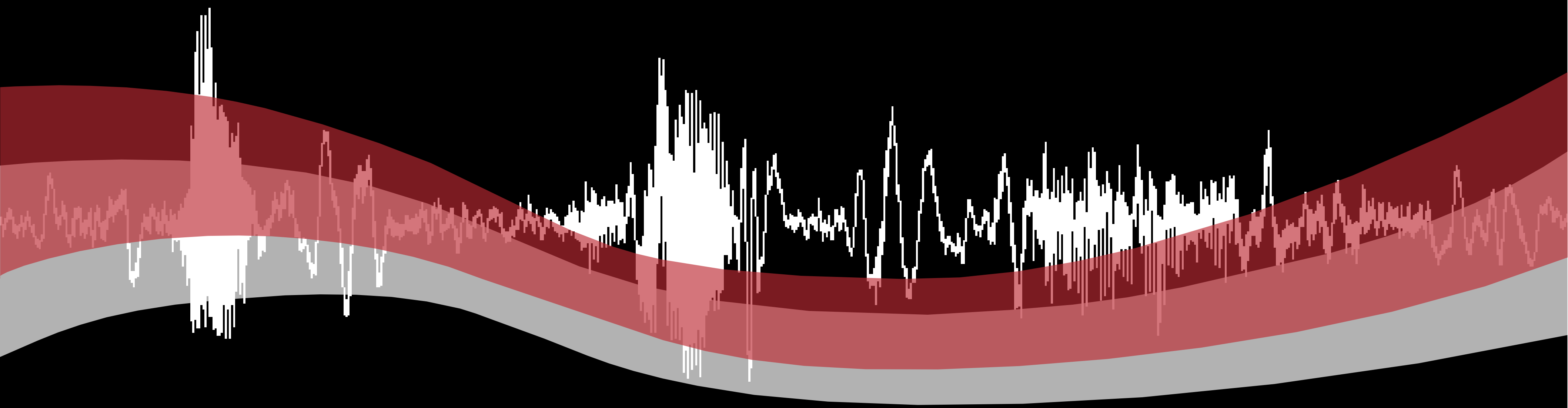
Single-channel source separation using non-negative matrix factorization



Mikkel N. Schmidt
Technical University of Denmark
mns@imm.dtu.dk
www.mikkelschmidt.dk

Agenda

Single-channel source separation

Non-negative matrix factorization

NMF 2-D deconvolution
(with Morten Mørup)

Speech separation using sparse NMF
(with Rasmus K. Olsson)

NMF with Gaussian process priors
(with Hans Laurberg)

Single-channel source separation

Additive model

Mixture → Sources → Noise

$$x(n) = \sum_{k=1}^{K} s_k(n) + e(n)$$

Under-determined problem: More information required

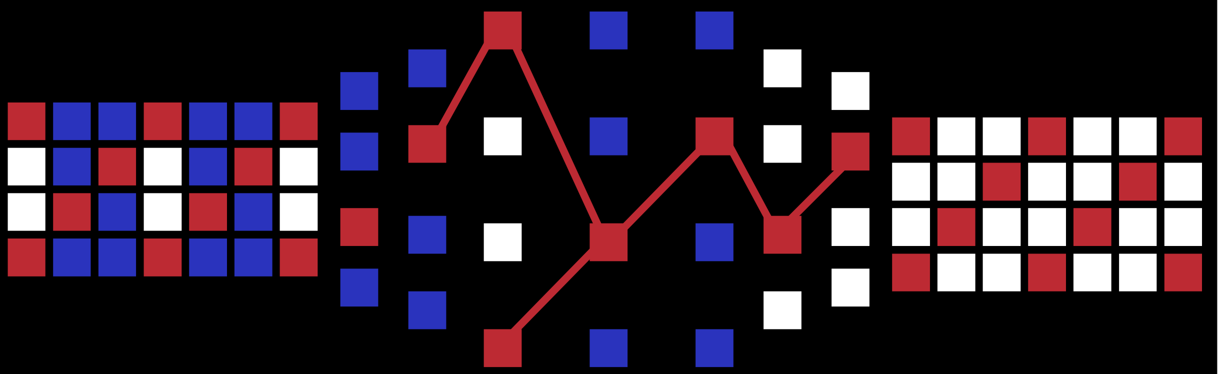
Example: Two-source noise-free

$$s_1(n) = \bar{s}(n) \qquad s_2(n) = x(n) - \bar{s}(n)$$

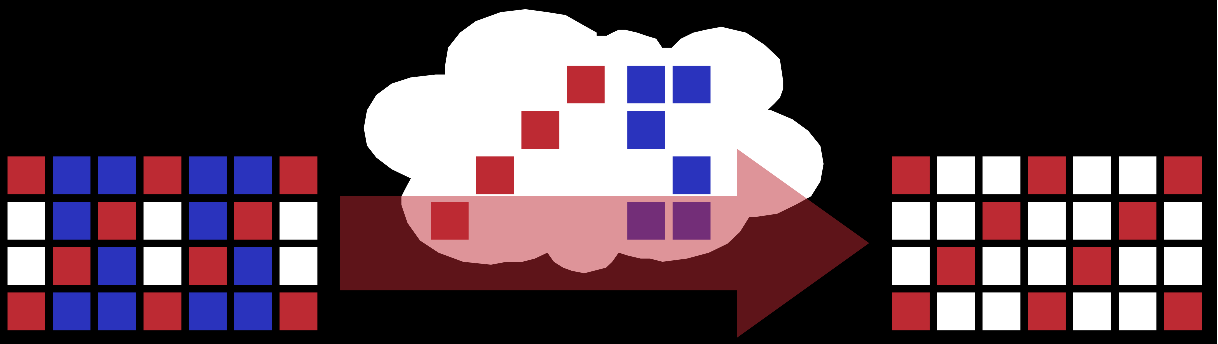
Approaches to single-channel source separation

