Bayesian matrix factorization with linear constraints

Mikkel N. Schmidt

UNIVERSITY OF CAMBRIDGE

Matrix factorization

- Matrix factorization

$$\underset{I \times J}{\boldsymbol{X}} = \underset{I \times K}{\boldsymbol{A}} \; \underset{K \times J}{\boldsymbol{B}} + \underset{I \times J}{\boldsymbol{N}} \qquad x_{ij} = \sum_{k=1}^{K} a_{ik} b_{kj} + n_{ij}$$

- Bi-linear model
- Specific assumptions / constraints / priors
  - Principal component analysis (PCA)
  - Probabilistic PCA and factor analysis
  - Independent component analysis (ICA)
  - Non-negative matrix factorization (NMF)

Motivation for linear constraints

- Intuitive way to specify prior information
- Can dramatically influence results
  - FA vs. NMF: Non-negativity constraint
    Can other constraints be equally powerful ?
- We develop a Bayesian framework for linearly constrained matrix factorization

Likelihood and noise prior

- Gaussian likelihood

$$p(\boldsymbol{X}|\boldsymbol{A}, \boldsymbol{B}) = \prod_{ij} \mathcal{N}\left(x_{ij} \middle| \sum_k a_{ik} b_{kj}, v_{ij}\right)$$

- Inverse-gamma noise variance prior

$$p(v_{ij}) = \mathcal{IG}(v_{ij}|\alpha, \beta)$$

Priors for matrices

Constrained Gaussian

$$p(\boldsymbol{a}) \propto \begin{cases} \mathcal{N}(\boldsymbol{\mu}_a, \boldsymbol{\Sigma}_a), & \text{if } \boldsymbol{Q}_a^\top \boldsymbol{a} \leq \boldsymbol{q}_a, \;\; \boldsymbol{R}_a^\top \boldsymbol{a} = \boldsymbol{r}_a, \\ 0, & \text{otherwise.} \end{cases}$$

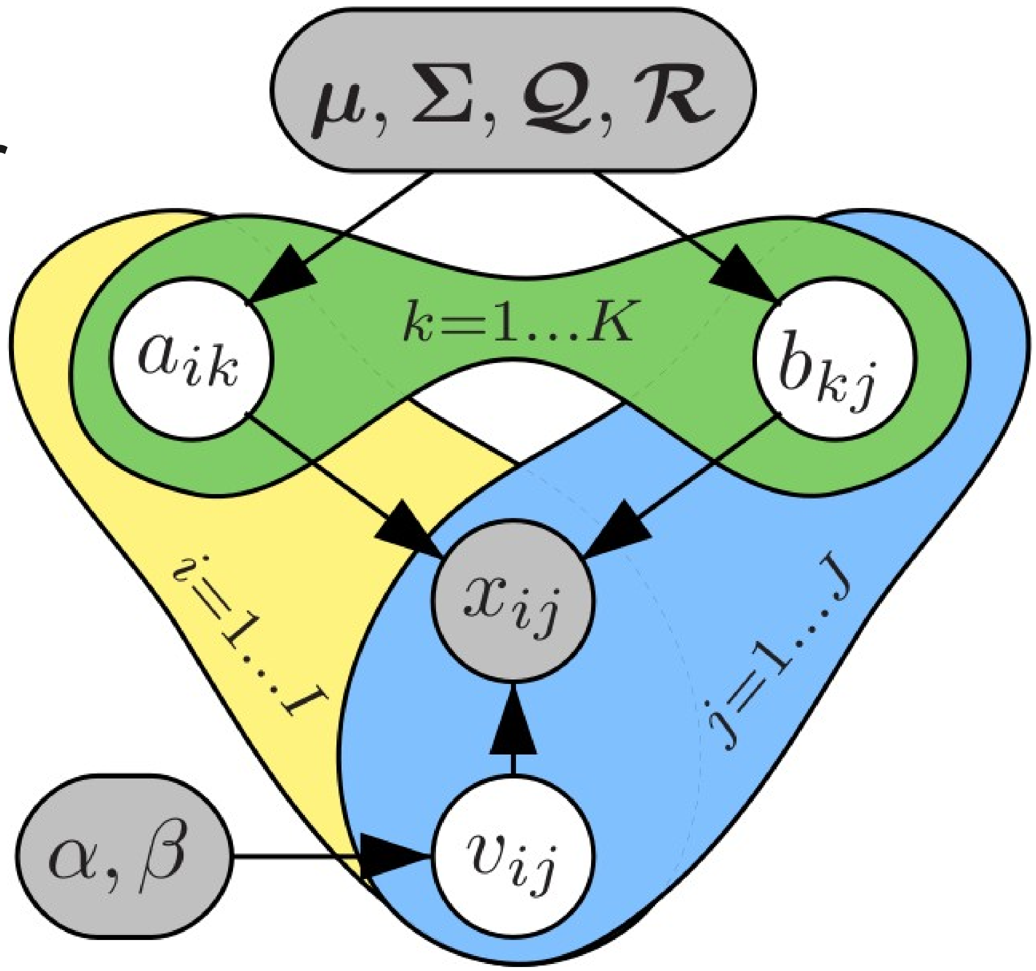
- Conjugate
  - Posterior conditional is also constrained Gaussian
- Normalization constant is intractable
  - Approximate inference

$\boldsymbol{a} = \mathrm{vec}(\boldsymbol{A})$

Inference

- Gibbs sampling
- Sample from posterior conditionals
  - Noise variance
    Inverse-gamma
  - Matrices ***A*** and ***B***
    Constrained Gaussian

