



# Bayesian non-negative matrix factorization

... or “How I learned to love the Gibbs sampler.”

Mikkel N. Schmidt

Technical University of Denmark



## Agenda

- **Bayesian inference recap**
- **Gibbs sampling**
- **Non-negative matrix factorization (NMF)**
- **Bayesian NMF using Gibbs sampling**



# Bayesian inference

*... a quick recap from yesterday*



# Bayesian inference

## 1. Formulate a generative model

- Likelihood  $p(\mathbf{X}|\boldsymbol{\theta})$
- Prior  $p(\boldsymbol{\theta})$

## 2. Observe data $\mathbf{X}$

## 3. Update your beliefs

$$\text{Posterior} \rightarrow p(\boldsymbol{\theta}|\mathbf{X}) = \frac{p(\mathbf{X}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{X})}$$

Likelihood      Prior  
Evidence



## Example: Bayesian prediction (1)

- **Problem: Given observed data  $\mathbf{X}$**   
**predict new data  $\tilde{\mathbf{X}}$**
- **Maximum a posteriori**

$$\hat{\theta} = \arg \max_{\theta} p(\theta | \mathbf{X}) \leftarrow \text{Posterior}$$
$$p(\tilde{\mathbf{X}} | \hat{\theta})$$

- **Predictive distribution**

$$p(\tilde{\mathbf{X}} | \mathbf{X}) = \int p(\tilde{\mathbf{X}} | \theta) p(\theta | \mathbf{X}) d\theta$$

Posterior



## Example: Bayesian prediction (2)

### ■ Predictive distribution

$$p(\tilde{\mathbf{X}}|\mathbf{X}) = \int p(\tilde{\mathbf{X}}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{X})d\boldsymbol{\theta}$$

### ■ Posterior sampling

$$p(\tilde{\mathbf{X}}|\mathbf{X}) \approx \sum_i p(\tilde{\mathbf{X}}|\boldsymbol{\theta}^{(i)})$$
$$\boldsymbol{\theta}^{(i)} \sim p(\boldsymbol{\theta}|\mathbf{X})$$

... but how do we draw samples from the posterior?



# Gibbs sampling

... draw samples from the posterior



## Gibbs sampling (1)

- **Problem: Draw samples from the posterior**

$$\{\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \dots, \boldsymbol{\theta}^{(I)}\} \quad \boldsymbol{\theta}^{(i)} \sim p(\boldsymbol{\theta} | \mathbf{X})$$

- **Partition parameters into components**

$$\boldsymbol{\theta} = [\theta_1, \theta_2, \dots, \theta_N]^\top$$

- **Sample the n'th component from conditional distribution given data and all other parameters**

$$\theta_n^{(i)} \sim p(\theta_n | \{\theta_{\setminus n}^{(i-1)}\}, \mathbf{X})$$



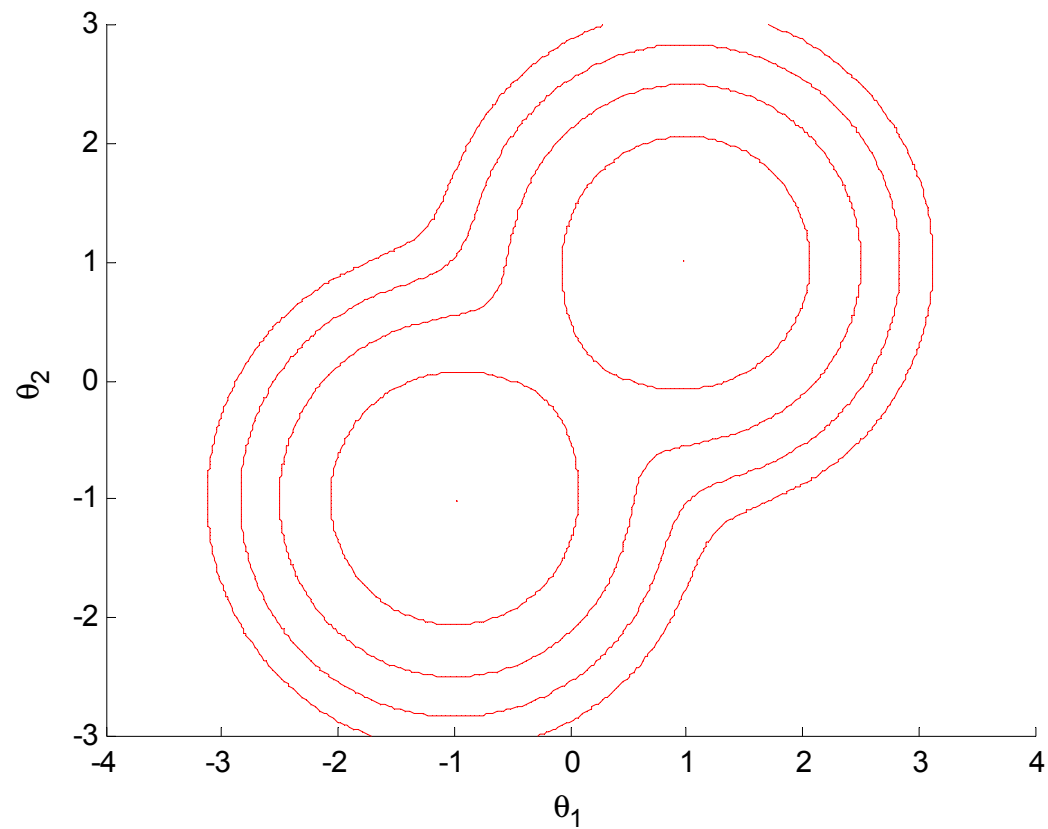
## Gibbs sampling (2)

