Non-negative Matrix Factorization: Algorithms, Extensions and Applications

Emmanouil Benetos

www.soi.city.ac.uk/∼sbbj660/

March 2013

CITY UNIVERSITY LONDON

Outline

CITY UNIVERSITY LONDON

Introduction

- Assumption: Perception of the whole is based on perception on its parts
- There is evidence for parts-based representations in the brain
- Algorithms for learning holistic representations: principal component analysis, singular value decomposition
- Allowing additive (and not subtractive) combinations leads to parts-based representations
- First work in the '90s: positive matrix factorization

CITY UNIVERSITY LONDON

Non-negative Matrix Factorization (1)

- Non-negative Matrix Factorization (NMF): first proposed by Lee and Seung in 1999
- Unsupervised algorithm for decomposing multivariate data
- Alternatively, a method for dimensionality reduction by factorizing a data matrix into a low rank decomposition
- Constraint: non-negativity of data
- This allows a parts-based representation because only additive combinations are allowed

D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, pp. 788-791, Oct. 1999.

CITY UNIVERSITY LONDON

Non-negative Matrix Factorization (2)

Model:

- Given a non-negative matrix V find non-negative matrix factors W and H such that:

$$V \approx WH \tag{1}$$

  where:
  - $V \in \mathbb{R}^{n \times m}$
  - $W \in \mathbb{R}^{n \times r}$
  - $H \in \mathbb{R}^{r \times m}$
- The rank r of the factorization is chosen as $(n + m)r < nm$, so that data is compressed
- The columns of H are in one-to-one correspondence with the columns of V. Thus WH can be interpreted as weighted sum of each of the basis vectors in W, the weights being the corresponding columns of H

CITY UNIVERSITY LONDON

Non-negative Matrix Factorization (3)

Cost Functions:

- To find an approximate factorization, cost functions that quantify the quality of the approximation need to be defined
- Euclidean distance:

$$||A - B||^2 = \sum_{ij} (A_{ij} - B_{ij})^2 \tag{2}$$

- Kullback-Leibler divergence (aka relative entropy):

$$D(A||B) = \sum_{ij} \left(A_{ij} \log \frac{A_{ij}}{B_{ij}} - A_{ij} + B_{ij} \right) \tag{3}$$

D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in Proc. *NIPS*, pp. 556-562, 2001.

CITY UNIVERSITY LONDON

Non-negative Matrix Factorization (4)

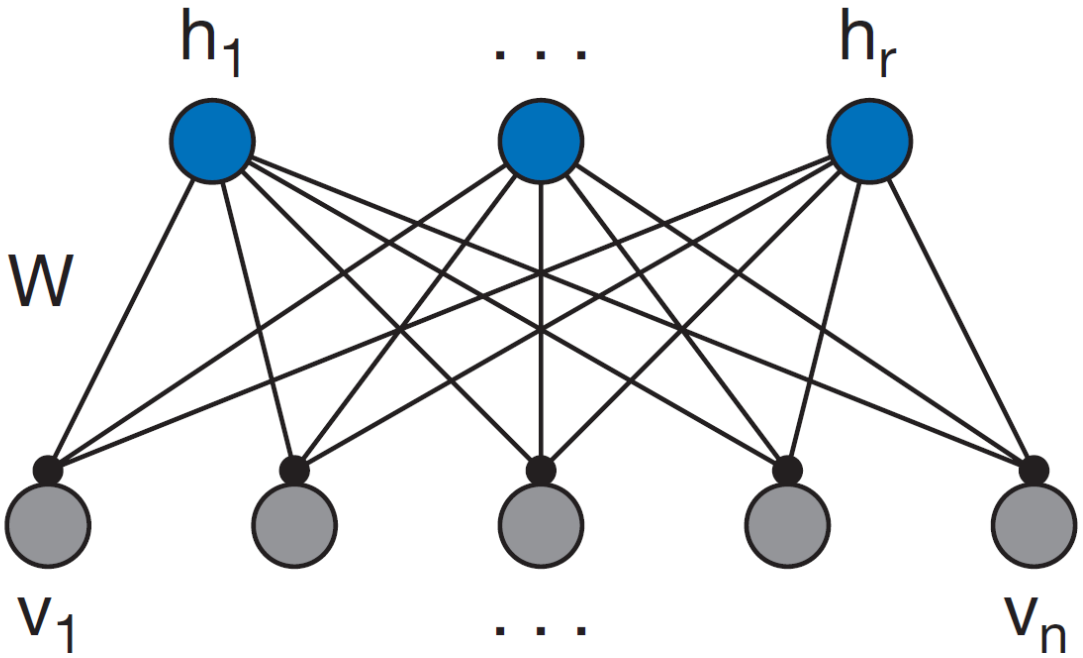
Algorithms:

- Problem 1:
  Minimize $||V - WH||^2$ wrt W and H, subject to $W, H \geq 0$
- Problem 2:
  Minimize $D(V||WH)$ wrt W and H, subject to $W, H \geq 0$
- W, H estimated using multiplicative update rules
- Alternative solutions: alternating least squares, gradient descent
- Other cost functions: Itakura-Saito distance, α-divergences, β-divergences, ϕ-divergences, Bregman divergences, ...
- Algorithms converge to a local minimum

CITY UNIVERSITY LONDON

Network representation

Network representation / generative model:



$$\langle v \rangle = Wh$$

Applications of NMF (1)

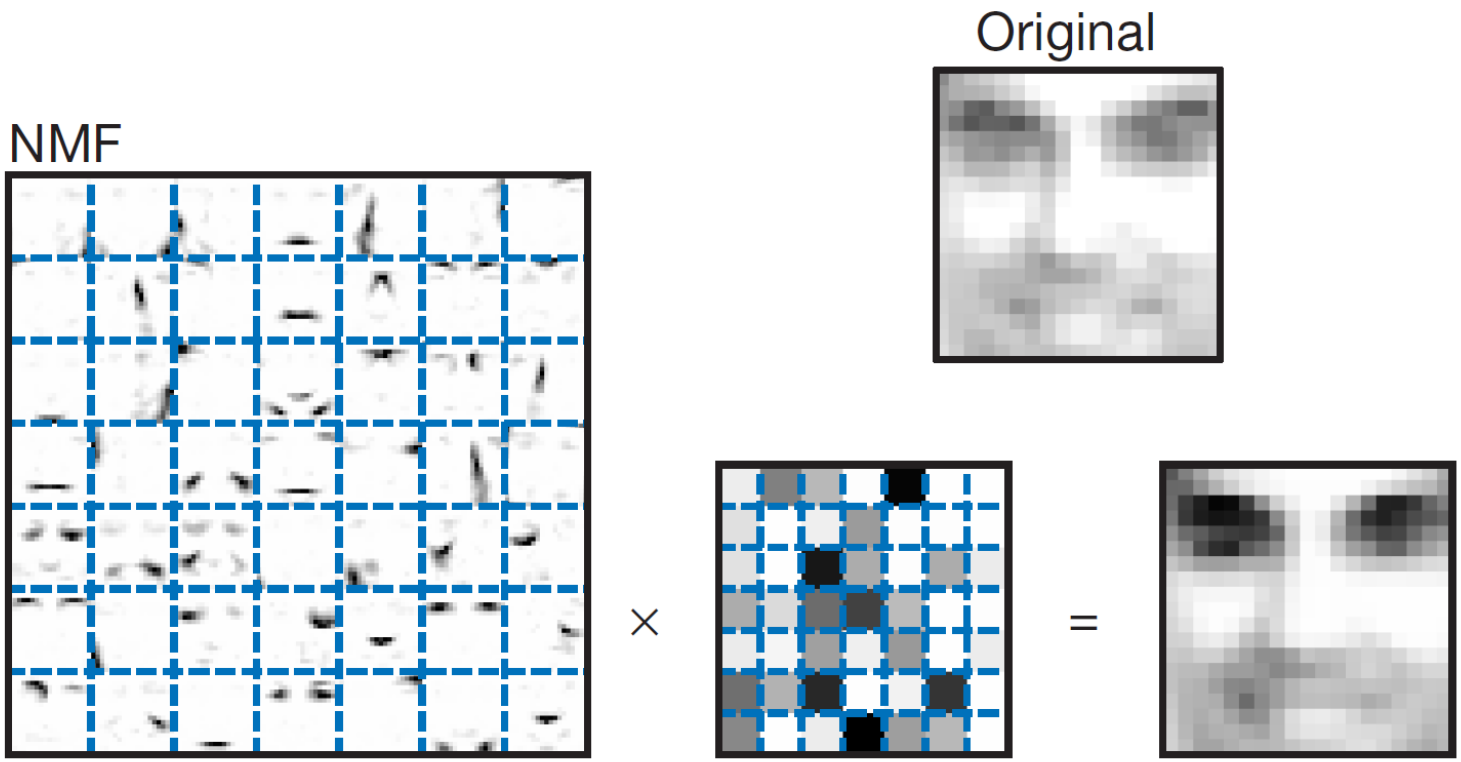
Applications:

- Detection, dimensionality reduction, clustering, classification, denoising, prediction
- Examples:
  - Digital image analysis
  - Text mining
  - Audio signal analysis
  - Bioinformatics

CITY UNIVERSITY LONDON

Applications of NMF (2)

Learning parts-based representation of faces:

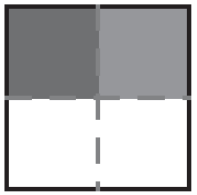


Applications of NMF (3)

