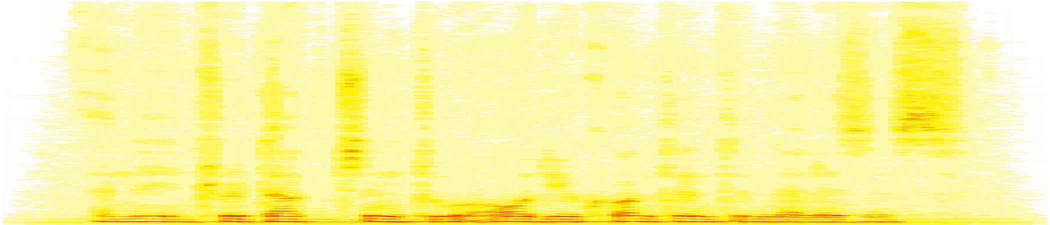
Introduction to Audio Content Analysis

Module 5.5: Non-negative Matrix Factorization for Fundamental Frequency Detection

alexander lerch



Georgia Tech Center for Music Technology
College of Design

introduction
overview

corresponding textbook section

Chapter 5 — Tonal Analysis: pp. 106

- **lecture content**
  - introduction to NMF
  - objective function and update rules
- **learning objectives**
  - describe the process of NMF
  - discuss the pros and cons of using NMF of polyphonic pitch detection
  - apply NMF to a simple audio file and interpret the results

introduction

overview

Georgia Tech Center for Music Technology College of Design

corresponding textbook section

Chapter 5 — Tonal Analysis: pp. 106

- **lecture content**
  - introduction to NMF
  - objective function and update rules
- **learning objectives**
  - describe the process of NMF
  - discuss the pros and cons of using NMF of polyphonic pitch detection
  - apply NMF to a simple audio file and interpret the results

non-negative matrix factorization

introduction

- **Non-negative Matrix Factorization (NMF)**
  Given a $m \times n$ matrix V, find a $m \times r$ matrix W and a $r \times n$ matrix H such that

$$V \approx WH$$

  - all matrices must be non-negative
  - rank r is usually smaller than m and n

- **advantage of non-negativity?**
  - additive model
  - relates to probability distributions
  - efficiency?

non-negative matrix factorization
introduction

- **Non-negative Matrix Factorization (NMF)**
  Given a $m \times n$ matrix V, find a $m \times r$ matrix W and a $r \times n$ matrix H such that

$$V \approx WH$$

  - all matrices must be non-negative
  - rank r is usually smaller than m and n
- **advantage of non-negativity?**
  - additive model
  - relates to probability distributions
  - efficiency?

non-negative matrix factorization

introduction

- **Non-negative Matrix Factorization (NMF)**
  Given a $m \times n$ matrix V, find a $m \times r$ matrix W and a $r \times n$ matrix H such that

$$V \approx WH$$

  - all matrices must be non-negative
  - rank r is usually smaller than m and n

- **advantage of non-negativity?**
  - additive model
  - relates to probability distributions
  - efficiency?

non-negative matrix factorization

introduction

- **Non-negative Matrix Factorization (NMF)**
  Given a $m \times n$ matrix V, find a $m \times r$ matrix W and a $r \times n$ matrix H such that

$$V \approx WH$$

  - all matrices must be non-negative
  - rank r is usually smaller than m and n

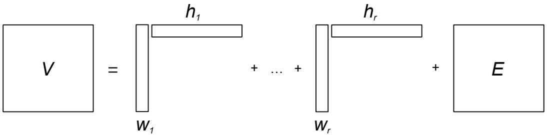
- **advantage of non-negativity?**
  - additive model
  - relates to probability distributions
  - efficiency?

non-negative matrix factorization

alternative formulation[1] to $V \approx WH$

$$V = \sum_{i=1}^{r} w_i \cdot h_i + E$$

- $V \in \mathbb{R}^{m \times n}$
- $W = [w_1, w_2, \ldots, w_r] \in \mathbb{R}^{m \times r}$
- $H = [h_1, h_2, \ldots, h_r]^T \in \mathbb{R}^{r \times n}$
- E is the error matrix



[1] A Cichocki, R Zdunek, A. Phan, *et al.*, *Nonnegative matrix and tensor factorizations: applications to exploratory multi-way data analysis and blind source separation*. John Wiley & Sons, 2009.

objective function

distance and divergence

Georgia Tech Center for Music Technology College of Design

- task: **iteratively minimize objective function** $D(V||WH)$
- typical distance measures ($B = WH$):
  - squared Euclidean distance:

$$D_{\mathrm{EU}}(V \parallel B) = \parallel V - B \parallel^2 = \sum_{ij} (V_{ij} - B_{ij})^2$$

  - generalized K-L divergence:

$$D_{\mathrm{KL}}(V \parallel B) = \sum_{ij} (V_{ij} \log \left(\frac{V_{ij}}{B_{ij}}\right) - V_{ij} + B_{ij})$$

  - others[2]: Bregman Divergence, Alpha-Divergence, Beta-Divergence, ...

[2] A Cichocki, R Zdunek, A. Phan, *et al.*, *Nonnegative matrix and tensor factorizations: applications to exploratory multi-way data analysis and blind source separation*. John Wiley & Sons, 2009.

objective function
distance and divergence

- task: **iteratively minimize objective function** $D(V||WH)$
- typical distance measures ($B = WH$):
  - squared Euclidean distance:

$$D_{\mathrm{EU}}(V \parallel B) = \parallel V - B \parallel^2 = \sum_{ij}(V_{ij} - B_{ij})^2$$

  - generalized K-L divergence:

$$D_{\mathrm{KL}}(V \parallel B) = \sum_{ij}\left(V_{ij} \log\left(\frac{V_{ij}}{B_{ij}}\right) - V_{ij} + B_{ij}\right)$$

  - others[2]: Bregman Divergence, Alpha-Divergence, Beta-Divergence, ...

