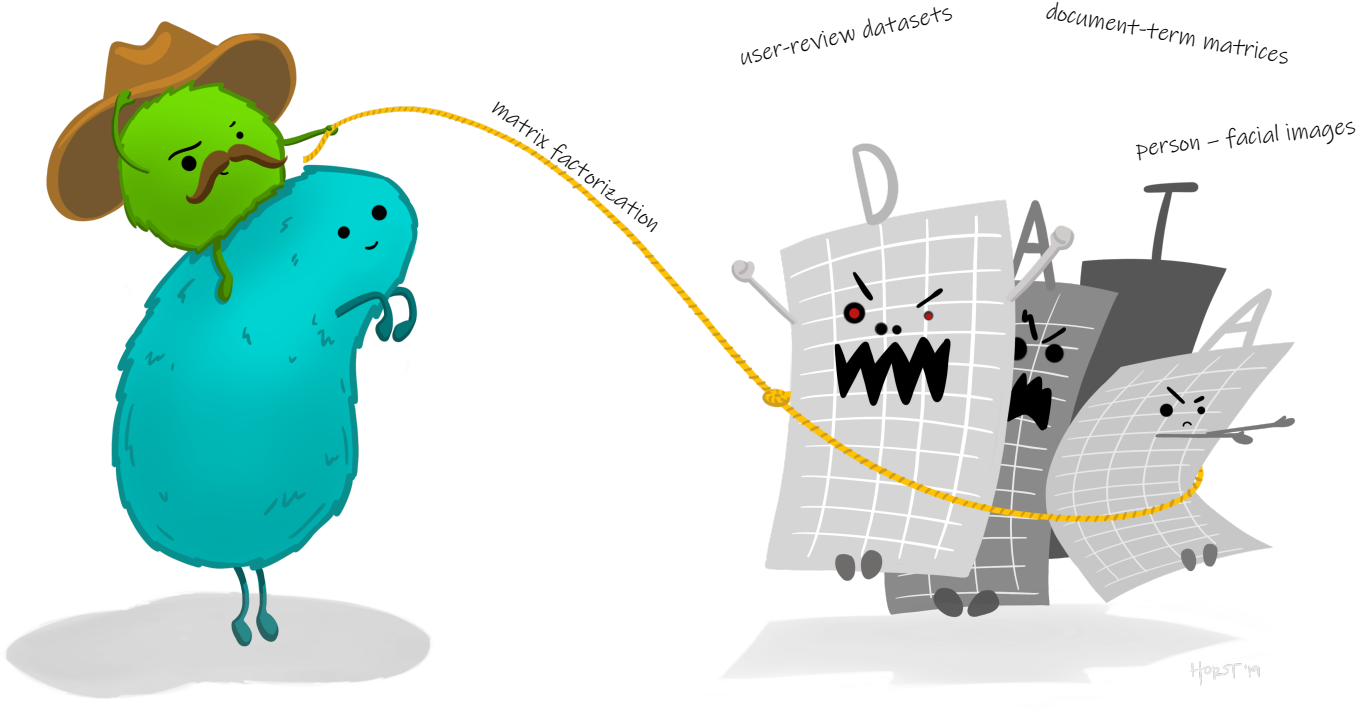
(Bayesian) non-negative matrix factorization

Greta Gašparac
Mentor: Assist. Prof. dr. Jana Faganeli Pucer

Feb 2021


matrix factorization
user-review datasets
document-term matrices
person – facial images
D
A
T
A
HORST'M

1. Non-negative Matrix Factorization

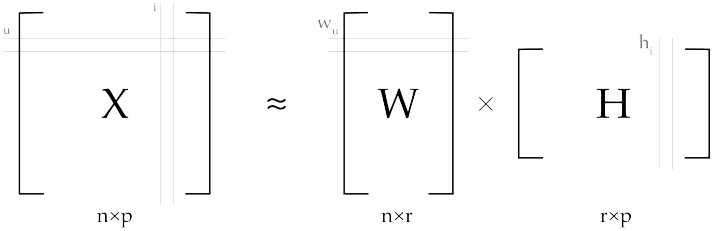
2. An Overview of Bayesian Statistics & Gibbs Sampling

3. Bayesian Non-negative Matrix Factorization

1. Non-negative Matrix Factorization

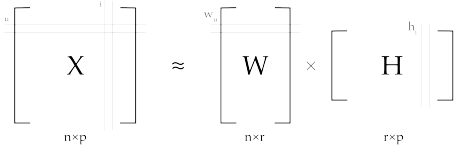
Non-negative matrix factorization (NMF)

- Belongs to the family of linear dimensionality reduction techniques based on matrix factorization.
- Popularised by Lee and Seung in 1999.
- Approximates our original matrix **X** with lower-rank matrices **W** and **H**, which are element-wise non-negative.
- The non-negativity constraint provides more intuitive factors.



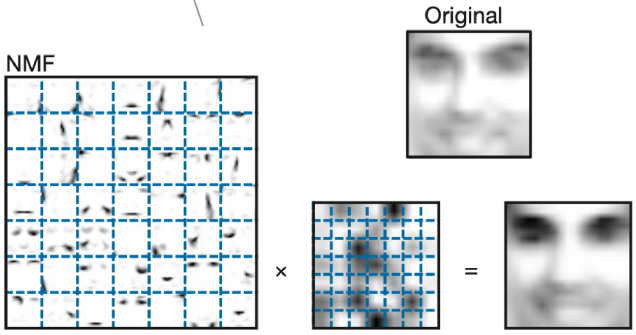
Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, *401*(6755), 788-791.

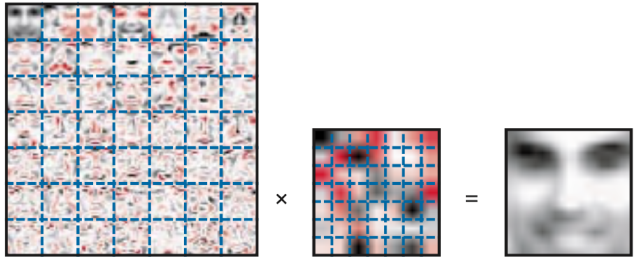
Example



localized features, parts-based representation

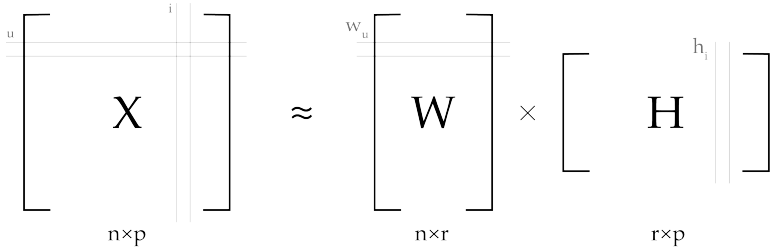
"eigenfaces", resemble distorted faces from the database





Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, *401*(6755), 788-791.