$$\theta_n^{(i)} \sim p(\theta_n | \{\theta_{\setminus n}^{(i-1)}\}, \mathbf{X})$$

- Sample from the (univariate) distribution of a single component,  $\theta_n$ , keeping all other components fixed
  - Converges to a sample from the full posterior
- ... but we must be able to sample from the conditionals

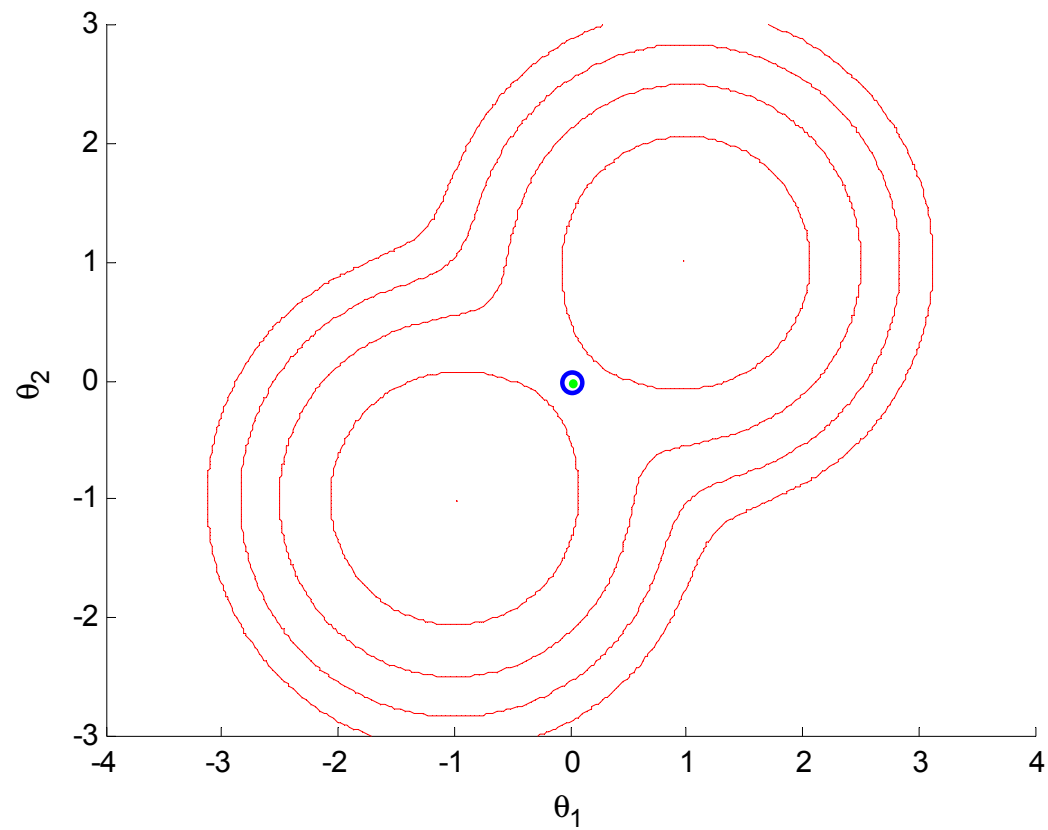


