Automatic Relevance Determination in Nonnegative Matrix Factorization (NMF)

Vincent Y. F. Tan [†] and Cédric Févotte [*]

[†]Laboratory for Information and Decision Systems (LIDS), Massachusetts Institute of Technology.

[*]Laboratoire Traitement et Communication de l'Information (LTCI) , CNRS - TELECOM ParisTech.

Massachusetts Institute of Technology

SPARS Workshop (Apr 6, 2009)

Introduction and Motivation

- Nonnegative matrix factorization (Lee and Seung 1999) is a popular technique for:
  1. Data analysis.
  2. Dimensionality reduction.
- NMF $\equiv$ non-subtractive, parts-based representation of nonnegative data.

Introduction and Motivation

- Nonnegative matrix factorization (Lee and Seung 1999) is a popular technique for:
  1. Data analysis.
  2. Dimensionality reduction.
- NMF $\equiv$ non-subtractive, parts-based representation of nonnegative data.
- Often, number of latent dimensions (or components) is assumed. Usually, this is not provided *a-priori*.

Introduction and Motivation

- We propose a Bayesian approach to estimate the latent dimensionality or model order.
- This is achieved by performing Automatic Relevance Determination (Mackay 1995).
- This has been used in Bayesian PCA (Bishop 1999) and sparse linear regression (Tipping 2001).

Related Work

- Generally little literature about model order selection in NMF.

Related Work

- Generally little literature about model order selection in NMF.
- Variational Bayesian methods have been proposed (Winther & Petersen 2007, Cemgil 2008) for NMF but such methods are usually computationally demanding.
    - Authors compute an approximation to the model evidence for every model order.

Related Work

- Generally little literature about model order selection in NMF.
- Variational Bayesian methods have been proposed (Winther & Petersen 2007, Cemgil 2008) for NMF but such methods are usually computationally demanding.
    - Authors compute an approximation to the model evidence for every model order.
- The work is somewhat similar to multiplicative sparse NMF algorithms with a sparsity ℓ_1 or ℓ_0 regularizer (Hoyer 2004, Mørup et al. 2008) added to the objective.

Notation – The NMF Model

Given a nonnegative data matrix

$$\mathbf{V} \in \mathbb{R}_{+}^{F \times N}.$$

Notation – The NMF Model

Given a nonnegative data matrix

$$\mathbf{V} \in \mathbb{R}_+^{F \times N}.$$

Task: Find two nonnegative matrices

1. Basis Matrix $\mathbf{W} \in \mathbb{R}_+^{F \times K}$
2. Activation Matrix $\mathbf{H} \in \mathbb{R}_+^{K \times N}$

such that

$$\mathbf{V} \approx \widehat{\mathbf{V}} = \mathbf{WH} = \sum_{k=1}^{K} \mathbf{w}_k h_k.$$

Notation – The NMF Model

Given a nonnegative data matrix

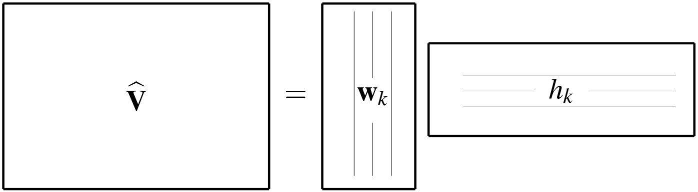
$$\mathbf{V} \in \mathbb{R}_+^{F \times N}.$$

Task: Find two nonnegative matrices

1. Basis Matrix $\mathbf{W} \in \mathbb{R}_+^{F \times K}$
2. Activation Matrix $\mathbf{H} \in \mathbb{R}_+^{K \times N}$

such that

$$\mathbf{V} \approx \widehat{\mathbf{V}} = \mathbf{WH} = \sum_{k=1}^{K} \mathbf{w}_k h_k.$$



Given a nonnegative data matrix

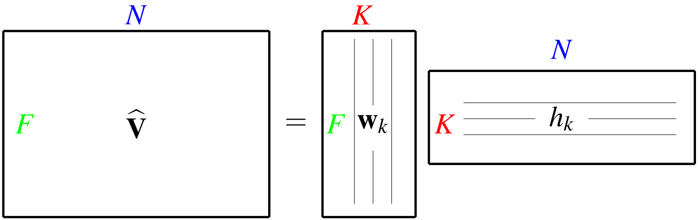
$$\mathbf{V} \in \mathbb{R}_+^{F \times N}.$$