Filtering

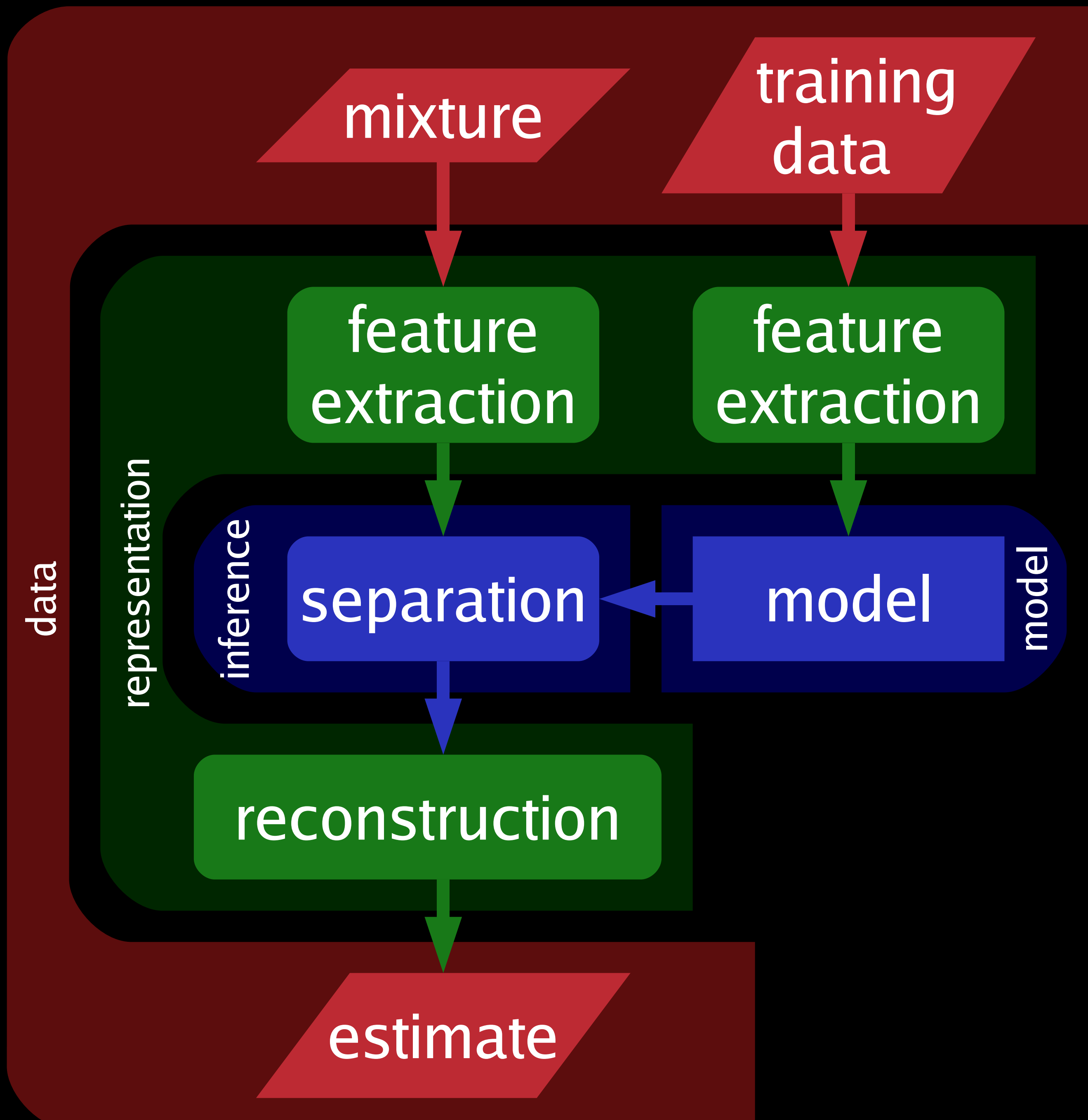
Decomposition and grouping



Source modelling



Model-based source-separation



Signal representation

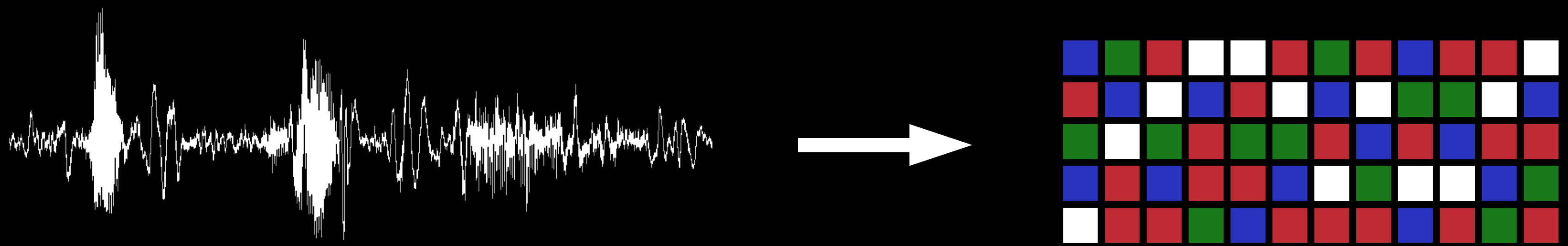
Emphasize desired characteristics

Introduce invariances

Allow assumptions of independence or exchangeability

Reduce dimensionality

Allow signal reconstruction



Model

Likelihood

$$p(\boldsymbol{x}|\boldsymbol{s}_1, \ldots, \boldsymbol{s}_K)$$

Prior

$$p(\boldsymbol{s}_1, \ldots, \boldsymbol{s}_K) = \prod_{k=1}^{K} p(\boldsymbol{s}_k)$$

- Mixing model
- Source model
  - Model building
  - Model training
  - Model adaptation
- Goals
  - Accurately model sources and mixing process
  - Enable efficient inference