What is NMF?



What makes a good latent representation?

Pick a measure to asses the quality of the approximation. Common choice: Frobenius norm.

$$||\boldsymbol{X} - \boldsymbol{WH}||_F^2 = \sum_{i,j}(\boldsymbol{X} - \boldsymbol{WH})_{i,j}^2$$

Optimisation problem for a rank *r* factorization:

$$\min_{W \geq 0,\, H \geq 0} ||X - WH||_F^2 + \_\_\_\_\_$$

Problems

NMF is ill-posed.

- No guarantee there exists a single best decomposition.
- Results depend on the initialization.
- *Solutions:* regularization terms in the objective function, multiple reruns.

Choice of *r*.

- Trade-off.
- *Approaches*: trial & error, SVD decomposition, domain expert knowledge.

NMF is NP-hard.

- Non-convex optimization problem.

NMF algorithms

Initialization
- Randomly
- Clustering
- SVD
- Use columns of X

Stopping criteria
- Value of objective function
- Time
- Maximum number of iterations
- Combination

Algorithm 1: Two-Block Coordinate Descent

Initialize matrices $\boldsymbol{W}^{(0)} \geq 0$ and $\boldsymbol{H}^{(0)} \geq 0$.
for $i = 1, 2, \ldots,$ *stopping criteria* **do**
  $\boldsymbol{W}^{(i)} = \text{update}(\boldsymbol{X}, \boldsymbol{H}^{(i-1)}, \boldsymbol{W}^{(i-1)})$
  $\boldsymbol{H}^{(i)} = \text{update}(\boldsymbol{X}, \boldsymbol{H}^{(i-1)}, \boldsymbol{W}^{(i)})$

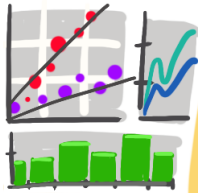
Multiplicative update rule proposed by Lee and Seung:

$$\boldsymbol{W} \leftarrow \boldsymbol{W} \circ \frac{\boldsymbol{X}\boldsymbol{H}^T}{\boldsymbol{W}\boldsymbol{H}\boldsymbol{H}^T},$$

$$\boldsymbol{H} \leftarrow \boldsymbol{H} \circ \frac{\boldsymbol{W}^T\boldsymbol{X}}{\boldsymbol{W}^T\boldsymbol{W}\boldsymbol{H}}$$

Applications

- Financial data mining

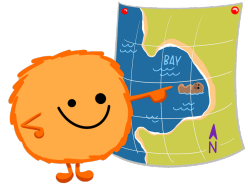


- Pattern detection in bioinformatics

- Recommender systems: The Netflix prize

- Natural language processing

- Environmetrics



de Fréin, R., Drakakis, K., Rickard, S., & Cichocki, A. (2008). Analysis of financial data using non-negative matrix factorization. In International Mathematical Forum (Vol. 3, No. 38, pp. 1853-1870). Journals of Hikari Ltd.
Brunet, J. P., Tamayo, P., Golub, T. R., & Mesirov, J. P. (2004). Metagenes and molecular pattern discovery using matrix factorization. Proceedings of the national academy of sciences, 101(12), 4164-4169.
The Netflix Prize: https://en.wikipedia.org/wiki/Netflix_Prize
Choo, J., Lee, C., Reddy, C. K., & Park, H. (2013). Utopian: User-driven topic modeling based on interactive nonnegative matrix factorization. IEEE transactions on visualization and computer graphics, 19(12), 1992-2001.
Paatero, P., & Tapper, U. (1994). Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. Environmetrics, 5(2), 111-126.

2. An Overview of Bayesian Statistics and Gibbs sampling

Bayesian Modeling

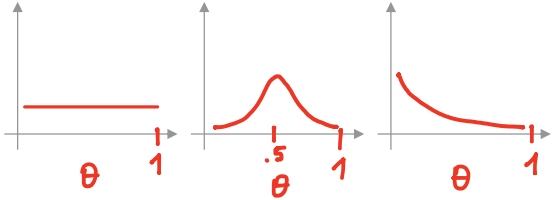
BAYES' THEOREM

$$\underbrace{p(\theta|y)}_{\text{POSTERIOR}} = \frac{\overbrace{p(y|\theta)}^{\text{LIKELIHOOD}} \cdot \overbrace{p(\theta)}^{\text{PRIOR}}}{\underbrace{p(y)}_{\int \ldots}} \propto p(y|\theta) \cdot p(\theta)$$

Coin toss example

Bayesian inference steps:

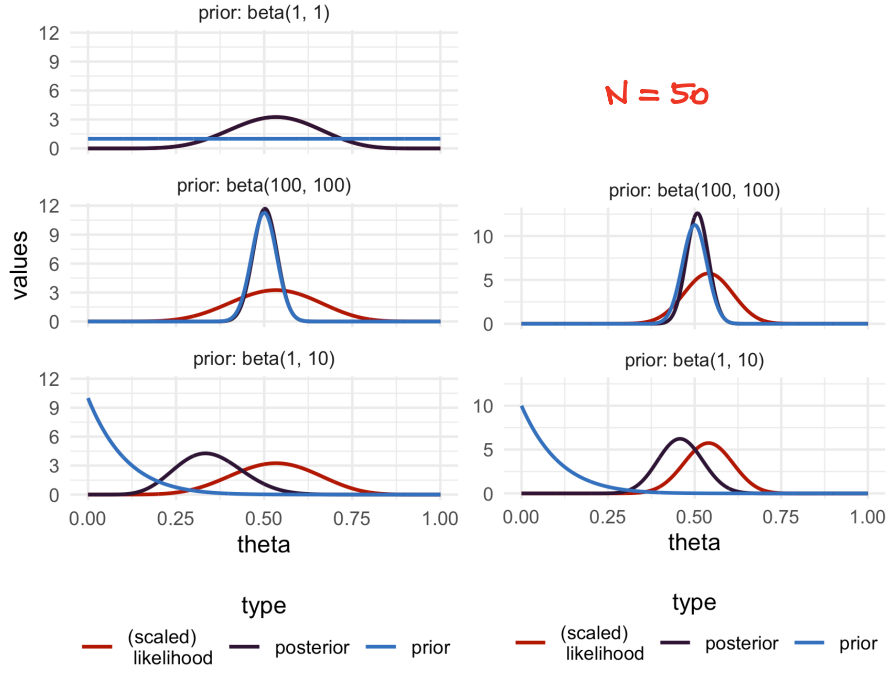
1. Specify prior.
2. Identify the observed data. HTTHHTHHHHTTHT
3. Construct a probabilistic model to represent the data.



$$p(\theta|y) = \frac{p(y|\theta) \cdot p(\theta)}{p(y)} \propto p(y|\theta) \cdot p(\theta)$$

N = 15

N = 50



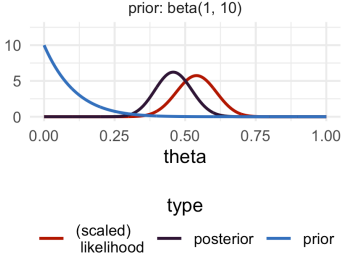
How do we get the posterior?