Task: Find two nonnegative matrices

1. Basis Matrix $\mathbf{W} \in \mathbb{R}_+^{F \times K}$
2. Activation Matrix $\mathbf{H} \in \mathbb{R}_+^{K \times N}$

such that

$$\mathbf{V} \approx \widehat{\mathbf{V}} = \mathbf{W}\mathbf{H} = \sum_{k=1}^{K} \mathbf{w}_k h_k.$$



Nonnegative Matrix Factorization

Usually, a cost function $D(\cdot|\cdot)$ is minimized, i.e.,

$$\min_{\mathbf{W},\mathbf{H}} \; D(\mathbf{V}|\mathbf{WH}) = \sum_{f=1}^{F} \sum_{n=1}^{N} d([\mathbf{V}]_{fn}|[\mathbf{WH}]_{fn})$$

Nonnegative Matrix Factorization

Usually, a cost function $D(\cdot|\cdot)$ is minimized, i.e.,

$$\min_{\mathbf{W},\mathbf{H}} \; D(\mathbf{V}|\mathbf{WH}) = \sum_{f=1}^{F}\sum_{n=1}^{N} d([\mathbf{V}]_{fn}|[\mathbf{WH}]_{fn})$$

where the cost (or divergence) d can be

$$d_{EUC}(x|y) = \frac{1}{2}(x-y)^2, \qquad \text{(Euclidean cost)}$$

or

$$d_{KL}(x|y) = x\log\left(\frac{x}{y}\right) - x + y. \qquad \text{(KL-divergence)}$$

Nonnegative Matrix Factorization

Usually, a cost function $D(\cdot|\cdot)$ is minimized, i.e.,

$$\min_{\mathbf{W},\mathbf{H}} \; D(\mathbf{V}|\mathbf{WH}) = \sum_{f=1}^{F}\sum_{n=1}^{N} d([\mathbf{V}]_{fn}|[\mathbf{WH}]_{fn})$$

where the cost (or divergence) d can be

$$d_{EUC}(x|y) = \frac{1}{2}(x-y)^2, \qquad \text{(Euclidean cost)}$$

or

$$d_{KL}(x|y) = x\log\left(\frac{x}{y}\right) - x + y. \qquad \text{(KL-divergence)}$$

Maximum Likelihood estimation of $\mathbf{W}$ and $\mathbf{H}$ corresponds to a particular noise model.

Main Idea

- Set up a statistical model.
- Place precision-like scale parameters or relevance weights

$$\boldsymbol{\beta} \triangleq (\beta_1, \ldots, \beta_K) \in \mathbb{R}_+^K$$

on columns of $\mathbf{W}$ and rows of $\mathbf{H}$.

Main Idea

- Set up a statistical model.
- Place precision-like scale parameters or relevance weights

$$\boldsymbol{\beta} \triangleq (\beta_1, \ldots, \beta_K) \in \mathbb{R}_+^K$$

on columns of $\mathbf{W}$ and rows of $\mathbf{H}$.
- Automatic Relevance Determination.

Main Idea

- Set up a statistical model.
- Place precision-like scale parameters or relevance weights

$$\boldsymbol{\beta} \triangleq (\beta_1, \ldots, \beta_K) \in \mathbb{R}_+^K$$

  on columns of $\mathbf{W}$ and rows of $\mathbf{H}$.
- Automatic Relevance Determination.

$$\widehat{\mathbf{V}} = \mathbf{w}_1 \, h_1 + \cdots + \mathbf{w}_K \, h_K$$

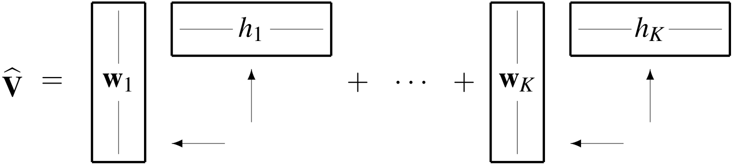
Main Idea

- Set up a statistical model.
- Place precision-like scale parameters or relevance weights

$$\boldsymbol{\beta} \triangleq (\beta_1, \ldots, \beta_K) \in \mathbb{R}_+^K$$

on columns of $\mathbf{W}$ and rows of $\mathbf{H}$.

- Automatic Relevance Determination.

