Distributionally Robust and Multi-Objective Nonnegative Matrix Factorization

Vincent Y. F. Tan (NUS)

Joint work with

Nicolas Gillis, Le Thi Khanh Hien and Valentin Leplat

(Université de Mons)

Group Meeting (March 2020)

Outline

Outline

Outline

Outline

Outline

Matrix Factorization Models

Data is usually in matrix form

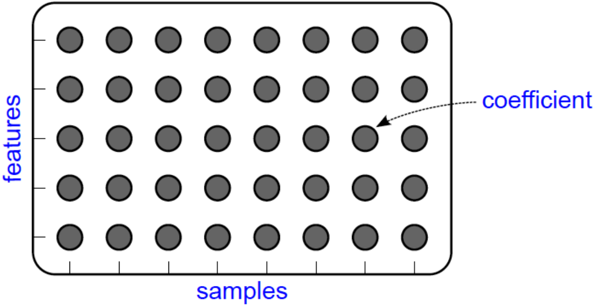


Figure reproduced from C. Févotte's slides

Matrix Factorization Models

Data is usually in matrix form

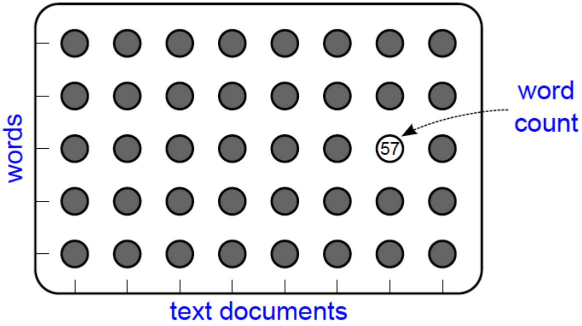


Figure reproduced from C. Févotte's slides

Matrix Factorization Models

- Dictionary Learning
- Low-Rank Approximation
- Factor Analysis
- Latent Semantic Modelling

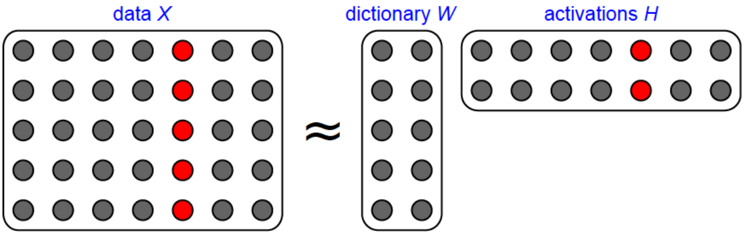


Figure reproduced from C. Févotte's slides

Non-Negative Matrix Factorization

- Non-Negative Matrix Factorization (NMF) is the task of approximating a given nonnegative matrix $X \in \mathbb{R}_+^{m \times n}$ such that

$$X \approx WH$$

where $W \in \mathbb{R}_+^{m \times r}$ and $H \in \mathbb{R}_+^{r \times n}$ are also nonnegative matrices.

Non-Negative Matrix Factorization

- Non-Negative Matrix Factorization (NMF) is the task of approximating a given nonnegative matrix $X \in \mathbb{R}_+^{m\times n}$ such that

$$X \approx WH$$

  where $W \in \mathbb{R}_+^{m\times r}$ and $H \in \mathbb{R}_+^{r\times n}$ are also nonnegative matrices.

- Usually $r \ll \min\{m, n\}$ so there is dimensionality reduction.

Non-Negative Matrix Factorization

- Non-Negative Matrix Factorization (NMF) is the task of approximating a given nonnegative matrix $X \in \mathbb{R}_+^{m\times n}$ such that

$$X \approx WH$$

where $W \in \mathbb{R}_+^{m\times r}$ and $H \in \mathbb{R}_+^{r\times n}$ are also nonnegative matrices.

- Usually $r \ll \min\{m, n\}$ so there is dimensionality reduction.
- Each column of $X(:,j) \in \mathbb{R}_+^m$ is a data point. Reconstructed via a linear combination of r basis elements given by the columns of W while the columns of H provide the weights

$$X(:,j) \approx \sum_{k=1}^{r} W(:,k)H(k,j), \qquad 1 \le j \le n$$

The Objective Function to be Minimized in NMF

- To ensure $X \approx WH$, we minimize an element-wise cost function

$$\min_{W,H \geq 0} \left[D(X, WH) = \sum_{i=1}^{m} \sum_{j=1}^{n} D\big(X_{i,j}, (WH)_{i,j}\big) \right]$$

The Objective Function to be Minimized in NMF

- To ensure $X \approx WH$, we minimize an element-wise cost function

$$\min_{W,H\geq 0}\left[D(X,WH)=\sum_{i=1}^{m}\sum_{j=1}^{n}D\big(X_{i,j},(WH)_{i,j}\big)\right]$$

