Beyond Separability in Nonnegative Matrix Factorization

Nicolas Nadisic[1,(2)]

25 May 2023 — Institut de Mathématiques de Bordeaux

[1]Ghent University, Belgium
[2]University of Mons, Belgium

Outline

1. Introduction — NMF and separability
2. Sparse separable NMF
3. Smoothed separable NMF

Introduction — NMF and separability

Our motivation

High-level motivations of this work:

- Extract underlying structures in data
- Better leverage a priori knowledge, notably nonnegativity, separability, and sparsity, to improve models
- Develop algorithms that are both guaranteed and computationally tractable

Starting point: Nonnegative matrix factorization (NMF)

Given $B \in \mathbb{R}_+^{m \times n}$ and $r \in \mathbb{N}$, find $A \in \mathbb{R}_+^{m \times r}$, and $X \in \mathbb{R}_+^{r \times n}$ such that $B \approx AX$.

Starting point: Nonnegative matrix factorization (NMF)

Given $B \in \mathbb{R}_+^{m \times n}$ and $r \in \mathbb{N}$, find $A \in \mathbb{R}_+^{m \times r}$, and $X \in \mathbb{R}_+^{r \times n}$ such that $B \approx AX$.

In practice, a common formulation is

Frobenius NMF

$$\min_{A \geq 0, X \geq 0} \|B - AX\|_F^2$$

Starting point: Nonnegative matrix factorization (NMF)

Given $B \in \mathbb{R}_+^{m \times n}$ and $r \in \mathbb{N}$, find $A \in \mathbb{R}_+^{m \times r}$, and $X \in \mathbb{R}_+^{r \times n}$ such that $B \approx AX$.

In practice, a common formulation is

Frobenius NMF

$$\min_{A \geq 0, X \geq 0} \|B - AX\|_F^2$$

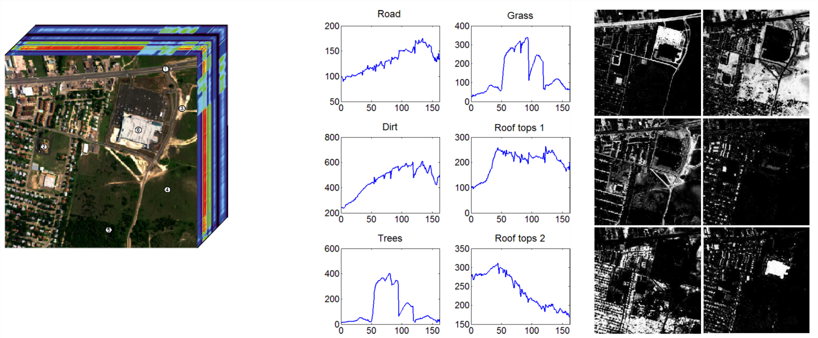
- Columns of A are features
- Columns of B are data points that can be expressed as additive linear combinations of features
- Entries of X represent the weight of each feature in each data point

Why nonnegativity?

- More interpretable factors (part-based representation)
- Naturally favors sparsity
- Is natural in many applications (image processing, hyperspectral unmixing, text mining, ...)

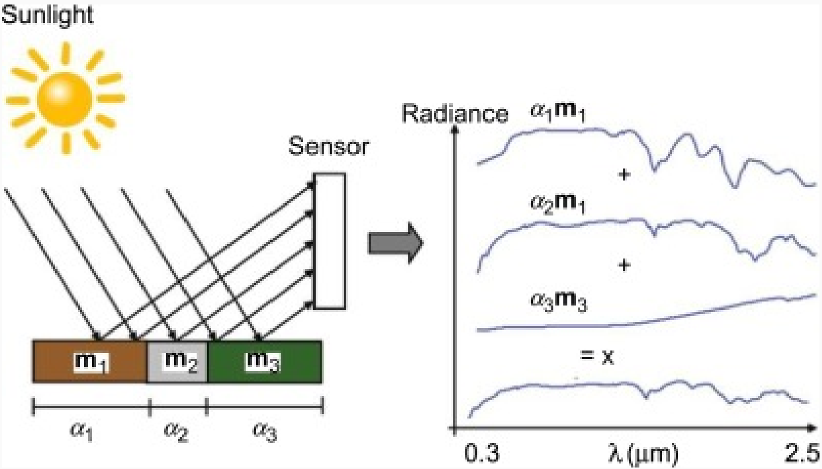
One application — Hyperspectral unmixing

$$\underbrace{B(:,j)}_{\substack{\text{spectral signature of}\\ \text{j-th pixel}}} \approx \sum_{p} \underbrace{A(:,p)}_{\substack{\text{spectral signature of}\\ \text{p-th material}}} \underbrace{X(p,j)}_{\substack{\text{abundance of p-th material}\\ \text{in j-th pixel}}}$$

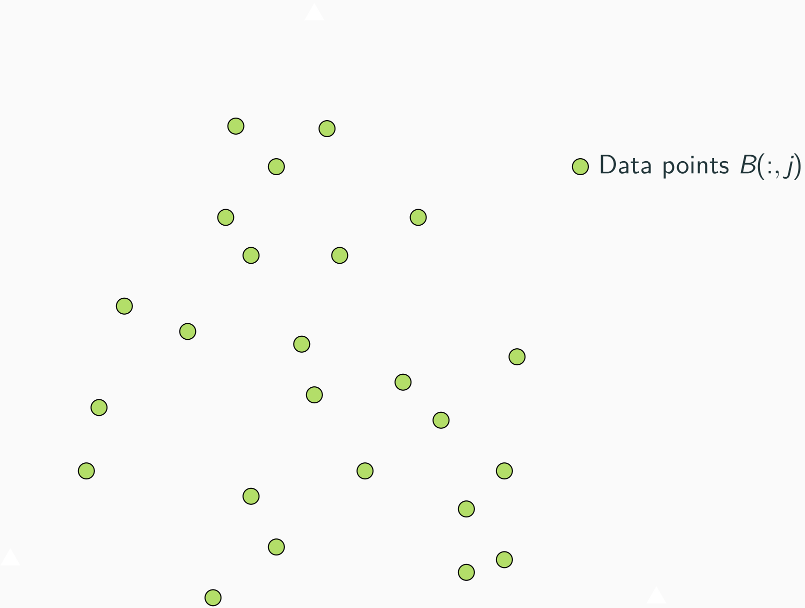


Images from J. Bioucas Dias and N. Gillis.

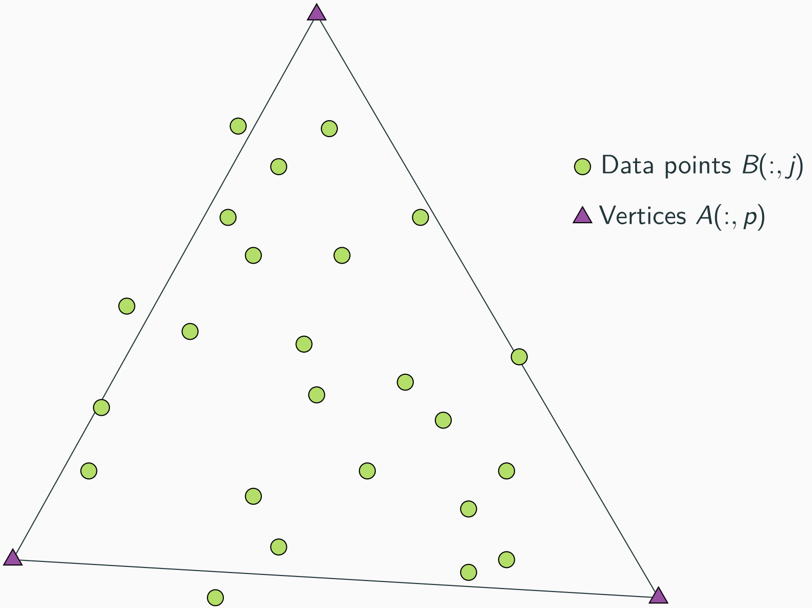
Linear mixing model



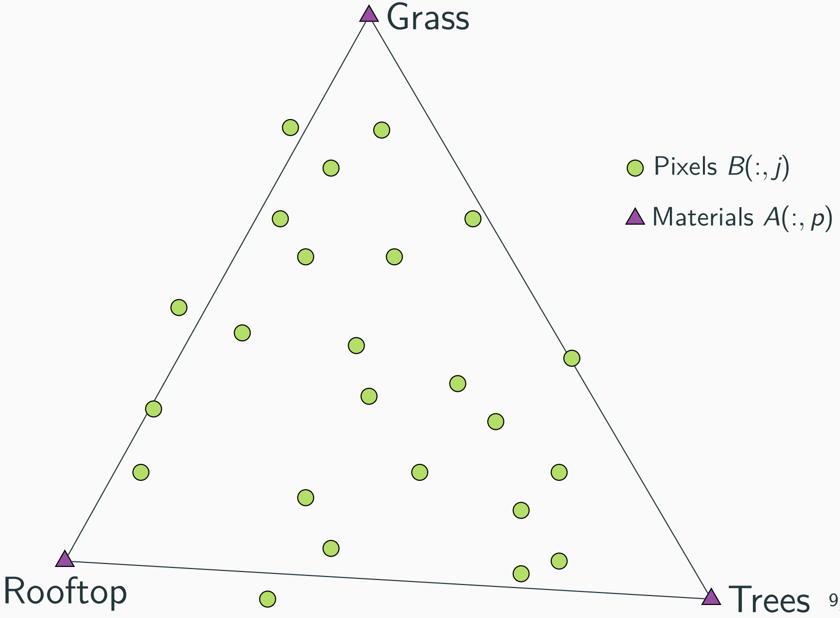
NMF Geometry ($B \approx AX$)



NMF Geometry ($B \approx AX$): cone / convex hull



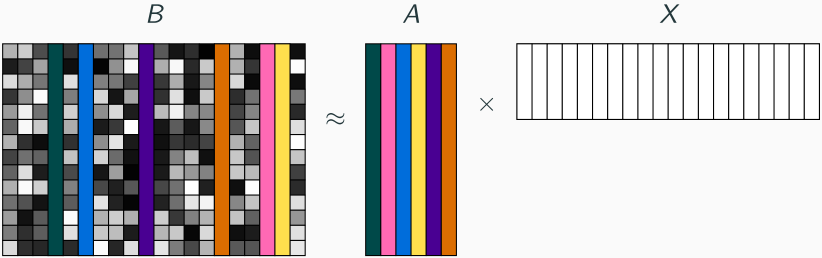
Application — Hyperspectral unmixing



The separability assumption

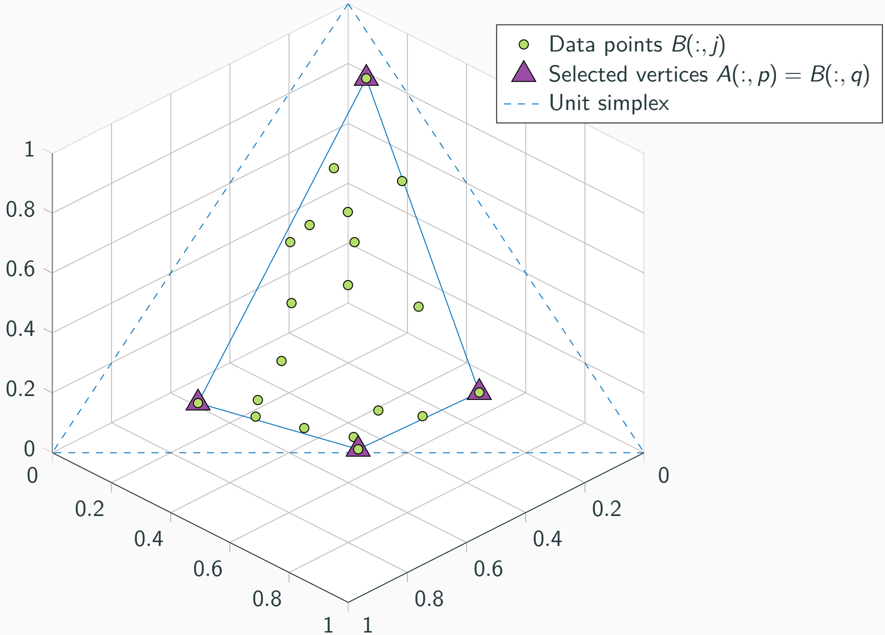
For each vertex, there exist at least one data point equal/close to this vertex