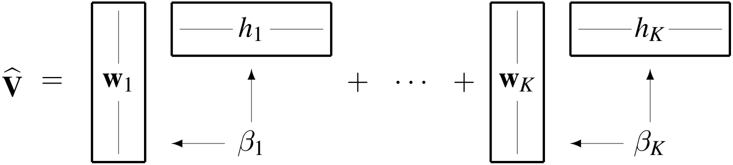


Main Idea

- Set up a statistical model.
- Place precision-like scale parameters or relevance weights

$$\boldsymbol{\beta} \triangleq (\beta_1, \ldots, \beta_K) \in \mathbb{R}_+^K$$

  on columns of $\mathbf{W}$ and rows of $\mathbf{H}$.
- Automatic Relevance Determination.



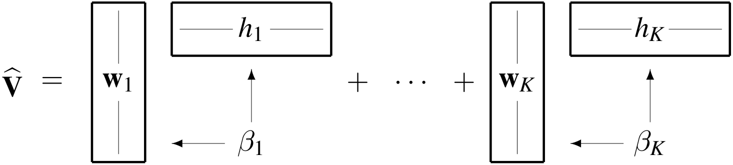
Main Idea

- Set up a statistical model.
- Place precision-like scale parameters or relevance weights

$$\boldsymbol{\beta} \triangleq (\beta_1, \ldots, \beta_K) \in \mathbb{R}_+^K$$

on columns of $\mathbf{W}$ and rows of $\mathbf{H}$.

- Automatic Relevance Determination.



- Perform inference.
- The number of β_k's that remain below a certain threshold is the model order.

Statistical Model – An Overview

Statistical Model – An Overview

- For the KL-divergence cost, find $\mathbf{W}^*, \mathbf{H}^*, \boldsymbol{\beta}^*$ such that the MAP criterion is optimized:

$$\min_{\mathbf{W},\mathbf{H},\boldsymbol{\beta}} \; C_{\mathrm{MAP}}(\mathbf{W},\mathbf{H},\boldsymbol{\beta}) \triangleq -\underbrace{\log p(\mathbf{W},\mathbf{H},\boldsymbol{\beta}|\mathbf{V})}_{\text{posterior}}.$$

Statistical Model – An Overview

- For the KL-divergence cost, find $\mathbf{W}^*, \mathbf{H}^*, \boldsymbol{\beta}^*$ such that the MAP criterion is optimized:

$$\min_{\mathbf{W},\mathbf{H},\boldsymbol{\beta}} \; C_{\mathrm{MAP}}(\mathbf{W},\mathbf{H},\boldsymbol{\beta}) \triangleq -\underbrace{\log p(\mathbf{W},\mathbf{H},\boldsymbol{\beta}|\mathbf{V})}_{\text{posterior}}.$$

where by Bayes' rule the posterior can be written as

$$-\underbrace{\log p(\mathbf{V}|\mathbf{W},\mathbf{H})}_{\text{likelihood}} - \underbrace{\log p(\mathbf{W}|\boldsymbol{\beta})}_{\text{prior on } \mathbf{W}} - \log \underbrace{p(\mathbf{H}|\boldsymbol{\beta})}_{\text{prior on } \mathbf{H}} - \log \underbrace{p(\boldsymbol{\beta}|a,b)}_{\text{prior on } \boldsymbol{\beta}}.$$

Statistical Model – An Overview

- For the KL-divergence cost, find $\mathbf{W}^*, \mathbf{H}^*, \boldsymbol{\beta}^*$ such that the MAP criterion is optimized:

$$\min_{\mathbf{W},\mathbf{H},\boldsymbol{\beta}} \quad C_{\mathrm{MAP}}(\mathbf{W}, \mathbf{H}, \boldsymbol{\beta}) \triangleq -\underbrace{\log p(\mathbf{W}, \mathbf{H}, \boldsymbol{\beta}|\mathbf{V})}_{\text{posterior}}.$$

