Minimum-volume Rank-deficient Nonnegative matrix factorizations

Valentin Leplat, Andersen M. S. Ang and Nicolas Gillis
University of Mons, Rue de Houdain 9, 7000 Mons, Belgium
{valentin.leplat, manshun.ang, nicolas.gillis}@umons.ac.be

European Research Council

UMONS Université de Mons

Abstract

In recent years, nonnegative matrix factorization (NMF) with volume regularization has been shown to be a powerful identifiable model; for example for hyperspectral unmixing, document classification, community detection and hidden Markov models. We show that minimum-volume NMF (min-vol NMF) can also be used when the basis matrix is rank deficient, which is a reasonable scenario for some real-world NMF problems (e.g., for unmixing multispectral images). We propose an alternating fast projected gradient method for minvol NMF and illustrate its use on rank-deficient NMF problems; namely a synthetic data set and a multispectral image.

Min-vol NMF

Given $X \in \mathbb{R}_+^{m \times n}$ and rank r,

$$[W \in \mathbb{R}_+^{m\times r}, H \in \mathbb{R}_+^{r\times n}] = \underset{W\geq 0, H(:,j)\in\Delta^r \forall j}{\text{argmin}} ||X - WH||_F^2 + \lambda \,\text{vol}(W), \quad (1)$$

where Δ^r is the r-dimensional unit simplex, λ is a parameter, $\text{vol}(W) = \text{logdet}(W^TW + \delta I)$ is a function that measures the volume of the columns of W.

- Meaning : look for W with minimum volume to make the solution unique
- Under conditions on $X = WH$, this model recovers the true underlying (W, H) that generated X. [2, 3, 4]

Rank-deficient case

- A key assumption in min-vol NMF: the basis matrix W is full rank ($\text{rank}(W) = r$).
- It may happen that W is not full column rank; for example when $\text{rank}(X) \neq \text{rank}_+(X)$. Example:

$$X = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{pmatrix}, \qquad \text{rank}(X) = 3 < \text{rank}_+(X) = 4. \quad (2)$$

The columns of X are the vertices of a square in a 2-dimensional subspace; see Fig. 2. This could also happen for example in multispectral imaging : #materials in the image > #spectral bands (i.e. $r > m$ so $\text{rank}(W) \leq m < r$).

Focus of this work : the rank-deficient scenario, that is, $\text{rank}(W) < r$.

Min-vol NMF in the rank-deficient case

min-vol NMF model $$\min_{W\geq 0, H(:,j)\in\Delta^r\forall j} ||X - WH||_F^2 + \lambda\,\text{logdet}(W^TW + \delta I) \quad (3)$$

Choice of the volume regularizer

- Common volume functions are $\det(W^TW)$ and $\text{logdet}(W^TW + \delta I)$.
- $\text{vol}(W) = \text{logdet}(W^TW + \delta I)$. Note $\sqrt{\det(W^TW)}/r!$ is the vol of the convex hull of the col. of W and the origin.
- As $\det(W^TW) = \Pi_{i=1}^r \sigma_i^2(W)$, the log term weight down large σ_i.
- If W is rank deficient, some $\sigma_i(W) = 0$ so $\det(W^TW) = 0$. So $\det(W^TW)$ cannot distinguish between different rank-deficient sol.
- As $\text{logdet}(W^TW + \delta I) = \sum_{i=1}^r \log(\sigma_i^2(W) + \delta)$, if W has one/more σ_i equal to zero, this measure still makes sense: among two rank-deficient sol. belonging to the same low-dimensional subspace, minimizing $\text{logdet}(W^TW + \delta I)$ will favor a solution whose convex hull has a smaller volume within that subspace as decreasing the non-zero $\sigma_i(W^TW + \delta I)$ will decrease $\text{logdet}(W^TW + \delta I)$.