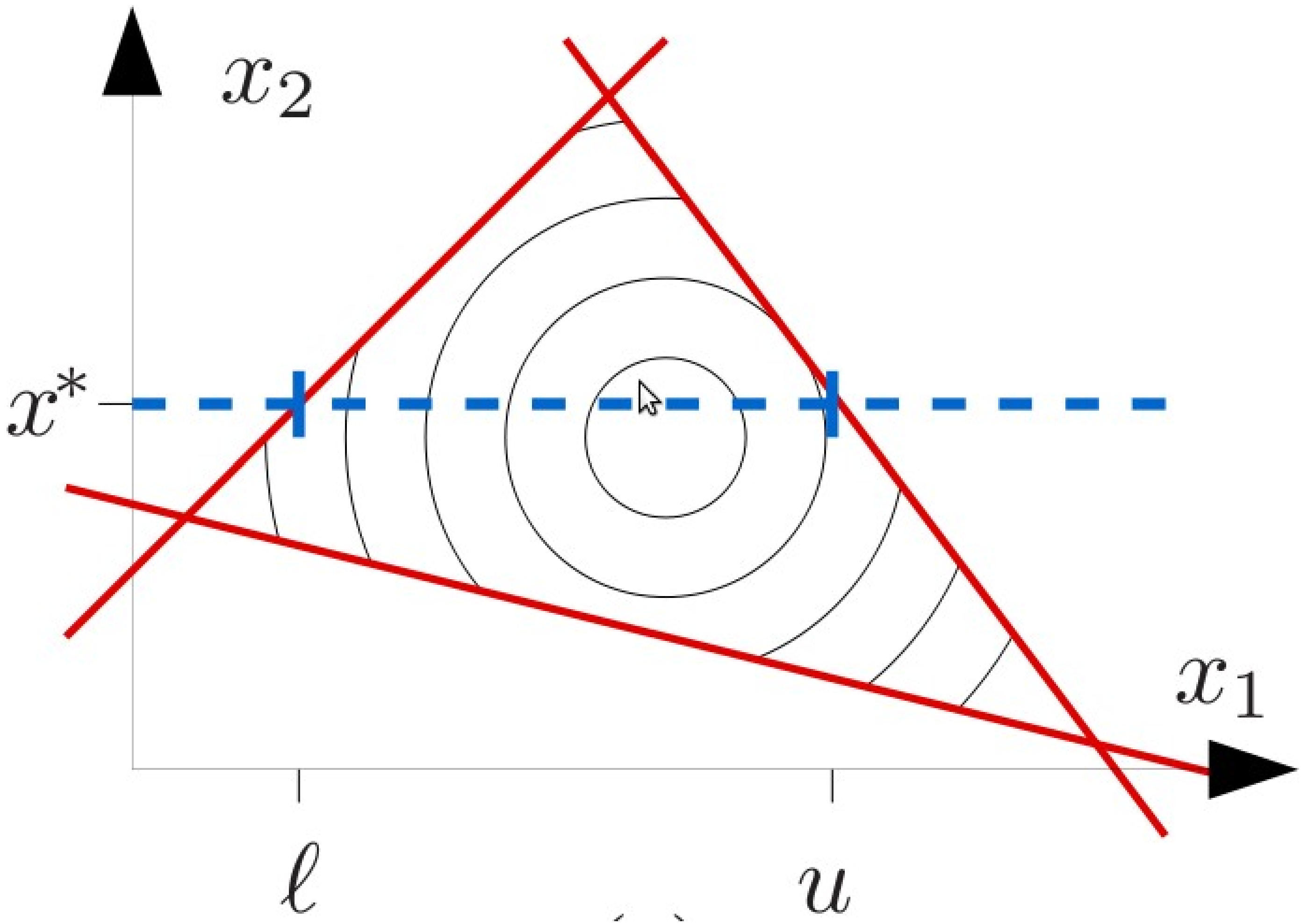
Graphical model



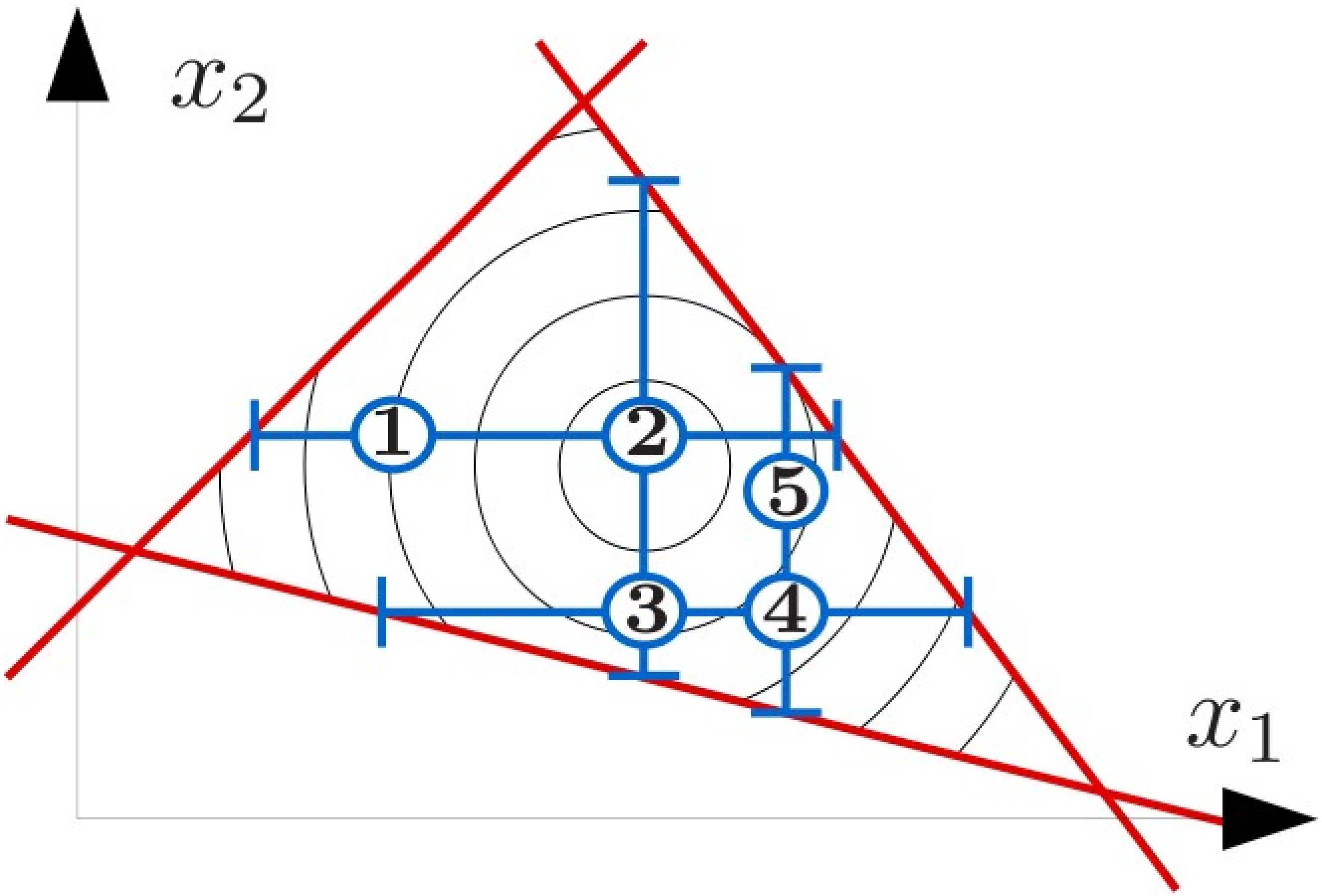
Sampling from constrained Gaussian

- Equality constraints
  - Project onto affine constraint subspace
    Gaussian with inequality constraints
- Inequality constraints
  - Gibbs sampler: Truncated Gaussian
- Truncated Gaussian
  - Rejection sampling (Geweke, 1991)
  - Inverse transform sampling
  - Slice sampling (Neal, 2003)

