22-27 MAY 2022

ICASSP 2022

SINGAPORE

IEEE Signal Processing Society

IEEE

T16: Recent advances in Nonnegative Matrix Factorization

Instructions to Zoom participants

zoom

Mute mic at all times

Turn off video at all times

Enter questions in chat box

T16: Recent advances in Nonnegative Matrix Factorization (Part 1)

Cédric Févotte

Recent advances in nonnegative matrix factorization

Part I: Generalities, optimization, regularization

Cédric Févotte
CNRS, Toulouse, France

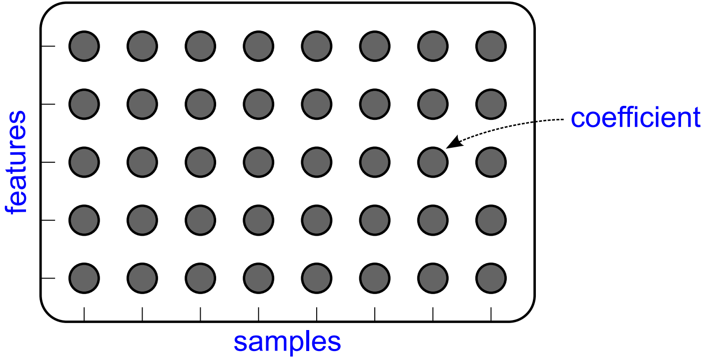
Vincent Y. F. Tan
National University of Singapore

ICASSP Tutorial
Singapore, May 2022

Outline

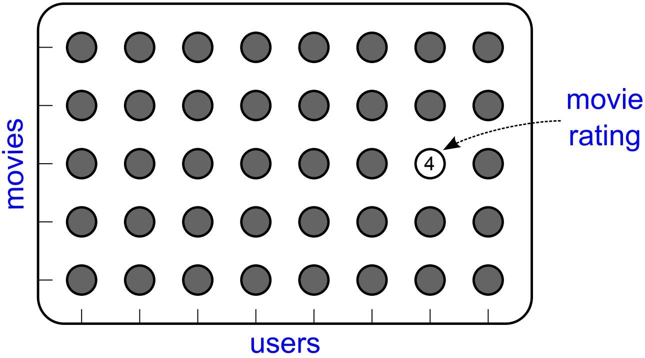
Matrix factorization models

Data often available in matrix form.



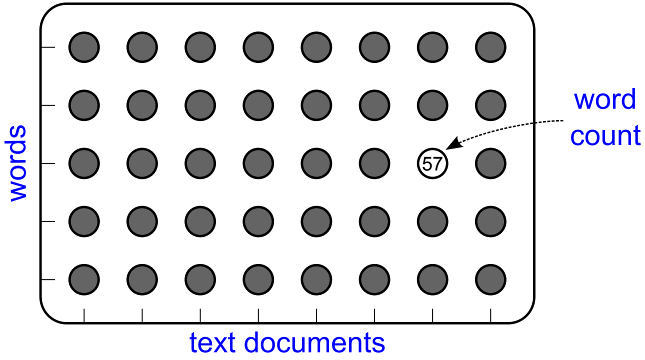
Matrix factorization models

Data often available in matrix form.



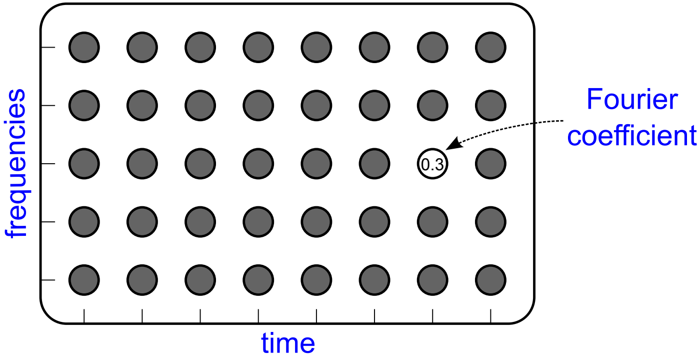
Matrix factorization models

Data often available in matrix form.



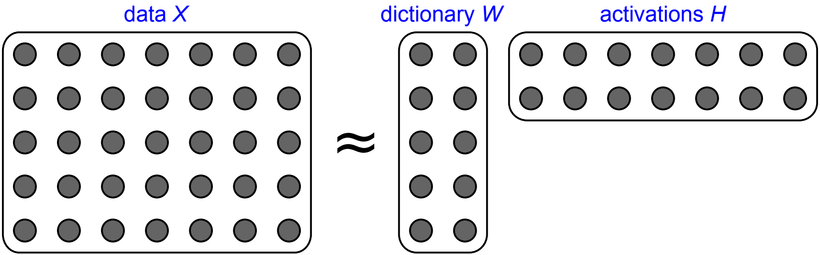
Matrix factorization models

Data often available in matrix form.



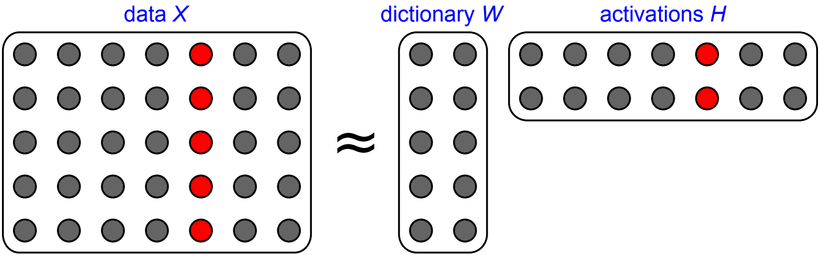
Matrix factorization models

$\approx$ **dictionary learning**
low-rank approximation
factor analysis
latent semantic analysis



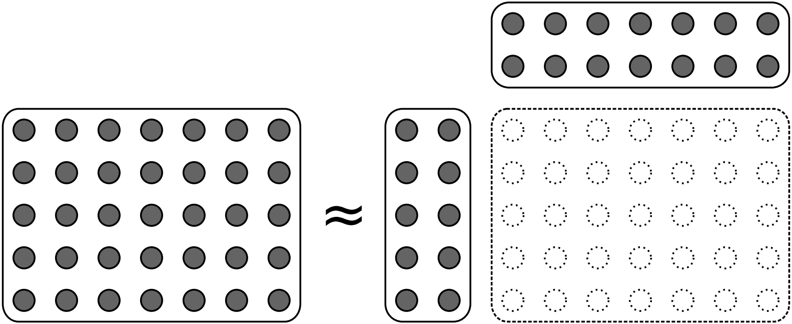
Matrix factorization models

$\approx$ **dictionary learning**
low-rank approximation
factor analysis
latent semantic analysis



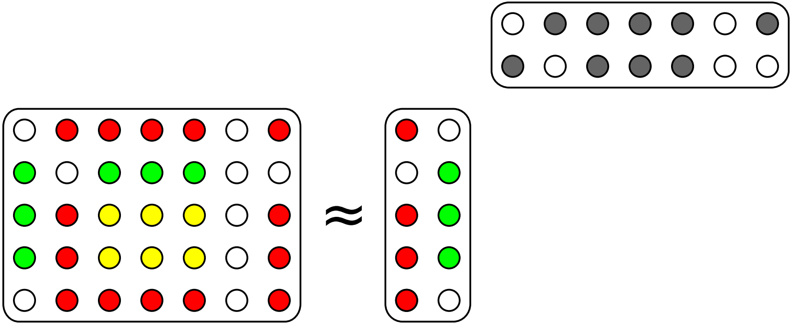
Matrix factorization models

for dimensionality reduction (coding, low-dimensional embedding)



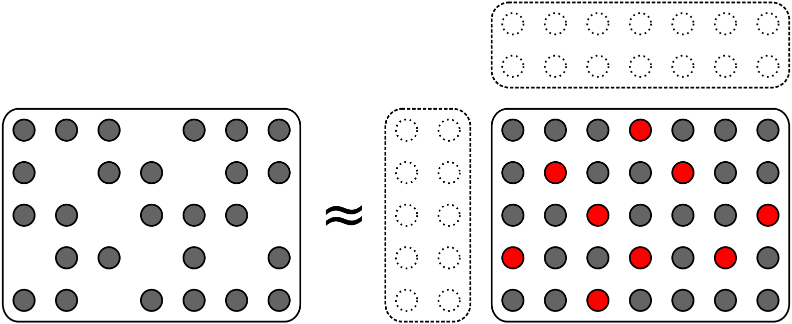
Matrix factorization models

for unmixing (source separation, latent topic discovery)



Matrix factorization models

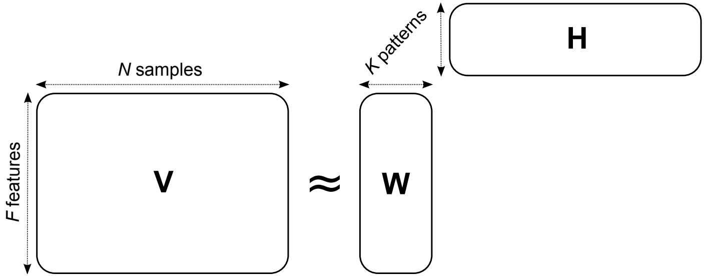
for interpolation (collaborative filtering, image inpainting)



Matrix factorization models

- simple generative & interpretable models, popular in unsupervised settings.
- used in many fields for a long time:
  - Principal component analysis PCA (Pearson, 1901)
  - Factor analysis (Spearman, 1904)
  - Latent semantic analysis LSA (Deerwester et al., 1988)
  - Independent component analysis ICA (Comon, 1994)
  - Nonnegative matrix factorization NMF (Lee & Seung, 1999)
  - Latent Dirichlet allocation LDA (Blei et al., 2003)
  - Sparse dictionary learning, e.g., K-SVD (Aharon et al., 2006)
- active topics:
  - design of nonconvex optimization algorithms with proven convergence
  - landscape analysis, search for global optima
  - conditions for identifiability
  - rank selection
  - probabilistic models & statistical approaches (e.g., integer-valued or binary data)

Nonnegative matrix factorization

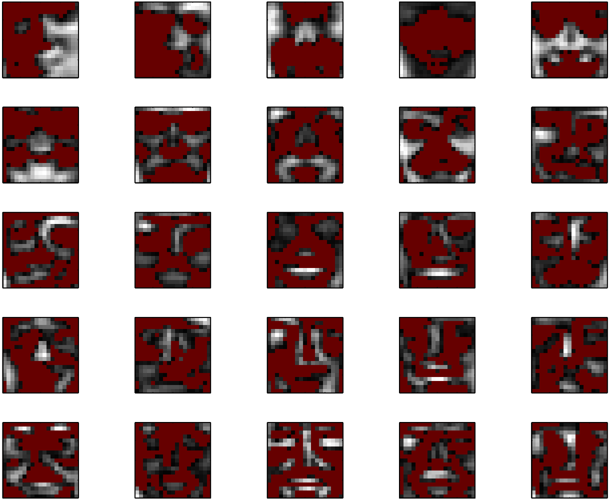


- data **V** and factors **W**, **H** have nonnegative entries.
- nonnegativity of **W** ensures interpretability of the dictionary, because patterns $\mathbf{w}_k$ and samples $\mathbf{v}_n$ belong to the same space.
- nonnegativity of **H** tends to produce part-based representations, because subtractive combinations are forbidden.

Early work by (Paatero and Tapper, 1994), landmark *Nature* paper by (Lee and Seung, 1999)

49 images among 2429 from MIT's CBCL face dataset

PCA dictionary with $K = 25$



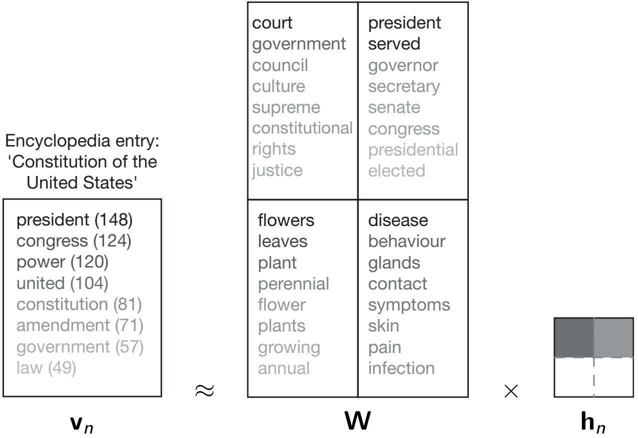
red pixels indicate negative values

NMF dictionary with $K = 25$

experiment reproduced from (Lee and Seung, 1999)

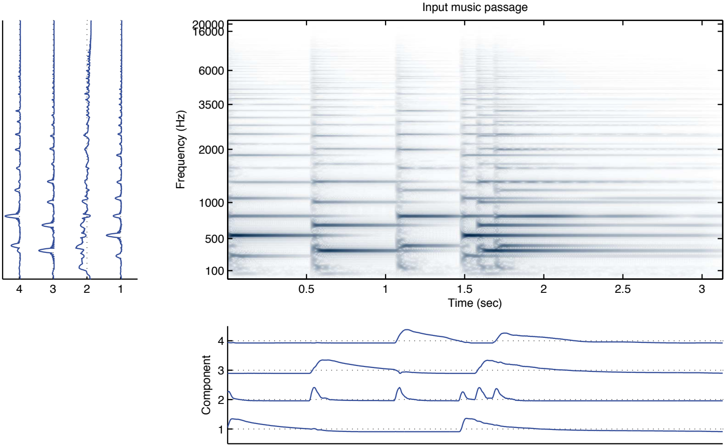
NMF for latent semantic analysis

(Lee and Seung, 1999; Hofmann, 1999)



reproduced from (Lee and Seung, 1999)

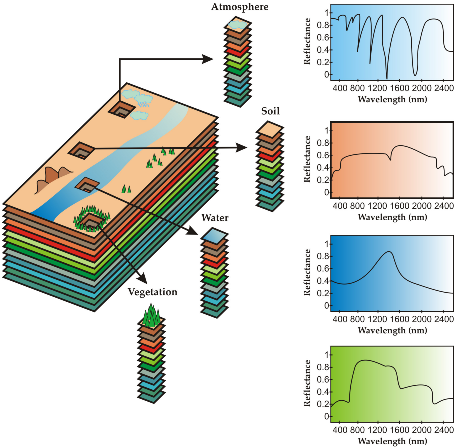
NMF for audio spectral unmixing
(Smaragdis and Brown, 2003)



reproduced from (Smaragdis, 2013)

NMF for hyperspectral unmixing

(Berry, Browne, Langville, Pauca, and Plemmons, 2007)



reproduced from (Bioucas-Dias et al., 2012)

Outline

NMF as a constrained minimization problem

Minimize a measure of fit between $\mathbf{V}$ and $\mathbf{WH}$, subject to nonnegativity:

$$\min_{\mathbf{W},\mathbf{H}\geq\mathbf{0}} D(\mathbf{V}|\mathbf{WH}) = \sum_{fn} d([\mathbf{V}]_{fn}|[\mathbf{WH}]_{fn}),$$

where $d(x|y)$ is a scalar cost function, e.g.,

- squared Euclidean distance (Paatero and Tapper, 1994; Lee and Seung, 2001)
- Kullback-Leibler divergence (Lee and Seung, 1999; Finesso and Spreij, 2006)
- Itakura-Saito divergence (Févotte, Bertin, and Durrieu, 2009)
- α-divergence (Cichocki et al., 2008)
- β-divergence (Cichocki et al., 2006; Févotte and Idier, 2011)
- Bregman divergences (Dhillon and Sra, 2005)
- and more in (Yang and Oja, 2011)

Regularization terms often added to $D(\mathbf{V}|\mathbf{WH})$ for sparsity, smoothness, etc.
Nonconvex problem.

Probabilistic models

- Let $\mathbf{V} \sim p(\mathbf{V}|\mathbf{WH})$ such that
  - $\mathrm{E}[\mathbf{V}|\mathbf{WH}] = \mathbf{WH}$
  - $p(\mathbf{V}|\mathbf{WH}) = \prod_{fn} p(v_{fn}|[\mathbf{WH}]_{fn})$
- then the following correspondences apply with

$$D(\mathbf{V}|\mathbf{WH}) = -\log p(\mathbf{V}|\mathbf{WH}) + \text{cst}$$

data support	distribution/noise	divergence	examples
real-valued	additive Gaussian	quadratic loss	many
integer	multinomial*	weighted KL	word counts
integer	Poisson	generalized KL	photon counts
nonnegative	multiplicative Gamma	Itakura-Saito	spectrogram
generally nonnegative	Tweedie	β-divergence	generalizes above models

*conditional independence over f does not apply

The β-divergence

A popular measure of fit in NMF (Basu et al., 1998; Cichocki and Amari, 2010)

$$d_\beta(x|y) \stackrel{\text{def}}{=} \begin{cases} \frac{1}{\beta\,(\beta-1)} \left(x^\beta + (\beta-1)\,y^\beta - \beta\,x\,y^{\beta-1}\right) & \beta \in \mathbb{R}\backslash\{0,1\} \\ x \log \frac{x}{y} + (y-x) & \beta = 1 \\ \frac{x}{y} - \log \frac{x}{y} - 1 & \beta = 0 \end{cases}$$

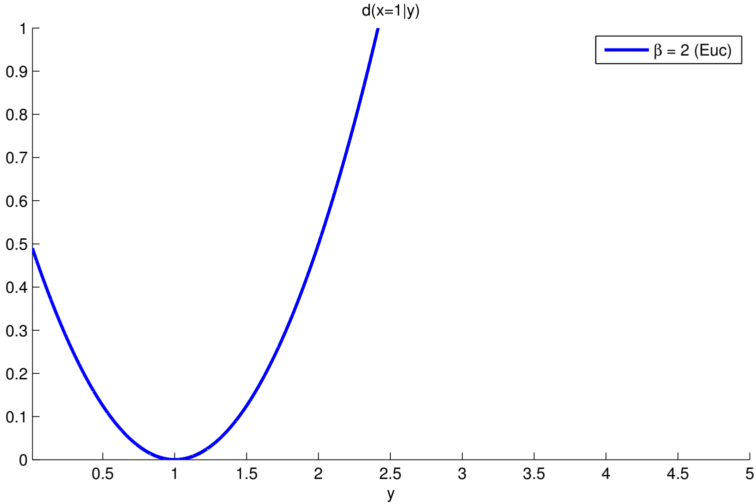
Special cases:

- squared Euclidean distance / quadratic loss ($\beta = 2$)
- generalized Kullback-Leibler (KL) divergence ($\beta = 1$)
- Itakura-Saito (IS) divergence ($\beta = 0$)

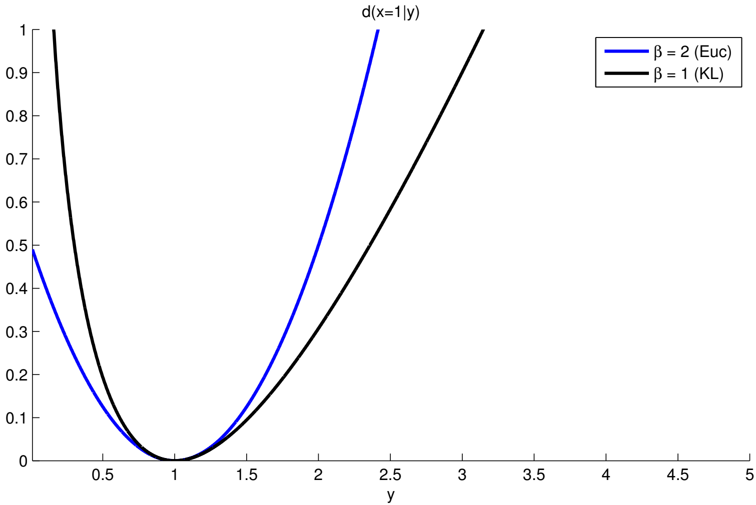
Properties:

- Homogeneity: $d_\beta(\lambda x|\lambda y) = \lambda^\beta d_\beta(x|y)$
- $d_\beta(x|y)$ is a convex function of y for $1 \leq \beta \leq 2$
- Bregman divergence

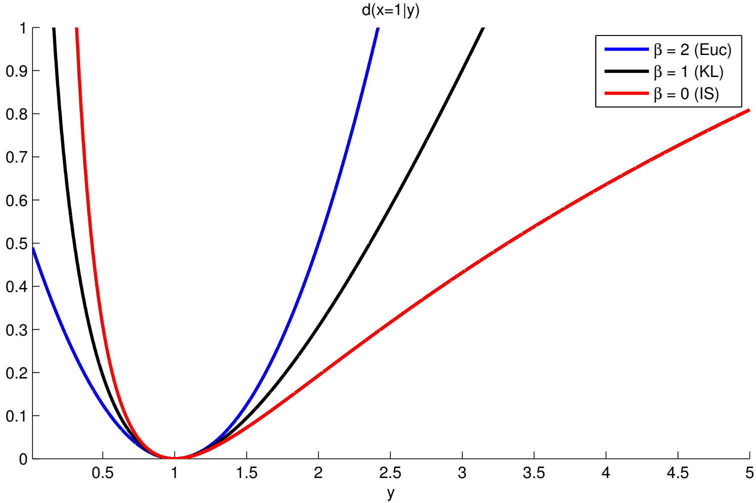
The β-divergence



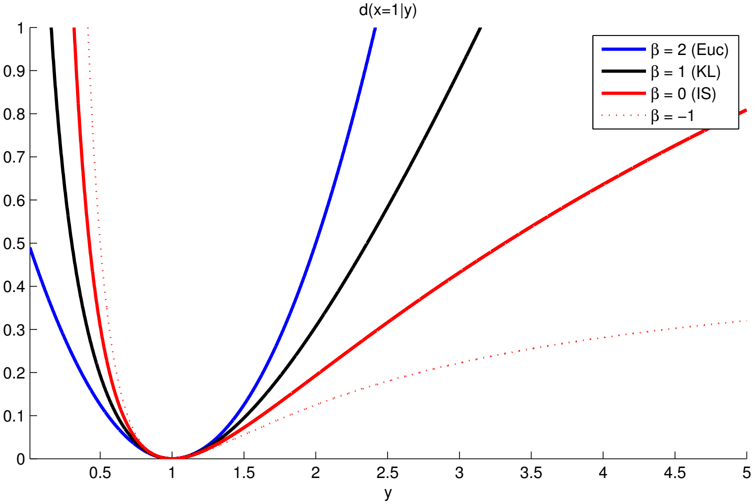
The β-divergence



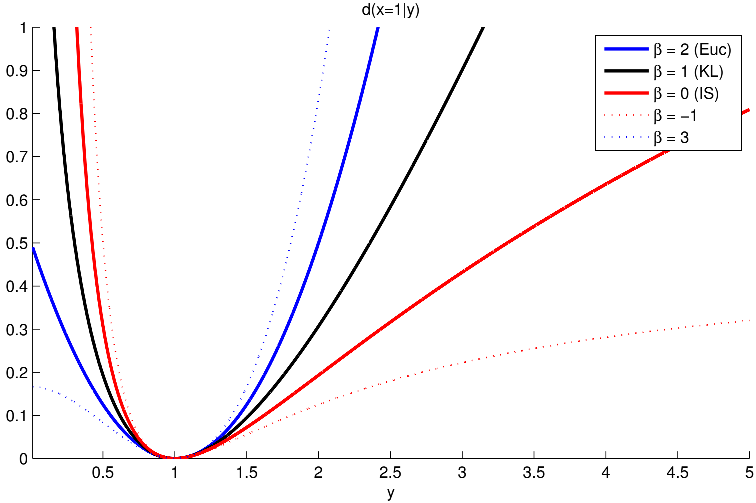
The β-divergence



The β-divergence



The β-divergence



A common NMF algorithm design: alternating methods

- Block-coordinate update of $\mathbf{H}$ given $\mathbf{W}^{(i-1)}$ and $\mathbf{W}$ given $\mathbf{H}^{(i)}$.
- Updates of $\mathbf{W}$ and $\mathbf{H}$ equivalent by transposition:

$$\mathbf{V} \approx \mathbf{W}\mathbf{H} \Leftrightarrow \mathbf{V}^T \approx \mathbf{H}^T\mathbf{W}^T$$

- Objective function separable in the columns of $\mathbf{H}$ or the rows of $\mathbf{W}$:

$$D(\mathbf{V}|\mathbf{W}\mathbf{H}) = \sum_n D(\mathbf{v}_n|\mathbf{W}\mathbf{h}_n)$$

- Essentially left with nonnegative linear regression:

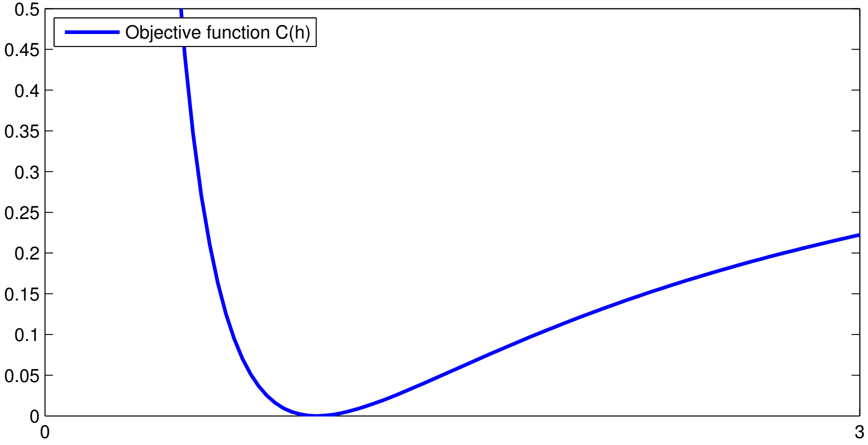
$$\min_{\mathbf{h} \geq \mathbf{0}} C(\mathbf{h}) \stackrel{\text{def}}{=} D(\mathbf{v}|\mathbf{W}\mathbf{h})$$

  Numerous references in the image restoration literature, e.g., (Richardson, 1972; Lucy, 1974; Daube-Witherspoon and Muehllehner, 1986; De Pierro, 1993)

Block-descent algorithm, nonconvex problem, initialization is an issue.