- Analytically

Problem: most of the models do not have nice conjugate posteriors!

$$p(\theta|y) = \frac{p(y|\theta) \cdot p(\theta)}{p(y)} \propto p(y|\theta) \cdot p(\theta)$$

- Using MCMC methods



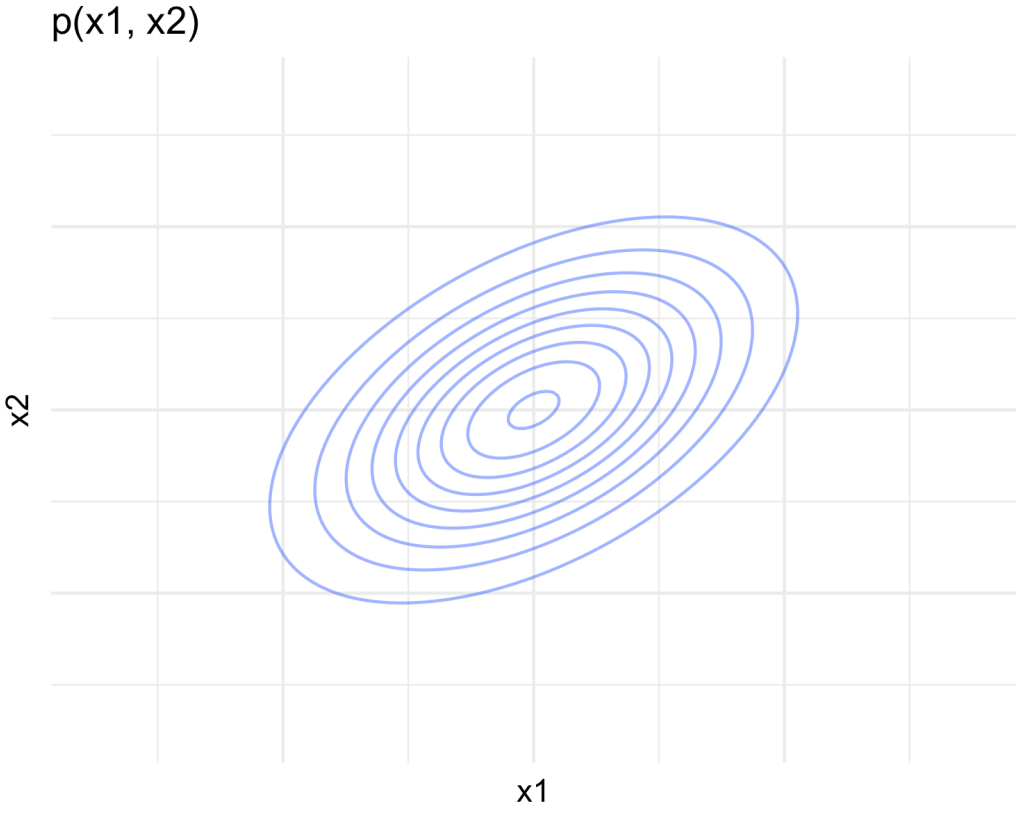
Gibbs Sampling

- Specific case of Metropolis-Hastings where the proposals are always accepted.
- Revolutionized the field of Bayesian statistics in the 80s and the 90s as it was the basis for the first generation of the Bayesian stats tools, such as BUGS, WinBUGS, JAGS.
- *Main idea behind:*
  a joint distribution $p(x_1, x_2, \ldots, x_n)$ can be charactarized by its complete set of full conditional distributions:
  $p(x_1 \mid x_2, x_3, \ldots, x_n)$, $p(x_2 \mid x_1, x_3, \ldots, x_n)$, …, $p(x_n \mid x_1, x_2, \ldots, x_{n-1})$.

Gibbs Sampling: bivariate normal example

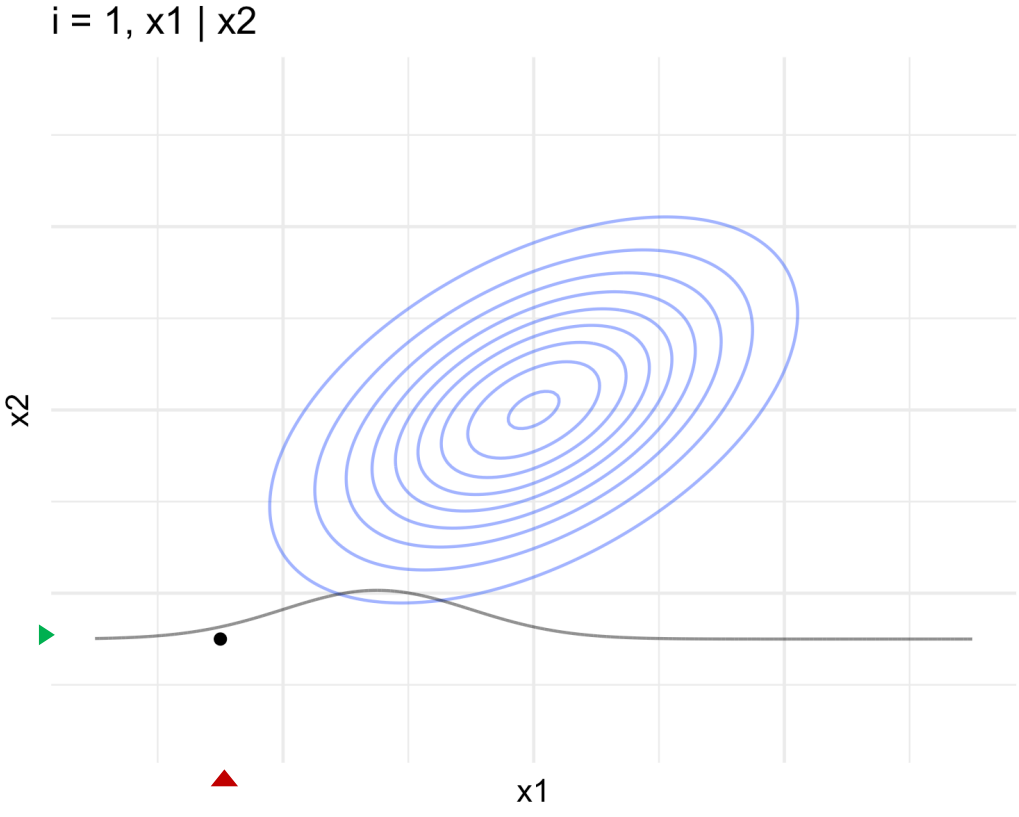
$$\boldsymbol{X} = \begin{pmatrix} x_A \\ x_B \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_A \\ \mu_B \end{pmatrix}, \begin{pmatrix} \Sigma_{AA} & \Sigma_{AB} \\ \Sigma_{BA} & \Sigma_{BB} \end{pmatrix}\right)$$

