Journal of the American Statistical Association

Taylor & Francis
Taylor & Francis Group

Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: https://amstat.tandfonline.com/loi/uasa20

Nonnegative Matrix Factorization Via Archetypal Analysis

Hamid Javadi & Andrea Montanari

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2019, VOL. 00, NO. 0, 1–12: Theory and Methods
https://doi.org/10.1080/10.1080/01621459.2019.1594832

Taylor & Francis
Taylor & Francis Group

Nonnegative Matrix Factorization Via Archetypal Analysis

Hamid Javadi[a] and Andrea Montanari[b]

[a]Department of Electrical and Computer Engineering, Rice University, Houston, TX; [b]Department of Electrical Engineering and Statistics, Stanford University, Stanford, CA

ABSTRACT
Given a collection of data points, nonnegative matrix factorization (NMF) suggests expressing them as convex combinations of a small set of "archetypes" with nonnegative entries. This decomposition is unique only if the true archetypes are nonnegative and sufficiently sparse (or the weights are sufficiently sparse), a regime that is captured by the separability condition and its generalizations.
In this article, we study an approach to NMF that can be traced back to the work of Cutler and Breiman [(1994), "Archetypal Analysis," *Technometrics*, 36, 338–347] and does not require the data to be separable, while providing a generally unique decomposition. We optimize a trade-off between two objectives: we minimize the distance of the data points from the convex envelope of the archetypes (which can be interpreted as an empirical risk), while also minimizing the distance of the archetypes from the convex envelope of the data (which can be interpreted as a data-dependent regularization). The archetypal analysis method of Cutler and Breiman is recovered as the limiting case in which the last term is given infinite weight. We introduce a "uniqueness condition" on the data which is necessary for identifiability. We prove that, under uniqueness (plus additional regularity conditions on the geometry of the archetypes), our estimator is robust. While our approach requires solving a nonconvex optimization problem, we find that standard optimization methods succeed in finding good solutions for both real and synthetic data. Supplementary materials for this article are available online

ARTICLE HISTORY
Received June 2017
Accepted January 2019

KEYWORDS
Dimensionality reduction;
Matrix factorization;
Separability.

1. Introduction

Unmixing convex combinations of a small number of signals—without knowing a priori the pure components—is a central statistical problem in a broad range of applications, from chemometrics (Paatero and Tapper 1994) to image processing (Lee and Seung 1999), topic modeling (Xu, Liu, and Gong 2003), clustering, and co-clustering (Long, Zhang, and Yu 2005; Wang et al. 2011; and Del Buono and Pio 2015). As an example, Figure 1 displays the infrared reflection spectra[1] of four molecules (caffeine, sucrose, lactose and trioctanoin) for wavenumbers between 1186 and 1530 cm^{-1} (data were retrieved from the NIST ChemistryWebBook dataset (Linstrom and Mallard 2017). Each spectrum is a vector $\boldsymbol{h}_{0,\ell} \in \mathbb{R}^d$, in $d = 87$ dimensions, whose components $((\boldsymbol{h}_{0,\ell})_1, \dots (\boldsymbol{h}_{0,\ell})_d)$ contain the reflection intensity of the ℓth molecule at different wavelengths. The index ℓ refers to the four molecules, whence $\ell \in \{1, \dots, 4\}$. We will refer to the vectors $\boldsymbol{h}_{0,1}, \dots, \boldsymbol{h}_{0,4}$ as to the "archetypes." If a mixture of these substances is analyzed, the resulting spectrum will be a convex combination of the four spectra $\boldsymbol{h}_{0,1}, \dots, \boldsymbol{h}_{0,4}$. The same situation arises in hyperspectral imaging (Ma et al. 2014), where the objective is to estimate the proportions of a certain number of analytes which depend on the spatial position in the image.

In order to mimic this setting, we generated $n = 250$, synthetic random convex combinations of the four spectra $\boldsymbol{h}_{0,1}$, …, $\boldsymbol{h}_{0,4}$, which we denote by $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n \in \mathbb{R}^d$. Each synthetic combination contains two or more of these four analytes. We then tried to reconstruct the archetype spectra from the $\boldsymbol{x}_i$'s.