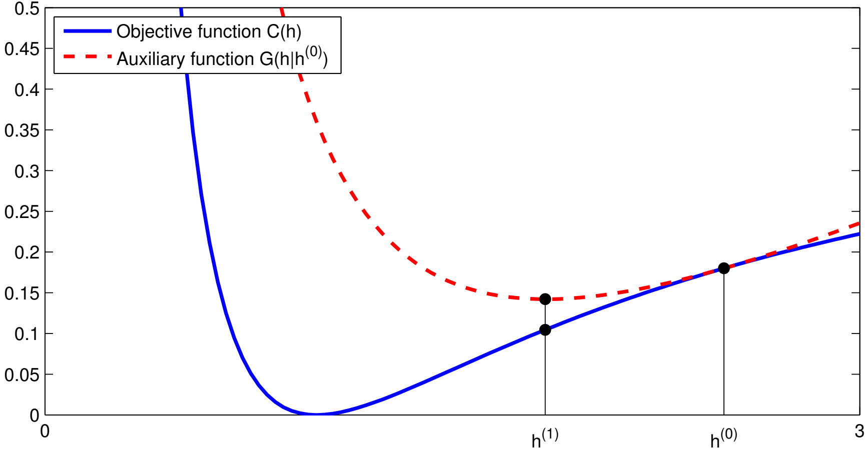
Majorization-minimization (MM)

Build $G(\mathbf{h}|\tilde{\mathbf{h}})$ such that $G(\mathbf{h}|\tilde{\mathbf{h}}) \geq C(\mathbf{h})$ and $G(\tilde{\mathbf{h}}|\tilde{\mathbf{h}}) = C(\tilde{\mathbf{h}})$.
Optimize (iteratively) $G(\mathbf{h}|\tilde{\mathbf{h}})$ instead of $C(\mathbf{h})$.



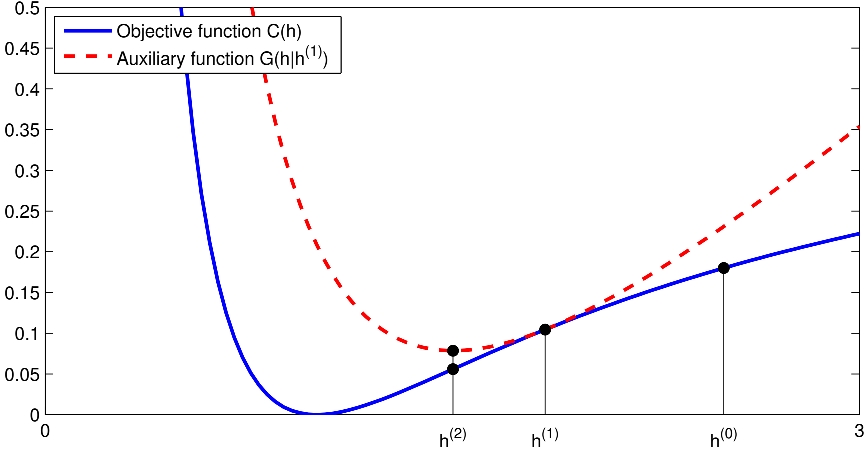
Majorization-minimization (MM)

Build $G(\mathbf{h}|\tilde{\mathbf{h}})$ such that $G(\mathbf{h}|\tilde{\mathbf{h}}) \geq C(\mathbf{h})$ and $G(\tilde{\mathbf{h}}|\tilde{\mathbf{h}}) = C(\tilde{\mathbf{h}})$.
Optimize (iteratively) $G(\mathbf{h}|\tilde{\mathbf{h}})$ instead of $C(\mathbf{h})$.



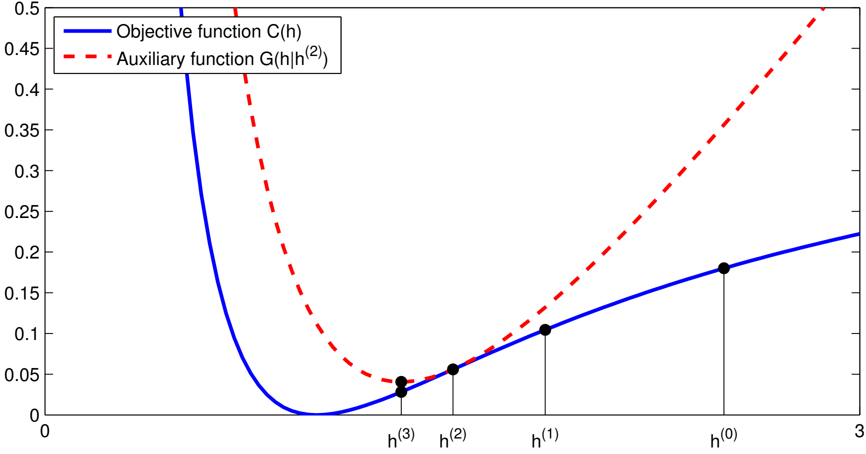
Majorization-minimization (MM)

Build $G(\mathbf{h}|\tilde{\mathbf{h}})$ such that $G(\mathbf{h}|\tilde{\mathbf{h}}) \geq C(\mathbf{h})$ and $G(\tilde{\mathbf{h}}|\tilde{\mathbf{h}}) = C(\tilde{\mathbf{h}})$.
Optimize (iteratively) $G(\mathbf{h}|\tilde{\mathbf{h}})$ instead of $C(\mathbf{h})$.



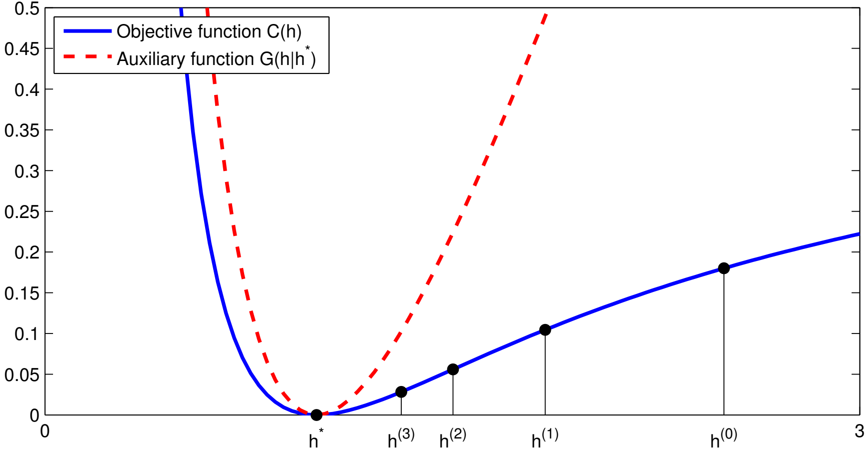
Majorization-minimization (MM)

Build $G(\mathbf{h}|\tilde{\mathbf{h}})$ such that $G(\mathbf{h}|\tilde{\mathbf{h}}) \geq C(\mathbf{h})$ and $G(\tilde{\mathbf{h}}|\tilde{\mathbf{h}}) = C(\tilde{\mathbf{h}})$.
Optimize (iteratively) $G(\mathbf{h}|\tilde{\mathbf{h}})$ instead of $C(\mathbf{h})$.



Majorization-minimization (MM)

Build $G(\mathbf{h}|\tilde{\mathbf{h}})$ such that $G(\mathbf{h}|\tilde{\mathbf{h}}) \geq C(\mathbf{h})$ and $G(\tilde{\mathbf{h}}|\tilde{\mathbf{h}}) = C(\tilde{\mathbf{h}})$.
Optimize (iteratively) $G(\mathbf{h}|\tilde{\mathbf{h}})$ instead of $C(\mathbf{h})$.



Majorization-minimization (MM)

- Finding a good & workable local majorization is the crucial point.
- Treating convex and concave terms separately with Jensen and tangent inequalities usually works. E.g.:

$$C_{\text{IS}}(\mathbf{h}) = \left[\sum_f \frac{v_f}{\sum_k w_{fk} h_k}\right] + \left[\sum_f \log\left(\sum_k w_{fk} h_k\right)\right] + cst$$

Majorization-minimization (MM)

- Finding a good & workable local majorization is the crucial point.
- Treating convex and concave terms separately with Jensen and tangent inequalities usually works. E.g.:

$$C_{\text{IS}}(\mathbf{h}) = \left[\sum_f \frac{v_f}{\sum_k w_{fk} h_k}\right] + \left[\sum_f \log\left(\sum_k w_{fk} h_k\right)\right] + cst$$

- In most cases, leads to nonnegativity-preserving multiplicative algorithms:

$$h_k = \tilde{h}_k \left(\frac{\nabla^-_{h_k} C(\tilde{\mathbf{h}})}{\nabla^+_{h_k} C(\tilde{\mathbf{h}})}\right)^\gamma$$

  - $\nabla_{h_k} C(\mathbf{h}) = \nabla^+_{h_k} C(\mathbf{h}) - \nabla^-_{h_k} C(\mathbf{h})$ and the two summands are nonnegative.
  - if $\nabla_{h_k} C(\tilde{\mathbf{h}}) > 0$, ratio of summands < 1 and h_k decreases.
  - γ is a divergence-specific scalar exponent.
- Details in (Nakano et al., 2010; Févotte and Idier, 2011; Yang and Oja, 2011)

Example: derivation for the Itakura-Saito divergence

- IS divergence ($\beta = 0$)

$$d_{\mathrm{IS}}(x|y) = \frac{x}{y} - \log\frac{x}{y} - 1$$

- Nonnegative linear regression with the IS divergence

$$\begin{aligned}\min_{\mathbf{h}\geq 0} C_{\mathrm{IS}}(\mathbf{h}) &= \sum_f d_{\mathrm{IS}}(v_f|[\mathbf{Wh}]_f) \\ &= \underbrace{\left[\sum_f \frac{v_f}{\sum_k w_{fk} h_k}\right]}_{C_1(\mathbf{h})\ \text{(convex)}} + \underbrace{\left[\sum_f \log\left(\sum_k w_{fk} h_k\right)\right]}_{C_2(\mathbf{h})\ \text{(concave)}} + cst\end{aligned}$$

Example: derivation for the Itakura-Saito divergence

- Majorization of $C_1(\mathbf{h})$ with Jensen's inequality.
  Let $f(x)$ be a convex function and $\boldsymbol{\lambda} \in \mathbb{R}_+^K$ with $\sum_k \lambda_k = 1$. Then:
$$f\left(\sum\nolimits_k \lambda_k h_k\right) \leq \sum\nolimits_k \lambda_k f(h_k).$$
- Let $\tilde{\mathbf{h}} \in \mathbb{R}_+^K$ be the current estimate, $\tilde{\mathbf{v}} = \mathbf{W}\tilde{\mathbf{h}}$ be the current approximation and
$$\lambda_{fk} = \frac{w_{fk}\tilde{h}_k}{\tilde{v}_f} = \frac{w_{fk}\tilde{h}_k}{\sum_j w_{fj}\tilde{h}_j} \quad \left(\text{note that } \sum\nolimits_k \lambda_{fk} = 1\right).$$
- Then, by convexity of $f(x) = x^{-1}$, we may write:
$$C_{\mathrm{IS}}(\mathbf{h}) = \sum\nolimits_f v_f \left(\sum\nolimits_k w_{fk} h_k\right)^{-1} = \sum\nolimits_f v_f \left(\sum\nolimits_k \lambda_{fk} \frac{w_{fk} h_k}{\lambda_{fk}}\right)^{-1}$$
$$\leq \sum\nolimits_{fk} v_f \frac{\lambda_{fk}^2}{w_{fk} h_k} = \sum\nolimits_{fk} w_{fk} \frac{v_f}{\tilde{v}_f^2} \frac{\tilde{h}_k^2}{h_k} = G_1(\mathbf{h}|\tilde{\mathbf{h}}).$$

Example: derivation for the Itakura-Saito divergence

- Majorization of $C_2(\mathbf{h})$ with the tangent inequality.
  Let $g(\mathbf{h})$ be a concave function then:

$$g(\mathbf{h}) \leq g(\tilde{\mathbf{h}}) + \nabla g(\tilde{\mathbf{h}})^\top (\mathbf{h} - \tilde{\mathbf{h}}) = \sum_k [\nabla g(\tilde{\mathbf{h}})]_k h_k + cst.$$

- Given $C_2(\mathbf{h}) = \sum_f \log \left(\sum_k w_{fk} h_k\right)$, we have:

$$[\nabla C_2(\tilde{\mathbf{h}})]_k = \nabla_{h_k} C_2(\tilde{\mathbf{h}}) = \sum_f \frac{w_{fk}}{\tilde{v}_f}.$$

- Finally, we may majorize $C_2(\mathbf{h})$ with:

$$G_2(\mathbf{h}|\tilde{\mathbf{h}}) = \sum_{fk} \frac{w_{fk}}{\tilde{v}_f} h_k + cst.$$

Example: derivation for the Itakura-Saito divergence

- In the end, we may majorize $C_{\text{IS}}(\mathbf{h})$ with:

$$\begin{aligned} G(\mathbf{h}|\tilde{\mathbf{h}}) &= G_1(\mathbf{h}|\tilde{\mathbf{h}}) + G_2(\mathbf{h}|\tilde{\mathbf{h}}) + cst \\ &= \sum_{fk} w_{fk} \left[\frac{v_f}{\tilde{v}_f^2} \frac{\tilde{h}_k^2}{h_k} + \frac{1}{\tilde{v}_f} h_k \right] + cst. \end{aligned}$$

- Smooth, convex and separable majorizer. Easily minimized by cancelling its gradient, leading to the MM-based multiplicative update

$$h_k = \tilde{h}_k \left(\frac{\sum_f w_{fk} v_f [\mathbf{W}\tilde{\mathbf{h}}]_f^{-2}}{\sum_f w_{fk} [\mathbf{W}\tilde{\mathbf{h}}]_f^{-1}} \right)^{\frac{1}{2}}.$$

- Algorithm known from (Cao et al., 1999). The $\frac{1}{2}$ exponent can be dropped using majorization-equalization (Févotte and Idier, 2011).

The multiplicative updates (MU) for NMF with β-divergence

- Alternating updates of $\mathbf{W}$ and $\mathbf{H}$.
- In standard practice, only one MM update applied to $\mathbf{W}$ and $\mathbf{H}$, rather than fully solving subproblems $\min_{\mathbf{W}\geq 0} D(\mathbf{V}|\mathbf{WH})$ and $\min_{\mathbf{H}} D(\mathbf{V}|\mathbf{WH})$.
- Leads to a valid descent algorithm with multiplicative updates given by:

$$\mathbf{H} \leftarrow \mathbf{H}. \left(\frac{\mathbf{W}^T \left[(\mathbf{WH})^{\cdot(\beta-2)} . \mathbf{V} \right]}{\mathbf{W}^T \left[\mathbf{WH} \right]^{\cdot(\beta-1)}} \right)^{\gamma(\beta)}$$

$$\mathbf{W} \leftarrow \mathbf{W}. \left(\frac{\left[(\mathbf{WH})^{\cdot(\beta-2)} . \mathbf{V} \right] \mathbf{H}^T}{\left[\mathbf{WH} \right]^{\cdot(\beta-1)} \mathbf{H}^T} \right)^{\gamma(\beta)}$$

- Very straightforward implementation, no hyperparameters!
- Nonnegativity is automatically preserved given positive initializations.
- Linear complexity per iteration.
- In practice, minimizing $D(\mathbf{V}+\epsilon|\mathbf{WH}+\epsilon)$ prevents from numerical issues.

Convergence of the iterates

- By design, we have convergence of the objective values $C(\mathbf{W},\mathbf{H}) = D(\mathbf{V}|\mathbf{WH})$.
- What about the iterates ? Only partial answers so far.
- A theoretical challenge arises from the lack of coercivity of the objective: $\|\mathbf{W}\|$ or $\|\mathbf{H}\| \to \infty \not\Rightarrow C(\mathbf{W},\mathbf{H}) \to \infty$.
- Due to the scale indeterminacy: $C(\mathbf{W}\boldsymbol{\Lambda}^{-1}, \boldsymbol{\Lambda}\mathbf{H}) = C(\mathbf{W},\mathbf{H})$, with $\boldsymbol{\Lambda} \to 0$.

Possible remedies (modified problems)

1) Impose $\mathbf{W} \geq \epsilon$, $\mathbf{H} \geq \epsilon$ (Takahashi et al., 2018; Hien and Gillis, 2021).
2) Slightly change the objective function to ensure coercivity (Zhao and Tan, 2018):

$$C(\mathbf{W},\mathbf{H}) = D(\mathbf{V}|\mathbf{WH}) + \epsilon\|\mathbf{W}\|_1 + \epsilon\|\mathbf{H}\|_1$$

MM results in adding ϵ at the denominator of the multiplicative updates.

Selecting β by matrix completion

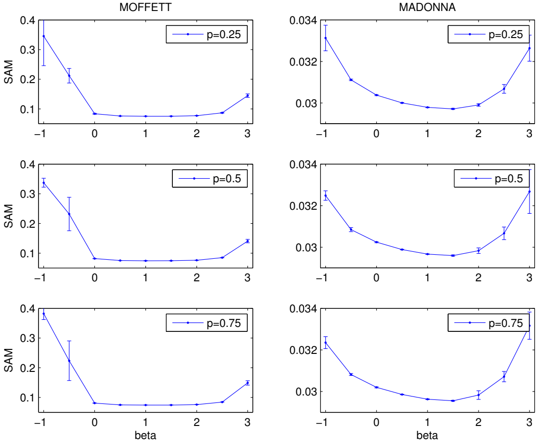
(Févotte and Dobigeon, 2015)

- **Data**: two unfolded hyperspectral cubes, $F \sim 150$, $N = 50 \times 50$
  - Aviris instrument over Moffett Field (CA), lake, soil & vegetation.
  - Hyspex/Madonna instrument over Villelongue (FR), forested area.
- a percentage of the pixels is randomly removed.
- **W** and **H** estimated from observed pixels (simple modification of MU).
- missing pixels are reconstructed from $\hat{\mathbf{V}} = \mathbf{WH}$.
- $K = 3$ ($\sim$ ground truth) and various values of β.
- evaluation using the average spectral angle mapper (aSAM):

$$\mathrm{aSAM}(\mathbf{V}) = \frac{1}{N} \sum_{n=1}^{N} \mathrm{acos} \left(\frac{\langle \mathbf{v}_n, \hat{\mathbf{v}}_n \rangle}{\|\mathbf{v}_n\| \|\hat{\mathbf{v}}_n\|} \right)$$

Selecting β by matrix completion

(Févotte and Dobigeon, 2015)



Recommended value $\beta \approx 1.5$ for these datasets
(compromise between Poisson and additive Gaussian noise).

Other alternating optimization methods

- MM-based multiplicative updates are a simple and competitive choice for many divergences (beyond β-divergences).
- More efficient options have been proposed for specific measures of fit, see books by Cichocki et al. (2009); Gillis (2020)

Quadratic loss (selection)

- Active-set methods (Kim and Park, 2011)
- Hierarchical alternating LS (Cichocki et al., 2007; Gillis and Glineur, 2012)
- Proximal gradient descent (Lin, 2007; Guan et al., 2012; Bolte et al., 2014)
- ADMM (Sun and Févotte, 2014; Huang et al., 2016)

Kullback-Leibler divergence (selection)

- Second-order coordinate descent methods (Hsieh and Dhillon, 2011)
- Hybrid Newton-type algorithms with line search and MU (Hien and Gillis, 2021)

Non-alternating methods (joint optimization)

- Optimize $C(\mathbf{W},\mathbf{H}) = D(\mathbf{V}|\mathbf{W},\mathbf{H})$ jointly in $\mathbf{W}$ and $\mathbf{H}$.
- Exciting line of research, driven by recent results in non-convex optimization. Possibly better optima and lower complexity.

Non-alternating methods (joint optimization)

- Optimize $C(\mathbf{W}, \mathbf{H}) = D(\mathbf{V}|\mathbf{W}, \mathbf{H})$ jointly in **W** and **H**.
- Exciting line of research, driven by recent results in non-convex optimization. Possibly better optima and lower complexity.

1) Proximal gradient algorithms with global smoothness constant ($\sim$Lipschitz) for the quadratic loss (Rakotomamonjy, 2013; Mukkamala and Ochs, 2019).

Non-alternating methods (joint optimization)

- Optimize $C(\mathbf{W}, \mathbf{H}) = D(\mathbf{V}|\mathbf{W}, \mathbf{H})$ jointly in $\mathbf{W}$ and $\mathbf{H}$.
- Exciting line of research, driven by recent results in non-convex optimization. Possibly better optima and lower complexity.

1) Proximal gradient algorithms with global smoothness constant ($\sim$Lipschitz) for the quadratic loss (Rakotomamonjy, 2013; Mukkamala and Ochs, 2019).
2) Joint MM algorithm for the β-divergence (Marmin, Goulart, and Févotte, 2021):
   - Global majorizer constructed using Jensen and tangent inequalities:

$$C(\mathbf{W}, \mathbf{H}) \leq G(\mathbf{W}, \mathbf{H}|\tilde{\mathbf{W}}, \tilde{\mathbf{H}})$$
$$C(\tilde{\mathbf{W}}, \tilde{\mathbf{H}}) = G(\tilde{\mathbf{W}}, \tilde{\mathbf{H}}|\tilde{\mathbf{W}}, \tilde{\mathbf{H}})$$

   - Global minimizer of G not available in closed form. G non-convex.
   - Alternate minimization of G leads to closed-form updates and new multiplicative rules. Important computational savings for some values of β (see paper).

Non-alternating methods (joint optimization)

- Optimize $C(\mathbf{W}, \mathbf{H}) = D(\mathbf{V}|\mathbf{W}, \mathbf{H})$ jointly in $\mathbf{W}$ and $\mathbf{H}$.
- Exciting line of research, driven by recent results in non-convex optimization. Possibly better optima and lower complexity.

1) Proximal gradient algorithms with global smoothness constant ($\sim$Lipschitz) for the quadratic loss (Rakotomamonjy, 2013; Mukkamala and Ochs, 2019).
2) Joint MM algorithm for the β-divergence (Marmin, Goulart, and Févotte, 2021):
   - Global majorizer constructed using Jensen and tangent inequalities:

$$C(\mathbf{W}, \mathbf{H}) \leq G(\mathbf{W}, \mathbf{H}|\tilde{\mathbf{W}}, \tilde{\mathbf{H}})$$
$$C(\tilde{\mathbf{W}}, \tilde{\mathbf{H}}) = G(\tilde{\mathbf{W}}, \tilde{\mathbf{H}}|\tilde{\mathbf{W}}, \tilde{\mathbf{H}})$$

   - Global minimizer of G not available in closed form. G non-convex.
   - Alternate minimization of G leads to closed-form updates and new multiplicative rules. Important computational savings for some values of β (see paper).
3) Second-order method for β-NMF based on efficient Hessian approximations and tricks to maintain semidefinite positivity (Vandecappelle et al., 2020).

Large-scale NMF

Online NMF

- Large number of samples $N >> F$.
- Update $\mathbf{W}$ as samples $\mathbf{v}_n$ become available.
- Vectors $\mathbf{h}_n$ act as latent variables, minimize

$$C(\mathbf{W}) = \sum_{n=1}^{N} \min_{\mathbf{h}_n \geq 0} D(\mathbf{v}_n | \mathbf{W}\mathbf{h}_n)$$

- Solved with online MM (Lefèvre et al., 2011b; Mairal, 2015; Zhao et al., 2017)

Stochastic NMF

- Large F and N.
- Online NMF with stochastic subsampling:

$$\min_{\mathbf{h}_n \geq 0} D(\mathbf{v}_n[\mathcal{I}] | \mathbf{W}[\mathcal{I}, :]\mathbf{h}_n)$$

where $\mathcal{I}$ is a random set of indices (Mensch et al., 2018).

Outline

Regularized NMF

- Induce prior information or desired structure on $\mathbf{H}$ (or $\mathbf{W}$) using penalty terms:

$$C(\mathbf{W}, \mathbf{H}) = D(\mathbf{V}|\mathbf{WH}) + S(\mathbf{H})$$

- MM algorithms are easily adapted to that setting:

$$D(\mathbf{V}|\mathbf{WH}) \qquad \leq G(\mathbf{H}|\tilde{\mathbf{H}}, \mathbf{W})$$

- Only the minimization step is changed.
- May however become intractable; sometimes $S(\mathbf{H})$ needs to be majorized itself.
- Similar to adjusting the proximal operator in proximal gradient descent.

Regularized NMF

- Induce prior information or desired structure on $\mathbf{H}$ (or $\mathbf{W}$) using penalty terms:

$$C(\mathbf{W}, \mathbf{H}) = D(\mathbf{V}|\mathbf{WH}) + S(\mathbf{H})$$

- MM algorithms are easily adapted to that setting:

$$D(\mathbf{V}|\mathbf{WH}) + S(\mathbf{H}) \leq G(\mathbf{H}|\tilde{\mathbf{H}}, \mathbf{W}) + S(\mathbf{H})$$

- Only the minimization step is changed.
- May however become intractable; sometimes $S(\mathbf{H})$ needs to be majorized itself.
- Similar to adjusting the proximal operator in proximal gradient descent.