Discovering semantic features in text:

court government council culture supreme constitutional rights justice	president served governor secretary senate congress presidential elected
flowers leaves plant perennial flower plants growing annual	disease behaviour glands contact symptoms skin pain infection

$\times$



$\approx$

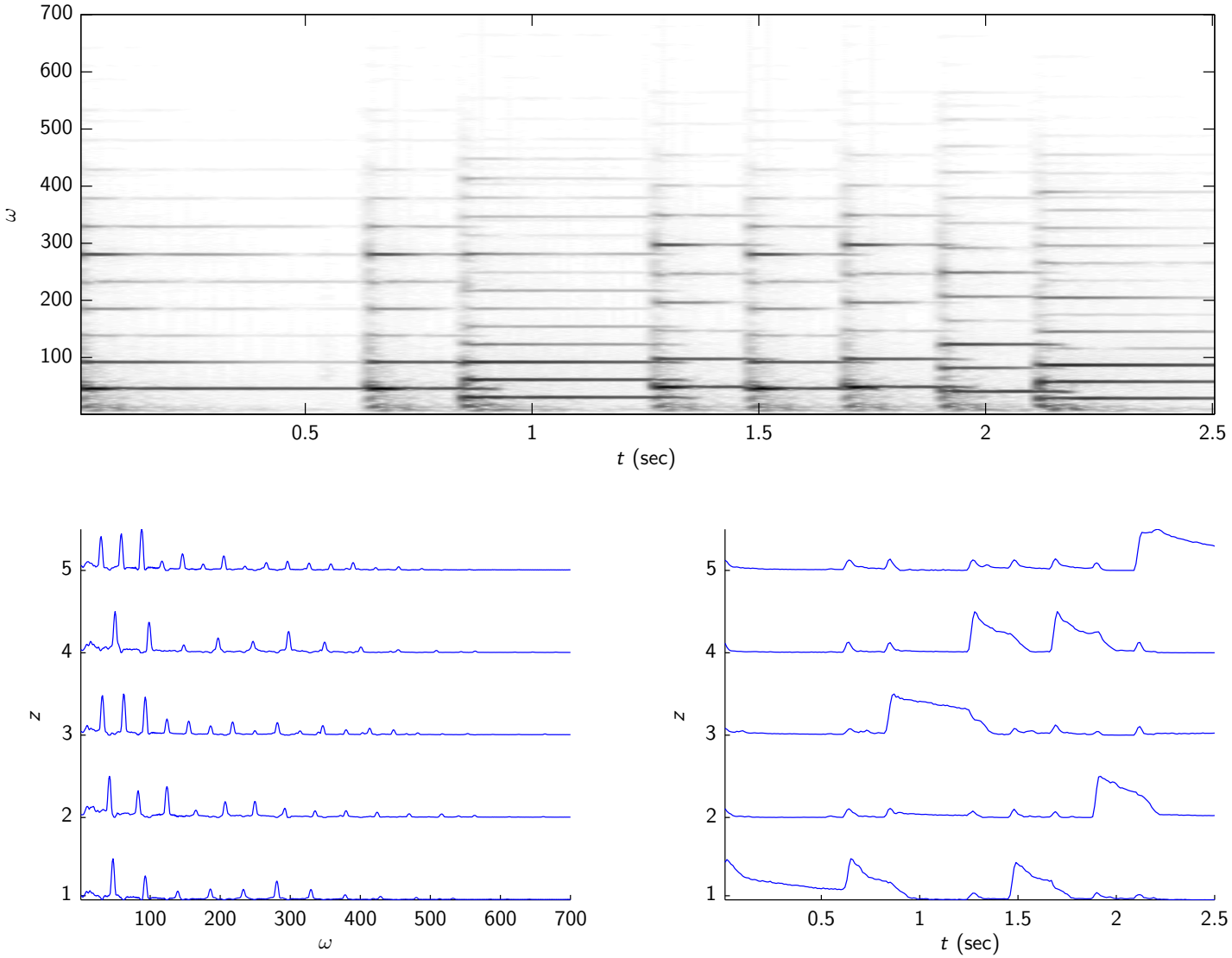
Encyclopedia entry:
'Constitution of the United States'

president (148) congress (124) power (120) united (104) constitution (81) amendment (71) government (57) law (49)

CITY UNIVERSITY LONDON

Applications of NMF (4)

Discovering musical notes in a recording:



Probabilistic Latent Semantic Analysis (1)

- In 1999, Hofmann proposed a technique for text processing and retrieval, called Probabilistic Latent Semantic Analysis (PLSA)
- ...which is also called Probabilistic Latent Semantic Indexing (PLSI)
- ...and also Probabilistic Latent Component Analysis (PLCA)!
- In fact, PLSA/PLSI/PLCA are the probabilistic counterparts of NMF using the KL divergence as a cost function
- This interpretation offers a framework that is easy to generalise and extend

T. Hofmann, "Learning the Similarity of Documents: an information-geometric approach to document retrieval and categorization," in Proc. *NIPS*, pp-914-920, 2000.