Gaussian with inequality constraints

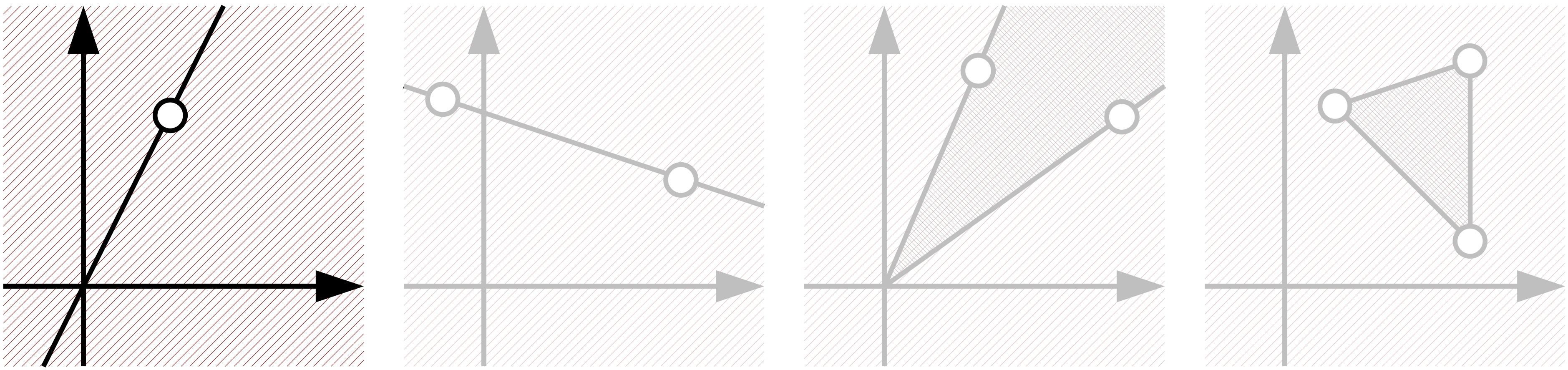


Gaussian with inequality constraints



Examples of model spaces

Linear subspace	Affine subspace	Polytopal cone	Polytope

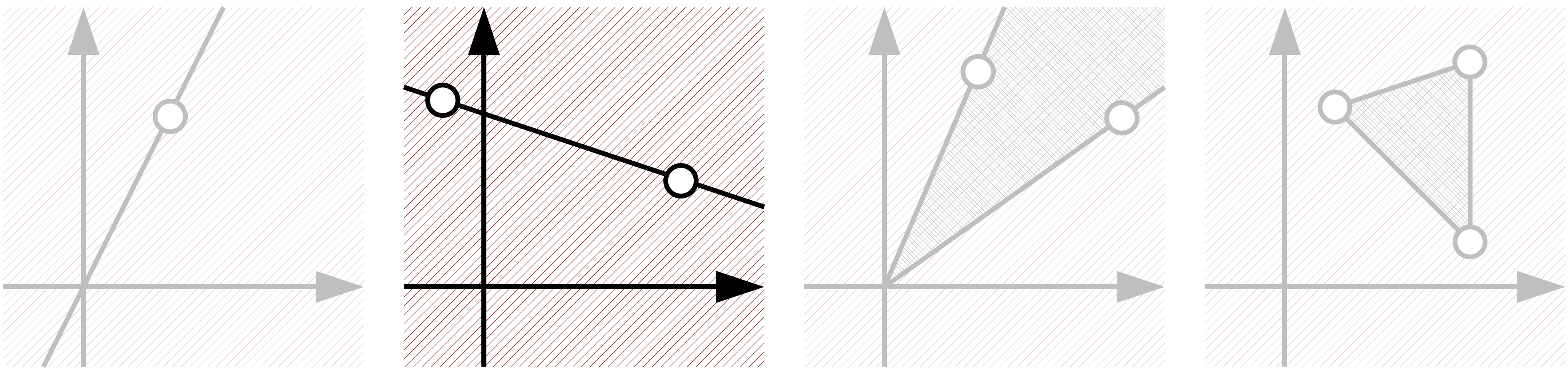


No constraints	$\sum_k b_{kj} = 1$	$b_{kj} \geq 0$	$b_{kj} \geq 0, \sum_k b_{kj} = 1$

- Basis vector
- Feasible region for basis vectors
- Feasible region for data vectors

Examples of model spaces

Linear subspace | Affine subspace | Polytopal cone | Polytope