Sparsity

- Promote zeros in **H** (or **W**), e.g,

$$S(\mathbf{H}) = \|\mathbf{H}\|_1 = \sum_{kn} h_{kn}, \quad S(\mathbf{H}) = \sum_{kn} \log(h_{kn} + \epsilon)$$

- Possibly with some group structure, e.g., cancel some rows of **H** (see Part II).
- Vast literature! Seminal paper by Hoyer (2004).
- Need to control $\|\mathbf{W}\|$ to avoid degenerate solutions $\|\mathbf{W}\| \to \infty$, $\|\mathbf{H}\| \to 0$.
- Because $C(\mathbf{W}\mathbf{\Lambda}^{-1}, \mathbf{\Lambda}\mathbf{H}) = D(\mathbf{V}|\mathbf{W}\mathbf{H}) + S(\mathbf{\Lambda}\mathbf{H})$, $S(\cdot)$ can be made arbitrary small.
- A common approach:

$$\min_{\mathbf{W},\mathbf{H}\geq 0} C(\mathbf{W}, \mathbf{H}) \quad \text{s.t.} \quad \forall k, \|\mathbf{w}_k\| = 1$$

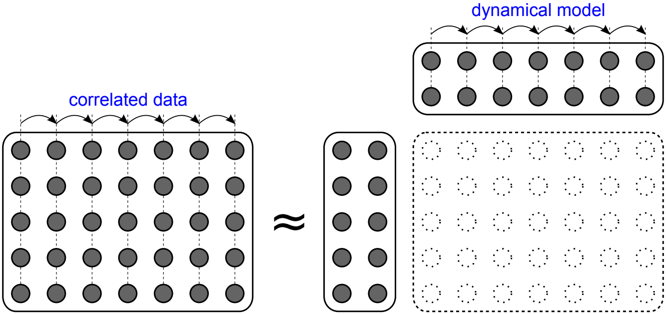
  - Change of variable (Eggert and Körner, 2004; Lefèvre et al., 2011a; Le Roux et al., 2015)
  - Lagrangian method (Leplat et al., 2021)

Smoothness

Impose temporal or spatial regularization, e.g.,

$$S(\mathbf{H}) = \sum_{kn} d(h_{kn} | h_{k(n-1)})$$

- Least squares penalization (Virtanen, 2007; Essid and Févotte, 2013)
- Gamma Markov chains (Smaragdis et al., 2014; Filstroff et al., 2021)

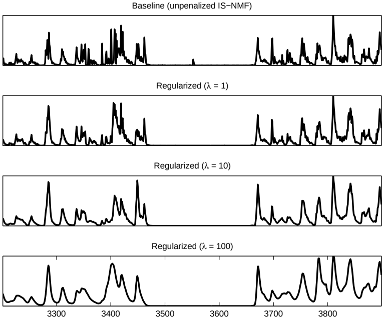


Smoothness

Impose temporal or spatial regularization, e.g.,

$$S(\mathbf{H}) = \sum_{kn} d(h_{kn} | h_{k(n-1)})$$

- Least squares penalization (Virtanen, 2007; Essid and Févotte, 2013)
- Gamma Markov chains (Smaragdis et al., 2014; Filstroff et al., 2021)



One row of **H** with increasing smoothness (Févotte, 2011)

Other common regularizers

- Orthogonal NMF: $\mathbf{H}\mathbf{H}^T = \mathbf{I}$.
  Essentially nonnegative clustering (Ding et al., 2006).
- Projective NMF: $\mathbf{H} = \mathbf{W}^T\mathbf{V}$.
  Essentially nonnegative PCA (Yang and Oja, 2010).
- Symmetric NMF: $\mathbf{H} = \mathbf{W}^T$.
  Popular in graph clustering (Kuang et al., 2012; Huang et al., 2013).
- Separable NMF: $\mathbf{W}$ is a subset of columns of $\mathbf{V}$.
  Very active research topic! (Donoho and Stodden, 2004; Arora et al., 2016)
- Archetypal NMF: $\mathbf{W}$ belongs to the column-range of $\mathbf{V}$.
  A relaxation of separable NMF (Ding et al., 2010; Chen et al., 2014).
- Minimum-volume NMF: penalize the aperture of $\mathbf{W}$.
  Very active research topic! (Miao and Qi, 2007; Chan et al., 2009)

Robust NMF for nonlinear hyperspectral unmixing
(Févotte and Dobigeon, 2015)

- Variants of the linear mixing model account for "non-linear" effects:

$$\mathbf{v}_n \approx \mathbf{W}\mathbf{h}_n + \mathbf{r}_n$$

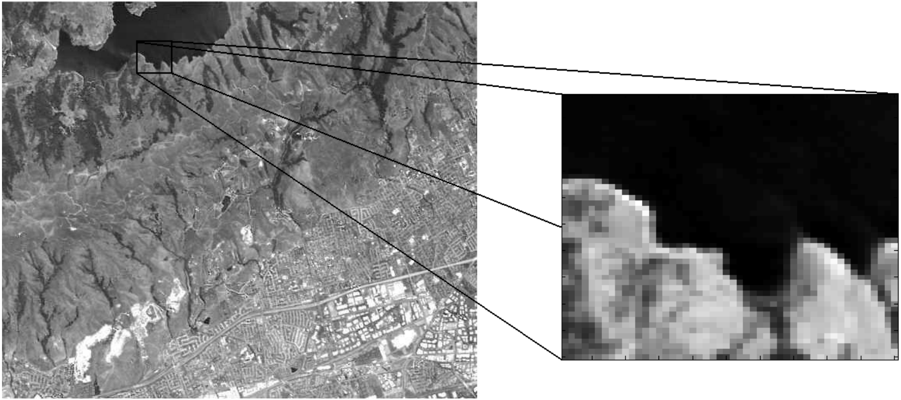
- Often, $\mathbf{r}_n$ has a parametric form, e.g., linear combination of quadratic components $\{\mathbf{w}_k \odot \mathbf{w}_j\}_{kj}$ (Nascimento and Bioucas-Dias, 2009; Fan et al., 2009; Altmann et al., 2012)
- Nonlinear effects usually affect few pixels only.
- We treat them as non-parametric sparse outliers.

$$\min_{\mathbf{W},\mathbf{H},\mathbf{R}\geq 0} D_\beta(\mathbf{V}|\mathbf{W}\mathbf{H} + \mathbf{R}) + \lambda\|\mathbf{R}\|_{2,1}$$

  where $\|\mathbf{R}\|_{2,1} = \sum_{n=1}^{N} \|\mathbf{r}_n\|_2$ induces sparsity at group level.
- A form of robust NMF (Candès et al., 2009)
- Optimized with majorization-minimization.

Robust NMF for nonlinear hyperspectral unmixing
(Févotte and Dobigeon, 2015)

Moffett Field data



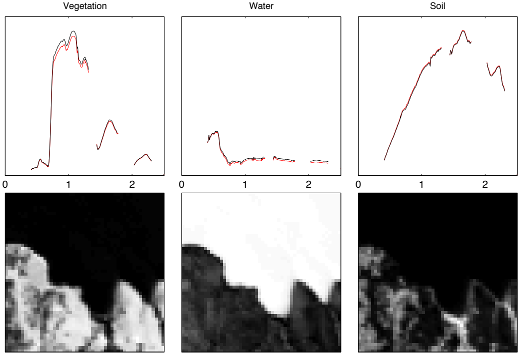
reproduced from (Dobigeon, 2007)

Robust NMF for nonlinear hyperspectral unmixing

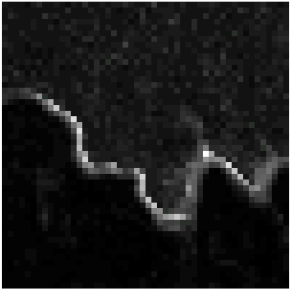
(Févotte and Dobigeon, 2015)

Unmixing results

spectral endmembers & activation maps
(**red:** $\beta = 1$, **black:** $\beta = 2$)



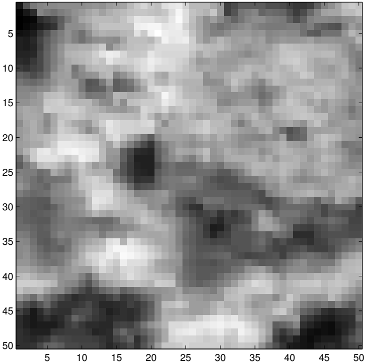
outlier energy $\{\|\mathbf{r}_n\|\}_n$
$(\beta = 1)$



Outlier term captures specific water/soil interactions.

Robust NMF for nonlinear hyperspectral unmixing
(Févotte and Dobigeon, 2015)

Villelongue/Madonna data (forested area)

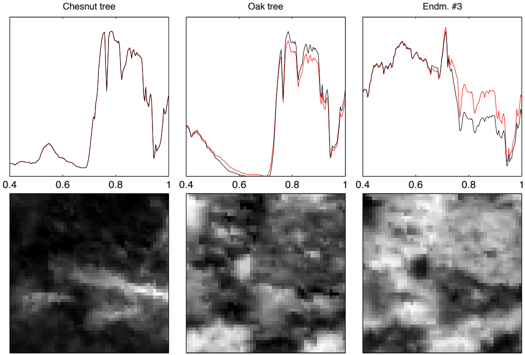


Robust NMF for nonlinear hyperspectral unmixing

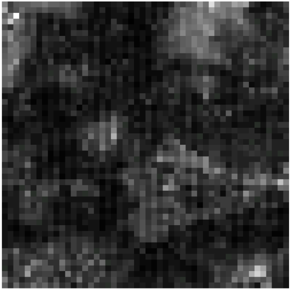
(Févotte and Dobigeon, 2015)

Unmixing results

spectral endmembers & activation maps
(**red:** $\beta = 1$, **black:** $\beta = 2$)



outlier energy $\{\|\mathbf{r}_n\|\}_n$
($\beta = 1$)

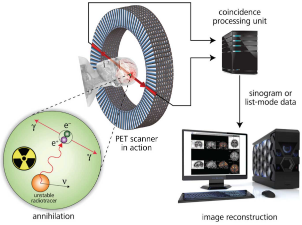


Outlier term seems to capture patterns due to sensor miscalibration.

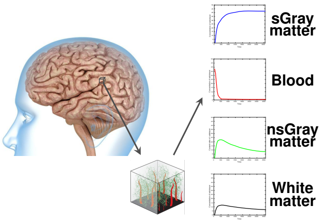
Factor analysis in dynamical PET

(Cavalcanti, Oberlin, Dobigeon, Févotte, Stute, Ribeiro, and Tauber, 2019)

- 3D functional imaging
- Observe the temporal evolution of the brain activity after injecting a radiotracer (biomarker of a specific compound).
- $\mathbf{v}_n$ is the time-activity curve (TAC) in voxel n.
- Neuroimaging: mixed contributions of 4 TAC signatures in each voxel.



Dynamic positron emission tomography



PET voxel decomposition

reproduced from (Cavalcanti, 2018)

Factor analysis in dynamical PET

(Cavalcanti, Oberlin, Dobigeon, Févotte, Stute, Ribeiro, and Tauber, 2019)

Mixing model

- the specific-binding TAC signature varies in space:

$$\begin{aligned}\mathbf{v}_n &\approx [\mathbf{w}_1 + \boldsymbol{\delta}_n] h_{1n} + \sum_{k=2}^{K} \mathbf{w}_k h_{kn} \\ &\approx [\mathbf{w}_1 + \mathbf{D}\mathbf{b}_n] h_{1n} + \sum_{k=2}^{K} \mathbf{w}_k h_{kn} \\ &\approx \mathbf{W}\mathbf{h}_n + h_{1n} \mathbf{D}\mathbf{b}_n\end{aligned}$$

- $\mathbf{D}$ is fixed and pre-trained using labeled or simulated data.

Estimation

$$\min_{\mathbf{W},\mathbf{H},\mathbf{B}\geq 0} D_\beta(\mathbf{V}|\mathbf{W}\mathbf{H} + \mathbf{1}\,\underline{\mathbf{h}}_1 \odot \mathbf{D}\mathbf{B}) + \lambda \|\mathbf{B}\|_{2,1}$$

Optimized with majorization-minimization.

Factor analysis in dynamical PET

(Cavalcanti, Oberlin, Dobigeon, Févotte, Stute, Ribeiro, and Tauber, 2019)

Unmixing results

- real dynamic PET image of a stroke subject injected with a tracer for neuroinflammation.
- MRI ground-truth region of the stroke.

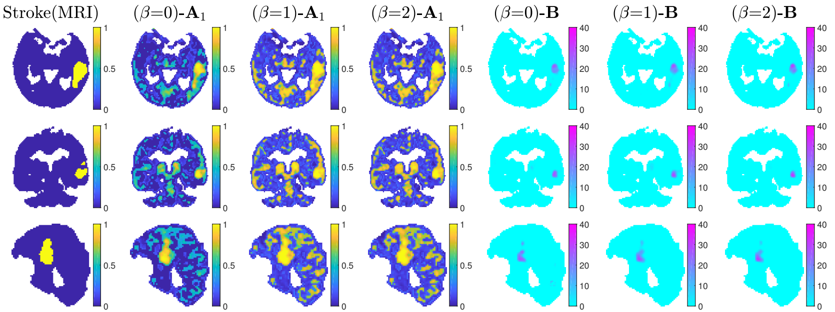


Fig.: Specific-binding activation (h_{1n}) and variability maps ($\|\mathbf{b}_n\|_{2,1}$) in three different planes and for three values of β

End of Part I

Half-time conclusions

- NMF has become a popular data processing tool over the last 20 years.
- Very suited to unmixing problems in unsupervised settings.
- Exciting non-convex optimization problem with non-Euclidean measures of fit.
- MM is a versatile algorithmic framework for NMF.
  - Simple multiplicative algorithms for the β-divergence and beyond.
  - Can be adapted to regularized NMF and variants.
  - More efficient algorithms exist for the quadratic loss.

Funding acknowledgement: *European Research Council, National Research Fondation Singapore, Agence Nationale de la Recherche France*

References I

Y. Altmann, A. Halimi, N. Dobigeon, and J.-Y. Tourneret. Supervised nonlinear spectral unmixing using a post-nonlinear mixing model for hyperspectral imagery. *IEEE Transactions on Image Processing*, 21(6): 3017–3025, June 2012.

S. Arora, R. Ge, R. Kannan, and A. Moitra. Computing a nonnegative matrix factorization—provably. *SIAM Journal on Computing*, 45(4):1582–1611, 2016.

A. Basu, I. R. Harris, N. L. Hjort, and M. C. Jones. Robust and efficient estimation by minimising a density power divergence. *Biometrika*, 85(3):549–559, Sep. 1998.

M. W. Berry, M. Browne, A. N. Langville, V. P. Pauca, and R. J. Plemmons. Algorithms and applications for approximate nonnegative matrix factorization. *Computational Statistics & Data Analysis*, 52(1):155–173, Sep. 2007.

J. M. Bioucas-Dias, A. Plaza, N. Dobigeon, M. Parente, Q. Du, P. Gader, and J. Chanussot. Hyperspectral unmixing overview: Geometrical, statistical, and sparse regression-based approaches. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 5(2):354–379, 2012.

J. Bolte, S. Sabach, and M. Teboulle. Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Mathematical Programming*, 146(1):459–494, 2014.

E. J. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *Journal of ACM*, 58(1): 1–37, 2009.

Y. Cao, P. P. B. Eggermont, and S. Terebey. Cross Burg entropy maximization and its application to ringing suppression in image reconstruction. *IEEE Transactions on Image Processing*, 8(2):286–292, Feb. 1999. doi: 10.1109/83.743861.

Y. C. Cavalcanti. *Factor analysis of dynamic PET images*. PhD thesis, Toulouse INP, 2018.

References II

Y. C. Cavalcanti, T. Oberlin, N. Dobigeon, C. Févotte, S. Stute, M. Ribeiro, and C. Tauber. Factor analysis of dynamic PET images: Beyond Gaussian noise. *IEEE Transactions on Medical Imaging*, 38(9):2231–2241, Sep. 2019. ISSN 0278-0062. doi: 10.1109/TMI.2019.2906828. URL `https://arxiv.org/pdf/1807.11455`.

T.-H. Chan, C.-Y. Chi, Y.-M. Huang, and W.-K. Ma. A convex analysis-based minimum-volume enclosing simplex algorithm for hyperspectral unmixing. *IEEE Transactions on Signal Processing*, 57(11):4418–4432, 2009.

Y. Chen, J. Mairal, and Z. Harchaoui. Fast and robust archetypal analysis for representation learning. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

A. Cichocki and S. Amari. Families of Alpha- Beta- and Gamma- divergences: Flexible and robust measures of similarities. *Entropy*, 12(6):1532–1568, June 2010.

A. Cichocki, R. Zdunek, and S. Amari. Csiszar's divergences for non-negative matrix factorization: Family of new algorithms. In *Proc. International Conference on Independent Component Analysis and Blind Signal Separation (ICA)*, pages 32–39, Charleston SC, USA, Mar. 2006.

A. Cichocki, R. Zdunek, and S.-i. Amari. Hierarchical ALS algorithms for nonnegative matrix and 3d tensor factorization. In *International Conference on Independent Component Analysis and Signal Separation*, 2007.

A. Cichocki, H. Lee, Y.-D. Kim, and S. Choi. Non-negative matrix factorization with α-divergence. *Pattern Recognition Letters*, 29(9):1433–1440, July 2008.

A. Cichocki, R. Zdunek, A. H. Phan, and S.-I. Amari. *Nonnegative matrix and tensor factorizations: Applications to exploratory multi-way data analysis and blind source separation*. John Wiley & Sons, 2009.

References III

M. Daube-Witherspoon and G. Muehllehner. An iterative image space reconstruction algorthm suitable for volume ECT. *IEEE Transactions on Medical Imaging*, 5(5):61 – 66, 1986. doi: 10.1109/TMI.1986.4307748.

A. R. De Pierro. On the relation between the ISRA and the EM algorithm for positron emission tomography. *IEEE Trans. Medical Imaging*, 12(2):328–333, 1993. doi: 10.1109/42.232263.

I. S. Dhillon and S. Sra. Generalized nonnegative matrix approximations with Bregman divergences. In *Advances in Neural Information Processing Systems (NIPS)*, 2005.

C. Ding, T. Li, W. Peng, and H. Park. Orthogonal nonnegative matrix t-factorizations for clustering. In *Proc. ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, pages 126–135. ACM, 2006.

C. H. Q. Ding, T. Li, and M. I. Jordan. Convex and semi-nonnegative matrix factorizations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(1):45 – 55, 2010. doi: 10.1109/TPAMI.2008.277.

D. Donoho and V. Stodden. When does non-negative matrix factorization give a correct decomposition into parts? In *Advances in Neural Information Processing Systems (NIPS)*, 2004.

J. Eggert and E. Körner. Sparse coding and NMF. In *Proc. IEEE International Joint Conference on Neural Networks*, pages 2529–2533, 2004.

S. Essid and C. Févotte. Smooth nonnegative matrix factorization for unsupervised audiovisual document structuring. *IEEE Transactions on Multimedia*, 15(2):415–425, Feb. 2013. doi: 10.1109/TMM.2012.2228474. URL `https://www.irit.fr/~Cedric.Fevotte/publications/journals/ieee_multimedia_smoothnmf.pdf`.

References IV

W. Fan, B. Hu, J. Miller, and M. Li. Comparative study between a new nonlinear model and common linear model for analysing laboratory simulated-forest hyperspectral data. *International Journal of Remote Sensing*, 30(11):2951–2962, June 2009.

C. Févotte. Majorization-minimization algorithm for smooth Itakura-Saito nonnegative matrix factorization. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Prague, Czech Republic, May 2011. URL https://www.irit.fr/~Cedric.Fevotte/publications/proceedings/icassp11a.pdf.

C. Févotte and N. Dobigeon. Nonlinear hyperspectral unmixing with robust nonnegative matrix factorization. *IEEE Transactions on Image Processing*, 24(12):4810–4819, Dec. 2015. doi: 10.1109/TIP.2015.2468177. URL https://www.irit.fr/~Cedric.Fevotte/publications/journals/tip2015.pdf.

C. Févotte and J. Idier. Algorithms for nonnegative matrix factorization with the beta-divergence. *Neural Computation*, 23(9):2421–2456, Sep. 2011. doi: 10.1162/NECO_a_00168. URL https://www.irit.fr/~Cedric.Fevotte/publications/journals/neco11.pdf.

C. Févotte, N. Bertin, and J.-L. Durrieu. Nonnegative matrix factorization with the Itakura-Saito divergence. With application to music analysis. *Neural Computation*, 21(3):793–830, Mar. 2009. doi: 10.1162/neco.2008.04-08-771. URL https://www.irit.fr/~Cedric.Fevotte/publications/journals/neco09_is-nmf.pdf.

L. Filstroff, O. Gouvert, C. Févotte, and O. Cappé. A comparative study of Gamma Markov chains for temporal non-negative factorization. *IEEE Transactions on Signal Processing*, 69:1614–1626, 2021. doi: 10.1109/TSP.2021.3060000. URL https://arxiv.org/pdf/2006.12843.

L. Finesso and P. Spreij. Nonnegative matrix factorization and I-divergence alternating minimization. *Linear Algebra and its Applications*, 416:270–287, 2006.

References V

N. Gillis. *Nonnegative Matrix Factorization*. SIAM, 2020.

N. Gillis and F. Glineur. Accelerated multiplicative updates and hierarchical ALS algorithms for nonnegative matrix factorization. *Neural Computation*, 24(4):1085–1105, 04 2012. ISSN 0899-7667. doi: 10.1162/NECO_a_00256. URL https://doi.org/10.1162/NECO_a_00256.

N. Guan, D. Tao, Z. Luo, and B. Yuan. NeNMF: An optimal gradient method for nonnegative matrix factorization. *IEEE Transactions on Signal Processing*, 60(6):2882–2898, 2012.

L. T. K. Hien and N. Gillis. Algorithms for nonnegative matrix factorization with the Kullback–Leibler divergence. *Journal of Scientific Computing*, 87(3):93, 2021.

T. Hofmann. Probabilistic latent semantic indexing. In *Proc. 22nd International Conference on Research and Development in Information Retrieval (SIGIR)*, 1999. URL http://www.cs.brown.edu/~th/papers/Hofmann-SIGIR99.pdf.

P. O. Hoyer. Non-negative matrix factorization with sparseness constraints. *Journal of Machine Learning Research*, 5:1457–1469, 2004.

C. J. Hsieh and I. S. Dhillon. Fast coordinate descent methods with variable selection for non-negative matrix factorization. In *Proc. 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 1064 – 1072, Aug. 2011.

K. Huang, N. D. Sidiropoulos, and A. Swami. Non-negative matrix factorization revisited: Uniqueness and algorithm for symmetric decomposition. *IEEE Transactions on Signal Processing*, 62(1):211–224, 2013.

K. Huang, N. D. Sidiropoulos, and A. P. Liavas. A flexible and efficient algorithmic framework for constrained matrix and tensor factorization. *IEEE Transactions on Signal Processing*, 64(19):5052–5065, 2016. doi: 10.1109/TSP.2016.2576427.

References VI

J. Kim and H. Park. Fast nonnegative matrix factorization: An active-set-like method and comparisons. *SIAM Journal on Scientific Computing*, 33:3261–3281, 2011.

D. Kuang, C. Ding, and H. Park. Symmetric nonnegative matrix factorization for graph clustering. In *Prco. SIAM International Conference on Data Mining*, pages 106–117, 2012.

J. Le Roux, F. J. Weninger, and J. R. Hershey. Sparse NMF–half-baked or well done? Technical report, Mitsubishi Electric Research Labs (MERL), 2015.

D. D. Lee and H. S. Seung. Learning the parts of objects with nonnegative matrix factorization. *Nature*, 401: 788–791, 1999.

D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. In *Advances in Neural and Information Processing Systems 13*, pages 556–562, 2001.

A. Lefèvre, F. Bach, and C. Févotte. Itakura-Saito nonnegative matrix factorization with group sparsity. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Prague, Czech Republic, May 2011a. URL `https://www.irit.fr/~Cedric.Fevotte/publications/proceedings/icassp11c.pdf`.

A. Lefèvre, F. Bach, and C. Févotte. Online algorithms for nonnegative matrix factorization with the Itakura-Saito divergence. In *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Mohonk, NY, Oct. 2011b. URL `https://www.irit.fr/~Cedric.Fevotte/publications/proceedings/waspaa11.pdf`.

V. Leplat, N. Gillis, and J. Idier. Multiplicative updates for nmf with β-divergences under disjoint equality constraints. *SIAM Journal on Matrix Analysis and Applications*, 42(2):730–752, 2021.

C.-J. Lin. Projected gradient methods for nonnegative matrix factorization. *Neural Computation*, 19: 2756–2779, 2007.

References VII

L. B. Lucy. An iterative technique for the rectification of observed distributions. *Astronomical Journal*, 79: 745–754, 1974. doi: 10.1086/111605.

J. Mairal. Incremental majorization-minimization optimization with application to large-scale machine learning. *SIAM Journal on Optimization*, 25(2):829–855, 2015.

A. Marmin, J. H. de M.. Goulart, and C. Févotte. Joint majorization-minimization for nonnegative matrix factorization with the beta-divergence. Technical report, arXiv, June 2021. URL https://arxiv.org/pdf/2106.15214.

A. Mensch, J. Mairal, B. Thirion, and G. Varoquaux. Stochastic subsampling for factorizing huge matrices. *IEEE Transactions on Signal Processing*, 66(1):113–128, 2018. doi: 10.1109/TSP.2017.2752697.

L. Miao and H. Qi. Endmember extraction from highly mixed data using minimum volume constrained nonnegative matrix factorization. *IEEE Transactions on Geoscience and Remote Sensing*, 45(3):765–777, 2007. ISSN 0196-2892. doi: 10.1109/TGRS.2006.888466.

M. C. Mukkamala and P. Ochs. Beyond alternating updates for matrix factorization with inertial bregman proximal gradient algorithms. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

M. Nakano, H. Kameoka, J. Le Roux, Y. Kitano, N. Ono, and S. Sagayama. Convergence-guaranteed multiplicative algorithms for non-negative matrix factorization with beta-divergence. In *Proc. IEEE International Workshop on Machine Learning for Signal Processing (MLSP'2010)*, Sep. 2010.

J. M. P. Nascimento and J. M. Bioucas-Dias. Nonlinear mixture model for hyperspectral unmixing. In *Proc. SPIE Image and Signal Processing for Remote Sensing XV*, 2009.

P. Paatero and U. Tapper. Positive matrix factorization : A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*, 5:111–126, 1994.

References VIII

A. Rakotomamonjy. Direct optimization of the dictionary learning problem. *IEEE Transactions on Signal Processing*, 61(12):5495–5506, 2013.

W. H. Richardson. Bayesian-based iterative method of image restoration. *Journal of the Optical Society of America*, 62:55–59, 1972.

P. Smaragdis. About this non-negative business. WASPAA keynote slides, 2013. URL `http://web.engr.illinois.edu/~paris/pubs/smaragdis-waspaa2013keynote.pdf`.

P. Smaragdis and J. C. Brown. Non-negative matrix factorization for polyphonic music transcription. In *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Oct. 2003.

P. Smaragdis, C. Févotte, G. Mysore, N. Mohammadiha, and M. Hoffman. Static and dynamic source separation using nonnegative factorizations: A unified view. *IEEE Signal Processing Magazine*, 31(3): 66–75, May 2014. doi: 10.1109/MSP.2013.2297715. URL `https://www.irit.fr/~Cedric.Fevotte/publications/journals/spm2014.pdf`.