$$X_A | X_B \sim N(\mu_A + \Sigma_{AB}\Sigma_{BB}^{-1}(x_B - \mu_B), \Sigma_{AA} - \Sigma_{AB}\Sigma_{BB}^{-1}\Sigma_{BA})$$



Gibbs Sampling: bivariate normal example

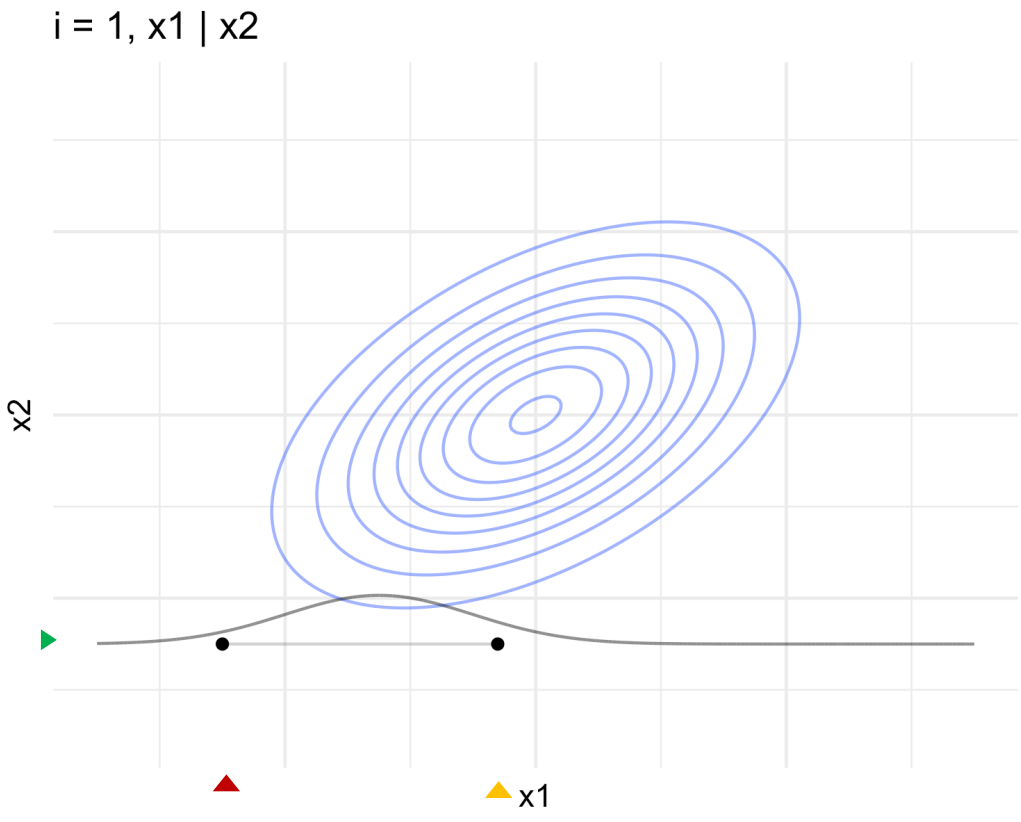
$X_0 = (\blacktriangle, \blacktriangleright)$



