Non-negative matrix factorization with Gaussian process priors

Mikkel N. Schmidt

Technical University of Denmark

Outline

- **Non-negative matrix factorization (NMF)**

  Examples of applications
  - DNA microarray analysis
  - Monaural audio separation
- **Gaussian processes (GP)**
- **NMF with GP-priors**

  Example of application
  - Chemical shift brain imaging

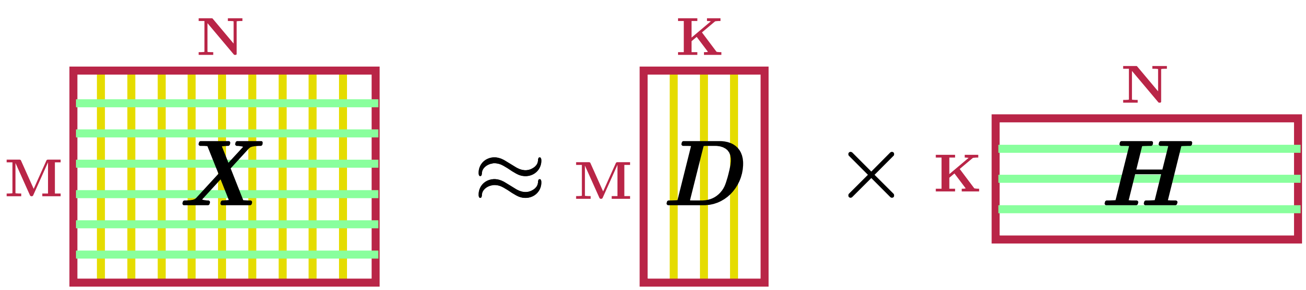
Non-negative matrix factorization

- **Non-negative bi-linear decomposition**

$$x_{i,j} \approx \sum_{k=1}^{K} d_{i,k} \cdot h_{k,j} \quad \text{s.t. } d_{i,k}, h_{k,j} \geq 0$$

- **In matrix notation**

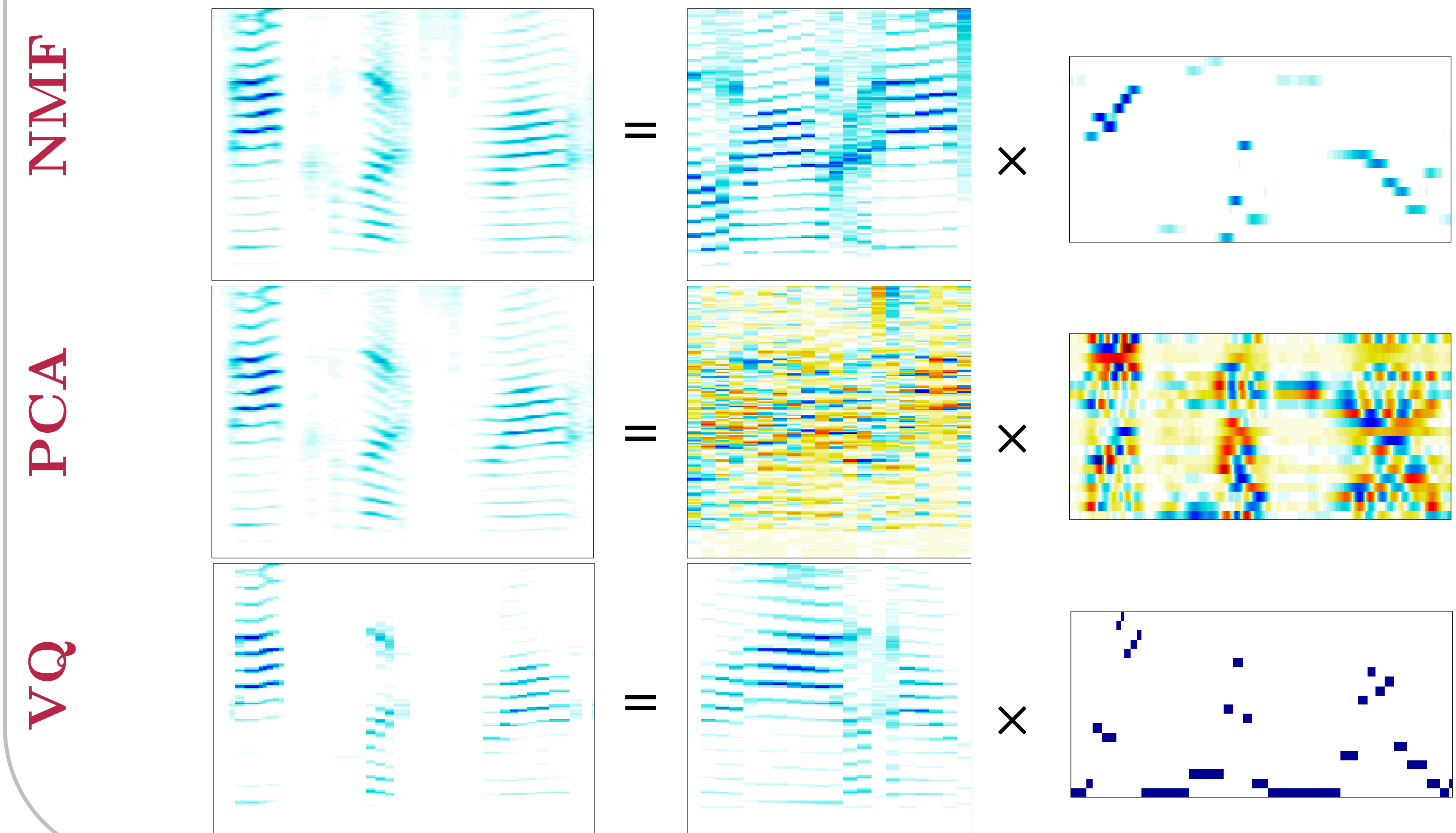
$$\boldsymbol{X} \approx \boldsymbol{D}\boldsymbol{H} \quad \text{s.t. } \boldsymbol{D}, \boldsymbol{H} \geq 0$$



Why non-negativity?

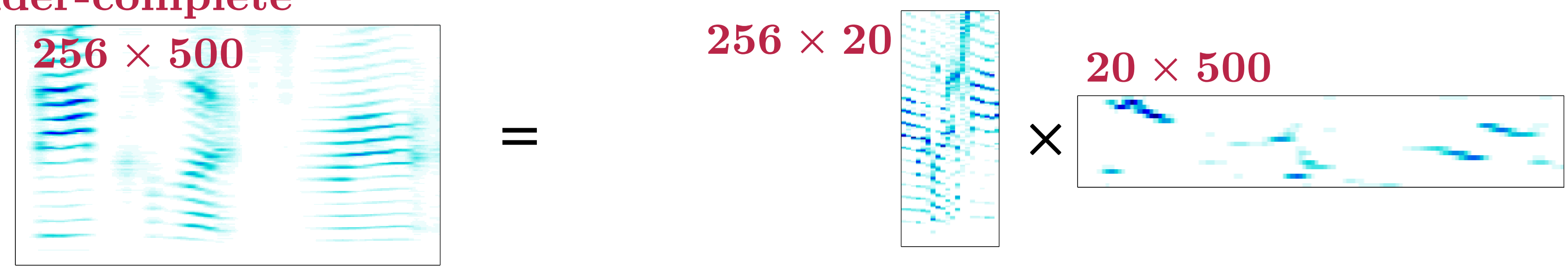
- **Many signals are non-negative by nature**
  - **Pixel intensities**
  - **Amplitude spectra**
  - **Occurrence counts**
  - **Discrete probabilities**
  - **etc.**
- **Non-subtractive model**
  - **No terms cancel out**
  - **Parts-based: The whole is modeled as a sum of parts**