No constraints

$\sum_k b_{kj} = 1$

$b_{kj} \geq 0$

$b_{kj} \geq 0, \sum_k b_{kj} = 1$

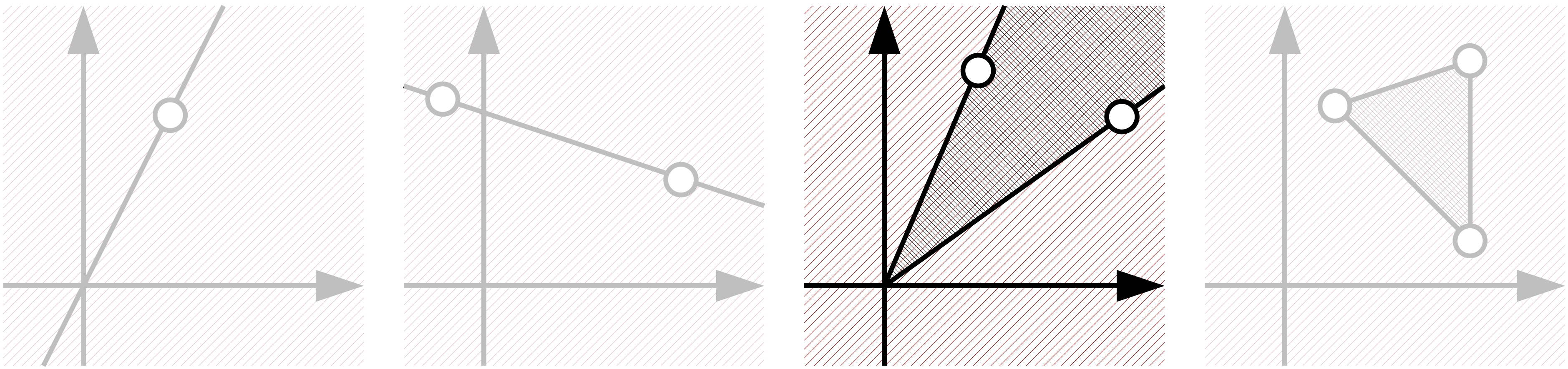
Basis vector

Feasible region for basis vectors

Feasible region for data vectors

Examples of model spaces

Linear subspace | Affine subspace | Polytopal cone | Polytope



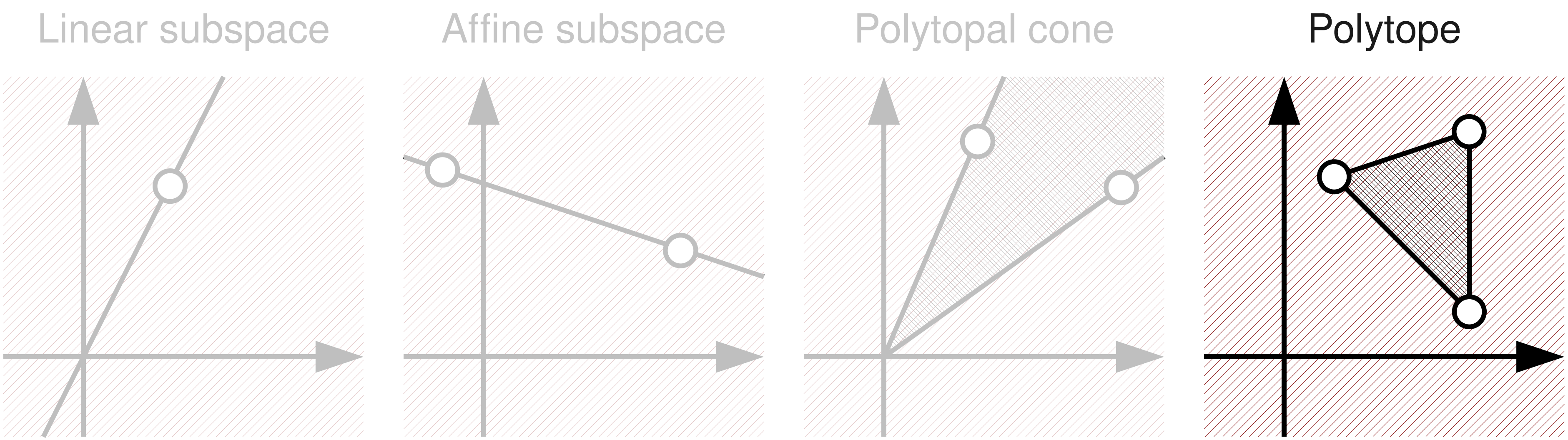
- Linear subspace: No constraints
- Affine subspace: $\sum_k b_{kj} = 1$
- Polytopal cone: $b_{kj} \geq 0$
- Polytope: $b_{kj} \geq 0, \sum_k b_{kj} = 1$