Gibbs Sampling: bivariate normal example

$X_0 = (\blacktriangle, \blacktriangleright)$

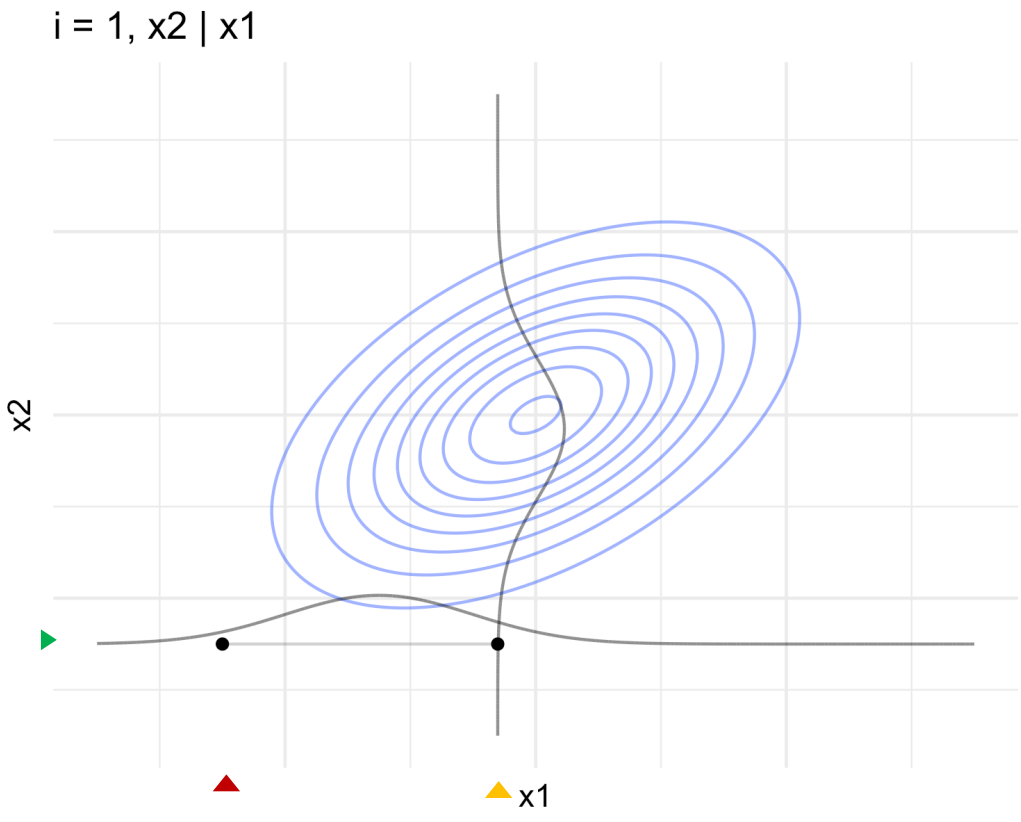
$X_1 = (\blacktriangle, \quad)$



Gibbs Sampling: bivariate normal example

$X_0 = (\blacktriangle, \blacktriangleright)$

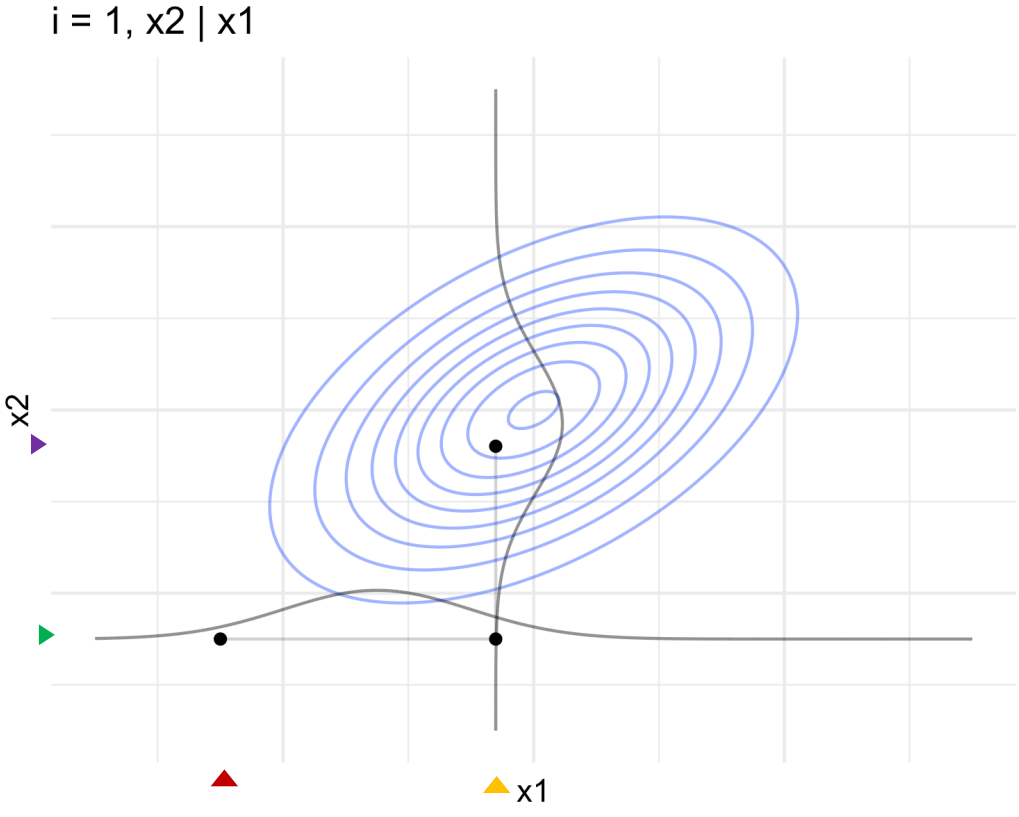
$X_1 = (\blacktriangle, \quad)$



Gibbs Sampling: bivariate normal example

$X_0 = (\blacktriangle, \blacktriangleright)$

$X_1 = (\blacktriangle, \blacktriangleright)$

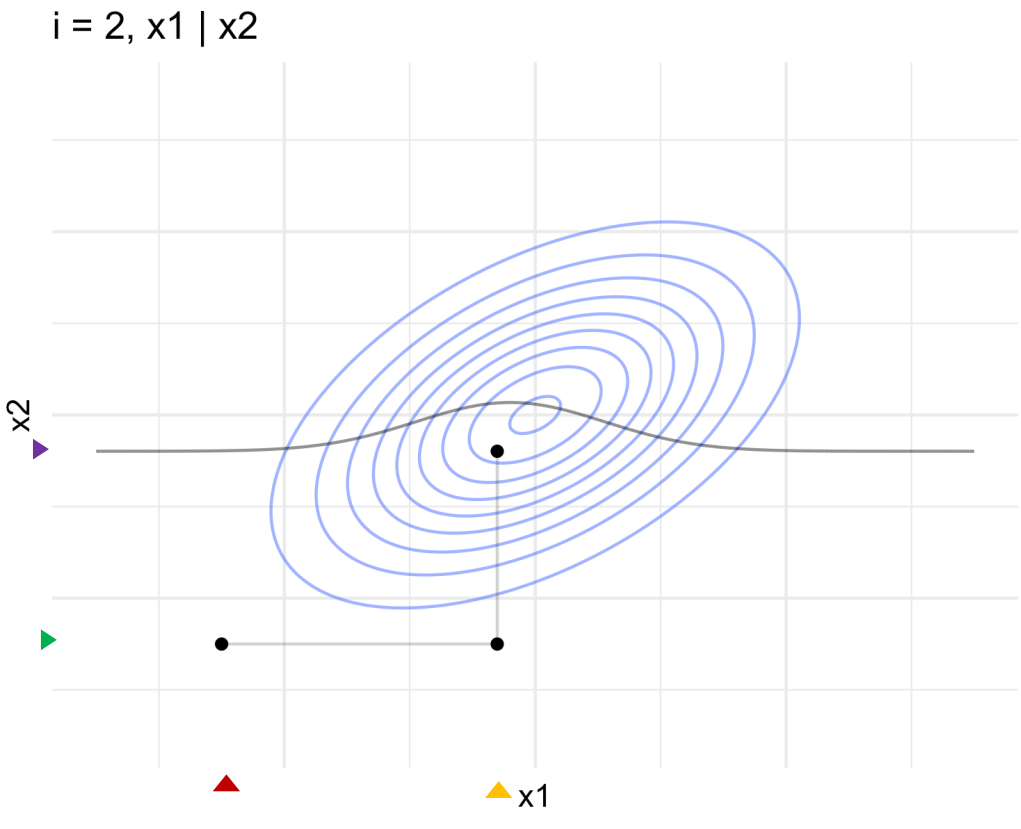


Gibbs Sampling: bivariate normal example

$X_0 = (\blacktriangle, \blacktriangleright)$