# Gibbs sampling example



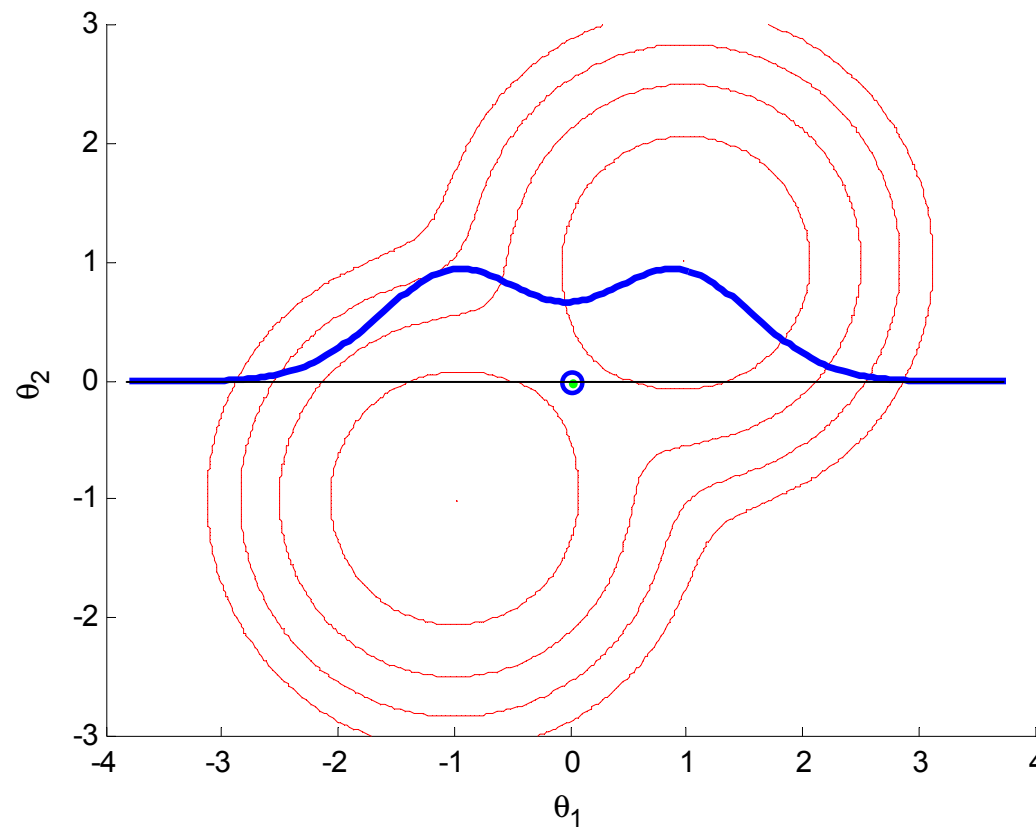


# Gibbs sampling example



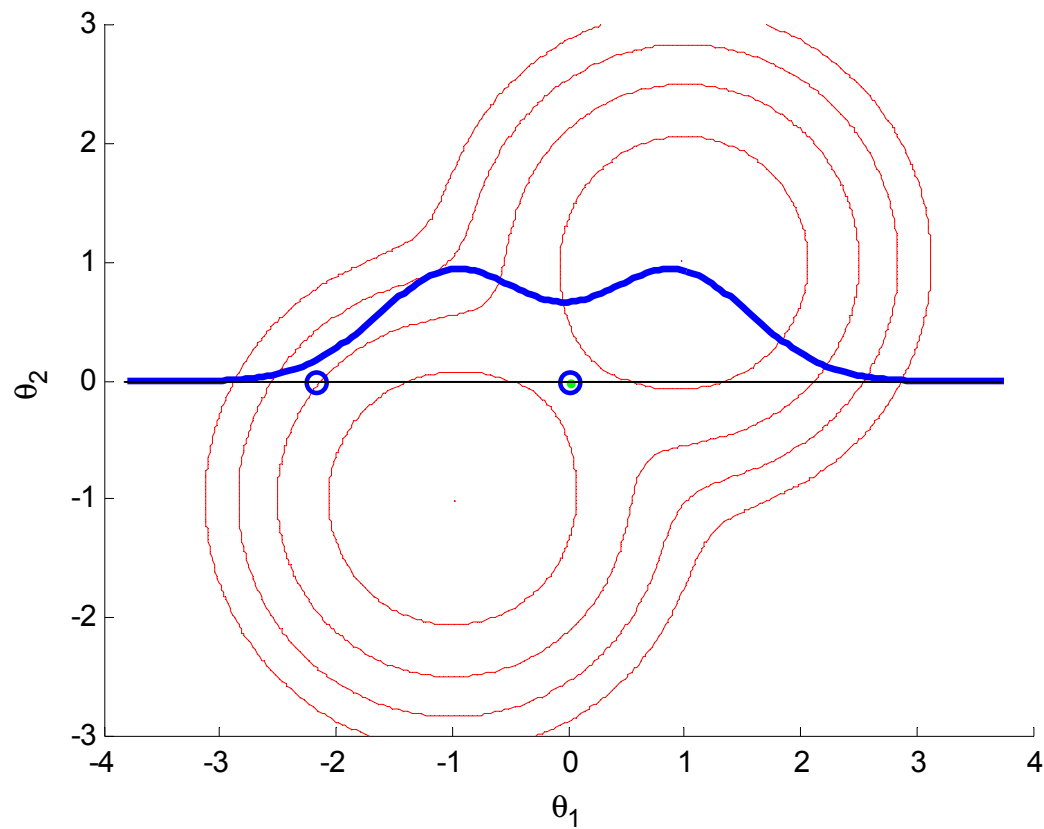


# Gibbs sampling example



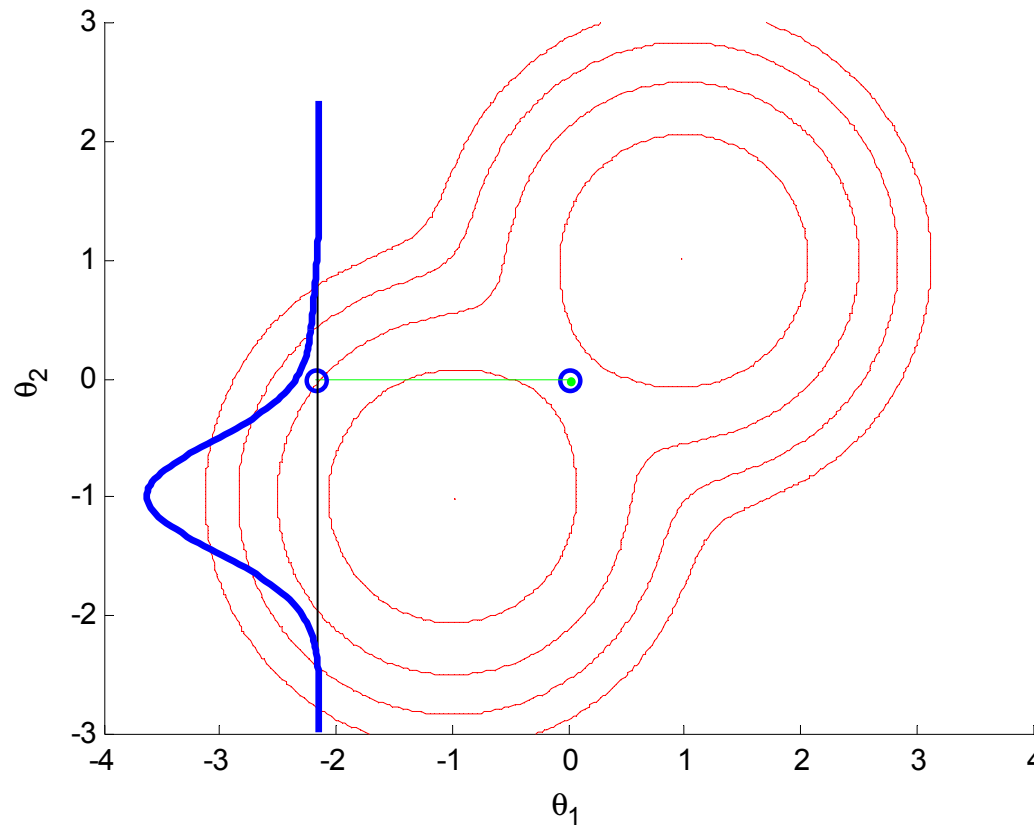


# Gibbs sampling example



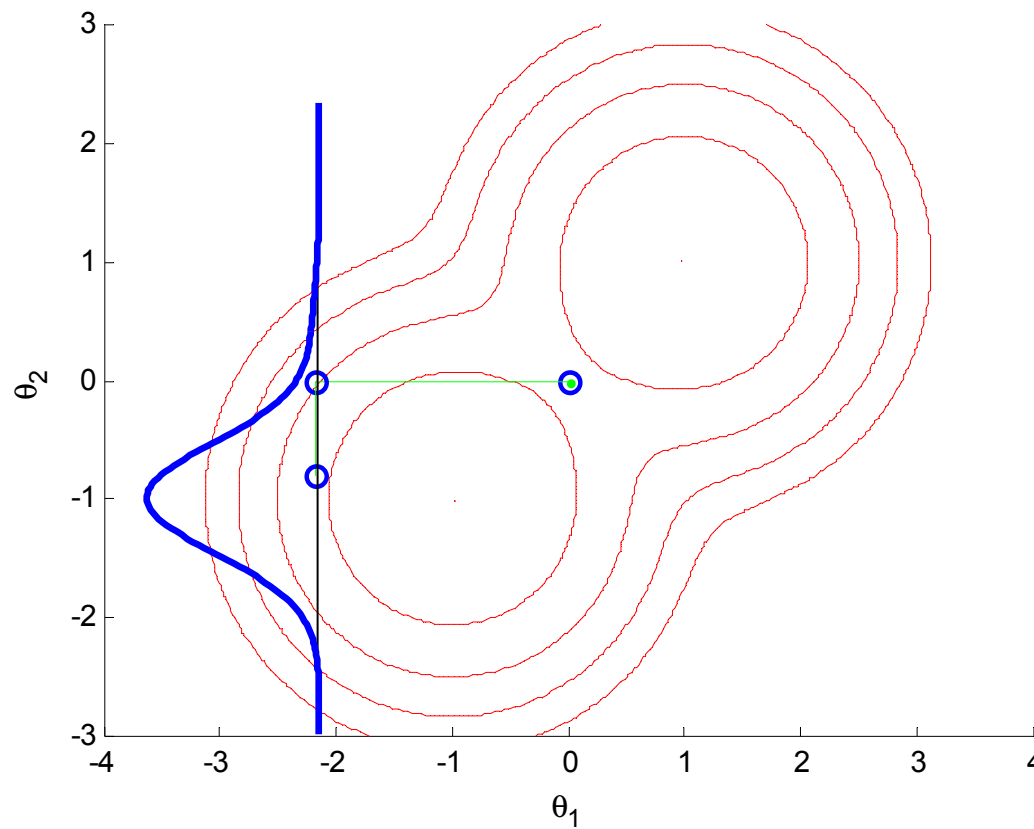


# Gibbs sampling example



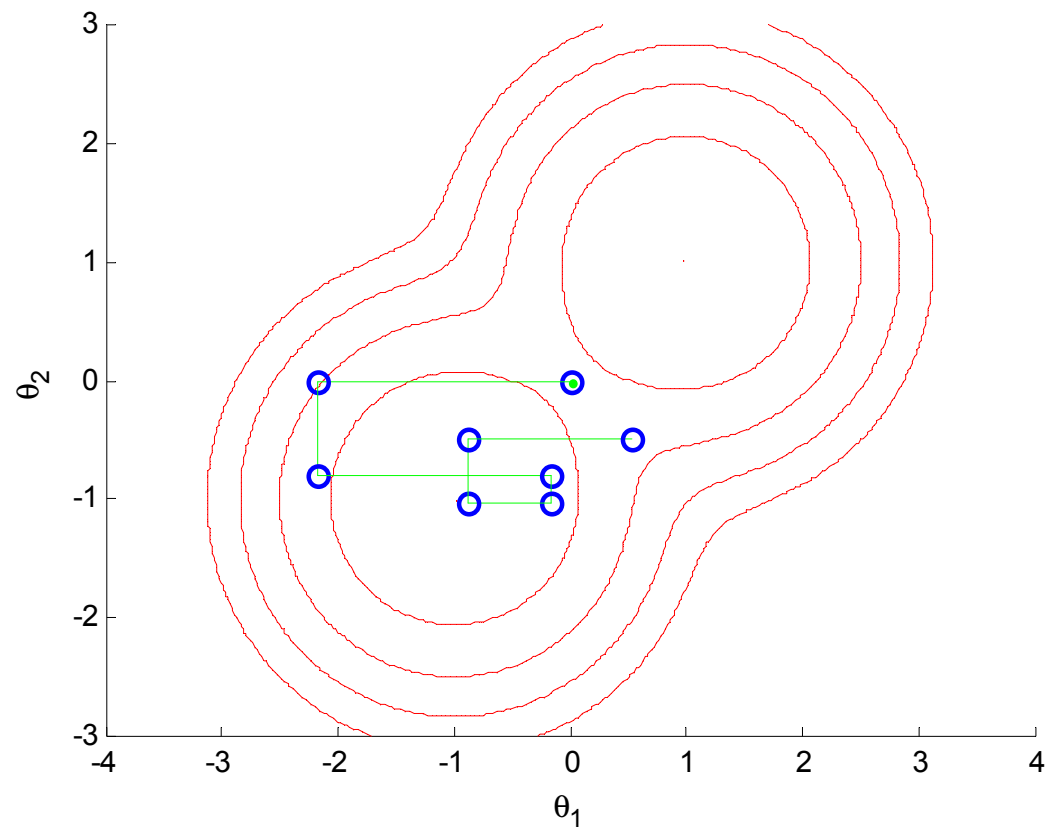


# Gibbs sampling example



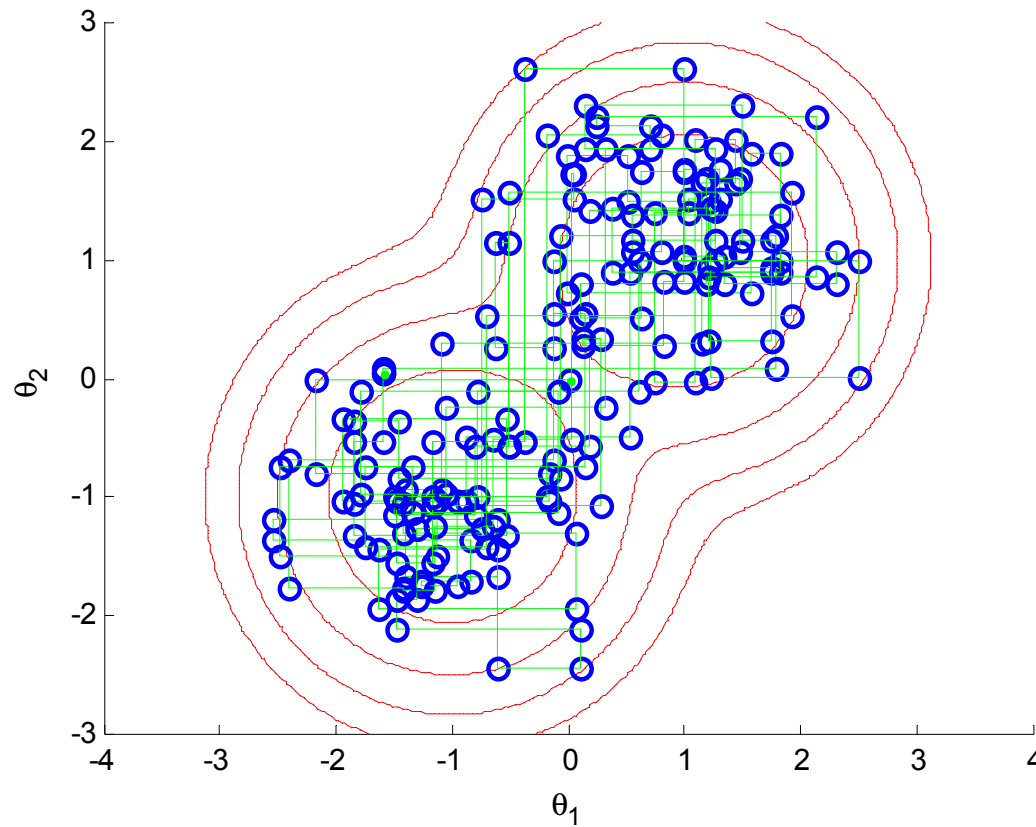


