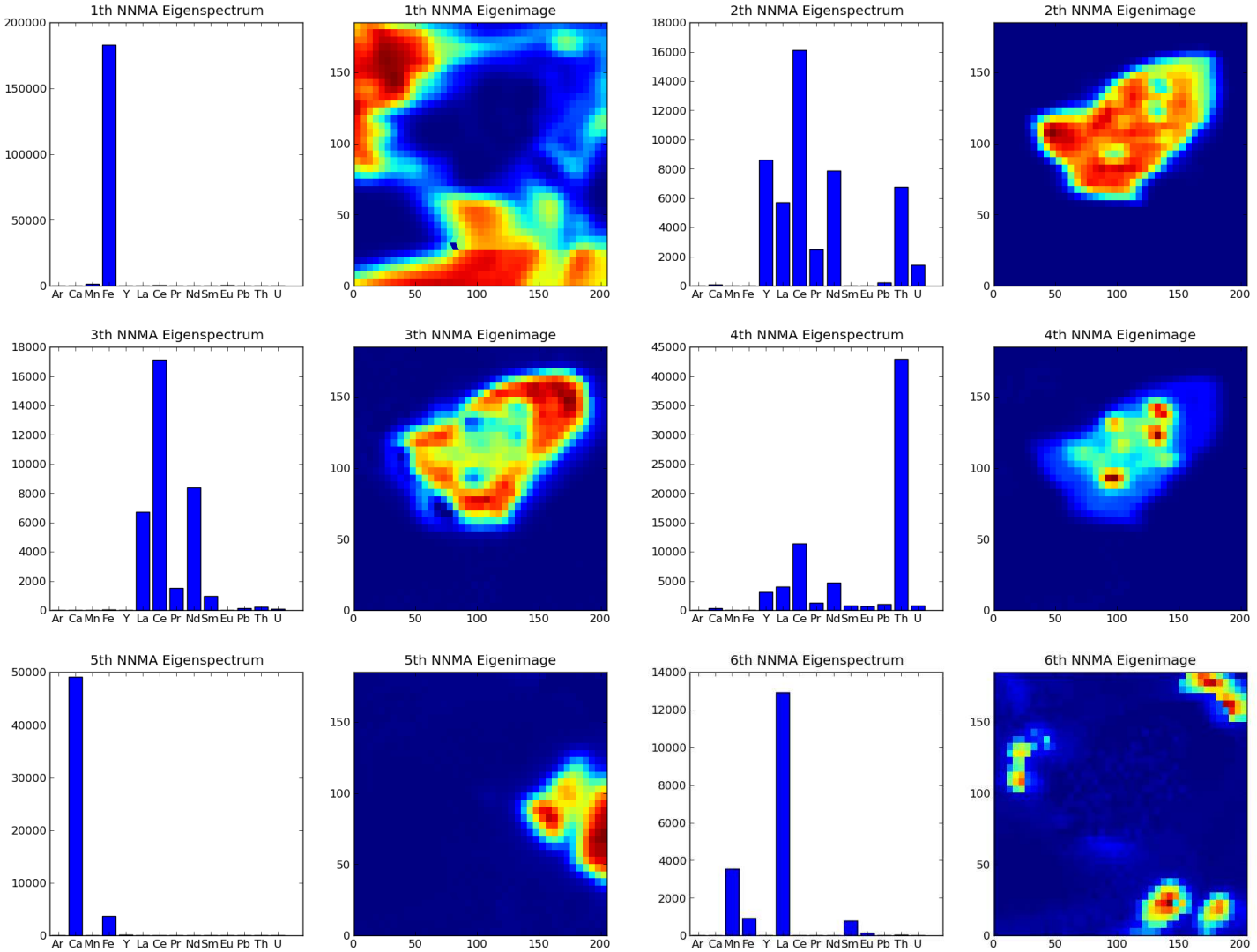


„**Learning the parts of objects by non-negative matrix factorization",**
by Lee & Seung, *Nature* 401, 788-791 (21 October 1999)

Example: Monazite NNMAed

Current interpretation:

1. Biotite
2. Monazit
3. Overgrowth at the monazit rim
4. Thorianite (ThO2)
5. Anorthite
6. Something exotic – most probably two phases



99.4% of data properly modelled, ~ each 0.6% under- and over-estimated

First conclusion

Advantages of NNMA

- Non-negativity restrain: Tries to **divide the data into similar parts**
- Does **not enforce orthogonality** of Eigenspectra
  (➔ no unnecessary demixing of entirely proper compounds)

Disadvantages

- Only **approximation** – needs significantly **longer!**
- You either need to **guess the number of Eigenspectra**,
  or you have to test several numbers.
- Solution is by definition **not** guaranteed to be **unique**
  ➔ check a couple of NNMA-runs from different starting position

➔ Can all be healed using more PC-power!

(estimated order for a real large dataset: ~ hours on a normal computer)

Things one should exploit

- If possible, **reduce the data first**! This could remove a lot of real noise.
  (going from full spectral data ➔ elementals maps or PCA-filtered data should yield at least a factor of 100-1000 in time)
- Number of factors: **Start with few factors**, add factors until satisfied.
- Starting position: **Use quick algorithms first!** Then refine.
  (maybe Non-negative ICA? Rather quick NNMA-algos like RRI?)
- **Initially reduce resolution** to approach minimum quicker.
  (can be done iteratively in any dimension)

- Especially trace elements: **Treat all elements on the same level ➔** apply weighting scheme before NNMA, reverse afterwards
  (e.g. Poisson weighting, work by Paul Kotula & Michael Keenan)

Acknowledgements

- **Karen Rickers**
  for the monazit-data
- **Tiziano Zito, Pietro Berkes** and **Niko Wilbert**
  for MDP (python module used to do PCA / ICA etc.)
- **Uwe Schmitt**
  for writing the python implementation
  of a whole bunch of different NNMA-algorithms
- **Armando Solé**
  for implementing Uwes NNMA-module
  along PCA in PyMca