D. L. Sun and C. Févotte. Alternating direction method of multipliers for non-negative matrix factorization with the beta-divergence. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy, May 2014. URL `https://www.irit.fr/~Cedric.Fevotte/publications/proceedings/icassp14a.pdf`.

N. Takahashi, J. Katayama, M. Seki, and J. Takeuchi. A unified global convergence analysis of multiplicative update rules for nonnegative matrix factorization. *Computational Optimization and Applications*, 71(1): 221–250, 2018.

M. Vandecappelle, N. Vervliet, and L. De Lathauwer. A second-order method for fitting the canonical polyadic decomposition with non-least-squares cost. *IEEE Transactions on Signal Processing*, 68:4454–4465, 2020.

References IX

T. Virtanen. Monaural sound source separation by non-negative matrix factorization with temporal continuity and sparseness criteria. *IEEE Transactions on Audio, Speech and Language Processing*, 15(3):1066–1074, Mar. 2007.

Z. Yang and E. Oja. Linear and nonlinear projective nonnegative matrix factorization. *IEEE Transactions on Neural Networks*, 21(5):734–749, 2010.

Z. Yang and E. Oja. Unified development of multiplicative algorithms for linear and quadratic nonnegative matrix factorization. *IEEE Transactions on Neural Networks*, 22:1878 – 1891, Dec. 2011. doi: http://dx.doi.org/10.1109/TNN.2011.2170094.

R. Zhao and V. Y. F. Tan. A unified convergence analysis of the multiplicative update algorithm for regularized nonnegative matrix factorization. *IEEE Transactions on Signal Processing*, 66(1):129–138, Jan 2018. ISSN 1053-587X. doi: 10.1109/TSP.2017.2757914.

R. Zhao, V. Y. Tan, and H. Xu. Online nonnegative matrix factorization with general divergences. In *Proc. AISTATS*, 2017.

22-27 MAY 2022
ICASSP 2022
SINGAPORE

IEEE Signal Processing Society

IEEE

T16: Recent advances in Nonnegative Matrix Factorization (Part 1)

Q&A

T16: Recent advances in Nonnegative Matrix Factorization

Break (1530-1600 UTC+8)

22-27 MAY 2022
ICASSP 2022
SINGAPORE

T16: Recent advances in Nonnegative Matrix Factorization (Part 2)

Vincent Tan

Recent Advances in Nonnegative Matrix Factorization

Part II: Extensions of NMF

Cédric Févotte
CNRS, Toulouse, France

Vincent Y. F. Tan
National University of Singapore

ICASSP Tutorial
Singapore, May 2022

Outline

Nonnegative rank selection by automatic relevance determination

Distributionally robust nonnegative matrix factorization

NMF in ranking models and sports analytics

PSDMF and links with phase retrieval and affine rank minimization

Outline

Nonnegative rank selection by automatic relevance determination

Distributionally robust nonnegative matrix factorization

NMF in ranking models and sports analytics

PSDMF and links with phase retrieval and affine rank minimization

Nonnegative Rank Selection by Automatic Relevance Determination (ARD)

Tan and Févotte (2013)

Nonnegative Rank Selection by Automatic Relevance Determination (ARD)

Tan and Févotte (2013)

- Recall that in NMF, one is given a data matrix $\mathbf{V} \in \mathbb{R}_+^{F \times N}$ and tries to find a dictionary matrix $\mathbf{W} \in \mathbb{R}_+^{F \times K}$ and coefficient matrix $\mathbf{H} \in \mathbb{R}_+^{K \times N}$ such that

$$\mathbf{V} \approx \hat{\mathbf{V}} = \mathbf{WH}.$$

Nonnegative Rank Selection by Automatic Relevance Determination (ARD)

Tan and Févotte (2013)

- Recall that in NMF, one is given a data matrix $\mathbf{V} \in \mathbb{R}_+^{F \times N}$ and tries to find a dictionary matrix $\mathbf{W} \in \mathbb{R}_+^{F \times K}$ and coefficient matrix $\mathbf{H} \in \mathbb{R}_+^{K \times N}$ such that

$$\mathbf{V} \approx \hat{\mathbf{V}} = \mathbf{WH}.$$

- Usually solved using a constrained minimization problem

$$\min_{\mathbf{W},\mathbf{H} \geq \mathbf{0}} D\big(\mathbf{V} \mid \mathbf{WH}\big) = \sum_{f=1}^{F} \sum_{n=1}^{N} d\big([\mathbf{V}]_{fn} \mid [\mathbf{WH}]_{fn}\big).$$

Nonnegative Rank Selection by Automatic Relevance Determination (ARD)

Tan and Févotte (2013)

- Recall that in NMF, one is given a data matrix $\mathbf{V} \in \mathbb{R}_+^{F \times N}$ and tries to find a dictionary matrix $\mathbf{W} \in \mathbb{R}_+^{F \times K}$ and coefficient matrix $\mathbf{H} \in \mathbb{R}_+^{K \times N}$ such that

$$\mathbf{V} \approx \hat{\mathbf{V}} = \mathbf{WH}.$$

- Usually solved using a constrained minimization problem

$$\min_{\mathbf{W},\mathbf{H} \geq \mathbf{0}} D\big(\mathbf{V} \mid \mathbf{WH}\big) = \sum_{f=1}^{F} \sum_{n=1}^{N} d\big([\mathbf{V}]_{fn} \mid [\mathbf{WH}]_{fn}\big).$$

- How to find the common/latent dimension K?

Nonnegative Rank Selection by Automatic Relevance Determination (ARD)

Tan and Févotte (2013)

- Recall that in NMF, one is given a data matrix $\mathbf{V} \in \mathbb{R}_+^{F \times N}$ and tries to find a dictionary matrix $\mathbf{W} \in \mathbb{R}_+^{F \times K}$ and coefficient matrix $\mathbf{H} \in \mathbb{R}_+^{K \times N}$ such that

$$\mathbf{V} \approx \hat{\mathbf{V}} = \mathbf{WH}.$$

- Usually solved using a constrained minimization problem

$$\min_{\mathbf{W},\mathbf{H} \geq \mathbf{0}} D(\mathbf{V} \mid \mathbf{WH}) = \sum_{f=1}^{F} \sum_{n=1}^{N} d\big([\mathbf{V}]_{fn} \mid [\mathbf{WH}]_{fn}\big).$$

- How to find the common/latent dimension K?
- If K is too large $\Longrightarrow$ Overfitting! K too small $\Longrightarrow$ Poor fit to model!

Nonnegative Rank Selection by Automatic Relevance Determination (ARD)

Tan and Févotte (2013)

- Recall that in NMF, one is given a data matrix $\mathbf{V} \in \mathbb{R}_+^{F \times N}$ and tries to find a dictionary matrix $\mathbf{W} \in \mathbb{R}_+^{F \times K}$ and coefficient matrix $\mathbf{H} \in \mathbb{R}_+^{K \times N}$ such that

$$\mathbf{V} \approx \hat{\mathbf{V}} = \mathbf{W}\mathbf{H}.$$

- Usually solved using a constrained minimization problem

$$\min_{\mathbf{W},\mathbf{H} \geq \mathbf{0}} D\big(\mathbf{V} \mid \mathbf{W}\mathbf{H}\big) = \sum_{f=1}^{F} \sum_{n=1}^{N} d\big([\mathbf{V}]_{fn} \mid [\mathbf{W}\mathbf{H}]_{fn}\big).$$

- How to find the common/latent dimension K?
- If K is too large $\implies$ Overfitting! K too small $\implies$ Poor fit to model!
- Solve this by automatic relevance determination (Bishop, 1999; Tipping, 2001)

Nonnegative Rank Selection by Automatic Relevance Determination (ARD)

Tan and Févotte (2013)

- Recall that in NMF, one is given a data matrix $\mathbf{V} \in \mathbb{R}_+^{F \times N}$ and tries to find a dictionary matrix $\mathbf{W} \in \mathbb{R}_+^{F \times K}$ and coefficient matrix $\mathbf{H} \in \mathbb{R}_+^{K \times N}$ such that

$$\mathbf{V} \approx \hat{\mathbf{V}} = \mathbf{WH}.$$

- Usually solved using a constrained minimization problem

$$\min_{\mathbf{W},\mathbf{H} \geq \mathbf{0}} D\big(\mathbf{V} \mid \mathbf{WH}\big) = \sum_{f=1}^{F} \sum_{n=1}^{N} d\big([\mathbf{V}]_{fn} \mid [\mathbf{WH}]_{fn}\big).$$

- How to find the common/latent dimension K?
- If K is too large $\implies$ Overfitting! K too small $\implies$ Poor fit to model!
- Solve this by automatic relevance determination (Bishop, 1999; Tipping, 2001)
- Natural extension of regularization ideas discussed by Cédric.

Probabilistic Model for ARD in NMF

- Assign each column of **W** and each row of **H** priors

$$\mathbf{W} = \begin{bmatrix} | & | & & | \\ \mathbf{w}_1 & \mathbf{w}_2 & \dots & \mathbf{w}_K \\ | & | & & | \end{bmatrix} \qquad \mathbf{H} = \begin{bmatrix} - & \underline{h}_1 & - \\ - & \underline{h}_2 & - \\ & \vdots & \\ - & \underline{h}_K & - \end{bmatrix}$$

Probabilistic Model for ARD in NMF

- Assign each column of **W** and each row of **H** priors

$$\mathbf{W} = \begin{bmatrix} | & | & & | \\ \mathbf{w}_1 & \mathbf{w}_2 & \dots & \mathbf{w}_K \\ | & | & & | \end{bmatrix} \qquad \mathbf{H} = \begin{bmatrix} - & \underline{h}_1 & - \\ - & \underline{h}_2 & - \\ & \vdots & \\ - & \underline{h}_K & - \end{bmatrix}$$

- Tie the k^{th} column $\mathbf{w}_k$ and the k^{th} row $\underline{h}_k$ together through a common relevance weight $\lambda_k \geq 0$.

Probabilistic Model for ARD in NMF

- Assign each column of $\mathbf{W}$ and each row of $\mathbf{H}$ priors

$$\mathbf{W} = \begin{bmatrix} | & | & & | \\ \mathbf{w}_1 & \mathbf{w}_2 & \dots & \mathbf{w}_K \\ | & | & & | \end{bmatrix} \qquad \mathbf{H} = \begin{bmatrix} - & \underline{h}_1 & - \\ - & \underline{h}_2 & - \\ & \vdots & \\ - & \underline{h}_K & - \end{bmatrix}$$

- Tie the k^{th} column $\mathbf{w}_k$ and the k^{th} row $\underline{h}_k$ together through a common relevance weight $\lambda_k \geq 0$.
- Maintain nonnegativity by choosing nonnegative priors, e.g.,
  - Half Gaussian, i.e.,

$$p(w_{fk} \mid \lambda_k) = \left(\frac{2}{\pi\lambda_k}\right)^{1/2} \exp\left(-\frac{w_{fk}^2}{2\lambda_k}\right) \qquad p(h_{kn} \mid \lambda_k) = \left(\frac{2}{\pi\lambda_k}\right)^{1/2} \exp\left(-\frac{h_{kn}^2}{2\lambda_k}\right).$$

  - Exponential

$$p(w_{fk} \mid \lambda_k) = \frac{1}{\lambda_k} \exp\left(-\frac{w_{fk}}{\lambda_k}\right) \qquad p(h_{kn} \mid \lambda_k) = \frac{1}{\lambda_k} \exp\left(-\frac{h_{kn}}{\lambda_k}\right)$$

  - Both these distributions are supported on $\mathbb{R}_+$.

Half Gaussian and Exponential

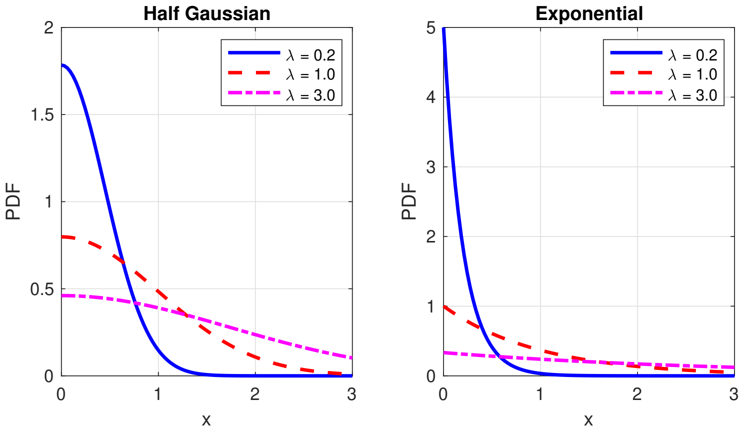
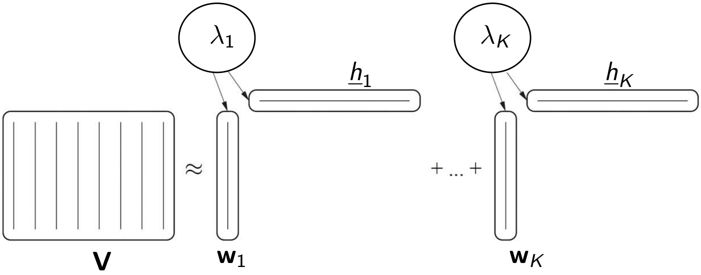


Figure: Half Gaussian and Exponential Distributions

Probabilistic Model for ARD in NMF



- λ_k is a common variance-like quantity.
- When $\lambda_k \downarrow 0$, $\|\mathbf{w}_k\|$ and $\|\underline{h}_k\|$ both tend to 0.
- The k^{th} component can be removed without compromising data fidelity.

Probabilistic Model for ARD in NMF

- Prior on common variance-like parameter λ_k is inverse-Gamma

$$p(\lambda_k; a, b) = \frac{b^a}{\Gamma(a)} \lambda_k^{-(a+1)} \exp\Big(-\frac{b}{\lambda_k}\Big),$$

where a and b are the shape and scale hyperparameters, respectively.

Probabilistic Model for ARD in NMF

- Prior on common variance-like parameter λ_k is inverse-Gamma

$$p(\lambda_k; a, b) = \frac{b^a}{\Gamma(a)} \lambda_k^{-(a+1)} \exp\Big(-\frac{b}{\lambda_k}\Big),$$

  where a and b are the shape and scale hyperparameters, respectively.

- Set a and b to be the same for all k.

Probabilistic Model for ARD in NMF

- Prior on common variance-like parameter λ_k is inverse-Gamma

$$p(\lambda_k; a, b) = \frac{b^a}{\Gamma(a)} \lambda_k^{-(a+1)} \exp\Big(-\frac{b}{\lambda_k}\Big),$$

  where a and b are the shape and scale hyperparameters, respectively.

- Set a and b to be the same for all k.

- The inverse-Gamma prior is chosen because it is conjugate to the variance-parameter in the Half Gaussian and the inverse rate parameter in the Exponential.

Probabilistic Model for ARD in NMF

- Prior on common variance-like parameter λ_k is inverse-Gamma

$$p(\lambda_k; a, b) = \frac{b^a}{\Gamma(a)} \lambda_k^{-(a+1)} \exp\Big(-\frac{b}{\lambda_k}\Big),$$

  where a and b are the shape and scale hyperparameters, respectively.

- Set a and b to be the same for all k.

- The inverse-Gamma prior is chosen because it is conjugate to the variance-parameter in the Half Gaussian and the inverse rate parameter in the Exponential.

- Leads to closed-form updates.

Probabilistic Model for ARD in NMF

- Prior on common variance-like parameter λ_k is inverse-Gamma

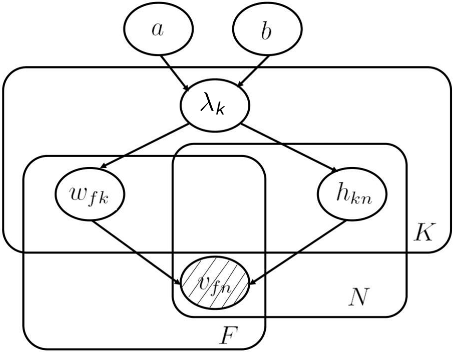
$$p(\lambda_k; a, b) = \frac{b^a}{\Gamma(a)} \lambda_k^{-(a+1)} \exp\Big(-\frac{b}{\lambda_k}\Big),$$

where a and b are the shape and scale hyperparameters, respectively.

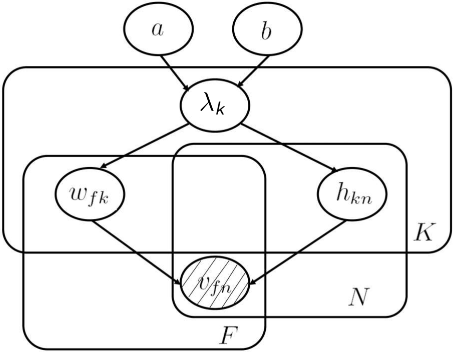
- Set a and b to be the same for all k.
- The inverse-Gamma prior is chosen because it is conjugate to the variance-parameter in the Half Gaussian and the inverse rate parameter in the Exponential.
- Leads to closed-form updates.
- Assume independence

$$p(\boldsymbol{\lambda}; a, b) = \prod_{k=1}^{K} p(\lambda_k; a, b).$$

Probabilistic Model for ARD in NMF



Probabilistic Model for ARD in NMF



- $\mathbf{V} = [v_{fn}]$ are observed;
- a, b are hyperparameters;
- Want to learn $\mathbf{W} = [w_{fn}]$ and $\mathbf{H} = [h_{kn}]$ and implicitly K, i.e.,

$$K = |\{k \in [K] : \lambda_k > \text{threshold}\}| .$$

Objective function for ARD in NMF

- Combining the prior and likelihood, the objective function (log-posterior) can be written as

$$\begin{aligned} C(\mathbf{W}, \mathbf{H}, \boldsymbol{\lambda}) &= -\log p(\mathbf{W}, \mathbf{H}, \boldsymbol{\lambda} \mid \mathbf{V}) \\ &\stackrel{c}{=} \frac{1}{\phi} D_\beta(\mathbf{V} \mid \mathbf{W}\mathbf{H}) + \sum_{k=1}^{K} \frac{1}{\lambda_k} \big(f(\mathbf{w}_k) + f(\underline{h}_k) + b \big) + c \log \lambda_k . \end{aligned}$$

Objective function for ARD in NMF

- Combining the prior and likelihood, the objective function (log-posterior) can be written as

$$C(\mathbf{W}, \mathbf{H}, \boldsymbol{\lambda}) = -\log p(\mathbf{W}, \mathbf{H}, \boldsymbol{\lambda} \mid \mathbf{V})$$
$$\stackrel{c}{=} \frac{1}{\phi} D_\beta(\mathbf{V} \mid \mathbf{WH}) + \sum_{k=1}^{K} \frac{1}{\lambda_k} \big(f(\mathbf{w}_k) + f(\underline{h}_k) + b \big) + c \log \lambda_k.$$

- Constant ϕ is the dispersion parameter (of the Tweedie distribution):
  - $\beta = 2$: Gaussian distribution and $\phi = \sigma^2$;
  - $\beta = 1$: Poisson distribution and $\phi = 1$;
  - $\beta = 0$: Gamma distribution and $\phi = 1/\alpha$ where α is the shape parameter;

Objective function for ARD in NMF

- Combining the prior and likelihood, the objective function (log-posterior) can be written as

$$
\begin{aligned}
C(\mathbf{W}, \mathbf{H}, \boldsymbol{\lambda}) &= -\log p(\mathbf{W}, \mathbf{H}, \boldsymbol{\lambda} \mid \mathbf{V}) \\
&\stackrel{c}{=} \frac{1}{\phi} D_\beta(\mathbf{V} \mid \mathbf{WH}) + \sum_{k=1}^{K} \frac{1}{\lambda_k} \big(f(\mathbf{w}_k) + f(\underline{h}_k) + b \big) + c \log \lambda_k .
\end{aligned}
$$

- Constant ϕ is the dispersion parameter (of the Tweedie distribution):
  - $\beta = 2$: Gaussian distribution and $\phi = \sigma^2$;
  - $\beta = 1$: Poisson distribution and $\phi = 1$;
  - $\beta = 0$: Gamma distribution and $\phi = 1/\alpha$ where α is the shape parameter;
- Constant c and function f depend on the likelihood model:
  - Half Gaussian model: $f(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|^2$ and $c = (F+N)/2 + a + 1$;
  - Exponent model: $f(\mathbf{x}) = \|\mathbf{x}\|_1$ and $c = F + N + a + 1$

Objective function for ARD in NMF

- Combining the prior and likelihood, the objective function (log-posterior) can be written as

$$C(\mathbf{W}, \mathbf{H}, \boldsymbol{\lambda}) = -\log p(\mathbf{W}, \mathbf{H}, \boldsymbol{\lambda} \mid \mathbf{V})$$
$$\stackrel{c}{=} \frac{1}{\phi} D_\beta(\mathbf{V} \mid \mathbf{WH}) + \sum_{k=1}^{K} \frac{1}{\lambda_k}\big(f(\mathbf{w}_k) + f(\underline{h}_k) + b\big) + c \log \lambda_k.$$

- Constant ϕ is the dispersion parameter (of the Tweedie distribution):
  - $\beta = 2$: Gaussian distribution and $\phi = \sigma^2$;
  - $\beta = 1$: Poisson distribution and $\phi = 1$;
  - $\beta = 0$: Gamma distribution and $\phi = 1/\alpha$ where α is the shape parameter;
- Constant c and function f depend on the likelihood model:
  - Half Gaussian model: $f(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|^2$ and $c = (F+N)/2 + a + 1$;
  - Exponent model: $f(\mathbf{x}) = \|\mathbf{x}\|_1$ and $c = F + N + a + 1$
- This cost has connections to reweighted ℓ_1 minimization (Candès et al., 2008) and group LASSO (Yuan and Lin, 2007).

Majorization-Minimization Algorithms for ℓ_2-ARD-NMF

- Using the MM ideas discussed by Cédric, we can derive updates for $\mathbf{W}$ and $\mathbf{H}$:

$$\mathbf{H} \longleftarrow \mathbf{H} \cdot \left(\frac{\mathbf{W}^\top [(\mathbf{W}\mathbf{H})^{\cdot(\beta-2)} \cdot \mathbf{V}]}{\mathbf{W}^\top [(\mathbf{W}\mathbf{H})]^{\cdot(\beta-1)} + \phi \mathbf{H}/\mathrm{repmat}(\boldsymbol{\lambda}, 1, N)} \right)^{\xi(\beta)}$$

$$\mathbf{W} \longleftarrow \mathbf{W} \cdot \left(\frac{[(\mathbf{W}\mathbf{H})^{\cdot(\beta-2)} \cdot \mathbf{V}]\mathbf{H}^\top}{[(\mathbf{W}\mathbf{H})]^{\cdot(\beta-1)}\mathbf{H}^\top + \phi \mathbf{W}/\mathrm{repmat}(\boldsymbol{\lambda}, F, 1)} \right)^{\xi(\beta)}$$

where

$$\xi(\beta) = \begin{cases} 1/(3-\beta) & \beta \le 2 \\ 1/(\beta-1) & \beta > 2 \end{cases}.$$

Majorization-Minimization Algorithms for ℓ_2-ARD-NMF

- Using the MM ideas discussed by Cédric, we can derive updates for $\mathbf{W}$ and $\mathbf{H}$:

$$\mathbf{H} \longleftarrow \mathbf{H} \cdot \left(\frac{\mathbf{W}^\top [(\mathbf{W}\mathbf{H})^{\cdot(\beta-2)} \cdot \mathbf{V}]}{\mathbf{W}^\top [(\mathbf{W}\mathbf{H})]^{\cdot(\beta-1)} + \phi \mathbf{H} / \mathrm{repmat}(\boldsymbol{\lambda}, 1, N)} \right)^{\xi(\beta)}$$

$$\mathbf{W} \longleftarrow \mathbf{W} \cdot \left(\frac{[(\mathbf{W}\mathbf{H})^{\cdot(\beta-2)} \cdot \mathbf{V}]\mathbf{H}^\top}{[(\mathbf{W}\mathbf{H})]^{\cdot(\beta-1)}\mathbf{H}^\top + \phi \mathbf{W} / \mathrm{repmat}(\boldsymbol{\lambda}, F, 1)} \right)^{\xi(\beta)}$$

  where

$$\xi(\beta) = \begin{cases} 1/(3-\beta) & \beta \le 2 \\ 1/(\beta-1) & \beta > 2 \end{cases}.$$

- The update for $\boldsymbol{\lambda}$ is

$$\lambda_k \longleftarrow \frac{\frac{1}{2}\|\mathbf{w}_k\|^2 + \frac{1}{2}\|\underline{h}_k\|^2 + b}{c} \qquad \forall\, k \in [K].$$

Estimating Hyperparameter b via the Method of Moments

- By the law of large numbers

$$\hat{\mu}_{\mathbf{V}} = \frac{1}{FN} \sum_{f',n'} v_{f'n'} \approx \mathbb{E}[v_{fn}] = \mathbb{E}[\hat{v}_{fn}] = \sum_{k} \mathbb{E}[w_{fk} h_{kn}].$$

Estimating Hyperparameter b via the Method of Moments

- By the law of large numbers

$$\hat{\mu}_{\mathbf{V}} = \frac{1}{FN} \sum_{f',n'} v_{f'n'} \approx \mathbb{E}[v_{fn}] = \mathbb{E}[\hat{v}_{fn}] = \sum_{k} \mathbb{E}[w_{fk} h_{kn}].$$

- Can compute $\mathbb{E}[\hat{v}_{fn}] = \sum_k \mathbb{E}[w_{fk} h_{kn}]$ in closed-form for the Half Gaussian and Exponential models using their moments:

$$\mathbb{E}[\hat{v}_{fn}] = \begin{cases} \dfrac{2Kb}{\pi(a-1)} & \text{Half Gaussian} \\ \dfrac{Kb^2}{(a-1)(a-2)} & \text{Exponential} \end{cases}$$

Estimating Hyperparameter b via the Method of Moments

- By the law of large numbers

$$\hat{\mu}_{\mathbf{V}} = \frac{1}{FN} \sum_{f',n'} v_{f'n'} \approx \mathbb{E}[v_{fn}] = \mathbb{E}[\hat{v}_{fn}] = \sum_k \mathbb{E}[w_{fk} h_{kn}].$$