- One choice for $D(\cdot,\cdot)$ is the β-divergence

$$D_\beta(x,y)=\begin{cases}\dfrac{x}{y}-\log\dfrac{x}{y}-1 & \text{for } \beta=0,\\ x\log\dfrac{x}{y}-x+y & \text{for } \beta=1,\\ \dfrac{1}{\beta(\beta-1)}\left(x^\beta+(\beta-1)y^\beta-\beta x y^{\beta-1}\right) & \text{for } \beta\neq 0,1.\end{cases}$$

The Objective Function to be Minimized in NMF

- To ensure $X \approx WH$, we minimize an element-wise cost function

$$\min_{W,H\geq 0}\left[D(X,WH)=\sum_{i=1}^{m}\sum_{j=1}^{n}D\big(X_{i,j},(WH)_{i,j}\big)\right]$$

- One choice for $D(\cdot,\cdot)$ is the β-divergence

$$D_\beta(x,y)=\begin{cases}\dfrac{x}{y}-\log\dfrac{x}{y}-1 & \text{for } \beta=0,\\ x\log\dfrac{x}{y}-x+y & \text{for } \beta=1,\\ \dfrac{1}{\beta(\beta-1)}\left(x^\beta+(\beta-1)y^\beta-\beta x y^{\beta-1}\right) & \text{for } \beta\neq 0,1.\end{cases}$$

- Note that if $\beta=2$, we have the quadratic cost $D_2(x,y)=\frac{1}{2}(x-y)^2$.

Statistical Models for NMF

- If $X_{i,j} = (WH)_{i,j} +$ Gaussian noise, then

$$-\log p(X_{i,j} \mid (WH)_{i,j}) \stackrel{c}{=} \frac{1}{2\sigma^2}\big((WH)_{i,j} - X_{i,j}\big)^2$$

then maximizing the log-likelihood $\equiv$ minimizing D_2 (Fro-NMF).

Statistical Models for NMF

- If $X_{i,j} = (WH)_{i,j} +$ Gaussian noise, then

$$-\log p(X_{i,j} \mid (WH)_{i,j}) \stackrel{c}{=} \frac{1}{2\sigma^2}\big((WH)_{i,j} - X_{i,j}\big)^2$$

then maximizing the log-likelihood $\equiv$ minimizing D_2 (Fro-NMF).

- If $X_{i,j} \sim \text{Poisson}((WH)_{i,j})$, then

$$-\log p(X_{i,j} \mid (WH)_{i,j}) = X_{i,j} \log \frac{X_{i,j}}{(WH)_{i,j}} + (WH)_{i,j} \stackrel{c}{=} D_1(X_{i,j}, (WH)_{i,j}),$$

then maximizing the log-likelihood $\equiv$ minimizing D_1 (KL-NMF).

Statistical Models for NMF

- If $X_{i,j} = (WH)_{i,j} +$ Gaussian noise, then

$$-\log p(X_{i,j} \mid (WH)_{i,j}) \stackrel{c}{=} \frac{1}{2\sigma^2}\big((WH)_{i,j} - X_{i,j}\big)^2$$

then maximizing the log-likelihood $\equiv$ minimizing D_2 (Fro-NMF).

- If $X_{i,j} \sim \mathrm{Poisson}((WH)_{i,j})$, then

$$-\log p(X_{i,j} \mid (WH)_{i,j}) = X_{i,j} \log \frac{X_{i,j}}{(WH)_{i,j}} + (WH)_{i,j} \stackrel{c}{=} D_1(X_{i,j}, (WH)_{i,j}),$$

then maximizing the log-likelihood $\equiv$ minimizing D_1 (KL-NMF).

- If $X_{i,j} = \mathrm{Gamma}(\alpha, (WH)_{i,j}/\alpha)$, then

$$-\log p(X_{i,j} \mid (WH)_{i,j}) = \frac{X_{i,j}}{(WH)_{i,j}} - \log \frac{X_{i,j}}{(WH)_{i,j}} - 1 = D_0(X_{i,j}, (WH)_{i,j}).$$

then maximizing the log-likelihood $\equiv$ minimizing D_0 (IS-NMF).

Applications, MO-NMF and DR-NMF

- Audio signal processing: $\beta \in \{0, 1\}$

Applications, MO-NMF and DR-NMF

- Audio signal processing: $\beta \in \{0, 1\}$
- Sparse document datasets: $\beta \in \{1, 2\}$

Applications, MO-NMF and DR-NMF

- Audio signal processing: $\beta \in \{0, 1\}$
- Sparse document datasets: $\beta \in \{1, 2\}$
- How to choose an appropriate β when given a new task? Say we consider $\beta \in \Omega$ where $\Omega \subset \mathbb{R}$ is a finite set.