NMF, PCA, and VQ

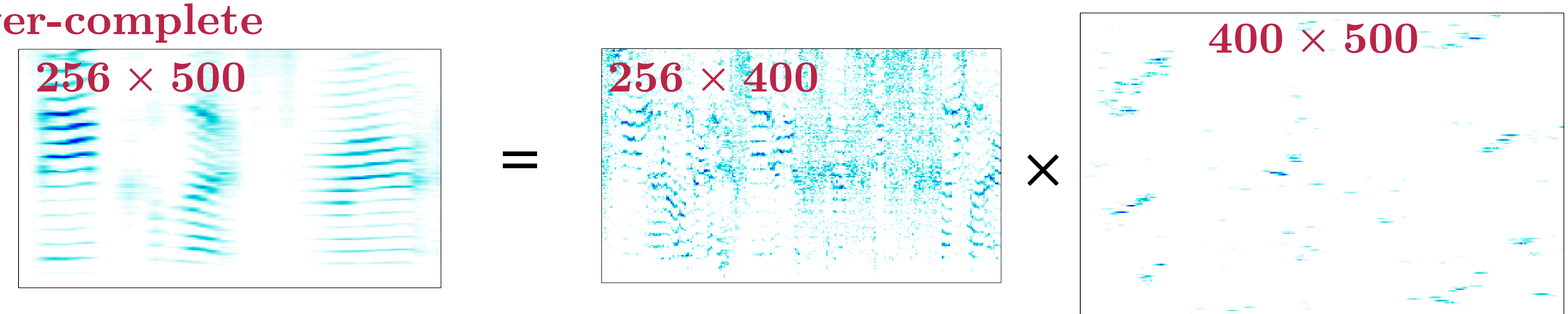


Over-complete decomposition

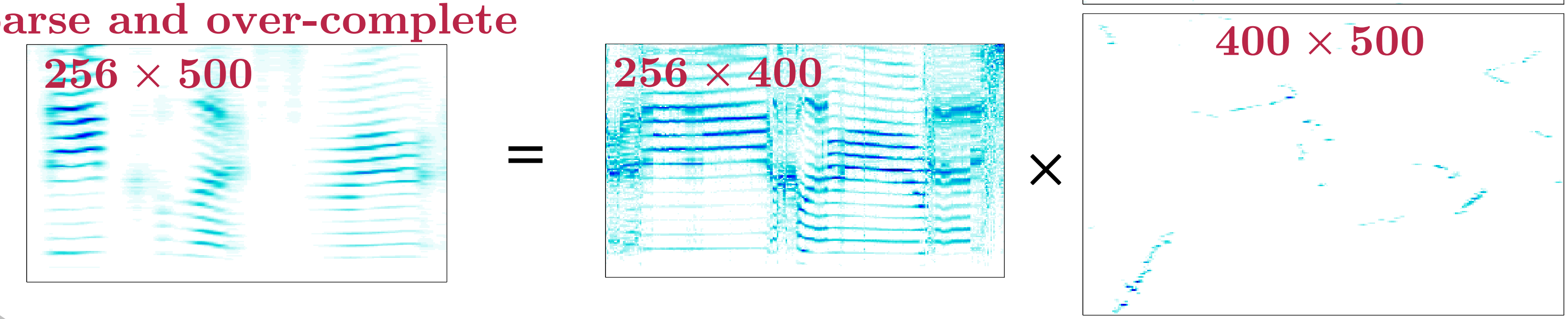
Under-complete



Over-complete



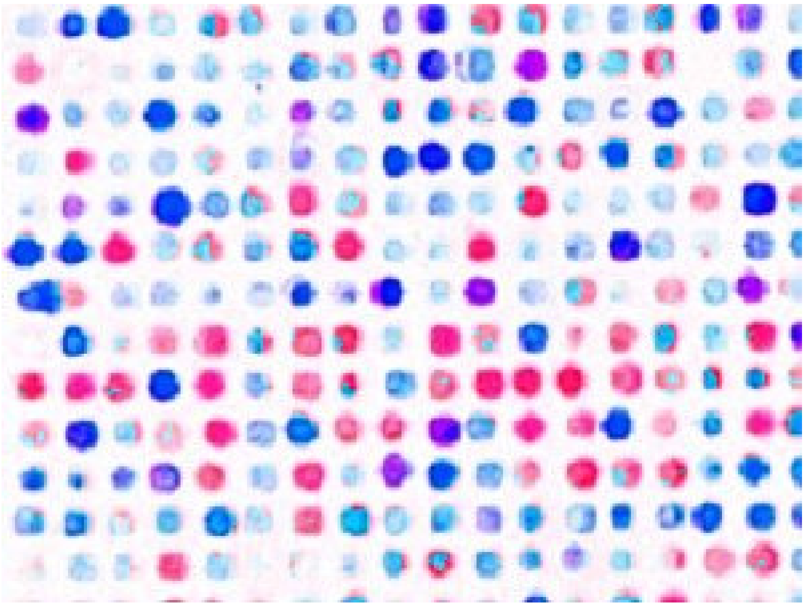
Sparse and over-complete



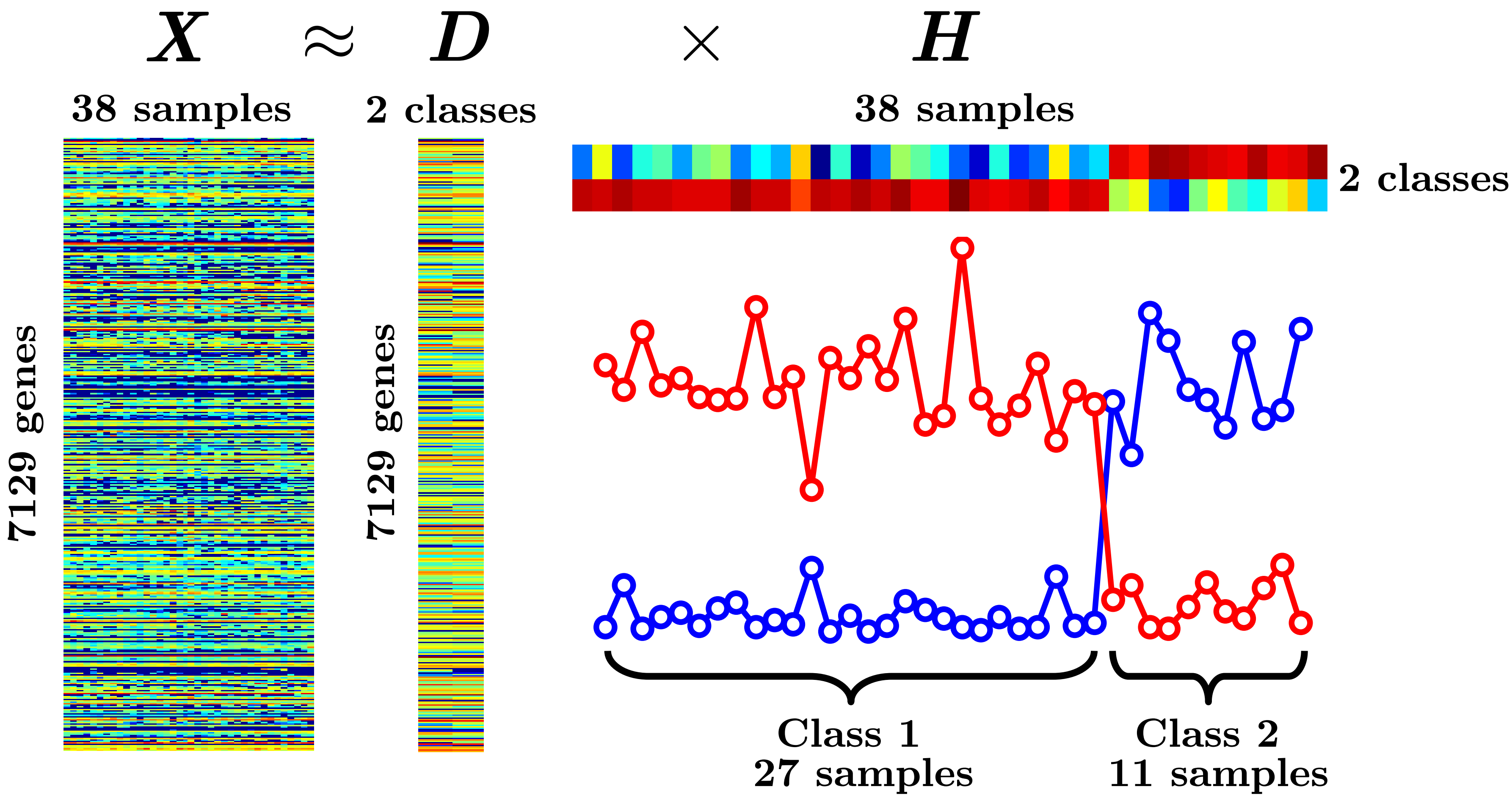
DNA microarray analysis

- **DNA microarray technology enables parallel analysis of thousands of genes**

- **Data can be represented in non-negative matrix e.g. gene $\times$ experiment**