⇔ pure-pixel assumption



⇔ There exists an index set $\mathcal{J}$ with $|\mathcal{J}| = r$ such that $B \approx B(:, \mathcal{J})X$

Geometry of separable NMF



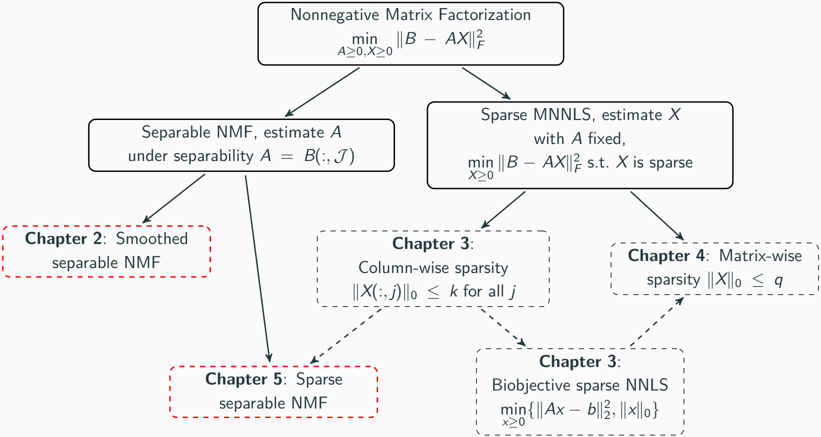
Separable NMF

- NMF is NP-hard in general (Vavasis 2010).
- Under the separability assumption, it's solvable in polynomial time (Arora et al. 2012).

Separable NMF is actually quite old

- Donoho and Stodden (2004) $\Rightarrow$ term "separability"
- Boardman, Kruse, and Green (1995) $\Rightarrow$ pure-pixel assumption in HSI
- Used since the 1970's in chemometrics

Overview of my PhD thesis



Sparse separable NMF

Sparse separable NMF

Presented in the article:

NN, Arnaud Vandaele, Jeremy E Cohen, and Nicolas Gillis (2020). "Sparse separable nonnegative matrix factorization". In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECMLPKDD)*, pp. 335–350.

Why? No work handles the underdetermined case with interior vertices, nor leverages sparsity

What? New model and exact algorithm for separable NMF with sparsity constraints, identifiability and complexity proofs

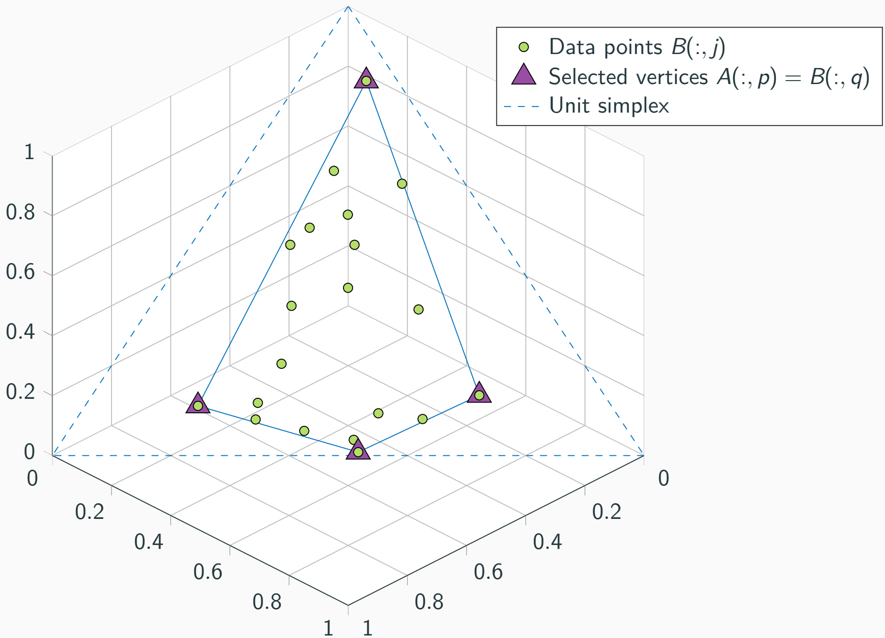
Starting point 1/2 — Separable NMF

Separability:

- The vertices are selected among the data points
- In hyperspectral unmixing, equivalent to pure-pixel assumption

Standard NMF model	$B = AX$
Separable NMF	$B = B(:, \mathcal{J})X$

Starting point 1/2 — Separable NMF



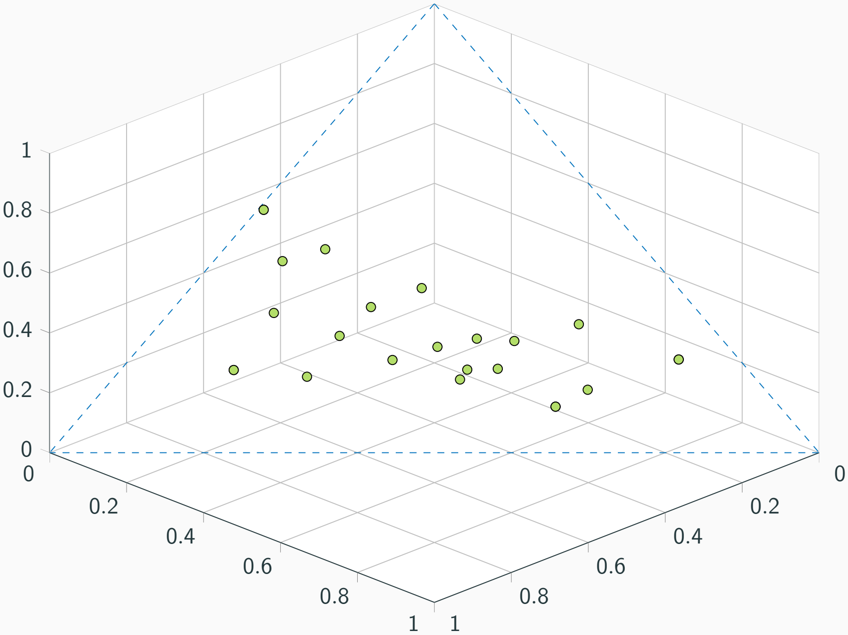
Algorithm for Separable NMF — SNPA

SNPA = Successive Nonnegative Projection Algorithm (Gillis 2014)

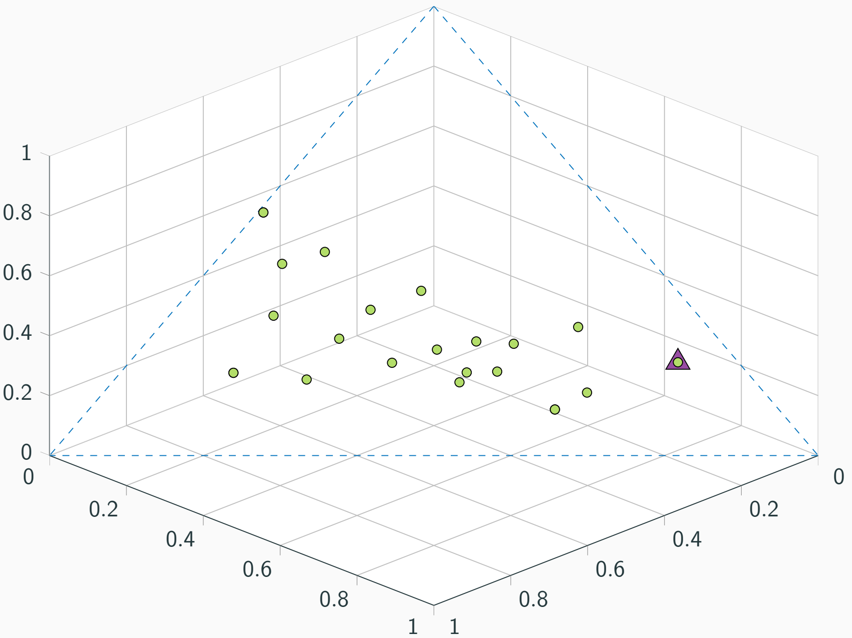
- Start with empty A, and residual $R = B$
- Alternate between
  - Greedy selection of one column of R to be added to A
  - Projection of R on the convex hull of the origin and columns of A
- Stop when reconstruction error $= 0$ (or $< \epsilon$)

(Condition: columns of B have unit ℓ_1-norm)