Applications, MO-NMF and DR-NMF

- Audio signal processing: $\beta \in \{0, 1\}$
- Sparse document datasets: $\beta \in \{1, 2\}$
- How to choose an appropriate β when given a new task? Say we consider $\beta \in \Omega$ where $\Omega \subset \mathbb{R}$ is a finite set.
- Multi-Objective NMF (MO-NMF)

$$\min_{W,H \geq 0} \{D_\beta(X, WH)\}_{\beta \in \Omega}$$

which is solved for a given probability vector $\lambda = (\lambda_\beta)_{\beta \in \Omega}$ using

$$\min_{W,H \geq 0} \left[D_\Omega^\lambda(X, WH) = \sum_{\beta \in \Omega} \lambda_\beta D_\beta(X, WH) \right]$$

Applications, MO-NMF and DR-NMF

- Audio signal processing: $\beta \in \{0, 1\}$
- Sparse document datasets: $\beta \in \{1, 2\}$
- How to choose an appropriate β when given a new task? Say we consider $\beta \in \Omega$ where $\Omega \subset \mathbb{R}$ is a finite set.
- Multi-Objective NMF (MO-NMF)

$$\min_{W,H \geq 0} \{D_\beta(X, WH)\}_{\beta \in \Omega}$$

which is solved for a given probability vector $\lambda = (\lambda_\beta)_{\beta \in \Omega}$ using

$$\min_{W,H \geq 0} \left[D_\Omega^\lambda(X, WH) = \sum_{\beta \in \Omega} \lambda_\beta D_\beta(X, WH) \right]$$

- Distributionally Robust NMF (DR-NMF)

$$\min_{W,H \geq 0} \max_{\beta \in \Omega} D_\beta(X, WH)$$

Scaling of the Objective

- For the family of β-divergences, it can be easily checked that

$$D_{\beta}(\alpha X, \alpha WH) = \alpha^{\beta} D_{\beta}(X, WH).$$

Scaling of the Objective

- For the family of β-divergences, it can be easily checked that

$$D_{\beta}(\alpha X, \alpha WH) = \alpha^{\beta} D_{\beta}(X, WH).$$

- Not desirable in practice as datasets are not properly scaled.

Scaling of the Objective

- For the family of β-divergences, it can be easily checked that

$$D_\beta(\alpha X, \alpha WH) = \alpha^\beta D_\beta(X, WH).$$

- Not desirable in practice as datasets are not properly scaled.

- Compute an approximate solution

$$(W_\beta, H_\beta) \approx \arg\min_{W,H \geq 0} D_\beta(X, WH), \quad \text{with error} \quad e_\beta = D_\beta(X, W_\beta H_\beta)$$

and define

$$\bar{D}_\beta(X, WH) = \frac{D_\beta(X, WH)}{e_\beta}, \quad \text{so that} \quad \bar{D}_\beta(X, W_\beta H_\beta) = 1.$$

Scaling of the Objective

- For the family of β-divergences, it can be easily checked that

$$D_\beta(\alpha X, \alpha WH) = \alpha^\beta D_\beta(X, WH).$$

- Not desirable in practice as datasets are not properly scaled.
- Compute an approximate solution

$$(W_\beta, H_\beta) \approx \arg\min_{W,H \geq 0} D_\beta(X, WH), \quad \text{with error} \quad e_\beta = D_\beta(X, W_\beta H_\beta)$$

and define

$$\bar{D}_\beta(X, WH) = \frac{D_\beta(X, WH)}{e_\beta}, \quad \text{so that} \quad \bar{D}_\beta(X, W_\beta H_\beta) = 1.$$

- Consider the optimization

$$\min_{W,H \geq 0} \bar{D}^\lambda_\Omega(X, WH), \quad \text{where} \quad \bar{D}^\lambda_\Omega(X, WH) = \sum_{\beta \in \Omega} \lambda_\beta \bar{D}_\beta(X, WH)$$

Outline

Multiplicative Updates Algorithm

- Let $f : \mathbb{R}^d \to \mathbb{R}$ be a differentiable function. Consider the general optimization problem with nonnegativity constraints

$$\min\{f(x) \ : \ x \geq 0\}.$$

Multiplicative Updates Algorithm

- Let $f : \mathbb{R}^d \to \mathbb{R}$ be a differentiable function. Consider the general optimization problem with nonnegativity constraints

$$\min\{f(x) \ : \ x \geq 0\}.$$

- Rescaled gradient descent method (with rescaling matrix B)

$$x^+ = x - B\nabla f(x)$$

Multiplicative Updates Algorithm

- Let $f : \mathbb{R}^d \to \mathbb{R}$ be a differentiable function. Consider the general optimization problem with nonnegativity constraints

$$\min\{f(x) \ : \ x \geq 0\}.$$

- Rescaled gradient descent method (with rescaling matrix B)

$$x^+ = x - B\nabla f(x)$$

- Say that $\nabla f(x) = \nabla_+ f(x) - \nabla_- f(x)$ where $\nabla_+ f(x) > 0$ and $\nabla_- f(x) > 0$.