Inference

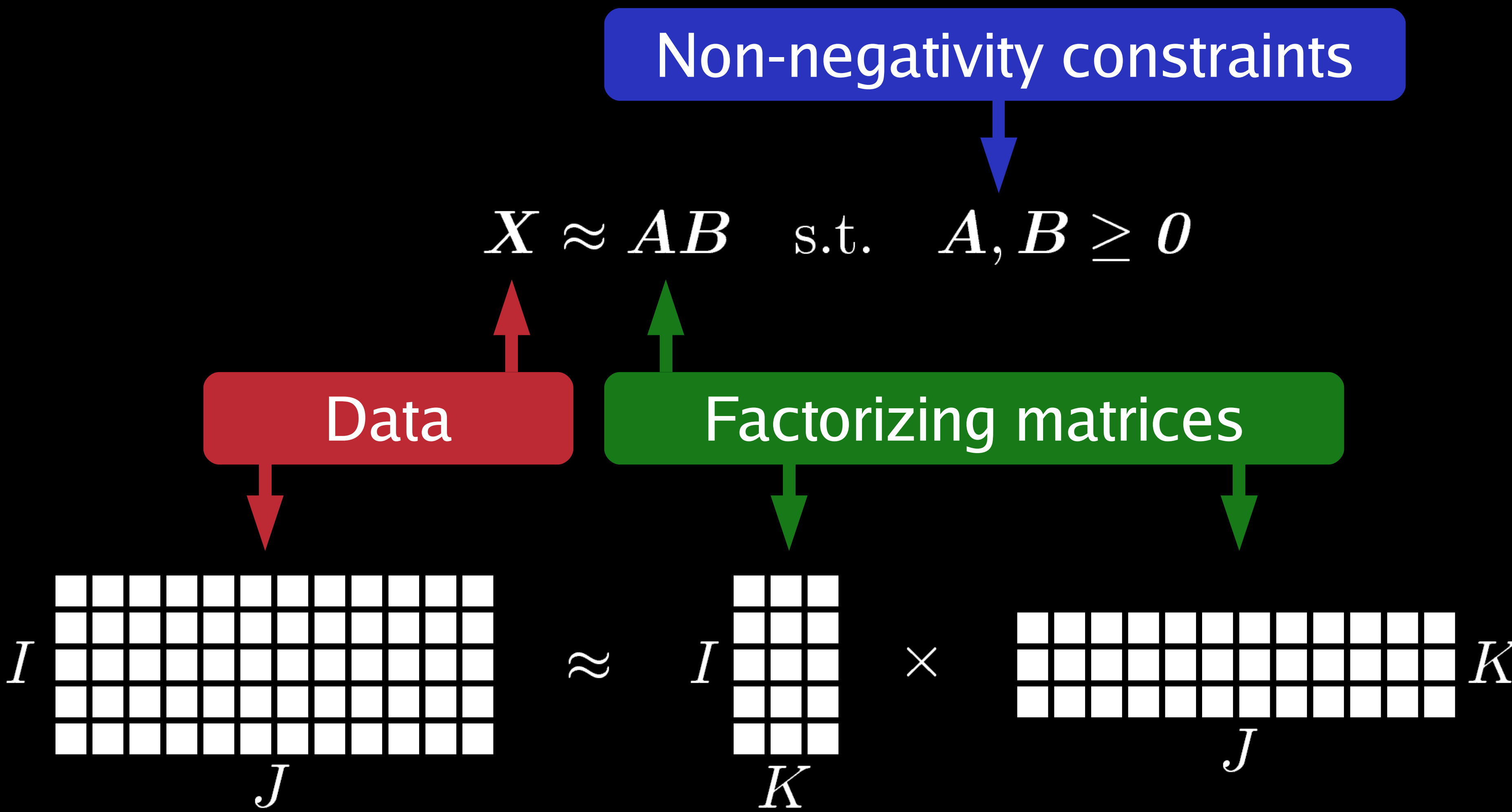
Posterior — Likelihood — Prior

$$p(\boldsymbol{s}_1, \ldots, \boldsymbol{s}_K | \boldsymbol{x}) \propto p(\boldsymbol{x} | \boldsymbol{s}_1, \ldots, \boldsymbol{s}_K) \prod_{k=1}^{K} p(\boldsymbol{s}_k)$$

Estimate sources:

- Maximum a posteriori, posterior mean, marginal MAP, etc.
- Solve optimization or integration problem

Non-negative matrix factorization

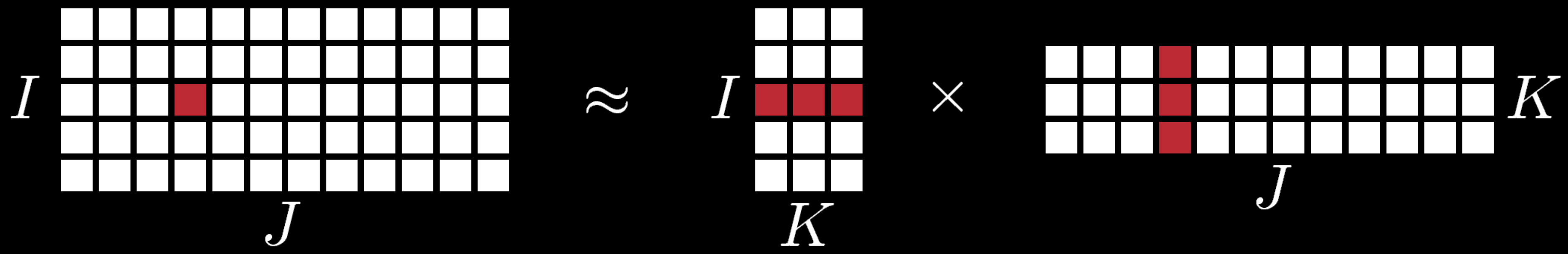


Non-negative matrix factorization

Non-negative bilinear model

$$X_{i,j} \approx \sum_{k=1}^{K} A_{i,k} B_{k,j}$$

Sum of products of non-negative variables



Why non-negativity?

Many signals are naturally non-negative

Pixel intensities

Amplitude spectra

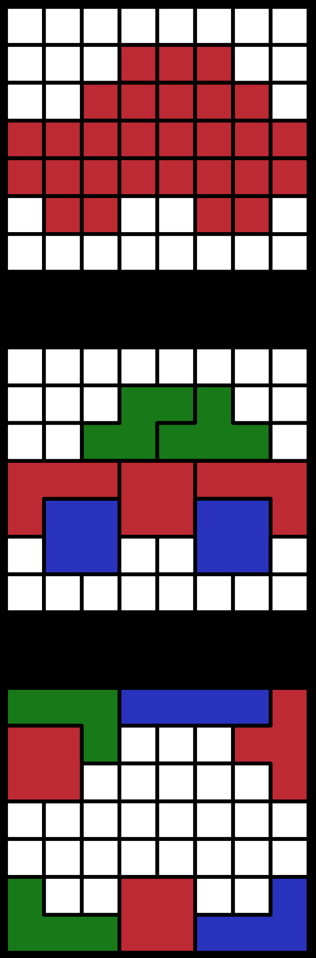
Occurrence counts

Discrete probabilities

Additive combination of features

No cancellations

Build the whole as a sum of parts



Basic computation

Constrained minimization problem

Divergence measure

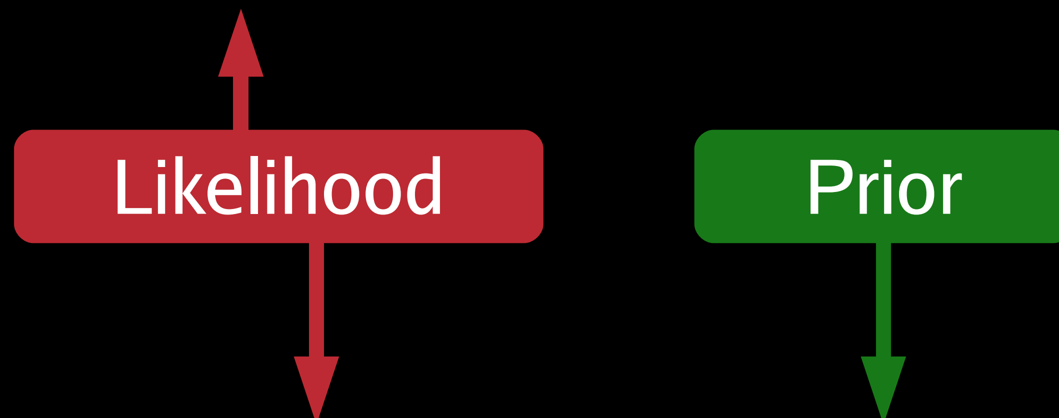
$$\{\boldsymbol{A}, \boldsymbol{B}\} = \arg \min_{\boldsymbol{A},\boldsymbol{B} \geq \boldsymbol{0}} \mathcal{D}(\boldsymbol{X}; \boldsymbol{A}, \boldsymbol{B})$$