Figure 1 displays the outcome of such reconstruction, whereby each column corresponds to a different reconstruction algorithm. We refer to Appendix A for further details and to Appendix F for comparison with seven other algorithms in the literature.

As illustrated by this example, this blind unmixing problem is often challenging, and existing approaches can be inaccurate. A key difficulty lies in the fact that—without further constraints—the problem is dramatically underdetermined. Given a set of valid archetypes $\{\boldsymbol{h}_{0,\ell}\}_{\ell \le r}$, any set $\{\boldsymbol{h}_\ell\}_{\ell \le r}$ whose convex hull contains the $\{\boldsymbol{h}_{0,\ell}\}_{\ell \le r}$ is also a solution of the problem. For instance, we can set $\boldsymbol{h}_\ell = \boldsymbol{h}_{0,\ell}$ for $\ell \le r-1$, and $\boldsymbol{h}_r = (1+s)\boldsymbol{h}_{0,r} - s\boldsymbol{h}_{0,1}$ for any $s \ge 0$, and obtain an equally good representation of the data.

Mathematically, we are given a set of data points $\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_n \in \mathbb{R}^d$, and want to represent them as convex combinations of a small set of vectors (the "archetypes" $\boldsymbol{h}_1, \dots, \boldsymbol{h}_\ell$):

$$\boldsymbol{x}_i \approx \sum_{\ell=1}^{r} w_{i,\ell} \boldsymbol{h}_\ell\,, \qquad w_{i,\ell} \ge 0, \quad \sum_{\ell=1}^{r} w_{i,\ell} = 1\,. \tag{1}$$

Figure 2 illustrates the geometry of this problem for a small synthetic example, with $d = 2$, $r = 3$, $n = 500$. Once again, any

CONTACT Hamid Javadi hrhakim@rice.edu Department of Electrical and Computer Engineering, Rice University, Houston, TX 77005; Andrea Montanari montanari@stanford.edu Department of Electrical Engineering and Statistics, Stanford University, Stanford, CA 94305.

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/r/JASA.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