$X_1 = (\blacktriangle, \blacktriangleright)$

$X_2 = (\quad, \quad)$

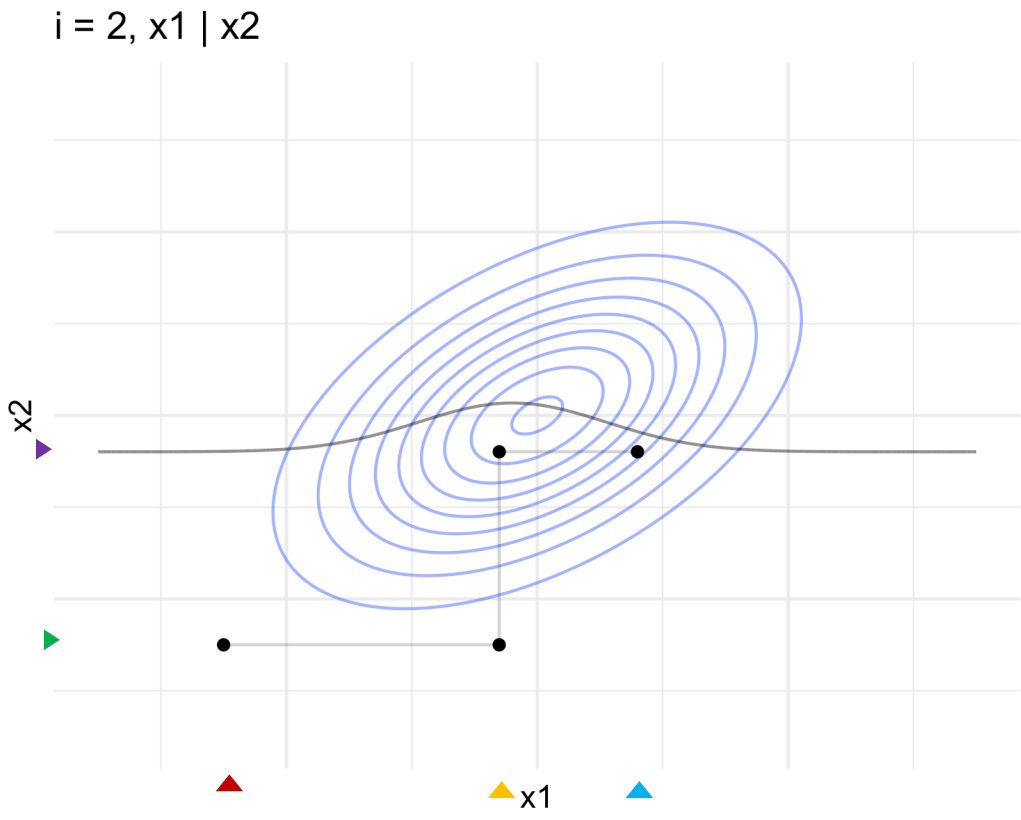


Gibbs Sampling: bivariate normal example

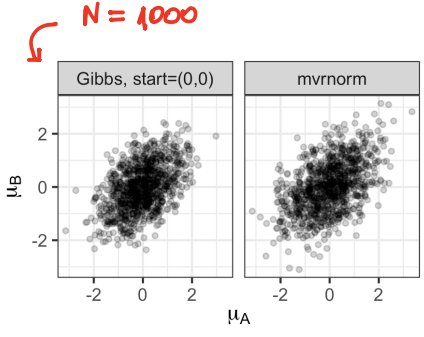
$X_0 = (\blacktriangle, \blacktriangleright)$

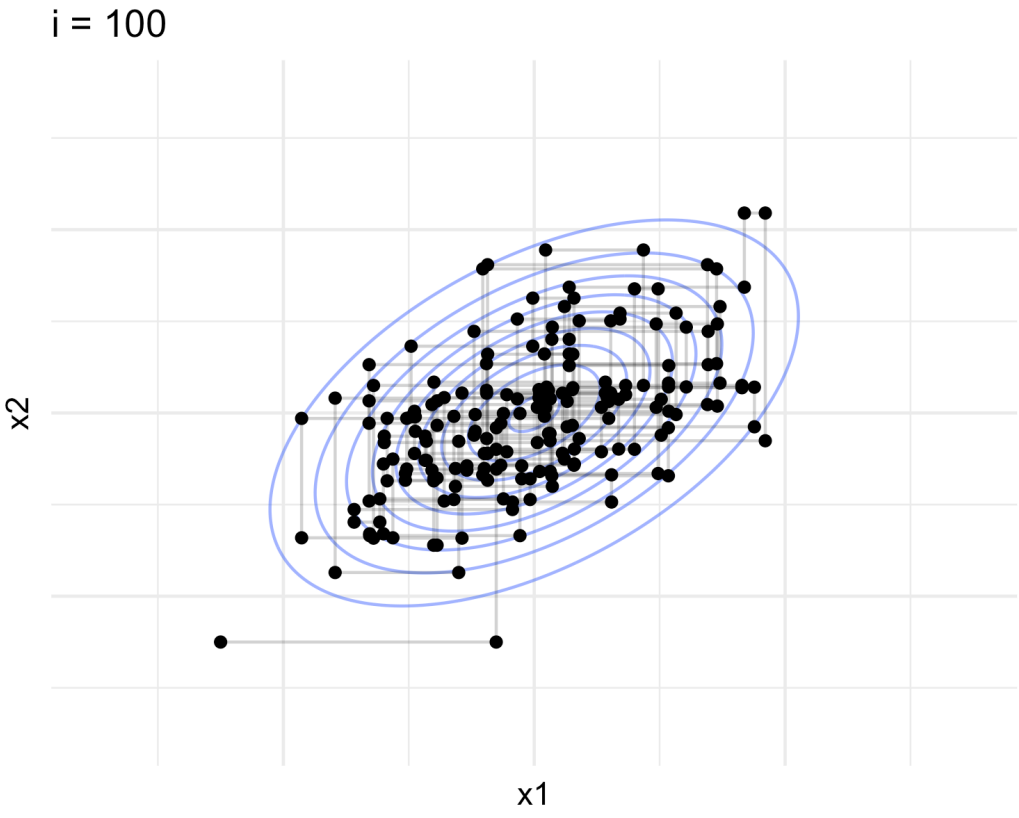
$X_1 = (\blacktriangle, \blacktriangleright)$

$X_2 = (\blacktriangle, \quad)$



Gibbs Sampling: bivariate normal example





Gibbs Sampling

- We use Gibbs sampling when the joint posterior distribution is difficult to sample from directly and full conditional distributions of the parameters are known.