Multiplicative Updates Algorithm

- Let $f : \mathbb{R}^d \to \mathbb{R}$ be a differentiable function. Consider the general optimization problem with nonnegativity constraints

$$\min\{f(x) \;:\; x \geq 0\}.$$

- Rescaled gradient descent method (with rescaling matrix B)

$$x^+ = x - B\nabla f(x)$$

- Say that $\nabla f(x) = \nabla_+ f(x) - \nabla_- f(x)$ where $\nabla_+ f(x) > 0$ and $\nabla_- f(x) > 0$. Taking $B_{ii} = x_i / \nabla_+ f(x)_i$, we obtain

$$x^+ = x - \frac{[x]}{[\nabla_+ f(x)]}(\nabla_+ f(x) - \nabla_- f(x)) = x \circ \frac{\nabla_- f(x)}{\nabla_+ f(x)}$$

Multiplicative Updates Algorithm

- Let $f : \mathbb{R}^d \to \mathbb{R}$ be a differentiable function. Consider the general optimization problem with nonnegativity constraints

$$\min\{f(x) \ : \ x \geq 0\}.$$

- Rescaled gradient descent method (with rescaling matrix B)

$$x^+ = x - B\nabla f(x)$$

- Say that $\nabla f(x) = \nabla_+ f(x) - \nabla_- f(x)$ where $\nabla_+ f(x) > 0$ and $\nabla_- f(x) > 0$. Taking $B_{ii} = x_i / \nabla_+ f(x)_i$, we obtain

$$x^+ = x - \frac{[x]}{[\nabla_+ f(x)]}(\nabla_+ f(x) - \nabla_- f(x)) = x \circ \frac{\nabla_- f(x)}{\nabla_+ f(x)}$$

- No tuning of step-sizes. If $x \geq 0$, then $x^+ \geq 0$ as well.

Application of MU Algorithm to DR-NMF

- Recall that for a fixed probability vector λ, we want to solve

$$\min_{W,H\geq 0} \bar{D}^{\lambda}_{\Omega}(X,WH), \quad \text{where} \quad \bar{D}^{\lambda}_{\Omega}(X,WH) = \sum_{\beta\in\Omega} \lambda_\beta \bar{D}_\beta(X,WH)$$

Application of MU Algorithm to DR-NMF

- Recall that for a fixed probability vector λ, we want to solve

$$\min_{W,H\geq 0} \bar{D}^{\lambda}_{\Omega}(X, WH), \quad \text{where} \quad \bar{D}^{\lambda}_{\Omega}(X, WH) = \sum_{\beta\in\Omega} \lambda_\beta \bar{D}_\beta(X, WH)$$

- Alternating minimization procedure: Min over H, then over W.

Application of MU Algorithm to DR-NMF

- Recall that for a fixed probability vector λ, we want to solve

$$\min_{W,H\geq 0} \bar{D}_{\Omega}^{\lambda}(X,WH), \quad \text{where} \quad \bar{D}_{\Omega}^{\lambda}(X,WH) = \sum_{\beta\in\Omega} \lambda_\beta \bar{D}_\beta(X,WH)$$

- Alternating minimization procedure: Min over H, then over W.

- For all β,

$$\nabla^H D_\beta(X,WH) = \nabla_+^H D_\beta(X,WH) - \nabla_-^H D_\beta(X,WH),$$

where ∇^H means gradient w.r.t. H.

Application of MU Algorithm to DR-NMF

- Recall that for a fixed probability vector λ, we want to solve

$$\min_{W,H\geq 0} \bar{D}_{\Omega}^{\lambda}(X,WH), \quad \text{where} \quad \bar{D}_{\Omega}^{\lambda}(X,WH) = \sum_{\beta\in\Omega} \lambda_\beta \bar{D}_\beta(X,WH)$$

- Alternating minimization procedure: Min over H, then over W.
- For all β,

$$\nabla^H D_\beta(X,WH) = \nabla_+^H D_\beta(X,WH) - \nabla_-^H D_\beta(X,WH),$$

where ∇^H means gradient w.r.t. H.

- After some tedious calculation,

$$\nabla_+^H D_\beta(X,WH) = W^T(WH)^{\circ(\beta-1)} \quad \text{and}$$
$$\nabla_-^H D_\beta(X,WH) = W^T\left((WH)^{\circ(\beta-2)} \circ X\right),$$

Application of MU Algorithm to DR-NMF

- Update H as follows:

$$H^+ = H \circ \frac{\left[\sum_{\beta\in\Omega} \lambda_\beta \big(\nabla_-^H \bar{D}_\beta(X, WH)\big)\right]}{\left[\sum_{\beta\in\Omega} \lambda_\beta \big(\nabla_+^H \bar{D}_\beta(X, WH)\big)\right]}.$$

