Robust nonnegative matrix factorisation with the $\beta$-divergence and applications in imaging

Cédric Févotte
Institut de Recherche en Informatique de Toulouse

Imaging & Machine Learning
Institut Henri Poincaré
April 2019
Outline

Generalities
  Matrix factorisation models
  Nonnegative matrix factorisation (NMF)

Optimisation for NMF
  Measures of fit
  Majorisation-minimisation

Applications in imaging
  Hyperspectral unmixing in remote sensing
  Factor analysis in dynamic PET
Data often available in matrix form.
Matrix factorisation models

Data often available in matrix form.
Matrix factorisation models

Data often available in matrix form.
Matrix factorisation models

Data often available in matrix form.
Matrix factorisation models

≈ dictionary learning
low-rank approximation
factor analysis
latent semantic analysis

Data $X$ ≈ Dictionary $W$ Activations $H$
Matrix factorisation models

≈ dictionary learning
≈ low-rank approximation
≈ factor analysis
≈ latent semantic analysis

\[
\approx \begin{array}{c}
\text{data } X \\
\text{dictionary } W \\
\text{activations } H
\end{array}
\]
Matrix factorisation models

for dimensionality reduction (coding, low-dimensional embedding)
Matrix factorisation models

for unmixing (source separation, latent topic discovery)
Matrix factorisation models

for interpolation (collaborative filtering, image inpainting)
Nonnegative matrix factorisation

- data \( V \) and factors \( W, H \) have nonnegative entries.
- nonnegativity of \( W \) ensures interpretability of the dictionary, because patterns \( w_k \) and samples \( v_n \) 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)
NMF for latent semantic analysis
(Lee and Seung, 1999; Hofmann, 1999)

Encyclopedia entry: 'Constitution of the United States'

president (148)
congress (124)
power (120)
united (104)
constitution (81)
amendment (71)
government (57)
law (49)

court
government
council
culture
supreme
constitutional
rights
justice

flowers
leaves
plant
perennial
flower
plants
growing
annual

disease
behaviour
glands
contact
symptoms
skin
pain
infection

\[ \mathbf{v}_n \approx \mathbf{W} \mathbf{h}_n \]

reproduced from (Lee and Seung, 1999)
NMF for audio spectral unmixing
(Smaragdis and Brown, 2003)

Non-Negative Matrix Factorization

All factors are positive-valued:

Resulting reconstruction is additive

Input music passage

Component Frequency

Time (sec)

Non-negative reproduced from (Smaragdis, 2013)
NMF for hyperspectral unmixing
(Berry, Browne, Langville, Pauca, and Plemmons, 2007)

Fig. 1. Hyperspectral imaging concept.

Introduction
Hyperspectral cameras [1]–[11] contribute significantly to earth observation and remote sensing [12], [13]. Their potential motivates the development of small, commercial, high spatial and spectral resolution instruments. They have also been used in food safety [14]–[17], pharmaceutical process monitoring and quality control [18]–[22], and biomedical, industrial, and biometric, and forensic applications [23]–[27]. HSCs can be built to function in many regions of the electro-magnetic spectrum. The focus here is on those covering the visible, near-infrared, and shortwave infrared spectral bands (in the range $0.3 \mu m$ to $2.5 \mu m$) [5]). Disregarding atmospheric effects, the signal recorded by an HSC at a pixel is a mixture of light scattered by substances located in the field of view [3]. Fig. 1 illustrates the measured data. They are organized into planes forming a data cube. Each plane corresponds to radiance acquired over a

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

Generalities
- Matrix factorisation models
- Nonnegative matrix factorisation (NMF)

Optimisation for NMF
- Measures of fit
- Majorisation-minimisation

Applications in imaging
- Hyperspectral unmixing in remote sensing
- Factor analysis in dynamic PET
NMF as a constrained minimisation problem

Minimise a measure of fit between $V$ and $WH$, subject to nonnegativity:

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

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)
- $\alpha$-divergence (Cichocki et al., 2008)
- $\beta$-divergence (Cichocki et al., 2006; Févotte and Idier, 2011)
- Bregman divergences (Dhillon and Sra, 2005)
- and more in (Yang and Oja, 2011)

Regularisation terms often added to $D(V|WH)$ for sparsity, smoothness, dynamics, etc.
Nonconvex problem.
Probabilistic models