Algorithm 4: Gibbs sampler

Initialize $\boldsymbol{x}^{(0)}$.

for $i = 1{:}N$ **do**

$x_1^{(i)} \sim p(x_1 | x_2^{(i-1)}, x_3^{(i-1)}, \ldots, x_n^{(i-1)})$

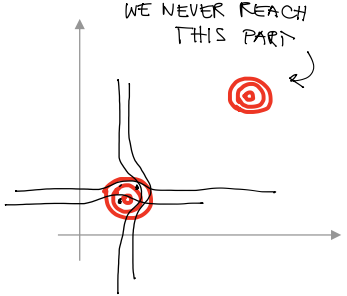
$x_2^{(i)} \sim p(x_2 | x_1^{(i)}, x_3^{(i-1)}, \ldots, x_n^{(i-1)})$

$\vdots$

$x_n^{(i)} \sim p(x_n | x_1^{(i)}, x_2^{(i)}, \ldots, x_{n-1}^{(i)})$

Output: $\{\boldsymbol{x}^{(i)}\}_{i=1}^{N}$

- Specific failure cases, diagnostics is important: check traceplots, autocorrelation, effective sample size.



3. Bayesian Non-negative Matrix Factorization

Bayesian NMF

- Bayesian matrix factorization approaches use a likelihood distribution to capture the noise in the data and place priors over factor matrices:

$$p(\boldsymbol{W}, \boldsymbol{H}, \boldsymbol{\theta} | \boldsymbol{X}) \propto p(\boldsymbol{X} | \boldsymbol{W}, \boldsymbol{H}, \boldsymbol{\theta}) p(\boldsymbol{W}, \boldsymbol{H}, \boldsymbol{\theta})$$

likelihood ≈ cost function
priors ≈ penalty terms

- Several approaches to Bayesian NMF:
  - Gibbs sampling with different prior-likelihood distribution combinations
  - Variational Bayesian inference
  - …
- Our agenda: *NMF Gibbs sampler* & *Iterated conditional modes (ICM)* algorithm

Schmidt, M. N., Winther, O., & Hansen, L. K. (2009, March). Bayesian non-negative matrix factorization. In *International Conference on Independent Component Analysis and Signal Separation* (pp. 540-547). Springer, Berlin, Heidelberg.

$$p(\boldsymbol{W}, \boldsymbol{H}, \boldsymbol{\theta} | \boldsymbol{X}) \propto p(\boldsymbol{X} | \boldsymbol{W}, \boldsymbol{H}, \boldsymbol{\theta}) p(\boldsymbol{W}, \boldsymbol{H}, \boldsymbol{\theta})$$

NMF Gibbs sampler

The NMF problem can be rewritten as: $\boldsymbol{X} = \boldsymbol{WH} + \boldsymbol{E}$, where $\boldsymbol{E}$ is a residual matrix.

LIKELIHOOD

We assume that the residuals $\boldsymbol{E}$ are **normally distributed** with mean 0 and variance σ^2.

$$p(\boldsymbol{X} | \boldsymbol{W}, \boldsymbol{H}, \sigma^2) = \prod_{n,p} \mathcal{N}(\boldsymbol{X}_{n,p}; (\boldsymbol{WH})_{n,p}, \sigma^2)$$

PRIORS

- $\boldsymbol{W}$ and $\boldsymbol{H}$ independently **exponentially distributed** with scales $\alpha_{\mathrm{n,r}}$, $\beta_{\mathrm{r,p.}}$
- The noise variance σ^2 has an **inverse gamma distribution** with shape k and scale θ.

$$p(\boldsymbol{W}) = \prod_{n,r} \mathcal{E}(\boldsymbol{W}_{n,r}; \alpha_{n,r})$$

$$p(\boldsymbol{H}) = \prod_{r,p} \mathcal{E}(\boldsymbol{H}_{r,p}; \beta_{r,p})$$

$$p(\sigma^2) = \mathcal{G}^{-1}(\sigma^2; k, \theta).$$

POSTERIOR $p(\boldsymbol{W}, \boldsymbol{H}, \sigma^2 \mid \boldsymbol{X})$

- Proportional to the product of likelihood and priors.
- However, we are interested in estimating the marginal density of the factors ... and we can't really integrate the posterior.

But we only need to know how to sample from the full conditionals!

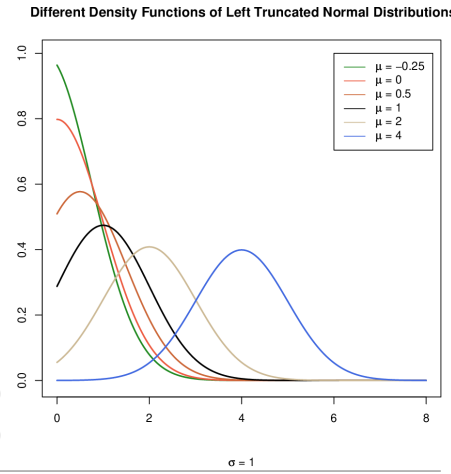
NMF Gibbs sampler

Full conditionals:

$$p(W_{n,r}|\boldsymbol{X}, \boldsymbol{W}_{\setminus(n,r)}, \boldsymbol{H}, \sigma^2) \propto p(\boldsymbol{X}|\boldsymbol{W}, \boldsymbol{H}, \sigma^2)p(\boldsymbol{W}) \propto \ldots \propto \mathcal{TN}(W_{n,r}; \mu_{W_{nr}}, \sigma^2_{W_{nr}})$$

$$p(H_{r,p}|\boldsymbol{X}, \boldsymbol{W}, \boldsymbol{H}_{\setminus(r,p)}, \sigma^2) \propto p(\boldsymbol{X}|\boldsymbol{W}, \boldsymbol{H}, \sigma^2)p(\boldsymbol{H}) \propto \ldots \propto \mathcal{TN}(H_{r,p}; \mu_{H_{rp}}, \sigma^2_{H_{rp}})$$

$$p(\sigma^2|\boldsymbol{X}, \boldsymbol{W}, \boldsymbol{H}) \propto \ldots \propto \mathcal{G}^{-1}(\sigma^2; k^*, \theta^*)$$



$$\mathcal{TN}(x; \mu, \sigma^2) = \begin{cases} \frac{\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)}{1-\Phi\left(-\frac{\mu}{\sigma}\right)} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases}$$

Math. derivation: Brouwer, T., Frellsen, J., & Lio, P. (2016). Fast Bayesian non-negative matrix factorisation and tri-factorisation. *arXiv preprint arXiv:1610.08127*.