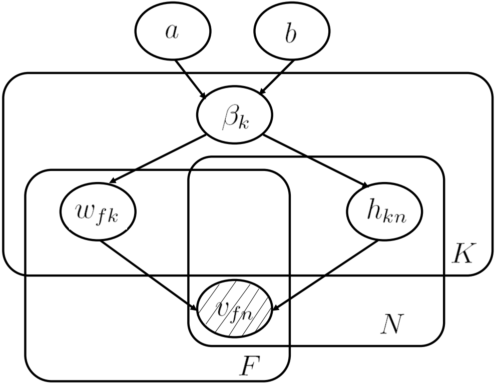
  where by Bayes' rule the posterior can be written as

$$-\underbrace{\log p(\mathbf{V}|\mathbf{W}, \mathbf{H})}_{\text{likelihood}} - \underbrace{\log p(\mathbf{W}|\boldsymbol{\beta})}_{\text{prior on } \mathbf{W}} - \log \underbrace{p(\mathbf{H}|\boldsymbol{\beta})}_{\text{prior on } \mathbf{H}} - \log \underbrace{p(\boldsymbol{\beta}|a, b)}_{\text{prior on } \boldsymbol{\beta}}.$$

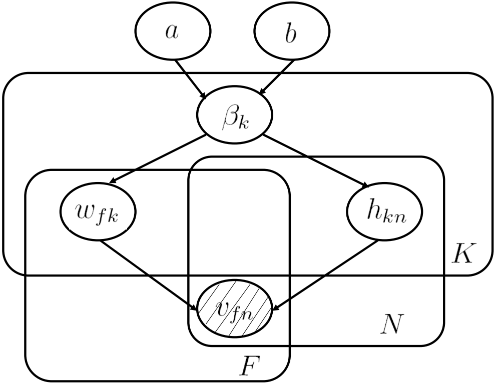
- Define the likelihood and priors and optimize $C_{\mathrm{MAP}}(\mathbf{W}, \mathbf{H}, \boldsymbol{\beta})$ efficiently.

Dependences between Variables

A Bayesian Network that describes our NMF statistical model.



A Bayesian Network that describes our NMF statistical model.



Need to specify:

- $p(\mathbf{V}|\mathbf{W}, \mathbf{H})$.
- $p(\mathbf{W}|\boldsymbol{\beta})$.
- $p(\mathbf{H}|\boldsymbol{\beta})$.
- $p(\boldsymbol{\beta}|a, b)$.

Bayesian NMF Model – Likelihood Model

- $p(\mathbf{V}|\mathbf{W}, \mathbf{H})$.
- Assume that the likelihood of an element of the matrix $\mathbf{V}$, denoted $p(v_{fn}|\hat{v}_{fn})$, is given by a Poisson with rate $\hat{v}_{fn}$.

Bayesian NMF Model – Likelihood Model

- $p(\mathbf{V}|\mathbf{W},\mathbf{H})$.
- Assume that the likelihood of an element of the matrix $\mathbf{V}$, denoted $p(v_{fn}|\hat{v}_{fn})$, is given by a Poisson with rate $\hat{v}_{fn}$.
- This corresponds to the log-likelihood of $\mathbf{V}$ given $\mathbf{W},\mathbf{H}$ as:

$$-\log p(\mathbf{V}|\mathbf{W},\mathbf{H}) \stackrel{c}{=} D_{KL}(\mathbf{V}|\mathbf{WH}).$$

- Maximizing log-likelihood $\equiv$ Minimizing KL-divergence.
- Free from hyperparameters.

Bayesian NMF Model – Prior Models on $\mathbf{W}$ and $\mathbf{H}$

- $p(\mathbf{W}|\beta)$ and $p(\mathbf{H}|\beta)$.
- Independent half-normal priors over each column k of $\mathbf{W}$ and row k of $\mathbf{H}$.

Bayesian NMF Model – Prior Models on $\mathbf{W}$ and $\mathbf{H}$

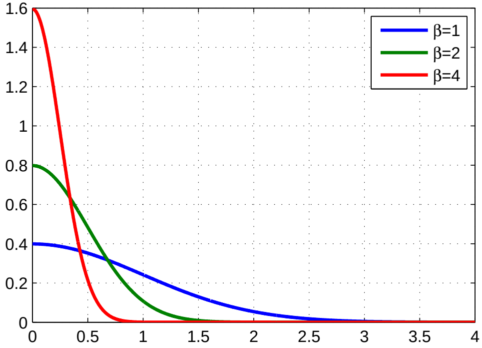
- $p(\mathbf{W}|\beta)$ and $p(\mathbf{H}|\beta)$.
- Independent half-normal priors over each column k of $\mathbf{W}$ and row k of $\mathbf{H}$.
- The priors are tied together through a single, common precision parameter β_k.

$$
\begin{aligned}
p(w_{fk}|\beta_k) &= \mathcal{HN}(w_{fk}|0, \beta_k^{-1}), \\
p(h_{kn}|\beta_k) &= \mathcal{HN}(h_{kn}|0, \beta_k^{-1}),
\end{aligned}
$$

where $\mathcal{HN}(w_{fk}|0, \beta_k^{-1})$ is the half-normal density with precision β_k.