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

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

<table>
<thead>
<tr>
<th>data support</th>
<th>distribution/noise</th>
<th>divergence</th>
<th>examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>real-valued</td>
<td>additive Gaussian</td>
<td>squared Euclidean</td>
<td>many</td>
</tr>
<tr>
<td>integer</td>
<td>multinomial*</td>
<td>weighted KL</td>
<td>word counts</td>
</tr>
<tr>
<td>integer</td>
<td>Poisson</td>
<td>generalised KL</td>
<td>photon counts</td>
</tr>
<tr>
<td>nonnegative</td>
<td>multiplicative Gamma</td>
<td>Itakura-Saito</td>
<td>spectrogram</td>
</tr>
<tr>
<td>generally</td>
<td>Tweedie</td>
<td>$\beta$-divergence</td>
<td>generalises</td>
</tr>
<tr>
<td>nonnegative</td>
<td></td>
<td></td>
<td>above models</td>
</tr>
</tbody>
</table>

*conditional independence over $f$ does not apply
The $\beta$-divergence

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

\[
d_\beta(x|y) \overset{\text{def}}{=} \begin{cases} 
\frac{1}{\beta(\beta-1)} (x^\beta + (\beta - 1) y^\beta - \beta x y^{\beta-1}) & \beta \in \mathbb{R}\setminus\{0, 1\} \\
x \log \frac{x}{y} + (y - x) & \beta = 1 \\
\frac{x}{y} - \log \frac{x}{y} - 1 & \beta = 0
\end{cases}
\]

Special cases:
- squared Euclidean distance ($\beta = 2$)
- generalised 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 $\beta$-divergence

$$d(x=1|y)$$

$\beta = 2$ (Euc)
The $\beta$-divergence
The $\beta$-divergence
The \( \beta \)-divergence

\[
d(x=1|y) \quad \beta \begin{cases} 2 \text{(Euc)} \\ 1 \text{(KL)} \\ 0 \text{(IS)} \\ -1 \end{cases}
\]
The $\beta$-divergence

\[
d(x=1|y)
\]

- $\beta = 2$ (Euc)
- $\beta = 1$ (KL)
- $\beta = 0$ (IS)
- $\beta = -1$
- $\beta = 3$
Common NMF algorithm design

- 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:
  \[
  \mathbf{V} \approx \mathbf{WH} \iff \mathbf{V}^T \approx \mathbf{H}^T \mathbf{W}^T
  \]
- Objective function separable in the columns of $\mathbf{H}$ or the rows of $\mathbf{W}$:
  \[
  D(\mathbf{V}|\mathbf{WH}) = \sum_n D(\mathbf{v}_n|\mathbf{Wh}_n)
  \]
- Essentially left with nonnegative linear regression:
  \[
  \min_{\mathbf{h} \geq 0} C(\mathbf{h}) \overset{\text{def}}{=} D(\mathbf{v}|\mathbf{Wh})
  \]

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, initialisation is an issue.
Build $G(h|\tilde{h})$ such that $G(h|\tilde{h}) \geq C(h)$ and $G(\tilde{h}|\tilde{h}) = C(\tilde{h})$.

Optimise (iteratively) $G(h|\tilde{h})$ instead of $C(h)$. 

**Objective function $C(h)$**

![Graph of Objective function $C(h)$]

-横轴代表变量范围
-纵轴代表函数值范围
-图中显示了函数的曲线
-图例说明了“Objective function $C(h)$”
Build $G(h|\tilde{h})$ such that $G(h|\tilde{h}) \geq C(h)$ and $G(\tilde{h}|\tilde{h}) = C(\tilde{h})$. Optimise (iteratively) $G(h|\tilde{h})$ instead of $C(h)$. 

![Graph showing objective function $C(h)$ and auxiliary function $G(h|h^{(0)})$.]
Majorisation-minimisation (MM)

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

![Graph](image-url)
Majorisation-minimisation (MM)

Build \( G(h|h) \) such that \( G(h|h) \geq C(h) \) and \( G(h|h) = C(h) \).

Optimise (iteratively) \( G(h|h) \) instead of \( C(h) \).
Majorisation-minimisation (MM)

Build $G(h|\tilde{h})$ such that $G(h|\tilde{h}) \geq C(h)$ and $G(\tilde{h}|\tilde{h}) = C(\tilde{h})$.