Constraints

Probabilistic formulation

Maximum likelihood

$$\mathcal{D}_{\mathrm{ML}} = -\log p(\boldsymbol{X}|\boldsymbol{A},\boldsymbol{B})$$



Maximum a posteriori

$$\mathcal{D}_{MAP} = -\log p(\boldsymbol{X}|\boldsymbol{A},\boldsymbol{B}) - \log p(\boldsymbol{A},\boldsymbol{B})$$

Divergence measures (Likelihoods)

Squared error (Lee and Seung, 1999)

Kullback-Leibler divergence (Lee and Seung, 1999)

Bregman's divergence (Dhillon and Sra, 2005)

Kompass' divergence (Kompass, 2007)

Csiszár's divergence (Cichocki et al., 2006)

Amari's alpha divergence (Cichocki et al., 2006)

Weighted versions of the above (Guillamet et al., 2001)

Distribution of the factors (Priors)

Sparsity (Hoyer, 2002)

Orthogonality (Ding et al., 2005)

Discriminative (Wang et al., 2004)

Smoothness (Virtanen, 2003)

Gaussian process (Schmidt and Laurberg, 2008)

Transformation invariance (Wersing et al., 2003)

Convolutive (Virtanen, 2004; Smaragdis, 2004)

2-D convolutive (Schmidt and Mørup, 2006)

Optimization strategies

DTU

Direct optimization

$$\{A, B\} \leftarrow \arg \min_{\boldsymbol{A},\boldsymbol{B} \geq \boldsymbol{0}} \mathcal{D}$$

Alternating optimization

repeat

$A \leftarrow \arg\min \mathcal{D}$ s.t. $A \geq \boldsymbol{0}$

$B \leftarrow \arg\min \mathcal{D}$ s.t. $B \geq \boldsymbol{0}$

until convergence

Alternating descent

repeat

$A \leftarrow A^*$ s.t. $\mathcal{D}^* \leq \mathcal{D}, A \geq \boldsymbol{0}$

$B \leftarrow B^*$ s.t. $\mathcal{D}^* \leq \mathcal{D}, B \geq \boldsymbol{0}$

until convergence

Non-convex

Can have convex subproblems

Algorithms

Projected least squares (Paatero, 1997)

Multiplicative updates (Lee and Seung, 1999)

Projected gradient descent (Lin, 2007)

Logarithmic barrier method (Lu and Wu, 2005)

Active set (Berry et al., 2006)

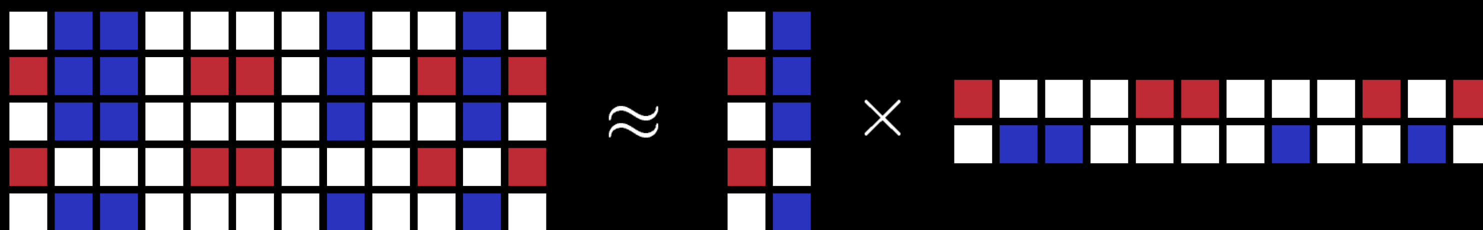
Quasi Newton (Kim et al., 2007)

Reparametrization (Cichocki et al., 2006)

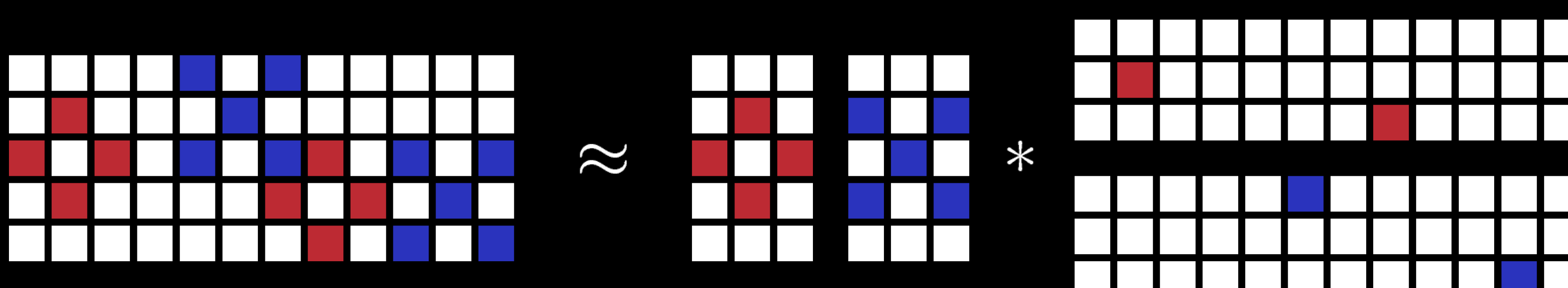
SOCP (Heiler and Schnörr, 2006)