- Least informative, High level of entropy.

Half-Normal Densities



- Half-normal densities with different precision parameters β.
- The larger the β, the "peakier" the density $\Rightarrow$ Less relevant components will be sparse.

Bayesian NMF Model – Prior Models on β

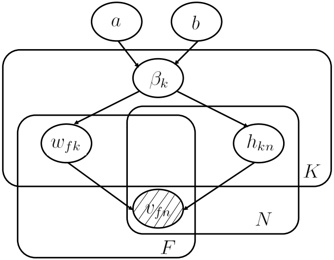
- Each precision parameter β_k is given a Gamma density:

$$p(\beta_k|a,b) = \mathcal{G}(\beta_k|a,b) = \frac{b^a}{\Gamma(a)}\beta_k^{a-1}\exp(-\beta_k b).$$

- This is the conjugate prior for β_k.

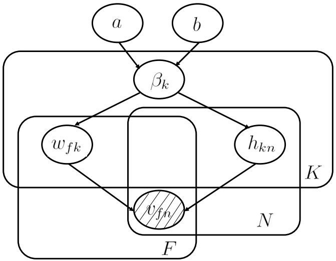
Recap: Dependences between Variables

Recap: Dependences between Variables



- $p(\mathbf{V}|\mathbf{W},\mathbf{H})$ – KL-div.
- $p(\mathbf{W}|\boldsymbol{\beta})$ – Half-normal.
- $p(\mathbf{H}|\boldsymbol{\beta})$ – Half-normal.
- $p(\boldsymbol{\beta}|a,b)$ – Gamma.

Recap: Dependences between Variables



- $p(\mathbf{V}|\mathbf{W}, \mathbf{H})$ – KL-div.
- $p(\mathbf{W}|\boldsymbol{\beta})$ – Half-normal.
- $p(\mathbf{H}|\boldsymbol{\beta})$ – Half-normal.
- $p(\boldsymbol{\beta}|a, b)$ – Gamma.

The MAP objective C_{MAP} is:

$$-\log p(\mathbf{W}, \mathbf{H}, \boldsymbol{\beta}|\mathbf{V}) \stackrel{c}{=} D_{KL}(\mathbf{V}|\mathbf{W}, \mathbf{H}) +$$

$$+\frac{1}{2}\sum_k \left[\left(\sum_f w_{fk}^2 + \sum_n h_{kn}^2 + 2b \right) \beta_k - (F + N - 2(a-1)) \log \beta_k \right].$$

Tradeoff involving the size of the β_k's.

Inference

- We have fully specified the Bayesian NMF statistical model.

Inference

- We have fully specified the Bayesian NMF statistical model.
- Inference is done using efficient multiplicative updates, which ensures positivity.
- To update a parameter θ (e.g. an element of $\mathbf{W}$)

$$\theta \leftarrow \theta \, \frac{[\nabla_\theta C_{\mathrm{MAP}}(\theta)]_+}{[\nabla_\theta C_{\mathrm{MAP}}(\theta)]_-} \, .$$

where

$$\nabla_\theta C_{\mathrm{MAP}}(\theta) = [\nabla_\theta C_{\mathrm{MAP}}(\theta)]_+ - [\nabla_\theta C_{\mathrm{MAP}}(\theta)]_- .$$

Please refer to our paper for the details.

Inference

- We have fully specified the Bayesian NMF statistical model.
- Inference is done using efficient multiplicative updates, which ensures positivity.
- To update a parameter θ (e.g. an element of $\mathbf{W}$)

$$\theta \leftarrow \theta \, \frac{[\nabla_\theta C_{\mathrm{MAP}}(\theta)]_+}{[\nabla_\theta C_{\mathrm{MAP}}(\theta)]_-} \, .$$

  where

$$\nabla_\theta C_{\mathrm{MAP}}(\theta) = [\nabla_\theta C_{\mathrm{MAP}}(\theta)]_+ - [\nabla_\theta C_{\mathrm{MAP}}(\theta)]_- .$$

  Please refer to our paper for the details.
- The model order is

$$K_{\mathrm{eff}} \triangleq |\{\beta_k \; : \; \beta_k < L\}|,$$

  and L can be found analytically.