CITY UNIVERSITY LONDON

Probabilistic Latent Semantic Analysis (2)

Model:

- Considering the entries of V as having been generated by a probability distribution $P(x_1, x_2)$, PLSA models $P(x_1, x_2)$ as a mixture of conditionally independent multinomial distributions:

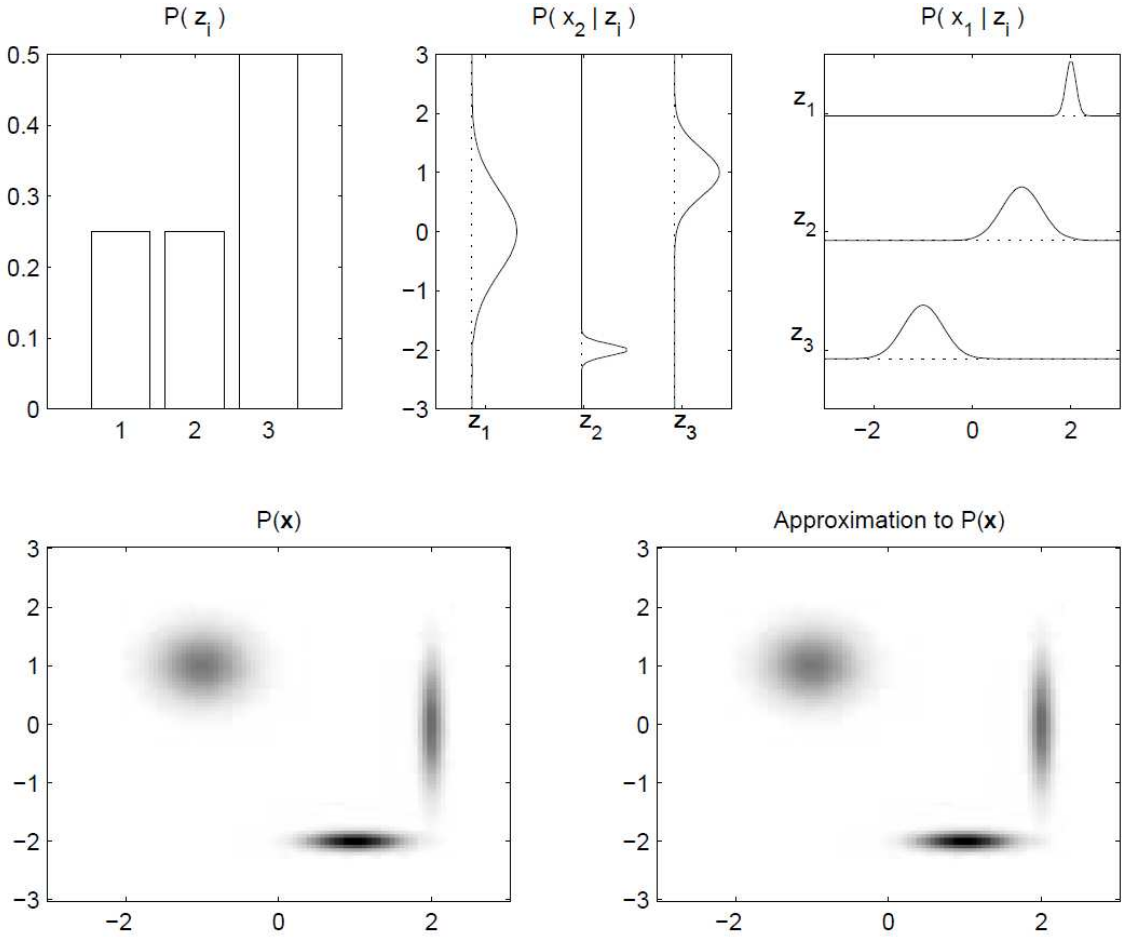
$$P(x_1, x_2) = \sum_z P(z)P(x_1|z)P(x_2|z) = P(x_1)\sum_z P(x_2|z)P(z|x_1) \tag{4}$$

- The 1st formulation is called symmetric, the 2nd asymmetric
- Parameters can be estimated using the Expectation-Maximization (EM) algorithm
- Essentially: H is $P(x_1|z)$ and W is $P(z)P(x_2|z)$

CITY UNIVERSITY LONDON

Probabilistic Latent Semantic Analysis (3)

Example of symmetric PLSA:



Convolutive Extensions (1)