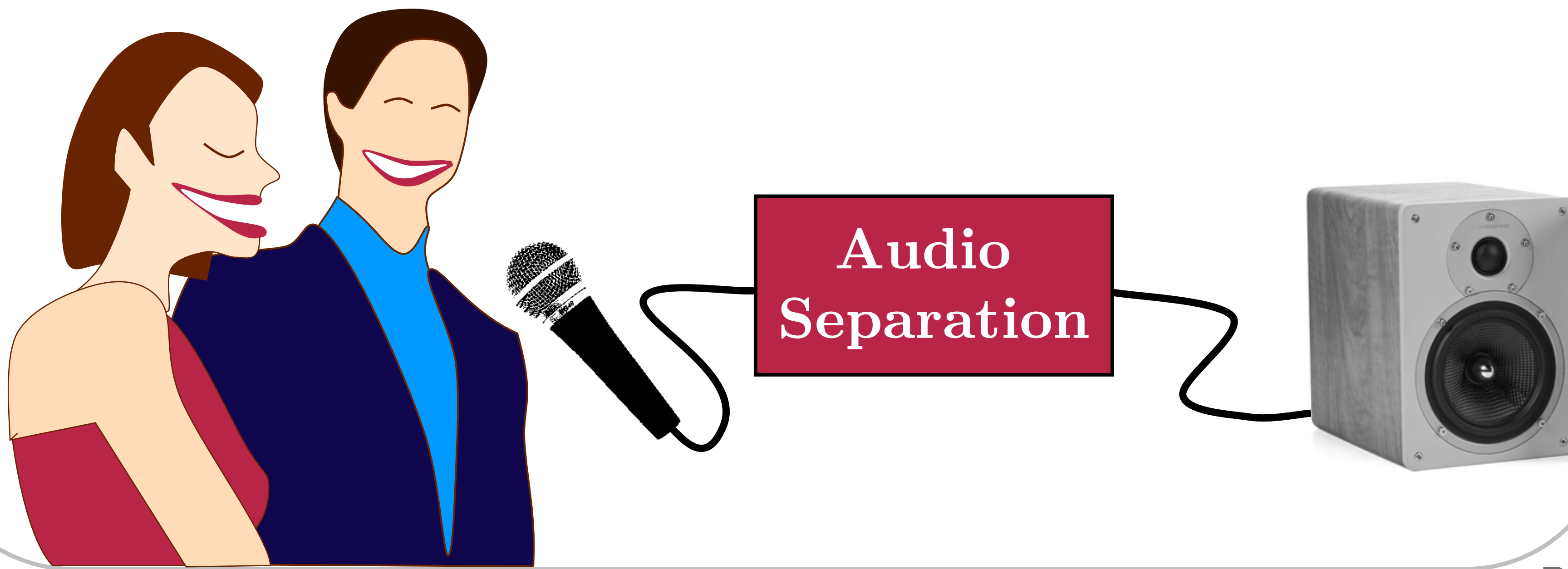


Two types of leukemia



Monaural audio separation

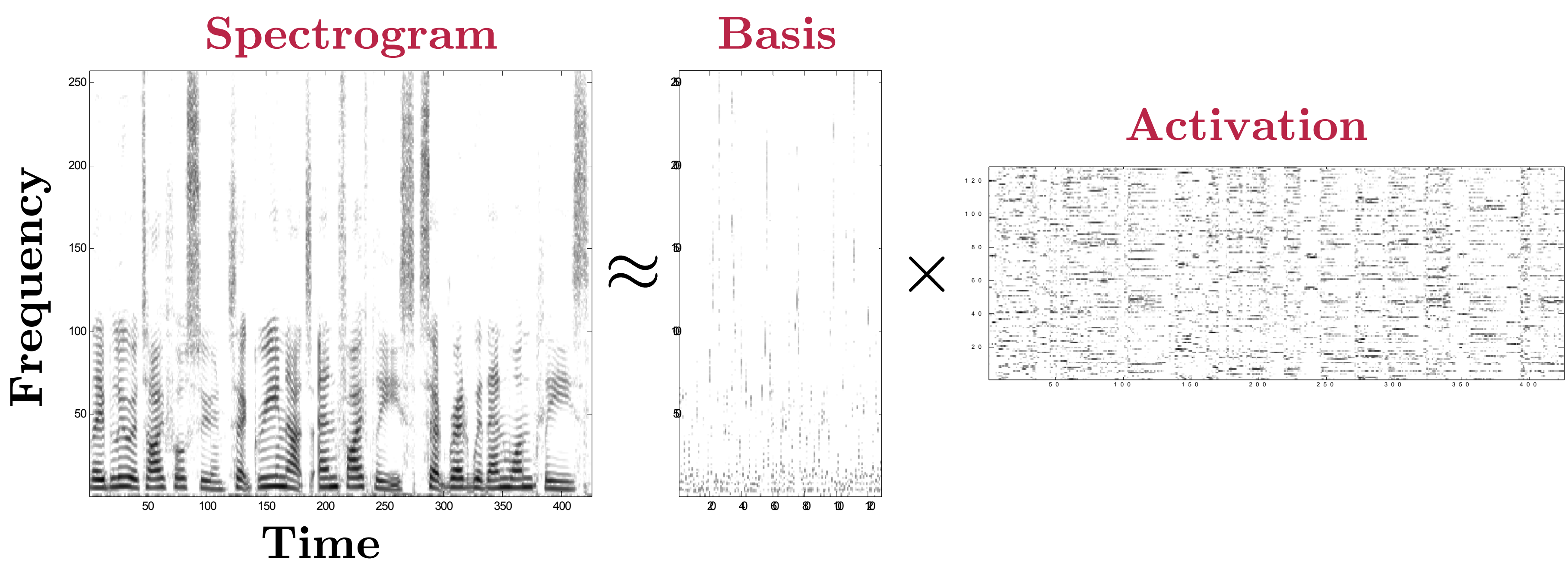
- **Problem:** Separate audio sources using only one-microphone recording of mixture



