Towards Tuning-Free Minimum-Volume Nonnegative Matrix Factorization

Duc Toan Nguyen, *Texas Christian University, Fort Worth, Texas*

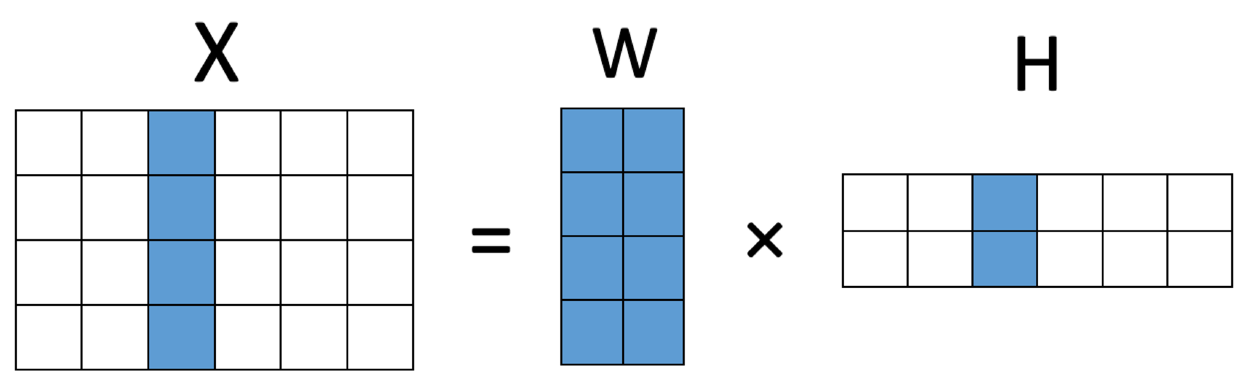
Advisor: Dr. Eric C. Chi, *Department of Statistics, Rice University, Houston, Texas*

Introduction

Nonnegative matrix factorization (NMF): Given a nonnegative matrix $\mathbf{X} \in \mathbb{R}^{m\times n}$ (all entries are nonnegative) and a target rank r that $0 < r \ll m$. We want to find a matrix factorization model

$$\mathbf{X} \approx \mathbf{WH},$$

where $\mathbf{W} \in \mathbb{R}^{m\times r}$ and $\mathbf{H} \in \mathbb{R}^{r\times n}$ are nonnegative.



Min-Vol Rank-Deficient NMF

Minimum-Volume Rank-Deficient NMF: In [3], Leplat et al. proposed the optimization problem for NMF

$$\min_{\mathbf{W},\mathbf{H}} \|\mathbf{X}-\mathbf{WH}\|_F^2 + \lambda \log\det(\mathbf{W}^\top\mathbf{W}+\delta\mathbf{I})$$
$$\text{s.t.} \quad (\mathbf{W},\mathbf{H}) \in \mathcal{S}$$

Notation:

- $\mathcal{S} = \{\boldsymbol{\theta} = (\mathbf{W},\mathbf{H}) | \mathbf{W},\mathbf{H} \geq 0$ and $\mathbf{1}^\top\mathbf{H}(:,j) \leq 1, \forall j\}$ (Constraint set)
- $\mathbf{H}(:,j)$ is the jth column of matrix $\mathbf{H}$

Motivation

In the paper [3], Leplat et al. briefly choose the tuning parameter λ as

$$\lambda = \tilde{\lambda}\frac{\|\mathbf{X}-\mathbf{W}_0\mathbf{H}_0\|_F^2}{\log\det(\mathbf{W}_0^\top\mathbf{W}_0+\delta\mathbf{I})}$$