Application of MU Algorithm to DR-NMF

- Update H as follows:

$$H^+ = H \circ \frac{\left[\sum_{\beta\in\Omega} \lambda_\beta \left(\nabla_-^H \bar{D}_\beta(X, WH)\right)\right]}{\left[\sum_{\beta\in\Omega} \lambda_\beta \left(\nabla_+^H \bar{D}_\beta(X, WH)\right)\right]}.$$

- Sometimes this may not result in a decrease in the objective, so we set $\gamma = 1$ and $H_1^+ = H^+$ and successively find γ such that while

$$\bar{D}_\Omega^\lambda(X, WH_\gamma^+) > \bar{D}_\Omega^\lambda(X, WH)$$

we reduce

$$\gamma = \frac{\gamma}{2}$$

and set

$$H_\gamma^+ = (1-\gamma)H + \gamma H^+.$$

Algorithm for DR-NMF

- For fixed λ, we have an MU algorithm to solve

$$\min_{W,H\geq 0} \bar{D}_{\Omega}^{\lambda}(X, WH), \quad \text{where} \quad \bar{D}_{\Omega}^{\lambda}(X, WH) = \sum_{\beta\in\Omega} \lambda_{\beta}\bar{D}_{\beta}(X, WH)$$

Algorithm for DR-NMF

- For fixed λ, we have an MU algorithm to solve

$$\min_{W,H\geq 0} \bar{D}_{\Omega}^{\lambda}(X, WH), \quad \text{where} \quad \bar{D}_{\Omega}^{\lambda}(X, WH) = \sum_{\beta\in\Omega} \lambda_{\beta} \bar{D}_{\beta}(X, WH)$$

- But we want to solve for $W, H \geq 0$ that minimizes

$$\max_{\beta\in\Omega} \bar{D}_{\beta}(X, WH) = \max_{\lambda\geq 0:\|\lambda\|_1=1} \sum_{\beta\in\Omega} \lambda_{\beta} \bar{D}_{\beta}(X, WH).$$

Algorithm for DR-NMF

- For fixed λ, we have an MU algorithm to solve

$$\min_{W,H\geq 0} \bar{D}_{\Omega}^{\lambda}(X, WH), \quad \text{where} \quad \bar{D}_{\Omega}^{\lambda}(X, WH) = \sum_{\beta\in\Omega} \lambda_\beta \bar{D}_\beta(X, WH)$$

- But we want to solve for $W, H \geq 0$ that minimizes

$$\max_{\beta\in\Omega} \bar{D}_\beta(X, WH) = \max_{\lambda\geq 0:\|\lambda\|_1=1} \sum_{\beta\in\Omega} \lambda_\beta \bar{D}_\beta(X, WH).$$

- So we want to solve

$$\min_{W,H\geq 0} \max_{\lambda\geq 0:\|\lambda\|_1=1} \sum_{\beta\in\Omega} \lambda_\beta \bar{D}_\beta(X, WH)$$

which is a min-max optimization problem.

Algorithm for DR-NMF

- For fixed λ, we have an MU algorithm to solve

$$\min_{W,H\geq 0} \bar{D}_{\Omega}^{\lambda}(X, WH), \quad \text{where} \quad \bar{D}_{\Omega}^{\lambda}(X, WH) = \sum_{\beta\in\Omega} \lambda_\beta \bar{D}_\beta(X, WH)$$

- But we want to solve for $W, H \geq 0$ that minimizes

$$\max_{\beta\in\Omega} \bar{D}_\beta(X, WH) = \max_{\lambda\geq 0:\|\lambda\|_1=1} \sum_{\beta\in\Omega} \lambda_\beta \bar{D}_\beta(X, WH).$$

- So we want to solve

$$\min_{W,H\geq 0} \max_{\lambda\geq 0:\|\lambda\|_1=1} \sum_{\beta\in\Omega} \lambda_\beta \bar{D}_\beta(X, WH)$$

  which is a min-max optimization problem.

- There are dual subgradient methods to solve this with convergence guarantees, but we found them to be slow.

Aggressive, Heuristic Algorithm for DR-NMF

- Initialize $\lambda_\beta = 1/|\Omega|$ for all $\beta \in \Omega$.

Aggressive, Heuristic Algorithm for DR-NMF

- Initialize $\lambda_\beta = 1/|\Omega|$ for all $\beta \in \Omega$.
- For each $k = 1, 2, \ldots$, we obtain $H^{(k+1)}$ using the MU algorithm with $W = W^{(k)}$ and $\lambda = \lambda^{(k)}$.

Aggressive, Heuristic Algorithm for DR-NMF

- Initialize $\lambda_\beta = 1/|\Omega|$ for all $\beta \in \Omega$.
- For each $k = 1, 2, \ldots$, we obtain $H^{(k+1)}$ using the MU algorithm with $W = W^{(k)}$ and $\lambda = \lambda^{(k)}$.
- We obtain $W^{(k+1)}$ using the MU algorithm with $H = H^{(k+1)}$ and $\lambda = \lambda^{(k)}$.