Choice of δ

- $\text{logdet}(W^TW + \delta I)$ is a non-convex surrogate for $\text{rank}(W)$.
- It is sharper than the nuclear norm for δ sufficiently small.
- So if one wants to promote rank-deficient solutions, δ should be small, say $\delta \leq 0.1$.
- δ should not be too small : (1) $WW^T + \delta I$ might be badly conditioned which makes the optimization problem harder to solve, (2) gives too much importance to zero singular values which might not be desirable.

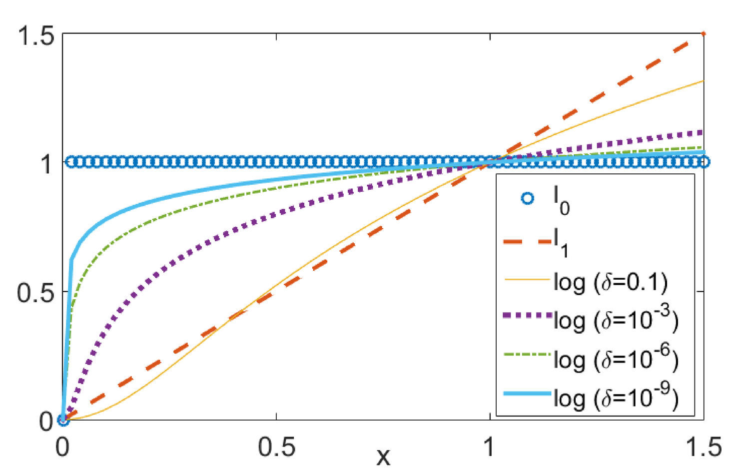

Figure 1: $\frac{\log(x^2+\delta)-\log\delta}{\log(1+\delta)-\log\delta}$, ℓ_1 norm and ℓ_0 norm.

Choice of λ

The choice of δ will influence the choice of λ : the smaller δ, the larger $|\text{logdet}(\delta)|$, to balance the two terms in the objective (3), λ should be smaller. For practical implementation, we initialize $W^{(0)} = X(:, \mathcal{K})$ where $\mathcal{K}$ is computed with the successive nonnegative projection algorithm (SNPA) that can handle the rank-deficient separable NMF problem. Note SNPA also provides the matrix $H^{(0)}$ so as to minimize $||X - W^{(0)}H^{(0)}||_F^2$ while $H^{(0)}(:, j) \in \Delta^r$ for all j. Finally, we will choose

$$\lambda = \tilde{\lambda}\frac{||X - W^{(0)}H^{(0)}||_F^2}{|\text{logdet}(W^{(0)T}W^{(0)} + \delta I)|},$$

where we recommend to choose $\tilde{\lambda}$ between 1 and 10^{-3} depending on the noise level (the noisier the input matrix, the larger λ should be).

Algorithm for min-vol NMF

Alternating minimization approach.

- On update H, use projected fast gradient method (PFGM)
- On update W, use PFGM applied on an strongly convex upper approximation of the objective function;

$$\ell(W) = ||X - WH||_F^2 + \lambda\,\text{logdet}(W^TW + \delta I) \leq 2\sum_{i=1}^n \left(\frac{1}{2}w_i^TAw_i - c_i^Tw_i\right) + b = \bar{\ell}(W),$$

where $Y = (Z^TZ + \delta I)^{-1}$ and $A = HH^T + \lambda Y$ are positive definite for $\delta, \lambda > 0$, $C = XH^T$, and b is a constant independent of W. Note $\bar{\ell}(W) = \ell(W)$ for $Z = W$.

Minimizing the upper bound $\bar{\ell}(W)$ of $\ell(W)$ requires to solve m independent strongly convex optimization problems with Hessian matrix A.

- PFGM has a linear rate of convergence $1 - \sqrt{\kappa^{-1}}$ where κ is the condition number of A.
- subproblem on H is not strongly cvx when W is rank deficient; PFGM converges sublinearly
- when W is rank deficient; $\frac{\lambda}{\delta} \leq \lambda_{\max}(A) \leq ||H||_2^2 + \frac{\lambda}{\delta}$ and as smaller δ gives larger the conditioning of A hence the slower will be the PFGM.