○ Basis vector

Feasible region for basis vectors

Feasible region for data vectors

Examples of model spaces



No constraints

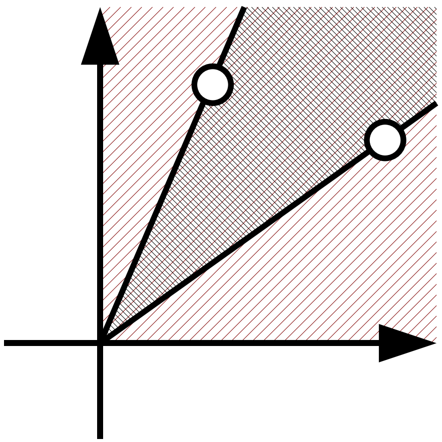
$$\sum_k b_{kj} = 1$$

$$b_{kj} \geq 0$$

$$b_{kj} \geq 0, \sum_k b_{kj} = 1$$

Examples of model spaces

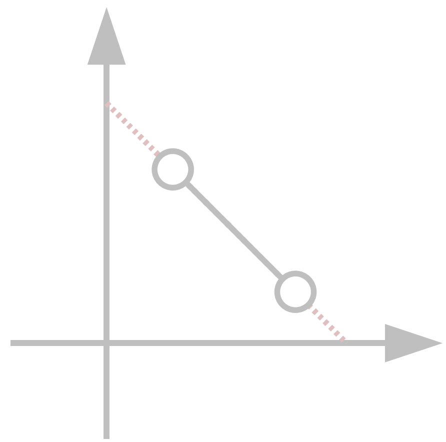
Polytopal cone in non.neg orthant



$a_{ik} \geq 0 \quad b_{kj} \geq 0$

Non-negative matrix factorization

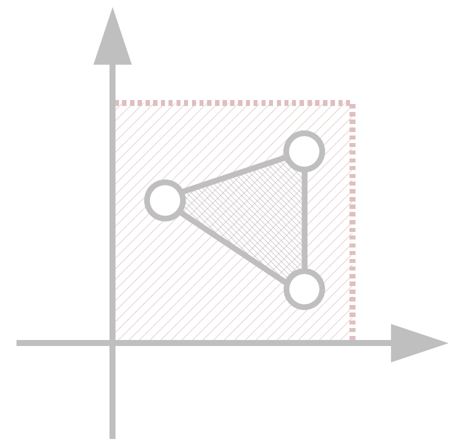
Polytope on unit simplex



$a_{ik} \geq 0, \sum_k a_{ik} = 1$

$b_{kj} \geq 0, \sum_k b_{kj} = 1$

Polytope in unit hypercube



$0 \leq a_{ik} \leq 1$

$b_{kj} \geq 0, \sum_k b_{kj} = 1$

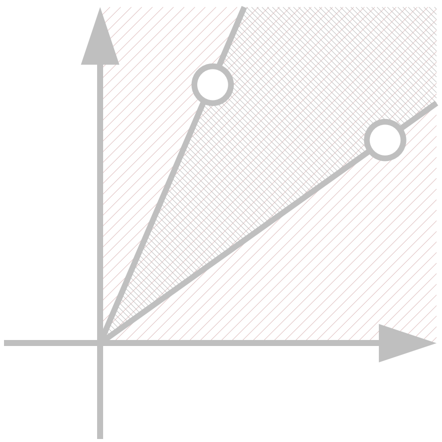
○ Source vector

Feasible region for source vectors

Feasible region for data vectors

Examples of model spaces

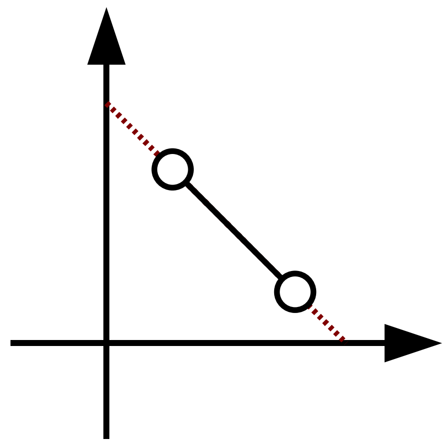
Polytopal cone in non.neg orthant



$$a_{ik} \geq 0 \quad b_{kj} \geq 0$$

Non-negative matrix factorization

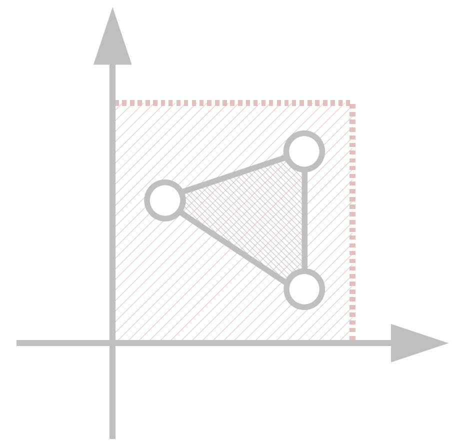
Polytope on unit simplex



$$a_{ik} \geq 0, \sum_k a_{ik} = 1$$