# Gibbs sampling example





# Gibbs sampling example





# Non-negative matrix factorization

... a parts-based bilinear decomposition



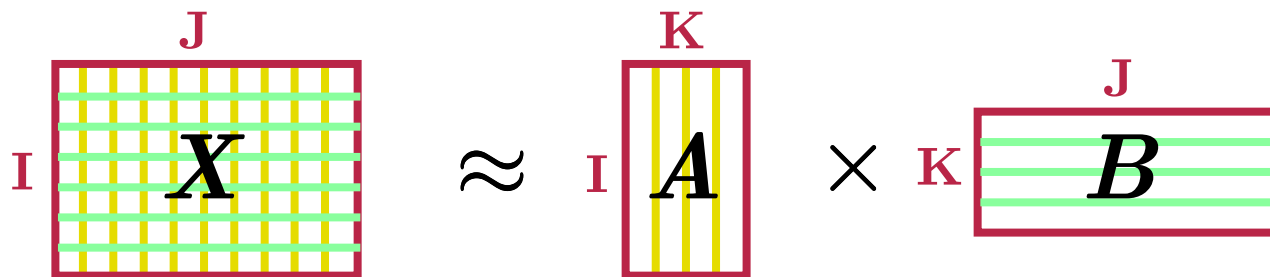
# Non-negative matrix factorization

## ■ Non-negative bilinear decomposition

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

## ■ In matrix notation

$$X \approx AB \quad \text{s.t. } A, B \geq 0$$





## 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**



## Computing the NMF

- Define cost function,  $\mathcal{D}(X; A, B)$ , that measures how well the data is approximated

- Minimize cost function

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

- Result: Matrices  $A$  and  $B$  that approximate data

... but this is not Bayesian



# Bayesian NMF

... sampling from the posterior distribution of  $A$  and  $B$



## Bayesian NMF

■ **Data:**  $X$ , **Parameters:**  $A$  and  $B$

■ **Likelihood:** Gaussian

$$p(\mathbf{X}|\mathbf{A}, \mathbf{B}) = \prod_{i,j} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x_{i,j} - [\mathbf{AB}]_{i,j})^2}{2\sigma^2}\right)$$