ARD for NMF with KL-divergence cost

Input : Nonnegative data $\mathbf{V}$, fixed hyperparameters a, b.

Output : $\boldsymbol{\beta}$, K_{eff}, $\mathbf{W}$ and $\mathbf{H}$ s.t. $\mathbf{V} \approx \hat{\mathbf{V}} = \mathbf{W}\,\mathbf{H}$.

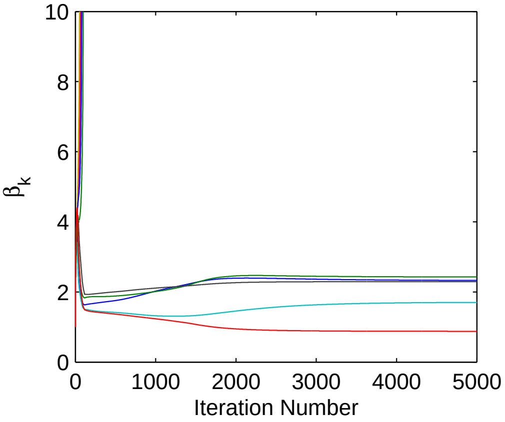
- Initialize $\mathbf{W}$ and $\mathbf{H}$ to nonnegative values.
- For $i = 1 : n_{iter}$
  - $\mathbf{H} \leftarrow \frac{\mathbf{H}}{\mathbf{W}^T \mathbf{1}_{F\times N} + \text{diag}(\boldsymbol{\beta})\mathbf{H}} \cdot \left[\mathbf{W}^T \left(\frac{\mathbf{V}}{\mathbf{W}\mathbf{H}}\right)\right]$
  - $\mathbf{W} \leftarrow \frac{\mathbf{W}}{\mathbf{1}_{F\times N}\mathbf{H}^T + \mathbf{W}\text{diag}(\boldsymbol{\beta})} \cdot \left[\left(\frac{\mathbf{V}}{\mathbf{W}\mathbf{H}}\right)\mathbf{H}^T\right]$
  - $\boldsymbol{\beta} \leftarrow \frac{F+N+2(\mathbf{a}-1)}{\mathbf{1}_{1\times F}(\mathbf{W}\cdot\mathbf{W}) + (\mathbf{H}\cdot\mathbf{H})\mathbf{1}_{N\times 1} + 2\mathbf{b}}$
- End For.
- Compute K_{eff}.

Linear in F, N, K.

A Synthetic Dataset

Generated $\mathbf{V} \in \mathbb{R}_+^{100 \times 1000}$ with effective dimensionality $K_{\text{eff}} = 5$.
Set $K = 10$ and ran inference.

Generated $\mathbf{V} \in \mathbb{R}_+^{100\times 1000}$ with effective dimensionality $K_{\text{eff}} = 5$.
Set $K = 10$ and ran inference.



- $K_{\text{eff}} = 5$ relevant components.
- $K - K_{\text{eff}} = 5$ irrelevant components.

The Swimmer Dataset

- Swimmer dataset (Donoho and Stodden 2003).
- $N = 256$ images each of size $F = 32 \times 32$.
- A figure with four moving parts (limbs), each able to exhibit four articulations.
- Fixed the shape parameter $a = 2$ and varied the scale b.

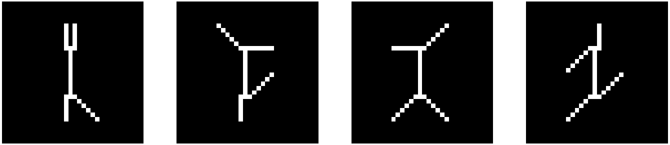


Figure: Sample images from the Swimmer dataset.