Amplitude spectrogram

- **Audio represented as a non-negative matrix**

$$\boldsymbol{X} \approx \boldsymbol{D}\boldsymbol{H}$$



Audio separation with NMF

1. **Learn a basis for each source**

$$\boldsymbol{D}_1, \boldsymbol{D}_2, \ldots, \boldsymbol{D}_N$$

2. **Compute activation for mixture**

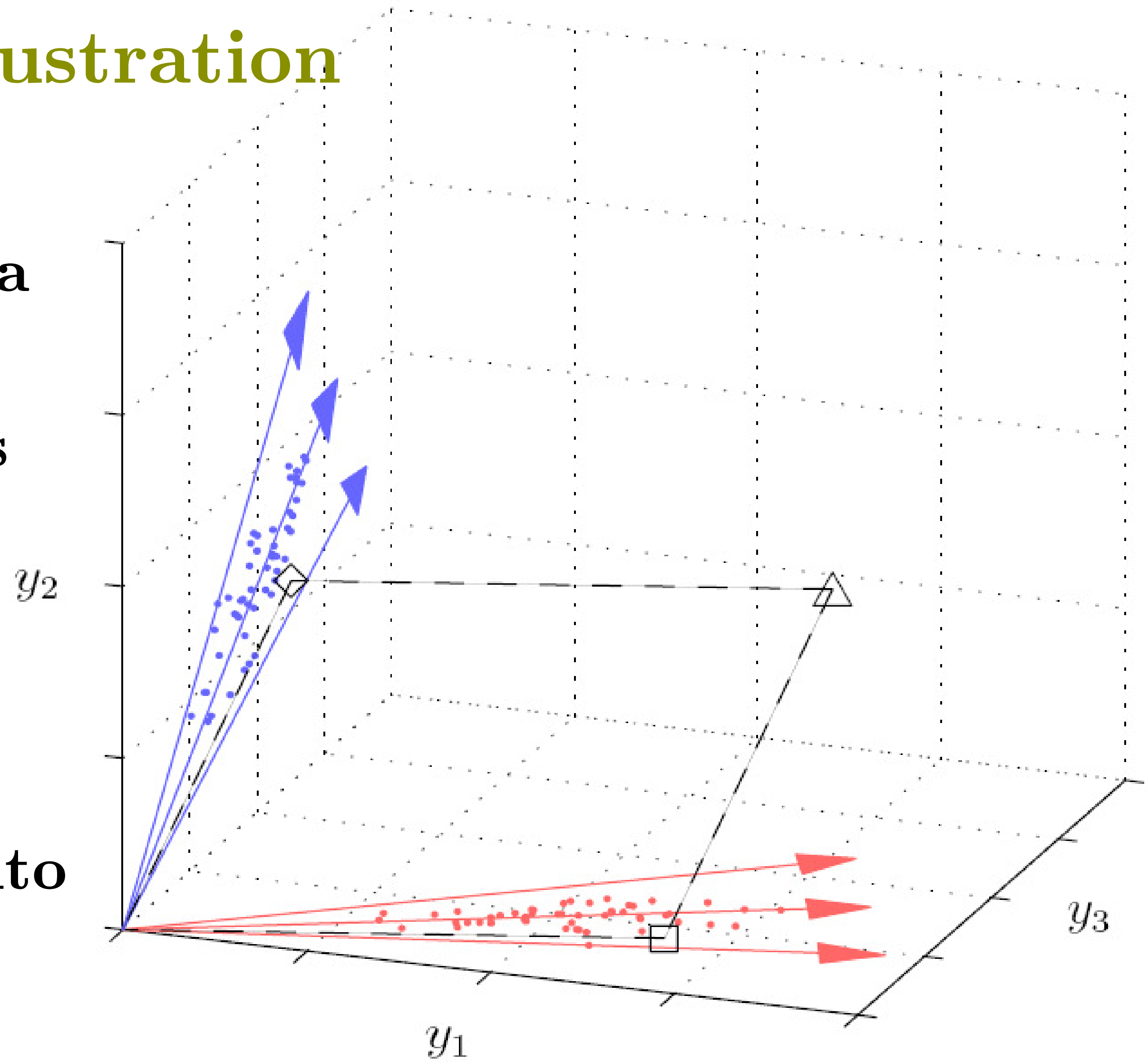
$$\boldsymbol{X} \approx \boldsymbol{D}\boldsymbol{H} = [\boldsymbol{D}_1, \boldsymbol{D}_2, \ldots, \boldsymbol{D}_N] \begin{bmatrix} \boldsymbol{H}_1 \\ \boldsymbol{H}_2 \\ \vdots \\ \boldsymbol{H}_N \end{bmatrix}$$

3. **Reconstruct each source separately**

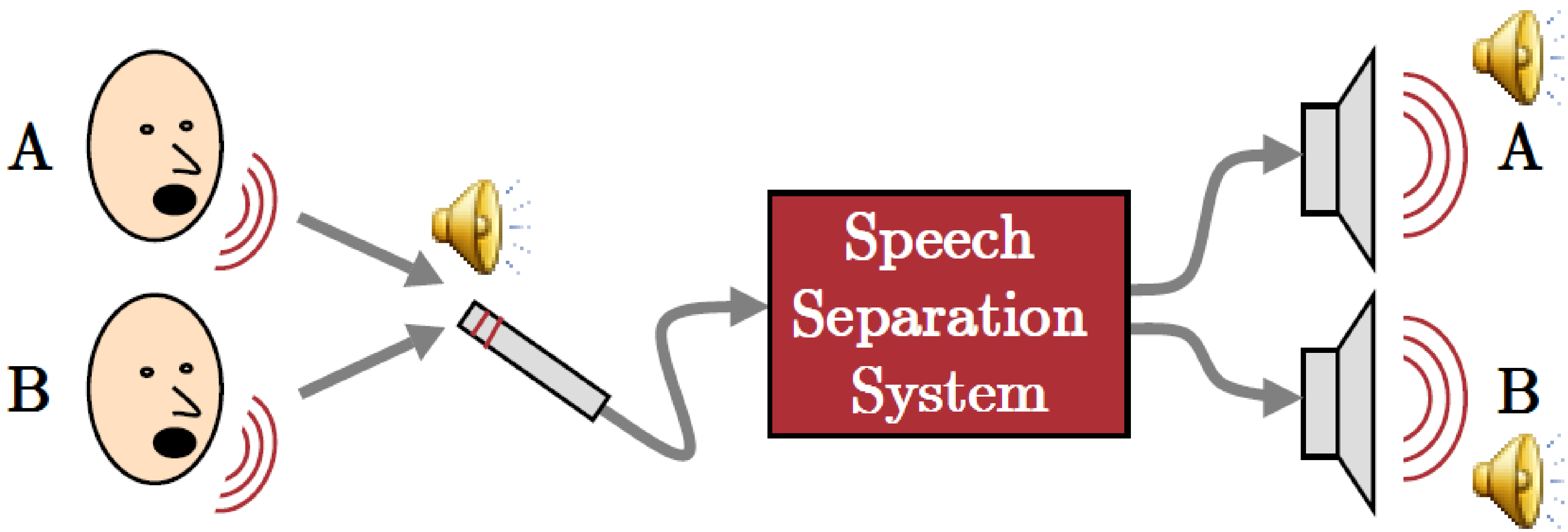
$$\widehat{\boldsymbol{X}}_n = \boldsymbol{D}_n \boldsymbol{H}_n$$

Subspace illustration

- **Training data**
- **Basis vectors**
- △ **Observed mixture**
- ◇ □ **Mappings onto subspaces**



Audio demonstration



Computing NMF

- **Minimize divergence between $\boldsymbol{X}$ and $\boldsymbol{DH}$, possibly plus some regularization**

$$\min_{\boldsymbol{D},\boldsymbol{H}\geq 0} \mathcal{D}(\boldsymbol{X},\boldsymbol{DH}) + \mathcal{R}(\boldsymbol{D},\boldsymbol{H})$$

- **Constrained minimization problem: A variety of algorithms for different divergence measures and regularizations**
- **Choice of divergence measure corresponds to assumptions about the data/noise/factor distribution**

Maximum likelihood NMF

Example: Gaussian i.i.d. noise

- **Likelihood function**

$$p(\boldsymbol{X}|\boldsymbol{D},\boldsymbol{H}) = \prod_{i,j} \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{(\boldsymbol{X}_{i,j} - [\boldsymbol{D}\boldsymbol{H}]_{i,j})^2}{2\sigma^2}\right)$$

- **Negative log-likelihood (serves as divergence)**

$$-\log p(\boldsymbol{X}|\boldsymbol{D},\boldsymbol{H}) \propto \sum_{i,j} (\boldsymbol{X}_{i,j} - [\boldsymbol{D}\boldsymbol{H}]_{i,j})^2$$

Maximum a posteriori NMF

- **Posterior distribution**

$$p(\boldsymbol{D},\boldsymbol{H}|\boldsymbol{X}) = \frac{p(\boldsymbol{X}|\boldsymbol{D},\boldsymbol{H})p(\boldsymbol{D},\boldsymbol{H})}{p(\boldsymbol{X})}$$