NMF 2-D deconvolution

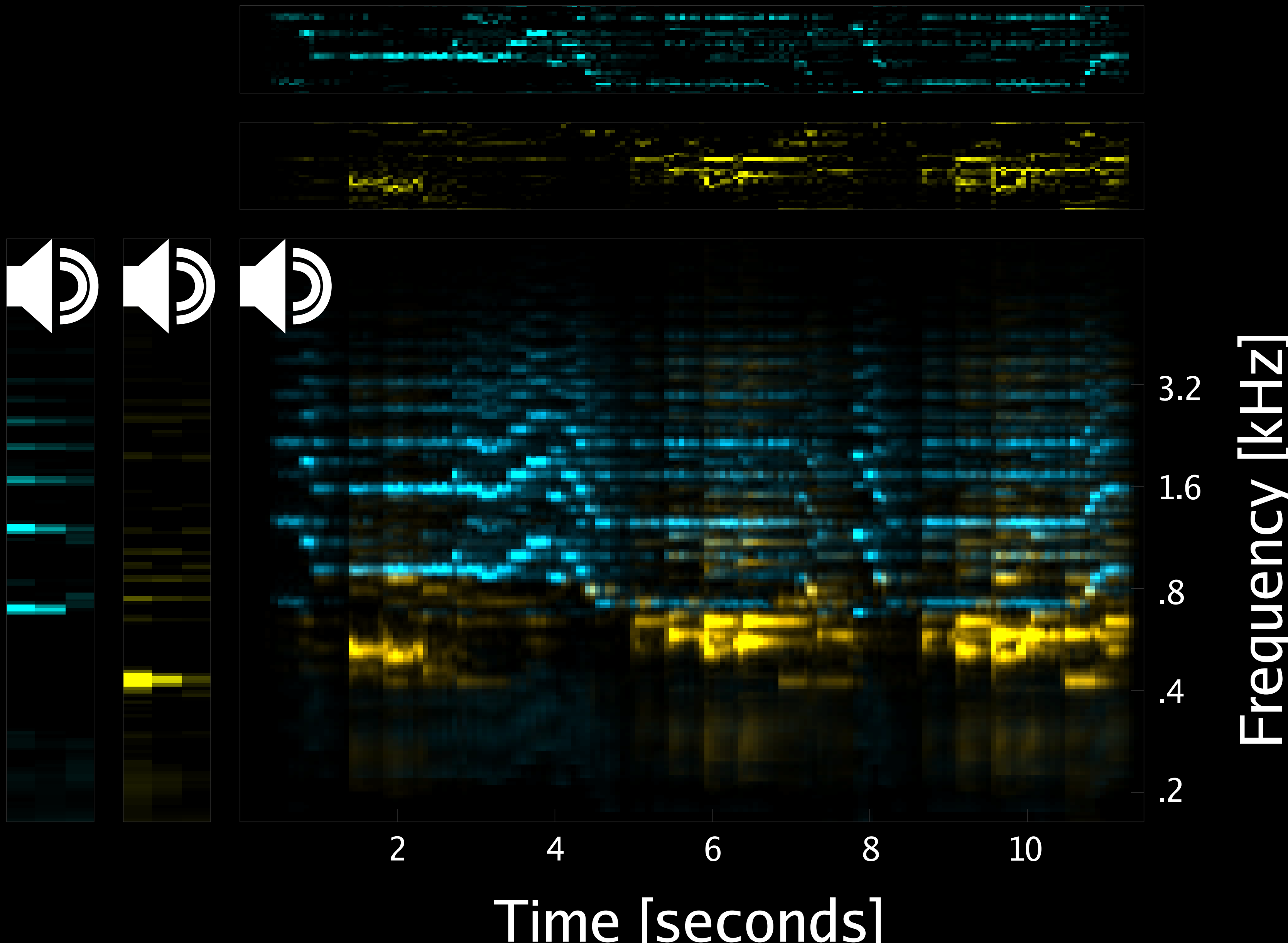
NMF $$X_{i,j} \approx \sum_{k=1}^{K} A_{i,k} B_{k,j}$$



NMF 2-D $$X_{i,j} \approx \sum_{t=1}^{T} \sum_{f=1}^{F} \sum_{k=1}^{K} A^{(t)}_{i-f,k} B^{(f)}_{k,j-t}$$



Example: Flute and guitar



Speech separation using sparse NMF

Training data

Learn basis for each source $\boldsymbol{X}_k \approx \boldsymbol{A}_k \boldsymbol{B}_k$

Mixture

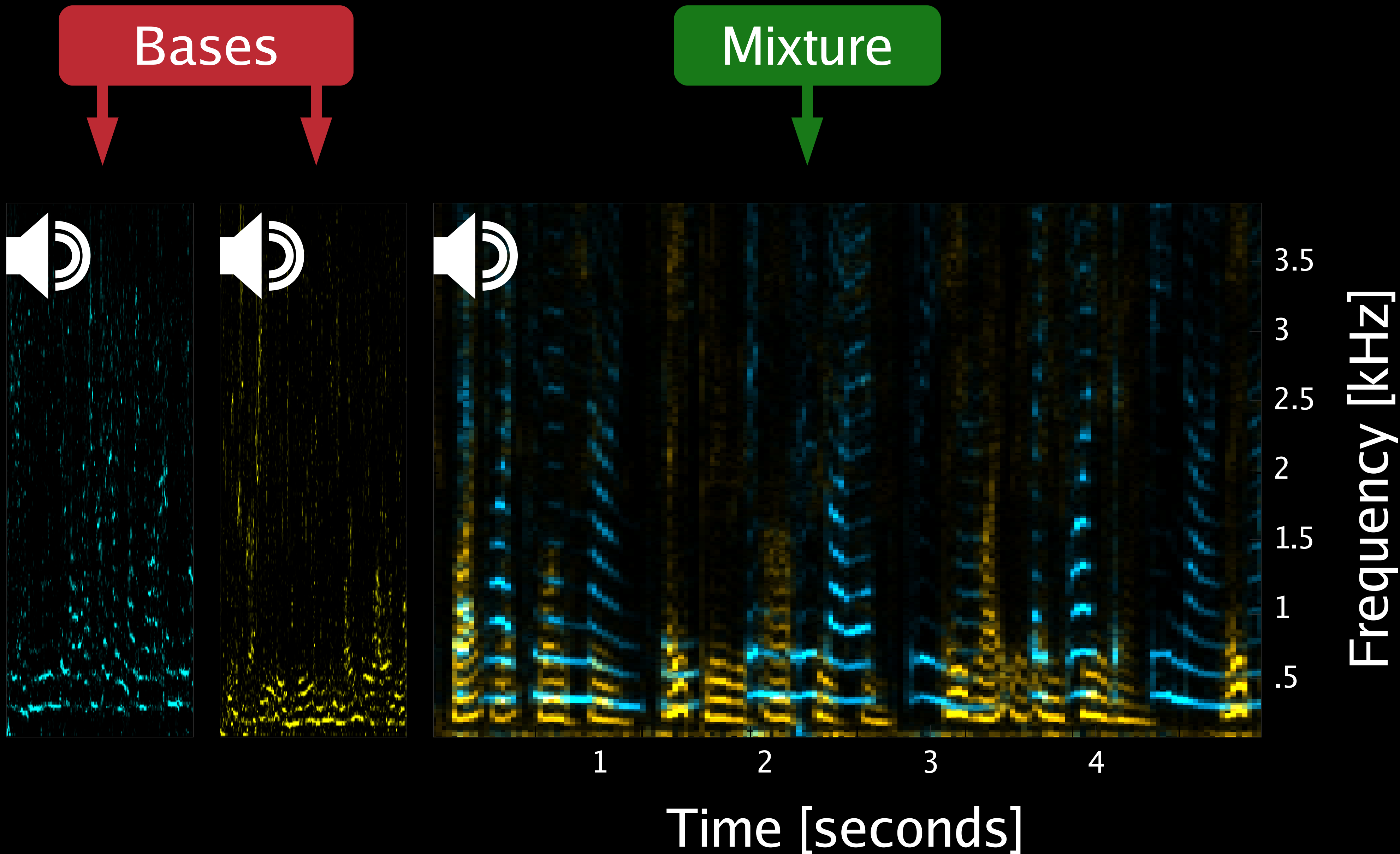
Basis

Decompose mixture $\boldsymbol{X} \approx [\boldsymbol{A}_1, \ldots, \boldsymbol{A}_K] \begin{bmatrix} \bar{\boldsymbol{B}}_1 \\ \vdots \\ \bar{\boldsymbol{B}}_K \end{bmatrix}$