such that

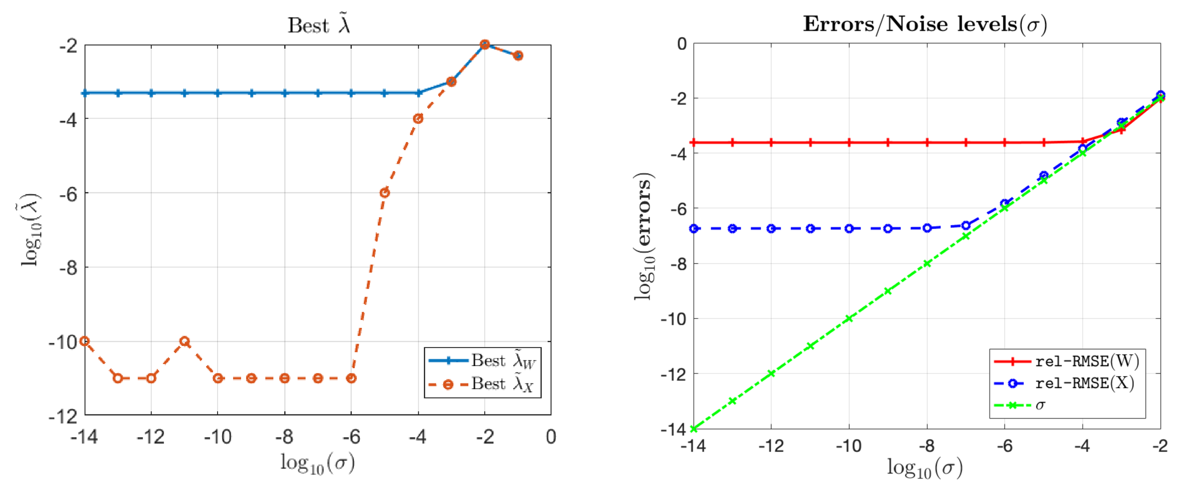
- $(\mathbf{W}_0,\mathbf{H}_0)$ is the initialization for $(\mathbf{W},\mathbf{H})$
- $\tilde{\lambda}$ **is between 1 and 10^{-3}** depends on the noise level

Question: What is the best $\tilde{\lambda}$ for each noise level?

Experiment with λ

- True $\mathbf{W}^\star = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{pmatrix}$
- True $\mathbf{H}^\star$: A stochastic 4×500 matrix generated by *Dirichlet distribution*
- True $\mathbf{X}^\star = \mathbf{W}^\star\mathbf{H}^\star$
- Noise level $\sigma = 10^{-i}$ for $i \in \{1,\cdots,14\}$, Simulated matrix $\mathbf{X} = \mathbf{X}^\star + \sigma\mathbf{N}$ ($\mathbf{N}$ is random, $n_{ij} \in [0,1]$)
- $\texttt{rel-RMSE}(\mathbf{X}) = \|\mathbf{X}^\star - \hat{\mathbf{X}}\|_F / \|\mathbf{X}^\star\|_F$
- $\texttt{rel-RMSE}(\mathbf{W}) = \|\mathbf{W}^\star - \hat{\mathbf{W}}\|_F / \|\mathbf{W}^\star\|_F$

For each noise level σ, use multiple $\tilde{\lambda}$ between 1.5 and 10^{-11}, choose the best results



(a) Best $\tilde{\lambda}$ (b) Smallest errors

Figure 1: Best results with different noise levels

With the best $\tilde{\lambda}$, the errors cannot get **under 10^{-8}**

Majorization-Minimization Variant for Min-Vol

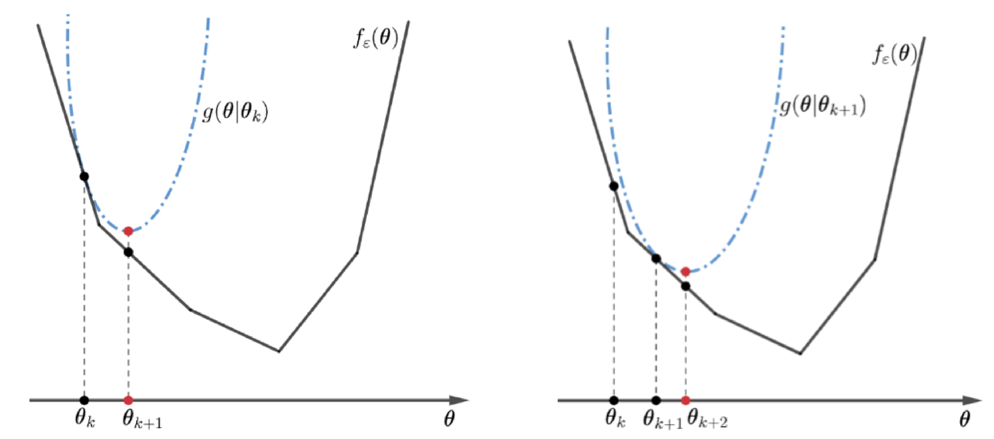
New problem (inspired by Square-Root Lasso)