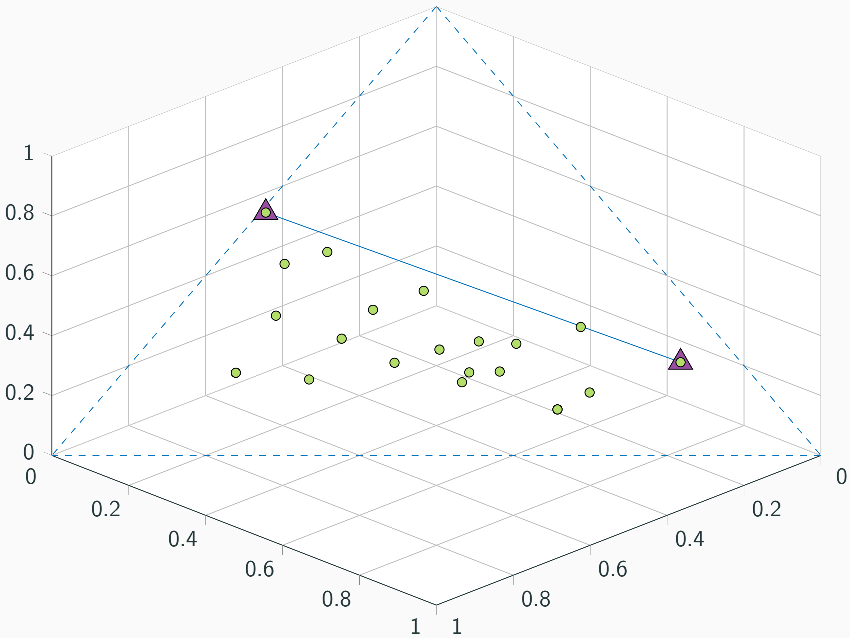
SNPA



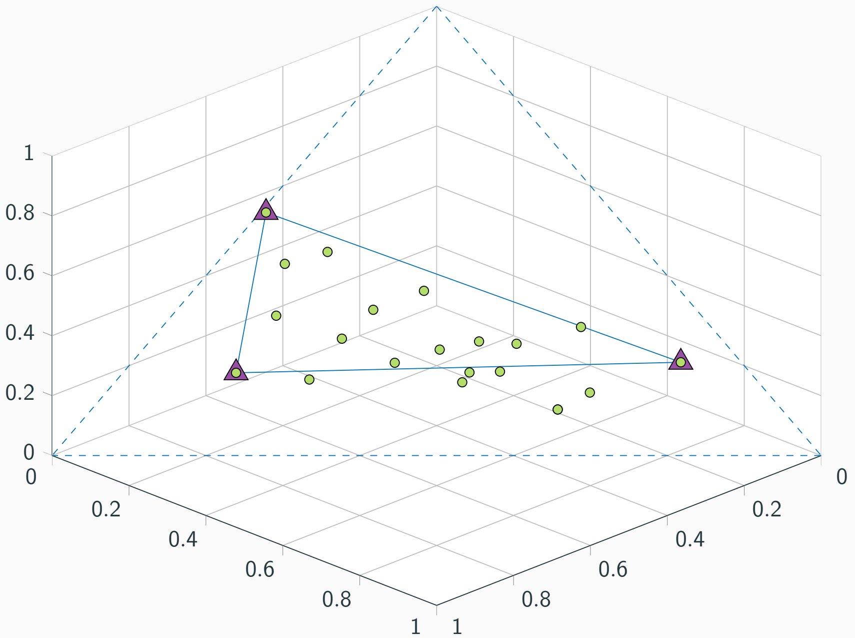
SNPA



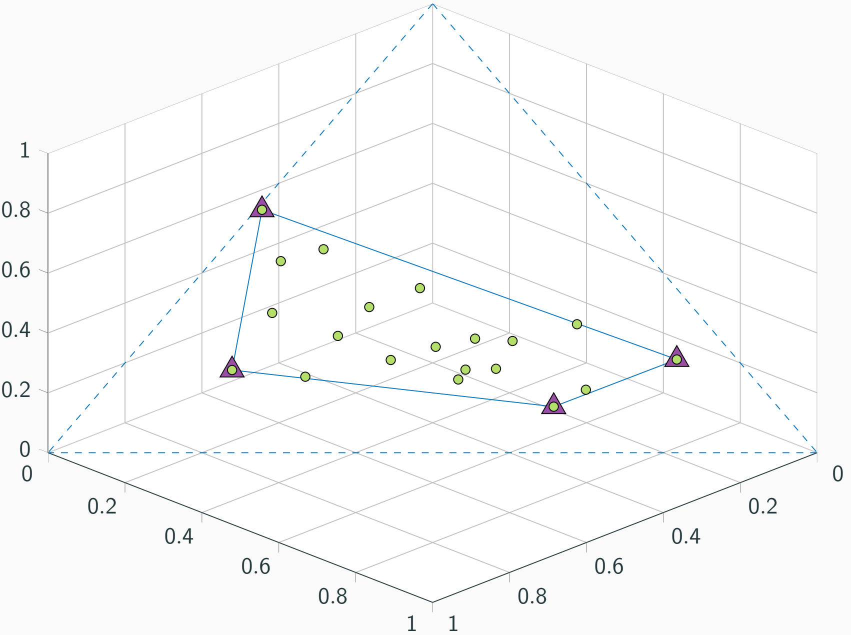
SNPA



SNPA



SNPA



A limitation of Separable NMF

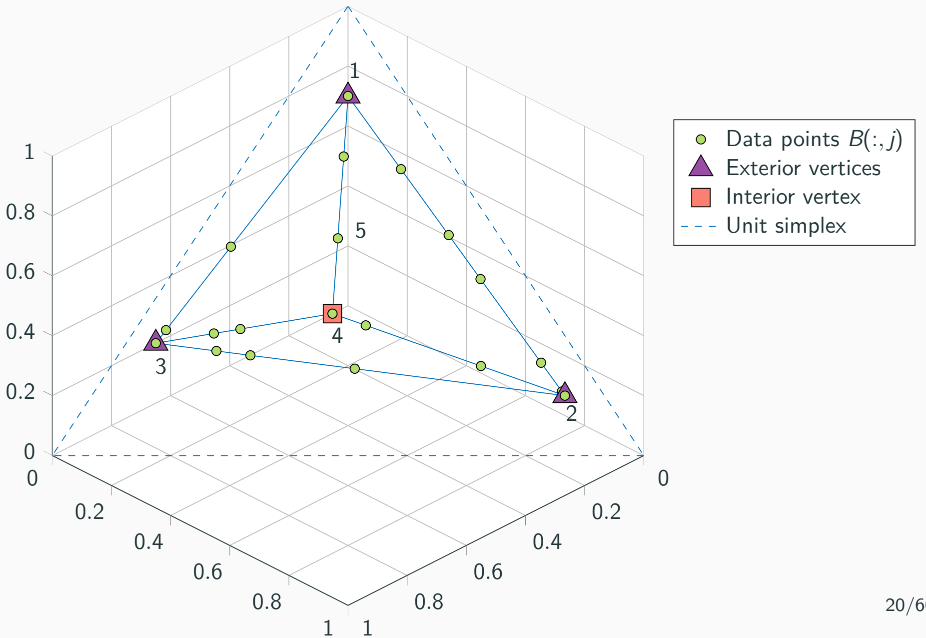
What if one column of A is a combination of others columns of A?

Ex: multispectral unmixing with $m < r$

$\rightarrow$ Interior vertex

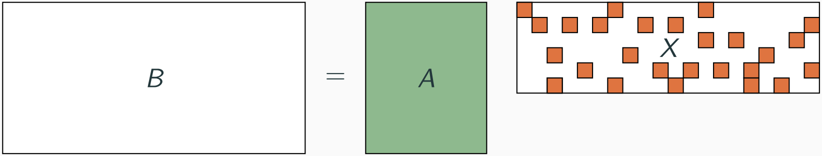
Not identifiable by separable NMF, because it belongs to the convex hull of the other vertices.

A limitation of Separable NMF



Starting point 2/2 — k-sparsity

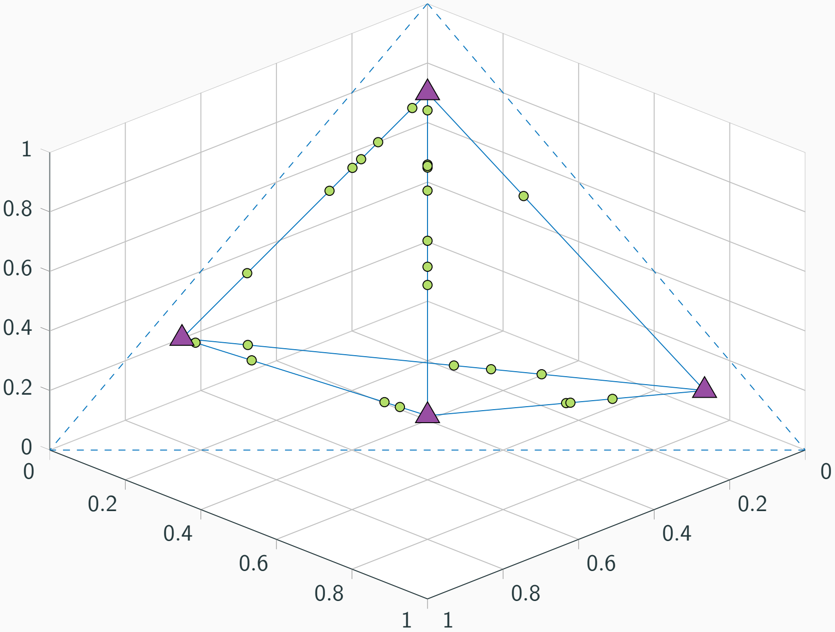
- $B \approx AX$ s.t. X is column-wise k-sparse
- Interpretation: a pixel expressed as a combination of at most k materials



k-sparse nonnegative least squares (NNLS)

$$\min_{X \geq 0} \|B - AX\|_F^2 \quad \text{s.t.} \quad \|X(:,j)\|_0 \leq k \text{ for all } j$$

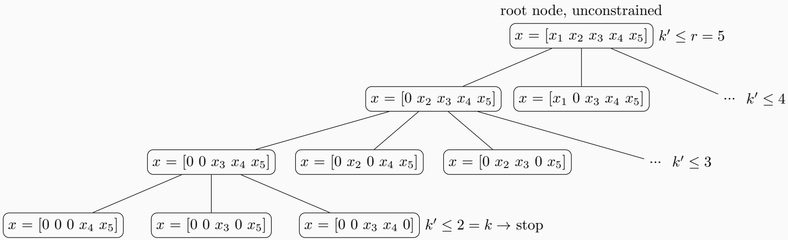
Starting point 2/2 — k-sparsity (ex. with $k = 2$)



Starting point 2/2 — k-sparsity

k-sparse NNLS is combinatorial, with $\binom{r}{k}$ possible combinations per column of X.

Previous work: a branch-and-bound algorithm to solve exactly k-sparse NNLS (NN, Vandaele, Gillis, et al. 2020).



Ex. of the BnB algorithm with $r = 5$ and $k = 2$

Sparse Separable NMF

Standard NMF model $B = AX$

Separable NMF $B = B(:, \mathcal{J})X$

Sparse sep NMF $B = B(:, \mathcal{J})X$ s.t. for all j, $\|X(:, j)\|_0 \leq k$

Our objective: handle situation separable NMF cannot, interior vertices and underdetermined cases, using a prior sparsity knowledge.

Our approach for SSNMF

Replace the projection step of SNPA, from projection on convex hull to projection on k-sparse hull, done with our BnB solver $\Rightarrow$ kSSNPA.

kSSNPA

- Identifies all interior vertices (non-selected points are never vertices)
- May also identify wrong vertices (explanation to come!)

$\Rightarrow$ kSSNPA can be seen as a screening technique to reduce the number of points to check.

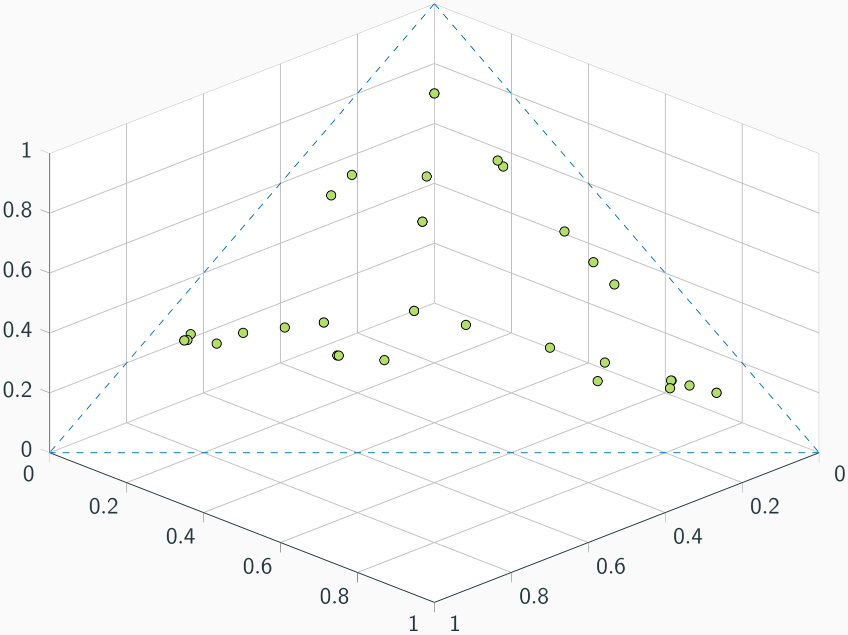
Our approach for SSNMF

In a nutshell, 3 steps:

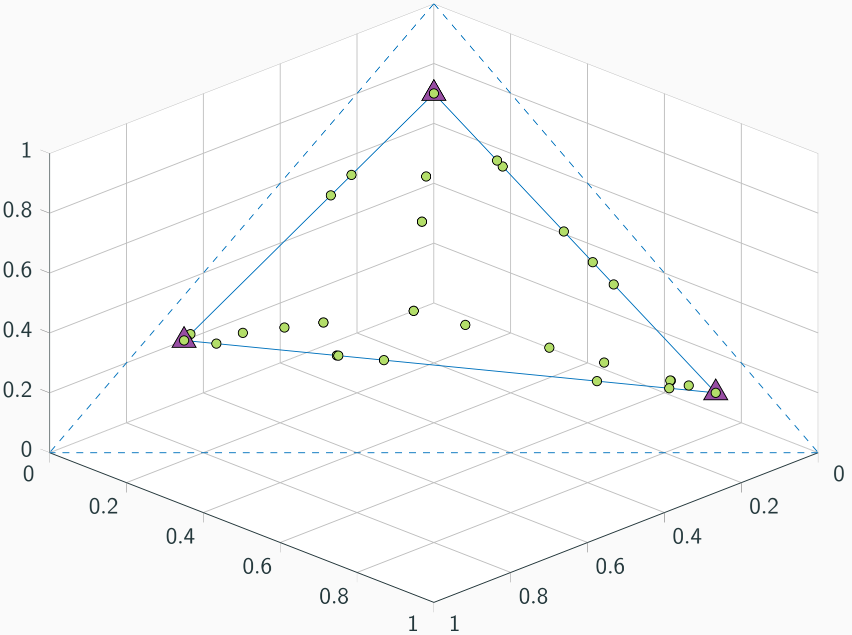
1. Identify exterior vertices with SNPA
2. Identify candidate interior vertices with kSSNPA
3. Discard bad candidates, those that are k-sparse combinations of other selected points (they cannot be vertices)

Our algorithm: BRASSENS Relies on Assumptions of Sparsity and Separability for Elegant NMF Solving.