- **Negative log posterior**

$$-\log p(\boldsymbol{D},\boldsymbol{H}|\boldsymbol{X}) \propto \underbrace{-\log p(\boldsymbol{X}|\boldsymbol{D},\boldsymbol{H})}_{\substack{\textbf{Likelihood}\\ \textbf{(divergence)}}} \underbrace{-\log p(\boldsymbol{D},\boldsymbol{H})}_{\substack{\textbf{Prior}\\ \textbf{(regularization)}}}$$

Bayesian NMF

- **Posterior distribution**

$$p(\boldsymbol{D},\boldsymbol{H}|\boldsymbol{X}) = \frac{p(\boldsymbol{X}|\boldsymbol{D},\boldsymbol{H})p(\boldsymbol{D},\boldsymbol{H})}{p(\boldsymbol{X})}$$

- **In general, integrals over the posterior are intractable ➔ approximate, e.g., using MCMC.**

Probabilistic model of factors

- **Standard NMF**

$$\boldsymbol{X} \approx \boldsymbol{D}\boldsymbol{H} \quad \text{s.t. } \boldsymbol{D}, \boldsymbol{H} \geq 0$$

  - Factors constrained to be non-negative
  - No other assumptions about the factors

- **Prior distribution over factors**

$$p(\boldsymbol{D}, \boldsymbol{H})$$

  - Prior distribution captures non-negativity as well as other properties, such as sparseness, smoothness, symmetries, etc.

Which prior distributions to use?

- **Distribution over non-negative reals**
  - **Rectified Gaussian: L_2 norm regularization**
  - **One-sided exponential: L_1 norm regularization**
- **Gaussian process mapped to the non-negative reals**
  - **Flexible, principled, and practical approach**
  - **Sparseness, smoothness, symmetries, etc.**

Gaussian Processes

- **A stochastic process which generates samples, x_i, such that any linear combination of x_i is Gaussian**
- **Characterized by its mean and covariance function**
- **Defines a distribution over functions**

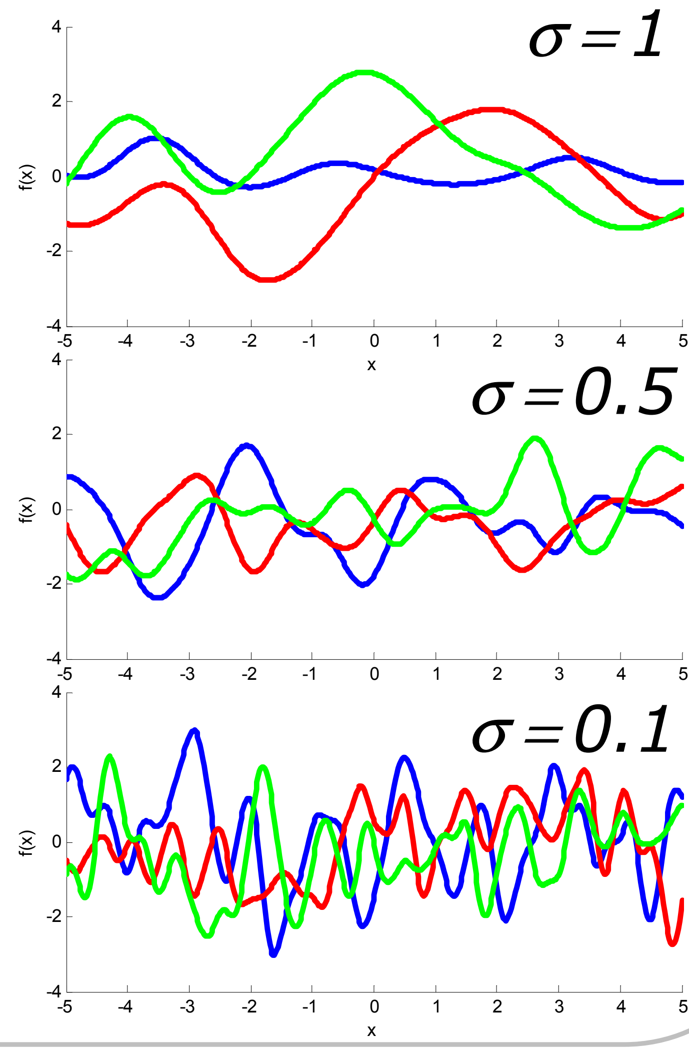
Example of Gaussian process

- **Mean function**

$$m(x) = 0$$

- **Covariance function**

$$k(x_1, x_2) = \exp\left(-\frac{(x_1 - x_2)^2}{2\sigma^2}\right)$$

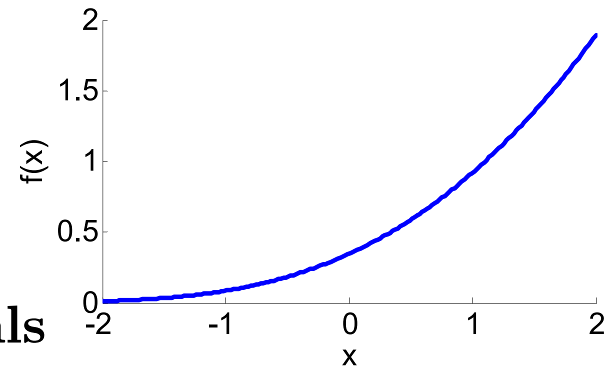


NMF with GP priors

$$\boldsymbol{X} \approx \boldsymbol{D}\boldsymbol{H}$$

$$\boldsymbol{D} = f_D(\boldsymbol{G_D}) \qquad \boldsymbol{H} = f_H(\boldsymbol{G_H})$$

- $\boldsymbol{G_D}$ and $\boldsymbol{G_H}$ are Gaussian processes
- Link functions
  - Strictly increasing
  - Maps the reals to the non-negative reals



Link Function

- **Map marginal distribution of Gaussian process to desired marginal distribution**

$$\boldsymbol{H} = f_H(\boldsymbol{G_H}) = P_H^{-1}\Big(P_{G_H}(\boldsymbol{G_H})\Big)$$

- **Example: Gaussian-to-Exponential**

$$f_H(\boldsymbol{G_H}) = -\frac{1}{\lambda}\log\left(\frac{1}{2} - \frac{1}{2}\Phi\left(\frac{\boldsymbol{G_H}}{\sqrt{2}\sigma_i}\right)\right)$$

Change of Variable

- **New variables δ and η**

$$\boldsymbol{D} = f_D(\boldsymbol{C}_D^\top \boldsymbol{\delta}) \qquad \boldsymbol{H} = f_H(\boldsymbol{C}_H^\top \boldsymbol{\eta})$$

  - **$\boldsymbol{C}$ is Cholesky decomposition of covariance matrix**
  - **Variables δ and η are i.i.d. Gaussian**

- **MAP estimate: $p(\delta, \eta | \boldsymbol{X})$**
  - **More robust and less local minima**
  - **Unconstrained optimization (use e.g. conj. grad.)**