Optimise (iteratively) $G(h|\tilde{h})$ instead of $C(h)$.
Majorisation-minimisation (MM)

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

\[
C_{IS}(h) = \left[ \sum_f \frac{v_f}{\sum_k w_{fk} h_k} \right] + \left[ \sum_f \log \left( \sum_k w_{fk} h_k \right) \right] + cst
\]
Majorisation-minimisation (MM)

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

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

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

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

- $\nabla h_k C(h) = \nabla^+_{h_k} C(h) - \nabla^-_{h_k} C(h)$ and the two summands are nonnegative.
- If $\nabla h_k C(\tilde{h}) > 0$, ratio of summands < 1 and $h_k$ goes left.
- $\gamma$ is a divergence-specific scalar exponent.

- Details in (Févotte and Idier, 2011; Yang and Oja, 2011; Zhao and Tan, 2018)
Outline

Generalities
Matrix factorisation models
Nonnegative matrix factorisation (NMF)

Optimisation for NMF
Measures of fit
Majorisation-minimisation

Applications in imaging
Hyperspectral unmixing in remote sensing
Factor analysis in dynamic PET
Hyperspectral model selection 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.
- $\mathbf{W}$ and $\mathbf{H}$ estimated with $K = 3$ ($\sim$ ground truth) and various values of $\beta$.
- missing pixels are reconstructed from $\hat{\mathbf{V}} = \mathbf{WH}$.
- evaluation using the average spectral angle mapper (aSAM):

$$\text{aSAM}(\mathbf{V}) = \frac{1}{N} \sum_{n=1}^{N} \text{acos} \left( \frac{\langle \mathbf{v}_n, \hat{\mathbf{v}}_n \rangle}{\|\mathbf{v}_n\| \|\hat{\mathbf{v}}_n\|} \right)$$
Hyperspectral model selection by matrix completion
(Févotte and Dobigeon, 2015)

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

- Optimised with majorisation-minimisation.
Moffett Field data

reproduced from (Dobigeon, 2007)
Nonlinear hyperspectral unmixing
(Févotte and Dobigeon, 2015)

Unmixing results

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

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

Outlier term captures specific water/soil interactions.
Nonlinear hyperspectral unmixing
(Févotte and Dobigeon, 2015)

Villelongue/Madonna data (forested area)
Unmixing results

spectral endmembers & activation maps

(red: $\beta = 1$, black: $\beta = 2$)

outlier energy $\{\|r_n\|\}_n$

($\beta = 1$)

Outlier term seems to capture patterns due to sensor miscalibration.
Outline

Generalities
  Matrix factorisation models
  Nonnegative matrix factorisation (NMF)

Optimisation for NMF
  Measures of fit
  Majorisation-minimisation

Applications in imaging
  Hyperspectral unmixing in remote sensing
  Factor analysis in dynamic PET
3D functional imaging

Observe the temporal evolution of the brain activity after injecting a radiotracer (biomarker of a specific compound).

\( v_n \) 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)
Mixing model

- the specific-binding TAC signature varies in space:

\[ v_n \approx [w_1 + \delta_n] h_{1n} + \sum_{k=2}^{K} w_k h_{kn} \]

\[ \approx [w_1 + Db_n] h_{1n} + \sum_{k=2}^{K} w_k h_{kn} \]

\[ \approx Wh_n + h_{1n} Db_n \]

- D is fixed and pre-trained using labeled or simulated data.

Estimation

\[ \min_{W,H,B\geq 0} \ D_\beta(V|WH + 1h_1 \odot DB) + \lambda \|B\|_{2,1} \]

Optimised with majorisation-minimisation.
**Factor analysis in dynamical PET**

**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 \( (\|b_n\|_{2,1}) \) in three different planes and for three values of \( \beta \)
Conclusions

- NMF can efficiently unmix composite data in imaging problems.
- Application-specific variants have been proposed.
- The $\beta$-divergence can be adjusted to the statistics of the noise.
- Majorisation-minimisation works well in this setting.

ERC-funded **postdoc positions** in machine learning & signal processing:
- Multimodal data processing for multimedia artistic creation (with Tim van Cruys)
- Learning with low-rank models (with Emmanuel Soubies)
- Bayesian deep learning (with Nicolas Dobigeon)

http://projectfactory.irit.fr/
SPARS 2019
Toulouse, France, July 1-4, 2019

Plenary speakers

Yuejie Chi, CMU
Emilie Chouzenoux, Univ. Paris-Est
Mark Davenport, Georgia Tech
Monika Dörfler, Univ. Vienna

Pier Luigi Dragotti, ICL
Bhaskar Rao, UC San Diego
Simon Thorpe, CNRS
Lenka Zdeborova, CNRS

Special talk by Michael I. Jordan, UC Berkeley

http://spars-workshop.org/


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)