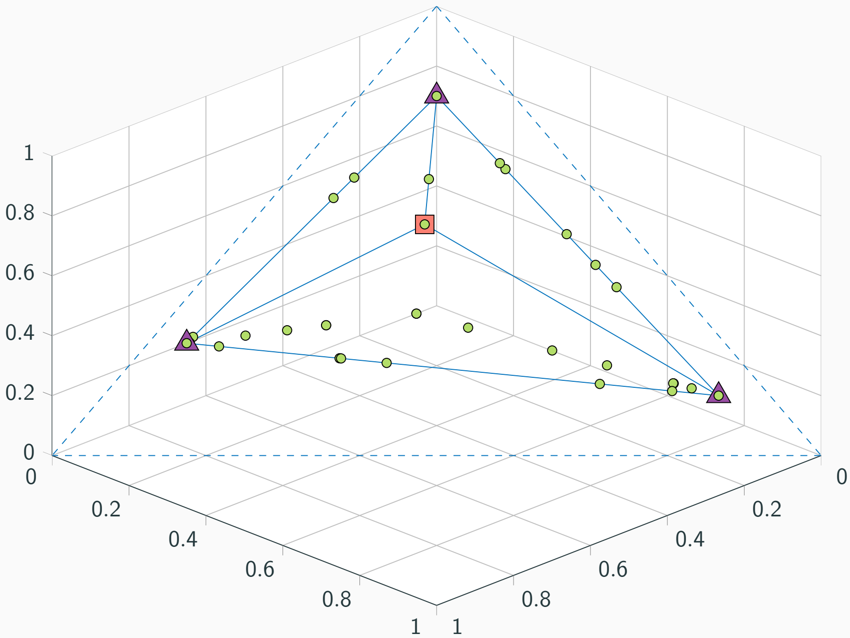
Brassens with sparsity $k = 2$



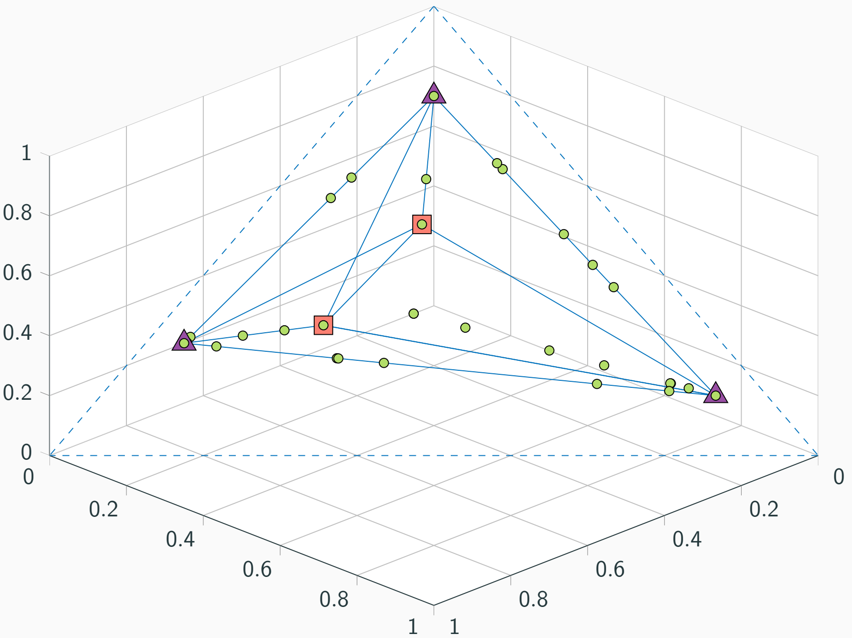
Brassens with sparsity $k = 2$



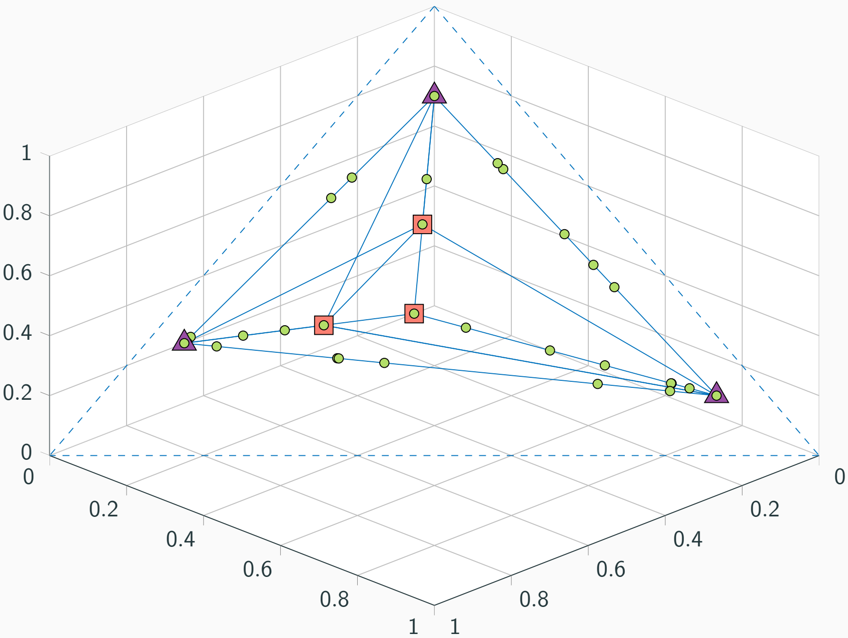
Brassens with sparsity $k = 2$



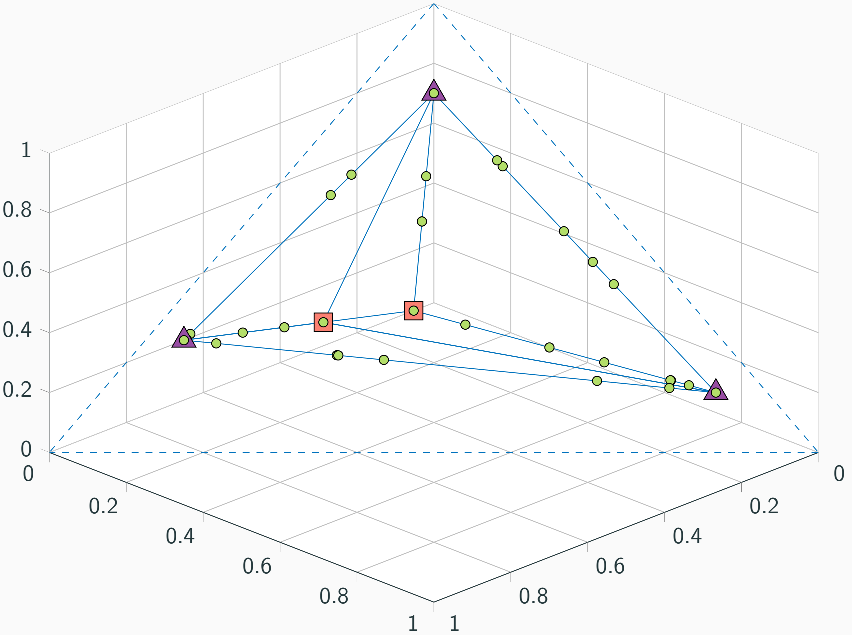
Brassens with sparsity $k = 2$



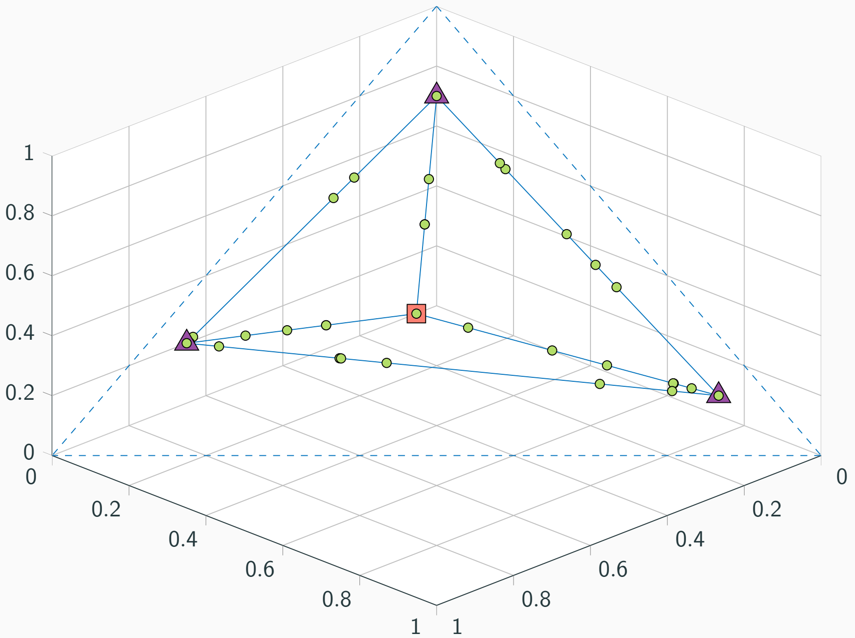
Brassens with sparsity $k = 2$



Brassens with sparsity $k = 2$



Brassens with sparsity $k = 2$



Complexity

- As opposed to Separable NMF, Sparse Separable NMF is NP-hard (proof in the paper and thesis)
- Hardness comes from the k-sparse projection
  - If k is a fixed constant, not NP-hard anymore
- Not too bad when r is small, with our BnB solver

Correctness

Assumption 1 No column of A is a nonnegative linear combination of k other columns of A.
⇒ necessary condition for recovery by Brassens

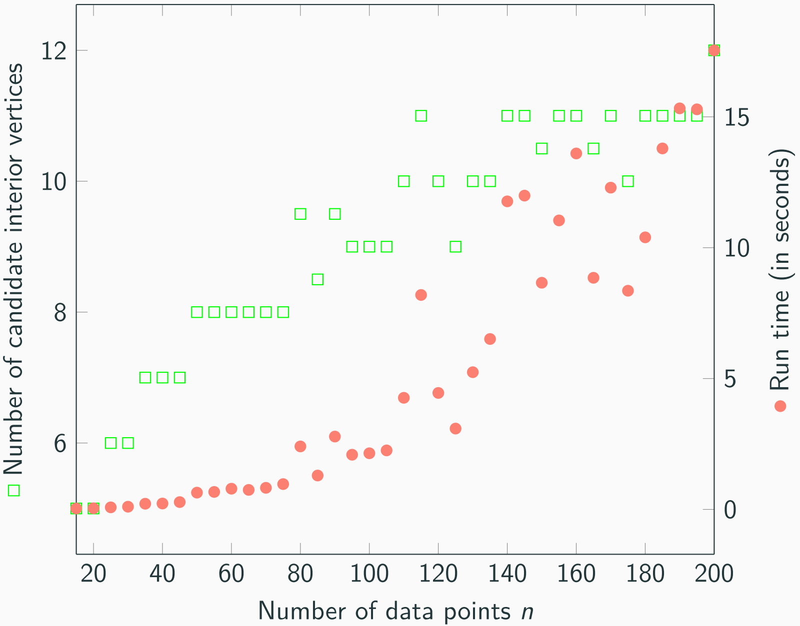
Assumption 2 No column of A is a nonnegative linear combination of k other columns of B.
⇒ sufficient condition for recovery by Brassens

If data points are k-sparse and generated at random, **Assumption 2** is true with probability one.