- Instead of identifying 1-D components, we might want to detect 2-D structures out of non-negative data
- The linear model is not good enough - we need a convolutive model
- Non-negative Matrix Factor Deconvolution (NMFD): Extracting shifted structures from non-negative data
- Model:

$$V \approx \sum_t W_t \overrightarrow{H}^t \tag{5}$$

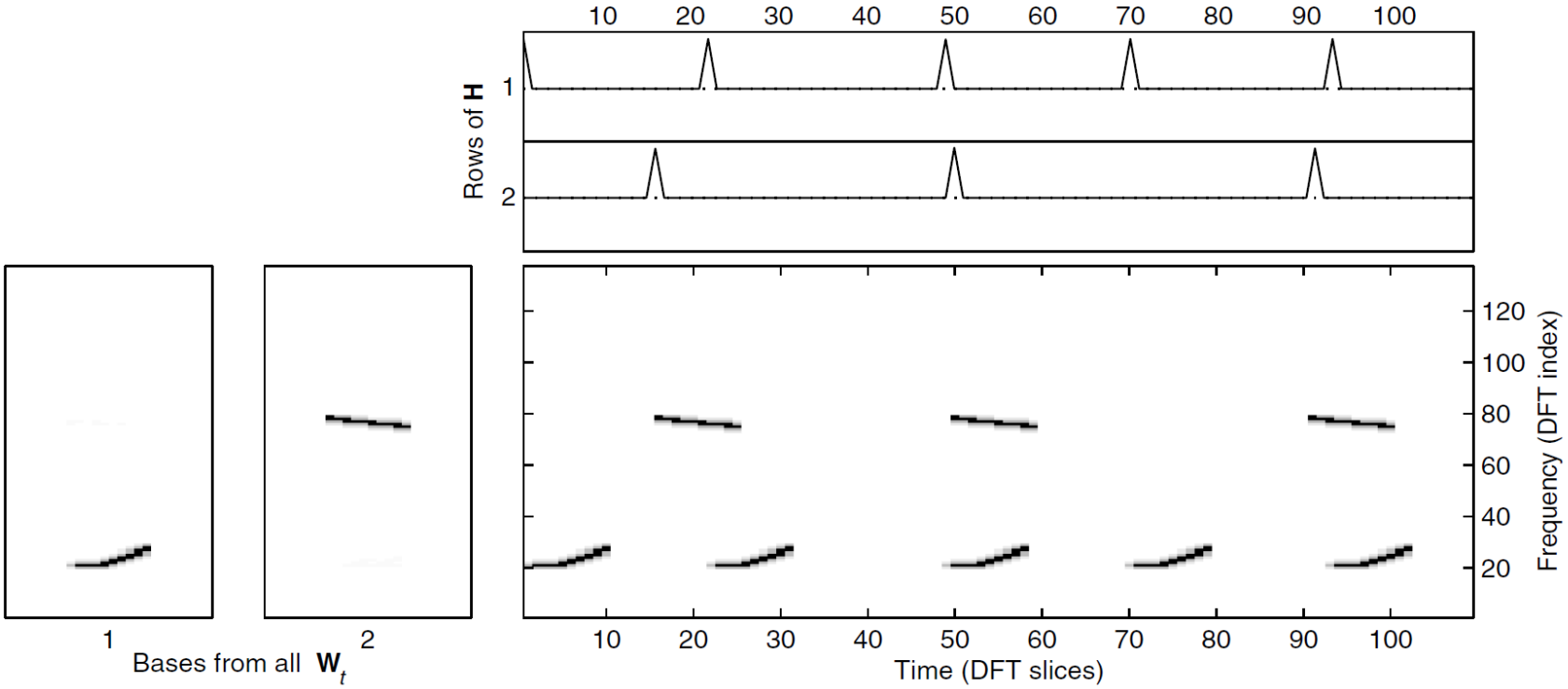
where: $V \in \mathbb{R}^{m \times n}$, $W_t \in \mathbb{R}^{m \times r}$, $H \in \mathbb{R}^{r \times n}$, and $\overrightarrow{H}^t$ shifts the columns of H by t spots to the right

P. Smaragdis, "Non-negative matrix factor deconvolution; extraction of multiple sound sources from monophonic inputs," in Proc. LVA/ICA, pp. 494-499, Sep. 2004.

CITY UNIVERSITY LONDON

Convolutive Extensions (2)

Example of NMFD:



Convolutive Extensions (3)

- Probabilistic counterpart of NMFD:
  Shift-invariant Probabilistic Latent Component Analysis
- Model (shift invariance across 1 dimension):

$$P(x, y) = \sum_{z} P(z) \sum_{\tau} P(x, \tau | z) P(y - \tau | z) \tag{6}$$

- Model (shift invariance across 2 dimensions):