Illustration of NMF with GP priors

- **Full matrix data**

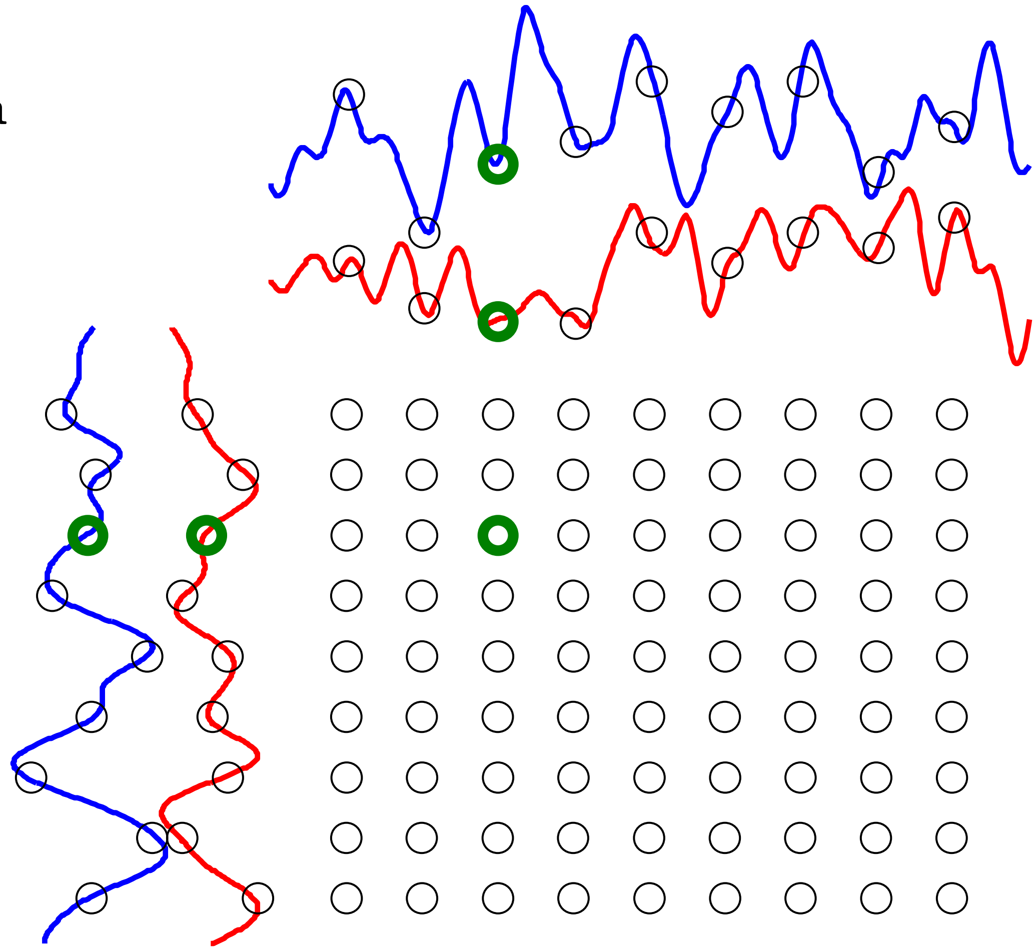


Illustration of NMF with GP priors

- **Full matrix data**
- **Missing values**

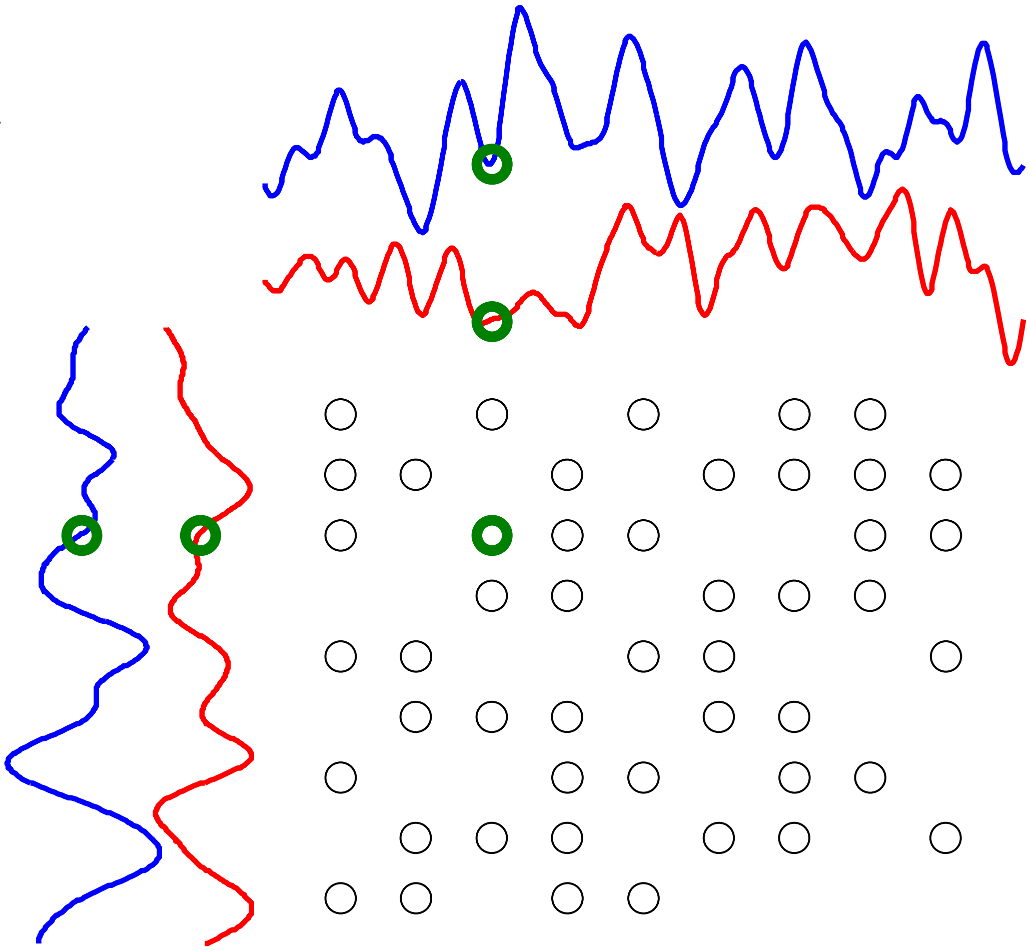
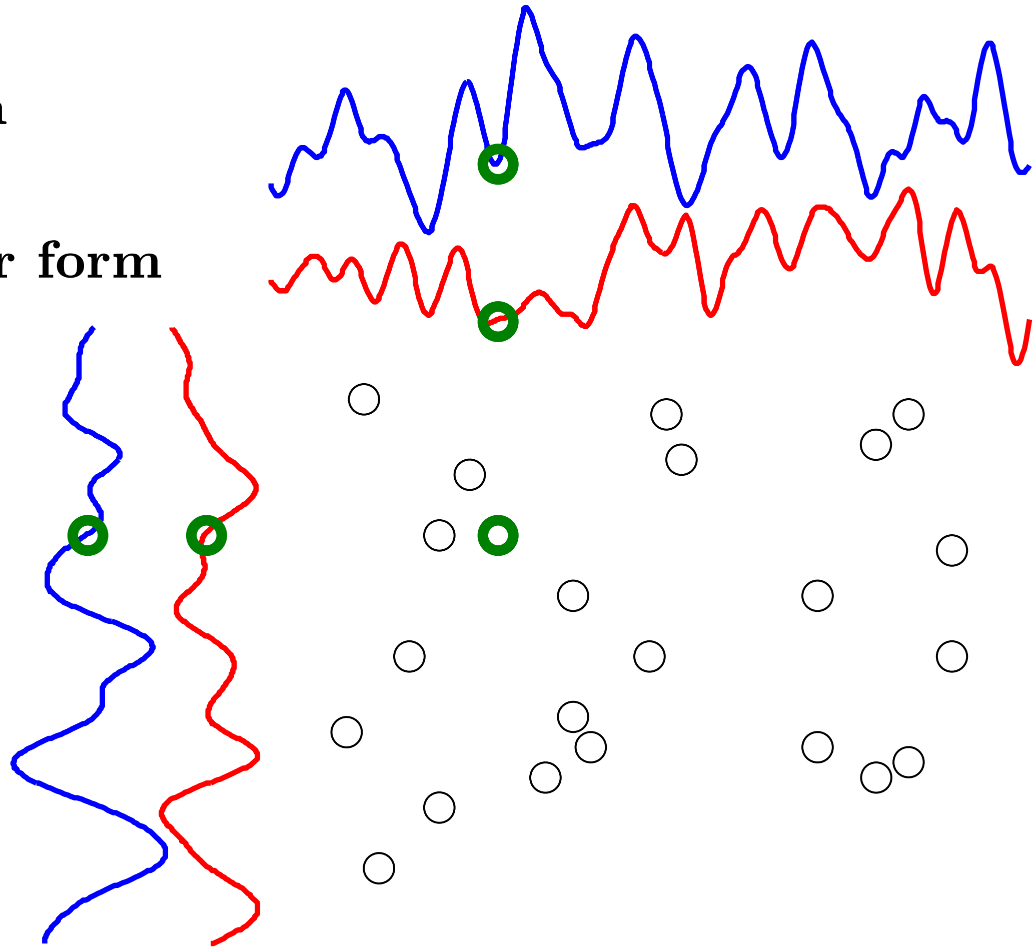
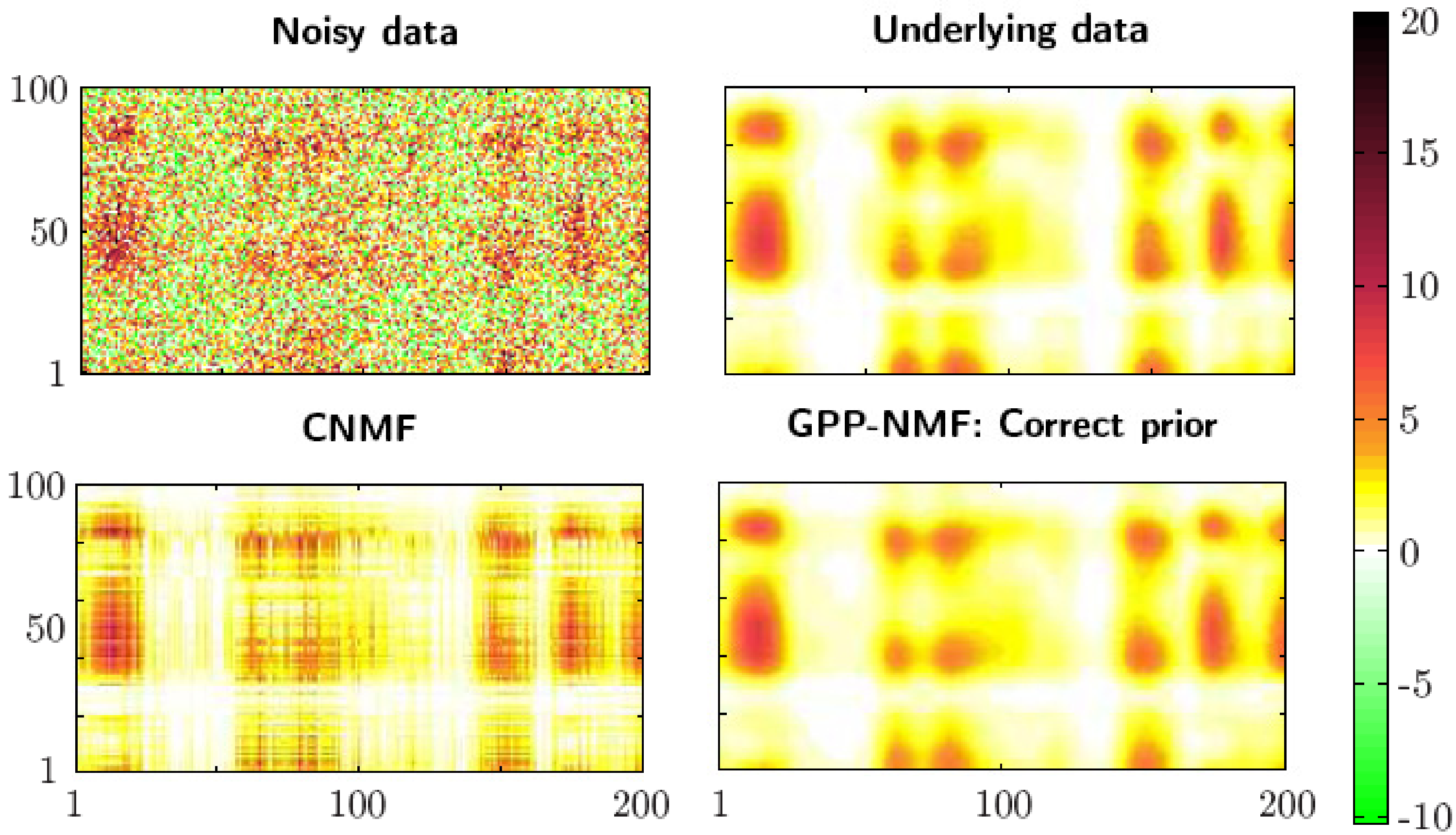