■ **Priors:** Exponential

$$p(\mathbf{A}) = \prod_{i,k} \alpha \cdot \exp(-\alpha \cdot a_{i,k}) \cdot u(a_{i,k})$$

$$p(\mathbf{B}) = \prod_{k,j} \beta \cdot \exp(-\beta \cdot b_{k,j}) \cdot u(b_{k,j})$$



## Bayesian NMF

■ **Posterior:** 
$$p(\mathbf{A}, \mathbf{B} | \mathbf{X}) = \frac{p(\mathbf{X} | \mathbf{A}, \mathbf{B}) p(\mathbf{A}) p(\mathbf{B})}{p(\mathbf{X})}$$

$$p(\mathbf{A}, \mathbf{B} | \mathbf{X}) = \frac{1}{p(\mathbf{X})} \cdot \prod_{i,j} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x_{i,j} - [\mathbf{AB}]_{i,j})^2}{2\sigma^2}\right) \\ \prod_{i,k} \alpha \cdot \exp(-\alpha \cdot a_{i,k}) \cdot u(a_{i,k}) \\ \prod_{k,j} \beta \cdot \exp(-\beta \cdot b_{k,j}) \cdot u(b_{k,j})$$

■ **Conditional distribution**

$$p(a_{i',k'} | \{a_{\setminus i', \setminus k'}\}, \mathbf{B}, \mathbf{X}) \propto p(\mathbf{A}, \mathbf{B} | \mathbf{X})$$



# Conditional distribution (1)

Variable Constants

$$p(a_{i',k'} | \{a_{\setminus i', \setminus k'}\}, \mathbf{B}, \mathbf{X})$$

$$\propto \frac{1}{p(\mathbf{X})} \cdot \prod_{k,j} \frac{1}{\sqrt{2\pi\alpha}} \exp\left(-\frac{(x_{i,j} - [\mathbf{AB}]_{i,j})^2}{2\sigma^2}\right)$$

$$\prod_{i,k} \alpha \cdot \exp(-\alpha \cdot a_{i,k}) \cdot u(a_{i,k})$$

$$\prod_{k,j} \beta \cdot \exp(-\beta \cdot b_{k,j}) \cdot u(b_{k,j})$$

$$\propto \prod_j \exp\left(-\frac{(x_{i',j} - [\mathbf{AB}]_{i',j})^2}{2\sigma^2}\right) \exp(-\alpha \cdot a_{i',k'}) \cdot u(a_{i',k'})$$



## Conditional distribution (2)

$$p(a_{i',k'} | \{a_{\setminus i', \setminus k'}\}, \mathbf{B}, \mathbf{X})$$

$$\propto \prod_j \exp\left(-\frac{(x_{i',j} - [\mathbf{AB}]_{i',j})^2}{2\sigma^2}\right) \exp(-\alpha \cdot a_{i',k'}) \cdot u(a_{i',k'})$$

$$= \exp\left(-\frac{\sum_j (x_{i',j} - \sum_{k \neq k'} a_{i',k} b_{k,j} - a_{i',k'} b_{k',j})^2 + 2\sigma^2 \alpha a_{i',k'}}{2\sigma^2}\right) \cdot u(a_{i',k'})$$

$$\propto \exp\left(-\frac{(a_{i',k'} - \mu_{a_{i',k'}})^2}{2\sigma_{a_{i',k'}}^2}\right) \cdot u(a_{i',k'})$$

Rectified Gaussian



## Sampling from a rectified Gaussian

### ■ Rectified Gaussian distribution

$$p(x) = \frac{1}{Z} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right) u(x)$$

### ■ The inverse cumulative distribution method:

Suppose  $y$  is uniform over  $(0,1)$ , then  $x$  and  $y$  are related by

$$x = P^{-1}(y) = \mu + \sqrt{2}\sigma \operatorname{erfc}^{-1}\left((1 - y)\operatorname{erfc}\left(-\frac{\mu}{\sqrt{2}\sigma}\right)\right)$$



# Putting it all together

## Algorithm

### Repeat

1. For each element in A, draw a sample
2. For each element in B, draw a sample
3. Save A and B

□

```

for  $m = 1$  to  $M$  do
     $C = BB^T$ 
     $D = XB^T$ 
    for  $n = 1$  to  $N$  do
         $A_{:,n} \leftarrow \mathcal{R}\left(\frac{D_{:,n} - A_{:, \setminus n} C_{\setminus n, n}}{C_{n, n}}, \frac{\sigma^2}{C_{n, n}}, \alpha_{:,n}\right)$ 
    end
     $E = A^T A$ 
     $F = A^T X$ 
    for  $n = 1$  to  $N$  do
         $B_{n,:} \leftarrow \mathcal{R}\left(\frac{F_{n,:} - E_{n, \setminus n} B_{\setminus n, :}}{E_{n, n}}, \frac{\sigma^2}{E_{n, n}}, \beta_{n,:}\right)$ 
    end
     $A^{(m)} \leftarrow A$ 
     $B^{(m)} \leftarrow B$ 
end
    
```