- Can compute $\mathbb{E}[\hat{v}_{fn}] = \sum_k \mathbb{E}[w_{fk} h_{kn}]$ in closed-form for the Half Gaussian and Exponential models using their moments:

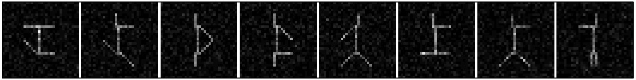
$$\mathbb{E}[\hat{v}_{fn}] = \begin{cases} \dfrac{2Kb}{\pi(a-1)} & \text{Half Gaussian} \\ \dfrac{Kb^2}{(a-1)(a-2)} & \text{Exponential} \end{cases}$$

- Can "invert" these relations to yield

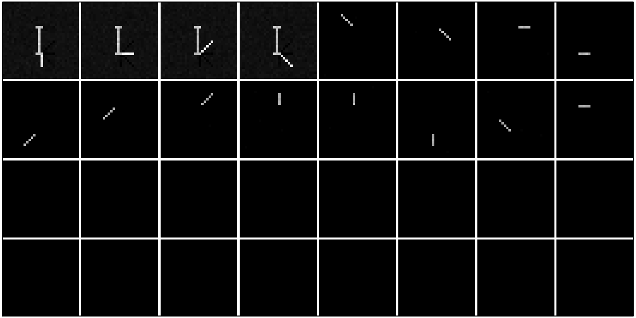
$$\hat{b} = \begin{cases} \dfrac{\pi(a-1)\hat{\mu}_{\mathbf{V}}}{2K} & \text{Half Gaussian} \\ \sqrt{\dfrac{(a-1)(a-2)\hat{\mu}_{\mathbf{V}}}{K}} & \text{Exponential} \end{cases}$$

Swimmer Decomposition Results

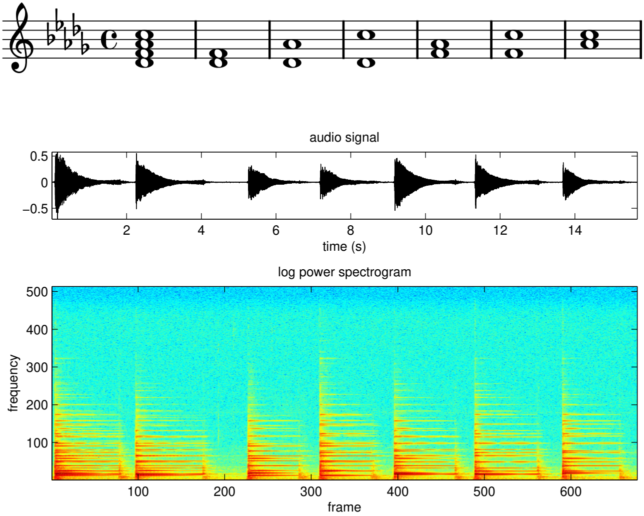
8 data samples (among 256)



Estimated **W** using exponential priors/ℓ_1 penalization



Audio Decomposition Results



Audio Decomposition Results

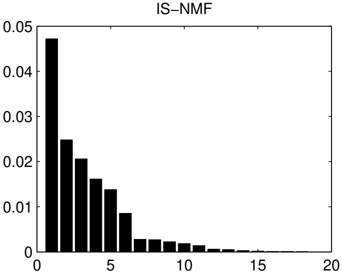


Figure: Histograms of standard deviation values of all $K = 18$ components produced by Itakura–Saito NMF and ARD Itakura–Saito NMF (with ℓ_2 penalization). ARD IS-NMF only retains the 6 "right" components

Audio Decomposition Results

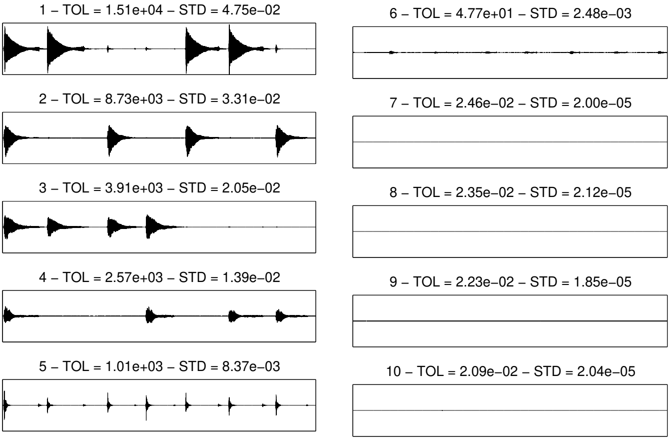


Figure: First 4 components extract the individual notes and the next 2 components extract the sound of hammer hitting the strings and the sound produced by the sustain pedal

Concluding Remarks from using ARD on NMF

- Introduced an Automatic Relevance Determination framework for learning the common/latent dimension K in NMF.

Concluding Remarks from using ARD on NMF

- Introduced an Automatic Relevance Determination framework for learning the common/latent dimension K in NMF.
- Simple, cheap and intuitive.

Concluding Remarks from using ARD on NMF

- Introduced an Automatic Relevance Determination framework for learning the common/latent dimension K in NMF.
- Simple, cheap and intuitive.
- Since its publication, ARD NMF (Tan and Févotte, 2013) has been used successfully in biology and genomics, among other scientific fields, e.g.,

[HTML] Comprehensive molecular characterization of muscle-invasive bladder cancer
AG Robertson, J Kim, H Al-Ahmadie, J Bellmunt, G Guo… - Cell, 2017 - Elsevier
We report a comprehensive analysis of 412 muscle-invasive bladder cancers characterized by multiple TCGA analytical platforms. Fifty-eight genes were significantly mutated, and the …
☆ Save 99 Cite Cited by 1453 Related articles All 24 versions

[HTML] Comprehensive and integrative genomic characterization of hepatocellular carcinoma
A Ally, M Balasundaram, R Carlsen, E Chuah, A Clarke… - Cell, 2017 - Elsevier
Liver cancer has the second highest worldwide cancer mortality rate and has limited therapeutic options. We analyzed 363 hepatocellular carcinoma (HCC) cases by whole …
☆ Save 99 Cite Cited by 1153 Related articles All 17 versions

[HTML] The repertoire of mutational signatures in human cancer
LB Alexandrov, J Kim, NJ Haradhvala, MN Huang… - Nature, 2020 - nature.com
Somatic mutations in cancer genomes are caused by multiple mutational processes, each of which generates a characteristic mutational signature 1. Here, as part of the Pan-Cancer …
☆ Save 99 Cite Cited by 1073 Related articles All 19 versions

Outline

Nonnegative rank selection by automatic relevance determination

Distributionally robust nonnegative matrix factorization

NMF in ranking models and sports analytics

PSDMF and links with phase retrieval and affine rank minimization

Distributionally Robust Nonnegative Matrix Factorization

(Gillis, Hien, Leplat, and Tan, 2022)

Distributionally Robust Nonnegative Matrix Factorization

(Gillis, Hien, Leplat, and Tan, 2022)

- The parameter β in D_β controls the noise statistics on **WH** (Tweedie distn);

Distributionally Robust Nonnegative Matrix Factorization

(Gillis, Hien, Leplat, and Tan, 2022)

- The parameter β in D_β controls the noise statistics on **WH** (Tweedie distn);
- If $v_{fn} = [\mathbf{WH}]_{fn} +$ Gaussian noise ($\beta = 2$), then

$$-\log p(v_{fn} \mid [\mathbf{WH}]_{fn}) \stackrel{c}{=} \frac{1}{2\sigma^2}\big([\mathbf{WH}]_{fn} - v_{fn}\big)^2,$$

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

Distributionally Robust Nonnegative Matrix Factorization

(Gillis, Hien, Leplat, and Tan, 2022)

- The parameter β in D_β controls the noise statistics on $\mathbf{WH}$ (Tweedie distn);
- If $v_{fn} = [\mathbf{WH}]_{fn} +$ Gaussian noise ($\beta = 2$), then

$$-\log p(v_{fn} \mid [\mathbf{WH}]_{fn}) \stackrel{c}{=} \frac{1}{2\sigma^2}\big([\mathbf{WH}]_{fn} - v_{fn}\big)^2,$$

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

- If $v_{fn} \sim$ Poisson($[\mathbf{WH}]_{fn}$) ($\beta = 1$), then

$$-\log p(v_{fn} \mid [\mathbf{WH}]_{fn}) = v_{fn} \log \frac{v_{fn}}{[\mathbf{WH}]_{fn}} + [\mathbf{WH}]_{fn},$$

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

Distributionally Robust Nonnegative Matrix Factorization

(Gillis, Hien, Leplat, and Tan, 2022)

- The parameter β in D_β controls the noise statistics on $\mathbf{WH}$ (Tweedie distn);
- If $v_{fn} = [\mathbf{WH}]_{fn} +$ Gaussian noise ($\beta = 2$), then

$$-\log p(v_{fn} \mid [\mathbf{WH}]_{fn}) \stackrel{c}{=} \frac{1}{2\sigma^2}\big([\mathbf{WH}]_{fn} - v_{fn}\big)^2,$$

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

- If $v_{fn} \sim$ Poisson($[\mathbf{WH}]_{fn}$) ($\beta = 1$), then

$$-\log p(v_{fn} \mid [\mathbf{WH}]_{fn}) = v_{fn} \log \frac{v_{fn}}{[\mathbf{WH}]_{fn}} + [\mathbf{WH}]_{fn},$$

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

- If $v_{fn} \sim$ Gamma($\alpha, [\mathbf{WH}]_{fn}/\alpha$) ($\beta = 0$), then

$$-\log p(v_{fn} \mid [\mathbf{WH}]_{fn}) = \frac{v_{fn}}{[\mathbf{WH}]_{fn}} - \log \frac{v_{fn}}{[\mathbf{WH}]_{fn}} - 1,$$

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

Applications, MO-NMF and DR-NMF

- Audio signal processing (Févotte et al., 2009; Virtanen, 2007): $\beta \in \{0, 1\}$

Applications, MO-NMF and DR-NMF

- Audio signal processing (Févotte et al., 2009; Virtanen, 2007): $\beta \in \{0, 1\}$
- Sparse document datasets (Chi and Kolda, 2012): $\beta \in \{1, 2\}$

Applications, MO-NMF and DR-NMF

- Audio signal processing (Févotte et al., 2009; Virtanen, 2007): $\beta \in \{0, 1\}$
- Sparse document datasets (Chi and Kolda, 2012): $\beta \in \{1, 2\}$
- How to choose an appropriate β when given a new task? Say we only consider $\beta \in \Omega$ where $\Omega \subset \mathbb{R}$ is a finite set, e.g., $\Omega = \{0, 1, 2\}$.

Applications, MO-NMF and DR-NMF

- Audio signal processing (Févotte et al., 2009; Virtanen, 2007): $\beta \in \{0, 1\}$
- Sparse document datasets (Chi and Kolda, 2012): $\beta \in \{1, 2\}$
- How to choose an appropriate β when given a new task? Say we only consider $\beta \in \Omega$ where $\Omega \subset \mathbb{R}$ is a finite set, e.g., $\Omega = \{0, 1, 2\}$.
- Multi-Objective NMF (MO-NMF)

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

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

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

Applications, MO-NMF and DR-NMF

- Audio signal processing (Févotte et al., 2009; Virtanen, 2007): $\beta \in \{0, 1\}$
- Sparse document datasets (Chi and Kolda, 2012): $\beta \in \{1, 2\}$
- How to choose an appropriate β when given a new task? Say we only consider $\beta \in \Omega$ where $\Omega \subset \mathbb{R}$ is a finite set, e.g., $\Omega = \{0, 1, 2\}$.
- Multi-Objective NMF (MO-NMF)

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

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

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

- Distributionally Robust NMF (DR-NMF)

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

Scaling of the Objective

- For the family of β-divergences,

$$D_\beta(\alpha \mathbf{V}, \alpha \mathbf{WH}) = \alpha^\beta D_\beta(\mathbf{V}, \mathbf{WH}) \qquad \forall\, \alpha > 0.$$

Scaling of the Objective

- For the family of β-divergences,

$$D_{\beta}(\alpha\mathbf{V}, \alpha\mathbf{WH}) = \alpha^{\beta} D_{\beta}(\mathbf{V}, \mathbf{WH}) \qquad \forall\, \alpha > 0.$$

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

Scaling of the Objective

- For the family of β-divergences,

$$D_\beta(\alpha \mathbf{V}, \alpha \mathbf{WH}) = \alpha^\beta D_\beta(\mathbf{V}, \mathbf{WH}) \qquad \forall\, \alpha > 0.$$

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

- Compute an approximate solution

$$(\mathbf{W}_\beta, \mathbf{H}_\beta) \approx \arg\min_{\mathbf{W},\mathbf{H} \geq \mathbf{0}} D_\beta(\mathbf{V}, \mathbf{WH}) \quad \text{with error} \quad e_\beta = D_\beta(\mathbf{V}, \mathbf{W}_\beta \mathbf{H}_\beta)$$

and define

$$\overline{D}_\beta(\mathbf{V}, \mathbf{WH}) = \frac{D_\beta(\mathbf{V}, \mathbf{WH})}{e_\beta} \quad \text{so that} \quad \overline{D}_\beta(\mathbf{V}, \mathbf{W}_\beta \mathbf{H}_\beta) = 1.$$

Scaling of the Objective

- For the family of β-divergences,

$$D_\beta(\alpha \mathbf{V}, \alpha \mathbf{WH}) = \alpha^\beta D_\beta(\mathbf{V}, \mathbf{WH}) \qquad \forall\, \alpha > 0.$$

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

- Compute an approximate solution

$$(\mathbf{W}_\beta, \mathbf{H}_\beta) \approx \underset{\mathbf{W},\mathbf{H} \geq \mathbf{0}}{\arg\min}\, D_\beta(\mathbf{V}, \mathbf{WH}) \quad \text{with error} \quad e_\beta = D_\beta(\mathbf{V}, \mathbf{W}_\beta \mathbf{H}_\beta)$$

and define

$$\overline{D}_\beta(\mathbf{V}, \mathbf{WH}) = \frac{D_\beta(\mathbf{V}, \mathbf{WH})}{e_\beta} \quad \text{so that} \quad \overline{D}_\beta(\mathbf{V}, \mathbf{W}_\beta \mathbf{H}_\beta) = 1.$$

- Consider the optimization problem

$$\min_{\mathbf{W},\mathbf{H} \geq \mathbf{0}} \overline{D}_\Omega^{\boldsymbol{\lambda}}(\mathbf{V}, \mathbf{WH}) \quad \text{where} \quad \overline{D}_\Omega^{\boldsymbol{\lambda}}(\mathbf{V}, \mathbf{WH}) = \sum_{\beta \in \Omega} \lambda_\beta \overline{D}_\beta(\mathbf{V}, \mathbf{WH}).$$

Multiplicative Update Algorithm

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

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

Multiplicative Update Algorithm

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

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

- Rescaled gradient descent method (with rescaling matrix $\mathbf{B}$)

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

Multiplicative Update Algorithm

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

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

- Rescaled gradient descent method (with rescaling matrix $\mathbf{B}$)

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

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

Multiplicative Update Algorithm

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

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

- Rescaled gradient descent method (with rescaling matrix $\mathbf{B}$)

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

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

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

Multiplicative Update Algorithm

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

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

- Rescaled gradient descent method (with rescaling matrix $\mathbf{B}$)

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

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

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

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

Application of MU Algorithm to DR-NMF

- Recall that for a fixed probability vector $\boldsymbol{\lambda}$, we want to solve

$$\min_{\mathbf{W},\mathbf{H}\geq 0} \overline{D}_{\Omega}^{\boldsymbol{\lambda}}(\mathbf{V},\mathbf{WH}), \quad \text{where} \quad \overline{D}_{\Omega}^{\boldsymbol{\lambda}}(\mathbf{V},\mathbf{WH}) = \sum_{\beta\in\Omega} \lambda_{\beta}\overline{D}_{\beta}(\mathbf{V},\mathbf{WH}).$$

Application of MU Algorithm to DR-NMF

- Recall that for a fixed probability vector $\boldsymbol{\lambda}$, we want to solve

$$\min_{\mathbf{W},\mathbf{H}\geq 0} \overline{D}_{\Omega}^{\boldsymbol{\lambda}}(\mathbf{V},\mathbf{WH}), \quad \text{where} \quad \overline{D}_{\Omega}^{\boldsymbol{\lambda}}(\mathbf{V},\mathbf{WH}) = \sum_{\beta\in\Omega} \lambda_\beta \overline{D}_\beta(\mathbf{V},\mathbf{WH}).$$

- Alternating minimization procedure: Minimize over $\mathbf{H}$, then over $\mathbf{W}$.

Application of MU Algorithm to DR-NMF

- Recall that for a fixed probability vector $\boldsymbol{\lambda}$, we want to solve

$$\min_{\mathbf{W},\mathbf{H}\geq 0} \overline{D}_{\Omega}^{\boldsymbol{\lambda}}(\mathbf{V},\mathbf{WH}), \quad \text{where} \quad \overline{D}_{\Omega}^{\boldsymbol{\lambda}}(\mathbf{V},\mathbf{WH}) = \sum_{\beta\in\Omega} \lambda_\beta \overline{D}_\beta(\mathbf{V},\mathbf{WH}).$$

- Alternating minimization procedure: Minimize over $\mathbf{H}$, then over $\mathbf{W}$.

- For all β,

$$\nabla^{\mathbf{H}} D_\beta(\mathbf{V},\mathbf{WH}) = \nabla^{\mathbf{H}}_{+} D_\beta(\mathbf{V},\mathbf{WH}) - \nabla^{\mathbf{H}}_{-} D_\beta(\mathbf{V},\mathbf{WH}),$$

where $\nabla^{\mathbf{H}}$ means gradient w.r.t. $\mathbf{H}$.

Application of MU Algorithm to DR-NMF

- Recall that for a fixed probability vector $\boldsymbol{\lambda}$, we want to solve

$$\min_{\mathbf{W},\mathbf{H}\geq 0} \overline{D}_{\Omega}^{\boldsymbol{\lambda}}(\mathbf{V},\mathbf{WH}), \quad \text{where} \quad \overline{D}_{\Omega}^{\boldsymbol{\lambda}}(\mathbf{V},\mathbf{WH}) = \sum_{\beta\in\Omega} \lambda_\beta \overline{D}_\beta(\mathbf{V},\mathbf{WH}).$$

- Alternating minimization procedure: Minimize over $\mathbf{H}$, then over $\mathbf{W}$.

- For all β,

$$\nabla^{\mathbf{H}} D_\beta(\mathbf{V},\mathbf{WH}) = \nabla_{+}^{\mathbf{H}} D_\beta(\mathbf{V},\mathbf{WH}) - \nabla_{-}^{\mathbf{H}} D_\beta(\mathbf{V},\mathbf{WH}),$$

  where $\nabla^{\mathbf{H}}$ means gradient w.r.t. $\mathbf{H}$.

- After some tedious calculation,

$$\nabla_{+}^{\mathbf{H}} D_\beta(\mathbf{V},\mathbf{WH}) = \mathbf{W}^\top (\mathbf{WH})^{\cdot(\beta-1)} \quad \text{and}$$
$$\nabla_{-}^{\mathbf{H}} D_\beta(\mathbf{V},\mathbf{WH}) = \mathbf{W}^\top \left((\mathbf{WH})^{\cdot(\beta-2)} \cdot \mathbf{V}\right),$$

Application of MU Algorithm to DR-NMF

- Update $\mathbf{H}$ as follows:

$$\mathbf{H}^{+} = \mathbf{H} \cdot \frac{\sum_{\beta\in\Omega} \lambda_\beta \left(\nabla_{-}^{\mathbf{H}} \overline{D}_\beta(\mathbf{V}, \mathbf{WH})\right)}{\sum_{\beta\in\Omega} \lambda_\beta \left(\nabla_{+}^{\mathbf{H}} \overline{D}_\beta(\mathbf{V}, \mathbf{WH})\right)}.$$

Application of MU Algorithm to DR-NMF

- Update $\mathbf{H}$ as follows:

$$\mathbf{H}^+ = \mathbf{H} \cdot \frac{\sum_{\beta \in \Omega} \lambda_\beta \big(\nabla_-^{\mathbf{H}} \overline{D}_\beta(\mathbf{V}, \mathbf{WH}) \big)}{\sum_{\beta \in \Omega} \lambda_\beta \big(\nabla_+^{\mathbf{H}} \overline{D}_\beta(\mathbf{V}, \mathbf{WH}) \big)}.$$

- Sometimes this may not result in a decrease in the objective, so we set $\gamma = 1$ and $\mathbf{H}_1^+ = \mathbf{H}^+$ and successively find γ such that while

$$\overline{D}_\Omega^{\boldsymbol{\lambda}}(\mathbf{V}, \mathbf{WH}_\gamma^+) > \overline{D}_\Omega^{\boldsymbol{\lambda}}(\mathbf{V}, \mathbf{WH})$$

we reduce

$$\gamma \longleftarrow \frac{\gamma}{2}$$

and set

$$\mathbf{H}_\gamma^+ = (1-\gamma)\mathbf{H} + \gamma \mathbf{H}^+.$$

Application of MU Algorithm to DR-NMF

- Update $\mathbf{H}$ as follows:

$$\mathbf{H}^{+} = \mathbf{H} \cdot \frac{\sum_{\beta \in \Omega} \lambda_\beta \left(\nabla_{-}^{\mathbf{H}} \overline{D}_\beta(\mathbf{V}, \mathbf{WH}) \right)}{\sum_{\beta \in \Omega} \lambda_\beta \left(\nabla_{+}^{\mathbf{H}} \overline{D}_\beta(\mathbf{V}, \mathbf{WH}) \right)}.$$

- Sometimes this may not result in a decrease in the objective, so we set $\gamma = 1$ and $\mathbf{H}_1^{+} = \mathbf{H}^{+}$ and successively find γ such that while

$$\overline{D}_\Omega^{\boldsymbol{\lambda}}(\mathbf{V}, \mathbf{WH}_\gamma^{+}) > \overline{D}_\Omega^{\boldsymbol{\lambda}}(\mathbf{V}, \mathbf{WH})$$

we reduce

$$\gamma \longleftarrow \frac{\gamma}{2}$$

and set

$$\mathbf{H}_\gamma^{+} = (1 - \gamma)\mathbf{H} + \gamma \mathbf{H}^{+}.$$

- But this tweak of γ is rarely needed.

Algorithm for DR-NMF

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

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

Algorithm for DR-NMF

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

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

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

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

Algorithm for DR-NMF

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

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

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

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

- So we want to solve

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

which is a min-max optimization problem.

Algorithm for DR-NMF

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

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

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

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

- So we want to solve

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

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.

Frank–Wolfe-type Algorithm for DR-NMF

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

Frank–Wolfe-type Algorithm for DR-NMF

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

Frank–Wolfe-type Algorithm for DR-NMF

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

Frank–Wolfe-type Algorithm for DR-NMF

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

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

  Update

$$\boldsymbol{\lambda}^{(t+1)} = (1 - \rho_t)\boldsymbol{\lambda}^{(t)} + \rho_t \boldsymbol{\lambda}_*^{(t)},$$

  where $\rho_t = 1/t$.

Frank–Wolfe-type Algorithm for DR-NMF

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

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

  Update

$$\boldsymbol{\lambda}^{(t+1)} = (1 - \rho_t)\boldsymbol{\lambda}^{(t)} + \rho_t \boldsymbol{\lambda}_*^{(t)},$$

  where $\rho_t = 1/t$.
- This is a Frank–Wolfe-type algorithm (FW would use $\rho_t = 2/(t+2)$).

Remarks on our Algorithm for DR-NMF

- Updates for $\mathbf{W}$ and $\mathbf{H}$ are meant to approximately minimize

$$(\mathbf{W}, \mathbf{H}) \mapsto \overline{D}_{\Omega}^{\boldsymbol{\lambda}^{(t)}}(\mathbf{V}, \mathbf{WH})$$

Remarks on our Algorithm for DR-NMF

- Updates for **W** and **H** are meant to approximately minimize

$$(\mathbf{W}, \mathbf{H}) \mapsto \overline{D}_{\Omega}^{\boldsymbol{\lambda}^{(t)}}(\mathbf{V}, \mathbf{W}\mathbf{H})$$

- For the update of $\boldsymbol{\lambda}$, notice that for all $\beta \in \Omega$

$$\overline{D}_{\beta^*}(\mathbf{V}, \mathbf{W}^{(t+1)}\mathbf{H}^{(t+1)}) \geq \overline{D}_{\beta}(\mathbf{V}, \mathbf{W}^{(t+1)}\mathbf{H}^{(t+1)}),$$

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

$$\boldsymbol{\lambda}_*^{(t)} = \arg\max \left\{ \overline{D}_{\beta}^{\boldsymbol{\lambda}}(\mathbf{V}, \mathbf{W}^{(t+1)}\mathbf{H}^{(t+1)}) : \boldsymbol{\lambda} \geq \mathbf{0}, \|\boldsymbol{\lambda}\|_1 = 1 \right\}.$$

Remarks on our Algorithm for DR-NMF

- Updates for $\mathbf{W}$ and $\mathbf{H}$ are meant to approximately minimize

$$(\mathbf{W}, \mathbf{H}) \mapsto \overline{D}_{\Omega}^{\boldsymbol{\lambda}^{(t)}}(\mathbf{V}, \mathbf{W}\mathbf{H})$$

- For the update of $\boldsymbol{\lambda}$, notice that for all $\beta \in \Omega$

$$\overline{D}_{\beta^*}(\mathbf{V}, \mathbf{W}^{(t+1)}\mathbf{H}^{(t+1)}) \geq \overline{D}_{\beta}(\mathbf{V}, \mathbf{W}^{(t+1)}\mathbf{H}^{(t+1)}),$$

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

$$\boldsymbol{\lambda}_*^{(t)} = \arg\max \left\{ \overline{D}_{\beta}^{\boldsymbol{\lambda}}(\mathbf{V}, \mathbf{W}^{(t+1)}\mathbf{H}^{(t+1)}) : \boldsymbol{\lambda} \geq \mathbf{0}, \|\boldsymbol{\lambda}\|_1 = 1 \right\}.$$

- 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

$$(\mathbf{W},\mathbf{H}) \mapsto \overline{D}_{\Omega}^{\boldsymbol{\lambda}^{(t)}}(\mathbf{V},\mathbf{WH})$$

- For the update of $\boldsymbol{\lambda}$, notice that for all $\beta \in \Omega$

$$\overline{D}_{\beta^*}(\mathbf{V},\mathbf{W}^{(t+1)}\mathbf{H}^{(t+1)}) \geq \overline{D}_{\beta}(\mathbf{V},\mathbf{W}^{(t+1)}\mathbf{H}^{(t+1)}),$$

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

$$\boldsymbol{\lambda}_*^{(t)} = \arg\max \left\{ \overline{D}_{\beta}^{\boldsymbol{\lambda}}(\mathbf{V},\mathbf{W}^{(t+1)}\mathbf{H}^{(t+1)}) : \boldsymbol{\lambda} \geq \mathbf{0}, \|\boldsymbol{\lambda}\|_1 = 1 \right\}.$$

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

- Forcing all β-divergences to decrease as well.