$$\min_{\mathbf{W},\mathbf{H}} f(\mathbf{W},\mathbf{H}) := \sqrt{\|\mathbf{X}-\mathbf{WH}\|_F^2} + \lambda\mathrm{vol}(\mathbf{W})$$
$$\text{s.t.} \quad (\mathbf{W},\mathbf{H}) \in \mathcal{S}$$

Modified problem ("smoothen" square-root term)

$$\min_{\mathbf{W},\mathbf{H}} f_\varepsilon(\mathbf{W},\mathbf{H}) := \sqrt{\|\mathbf{X}-\mathbf{WH}\|_F^2+\varepsilon} + \lambda\mathrm{vol}(\mathbf{W})$$
$$\text{s.t.} \quad (\mathbf{W},\mathbf{H}) \in \mathcal{S}$$

where $\mathrm{vol}(\mathbf{W}) = \log\det(\mathbf{W}^\top\mathbf{W}+\delta\mathbf{I})$



Choose a surrogate function $g(\boldsymbol{\theta}|\boldsymbol{\theta}_k)$ for $f_\varepsilon(\boldsymbol{\theta})$ s.t.

$$f_\varepsilon(\boldsymbol{\theta}) \leq g(\boldsymbol{\theta}|\boldsymbol{\theta}_k), \text{ for all } \boldsymbol{\theta} \in \mathcal{S}.$$
$$f_\varepsilon(\boldsymbol{\theta}_k) = g(\boldsymbol{\theta}_k|\boldsymbol{\theta}_k)$$

Square-root Min-Vol

Algorithm 1 Square-Root Min-Vol NMF

Input: $\mathbf{X} \in \mathbb{R}_+^{m\times n}$, target rank r, λ, δ, ε.
Output: $\mathbf{W} \in \mathbb{R}_+^{m\times r}, \mathbf{H} \in \mathbb{R}_+^{r\times n}$ in $\mathcal{S}$

$(\mathbf{W}_1,\mathbf{H}_1) = \mathbf{SNPA}(\mathbf{X},r)$
$\lambda_1 = \lambda$
for $k = 1,\ldots$ **do**
  $(\mathbf{W}_{k+1},\mathbf{H}_{k+1}) = \mathbf{MinVol}(\mathbf{X},r,[\mathbf{W}_k,\mathbf{H}_k,\lambda_k,\delta])$
  $\lambda_{k+1} \leftarrow (2\sqrt{\|\mathbf{X}-\mathbf{W}_{k+1}\mathbf{H}_{k+1}\|_F^2+\varepsilon})\lambda$
end for

SNPA: successive nonnegative projection algorithm [1]