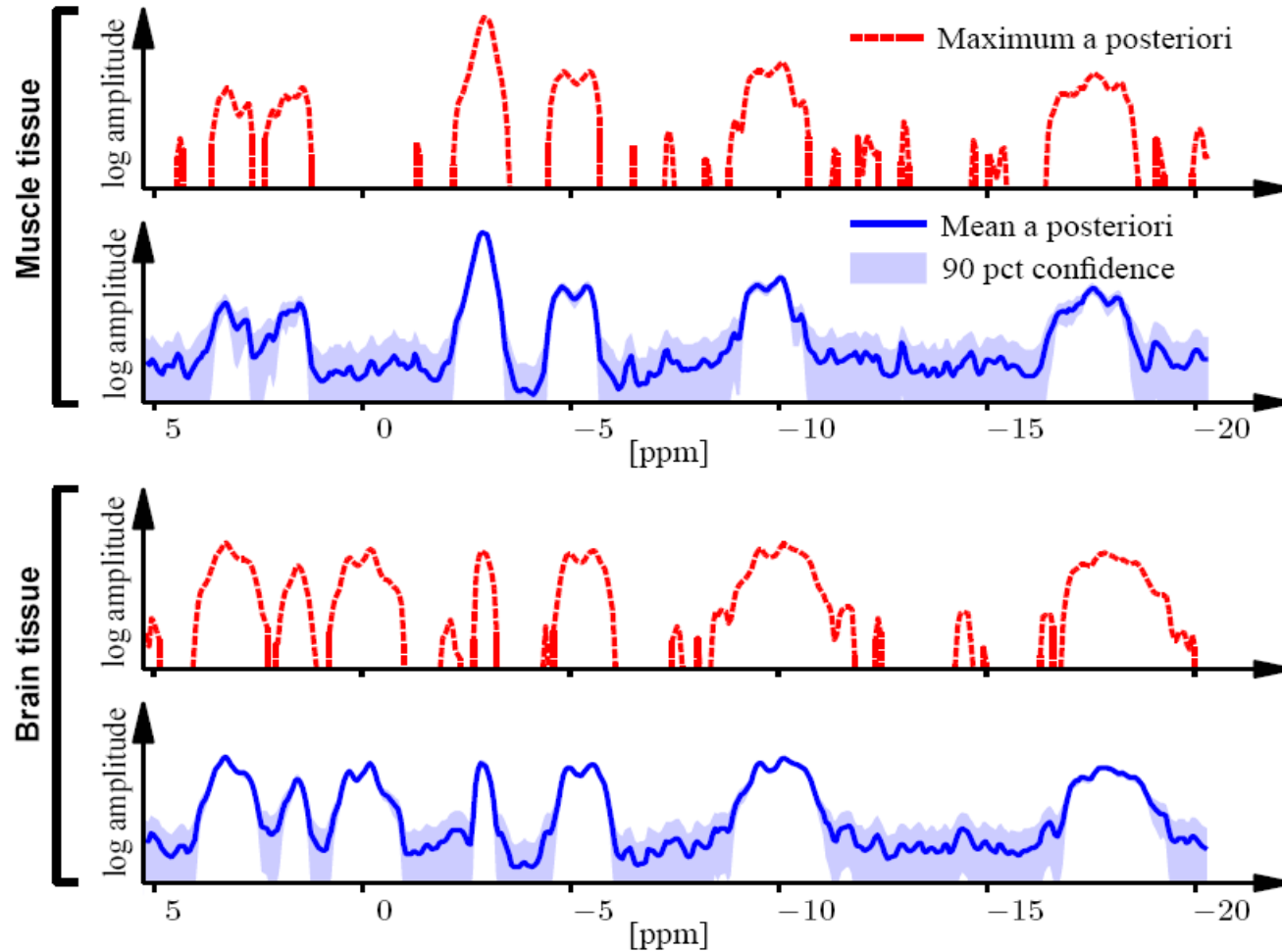


## Example: Chemical shift brain imaging

- **Dataset: 369-dimensional spectra measured at 512 positions in human head.**
- **Two components: brain- and muscle tissue.**
- **Sample: 40,000 points**



# Example: Chemical shift brain imaging





# Questions