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 do not know this, we can compare DR-NMF, KL-NMF and Frobenius-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 do not know this, we can compare DR-NMF, KL-NMF and Frobenius-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 do not know this, we can compare DR-NMF, KL-NMF and Frobenius-NMF
- Use these NMF methods for clustering (topic modeling)
- Clustering accuracy

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

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

Figure: Clustering accuracies of various methods

Dense Time-Frequency Matrices of Audio Signals

- Use the data set piano_Mary

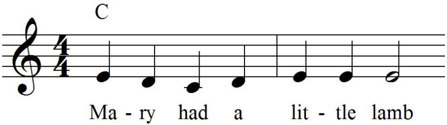


Figure: Musical score of "Mary had a little lamb". The notes are activated as follows: E_4, D_4, C_4, D_4, E_4, E_4, E_4.

- Use the data set piano_Mary

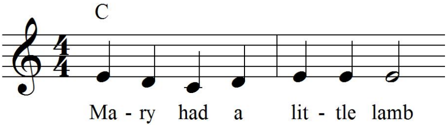


Figure: Musical score of "Mary had a little lamb". The notes are activated as follows: E_4, D_4, C_4, D_4, E_4, E_4, E_4.

- 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

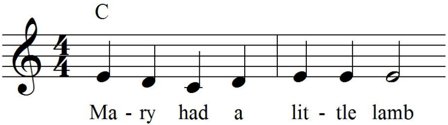


Figure: Musical score of "Mary had a little lamb". The notes are activated as follows: E_4, D_4, C_4, D_4, E_4, E_4, E_4.

- 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

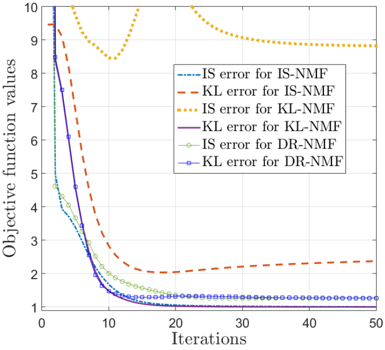


Figure: Evolution of scaled β-divergences

No Added Noise

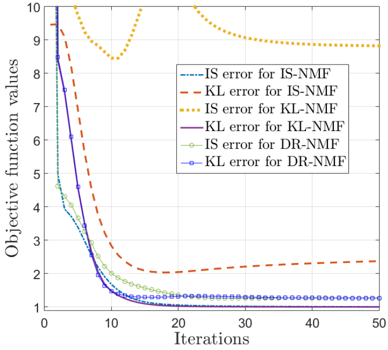


Figure: Evolution of scaled β-divergences

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

No Added Noise

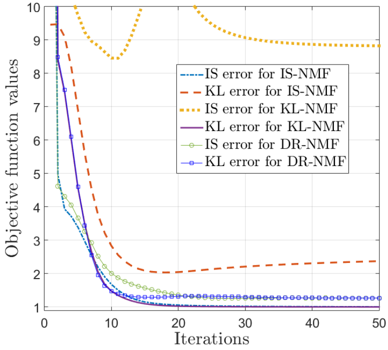


Figure: Evolution of scaled β-divergences

- 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

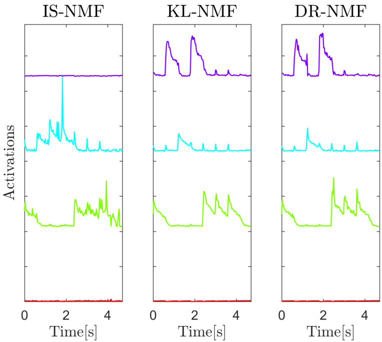


Figure: IS-NMF, KL-NMF, and DR-NMF with $\Omega = \{0, 1\}$ in Poisson noise.

Added Poisson Noise

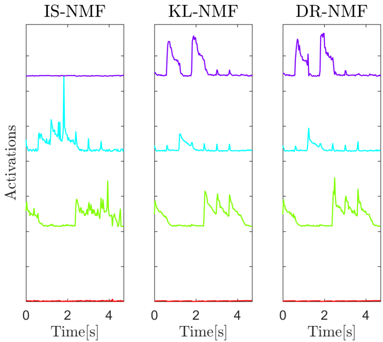


Figure: IS-NMF, KL-NMF, and DR-NMF with $\Omega = \{0, 1\}$ in Poisson noise.

- Rows of $\mathbf{H}$ are recovered successfully.
- C_4 is activated once, D_4 twice and E_4 four times.

Outline

Nonnegative rank selection by automatic relevance determination

Distributionally robust nonnegative matrix factorization

NMF in ranking models and sports analytics

PSDMF and links with phase retrieval and affine rank minimization

Using Nonnegative Matrix Factorization in Ranking Models for Sports Analytics

(Xia, Tan, Filstroff, and Févotte, 2019)

Using Nonnegative Matrix Factorization in Ranking Models for Sports Analytics

(Xia, Tan, Filstroff, and Févotte, 2019)

Rod Laver
3 Grand Slam

Roger Federer
3 Grand Slam

Novak Djokovic
5 Grand Slam

1960s

2005 2006

2010 2011

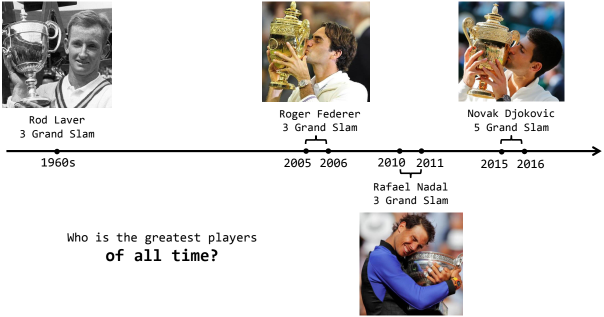
2015 2016

Rafael Nadal
3 Grand Slam

Who is the greatest players
of all time?

Using Nonnegative Matrix Factorization in Ranking Models for Sports Analytics

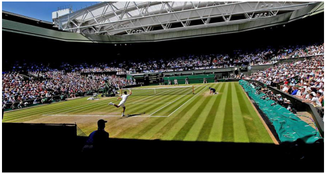
(Xia, Tan, Filstroff, and Févotte, 2019)



Who is the greatest of all time (GOAT)?

What could be a Pertinent Latent Variable?

What could be a Pertinent Latent Variable?



Wimbledon
Grass Outdoors

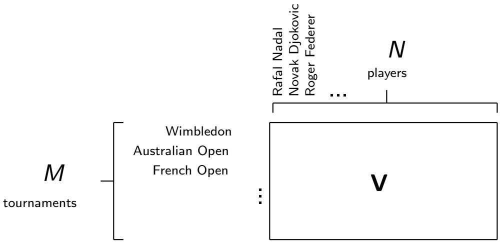
Australian Open
Hard Outdoors

French Open
Clay Outdoors

US Open
Hard Outdoors

Ranking Tennis Players with Latent Variables

Ranking Tennis Players with Latent Variables



Ranking Tennis Players with Latent Variables

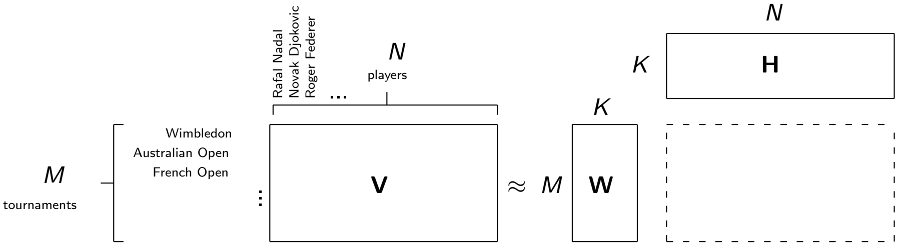


Figure: The hybrid BTL-NMF Model

Ranking Tennis Players with Latent Variables

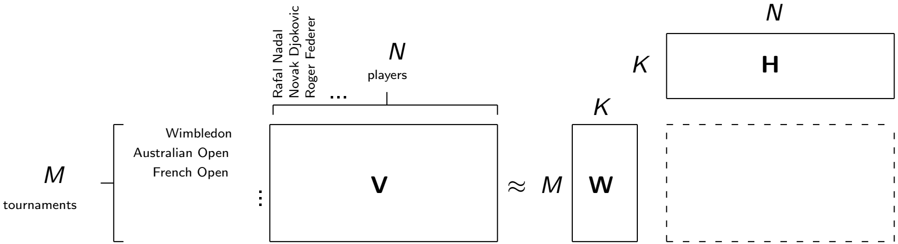


Figure: The hybrid BTL-NMF Model

- Bradley–Terry–Luce (Bradley and Terry, 1952; Luce, 1959) ranking model:

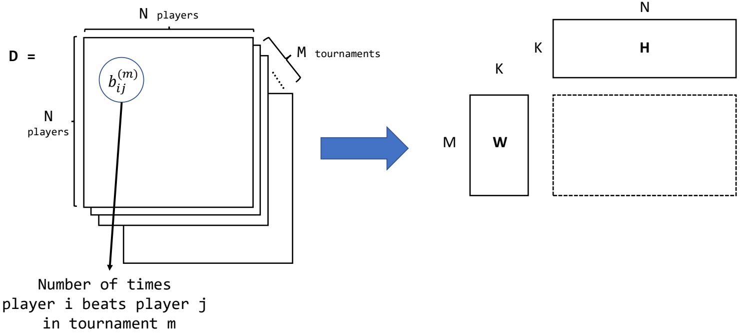
$$\Pr\left(\text{player } i \text{ beats player } j \mid \text{tournament } m\right) = \frac{\lambda_{mi}}{\lambda_{mi} + \lambda_{mj}}$$

- λ_{mi}: Skill level of player i in tournament m.

Data Collected and Likelihood Function

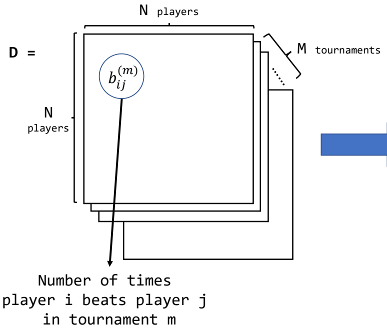
GIVEN: D ~ M (N x N) matrices

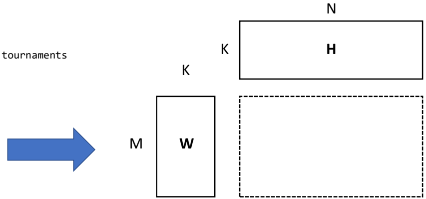
Output: W ~ (M x K) matrix
H ~ (K x N) matrix



Data Collected and Likelihood Function

GIVEN: D ~ M (N x N) matrices → **Output: W ~ (M x K) matrix**
H ~ (K x N) matrix





Likelihood Function

$$L(\mathbf{W}, \mathbf{H} \mid \mathcal{D} = \{b_{ij}^{(m)}\})$$

$$= \prod_{m=1}^{M} \prod_{(i,j)} \left(\underbrace{\frac{[\mathbf{WH}]_{mi}}{[\mathbf{WH}]_{mi} + [\mathbf{WH}]_{mj}}}_{\text{Prob. } i \text{ beats } j \text{ in tourn. } m} \right)^{b_{ij}^{(m)}}$$

Objective Function to be Minimized

- Take the negative log of the likelihood to get the following objective function

$$\begin{aligned}
&\operatorname*{arg\,min}_{\mathbf{W},\mathbf{H}\geq\mathbf{0}} f(\mathbf{W},\mathbf{H} \mid \mathcal{D}) = -\log L\big(\mathbf{W},\mathbf{H} \mid \mathcal{D}\big) \\
&\equiv \operatorname*{arg\,min}_{\mathbf{W},\mathbf{H}\geq\mathbf{0}} \sum_{m=1}^{M} \sum_{(i,j)\in\mathcal{P}_m} b_{ij}^{(m)} \left[-\log([\mathbf{WH}]_{mi}) + \log([\mathbf{WH}]_{mi} + [\mathbf{WH}]_{mj})\right],
\end{aligned}$$

where $\mathcal{P}_m$ is the set of games that i and j played in tournament m.

Objective Function to be Minimized

- Take the negative log of the likelihood to get the following objective function

$$
\begin{aligned}
&\arg\min_{\mathbf{W},\mathbf{H}\geq\mathbf{0}} f(\mathbf{W},\mathbf{H} \mid \mathcal{D}) = -\log L\big(\mathbf{W},\mathbf{H} \mid \mathcal{D}\big) \\
&\equiv \arg\min_{\mathbf{W},\mathbf{H}\geq\mathbf{0}} \sum_{m=1}^{M} \sum_{(i,j)\in\mathcal{P}_m} b_{ij}^{(m)} \left[-\log([\mathbf{WH}]_{mi}) + \log([\mathbf{WH}]_{mi} + [\mathbf{WH}]_{mj})\right],
\end{aligned}
$$

  where $\mathcal{P}_m$ is the set of games that i and j played in tournament m.

- Unfortunately, this objective function is not convex in $(\mathbf{W},\mathbf{H})$.

Objective Function to be Minimized

- Take the negative log of the likelihood to get the following objective function

$$\begin{aligned}
&\arg\min_{\mathbf{W},\mathbf{H}\geq\mathbf{0}} f(\mathbf{W},\mathbf{H} \mid \mathcal{D}) = -\log L\big(\mathbf{W},\mathbf{H} \mid \mathcal{D}\big) \\
\equiv\ &\arg\min_{\mathbf{W},\mathbf{H}\geq\mathbf{0}} \sum_{m=1}^{M} \sum_{(i,j)\in\mathcal{P}_m} b_{ij}^{(m)} \left[-\log([\mathbf{WH}]_{mi}) + \log([\mathbf{WH}]_{mi} + [\mathbf{WH}]_{mj})\right],
\end{aligned}$$

  where $\mathcal{P}_m$ is the set of games that i and j played in tournament m.

- Unfortunately, this objective function is not convex in $(\mathbf{W},\mathbf{H})$.

- Majorization-Minimization (MM) comes to the rescue again!

Objective Function to be Minimized

- Take the negative log of the likelihood to get the following objective function

$$\operatorname*{arg\,min}_{\mathbf{W},\mathbf{H}\geq\mathbf{0}} f(\mathbf{W},\mathbf{H} \mid \mathcal{D}) = -\log L\big(\mathbf{W},\mathbf{H} \mid \mathcal{D}\big)$$

$$\equiv \operatorname*{arg\,min}_{\mathbf{W},\mathbf{H}\geq\mathbf{0}} \sum_{m=1}^{M} \sum_{(i,j)\in\mathcal{P}_m} b_{ij}^{(m)} \left[-\log([\mathbf{WH}]_{mi}) + \log([\mathbf{WH}]_{mi} + [\mathbf{WH}]_{mj})\right],$$

where $\mathcal{P}_m$ is the set of games that i and j played in tournament m.

- Unfortunately, this objective function is not convex in $(\mathbf{W},\mathbf{H})$.
- Majorization-Minimization (MM) comes to the rescue again!
- Main ideas: For any concave function g (tangent inequality),

$$g(\mathbf{y}) \leq g(\mathbf{x}) + \nabla g(\mathbf{x})^\top(\mathbf{y}-\mathbf{x})$$

and Jensen's inequality for the convex function $t \mapsto -\log t$.

Majorization-Minimization Updates

- After some straightforward but tedious algebra, we can construct two auxiliary functions $u_1(\mathbf{W}, \tilde{\mathbf{W}} \mid \mathbf{H})$ and $u_2(\mathbf{H}, \tilde{\mathbf{H}} \mid \mathbf{W})$ that majorize the objective function

$$f(\mathbf{W}, \mathbf{H} \mid \mathcal{D}) = -\log L\big(\mathbf{W}, \mathbf{H} \mid \mathcal{D}\big).$$

Majorization-Minimization Updates

- After some straightforward but tedious algebra, we can construct two auxiliary functions $u_1(\mathbf{W}, \tilde{\mathbf{W}} \mid \mathbf{H})$ and $u_2(\mathbf{H}, \tilde{\mathbf{H}} \mid \mathbf{W})$ that majorize the objective function

$$f(\mathbf{W}, \mathbf{H} \mid \mathcal{D}) = -\log L\big(\mathbf{W}, \mathbf{H} \mid \mathcal{D}\big).$$

- Implement

$$\mathbf{W}^{(t+1)} = \operatorname*{arg\,min}_{\mathbf{W} \geq \mathbf{0}} u_1(\mathbf{W}, \mathbf{W}^{(t)} \mid \mathbf{H}^{(t)})$$
$$\mathbf{H}^{(t+1)} = \operatorname*{arg\,min}_{\mathbf{H} \geq \mathbf{0}} u_2(\mathbf{H}, \mathbf{H}^{(t)} \mid \mathbf{W}^{(t+1)})$$

Majorization-Minimization Updates

- Update for $\mathbf{W}$:

$$w_{mk} \longleftarrow \frac{\sum\limits_{(i,j)\in\mathcal{P}_m} b_{ij}^{(m)} \frac{w_{mk}h_{ki}}{[\mathbf{WH}]_{mi}}}{\sum\limits_{(i,j)\in\mathcal{P}_m} b_{ij}^{(m)} \frac{h_{ki}+h_{kj}}{[\mathbf{WH}]_{mi}+[\mathbf{WH}]_{mj}}}.$$

Majorization-Minimization Updates

- Update for $\mathbf{W}$:

$$w_{mk} \longleftarrow \frac{\sum\limits_{(i,j)\in\mathcal{P}_m} b_{ij}^{(m)} \frac{w_{mk} h_{ki}}{[\mathbf{WH}]_{mi}}}{\sum\limits_{(i,j)\in\mathcal{P}_m} b_{ij}^{(m)} \frac{h_{ki}+h_{kj}}{[\mathbf{WH}]_{mi}+[\mathbf{WH}]_{mj}}}.$$

- Update for $\mathbf{H}$:

$$h_{ki} \longleftarrow \frac{\sum\limits_{m} \sum\limits_{j\neq i:(i,j)\in\mathcal{P}_m} b_{ij}^{(m)} \frac{w_{mk} h_{ki}}{[\mathbf{WH}]_{mi}}}{\sum\limits_{m} \sum\limits_{j\neq i:(i,j)\in\mathcal{P}_m} \left(b_{ij}^{(m)} + b_{ji}^{(m)}\right) \frac{w_{mk}}{[\mathbf{WH}]_{mi}+[\mathbf{WH}]_{mj}}}.$$

Majorization-Minimization Updates

- Update for $\mathbf{W}$:

$$w_{mk} \longleftarrow \frac{\sum_{(i,j)\in\mathcal{P}_m} b_{ij}^{(m)} \frac{w_{mk} h_{ki}}{[\mathbf{WH}]_{mi}}}{\sum_{(i,j)\in\mathcal{P}_m} b_{ij}^{(m)} \frac{h_{ki}+h_{kj}}{[\mathbf{WH}]_{mi}+[\mathbf{WH}]_{mj}}}.$$

- Update for $\mathbf{H}$:

$$h_{ki} \longleftarrow \frac{\sum_{m} \sum_{j\neq i:(i,j)\in\mathcal{P}_m} b_{ij}^{(m)} \frac{w_{mk} h_{ki}}{[\mathbf{WH}]_{mi}}}{\sum_{m} \sum_{j\neq i:(i,j)\in\mathcal{P}_m} \left(b_{ij}^{(m)} + b_{ji}^{(m)}\right) \frac{w_{mk}}{[\mathbf{WH}]_{mi}+[\mathbf{WH}]_{mj}}}.$$

- Simple, fuss-free updates.
- Used a few other hacks to ensure normalization and no divide by 0 errors.
- Under the right conditions, can prove convergence guarantees to "stationary points" (Zhao and Tan, 2018).

Data Collection

N = 20

Australian Open
Roland Garros
Wimbledon
US Open
Indian Wells Masters
Madrid Open
Miami Open
Monte-Carlo Masters
Pairs Masters
Italian Open
Canada Masters
Cincinnati Masters
Shanghai Masters
ATP Finals

M = 14

4 Grand Slam + 10 Most Famous ATP tournaments

Top 20 players who both

Rafael Nadal
Novak Djokovic
David Ferrer
Tomas Berdych
Roger Federer
Andy Murray
Fernando Verdasco
Philipp Kohlschreiber
Richard Gasquet
Gilles Simon
Stan Wawrinka
Jo-Wilfried Tsonga
Marin Cilic
Feliciano Lopez
John Isner
Nicolas Almagro
Juan Martin del Potro
Gael Monfils
Milos Raonic
Kei Nishikori

Have the highest number of participation in the 14 tournaments from 2007-2017 $\cap$ Have the highest total number of matches played from 2007-2017

Data Collection

Name	Rafael Nadal	Novak Djokovic	David Ferrer	Tomas Berdych	Roger Federer	Andy Murray	Fernando Verdasco	Philipp Kohlschreiber	Richard Gasquet
Rafael Nadal	0	0	0	1	3	1	1	2	0
Novak Djokovic	1	0	3	2	3	5	2	0	0
David Ferrer	1	0	0	0	0	0	0	0	1
Tomas Berdych	1	0	0	0	0	0	1	1	1
Roger Federer	1	1	0	4	0	2	0	0	0
Andy Murray	1	0	2	1	1	0	1	0	0
Fernando Verdasco	1	0	0	0	0	1	0	0	0
Philipp Kohlschreiber	0	0	0	0	0	0	0	0	0
Richard Gasquet	0	0	0	0	0	0	0	0	0

Zeros on the diagonal ($b_{ii}^{(m)} = 0$)

Non-zero

True zeros ($b_{ij}^{(m)} = 0, b_{ji}^{(m)} > 0$)

Missing data ($b_{ij}^{(m)} = b_{ji}^{(m)} = 0$)

	Male	
Total Entries	$14 \times 20 \times 20 = 5600$	
	Number	Percentage
Non-zero	1024	18.30%
Zeros on the diagonal	280	5.00%
Missing data	3478	62.10%
True zeros	818	14.60%

Results on Tournaments for Men's Dataset

Tournaments	Row Normalization: non-clay	Row Normalization: clay	Column Normalization		
Australian Open	5.77E-01	4.23E-01	1.15E-01	7.66E-02	
French Open	3.44E-01	6.56E-01	8.66E-02	1.50E-01	1
Wimbledon	6.43E-01	3.57E-01	6.73E-02	3.38E-02	
US Open	5.07E-01	4.93E-01	4.62E-02	4.06E-02	
Indian Wells Masters	6.52E-01	3.48E-01	1.34E-01	6.50E-02	
Madrid Open	3.02E-01	6.98E-01	6.43E-02	1.34E-01	3
Miami Open	5.27E-01	4.73E-01	4.95E-02	4.02E-02	
Monte-Carlo Masters	1.68E-01	8.32E-01	2.24E-02	1.01E-01	4
Paris Masters	1.68E-01	8.32E-01	1.29E-02	5.76E-02	
Italian Open	0.00E-00	1.00E-00	1.82E-104	1.36E-01	2
Canadian Open	1.00E-00	0.00E-00	1.28E-01	1.78E-152	
Cincinnati Masters	5.23E-01	4.77E-01	1.13E-01	9.36E-02	
Shanghai Masters	7.16E-01	2.84E-01	1.13E-01	4.07E-02	
The ATP Finals	5.72E-01	4.28E-01	4.59E-02	3.11E-02	

Latent variable discovered to be "surface type"

Results on Player Rankings by Latent Variable

	Players	non-clay	clay	Total Matches
		matrix $\mathbf{H}^T$		
Hard Court player →	Novak Djokovic	1.20E-01	9.98E-02	283
Clay player →	Rafael Nadal	2.48E-02	1.55E-01	241
Grass player →	Roger Federer	1.15E-01	2.34E-02	229
Non-clay player →	Andy Murray	7.57E-02	8.43E-03	209
	Tomas Berdych	0.00E-00	3.02E-02	154
	David Ferrer	6.26E-40	3.27E-02	147
Clay player →	Stan Wawrinka	2.93E-55	4.08E-02	141
	Jo-Wilfried Tsonga	3.36E-02	2.71E-03	121
	Richard Gasquet	5.49E-03	1.41E-02	102
	Juan Martin del Potro	2.90E-02	1.43E-02	101
	Marin Cilic	2.12E-02	0.00E-00	100
	Fernando Verdasco	1.36E-02	8.79E-03	96
	Kei Nishikori	7.07E-03	2.54E-02	94
	Gilles Simon	1.32E-02	4.59E-03	83
	Milos Raonic	1.45E-02	7.25E-03	78
	Philipp Kohlschreiber	2.18E-06	5.35E-03	76
	John Isner	2.70E-03	1.43E-02	78
	Feliciano Lopez	1.43E-02	3.31E-03	75
	Gael Monfils	3.86E-21	1.33E-02	70
	Nicolas Almagro	6.48E-03	6.33E-06	60

Figure: Players rankings according to discovered latent variable – “surface type”

Results on Player Rankings by Tournament

Tournament	Novak Djokovic	Rafael Nadal	Roger Federer	Andy Murray	Stan Wawrinka
Australian Open	2.16E-02	1.54E-02	1.47E-02	9.13E-03	3.34E-03
French Open	1.39E-02	1.43E-02	7.12E-03	4.11E-03	3.48E-03 5
Wimbledon	2.63E-02	1.66E-02	1.91E-02	1.20E-02	3.39E-03
US Open	1.17E-02	9.42E-03	7.38E-03	4.51E-03	2.13E-03
Indian Wells Masters	2.29E-02	1.42E-02	1.68E-02	1.06E-02	2.88E-03
Madrid Open	1.38E-02	1.51E-02	6.63E-03	3.75E-03	3.72E-03 4
Miami Open	2.95E-02	2.30E-02	1.90E-02	1.17E-02	5.15E-03 1
Monte-Carlo Masters	1.19E-02	1.53E-02	4.46E-03	2.27E-03	3.92E-03 3
Paris Masters	7.29E-03	9.37E-03	2.73E-03	1.39E-03	2.40E-03
Italian Open	1.19E-02	1.84E-02	2.78E-03	1.00E-03	4.87E-03 2
Canadian Open	1.16E-02	2.40E-03	1.11E-02	7.32E-03	2.42E-51
Cincinnati Masters	1.82E-02	1.43E-02	1.17E-02	7.17E-03	3.20E-03
Shanghai Masters	8.12E-03	4.38E-03	6.29E-03	4.01E-03	8.24E-04
The ATP Finals	1.13E-02	8.13E-03	7.63E-03	4.74E-03	1.77E-03

Figure: Players' skill levels according to tournaments

Concluding Remarks for NMF for Sports Analytics

- Developed a statistical model that is a hybrid between the BTL ranking/pairwise comparison model

$$\Pr\left(\text{player } i \text{ beats player } j \mid \text{tournament } m\right) = \frac{\lambda_{mi}}{\lambda_{mi} + \lambda_{mj}}$$

and nonnegative matrix factorization (NMF).