Min-vol NMF using alternating PFGM

- Initialize (W, H) using SNPA, let $\lambda = \tilde{\lambda}\frac{||X-WH||_F^2}{\text{logdet}(W^TW+\delta I)}$.
- For $k = 1, 2, \ldots$
  - Let $A = HH^T + \lambda(W^TW + \delta I)^{-1}$ and $C = XH^T$.
  - Perform a few steps of PFGM on the problem $\min_{U\geq 0}\frac{1}{2}\langle U^TU, A\rangle - \langle U, C\rangle$, with initialization $U = W$
  - Perform a few steps of PFGM on problem $\min_{H(:,j)\in\Delta^r\forall j} ||X - WH||_F^2$

Numerical Experiments

Synthetic data set. $X = WH \in \mathbb{R}^{4\times 500}$ constructed with W as matrix from (2) so $\text{rank}(W) = 3 < r = 4$, and each col. of H is distributed using the Dirichlet distribution of parameter $(0.1, \ldots, 0.1)$. Each col. of H with an entry larger 0.8 is resampled as long as this condition does not hold. This guarantees that no data point is close to a col. of W (this is sometimes referred to as the purity index). As observed on Fig. 2, proposed algorithm is able to perfectly recover the true col. of W.

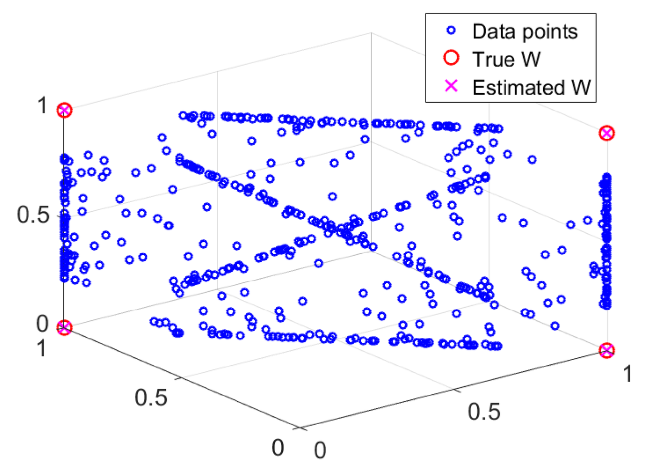

Figure 2: Synthetic data set and recovery. (Only the first three entries of each four-dimensional vector are displayed.)

Fig. 3 illustrates the same experiment where noise is added to $X = \max(0, WH + N)$ where $N = \epsilon$ randn(m,n) in Matlab notation (i.i.d. Gaussian distribution of mean zero and standard deviation ϵ). Note that the average of the entries of X is 0.5 (each col. is a linear combination of the col. of W, with weights summing to one). Fig. 3 displays the average over 20 randomly generated matrices X of the relative error $d(W, \tilde{W}) = \frac{||W-\tilde{W}||_F}{||W||_F}$ where $\tilde{W}$ is the solution computed by Alg. depending on the noise level ϵ. This illustrates that min-vol NMF is robust against noise since the $d(W, \tilde{W})$ is smaller than 1% for $\epsilon \leq 1\%$.

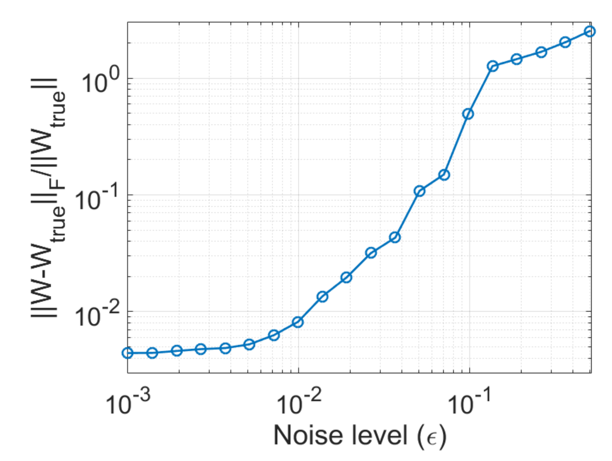

Figure 3: Evolution of the recovery of the true W depending on the noise $N = \epsilon$ rand(m,n) using Alg. ($\tilde{\lambda} = 0.01$, $\delta = 0.1$, maxiter = 100).