Results

With the same experiment set-up

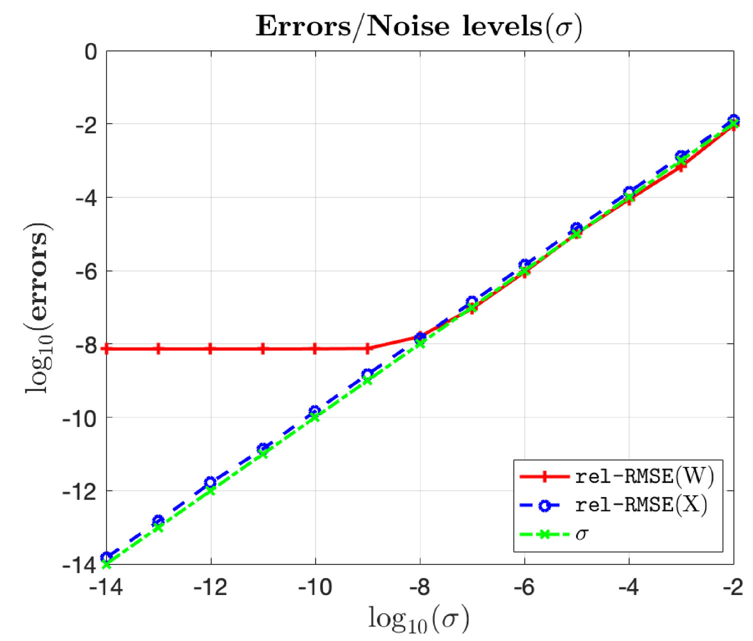


Figure 2: For **Square-Root Min-Vol**, both errors are much smaller, and $\texttt{rel-RMSE}(\mathbf{X}) = 10^{-14}$!

Applications

Applications: Hyperspectral unmixing (HU), topic modeling, representation learning.

Face Representation Learning:

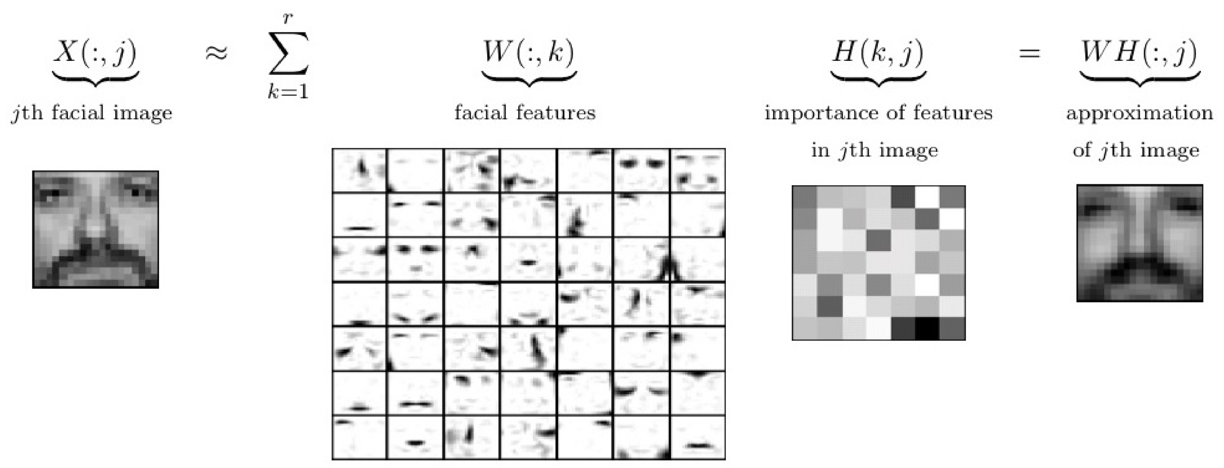


Figure 3: NMF for Representation Learning [2]. In matrix $\mathbf{X}$, each column represents a face image, while in matrix $\mathbf{W}$, each column represents a facial feature (learned basis).

Conclusion & Acknowledgement

Conclusion:

- The efficiency of **MinVol** algorithm depends on the initial λ choice
- The **MinVol** algorithm can be improved by the **Square-Root Min-Vol NMF**

Acknowledgement: This work was conducted in the 2023 REU STAT-DATASCI program at the Department of Statistics at Rice University and partially funded by a grant from the National Institute of General Medical Sciences (R01GM135928: EC).

References

[1] Nicolas Gillis. "Successive nonnegative projection algorithm for robust nonnegative blind source separation". In: *SIAM Journal on Imaging Sciences* 7.2 (2014), pp. 1420–1450.

[2] Nicolas Gillis. "The why and how of nonnegative matrix factorization". In: *Regularization, optimization, kernels, and support vector machines* 12.257 (2014), pp. 257–291.

[3] Valentin Leplat, Andersen M.S. Ang, and Nicolas Gillis. "Minimum-volume Rank-deficient Nonnegative Matrix Factorizations". In: *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2019, pp. 3402–3406.