Concluding Remarks for NMF for Sports Analytics

- Developed a statistical model that is a hybrid between the BTL ranking/pairwise comparison model

$$\Pr\left(\text{player } i \text{ beats player } j \mid \text{tournament } m\right) = \frac{\lambda_{mi}}{\lambda_{mi} + \lambda_{mj}}$$

  and nonnegative matrix factorization (NMF).

- Developed simple update rules based on MM that are easy to implement and have convergence guarantees.

Concluding Remarks for NMF for Sports Analytics

- Developed a statistical model that is a hybrid between the BTL ranking/pairwise comparison model

$$\Pr\left(\text{player } i \text{ beats player } j \mid \text{tournament } m\right) = \frac{\lambda_{mi}}{\lambda_{mi} + \lambda_{mj}}$$

  and nonnegative matrix factorization (NMF).

- Developed simple update rules based on MM that are easy to implement and have convergence guarantees.

- Confirms our intuition that court surface is a pertinent latent variable.

Concluding Remarks for NMF for Sports Analytics

- Developed a statistical model that is a hybrid between the BTL ranking/pairwise comparison model

$$\Pr\left(\text{player } i \text{ beats player } j \mid \text{tournament } m\right) = \frac{\lambda_{mi}}{\lambda_{mi} + \lambda_{mj}}$$

  and nonnegative matrix factorization (NMF).

- Developed simple update rules based on MM that are easy to implement and have convergence guarantees.

- Confirms our intuition that court surface is a pertinent latent variable.

- Ranked players according to the discovered latent variable (court surface) over a time window of 10 years.

Outline

Nonnegative rank selection by automatic relevance determination

Distributionally robust nonnegative matrix factorization

NMF in ranking models and sports analytics

PSDMF and links with phase retrieval and affine rank minimization

Positive Semidefinite Matrix Factorization (PSDMF)

(Fiorini, Massar, Pokutta, Tiwary, and Wolf, 2012; Gouveia, Parrilo, and Thomas, 2013)

Positive Semidefinite Matrix Factorization (PSDMF)

(Fiorini, Massar, Pokutta, Tiwary, and Wolf, 2012; Gouveia, Parrilo, and Thomas, 2013)

- Given a nonnegative matrix $\mathbf{V} \in \mathbb{R}_+^{F \times N}$, find (symmetric) $K \times K$ positive semidefinite (PSD) matrices $\mathbf{W}_f, f = 1, \ldots, F$ and $\mathbf{H}_n, n = 1, \ldots, N$ such that

$$v_{fn} = [\mathbf{V}]_{fn} = \underbrace{\langle \mathbf{W}_f, \mathbf{H}_n \rangle}_{\text{matrix inner product}} = \underbrace{\mathrm{Tr}(\mathbf{W}_f \mathbf{H}_n)}_{\text{trace}}.$$

Positive Semidefinite Matrix Factorization (PSDMF)

(Fiorini, Massar, Pokutta, Tiwary, and Wolf, 2012; Gouveia, Parrilo, and Thomas, 2013)

- Given a nonnegative matrix $\mathbf{V} \in \mathbb{R}_+^{F \times N}$, find (symmetric) $K \times K$ positive semidefinite (PSD) matrices $\mathbf{W}_f, f = 1, \ldots, F$ and $\mathbf{H}_n, n = 1, \ldots, N$ such that

$$v_{fn} = [\mathbf{V}]_{fn} = \underbrace{\langle \mathbf{W}_f, \mathbf{H}_n \rangle}_{\text{matrix inner product}} = \underbrace{\mathrm{Tr}(\mathbf{W}_f \mathbf{H}_n)}_{\text{trace}}.$$

- The PSD rank of $\mathbf{V}$ is smallest K such that $\mathbf{V}$ admits an exact PSD factorization.

Positive Semidefinite Matrix Factorization (PSDMF)

(Fiorini, Massar, Pokutta, Tiwary, and Wolf, 2012; Gouveia, Parrilo, and Thomas, 2013)

- Given a nonnegative matrix $\mathbf{V} \in \mathbb{R}_+^{F \times N}$, find (symmetric) $K \times K$ positive semidefinite (PSD) matrices $\mathbf{W}_f, f = 1, \ldots, F$ and $\mathbf{H}_n, n = 1, \ldots, N$ such that

$$v_{fn} = [\mathbf{V}]_{fn} = \underbrace{\langle \mathbf{W}_f, \mathbf{H}_n \rangle}_{\text{matrix inner product}} = \underbrace{\mathrm{Tr}(\mathbf{W}_f \mathbf{H}_n)}_{\text{trace}}.$$

- The PSD rank of $\mathbf{V}$ is smallest K such that $\mathbf{V}$ admits an exact PSD factorization.

- If $\{\mathbf{W}_f\}$ and $\{\mathbf{H}_n\}$ are diagonal, let

$$\mathbf{w}_f = \mathrm{diag}\,(\mathbf{W}_f) \in \mathbb{R}_+^K, \quad \text{and} \quad \mathbf{h}_n = \mathrm{diag}\,(\mathbf{H}_n) \in \mathbb{R}_+^K,$$

then

$$v_{fn} = [\mathbf{V}]_{fn} = \underbrace{\langle \mathbf{w}_f, \mathbf{h}_n \rangle}_{\text{vector inner product}} = \sum_k w_{fk} h_{kn}.$$

PSDMF reduces to NMF!

PSDMF and PSD Rank

- Extension linking NMF with geometric and linear constraints in linear programming (Yannakakis, 1991)

PSDMF and PSD Rank

- Extension linking NMF with geometric and linear constraints in linear programming (Yannakakis, 1991)
- The smallest number K such that a polytope can be written as a projection (a "shadow") of a spectrahedron of size K (an affine slice of the cone of $K \times K$ positive semidefnite matrices $\mathbb{S}_+^K$) is equal to the PSD rank of a slack matrix of the original polytope.

- Extension linking NMF with geometric and linear constraints in linear programming (Yannakakis, 1991)
- The smallest number K such that a polytope can be written as a projection (a "shadow") of a spectrahedron of size K (an affine slice of the cone of $K \times K$ positive semidefnite matrices $\mathbb{S}_+^K$) is equal to the PSD rank of a slack matrix of the original polytope.
- Example: Slack matrix of the square.

$$S_4 = \begin{bmatrix} 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 \end{bmatrix},$$

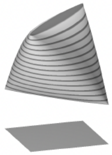


Figure: From Averkov et al. (2018)

$\text{rank}(S_4) = 3$, $\text{nn-rank}(S_4) = 4$ and $\text{psd-rank}(S_4) = 3$, a spectrahedron in $\mathbb{S}_+^3$.

Other Motivations for PSDMF

- Of fundamental importance in various fields:
  - Combinatorial optimization (Gouveia et al., 2013; Fawzi et al., 2015);
  - Quantum information theory (Fiorini et al., 2012; Fawzi et al., 2015);
  - Quantum communications and quantum computing (Jain et al., 2013; van Apeldoorn et al., 2020);
  - Probabilistic modeling (Glasser et al., 2019);
  - Quantum-based recommendation systems (Stark, 2016).

Other Motivations for PSDMF

- Of fundamental importance in various fields:
  - Combinatorial optimization (Gouveia et al., 2013; Fawzi et al., 2015);
  - Quantum information theory (Fiorini et al., 2012; Fawzi et al., 2015);
  - Quantum communications and quantum computing (Jain et al., 2013; van Apeldoorn et al., 2020);
  - Probabilistic modeling (Glasser et al., 2019);
  - Quantum-based recommendation systems (Stark, 2016).

- Connection to quantum is because quantum measurements $\{\mathbf{M}_i\}$, known as positve operator valued measures (POVMs) are PSD and sum to the identity

$$\sum_i \mathbf{M}_i = \mathbf{I}.$$

Other Motivations for PSDMF

- Of fundamental importance in various fields:
  - Combinatorial optimization (Gouveia et al., 2013; Fawzi et al., 2015);
  - Quantum information theory (Fiorini et al., 2012; Fawzi et al., 2015);
  - Quantum communications and quantum computing (Jain et al., 2013; van Apeldoorn et al., 2020);
  - Probabilistic modeling (Glasser et al., 2019);
  - Quantum-based recommendation systems (Stark, 2016).
- Connection to quantum is because quantum measurements $\{\mathbf{M}_i\}$, known as positve operator valued measures (POVMs) are PSD and sum to the identity

$$\sum_i \mathbf{M}_i = \mathbf{I}.$$

- We are mainly concerned with algorithms and approximate factorization.

Objective Function

- Consider the PSMDF model

$$v_{fn} = [\mathbf{V}]_{fn} = \langle \mathbf{W}_f, \mathbf{H}_n \rangle = \mathrm{Tr}(\mathbf{W}_f \mathbf{H}_n)$$

Objective Function

- Consider the PSMDF model

$$v_{fn} = [\mathbf{V}]_{fn} = \langle \mathbf{W}_f, \mathbf{H}_n \rangle = \mathrm{Tr}(\mathbf{W}_f \mathbf{H}_n)$$

- PSD matrices $\{\mathbf{W}_f\}_{f=1}^{F}$ and $\{\mathbf{H}_n\}_{n=1}^{N}$ can be estimated by minimizing a quadratic objective function (Stark, 2016; Vandaele et al., 2018):

$$g(\{\mathbf{W}_f\}_{f=1}^{F}, \{\mathbf{H}_n\}_{n=1}^{N}) = \frac{1}{2} \sum_{f,n} \left(v_{fn} - \mathrm{Tr}(\mathbf{W}_f \mathbf{H}_n)\right)^2$$

Objective Function

- Consider the PSMDF model

$$v_{fn} = [\mathbf{V}]_{fn} = \langle \mathbf{W}_f, \mathbf{H}_n \rangle = \mathrm{Tr}(\mathbf{W}_f \mathbf{H}_n)$$

- PSD matrices $\{\mathbf{W}_f\}_{f=1}^{F}$ and $\{\mathbf{H}_n\}_{n=1}^{N}$ can be estimated by minimizing a quadratic objective function (Stark, 2016; Vandaele et al., 2018):

$$g(\{\mathbf{W}_f\}_{f=1}^{F}, \{\mathbf{H}_n\}_{n=1}^{N}) = \frac{1}{2} \sum_{f,n} \left(v_{fn} - \mathrm{Tr}(\mathbf{W}_f \mathbf{H}_n)\right)^2$$

- For fixed $\{\mathbf{W}_f\}_{f=1}^{F}$, g is convex in $\{\mathbf{H}_n\}_{n=1}^{N}$ and vice versa (Vandaele et al., 2018).

Objective Function

- Consider the PSMDF model

$$v_{fn} = [\mathbf{V}]_{fn} = \langle \mathbf{W}_f, \mathbf{H}_n \rangle = \mathrm{Tr}(\mathbf{W}_f \mathbf{H}_n)$$

- PSD matrices $\{\mathbf{W}_f\}_{f=1}^F$ and $\{\mathbf{H}_n\}_{n=1}^N$ can be estimated by minimizing a quadratic objective function (Stark, 2016; Vandaele et al., 2018):

$$g(\{\mathbf{W}_f\}_{f=1}^F, \{\mathbf{H}_n\}_{n=1}^N) = \frac{1}{2} \sum_{f,n} (v_{fn} - \mathrm{Tr}(\mathbf{W}_f \mathbf{H}_n))^2$$

- For fixed $\{\mathbf{W}_f\}_{f=1}^F$, g is convex in $\{\mathbf{H}_n\}_{n=1}^N$ and vice versa (Vandaele et al., 2018).

- Other objective functions are possible (Glasser et al., 2019; Basu et al., 2016; Lahat and Févotte, 2021)

Alternating Minimization

- Minimize the objective function g w.r.t. $\{\mathbf{H}_n\}_{n=1}^{N}$ for fixed $\{\mathbf{W}_f\}_{f=1}^{F}$, i.e.,

$$\{\mathbf{H}_n^{+}\}_{n=1}^{N} = \arg\min_{\{\mathbf{H}_n\}_{n=1}^{N}} g(\{\mathbf{W}_f\}_{f=1}^{F}, \{\mathbf{H}_n\}_{n=1}^{N})$$

Alternating Minimization

- Minimize the objective function g w.r.t. $\{\mathbf{H}_n\}_{n=1}^N$ for fixed $\{\mathbf{W}_f\}_{f=1}^F$, i.e.,

$$\{\mathbf{H}_n^+\}_{n=1}^N = \underset{\{\mathbf{H}_n\}_{n=1}^N}{\arg\min}\, g(\{\mathbf{W}_f\}_{f=1}^F, \{\mathbf{H}_n\}_{n=1}^N)$$

- Change roles, i.e.,

$$\{\mathbf{W}_f^+\}_{f=1}^F = \underset{\{\mathbf{W}_f\}_{f=1}^F}{\arg\min}\, g(\{\mathbf{W}_f\}_{f=1}^F, \{\mathbf{H}_n^+\}_{n=1}^N)$$

Alternating Minimization

- Minimize the objective function g w.r.t. $\{\mathbf{H}_n\}_{n=1}^N$ for fixed $\{\mathbf{W}_f\}_{f=1}^F$, i.e.,

$$\{\mathbf{H}_n^+\}_{n=1}^N = \underset{\{\mathbf{H}_n\}_{n=1}^N}{\arg\min}\, g(\{\mathbf{W}_f\}_{f=1}^F, \{\mathbf{H}_n\}_{n=1}^N)$$

- Change roles, i.e.,

$$\{\mathbf{W}_f^+\}_{f=1}^F = \underset{\{\mathbf{W}_f\}_{f=1}^F}{\arg\min}\, g(\{\mathbf{W}_f\}_{f=1}^F, \{\mathbf{H}_n^+\}_{n=1}^N)$$

- Repeat until a stopping criterion is achieved;

Alternating Minimization

- Minimize the objective function g w.r.t. $\{\mathbf{H}_n\}_{n=1}^{N}$ for fixed $\{\mathbf{W}_f\}_{f=1}^{F}$, i.e.,

$$\{\mathbf{H}_n^+\}_{n=1}^{N} = \arg\min_{\{\mathbf{H}_n\}_{n=1}^{N}} g(\{\mathbf{W}_f\}_{f=1}^{F}, \{\mathbf{H}_n\}_{n=1}^{N})$$

- Change roles, i.e.,

$$\{\mathbf{W}_f^+\}_{f=1}^{F} = \arg\min_{\{\mathbf{W}_f\}_{f=1}^{F}} g(\{\mathbf{W}_f\}_{f=1}^{F}, \{\mathbf{H}_n^+\}_{n=1}^{N})$$

- Repeat until a stopping criterion is achieved;

- Several other algorithms had been independently developed by Vandaele et al. (2018), Basu et al. (2016), Glasser et al. (2019) and Stark (2016) based on this alternating approach.

Decrease Objective Separately w.r.t. each $\mathbf{H}_n$

- Focus on the first problem:

$$\{\mathbf{H}_n^+\}_{n=1}^N = \arg\min_{\{\mathbf{H}_n\}_{n=1}^N} g(\{\mathbf{W}_f\}_{f=1}^F, \{\mathbf{H}_n\}_{n=1}^N)$$

Decrease Objective Separately w.r.t. each $\mathbf{H}_n$

- Focus on the first problem:

$$\{\mathbf{H}_n^+\}_{n=1}^N = \arg\min_{\{\mathbf{H}_n\}_{n=1}^N} g(\{\mathbf{W}_f\}_{f=1}^F, \{\mathbf{H}_n\}_{n=1}^N)$$

- Objective function can be written as a sum of N terms

$$g(\{\mathbf{W}_f\}_{f=1}^F, \{\mathbf{H}_n\}_{n=1}^N) = \sum_{n=1}^N g_n(\{\mathbf{W}_f\}_{f=1}^F, \mathbf{H}_n)$$

where

$$g_n(\{\mathbf{W}_f\}_{f=1}^F, \mathbf{H}_n) = \frac{1}{2}\sum_{f=1}^F (v_{fn} - \mathrm{Tr}(\mathbf{W}_f \mathbf{H}_n))^2 = \frac{1}{2}\left\|\mathbf{v}_n - \mathcal{W}(\mathbf{H}_n)\right\|^2$$

and $\mathcal{W}(\mathbf{H}) = [\langle \mathbf{W}_1, \mathbf{H}\rangle, \ldots, \langle \mathbf{W}_F, \mathbf{H}\rangle]^\top \in \mathbb{R}^F$ for any matrix $\mathbf{H}$.

Link with Affine Rank Minimization and Phase Retrieval

(Lahat, Lang, Tan, and Févotte, 2021; Lahat and Févotte, 2021)

- For a specific $n = 1, \ldots, N$, estimating $\mathbf{H}_n$ is tantamount to

$$\min_{\mathbf{H}_n} \quad \|\mathbf{v}_n - \mathcal{W}(\mathbf{H}_n)\|^2 \quad \text{subject to} \quad \mathbf{H}_n \in \mathbb{S}_+^K$$

  and possibly a rank constraint on $\mathbf{H}_n$;

Link with Affine Rank Minimization and Phase Retrieval

(Lahat, Lang, Tan, and Févotte, 2021; Lahat and Févotte, 2021)

- For a specific $n = 1, \ldots, N$, estimating $\mathbf{H}_n$ is tantamount to

$$\min_{\mathbf{H}_n} \ \|\mathbf{v}_n - \mathcal{W}(\mathbf{H}_n)\|^2 \quad \text{subject to} \quad \mathbf{H}_n \in \mathbb{S}_+^K$$

  and possibly a rank constraint on $\mathbf{H}_n$;

- Because in PSDMF, one typically imposes low rank constraints on $\mathbf{W}_f$ and $\mathbf{H}_n$ too ("inner ranks" are small);

Link with Affine Rank Minimization and Phase Retrieval

(Lahat, Lang, Tan, and Févotte, 2021; Lahat and Févotte, 2021)

- For a specific $n = 1, \dots, N$, estimating $\mathbf{H}_n$ is tantamount to

$$\min_{\mathbf{H}_n} \; \|\mathbf{v}_n - \mathcal{W}(\mathbf{H}_n)\|^2 \quad \text{subject to} \quad \mathbf{H}_n \in \mathbb{S}_+^K$$

  and possibly a rank constraint on $\mathbf{H}_n$;

- Because in PSDMF, one typically imposes low rank constraints on $\mathbf{W}_f$ and $\mathbf{H}_n$ too ("inner ranks" are small);

- This optimization is known as affine rank minimization (Recht et al., 2010; Jain et al., 2010);

Link with Affine Rank Minimization and Phase Retrieval

(Lahat, Lang, Tan, and Févotte, 2021; Lahat and Févotte, 2021)

- For a specific $n = 1, \ldots, N$, estimating $\mathbf{H}_n$ is tantamount to

$$\min_{\mathbf{H}_n} \quad \|\mathbf{v}_n - \mathcal{W}(\mathbf{H}_n)\|^2 \quad \text{subject to} \quad \mathbf{H}_n \in \mathbb{S}_+^K$$

  and possibly a rank constraint on $\mathbf{H}_n$;

- Because in PSDMF, one typically imposes low rank constraints on $\mathbf{W}_f$ and $\mathbf{H}_n$ too ("inner ranks" are small);

- This optimization is known as affine rank minimization (Recht et al., 2010; Jain et al., 2010);

- If $\text{rank}(\mathbf{W}_f) = 1$ for all $f = 1, \ldots, F$, and $\text{rank}(\mathbf{H}_n) = 1$, this is known as phase retrieval (Candès et al., 2015).

Affine Rank Minimization and Hard Thresholding

(Lahat, Lang, Tan, and Févotte, 2021)

- Many algorithms for affine rank minimization and phase retrieval;

Affine Rank Minimization and Hard Thresholding

(Lahat, Lang, Tan, and Févotte, 2021)

- Many algorithms for affine rank minimization and phase retrieval;
- Use them for PSDMF to solve the subproblem to update $\{\mathbf{H}_n\}_{n=1}^{N}$ given $\mathcal{W}$.

Affine Rank Minimization and Hard Thresholding

(Lahat, Lang, Tan, and Févotte, 2021)

- Many algorithms for affine rank minimization and phase retrieval;
- Use them for PSDMF to solve the subproblem to update $\{\mathbf{H}_n\}_{n=1}^N$ given $\mathcal{W}$.
- Projection of a matrix onto the set of $K \times K$ PSD matrices of rank $\leq R$ is denoted as $\mathrm{H}_{\mathbb{S}_+^K,R}(\cdot)$, also known as hard thresholding;

Affine Rank Minimization and Hard Thresholding

(Lahat, Lang, Tan, and Févotte, 2021)

- Many algorithms for affine rank minimization and phase retrieval;
- Use them for PSDMF to solve the subproblem to update $\{\mathbf{H}_n\}_{n=1}^{N}$ given $\mathcal{W}$.
- Projection of a matrix onto the set of $K \times K$ PSD matrices of rank $\leq R$ is denoted as $\mathrm{H}_{\mathbb{S}_+^K, R}(\cdot)$, also known as hard thresholding;
- Can be computed as (Jain et al., 2010; Tu et al., 2016)

$$\mathrm{H}_{\mathbb{S}_+^K, R}(\mathbf{H}) = \mathbf{U}\mathbf{\Lambda}_R\mathbf{U}^\top$$

  where

  - $\mathbf{\Lambda}_R \in \mathbb{R}^{R \times R}$ is a diagonal nonnegative matrix with the R largest nonnegative eigenvalues of $\mathbf{H}$ on its main diagonal;
  - the columns of $\mathbf{U} \in \mathbb{R}^{K \times R}$ are the eigenvectors of $\mathbf{W}$ associated with these R largest nonnegative eigenvalues

Affine Rank Minimization and Hard Thresholding

(Lahat, Lang, Tan, and Févotte, 2021)

- Many algorithms for affine rank minimization and phase retrieval;
- Use them for PSDMF to solve the subproblem to update $\{\mathbf{H}_n\}_{n=1}^N$ given $\mathcal{W}$.
- Projection of a matrix onto the set of $K \times K$ PSD matrices of rank $\leq R$ is denoted as $\mathrm{H}_{\mathbb{S}_+^K,R}(\cdot)$, also known as hard thresholding;
- Can be computed as (Jain et al., 2010; Tu et al., 2016)

$$\mathrm{H}_{\mathbb{S}_+^K,R}(\mathbf{H}) = \mathbf{U}\mathbf{\Lambda}_R\mathbf{U}^\top$$

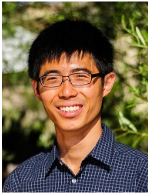
  where
  - $\mathbf{\Lambda}_R \in \mathbb{R}^{R\times R}$ is a diagonal nonnegative matrix with the R largest nonnegative eigenvalues of $\mathbf{H}$ on its main diagonal;
  - the columns of $\mathbf{U} \in \mathbb{R}^{K\times R}$ are the eigenvectors of $\mathbf{W}$ associated with these R largest nonnegative eigenvalues
- Can also use singular value projection; see Lahat et al. (2021) for details.

Majorization-Minimization Algorithm for PSDMF
(Soh and Varvitsiotis, 2021)

- While links to phase retrieval and affine rank minimization are nice, theoretical guarantees (e.g., convergence guarantees) are lacking. ☹

Majorization-Minimization Algorithm for PSDMF
(Soh and Varvitsiotis, 2021)

- While links to phase retrieval and affine rank minimization are nice, theoretical guarantees (e.g., convergence guarantees) are lacking. ☹
- Would be good to develop a multiplicative update-type algorithm based on majorization-minimization (MM). ☺



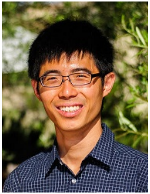
Y. S. Soh (NUS Math)

A. Vavitsiotis (SUTD)

Majorization-Minimization Algorithm for PSDMF

(Soh and Varvitsiotis, 2021)

- While links to phase retrieval and affine rank minimization are nice, theoretical guarantees (e.g., convergence guarantees) are lacking. ☹
- Would be good to develop a multiplicative update-type algorithm based on majorization-minimization (MM). ☺



Y. S. Soh (NUS Math)

A. Vavitsiotis (SUTD)

“Slides” below borrowed from Y. S. Soh with permission and with thanks.