Reconstruct each source $\hat{\boldsymbol{X}}_k \approx \boldsymbol{A}_k \bar{\boldsymbol{B}}_k$

Source estimate

Example: Two speakers



NMF with Gaussian process priors

Data

$$\boldsymbol{X} \approx \boldsymbol{A}\boldsymbol{B}$$

Gaussian processes

$$A = f_A(\boldsymbol{G}_A) \quad \boldsymbol{B} = f_B(\boldsymbol{G}_B)$$

Link functions

GP: General distribution over functions

Link function: Non-linear map to non-negative reals

Gaussian process

Distribution over functions

Characterized by

Mean function

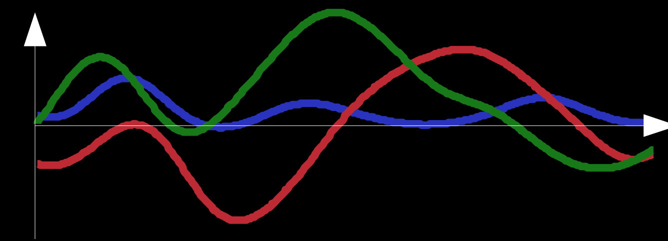
Covariance function

Example

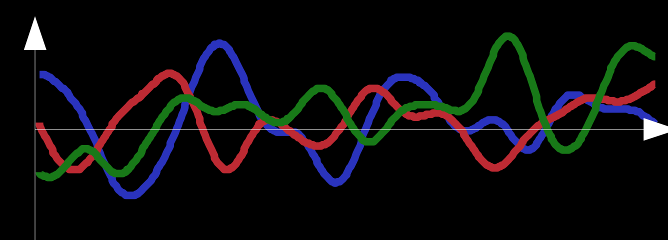
$$m(x) = 0$$

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

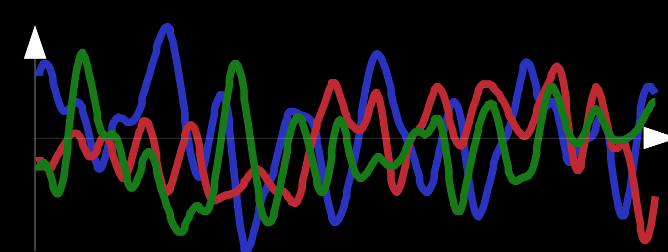
$\sigma = 1$



$\sigma = .5$



$\sigma = .1$



DTU

Change of variable

Cholesky decomposition

$$A = f_A(C_A^\top \boldsymbol{\alpha}) \qquad B = f_B(C_B^\top \boldsymbol{\beta})$$