Image credits: : https://www.semanticscholar.org/paper/Parameter-Estimation-and-Hypothesis-Testing-for-the-Hattaway/e8dacbce1f5b418b572656ac4f59e49f6129ce5f/figure/0

NMF Gibbs sampler

Efficient implementation: the elements in each column of $\boldsymbol{W}$ (row of $\boldsymbol{H}$) are conditionally independent, so we can sample an entire column (row) simultaneously.

Algorithm 3: NMF Gibbs sampler

Initialize $\boldsymbol{W}^{(0)}$ and $\boldsymbol{H}^{(0)}$.
for $i = 1{:}N$ **do**
  $\boldsymbol{W} \leftarrow \boldsymbol{W}^{(i-1)}$, $\boldsymbol{H} \leftarrow \boldsymbol{H}^{(i-1)}$
  for $j = 1$ *to* r **do**
    $\boldsymbol{W}_{:,j} \leftarrow \mathcal{TN}(\mu_{W_{:,j}}, \sigma^2_{W_{:,j}})$
  $\sigma^2 \leftarrow \mathcal{G}^{-1}(k^*, \theta_i^*)$
  for $j = 1$ *to* r **do**
    $\boldsymbol{H}_{j,:} \leftarrow \mathcal{TN}(\mu_{H_{j,:}}, \sigma^2_{H_{j,:}})$
  $\boldsymbol{W}^{(i)} \leftarrow \boldsymbol{W}$, $\boldsymbol{H}^{(i)} \leftarrow \boldsymbol{H}$
Output: $\{\boldsymbol{W}^{(i)}, \boldsymbol{H}^{(i)}\}_{i=1}^{N}$

Iterated conditional modes (ICM)

- Maximum-a-posteriori (MAP) estimate: $\theta_{MAP} = \max_{\theta} p(\theta|\boldsymbol{X})$
- For random variables $A \sim \mathcal{G}^{-1}(a, b)$ and $B \sim \mathcal{TN}(\mu, \sigma^2)$ the modes are $\frac{b}{a+1}$ and $\max(0, \mu)$, respectively.

Algorithm 4: ICM

Initialize $\boldsymbol{W}$ and $\boldsymbol{H}$.
$i = 1$
while *stopping criteria not reached* **do**
  for $j = 1$ *to* r **do**
    $\boldsymbol{W}_{:,j} \leftarrow \max(0, \mu_{W_{:,j}})$
  $\sigma^2 = \frac{\theta_i^*}{k^*+1}$
  for $j = 1$ *to* r **do**
    $\boldsymbol{H}_{j,:} \leftarrow \max(0, \mu_{H_{j,:}})$
  $i = i + 1$
Output: $\boldsymbol{W}, \boldsymbol{H}$

To Bayes or not to Bayes

Experiments

Convergence:

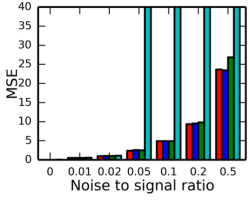
- Iterations: track the convergence rate of the error on the data against the number of iterations taken.
- Time: timing each method 10 times and taking the average.

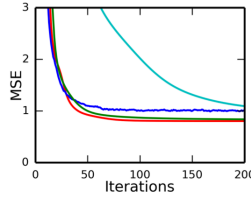
Missing data:

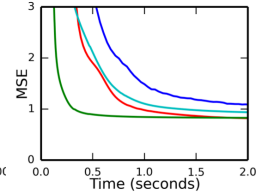
- Testing the ability of the model to recover missing values as the fraction of unknown entries increases.

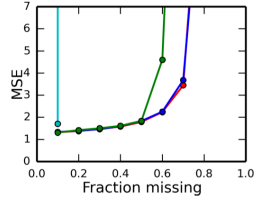
Noise:

- Adding different levels of Gaussian noise to the data.









Method	Estimate	Requires sampling	Speed of convergence	Robustness
Non-probabilistic	point	no	medium	low
ICM	point (MAP)	no	high	medium
Gibbs	full posterior	yes	low	high

Brouwer, T., Frellsen, J., & Lio, P. (2016). Fast Bayesian non-negative matrix factorisation and tri-factorisation. *arXiv preprint arXiv:1610.08127*.

To Bayes or not to Bayes

Choice of r

- In the Bayesian framework model selection can be performed by evaluating the marginal likelihood, $p(\boldsymbol{X})$.

Chib's method

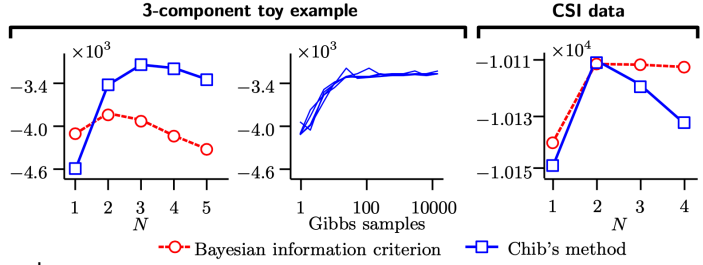
- Can be applied to Gibbs sampler output and is based on the relation: $p(\boldsymbol{X}) = \dfrac{p(\boldsymbol{X}|\theta)p(\theta)}{p(\theta|\boldsymbol{X})}$

$$\log p(\boldsymbol{X}) = \log p(\boldsymbol{X}|\theta^*) + \log p(\theta^*) - \log p(\theta^*|\boldsymbol{X})$$

- We can partition θ in arbitrary number of blocks, the only requirement: full-conditional sampling from each block needs to be possible!

Example

- 3-component toy example:
  - Generated data: two random unit mean i.i.d. exponentialy distributed matrices, where n = 100, p = 20, r = 3.
  - Added noise.
  - Computed marginal likelihood and BIC for r between 1 and 5.
- CSI data: chemical shift imaging data containing 2 spectral components.



Schmidt, M. N., Winther, O., & Hansen, L. K. (2009, March). Bayesian non-negative matrix factorization. In *International Conference on Independent Component Analysis and Signal Separation* (pp. 540-547). Springer, Berlin, Heidelberg.

Chib, S. (1995). Marginal likelihood from the Gibbs output. *Journal of the american statistical association*, *90*(432), 1313-1321.

Conclusion

- NMF is a dimensionality reduction method, providing more intuitive and meaningful interpretation, which is why it is so attractive for many different domains.

- Non-probabilistic vs. Bayesian:
  - Trade-offs in convergence, robustness to noise & sparsity.
  - Bayesian approach solves some of the issues of the non-probabilistic NMF.

- Nowadays we have more advanced methods for tasks, such as recommendation, however, when we are interested in interpretability, we still utilize methods based on the ideas presented today.

Materials & code: https://github.com/ggapac/b-nmf
Contact: greta.gasparac@fri.uni-lj.si