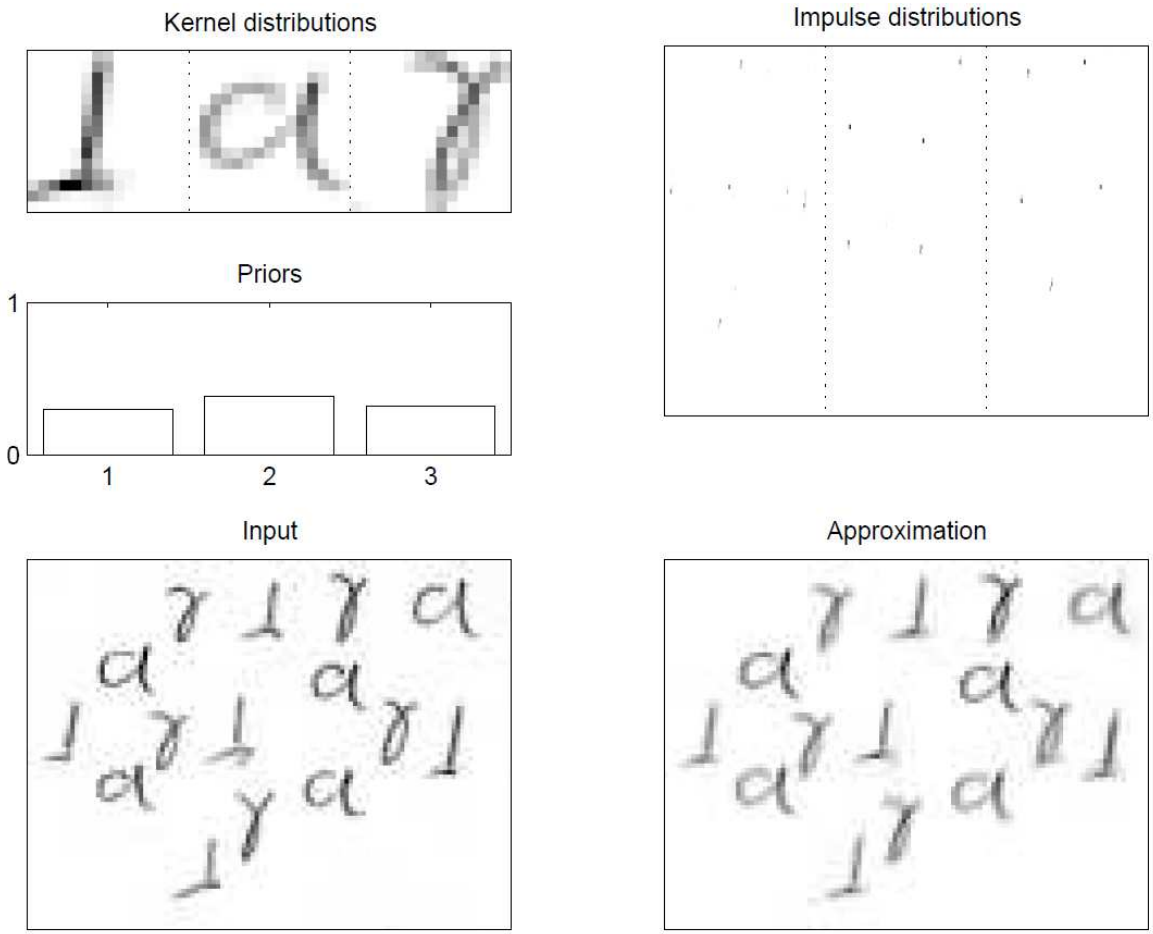
$$P(x, y) = \sum_{z} P(z) \sum_{\tau_x} \sum_{\tau_y} P(\tau_x, \tau_y | z) P(x - \tau_x, y - \tau_y | z) \tag{7}$$

P. Smaragdis, B. Raj, M. V. S. Shashanka, "Sparse and shift-invariant feature extraction from non-negative data," in Proc. *ICASSP*, pp. 2069-2072, 2008.