The Swimmer Dataset – Regularization Path

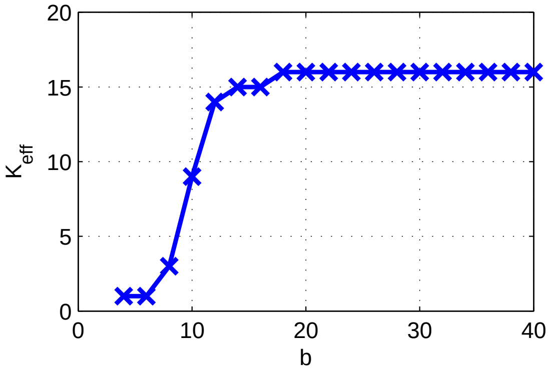


Figure: K_{eff} against b, the scale parameter of the Gamma prior.

The Swimmer Dataset – Basis Images

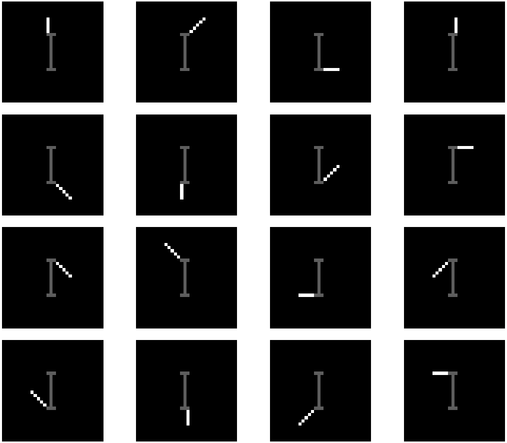


Figure: The 16 limb positions are correctly recovered.

MIT CBCL Faces Dataset $N = 2429$, $F = 19 \times 19$

The MIT CBCL Dataset – Regularization Path

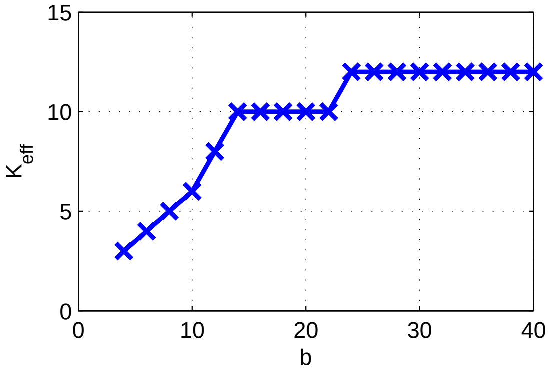


Figure: K_{eff} against b, the scale parameter of the Gamma prior.

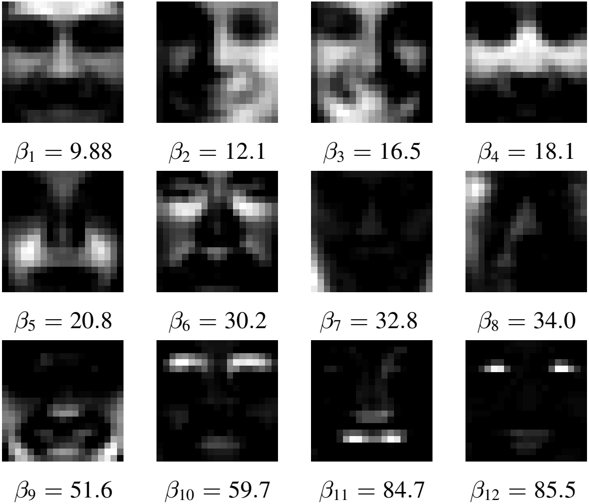


Figure: Basis images and corresponding relevance weights $\{\beta_k\}$.

Conclusions

- Bayesian approach that performs model order selection for NMF by borrowing ideas from ARD.
- Computationally cheap.

Conclusions

- Bayesian approach that performs model order selection for NMF by borrowing ideas from ARD.
- Computationally cheap.
- Identify components that are ‘relevant’ for modeling the data.
- Experiments show that we are able to recover the latent dimensionality of synthetic and real data.

Extensions

- Different cost functions (Euclidean, Itakura-Saito).
- Different prior models.
- Nonnegative Tensor Factorization.

$$\widehat{\mathbf{V}} = \sum_{k=1}^{K} \mathbf{w}_k^{(1)} \circ \mathbf{w}_k^{(2)} \circ h_k.$$

Questions and Comments

- Thank you for your kind attention.
- Matlab© Code can be found online at http://web.mit.edu/vtan/www/spars09
- Authors can be reached at
  - Vincent Tan: http://web.mit.edu/vtan/www/
  - Cédric Févotte: http://www.tsi.enst.fr/~fevotte/