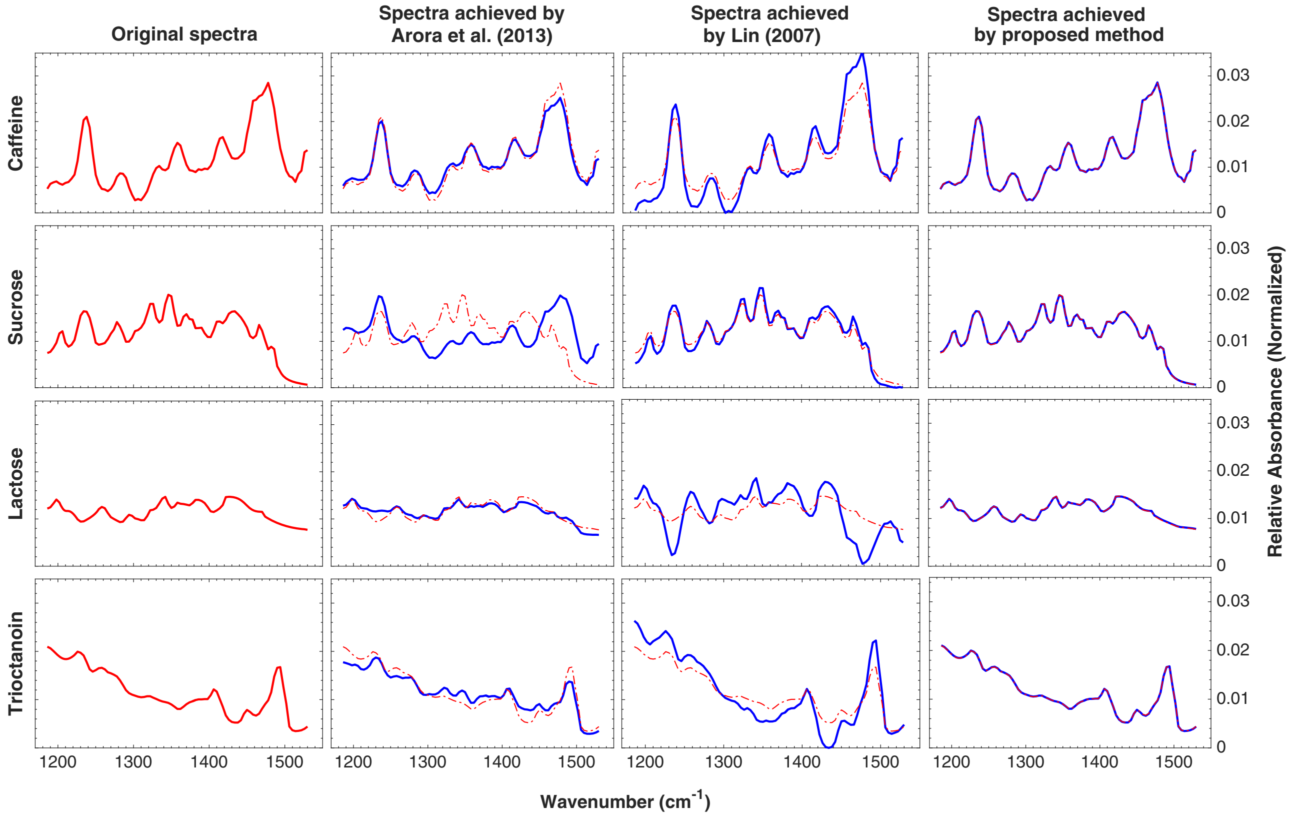


Figure 1. Left column: Infrared reflection spectra of four molecules. Subsequent columns: Spectra estimated from $n = 250$ spectra of mixtures of the four original substances (synthetic data generated by taking random convex combinations of the pure spectra, see Appendix A for details). Each column reports the results obtained with a different estimator: continuous blue lines correspond to the reconstructed spectra; dashed red lines correspond to the ground truth.

set of three vectors whose convex hull contains the data points is a valid solution of the problem. The central question is therefore: *How should we constrain the decomposition (1) in such a way that it is generally unique (up to permutations of the r archetypes)?*

Since the seminal work of Paatero and Tapper (1994), Paatero (1997), and of Lee and Seung (1999, 2001), a large amount of work has addressed this question by making the assumption that the archetypes are componentwise nonnegative $\boldsymbol{h}_\ell \geq 0$. Among other applications, the nonnegativity constraint is justified in chemometrics (reflection or absorption spectra are nonnegative), and topic modeling (in this case archetypes correspond to topics, which are represented as probability distributions over words). This formulation has become popular as nonnegative matrix factorization (NMF).

Under the nonnegativity constraint $\boldsymbol{h}_\ell \geq 0$ the roles of weights and archetypes become interchangeable. This is easily seen in matrix notation. We represent the data as a matrix $\boldsymbol{X} \in \mathbb{R}^{n \times d}$ whose ith row is the vector $\boldsymbol{x}_i$, the weights by a matrix $\boldsymbol{W} = (w_{i,\ell})_{i \leq n, \ell \leq r} \in \mathbb{R}^{n \times r}$ and the archetypes by a matrix $\boldsymbol{H} = (h_{\ell,j})_{\ell \leq r, j \leq d} \in \mathbb{R}^{r \times d}$. When $\boldsymbol{h}_\ell \geq 0$, it is well known that—without loss of generality—we can assume $\sum_{\ell=1}^{r} h_{\ell,i} = 1$. The factorization (1) takes therefore the form $\boldsymbol{X} = \boldsymbol{W}\boldsymbol{H}$, and we can equivalently consider $\boldsymbol{W}$ as weights and $\boldsymbol{H}$ as archetypes or vice versa (by transposing $\boldsymbol{X}$).