Aggressive, Heuristic Algorithm for DR-NMF

- Initialize $\lambda_\beta = 1/|\Omega|$ for all $\beta \in \Omega$.
- For each $k = 1, 2, \ldots$, we obtain $H^{(k+1)}$ using the MU algorithm with $W = W^{(k)}$ and $\lambda = \lambda^{(k)}$.
- We obtain $W^{(k+1)}$ using the MU algorithm with $H = H^{(k+1)}$ and $\lambda = \lambda^{(k)}$.
- Let $\beta^* \in \arg\max_{\beta\in\Omega} \bar{D}_\beta(X, W^{(k+1)} H^{(k+1)})$ and

$$(\lambda_*^{(k)})_\beta = \begin{cases} 1 & \text{if } \beta = \beta^*, \\ 0 & \text{if } \beta \neq \beta^*. \end{cases}$$

Update

$$\lambda^{(k+1)} = \lambda^{(k)} + \underbrace{\rho_k}_{:=1/k} \lambda_*^{(k)}, \quad \text{then normalize} \quad \lambda^{(k+1)} \leftarrow \frac{\lambda^{(k+1)}}{\|\lambda^{(k+1)}\|_1}.$$

Remarks on our Algorithm for DR-NMF

- Updates for W and H are meant to approximately minimize

$$(W,H) \mapsto \bar{D}_{\Omega}^{\lambda^{(k)}}(X, WH)$$

Remarks on our Algorithm for DR-NMF

- Updates for W and H are meant to approximately minimize

$$(W, H) \mapsto \bar{D}_{\Omega}^{\lambda^{(k)}}(X, WH)$$

- For the update of λ, notice that for all $\beta \in \Omega$

$$\bar{D}_{\beta^*}(X, W^{(k+1)}H^{(k+1)}) \geq \bar{D}_{\beta}(X, W^{(k+1)}H^{(k+1)}),$$

and since $\lambda \mapsto \bar{D}_{\beta}^{\lambda}$ is linear, we have

$$\lambda_*^{(k)} = \arg\max \left\{ \bar{D}_{\beta}^{\lambda}(X, W^{(k+1)}H^{(k+1)}) : \lambda \geq 0, \|\lambda\|_1 = 1 \right\}.$$

Remarks on our Algorithm for DR-NMF

- Updates for W and H are meant to approximately minimize

$$(W,H) \mapsto \bar{D}_{\Omega}^{\lambda^{(k)}}(X, WH)$$

- For the update of λ, notice that for all $\beta \in \Omega$

$$\bar{D}_{\beta^*}(X, W^{(k+1)}H^{(k+1)}) \geq \bar{D}_{\beta}(X, W^{(k+1)}H^{(k+1)}),$$

  and since $\lambda \mapsto \bar{D}_{\beta}^{\lambda}$ is linear, we have

$$\lambda_*^{(k)} = \arg\max \left\{ \bar{D}_{\beta}^{\lambda}(X, W^{(k+1)}H^{(k+1)}) : \lambda \geq 0, \|\lambda\|_1 = 1 \right\}.$$

- The β^*-divergence is given the most importance at the next iteration

Remarks on our Algorithm for DR-NMF

- Updates for W and H are meant to approximately minimize

$$(W,H) \mapsto \bar{D}_{\Omega}^{\lambda^{(k)}}(X, WH)$$

- For the update of λ, notice that for all $\beta \in \Omega$

$$\bar{D}_{\beta^*}(X, W^{(k+1)}H^{(k+1)}) \geq \bar{D}_{\beta}(X, W^{(k+1)}H^{(k+1)}),$$

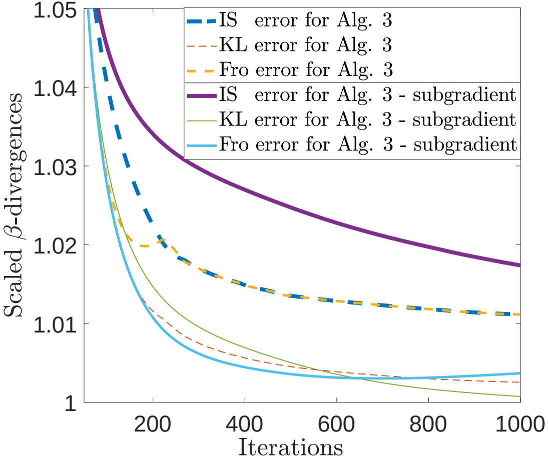
and since $\lambda \mapsto \bar{D}_{\beta}^{\lambda}$ is linear, we have

$$\lambda_*^{(k)} = \arg\max \left\{ \bar{D}_{\beta}^{\lambda}(X, W^{(k+1)}H^{(k+1)}) : \lambda \geq 0, \|\lambda\|_1 = 1 \right\}.$$