$$b_{kj} \geq 0, \sum_k b_{kj} = 1$$

Polytope in unit hypercube



$$0 \leq a_{ik} \leq 1$$

$$b_{kj} \geq 0, \sum_k b_{kj} = 1$$

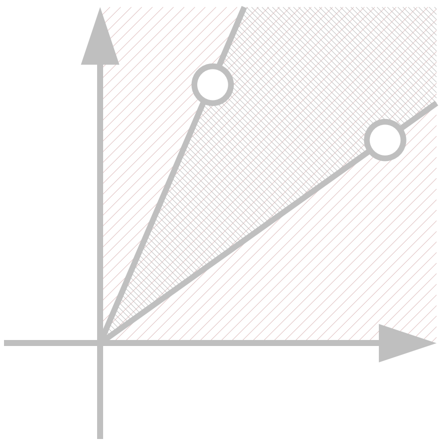
○ Source vector

Feasible region for source vectors

Feasible region for data vectors

Examples of model spaces

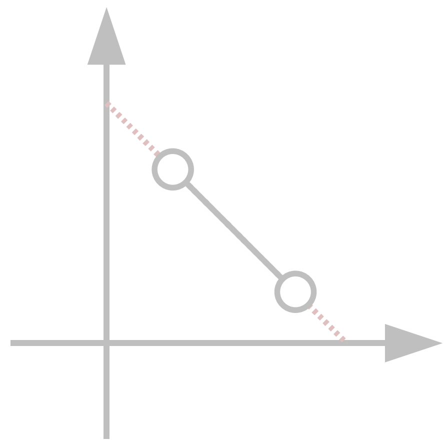
Polytopal cone in non.neg orthant



$$a_{ik} \geq 0 \quad b_{kj} \geq 0$$

Non-negative matrix factorization

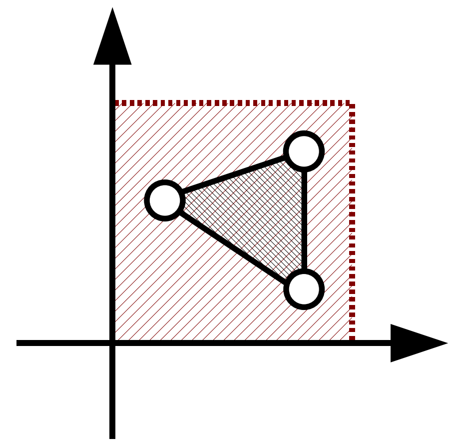
Polytope on unit simplex



$$a_{ik} \geq 0, \sum_k a_{ik} = 1$$

$$b_{kj} \geq 0, \sum_k b_{kj} = 1$$

Polytope in unit hypercube



$$0 \leq a_{ik} \leq 1$$

$$b_{kj} \geq 0, \sum_k b_{kj} = 1$$

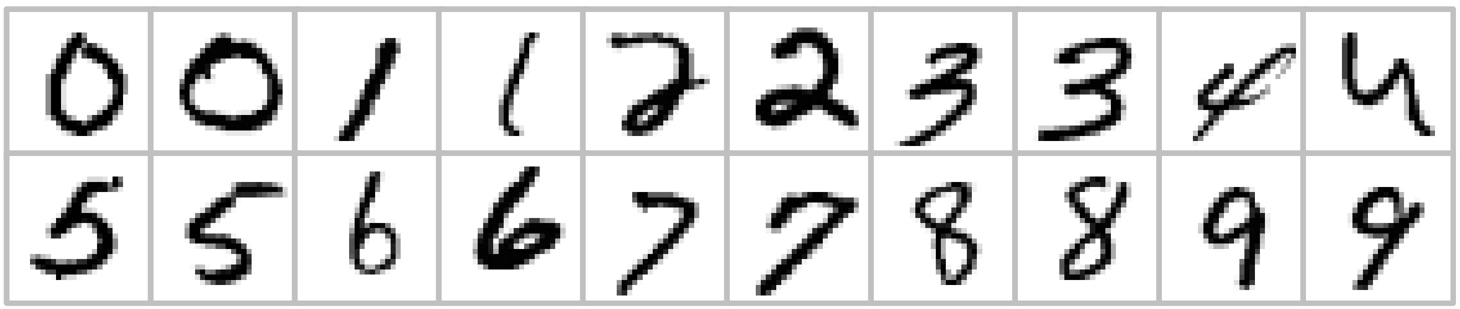
Source vector

Feasible region for source vectors

Feasible region for data vectors

Simulation experiments

- MNIST handwritten digits



  - Grayscale images
  - 28 x 28 pixels

Mixture dataset

- Mixtures of two handwritten digits

- Images added and normalized
  - 8000 unique images, 4000 mixed images
  - Data matrix: 784 x 4000
- Compute interpretable features that explain data