Illustration of NMF with GP priors

- Full matrix data
- Missing values
- General bi-linear form

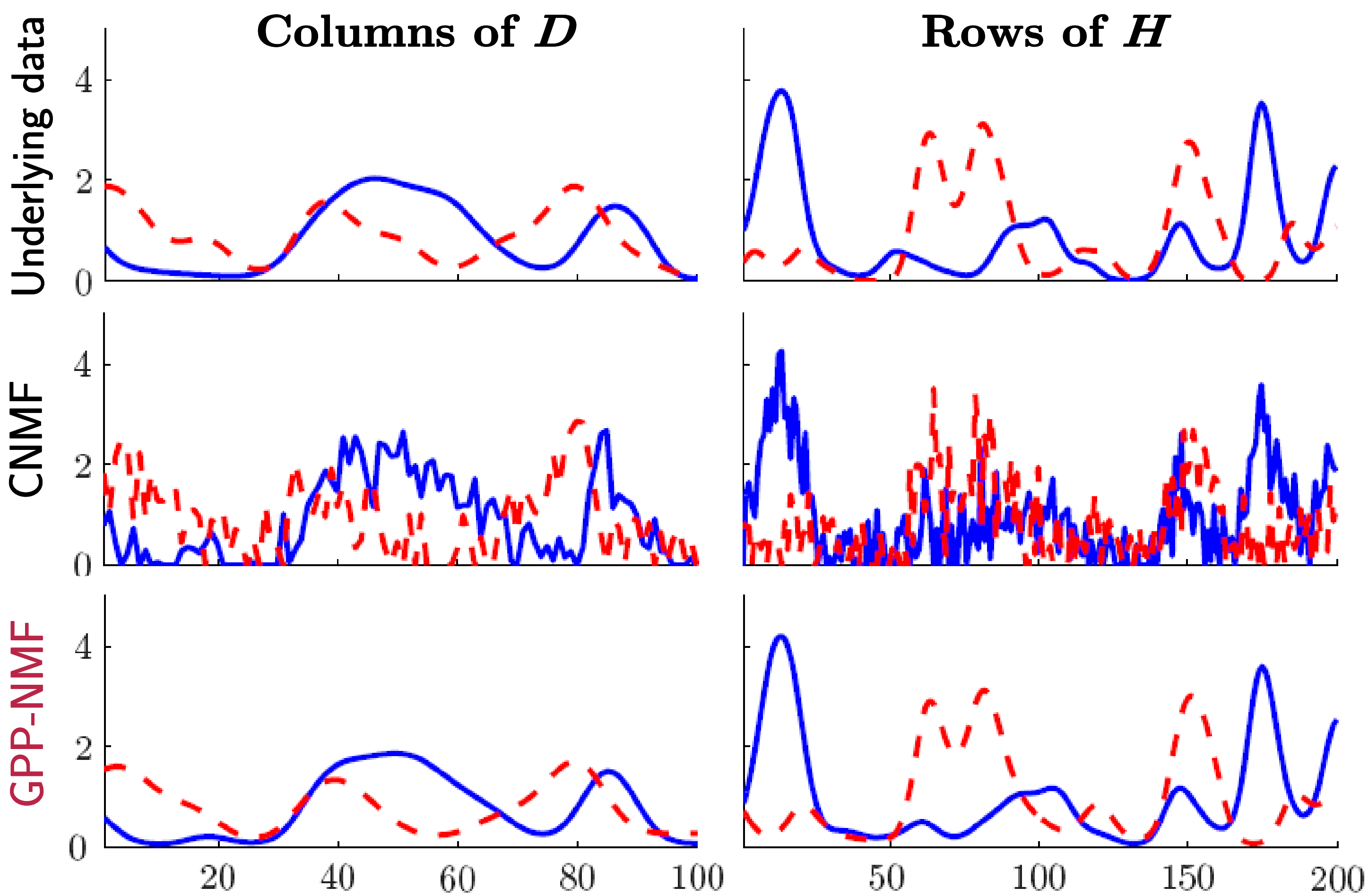


Toy Example



Noisy data
Underlying data
CNMF
GPP-NMF: Correct prior
100
50
1
1
100
200
20
15
10
5
0
-5
-10