Experiments

- Experiments on synthetic datasets with interior vertices
- Experiment on underdetermined multispectral unmixing (Urban image, 309×309 pixels, limited to $m = 3$ spectral bands, and we search for $r = 5$ materials)
- No other algorithm can tackle SSNMF, so comparisons are limited

XP Synthetic: 3 exterior and 2 interior vertices, n grows



XP Synthetic 2: dimensions grow

m	n	r	k	Number of candidates	Run time in seconds
3	25	5	2	5.5	0.26
4	30	6	3	8.5	3.30
5	35	7	4	9.5	38.71
6	40	8	5	13	395.88

Conclusion from experiments:

- kSSNPA is efficient to select few candidates
- Still, Brassens does not scale well :(

XP on 3-bands Urban dataset with $r = 5$

SNPA

Grass+Trees +Rooftops | Rooftops 1 | Dirt+Road +Rooftops | Dirt+Grass | Rooftops 1 +Dirt+Road

BRASSENS (finds 1 interior point)

Grass+Trees | Rooftops 1 | Road | Rooftops+Road | Dirt+Grass

Conclusion

Sparse Separable NMF, a new model that combine constraints of separability and k-sparsity:

- Can handle some cases that Separable NMF cannot handle, such as interior vertices in underdetermined problems
- We proved it is NP-hard (unlike Sep NMF), but actually "not so hard" for small r
- It is provably solved by our algorithm Brassens under mild assumptions

Limitations:

- Brassens does not scale well
- Theoretical results limited to the noiseless case
- Limited robustness to noise

Smoothed separable NMF

Smoothed separable NMF

Presented in the article:

NN, Nicolas Gillis, and Christophe Kervazo (2021). "Smoothed separable nonnegative matrix factorization". In: *preprint arXiv:2110.05528*.

Why? Separable NMF is popular and powerful but algorithms do not leverage the presence of multiple pure data points (only one does so, and it has limitations)

What? Two smoothed separable NMF algorithms that outperform the state of the art

Model 1: Separable NMF (reminder)

Separability assumption

There exists an index set $\mathcal{J}$ with $|\mathcal{J}| = r$ such that

$$B = B(:, \mathcal{J})X + N$$

(where N is bounded noise)

Interpretation: for each vertex, there exist at least one data point equal to this vertex $\Leftrightarrow$ pure-pixel assumption

Model 1: Separable NMF (reminder)

Separability assumption

There exists an index set $\mathcal{J}$ with $|\mathcal{J}| = r$ such that

$$B = B(:, \mathcal{J})X + N$$

(where N is bounded noise)

Interpretation: for each vertex, there exist at least one data point equal to this vertex $\Leftrightarrow$ pure-pixel assumption

Algorithms: here we focus on two greedy algorithms

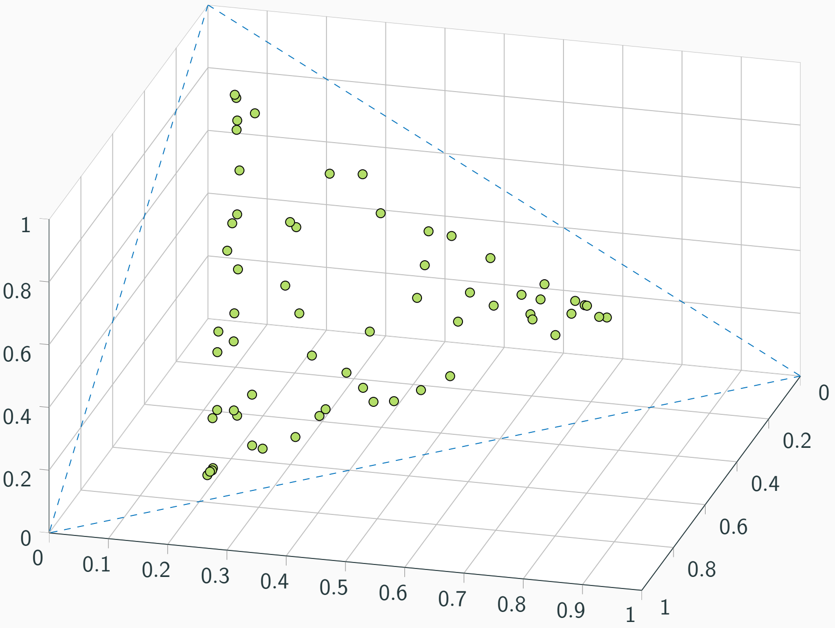
- VCA: Vertex Component Analysis (Nascimento et al. 2005)
- SPA: Successive Projection Algorithm (Araújo et al. 2001)

VCA in a nutshell (Nascimento et al. 2005)

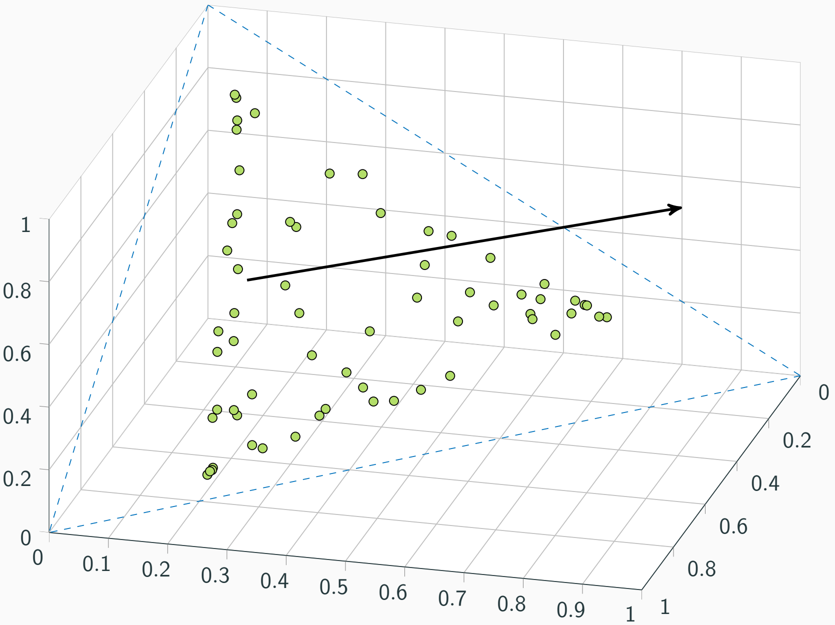
- Greedy selection of vertices
- Random orthogonal projections

Advantage: randomized algorithm, can be run several times to keep the best solution
Issue: not provably robust to noise