CITY UNIVERSITY LONDON

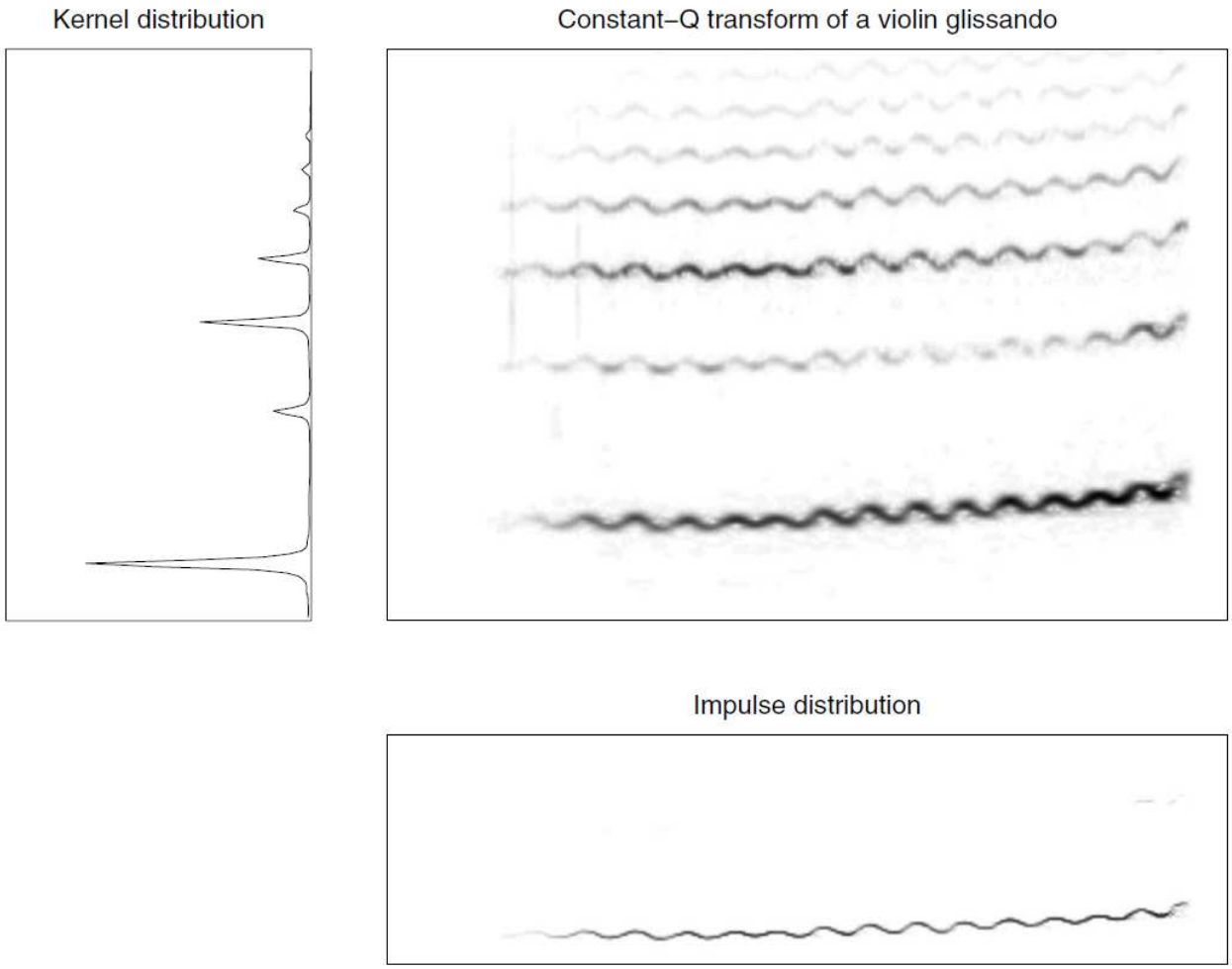
Convolutive Extensions (4)

Example of Shift-invariant PLCA for handwriting recognition:



Convolutive Extensions (5)

Example of Shift-invariant PLCA for musical pitch detection:



Tensorial Extensions (1)

- What if we want to decompose multidimensional data (i.e. tensors)?
- NMF extends to Non-negative Tensor Factorization (NTF)
- Model:

$$\mathcal{V} \approx \sum_{j=1}^{k} \mathbf{u}_1^j \otimes \mathbf{u}_2^j \otimes \ldots \otimes \mathbf{u}_n^j \tag{8}$$