From NMF to PSDMF

Lee and Seung (1999) update rule writes

$$\mathbf{h} \longleftarrow \mathbf{h} \cdot \frac{\mathbf{W}^\top \mathbf{v}}{\mathbf{W}^\top \mathbf{W} \mathbf{v}}.$$

From NMF to PSDMF

Lee and Seung (1999) update rule writes

$$\mathbf{h} \longleftarrow \mathbf{h} \cdot \frac{\mathbf{W}^\top \mathbf{v}}{\mathbf{W}^\top \mathbf{W} \mathbf{v}}.$$

Embed **h** as a diagonal matrix.

$$\begin{pmatrix} \ddots & & \\ & h_k & \\ & & \ddots \end{pmatrix} \longleftarrow \begin{pmatrix} \ddots & & \\ & \frac{(\mathbf{W}^\top \mathbf{v})_k}{(\mathbf{W}^\top \mathbf{W}\mathbf{h})_k} & \\ & & \ddots \end{pmatrix} \begin{pmatrix} \ddots & & \\ & h_k & \\ & & \ddots \end{pmatrix}$$

nng vectors $\cong$ diagonal PSD matrices

From NMF to PSDMF

Re-arrange

$$\begin{pmatrix} \ddots & & \\ & h_k & \\ & & \ddots \end{pmatrix} \longleftarrow \begin{pmatrix} \ddots & & \\ & \frac{(\mathbf{W}^\top \mathbf{v})_k}{(\mathbf{W}^\top \mathbf{W}\mathbf{h})_k} & \\ & & \ddots \end{pmatrix} \begin{pmatrix} \ddots & & \\ & h_k & \\ & & \ddots \end{pmatrix}$$

$$= \begin{pmatrix} \ddots & & \\ & \frac{h_k}{(\mathbf{W}^\top \mathbf{W}\mathbf{h})_k} & \\ & & \ddots \end{pmatrix} \begin{pmatrix} \ddots & & \\ & (\mathbf{W}^\top \mathbf{v})_k & \\ & & \ddots \end{pmatrix}$$

PSD Factorization

$$\underbrace{\mathbf{H}}_{\text{matrix}} \leftarrow \underbrace{T}_{\text{operator}} (\underbrace{\mathcal{W}^\top \mathbf{v}}_{\text{matrix}})$$

From NMF to PSDMF

Re-arrange

$$\begin{pmatrix} \ddots & & \\ & h_k & \\ & & \ddots \end{pmatrix} \longleftarrow \begin{pmatrix} \ddots & & \\ & \frac{(\mathbf{W}^\top \mathbf{v})_k}{(\mathbf{W}^\top \mathbf{W}\mathbf{h})_k} & \\ & & \ddots \end{pmatrix} \begin{pmatrix} \ddots & & \\ & h_k & \\ & & \ddots \end{pmatrix}$$

$$= \begin{pmatrix} \ddots & & \\ & \frac{h_k}{(\mathbf{W}^\top \mathbf{W}\mathbf{h})_k} & \\ & & \ddots \end{pmatrix} \begin{pmatrix} \ddots & & \\ & (\mathbf{W}^\top \mathbf{v})_k & \\ & & \ddots \end{pmatrix}$$

PSD Factorization

$$\underbrace{\mathbf{H}}_{\text{matrix}} \leftarrow \underbrace{T}_{\text{operator}} (\underbrace{\mathcal{W}^\top \mathbf{v}}_{\text{matrix}})$$

Find: operator T that is (i) simple, (ii) preserves PSD-ness, (iii) generalizes averaging of $\mathbf{H}$ and $([\mathcal{W}^T\mathcal{W}](\mathbf{H}))^{-1}$.

From NMF to PSDMF

- The analogue of diagonal scaling is conjugation

$$\mathbf{W} \longleftarrow \mathbf{M}(\mathcal{W}^\top \mathbf{v})\mathbf{M}$$

where

$$\mathbf{M} = \text{Geometric mean}\big(\mathbf{H}, ([\mathcal{W}^T\mathcal{W}](\mathbf{H}))^{-1}\big)$$

From NMF to PSDMF

- The analogue of diagonal scaling is conjugation

$$\mathbf{W} \longleftarrow \mathbf{M}(\mathcal{W}^\top \mathbf{v})\mathbf{M}$$

where

$$\mathbf{M} = \text{Geometric mean}\big(\mathbf{H}, ([\mathcal{W}^T\mathcal{W}](\mathbf{H}))^{-1}\big)$$

- Preserves PSD-ness and is simple.

Geometric Mean of Two Positive Definite (PD) Matrices

- **Definition**: The matrix geometric mean of two PD matrices $\mathbf{C}$ and $\mathbf{D}$ is

$$\mathbf{C}\#\mathbf{D} := \mathbf{C}^{1/2}(\mathbf{C}^{-1/2}\mathbf{D}\mathbf{C}^{-1/2})^{1/2}\mathbf{C}^{1/2}$$

  Generalizes the geometric mean $\sqrt{cd}$ of two positive numbers c and d.

Geometric Mean of Two Positive Definite (PD) Matrices

- **Definition**: The matrix geometric mean of two PD matrices $\mathbf{C}$ and $\mathbf{D}$ is

$$\mathbf{C}\#\mathbf{D} := \mathbf{C}^{1/2}(\mathbf{C}^{-1/2}\mathbf{D}\mathbf{C}^{-1/2})^{1/2}\mathbf{C}^{1/2}$$

  Generalizes the geometric mean $\sqrt{cd}$ of two positive numbers c and d.

- **Equivalent Definition**: Unique PD solution $\mathbf{X}^*$ to the Riccati equation

$$\mathbf{X}\mathbf{C}^{-1}\mathbf{X} = \mathbf{D}.$$

Geometric Mean of Two Positive Definite (PD) Matrices

- **Definition**: The matrix geometric mean of two PD matrices $\mathbf{C}$ and $\mathbf{D}$ is

$$\mathbf{C}\#\mathbf{D} := \mathbf{C}^{1/2}(\mathbf{C}^{-1/2}\mathbf{D}\mathbf{C}^{-1/2})^{1/2}\mathbf{C}^{1/2}$$

  Generalizes the geometric mean $\sqrt{cd}$ of two positive numbers c and d.

- **Equivalent Definition**: Unique PD solution $\mathbf{X}^*$ to the Riccati equation

$$\mathbf{X}\mathbf{C}^{-1}\mathbf{X} = \mathbf{D}.$$

- $\mathbf{C}\#\mathbf{D}$ is the midpoint of the geodesic joining $\mathbf{C}$ and $\mathbf{D}$ on the manifold of PD matrices.

Geometric Mean of Two Positive Definite (PD) Matrices

- **Definition**: The matrix geometric mean of two PD matrices $\mathbf{C}$ and $\mathbf{D}$ is

$$\mathbf{C}\#\mathbf{D} := \mathbf{C}^{1/2}(\mathbf{C}^{-1/2}\mathbf{D}\mathbf{C}^{-1/2})^{1/2}\mathbf{C}^{1/2}$$

  Generalizes the geometric mean $\sqrt{cd}$ of two positive numbers c and d.

- **Equivalent Definition**: Unique PD solution $\mathbf{X}^*$ to the Riccati equation

$$\mathbf{X}\mathbf{C}^{-1}\mathbf{X} = \mathbf{D}.$$

- $\mathbf{C}\#\mathbf{D}$ is the midpoint of the geodesic joining $\mathbf{C}$ and $\mathbf{D}$ on the manifold of PD matrices.

- Fun facts:

$$\mathbf{C}\#\mathbf{D} = \mathbf{D}\#\mathbf{C} \quad \text{and} \quad (\mathbf{C}\#\mathbf{D})^{-1} = \mathbf{C}^{-1}\#\mathbf{D}^{-1}.$$

Multiplicative-Type Algorithm for PSDMF

Recall for fixed $\mathbf{W}_f \in \mathbb{S}_+^K, f = 1, \ldots, F$, we aim to solve

$$\min_{\mathbf{H}} \quad \left\| \mathbf{v} - \mathcal{W}(\mathbf{H}) \right\|^2 \quad \text{subject to} \quad \mathbf{H} \in \mathbb{S}_+^K$$

where $\mathcal{W}(\mathbf{H}) = \left[\langle \mathbf{W}_1, \mathbf{H} \rangle, \ldots, \langle \mathbf{W}_F, \mathbf{H} \rangle\right]^\top \in \mathbb{R}^F$.

Multiplicative-Type Algorithm for PSDMF

Recall for fixed $\mathbf{W}_f \in \mathbb{S}_+^K, f = 1, \ldots, F$, we aim to solve

$$\min_{\mathbf{H}} \quad \|\mathbf{v} - \mathcal{W}(\mathbf{H})\|^2 \quad \text{subject to} \quad \mathbf{H} \in \mathbb{S}_+^K$$

where $\mathcal{W}(\mathbf{H}) = [\langle \mathbf{W}_1, \mathbf{H} \rangle, \ldots, \langle \mathbf{W}_F, \mathbf{H} \rangle]^\top \in \mathbb{R}^F$.

Theorem (Soh and Varvitsiotis (2021))

The objective function $\|\mathbf{v} - \mathcal{W}(\mathbf{H})\|$ is non-increasing under the update rule

$$\mathbf{H}^+ = \mathbf{M}(\mathcal{W}^\top \mathbf{v})\mathbf{M}, \quad \textit{where} \quad \mathbf{M} = \left([\mathcal{W}^\top \mathcal{W}](\mathbf{H})\right)^{-1} \# (\mathbf{H})$$

Furthermore, if initialized with a PD matrix, the subsequent iterates remain PD.

Multiplicative-Type Algorithm for PSDMF

Recall for fixed $\mathbf{W}_f \in \mathbb{S}_+^K, f = 1, \ldots, F$, we aim to solve

$$\min_{\mathbf{H}} \quad \|\mathbf{v} - \mathcal{W}(\mathbf{H})\|^2 \quad \text{subject to} \quad \mathbf{H} \in \mathbb{S}_+^K$$

where $\mathcal{W}(\mathbf{H}) = [\langle \mathbf{W}_1, \mathbf{H} \rangle, \ldots, \langle \mathbf{W}_F, \mathbf{H} \rangle]^\top \in \mathbb{R}^F$.

Theorem (Soh and Varvitsiotis (2021))

The objective function $\|\mathbf{v} - \mathcal{W}(\mathbf{H})\|$ *is non-increasing under the update rule*

$$\mathbf{H}^+ = \mathbf{M}(\mathcal{W}^\top \mathbf{v})\mathbf{M}, \quad \textit{where} \quad \mathbf{M} = \left([\mathcal{W}^\top \mathcal{W}](\mathbf{H})\right)^{-1} \#(\mathbf{H})$$

Furthermore, if initialized with a PD matrix, the subsequent iterates remain PD.

Reduces to Lee and Seung (1999) update in the diagonal case, i.e.,

$$\mathbf{h}^+ = \mathbf{h} \cdot \frac{\mathbf{W}^\top \mathbf{v}}{\mathbf{W}^\top \mathbf{W} \mathbf{v}}.$$

Multiplicative-Type Algorithm for PSDMF

Matrix Multiplicative Update (MMU) Algorithm (Soh and Varvitsiotis, 2021):

- Input: A matrix $\mathbf{V} \in \mathbb{R}_+^{F \times N}$ and parameter $K \in \mathbb{N}$;

Multiplicative-Type Algorithm for PSDMF

Matrix Multiplicative Update (MMU) Algorithm (Soh and Varvitsiotis, 2021):

- Input: A matrix $\mathbf{V} \in \mathbb{R}_+^{F \times N}$ and parameter $K \in \mathbb{N}$;
- Output $\{\mathbf{W}_f\}_{f=1}^F, \{\mathbf{H}_n\}_{n=1}^N \subset \mathbb{S}_+^K$ such that $v_{fn} \approx \langle \mathbf{W}_f, \mathbf{H}_n \rangle$ for all f, n;

Multiplicative-Type Algorithm for PSDMF

Matrix Multiplicative Update (MMU) Algorithm (Soh and Varvitsiotis, 2021):

- Input: A matrix $\mathbf{V} \in \mathbb{R}_+^{F\times N}$ and parameter $K \in \mathbb{N}$;
- Output $\{\mathbf{W}_f\}_{f=1}^F, \{\mathbf{H}_n\}_{n=1}^N \subset \mathbb{S}_+^K$ such that $v_{fn} \approx \langle \mathbf{W}_f, \mathbf{H}_n \rangle$ for all f, n;
- While stopping criterion not satisfied, do

$$\mathbf{W}_f \longleftarrow \mathbf{N}_f(\mathcal{H}^\top \mathbf{v}_{f,:})\mathbf{N}_f \quad \text{where} \quad \mathbf{N}_f = \big([\mathcal{H}^\top\mathcal{H}](\mathbf{W}_f)\big)^{-1} \# (\mathbf{W}_f)$$

and

$$\mathbf{H}_n \longleftarrow \mathbf{M}_n(\mathcal{W}^\top \mathbf{v}_{:,n})\mathbf{M}_n \quad \text{where} \quad \mathbf{M}_n = \big([\mathcal{W}^\top\mathcal{W}](\mathbf{H}_n)\big)^{-1} \# (\mathbf{H}_n).$$

Multiplicative-Type Algorithm for PSDMF

Matrix Multiplicative Update (MMU) Algorithm (Soh and Varvitsiotis, 2021):

- Input: A matrix $\mathbf{V} \in \mathbb{R}_+^{F \times N}$ and parameter $K \in \mathbb{N}$;
- Output $\{\mathbf{W}_f\}_{f=1}^F, \{\mathbf{H}_n\}_{n=1}^N \subset \mathbb{S}_+^K$ such that $v_{fn} \approx \langle \mathbf{W}_f, \mathbf{H}_n \rangle$ for all f, n;
- While stopping criterion not satisfied, do

$$\mathbf{W}_f \longleftarrow \mathbf{N}_f(\mathcal{H}^\top \mathbf{v}_{f,:})\mathbf{N}_f \quad \text{where} \quad \mathbf{N}_f = \left([\mathcal{H}^\top \mathcal{H}](\mathbf{W}_f)\right)^{-1} \#(\mathbf{W}_f)$$

and

$$\mathbf{H}_n \longleftarrow \mathbf{M}_n(\mathcal{W}^\top \mathbf{v}_{:,n})\mathbf{M}_n \quad \text{where} \quad \mathbf{M}_n = \left([\mathcal{W}^\top \mathcal{W}](\mathbf{H}_n)\right)^{-1} \#(\mathbf{H}_n).$$

Properties of MMU:

- Always operates in interior of PSD cone (no projection needed);
- Geometric interpretation of trajectory;
- Recovers classical MU (Lee and Seung, 1999) if matrices are diagonal.

Proof Idea

- Write down candidate auxiliary function inspired by Taylor's theorem;

Proof Idea

- Write down candidate auxiliary function inspired by Taylor's theorem;
- Show it dominates square loss, this reduces to

$$\mathbf{M} \otimes \mathbf{M} - \mathcal{W}^\top \mathcal{W} \succcurlyeq 0,$$

where $\mathbf{M} = \left([\mathcal{W}^\top \mathcal{W}](\mathbf{H})\right)^{-1} \#(\mathbf{H})$ is the matrix geometric mean;

Proof Idea

- Write down candidate auxiliary function inspired by Taylor's theorem;
- Show it dominates square loss, this reduces to

$$\mathbf{M} \otimes \mathbf{M} - \mathcal{W}^\top \mathcal{W} \succcurlyeq 0,$$

  where $\mathbf{M} = \left([\mathcal{W}^\top \mathcal{W}](\mathbf{H})\right)^{-1} \#(\mathbf{H})$ is the matrix geometric mean;
- Pre-multiply with $\mathbf{H}^{-1/2}$, reduces to $\mathbf{H} = \mathbf{I}$;

Proof Idea

- Write down candidate auxiliary function inspired by Taylor's theorem;
- Show it dominates square loss, this reduces to

$$\mathbf{M} \otimes \mathbf{M} - \mathcal{W}^\top \mathcal{W} \succcurlyeq 0,$$

  where $\mathbf{M} = \left([\mathcal{W}^\top \mathcal{W}](\mathbf{H})\right)^{-1} \#(\mathbf{H})$ is the matrix geometric mean;
- Pre-multiply with $\mathbf{H}^{-1/2}$, reduces to $\mathbf{H} = \mathbf{I}$;
- Apply Cauchy–Schwarz inequality

$$\mathrm{Tr}(\mathbf{X}^2)\mathrm{Tr}(\mathbf{Y}^2) \geq \mathrm{Tr}(\mathbf{X}\mathbf{Y})^2$$

  and a consequence of Lieb's concavity theorem (Lieb, 1973)

$$\left(\sum_i \mathbf{X}_i^{1/2}\right) \otimes \left(\sum_i \mathbf{X}_i^{1/2}\right) \preccurlyeq \left(\sum_i \mathbf{X}_i\right)^{1/2} \otimes \left(\sum_i \mathbf{X}_i\right)^{1/2}.$$

Concluding Remarks for PSDMF

- PSDMF (Gouveia et al., 2013; Fiorini et al., 2012; Vandaele et al., 2018) is a generalization of NMF

$$
\begin{aligned}
\text{(PSDMF)} \qquad v_{fn} &= \langle \mathbf{W}_f, \mathbf{H}_n \rangle, \qquad & \mathbf{W}_f, \mathbf{H}_n \succcurlyeq 0, \\
\text{(NMF)} \qquad v_{fn} &= \langle \mathbf{w}_f, \mathbf{h}_n \rangle, \qquad & \mathbf{w}_f, \mathbf{h}_n \geq \mathbf{0}.
\end{aligned}
$$

Concluding Remarks for PSDMF

- PSDMF (Gouveia et al., 2013; Fiorini et al., 2012; Vandaele et al., 2018) is a generalization of NMF

$$
\begin{aligned}
&(\text{PSDMF}) \qquad v_{fn} = \langle \mathbf{W}_f, \mathbf{H}_n \rangle, \qquad \mathbf{W}_f, \mathbf{H}_n \succcurlyeq 0, \\
&(\text{NMF}) \qquad v_{fn} = \langle \mathbf{w}_f, \mathbf{h}_n \rangle, \qquad \mathbf{w}_f, \mathbf{h}_n \geq \mathbf{0}.
\end{aligned}
$$

- Can use signal processing primitives such as phase retrieval and affine rank minimization within an alternating minimzation framework to find $\{\mathbf{W}_f\}$ and $\{\mathbf{H}_n\}$ (Lahat et al., 2021);

Concluding Remarks for PSDMF

- PSDMF (Gouveia et al., 2013; Fiorini et al., 2012; Vandaele et al., 2018) is a generalization of NMF

$$\begin{aligned} &(\text{PSDMF}) \qquad v_{fn} = \langle \mathbf{W}_f, \mathbf{H}_n \rangle, \qquad \mathbf{W}_f, \mathbf{H}_n \succcurlyeq 0, \\ &(\text{NMF}) \qquad v_{fn} = \langle \mathbf{w}_f, \mathbf{h}_n \rangle, \qquad \mathbf{w}_f, \mathbf{h}_n \geq \mathbf{0}. \end{aligned}$$

- Can use signal processing primitives such as phase retrieval and affine rank minimization within an alternating minimzation framework to find $\{\mathbf{W}_f\}$ and $\{\mathbf{H}_n\}$ (Lahat et al., 2021);
- Even better, use majorization-minimization (MM) in the space of PD matrices (Soh and Varvitsiotis, 2021);

Concluding Remarks for PSDMF

- PSDMF (Gouveia et al., 2013; Fiorini et al., 2012; Vandaele et al., 2018) is a generalization of NMF

$$
\begin{aligned}
\text{(PSDMF)} \qquad v_{fn} &= \langle \mathbf{W}_f, \mathbf{H}_n \rangle, \qquad & \mathbf{W}_f, \mathbf{H}_n \succcurlyeq 0, \\
\text{(NMF)} \qquad v_{fn} &= \langle \mathbf{w}_f, \mathbf{h}_n \rangle, \qquad & \mathbf{w}_f, \mathbf{h}_n \geq \mathbf{0}.
\end{aligned}
$$

- Can use signal processing primitives such as phase retrieval and affine rank minimization within an alternating minimzation framework to find $\{\mathbf{W}_f\}$ and $\{\mathbf{H}_n\}$ (Lahat et al., 2021);
- Even better, use majorization-minimization (MM) in the space of PD matrices (Soh and Varvitsiotis, 2021);
- Other extensions to symmetric cones, including SOCPs.

References I

G. Averkov, V. Kaibel, and S. Weltge. Maximum semidefinite and linear extension complexity of families of polytopes. Math. Program., 167(2):381–394, 2018.

A. Basu, M. Dinitz, and X. Li. Computing approximate PSD factorizations. In Proc. APPROX/RANDOM, volume 60, pages 2:1–2:12, 2016.

C. M. Bishop. Bayesian PCA. In Advances of Neural Information Processing Systems (NIPS), 1999.

R. Bradley and M. Terry. Rank analysis of incomplete block designs I: The method of paired comparisons. Biometrika, 35:324–345, 1952.

E. Candès, Y. C. Eldar, T. Strohmer, and V. Voroninski. Phase retrieval via matrix completion. SIAM Review, 57(2):225–251, 2015.

E. J. Candès, M. B. Wakin, and S. P. Boyd. Enhancing sparsity by reweighted ℓ_1 minimization. J. Fourier Analysis and Applications, 14:877–905, 2008.

E. C. Chi and T. G. Kolda. On tensors, sparsity, and nonnegative factorizations. SIAM Journal on Matrix Analysis and Applications, 33(4):1272–1299, 2012.

H. Fawzi, J. Gouveia, P. A. Parrilo, R. Z. Robinson, and R. R. Thomas. Positive semidefinite rank. Math. Program., 153(1):133–177, 2015.

C. Févotte, N. Bertin, and J.-L. Durrieu. Nonnegative matrix factorization with the Itakura-Saito divergence. With application to music analysis. Neural Computation, 21(3):793–830, Mar. 2009. doi: 10.1162/neco.2008.04-08-771. URL `https://www.irit.fr/~Cedric.Fevotte/publications/journals/neco09_is-nmf.pdf`.

S. Fiorini, S. Massar, S. Pokutta, H. R. Tiwary, and R. D. Wolf. Linear vs. semidefinite extended formulations: exponential separation and strong lower bounds. In Proc. STOC, pages 95–106, 2012.

References II

N. Gillis, L. T. K. Hien, V. Leplat, and V. Y. F. Tan. Distributionally Robust and Multi-Objective Nonnegative Matrix Factorization . IEEE Trans. on Pattern Analysis and Machine Intelligence, 2022.

I. Glasser, R. Sweke, N. Pancotti, J. Eisert, and I. Cirac. Expressive power of tensor-network factorizations for probabilistic modeling. In Proc. Adv. Neural Inf. Process. Syst., pages 1496–1508, 2019.

J. Gouveia, P. A. Parrilo, and R. R. Thomas. Lifts of convex sets and cone factorizations. Mathematics of Operations Research, 38(2):248–264, 2013.

P. Jain, R. Meka, and I. S. Dhillon. Guaranteed rank minimization via singular value projection. In Advances of Neural Information Processing Systems (NIPS), pages 935–945, 2010.

R. Jain, Y. Shi, Z. Wei, and S. Zhang. Efficient protocols for generating bipartite classical distributions and quantum states. IEEE Transactions on Information Theory, 59(8):5171–5178, 2013.

D. Lahat and C. Févotte. Positive semidefinite matrix factorization based on truncated Wirtinger flow. In Proc. Eusipco, 2021.

D. Lahat, Y. Lang, V. Y. F. Tan, and C. Févotte. Positive semidefinite matrix factorization: A connection with phase retrieval and affine rank minimization. IEEE Transactions on Signal Processing, 69:3059–3074, 2021. doi: 10.1109/TSP.2021.3071293. URL https://arxiv.org/pdf/2007.12364.

D. D. Lee and H. S. Seung. Learning the parts of objects with nonnegative matrix factorization. Nature, 401: 788–791, 1999.

E. H. Lieb. Convex trace functions and the Wigner–Yanase–Dyson conjecture. Advances in Mathematics, 11 (3):267–288, 1973.

R. Luce. Individual choice behavior: A theoretical analysis. Wiley, 1959.

References III

B. Recht, M. Fazel, and P. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. SIAM Review, 52(3):471–501, 2010.

Y. S. Soh and A. Varvitsiotis. A non-commutative extension of Lee-Seung's algorithm for positive semidefinite factorizations. In Advances in Neural Processing Systems (NeurIPS), 2021.

C. J. Stark. Recommender systems inspired by the structure of quantum theory. arXiv 1691.06035, 2016.

V. Y. F. Tan and C. Févotte. Automatic relevance determination in nonnegative matrix factorization with the β-divergence. IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(7):1592–1605, 2013.

M. E. Tipping. Sparse Bayesian Learning and the Relevance Vector Machine. Journal of Machine Learning Research, 1:211–244, 2001.

S. Tu, R. Boczar, M. Simchowitz, M. Soltanolkotabi, and B. Recht. Low-rank solutions of linear matrix equations via Procrustes flow. In Proc. Intl. Conf. on Machine Learning (ICML), pages 964–073, 2016.

J. van Apeldoorn, A. Gilyén, S. Gribling, and R. deWolf. Quantum SDPSolvers: Better upper and lower bounds. Quantum, 4(230), Feb 2020.

A. Vandaele, F. Glineur, and N. Gillis. Algorithms for positive semidefinite factorization. Computational Optimization and Applications, 71(1):193–219, 2018.

T. Virtanen. Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria. IEEE Transactions on Audio, Speech, and Language Processing, 15(3):1066–1074, 2007.

References IV

R. Xia, V. Y. F. Tan, L. Filstroff, and C. Févotte. A ranking model motivated by nonnegative matrix factorization with applications to tennis tournaments. In Proc. European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECMLPKDD), Sep. 2019. URL https://arxiv.org/abs/1903.06500.

M. Yannakakis. Expressing combinatorial optimization problems by linear programs. J. Comput. Syst. Sci, 43 (3):441–466, 1991.

M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. J. Royal Statistical Soc., 68(1):49–67, 2007.

R. Zhao and V. Y. F. Tan. A unified convergence analysis of the multiplicative update algorithm for regularized nonnegative matrix factorization. IEEE Transactions on Signal Processing, 66(1):129–138, Jan 2018. ISSN 1053-587X. doi: 10.1109/TSP.2017.2757914.

22-27 MAY 2022

ICASSP 2022

SINGAPORE

IEEE Signal Processing Society

IEEE

T16: Recent advances in Nonnegative Matrix Factorization (Part 2)

Q&A

22-27 MAY 2022

ICASSP 2022

SINGAPORE

IEEE Signal Processing Society

IEEE

T16: Recent advances in Nonnegative Matrix Factorization

End