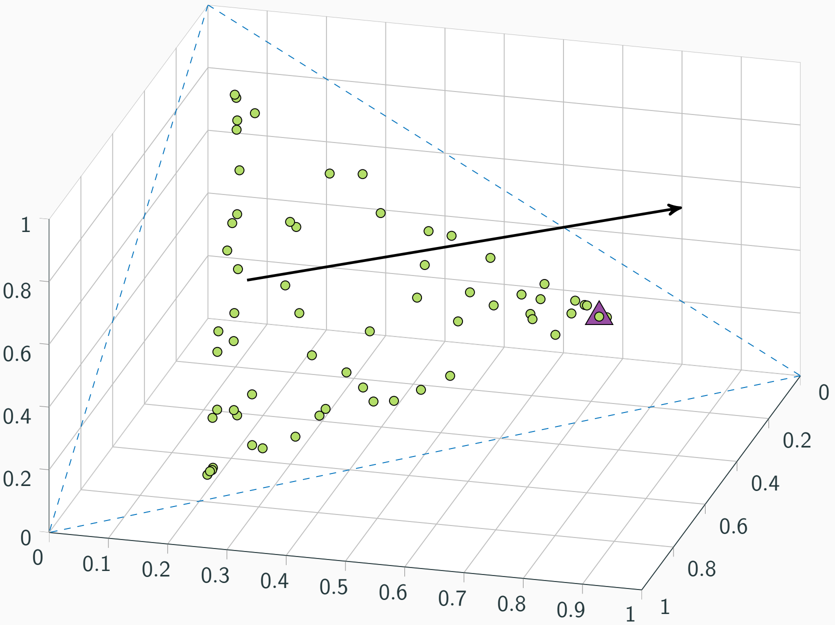
VCA — Animation



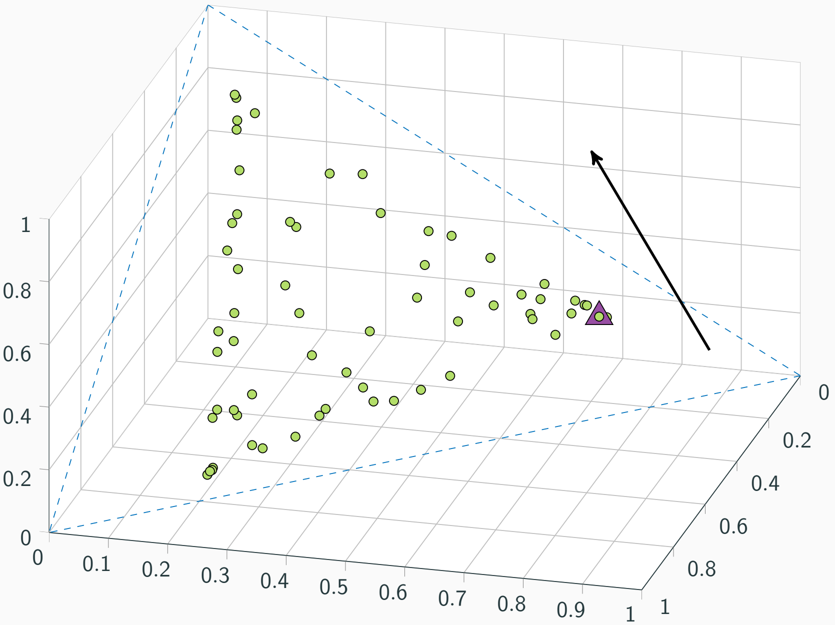
VCA — Animation



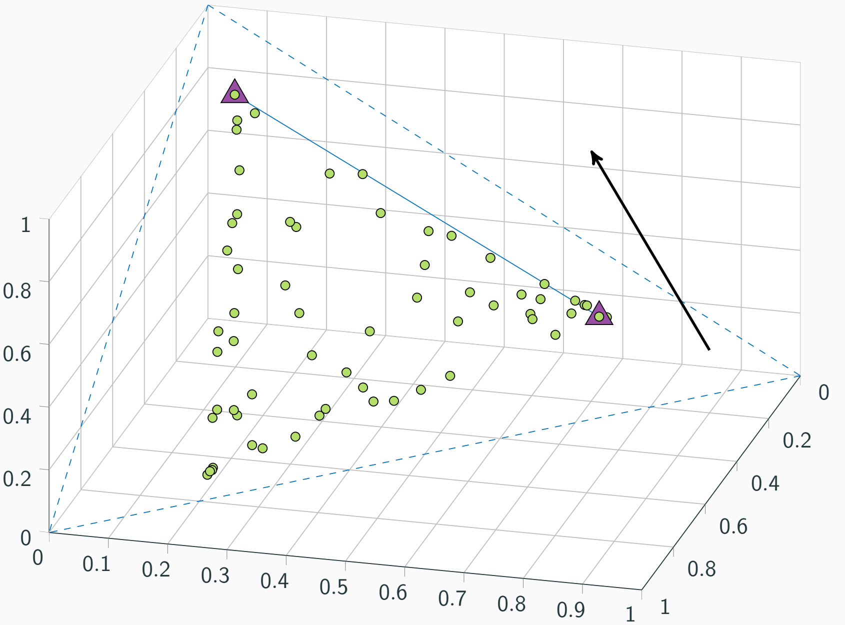
VCA — Animation



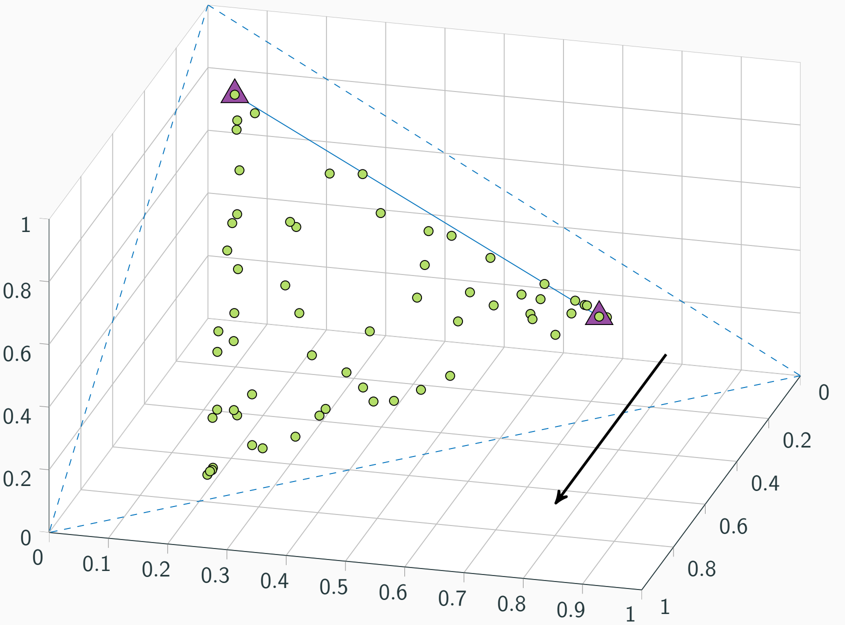
VCA — Animation



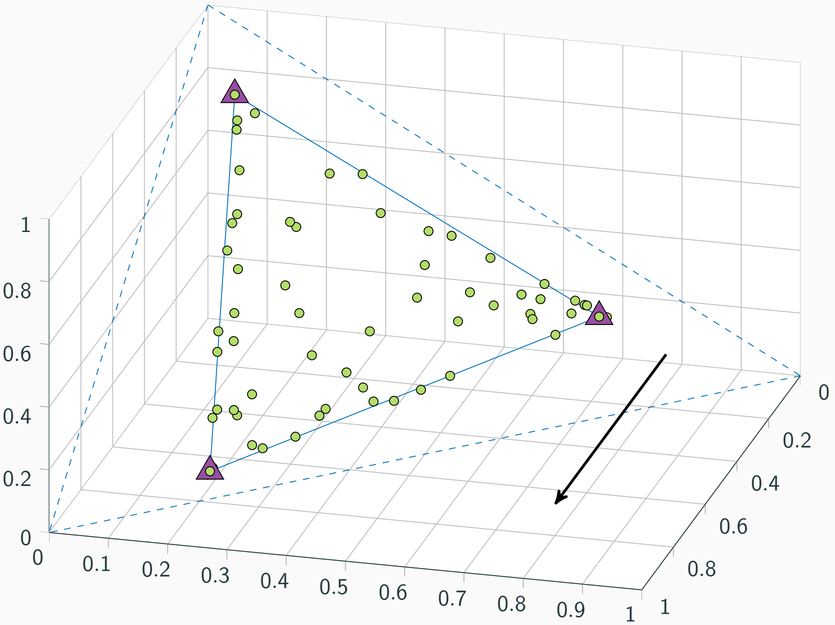
VCA — Animation



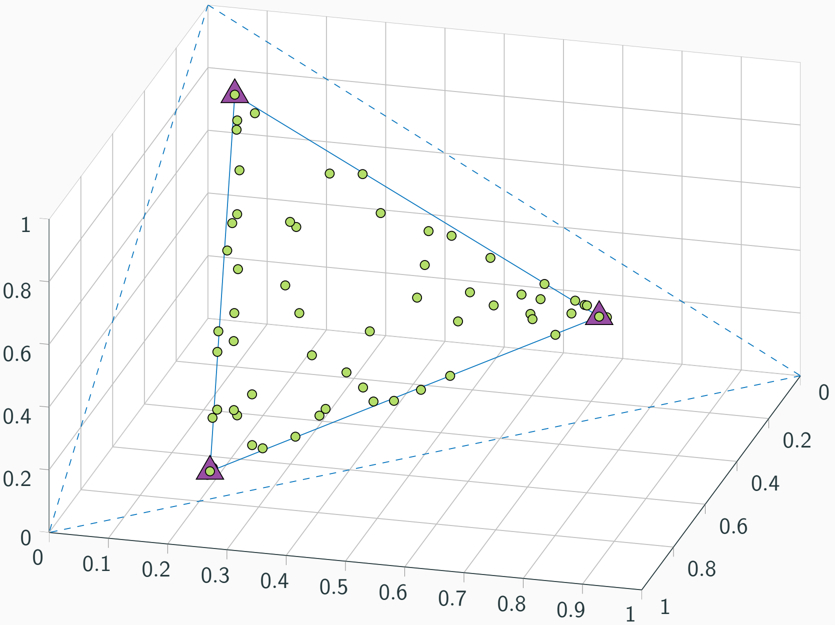
VCA — Animation



VCA — Animation



VCA — Animation

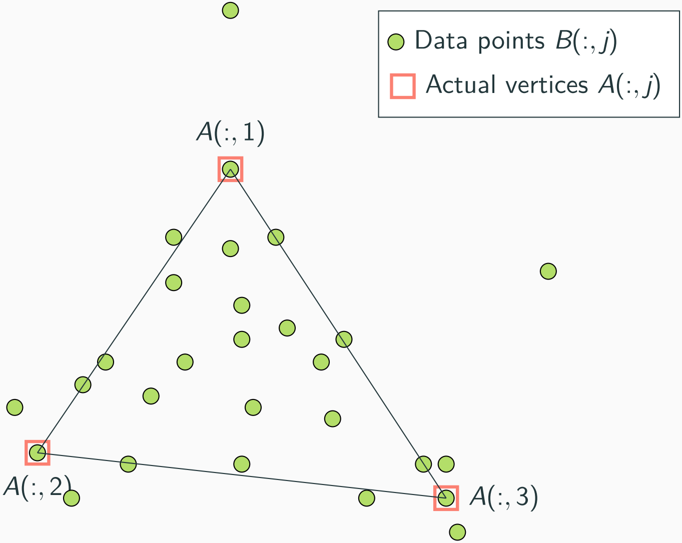


SPA in a nutshell (Araújo et al. 2001)

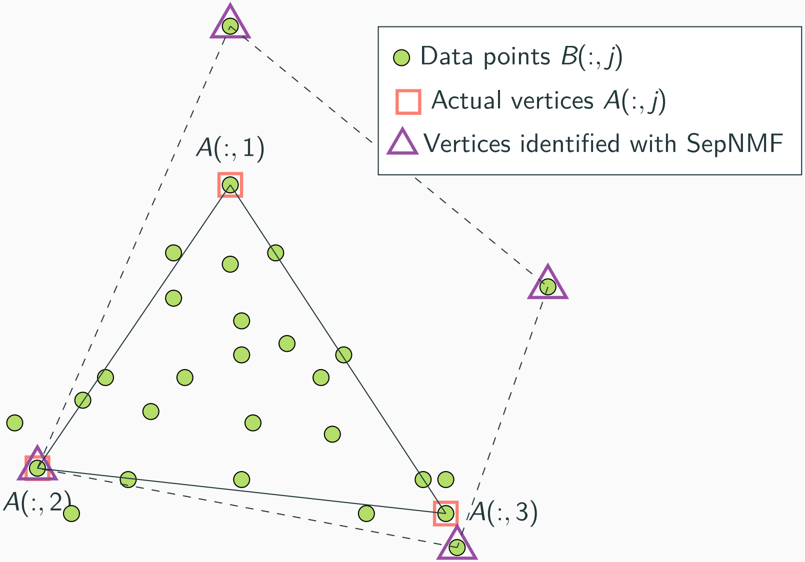
- Similar to VCA
- Orthogonal projection with no randomness
- Selects the column of the residual with highest ℓ_2-norm

Advantage: provably robust to noise (column-wise bounds for N)
Issue: deterministic

Issues of Separable NMF: outliers, extreme points



Issues of Separable NMF: outliers, extreme points



Model 2: Proximal latent points (Bhattacharyya et al. 2020)

Proximal latent points assumption

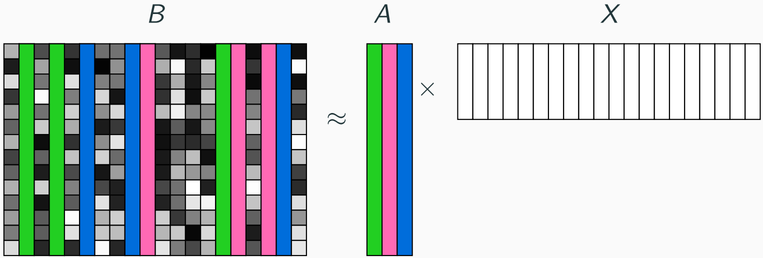
There exists r index sets, $\mathcal{K}_k$ for $k = 1, 2, \ldots, r$, of cardinality at least $p = \delta n$ such that

$$\|AX(:, j) - A(:, k)\|_2 \leq \frac{4\sigma}{\delta} \text{ for all } j \in \mathcal{K}_k,$$

for some $\delta \in \left[\frac{1}{n}, \frac{1}{r}\right]$ and $\sigma > 0$

Model 2: Proximal latent points (Bhattacharyya et al. 2020)

Interpretation: Each vertex has at least p data points close to it.

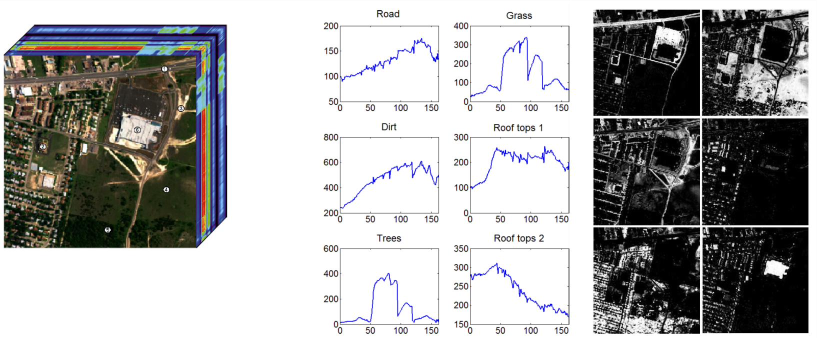


Model 2: Proximal latent points (Bhattacharyya et al. 2020)

- Assumption is stronger than separability, but it allows more noise, and is realistic in practice.
- The proposed Algorithm to Learn a Latent Simplex (ALLS) has practical issues.

Proximal latent points in hyperspectral unmixing

$$\underbrace{B(:,j)}_{\substack{\text{spectral signature of}\\ \text{j-th pixel}}} \approx \sum_{p} \underbrace{A(:,p)}_{\substack{\text{spectral signature of}\\ \text{p-th material}}} \underbrace{X(p,j)}_{\substack{\text{abundance of p-th material}\\ \text{in j-th pixel}}}$$



Images from J. Bioucas Dias and N. Gillis.