Multispectral image. San Diego airport is a hyperspectral image (HSI) : 158 clean bands, 400×400 pixels for each spectral image. Mainly 3 types of materials: road surfaces, roofs and vegetation. The image can be well approximated using r=8. As we are interested in $\text{rank}(W) < r$, we pick m=5 spectral band using successive projection algorithm (Gram-Schmidt with column pivoting) applied on X^T. This provides bands that are representative and we are factoring 5-by-160000 matrix using a r=8. Here we used $\tilde{\lambda}$=0.1 and 1000 iterations. From the initial solution provided by SNPA, min-vol NMF reduce error $||X - WH||_F$ by a factor of 11.7 while term $\text{logdet}(W^TW + \delta I)$ only increases by a factor of 1.06. Final relative error $\frac{||X-WH||_F}{||X||_F} = 0.2\%$.

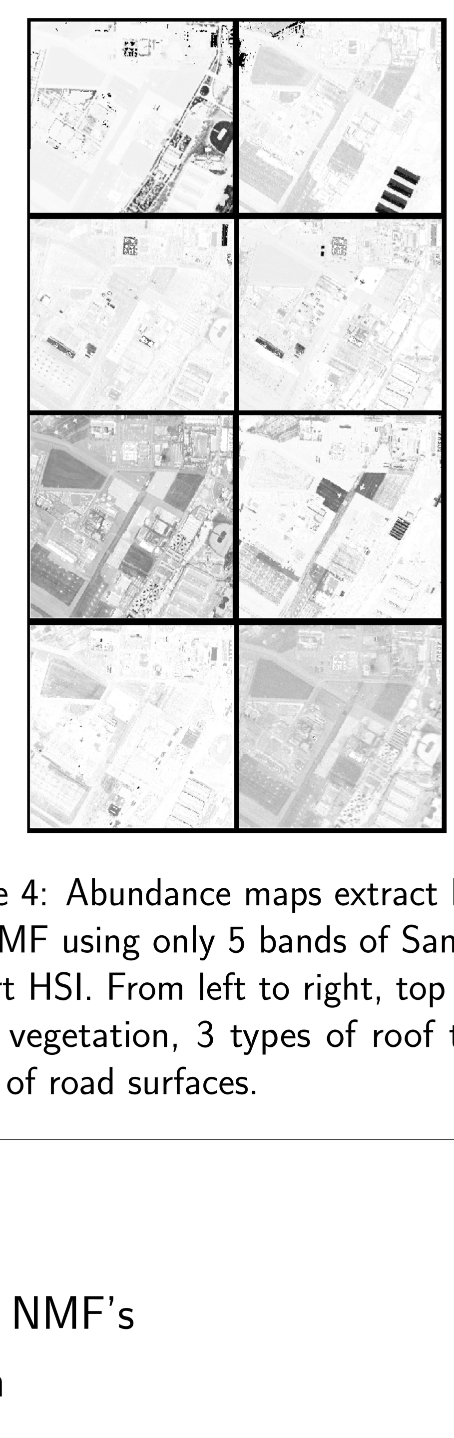

Figure 4: Abundance maps extract by min-vol NMF using only 5 bands of San Diego airport HSI. From left to right, top to bottom: vegetation, 3 types of roof tops, 4 types of road surfaces.

Conclusion

- min-vol NMF can be used meaningfully for rank-deficient NMF's
- We proposed an efficient algorithm to tackle this problem
- Open questions
  - Under which conditions can we prove the identifiability of min-vol NMF in the rank-deficient case ?
  - Can we prove robustness to noise of such techniques? (The question is also open for the full-rank case.)
  - Can we design faster and more robust algorithms? And algorithms taking advantage of the fact that the solution is rank-deficient?