- The β^*-divergence is given the most importance at the next iteration

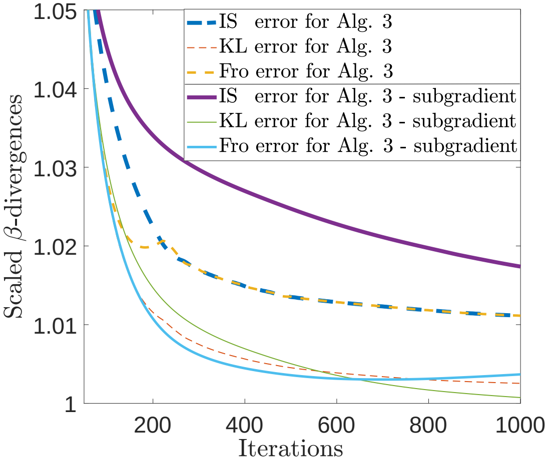
- Forcing all β-divergences to decrease as well.

Comparison to Dual-Subgradient-Based Algorithm



Evolution of the scaled β-divergences

Comparison to Dual-Subgradient-Based Algorithm



Evolution of the scaled β-divergences

Aggressive heuristic $\Longrightarrow$ $\max_{\beta\in\Omega} \bar{D}_\beta(X, W^{(k)}H^{(k)}) \leq 1+\epsilon$ faster.

Outline

Sparse Document Data Sets

- For sparse data sets, one often chooses $\beta \in \Omega = \{1, 2\}$

Sparse Document Data Sets

- For sparse data sets, one often chooses $\beta \in \Omega = \{1, 2\}$
- For sparse word-count datasets, Poisson noise is the most appropriate

Sparse Document Data Sets

- For sparse data sets, one often chooses $\beta \in \Omega = \{1, 2\}$
- For sparse word-count datasets, Poisson noise is the most appropriate
- But say we don't know this, we can compare DR-NMF, KL-NMF and Fro-NMF

Sparse Document Data Sets

- For sparse data sets, one often chooses $\beta \in \Omega = \{1, 2\}$
- For sparse word-count datasets, Poisson noise is the most appropriate
- But say we don't know this, we can compare DR-NMF, KL-NMF and Fro-NMF
- Use these NMF methods for clustering (topic modeling)

Sparse Document Data Sets

- For sparse data sets, one often chooses $\beta \in \Omega = \{1, 2\}$
- For sparse word-count datasets, Poisson noise is the most appropriate
- But say we don't know this, we can compare DR-NMF, KL-NMF and Fro-NMF
- Use these NMF methods for clustering (topic modeling)
- Clustering accuracy

$$\text{accuracy}(\{\tilde{C}_i\}_{i=1}^r) := \min_{\pi:[r]\to[r]} \frac{1}{m} \sum_{i=1}^{r} |C_i \cap \tilde{C}_{\pi(i)}|$$

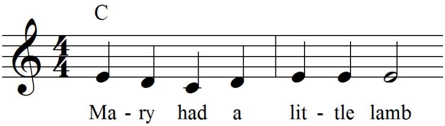
Sparse Document Data Sets

data set	number of classes	Clustering accuracy (%) KL-NMF	Fro-NMF	DR-NMF
NG20	20	**50.15**	17.78	<u>27.60</u>
NG3SIM	3	<u>59.07</u>	34.29	**68.05**
classic	4	**65.53**	49.21	<u>58.98</u>
ohscal	10	**41.54**	35.71	<u>40.23</u>
k1b	6	54.40	**73.50**	<u>62.35</u>
hitech	6	41.03	**48.28**	<u>41.68</u>
reviews	5	**78.10**	45.24	<u>75.33</u>
sports	7	<u>53.48</u>	49.24	**62.60**
la1	6	**70.69**	45.47	<u>66.67</u>
la12	6	**71.24**	47.91	<u>67.75</u>
la2	6	**70.34**	51.58	<u>68.62</u>
tr11	9	**52.90**	46.38	<u>46.62</u>
tr23	6	30.39	**39.71**	<u>34.80</u>
tr41	10	**60.25**	35.31	<u>49.20</u>
tr45	10	**56.67**	<u>38.12</u>	31.59
Average		57.05	43.85	53.47

Clustering accuracies of various methods

Dense Time-Frequency Matrices of Audio Signals

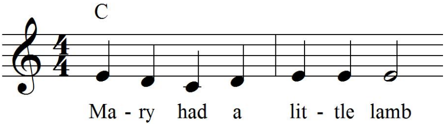
- Use the data set piano_Mary



Musical score of "Mary had a little lamb". The notes are activated as follows: E_4, D_4, C_4, D_4, E_4, E_4, E_4.

Dense Time-Frequency Matrices of Audio Signals

- Use the data set piano_Mary

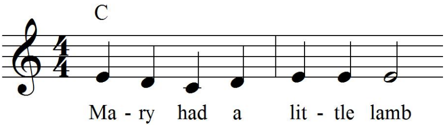


Musical score of "Mary had a little lamb". The notes are activated as follows: E_4, D_4, C_4, D_4, E_4, E_4, E_4.