UNIVERSITY OF CAMBRIDGE

Computational and Biological Learning

Linear constraints

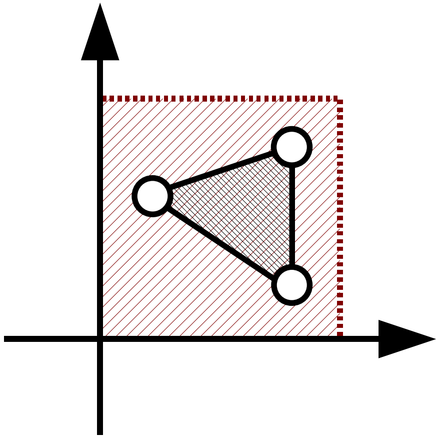
A

- Between zero and one
  - Allows interpretation as image

B

- Non-negative
  - Only additive combinations
- Sum-to-unity
  - Negative correlation: Compete, not collaborate

Polytope in unit hypercube



$$0 \leq a_{ik} \leq 1$$

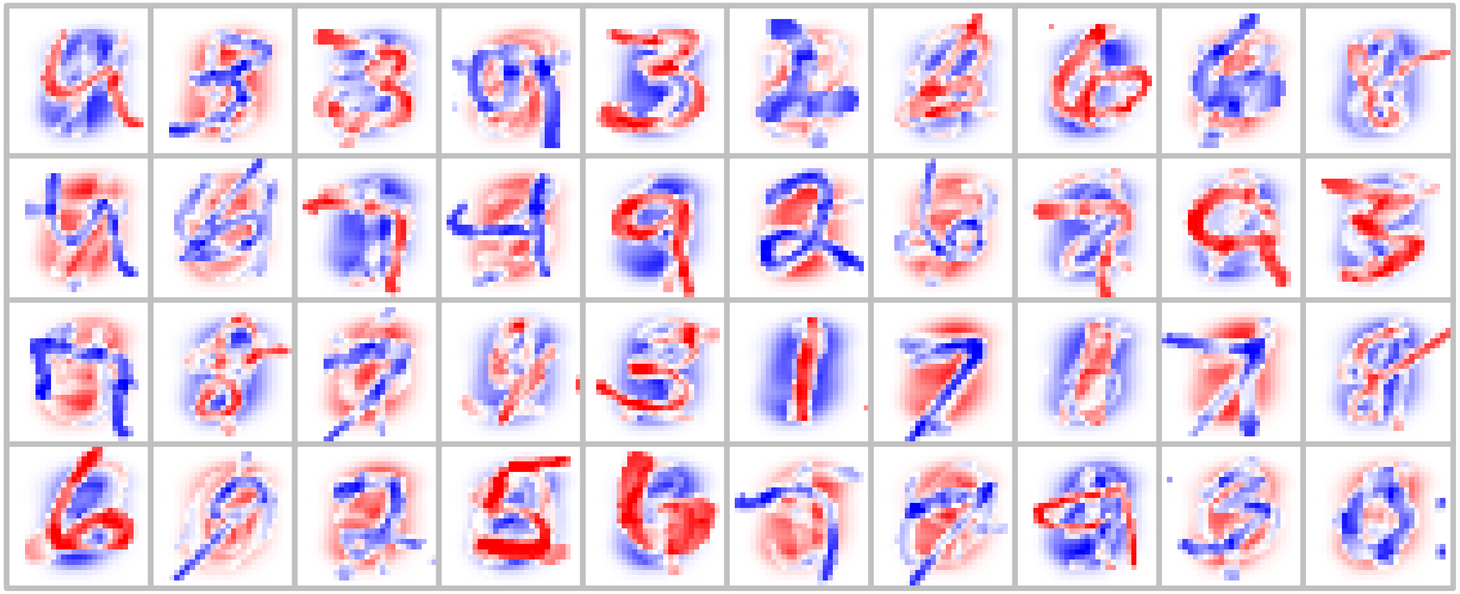
$$b_{kj} \geq 0, \sum_k b_{kj} = 1$$

Experiment details

- Priors
  - Isotropic noise model
  - Standard Gaussian priors
- 40 features (4 exemplars per digit)
- 10,000 Gibbs samples
- Comparison to ICA and NMF

Independent component analysis

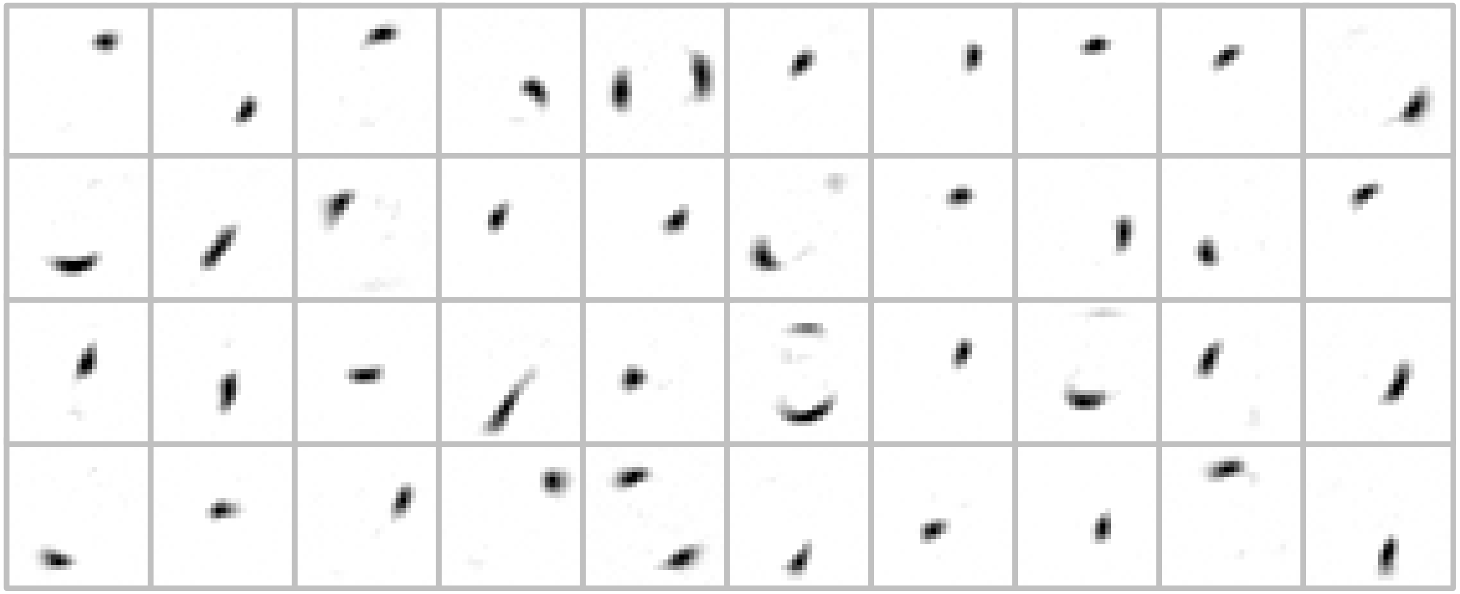
FastICA (Hyvärinen, 1999)