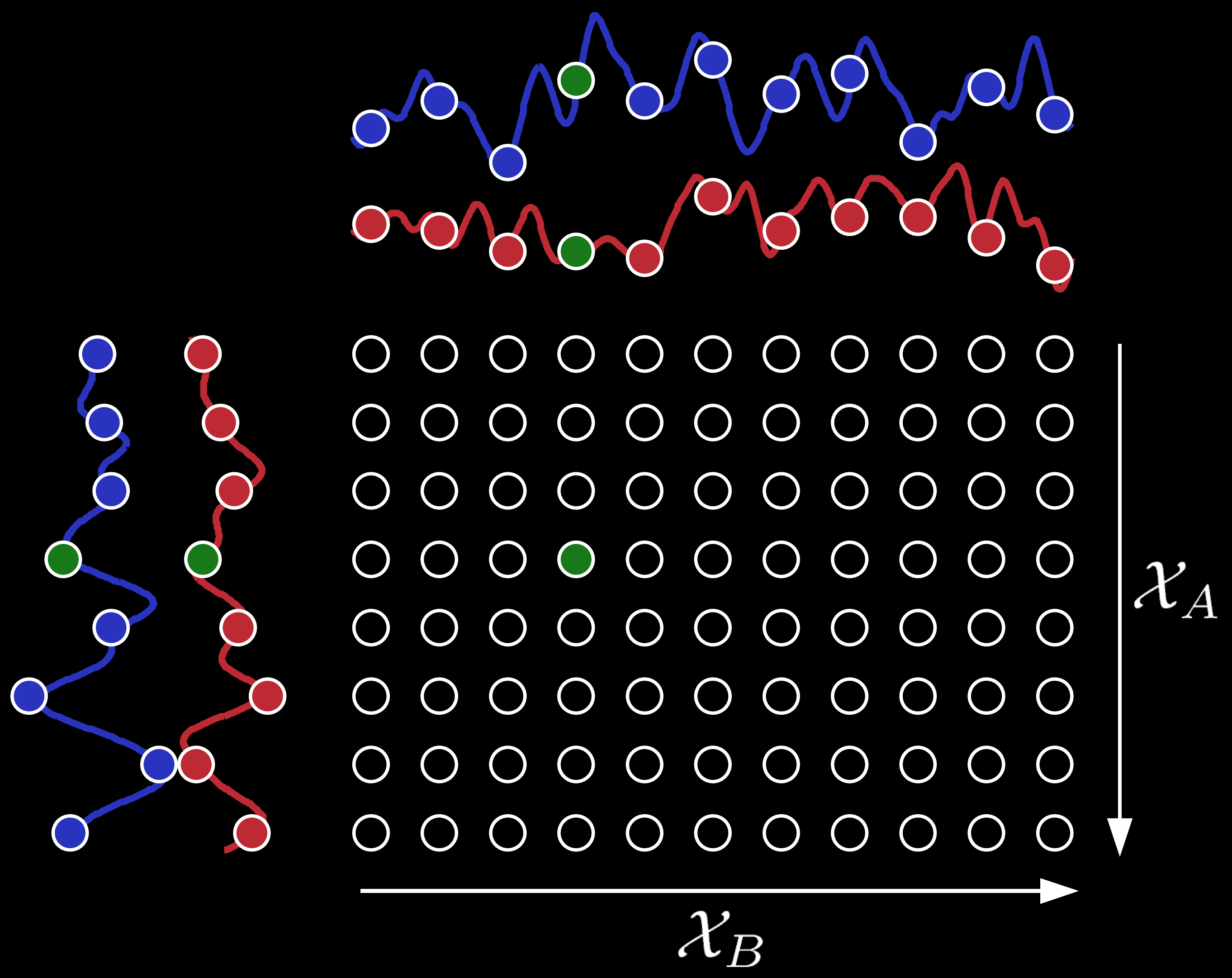
I.I.D. standard normal

Same model, different parametrization

Parameters are a priori uncorrelated

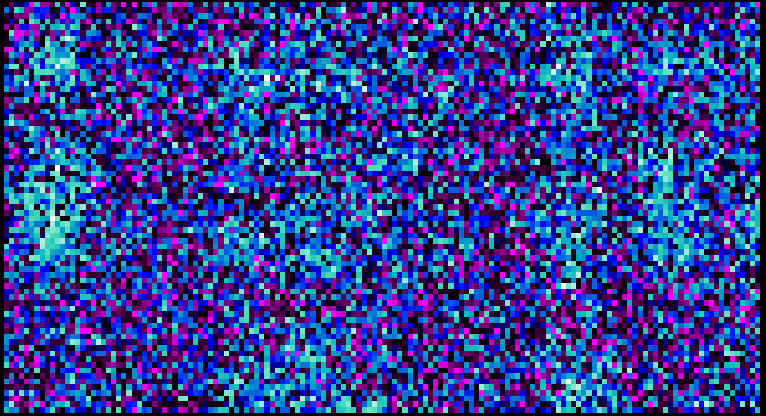
Empirically better optimization properties