The decomposition (1) is unique provided that the archetypes or the weights are sufficiently sparse. This point was clarified by Donoho and Stodden (2003), who introduced a separability condition that implies uniqueness. The nonnegative archetypes $\boldsymbol{h}_1, \cdots, \boldsymbol{h}_r$ are separable if, for each $\ell \in [r]$ there exists an index $i(\ell) \in [d]$ such that $(\boldsymbol{h}_\ell)_{i(\ell)} = 1$, and $(\boldsymbol{h}_{\ell'})_{i(\ell)} = 0$ for all $\ell' \neq \ell$. Due to the interchangeability of weights and archetypes discussed above, we can consider an equivalent "weight separability" condition. This requires that for $\ell \in [r]$ there exists an index $i(\ell) \in [n]$ such that $w_{i(\ell),\ell} = 1$, and $w_{i(\ell),\ell'} = 0$ for all $\ell' \neq \ell$. Since "archetype separability" and "weight separability" are mathematically equivalent we shall focus hereafter on the latter (with the caveat that the two are connected by a transposition of $\boldsymbol{X}$).

Weight separability has a simple geometric interpretation: the data are separable if for each archetype $\boldsymbol{h}_\ell$ there is at least one data point $\boldsymbol{x}_i$ such that $\boldsymbol{x}_i = \boldsymbol{h}_\ell$. A copious literature has developed algorithms for nonnegative matrix factorization under separability condition or its generalizations (Donoho and Stodden 2003; Arora et al. 2012; Recht et al. 2012; Arora et al. 2013; Ge and Zou 2015).

Of course, this line of work has a drawback: in practice we do not know whether the data are separable. (We refer to Section 5 for a comparison with Ge and Zou (2015), which relaxes the separability assumption.) Furthermore, there are many cases in which the archetypes $\boldsymbol{h}_1, \ldots, \boldsymbol{h}_r$ are not necessarily nonnegative. For instance, in spike sorting, the data $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ are electrophysiological measurements of neural activity at n positions, and the archetypes $\boldsymbol{h}_1, \ldots, \boldsymbol{h}_r$ correspond to waveforms associated to different neurons (Roux, Cheveigné, and Parra 2009). In very high-dimensional applications, the archetypes $\boldsymbol{h}_\ell$ are nonnegative, but—in order to reduce complexity—the data $\{\boldsymbol{x}_i\}_{i \leq n}$ are replaced by a random low-dimensional projection (Kim, Sra, and Dhillon 2008; Wang and Li 2010). The projected