- Considered no added noise and adding Poisson noise to the music piece

Dense Time-Frequency Matrices of Audio Signals

- Use the data set piano_Mary

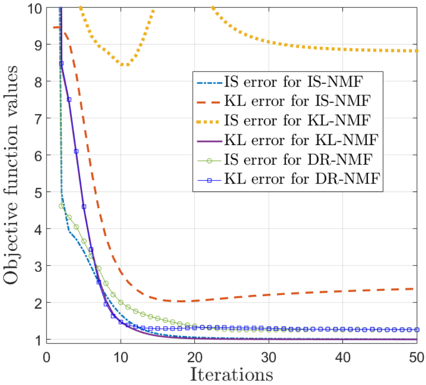


Musical score of "Mary had a little lamb". The notes are activated as follows: E_4, D_4, C_4, D_4, E_4, E_4, E_4.

- Considered no added noise and adding Poisson noise to the music piece

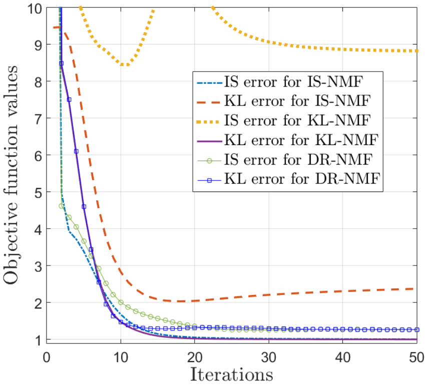
- Tested in DR-NMF (with $\Omega = \{0, 1\}$), IS-NMF ($\beta = 0$) and KL-NMF ($\beta = 1$)

No Added Noise



Evolution of scaled β-divergences

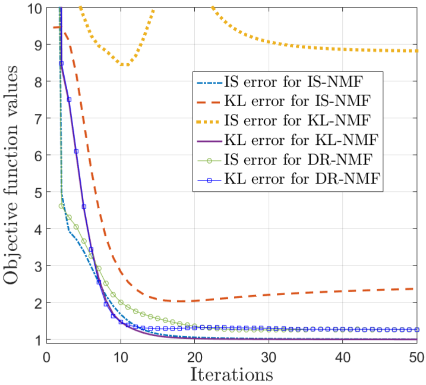
No Added Noise



Evolution of scaled β-divergences

- DR-NMF is able to compute a model with low IS- and KL-error

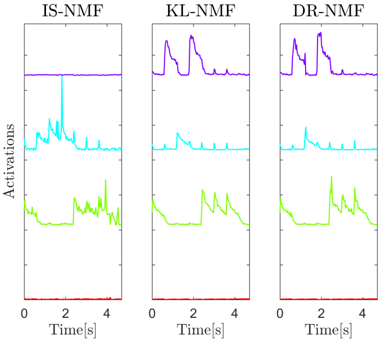
No Added Noise



Evolution of scaled β-divergences

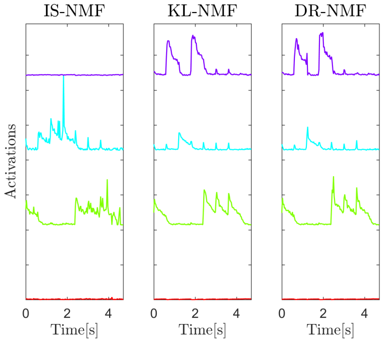
- DR-NMF is able to compute a model with low IS- and KL-error
- KL-NMF has IS-error 9 times that of IS-NMF

Added Poisson Noise



Comparative study of NMF with IS- and KL-divergences, and DR-NMF with $\Omega = \{0, 1\}$ and Poisson noise.

Added Poisson Noise



Comparative study of NMF with IS- and KL-divergences, and DR-NMF with $\Omega = \{0, 1\}$ and Poisson noise.

- Rows of H recovered successfully.
- C_4 is activated once, D_4 twice and E_4 four times.

Outline

Conclusion and Future Work

- Proposed a Multi-Objective and Distributionally Robust variant of NMF

Conclusion and Future Work

- Proposed a Multi-Objective and Distributionally Robust variant of NMF
- Works exceedingly well in practice (audio, document data sets) without knowledge of β

Conclusion and Future Work

- Proposed a Multi-Objective and Distributionally Robust variant of NMF
- Works exceedingly well in practice (audio, document data sets) without knowledge of β
- Prove convergence guarantees for our algorithm (there are convergence guarantees for the slow dual subgradient method)

Conclusion and Future Work

- Proposed a Multi-Objective and Distributionally Robust variant of NMF
- Works exceedingly well in practice (audio, document data sets) without knowledge of β
- Prove convergence guarantees for our algorithm (there are convergence guarantees for the slow dual subgradient method)
- Full paper here (https://arxiv.org/abs/1901.10757).