Illustration of NMF with GP priors

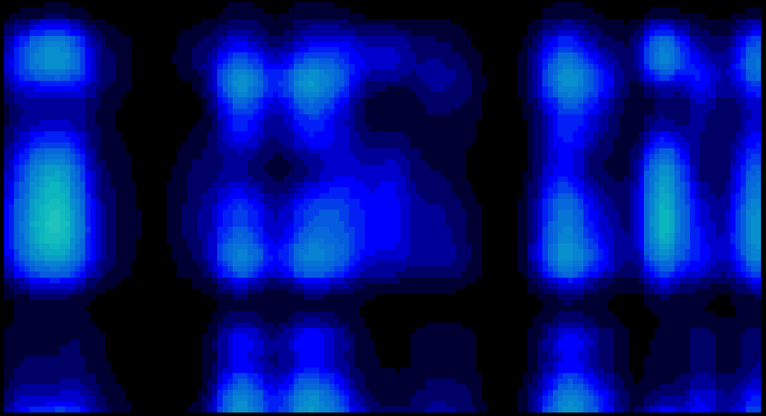


Example: Toy problem

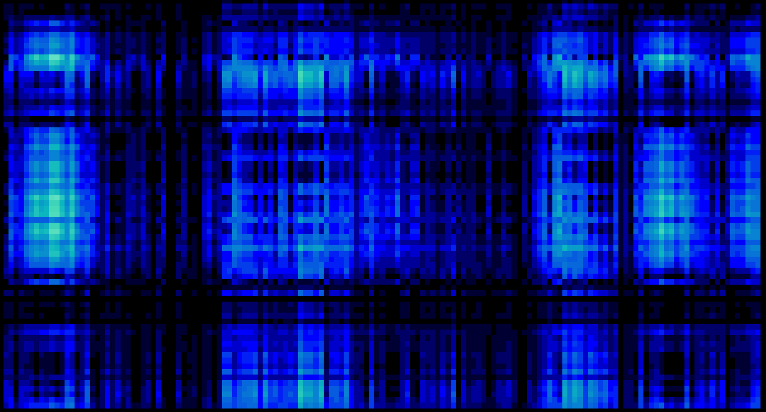
Noisy data



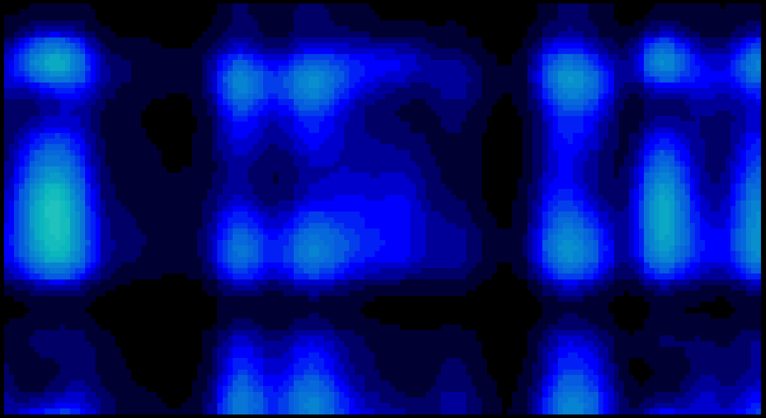
Underlying data



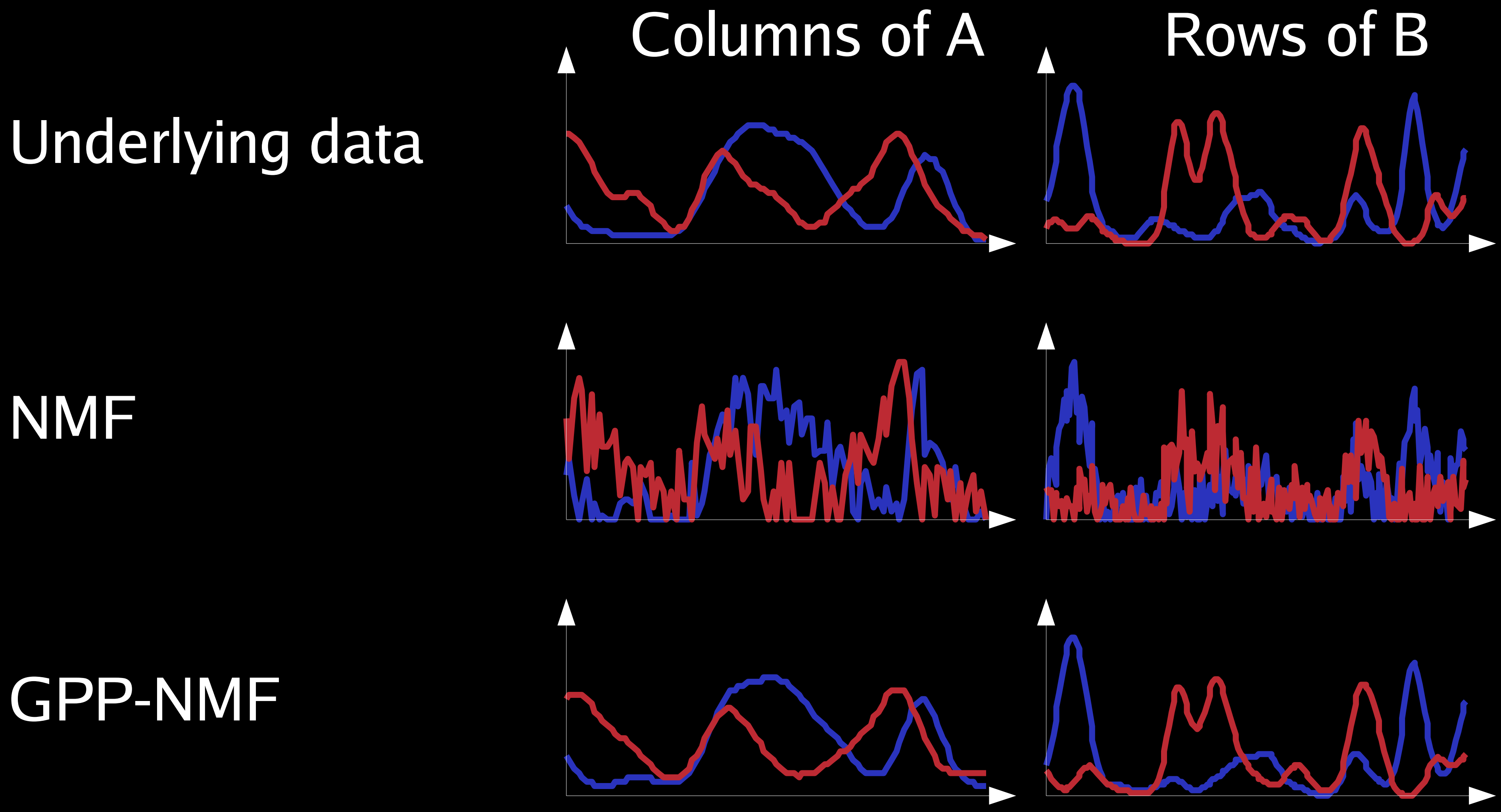
NMF



GPP-NMF



Example: Toy problem



Example: Chemical shift brain imaging

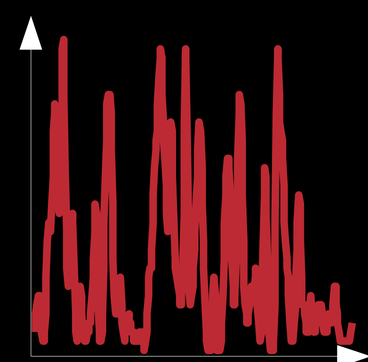
Data: 369 point spectra @ 512 points ($8 \times 8 \times 8$ grid) in human head

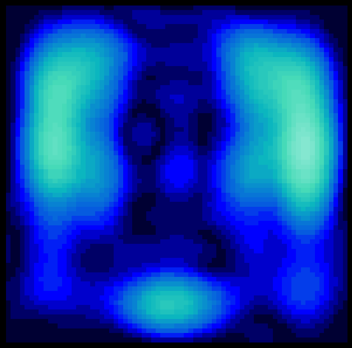
Task: Distinguish between brain and muscle tissue

Spectra prior: Smooth exponential distribution

Activation prior: 3-D smooth, exponential distribution, left-to-right symmetric

Random draw from prior

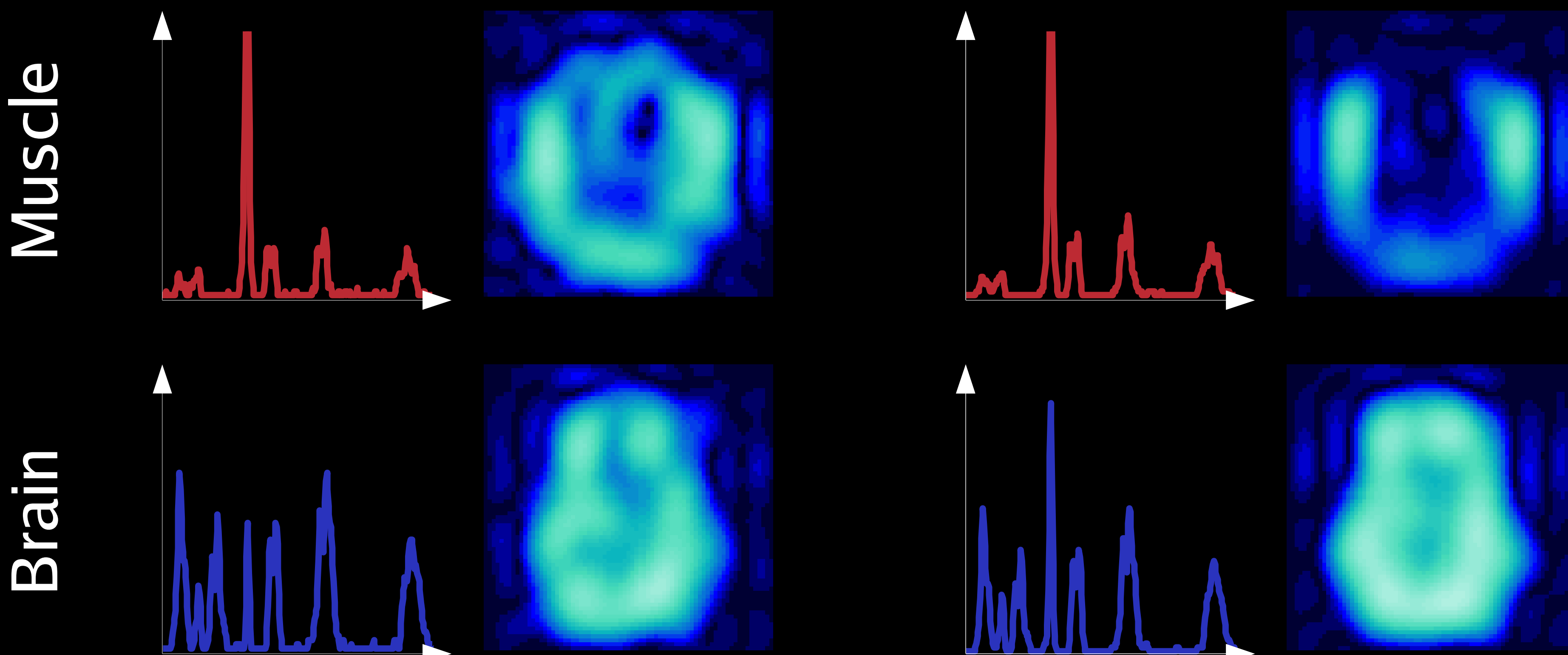




Example: Chemical shift brain imaging

Non-negative matrix factorization

NMF with Gaussian process priors



Conclusions

Single-channel source separation is a difficult problem that occurs in many areas

Model based source separation provides a principled approach to solving the problem

NMF and its extensions are useful models for single-channel source separation

Gaussian processes can be used as a general framework for incorporating prior information in NMF