Toy Example

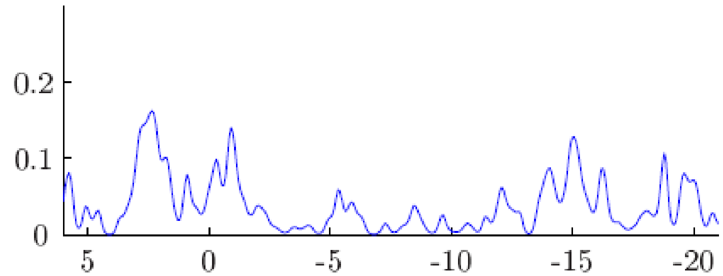

Columns of D
Rows of H
Underlying data
CNMF
GPP-NMF
0
2
4
20
40
60
80
100
50
100
150
200

Chemical Shift Brain Imaging

- **Data:** 369 point chemical shift spectra measured at 512 positions (8×8×8 grid) in human skull.

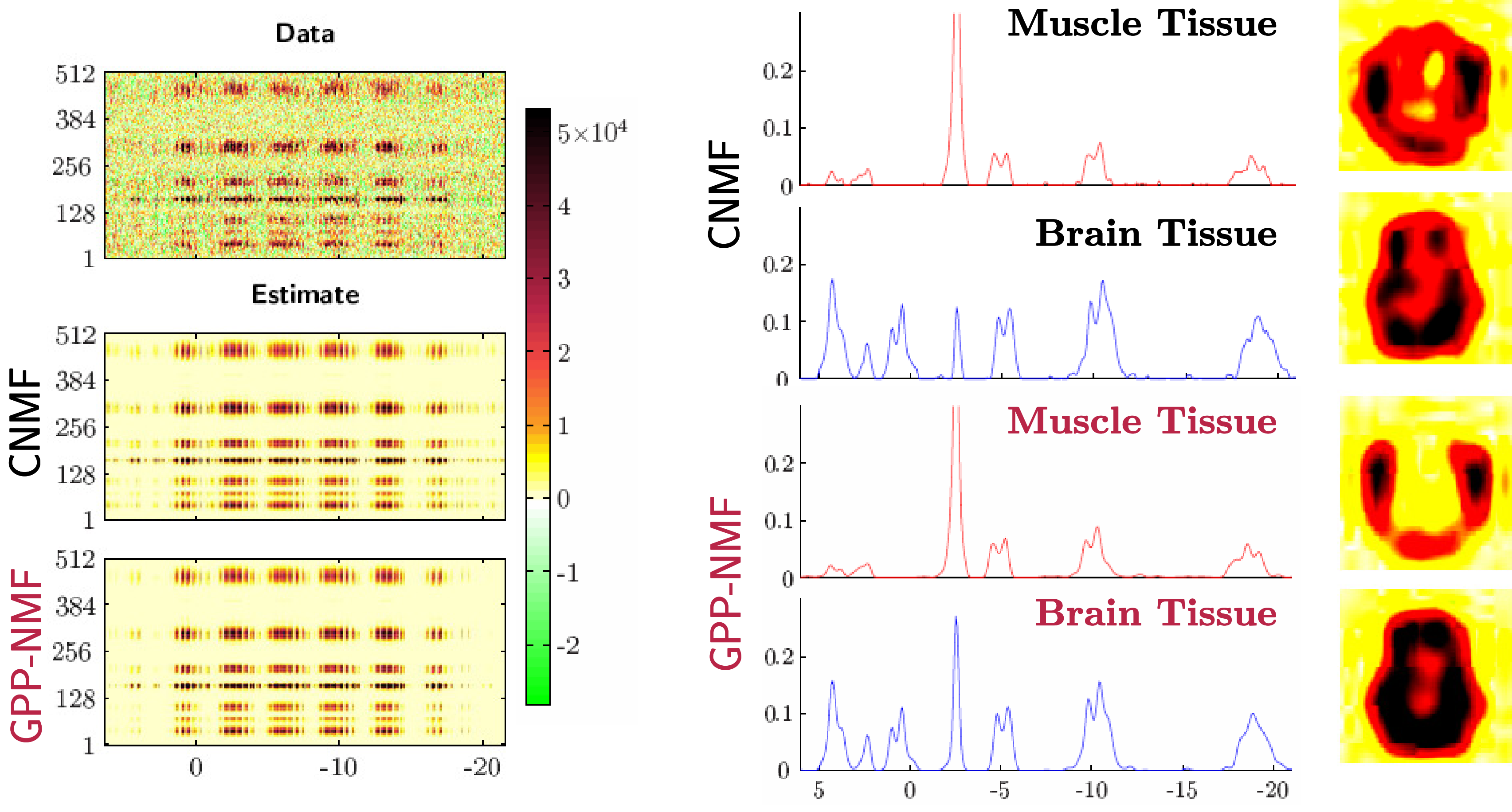
- **Task:** Distinguish between brain and muscle tissue

- **Prior for spectra:** Smooth, exponentially distributed.



- **Prior for activations in skull:** Smooth in 3D, exponentially distributed, left-to-right symmetric.

Chemical Shift Brain Imaging



Conclusions

- **NMF**
  - **General and versatile method**
  - **Can be used to analyze a variety of problems including**
    - **DNA microarray analysis**
    - **Audio signal separation**
- **NMF with GP-priors**
  - **Extends the NMF framework by adding prior information**
  - **Can improve the quality of non-negative factorizations**

Future work

- **Full Bayesian treatment of the model (MCMC)**
- **Learn parameters of kernel function**
- **Learn link functions from data**
- **Learn number of components (nonparametric Bayes)**

References

Non-negative Matrix Factorization

- D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. Nature, 401(6755):788–791, 1999.
- P. Paatero and U. Tapper. Positive matrix factorization: A nonnegative factor model with optimal utilization of error-estimates of data values. Environmetrics, 5:111–126, 1994.
- Michael W. Berry, Murray Browne, Amy N. Langville, V. Paul Pauca, and Robert J. Plemmons. Algorithms and applications for approximate nonnegative matrix factorization. Computational Statistics and Data Analysis, 2006.

DNA Microarray Analysis

- Jean-Philippe Brunet, Pablo Tamayo, Todd R. Golub, and Jill P. Mesirov. Metagenes and molecular pattern discovery using matrix factorization. Proceedings of the National Academy of Sciences (PNAS), 101(12):4164–4169, Mar 2004.

References

Speech Separation with NMF

- Mikkel N. Schmidt and Rasmus K. Olsson, Linear Regression on Sparse Features for Single-Channel Speech Separation. Applications of Signal Processing to Audio and Acoustics, IEEE Workshop on, 2007.
- Mikkel N. Schmidt and Rasmus K. Olsson, Single-Channel Speech Separation using Sparse Non-Negative Matrix Factorization. Spoken Language Processing, ISCA International Conference on, 2006.

Gaussian Processes

- Carl Edward Rasmussen and Christopher K. I. Williams, Gaussian Processes for Machine Learning. MIT Press, 2006.

NMF with GP-priors

- Mikkel N. Schmidt and Hans Laurberg, Non-negative matrix factorization with Gaussian process priors. Computational Intelligence and Neuroscience, 2008.