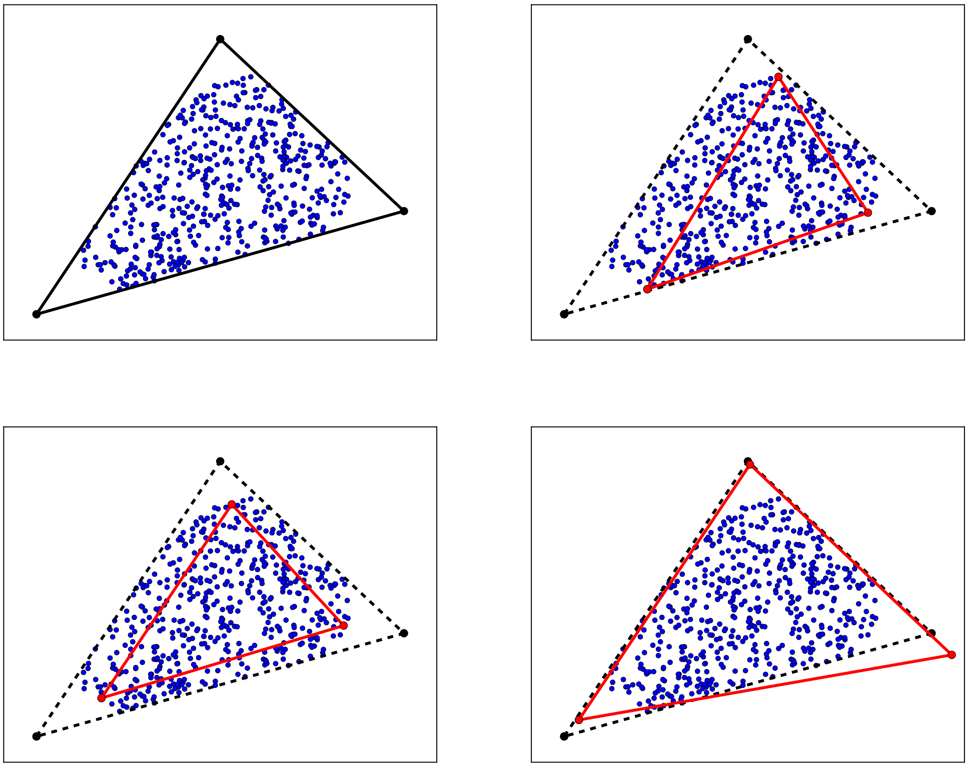


Figure 2. Toy example of archetype reconstruction. Top left: data points (blue) are generated as random linear combinations of $r = 3$ archetypes in $d = 2$ dimensions (black, see Appendix A for details). Top right (red): Initialization using the algorithm of Arora et al. (2013). Bottom left (red): Output of the alternate minimization algorithm of Cutler and Breiman (1994) with initialization form the previous frame. Bottom right (red): Alternate minimization algorithm to compute the estimator (6), with $\lambda = 0.0166$.

archetypes lose the nonnegativity property. Finally, the decomposition (1) is generally nonunique, even under the constraint $\boldsymbol{h}_\ell \geq 0$. This is illustrated, again, in Figure 1: all the spectra are strictly positive, and hence we can find archetypes $\boldsymbol{h}_1, \dots, \boldsymbol{h}_4$ that are still nonnegative and whose convex envelope contains $\boldsymbol{h}_{0,1}, \dots, \boldsymbol{h}_{0,4}$.

Since NMF is generally underdetermined, standard methods fail in such applications, as illustrated in Figure 1. The third column of Figure 1 uses a projected gradient algorithm from Lin (2007) to solve the problem

$$\begin{aligned} &\text{minimize} \quad && \|\boldsymbol{X} - \boldsymbol{W}\boldsymbol{H}\|_F^2\,, \\ &\text{subject to} \quad && \boldsymbol{W} \geq 0, \quad \boldsymbol{H} \geq 0\,. \end{aligned} \tag{2}$$

Empirically, projected gradient converges to a point with very small fitting error $\|\boldsymbol{X} - \boldsymbol{W}\boldsymbol{H}\|_F^2$, but the reconstructed spectra (rows of $\boldsymbol{H}$) are inaccurate. The second column in the same figure shows the spectra reconstructed using an algorithm from Arora et al. (2013), which assumes separability: as expected, the reconstruction is not accurate.

Several earlier works addressed the nonuniqueness problem in classical nonnegative matrix factorization. Among others, Miao and Qi (2007) penalize a matrix of archetypes $\boldsymbol{H}$ by the corresponding volume and try to find the set of archetypes $\boldsymbol{h}_1, \dots, \boldsymbol{h}_\ell$ whose convex hull contains data points and has the least volume. To the best of our knowledge, none of these works establishes robustness of the proposed methods.

In a less widely known paper, Cutler and Breiman (1994) addressed the same problem using what they call "archetypal analysis." Archetypal analysis presents two important differences with respect to standard NMF: (1) The archetypes $\boldsymbol{h}_\ell$ are not necessarily required to be nonnegative (although this constraint can be easily incorporated); (2) The underdetermination of the decomposition (1) is addressed by requiring that the archetypes belong to the convex hull of the data points: $\boldsymbol{h}_\ell \in \text{conv}(\{\boldsymbol{x}_i\}_{i \leq n})$. These ideas were further developed in the work of Mørup and Hansen (2012), which however does not provide theoretical analysis in the presence of noise.