ALLS in a nutshell

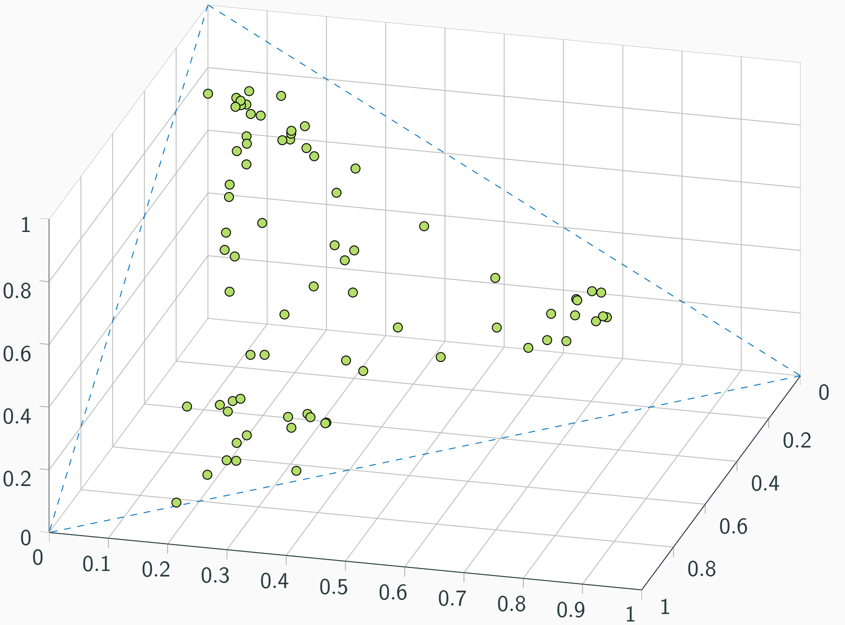
- Similar to VCA
- Averages p data points instead of selecting one

Advantage:

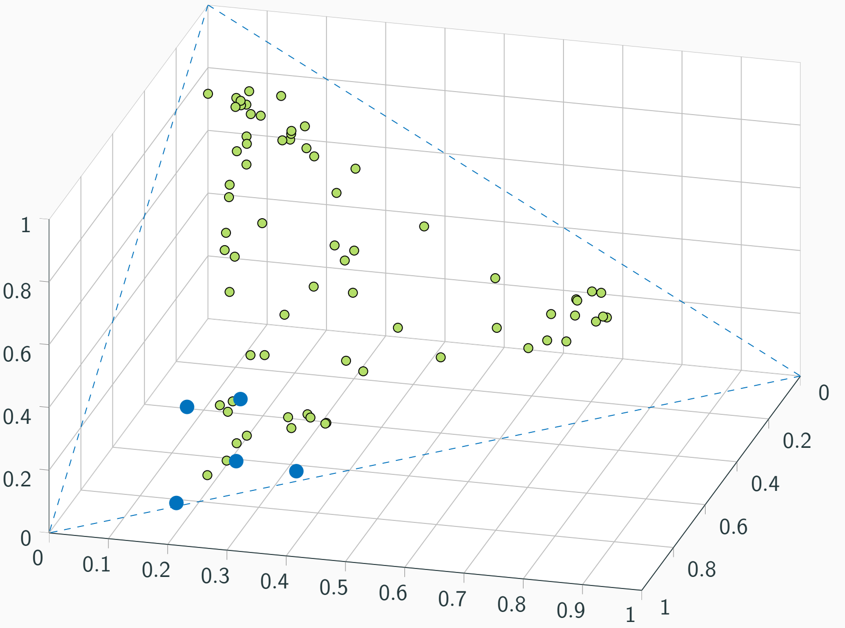
- Probabilistic robustness to noise, depending on spectral norm of N
- More robust to outliers

Conceptually, ALLS is equivalent to applying VCA on the smoothed data set consisting of the $\binom{n}{p}$ points which are the averages of all possible combinations of p data points of B

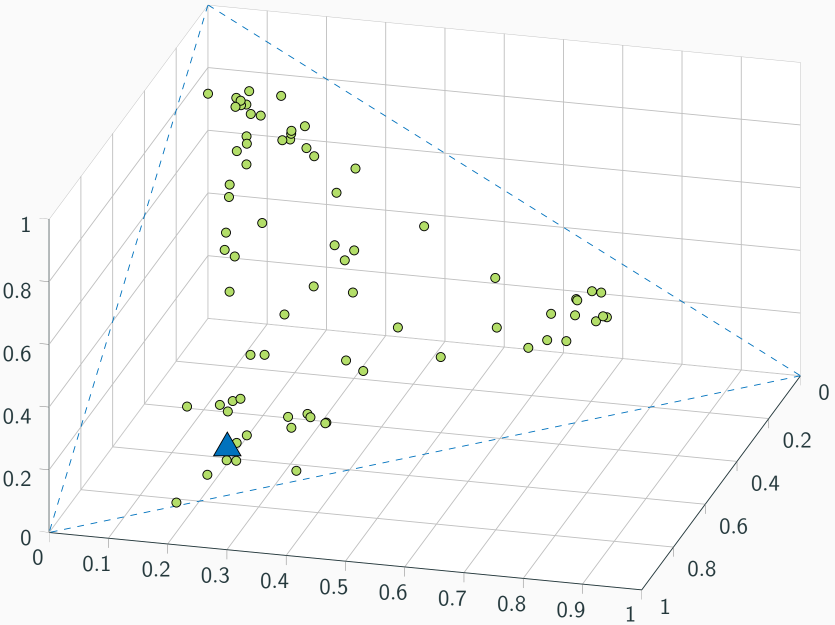
ALLS — Animation ($p = 5$)



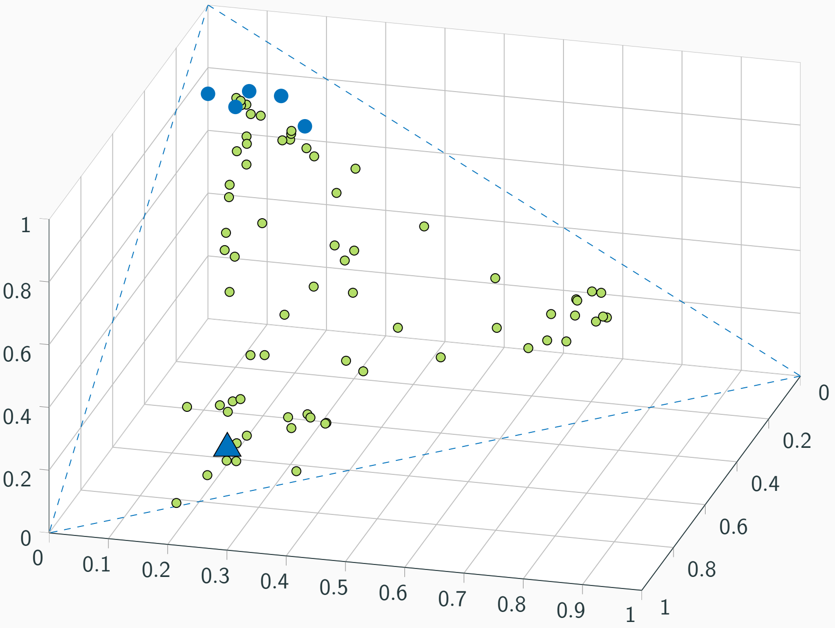
ALLS — Animation ($p = 5$)



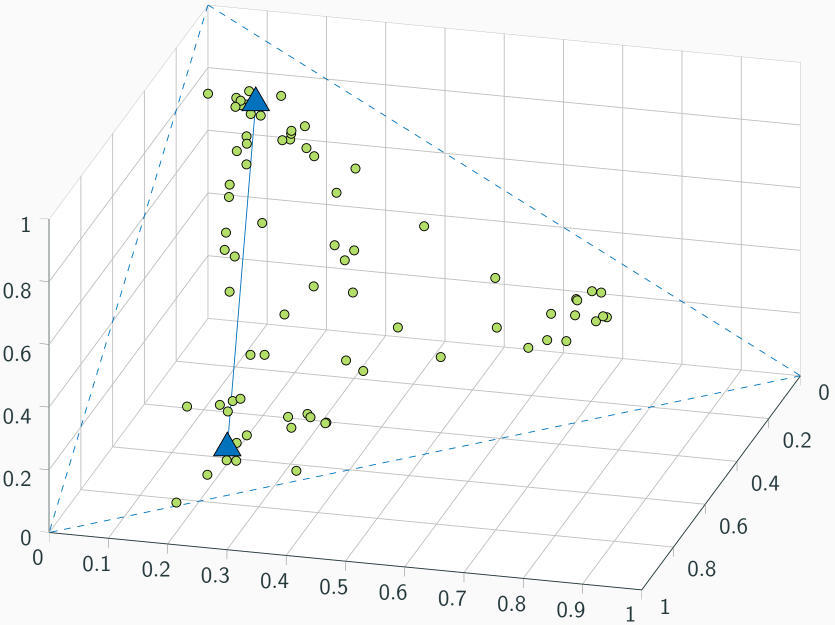
ALLS — Animation ($p = 5$)



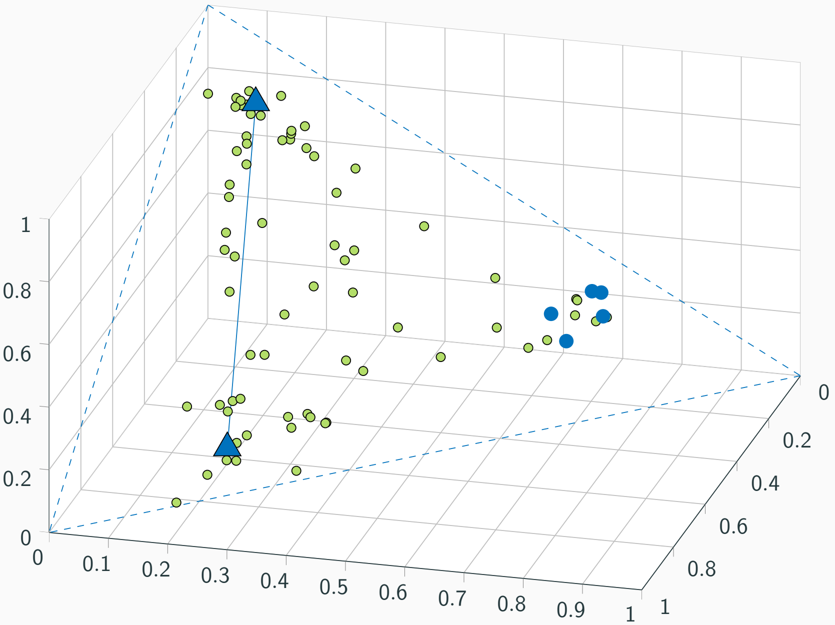
ALLS — Animation ($p = 5$)



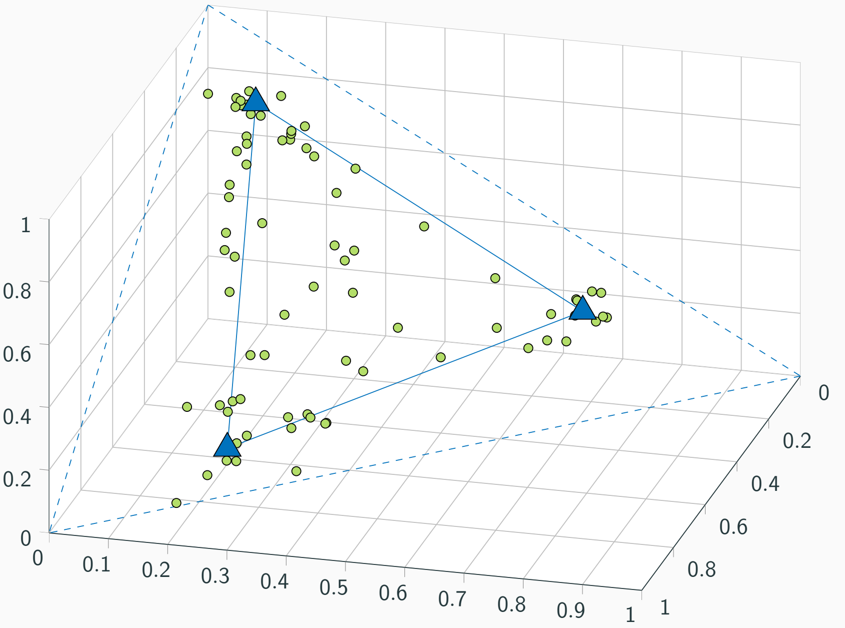
ALLS — Animation ($p = 5$)