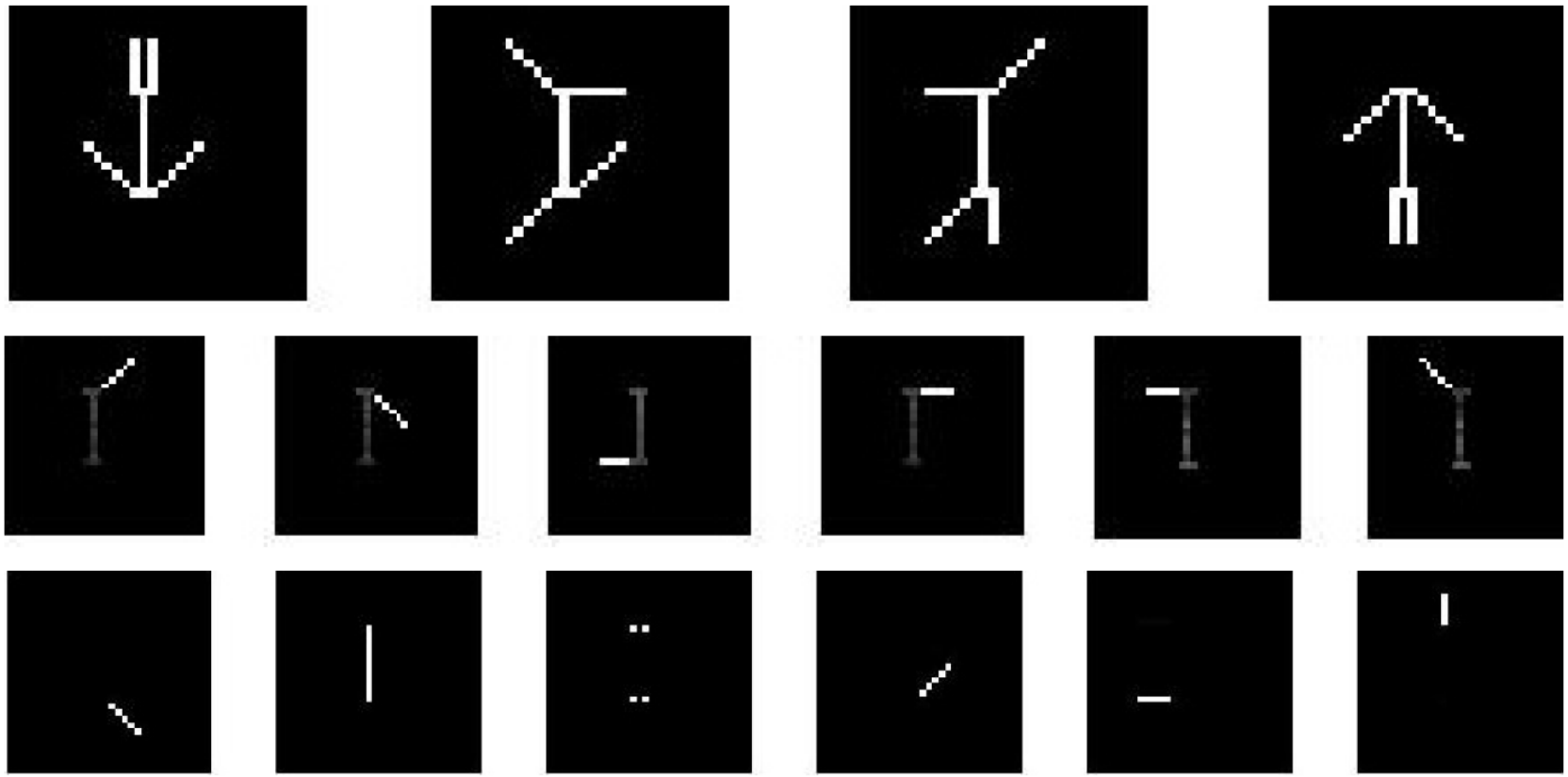
  where $\mathcal{V}$ is an n-way array. The approximation tensor has rank k.
- As in NMF, there are cost functions and update rules for estimating the nk vectors $\mathbf{u}_i^j$

A. Shashua and T. Hazan, "Non-negative tensor factorization with applications to statistics and computer vision," in Proc. *ICML*, pp. 792-799, 2005.

CITY UNIVERSITY LONDON

Tensorial Extensions (2)

Comparison of extracted NMF bases (2nd row) with NTF bases (3rd row):



Tensorial Extensions (3)

- Probabilistic counterpart:
  Probabilistic Latent Component Analysis (PLCA)
- Model:

$$P(\mathbf{x}) = \sum_{z} P(z) \prod_{j=1}^{N} P(x_j|z) \tag{9}$$

  where $P(\mathbf{x})$ is an N-dimensional distribution of the random variable $\mathbf{x} = x_1, x_2, \cdots, x_N$.
- Unknown parameters estimated using EM algorithm

M. Shashanka, B. Raj, and P. Smaragdis, "Probabilistic latent variable models as nonnegative factorizations," *Computational Intelligence and Neuroscience*, Article ID 947438, 2008.

CITY UNIVERSITY LONDON

More Extensions...

- Incorporate sparsity constraints into NMF/PLSA
- Incorporate prior distributions into PLCA models (incorporating information into the problem)
- Keeping fixed bases: supervised algorithms!
- Temporal constraints for modeling time-series: e.g. Non-negative Hidden Markov Model (N-HMM)
- More complex models based on NMF/PLSA: introduce more latent variables
- What if we don't know the size of r? Non-parametric techniques (e.g. GaP-NMF)
- For analysing spectra without losing phase: Complex NMF, High-Resolution NMF

CITY UNIVERSITY LONDON

Resources

- Matlab Statistics Toolbox
- NMFlib (Matlab): http://www.ee.columbia.edu/~grindlay/code.html
- NMF-CUDA (OpenMP): https://github.com/ebattenberg/nmf-cuda
- NMFLAB (Matlab): http://www.bsp.brain.riken.go.jp/ICALAB/nmflab.html
- NMFN (R): http://cran.r-project.org/web/packages/NMFN/index.html
- NMF-DTU (Matlab): http://cogsys.imm.dtu.dk/toolbox/nmf/index.html
- nima (Python): http://nimfa.biolab.si/
- libNMF (C): http://www.univie.ac.at/rlcta/software/
- ITL Toolbox (C++): http://www.insight-journal.org/browse/publication/152
- NNMA (C++): http://people.kyb.tuebingen.mpg.de/suvrit/work/progs/nnma.html