[2] A Cichocki, R Zdunek, A. Phan, *et al.*, *Nonnegative matrix and tensor factorizations: applications to exploratory multi-way data analysis and blind source separation*. John Wiley & Sons, 2009.

objective function

gradient descent

- minimization of objective function

- **gradient descent**: minimum can be found as zero of derivative
  - 2D example: given a function $f(x_1, x_2)$, find the minimum $x_1 = a$ and $x_2 = b$

    1. initialize $x_i(0)$ with random numbers
    2. update points iteratively:

$$x_i(n+1) = x_i(n) - \alpha \cdot \frac{\partial f}{\partial x_i}, \quad i = [1, 2]$$

$\Rightarrow$ as iteration number n increases, x_1, x_2 will be closer to a, b.

objective function
gradient descent

- minimization of objective function
- **gradient descent**: minimum can be found as zero of derivative
  - 2D example: given a function $f(x_1, x_2)$, find the minimum $x_1 = a$ and $x_2 = b$
    1. initialize $x_i(0)$ with random numbers
    2. update points iteratively:

$$x_i(n+1) = x_i(n) - \alpha \cdot \frac{\partial f}{\partial x_i}, \quad i = [1, 2]$$

$\Rightarrow$ as iteration number n increases, x_1, x_2 will be closer to a, b.

objective function

additive vs. multiplicative update rules

Georgia Tech Center for Music Technology College of Design

optimization of objective function[3] $D_{\mathrm{EU}}(V \parallel WH) = \parallel V - WH \parallel^2$

- **additive** update rules:

$$H \leftarrow H + \alpha \frac{\partial J}{\partial H} = H + \alpha[(W^T V) - (W^T WH)]$$

$$W \leftarrow W + \alpha \frac{\partial J}{\partial W} = W + \alpha[(VH^T) - (WHH^T)]$$

- **multiplicative** update rules:

$$H \leftarrow H \frac{(W^T V)}{(W^T WH)}$$

$$W \leftarrow W \frac{(VH^T)}{(WHH^T)}$$

[3] D Seung and L Lee, "Algorithms for non-negative matrix factorization," in *Advances in neural information processing systems*, 2001, pp. 556–562. [Online]. Available: http://www.public.asu.edu/~jye02/CLASSES/Fall-2007/NOTES/lee01algorithms.pdf.

objective function

additive vs. multiplicative update rules

Georgia Tech Center for Music Technology College of Design

optimization of objective function[3] $D_{\mathrm{EU}}(V \parallel WH) = \parallel V - WH \parallel^2$

- **additive** update rules:

$$H \leftarrow H + \alpha \frac{\partial J}{\partial H} = H + \alpha[(W^T V) - (W^T WH)]$$

$$W \leftarrow W + \alpha \frac{\partial J}{\partial W} = W + \alpha[(VH^T) - (WHH^T)]$$

- **multiplicative** update rules:

$$H \leftarrow H \frac{(W^T V)}{(W^T WH)}$$

$$W \leftarrow W \frac{(VH^T)}{(WHH^T)}$$

[3] D Seung and L Lee, "Algorithms for non-negative matrix factorization," in *Advances in neural information processing systems*, 2001, pp. 556–562. [Online]. Available: http://www.public.asu.edu/~jye02/CLASSES/Fall-2007/NOTES/lee01algorithms.pdf.

objective function

additional cost function constraints

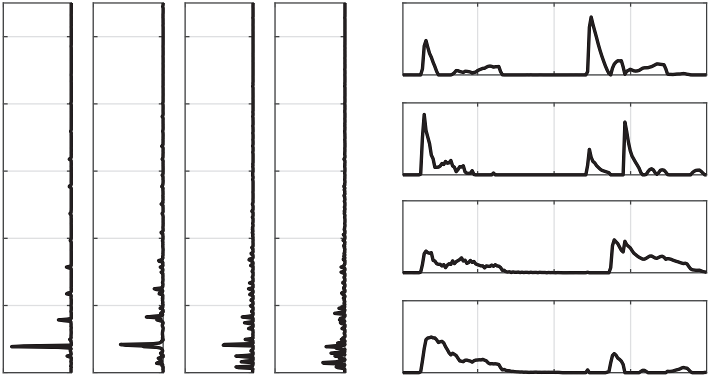
- additional penalty terms (regularization terms) may be added to objective function
- example: sparsity in W or H

$$D = \| V - WH \|^2 + \alpha J_{\mathrm{W}}(W) + \beta J_{\mathrm{H}}(H)$$

  - α, β: coefficients for controlling degree of sparsity
  - J_{W} and J_{H}: typically L_1, L_2 norm

example
template extraction

- unsupervised extraction of templates and activations
- input audio:



matlab source: matlab/displayNmfTemplates.m

summary
lecture content

Georgia Tech Center for Music Technology College of Design

- **non-negative matrix factorization**
  - iterative process minimizing an objective function
  - split a matrix into a template matrix and an activation matrix
- **NMF for pitch tracking**
  - input usually magnitude spectrogram
    - templates: spectra of notes/sounds
    - activation: loudness/trigger of these sound