- Complex patterns, dominated by single digit
- Negative values: No image interpretation

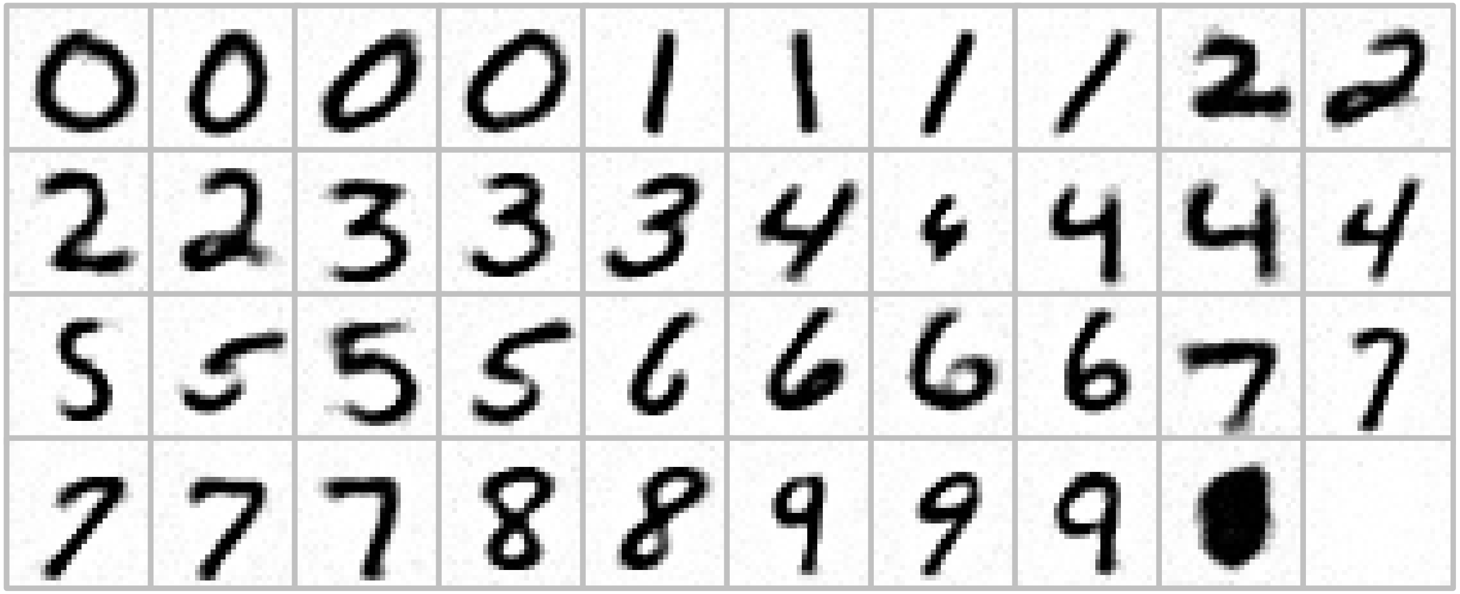
Non-negative matrix factorization

Multiplicative updates (Lee and Seung, 1999)



- Parts-based, sparse features
- Clear interpretation as image features

Linearly constrained matrix factorization



- Features resemble handwritten digits
  - One is a black blob. One is all white

Conclusions

- We have presented
  - Matrix factorization with linear constraints
  - Inference via Gibbs sampling
- We have demonstrated
  - Constraints dramatically influence results
  - Useful for unsupervised source separation

Thank you

References

Geweke (1991), Efficient Simulation from the Multivariate Normal and Student-t Distributions Subject to Linear Constraints and the Evaluation of Constraint Probabilities Computer Sciences and Statistics, Proceedings the 23rd Symposium on the Interface between, 571-578

Hyvärinen (1999), Fast and Robust Fixed-Point Algorithms for Independent Component Analysis. IEEE Transactions on Neural Networks 10(3):626-634

Lee and Seung (1999), Learning the parts of objects by non-negative matrix factorization Nature, 401, 788-791

Neal (2003), Slice sampling, Annals of Statistics, 31, 705-76

Schmidt (2009), Linearly constrained Bayesian matrix factorization for blind source separation, Submitted