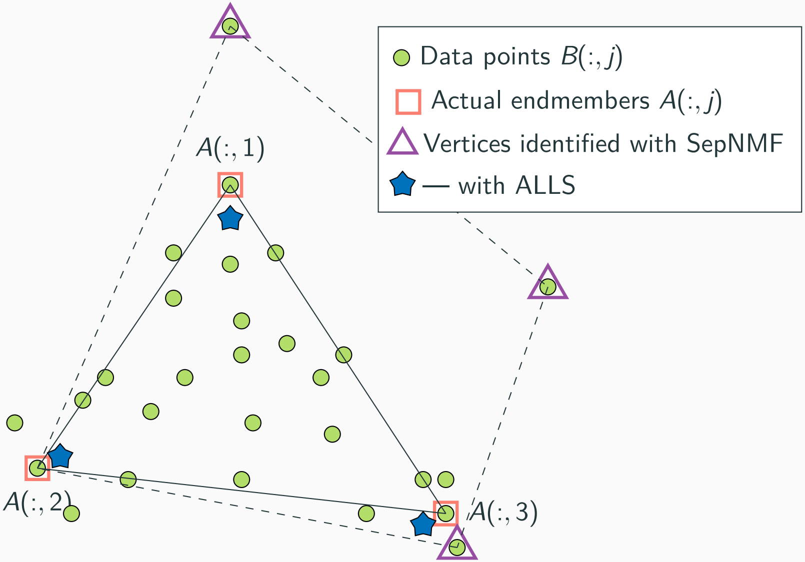
ALLS — Animation ($p = 5$)



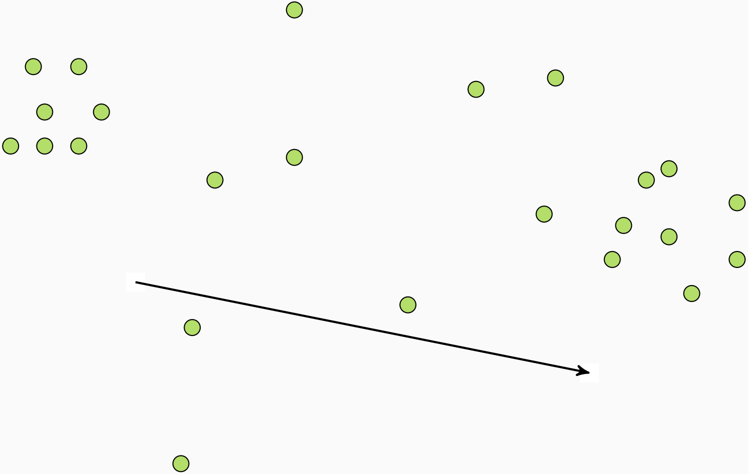
ALLS — Animation ($p = 5$)



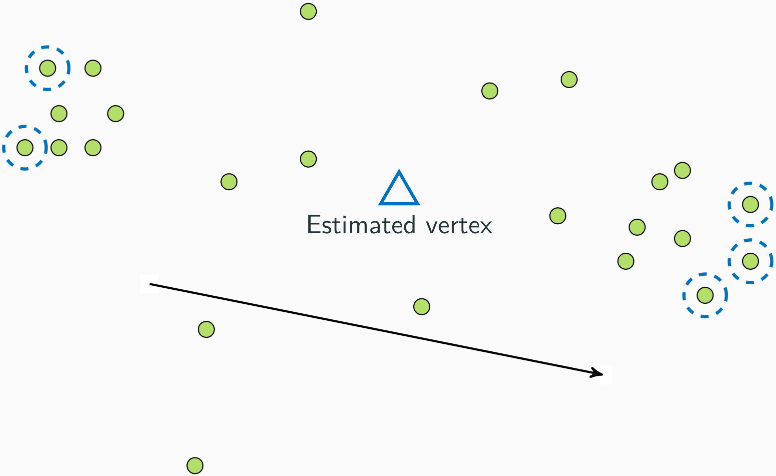
Advantage of the proximal latent points assumption



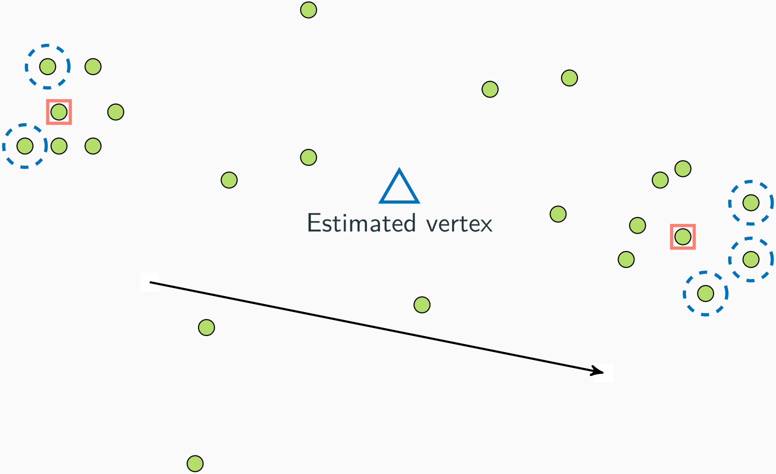
Issue of ALLS 1/2: absolute value (ex with $p = 5$)



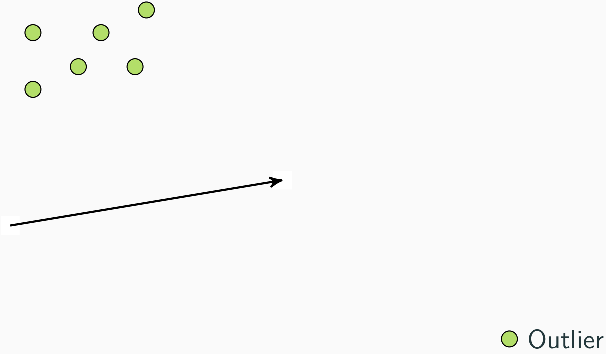
Issue of ALLS 1/2: absolute value (ex with $p = 5$)



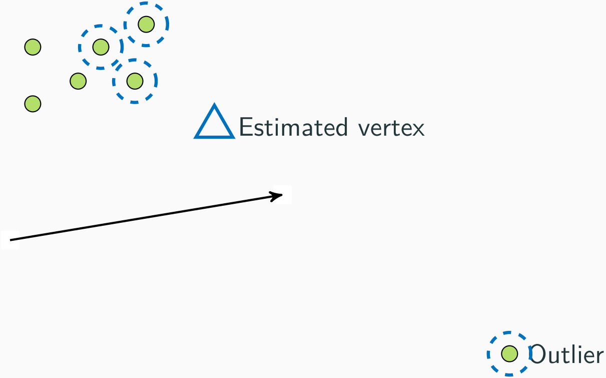
Issue of ALLS 1/2: absolute value (ex with $p = 5$)



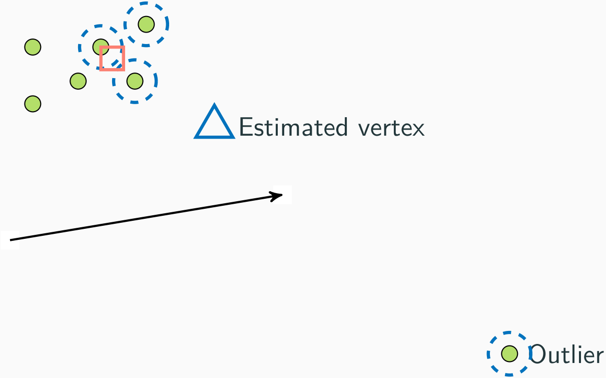
Issue of ALLS 2/2: mean aggregation (ex with $p = 4$)



Issue of ALLS 2/2: mean aggregation (ex with $p = 4$)



Issue of ALLS 2/2: mean aggregation (ex with $p = 4$)



Our contribution

- Smoothed variants of algorithms VCA and SPA that leverage the proximal latent points assumption $\Rightarrow$ SVCA and SSPA
- Aggregates p data points to find each vertex
- Best of both worlds
- Empirically better than VCA, SPA, and ALLS

Smoothed VCA (SVCA)

The best of both worlds!

Similar to ALLS, but:

- Instead of selecting the p entries maximizing the absolute value of u_k, we take the p indices maximizing (resp. minimizing) u_k if the median of the p largest values of u_k is larger (resp. smaller) than the absolute value of the median of the p smallest values of u_k.
- Instead of the mean, we use the median to aggregate points

Robustness results of ALLS apply to SVCA!

Smoothed SPA (SSPA)

Similar to SVCA, but we replace the random direction in the selection step by the column of the residual $P^{\perp}B$ with maximum ℓ_2-norm

Provably robust for $p = 1$ (SPA), we don't know for $p > 1$

Experiment: unmixing of hyperspectral image Urban

SPA

Smoothed SPA

More experiments

See preprint :)

Conclusion

- New assumption is stronger, but often true in real-world datasets
- Empirically, smoothed algorithms perform better than VCA, SPA, and ALLS
- More robust to outliers and noise
- Good way to handle spectral variability?

Future research

Algorithm:

- Strategy to find the best p automatically
- Different p for every endmember
- Other aggregation methods

Theory:

- Identifiability and uniqueness of solution
- Robustness to noise, recovery guarantees

References i

- Vavasis, Stephen A. (2010). "On the Complexity of Nonnegative Matrix Factorization". In: *SIAM Journal on Optimization* 20.3, pp. 1364–1377.
- Arora, Sanjeev, Rong Ge, Ravindran Kannan, and Ankur Moitra (2012). "Computing a Nonnegative Matrix Factorization – Provably". In: *Proceedings of the Forty-Fourth Annual ACM Symposium on Theory of Computing*. Association for Computing Machinery, pp. 145–162.
- NN, Arnaud Vandaele, Jeremy E Cohen, and Nicolas Gillis (2020). "Sparse separable nonnegative matrix factorization". In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECMLPKDD)*, pp. 335–350.
- Gillis, Nicolas (2014). "Successive Nonnegative Projection Algorithm for Robust Nonnegative Blind Source Separation". In: *SIAM Journal on Imaging Sciences* 7.2, pp. 1420–1450.

- NN, Arnaud Vandaele, Nicolas Gillis, and Jeremy E Cohen (2020). "Exact sparse nonnegative least squares". In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5395–5399.
- NN, Nicolas Gillis, and Christophe Kervazo (2021). "Smoothed separable nonnegative matrix factorization". In: *preprint arXiv:2110.05528*.
- Nascimento, José MP and José M Bioucas-Dias (2005). "Vertex component analysis: A fast algorithm to unmix hyperspectral data". In: *IEEE Transactions on Geoscience and Remote Sensing* 43.4, pp. 898–910.
- Araújo, U.M.C., B.T.C. Saldanha, R.K.H Galvão, T. Yoneyama, H.C. Chame, and V. Visani (2001). "The successive projections algorithm for variable selection in spectroscopic multicomponent analysis". In: *Chemometrics and Intelligent Laboratory Systems* 57.2, pp. 65–73.

Bhattacharyya, Chiranjib and Ravindran Kannan (2020). "Finding a latent k–simplex in $O^*(k \cdot \text{nnz}(\text{data}))$ time via subset smoothing". In: *Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms*. SIAM, pp. 122–140.

Thanks!

Contact: `nicolas.nadisic@ugent.be`

Website: `http://nicolasnadisic.xyz`