In applications, the condition $\boldsymbol{h}_\ell \in \text{conv}(\{\boldsymbol{x}_i\}_{i \leq n})$ enforced by Cutler and Breiman (1994) is too strict. This article builds on the ideas of Cutler and Breiman (1994) to propose a formulation of NMF that is uniquely defined (barring degenerate cases) and provides a useful notion of optimality. In particular, we present the following contributions.

Archetypal reconstruction. We propose to reconstruct the archetypes $\boldsymbol{h}_1, \dots, \boldsymbol{h}_r$ by optimizing a combination of two objectives. On the one hand, we minimize the error in the decomposition (1). This amounts to minimizing the distance between the data points and the convex hull of the archetypes. On the other hand, we minimize the distance of the archetypes from the convex hull of data points. This relaxes the original condition imposed in Cutler and Breiman (1994) which required the archetypes to lie in $\text{conv}(\{\boldsymbol{x}_i\})$, and allows to treat nonseparable data.

Robustness guarantee. We next assume that the decomposition (1) approximately holds for some "true" archetypes $\boldsymbol{h}_\ell^0$ and weights $w_{i,\ell}^0$, namely $\boldsymbol{x}_i = \boldsymbol{x}_i^0 + \boldsymbol{z}_i$, where $\boldsymbol{x}_i^0 = \sum_{\ell=1}^r w_{i,\ell}^0 \boldsymbol{h}_\ell^0$ and $\boldsymbol{z}_i$ captures unexplained effects. We introduce a "uniqueness condition" on the data $\{\boldsymbol{x}_i^0\}_{i \leq n}$ which is necessary for exactly recovering the archetypes from the noiseless data. We prove that, under uniqueness (plus additional regularity conditions on the geometry of the archetypes), our estimator is robust. Namely it outputs archetypes $\{\hat{\boldsymbol{h}}_\ell\}_{\ell\le r}$ whose distance from the true ones $\{\boldsymbol{h}^0_\ell\}_{\ell\le r}$ (in a suitable metric) is controlled by $\sup_{i\le n}\|\boldsymbol{z}_i\|_2$.

Algorithms. Our approach reconstructs the archetypes $\boldsymbol{h}_1,\dots,\boldsymbol{h}_r$ by minimizing a nonconvex risk function $\mathscr{R}_\lambda(\boldsymbol{H})$. We propose three descent algorithms that appear to perform well on realistic instances of the problem. In particular, Section 4 introduces a proximal alternating linearized minimization algorithm (PALM) that is guaranteed to converge to critical points of the risk function. Appendix E discusses two alternative approaches. One possible explanation for the success of such descent algorithms is that reasonably good initializations can be constructed using spectral methods, or approximating the data as separable, cf. Section 4.1. We defer a study of global convergence of this two-stage approach to future work.

2. An Archetypal Reconstruction Approach

Let $\mathcal{Q}\subseteq\mathbb{R}^d$ be a convex set and $D:\mathcal{Q}\times\mathcal{Q}\to\mathbb{R}$, $(\boldsymbol{x},\boldsymbol{y})\mapsto D(\boldsymbol{x};\boldsymbol{y})$ a loss function on $\mathcal{Q}$. For a point $\boldsymbol{u}\in\mathcal{Q}$ and a matrix $\boldsymbol{V}\in\mathbb{R}^{m\times d}$, with rows $\boldsymbol{v}_1,\dots,\boldsymbol{v}_m\in\mathcal{Q}$, we let

$$\mathscr{D}(\boldsymbol{u};\boldsymbol{V})\equiv\min\Big\{D\big(\boldsymbol{u};\boldsymbol{V}^{\mathsf{T}}\boldsymbol{\pi}\big)\;:\quad\boldsymbol{\pi}\in\Delta^m\Big\},\tag{3}$$

$$\Delta^m\equiv\Big\{\boldsymbol{x}\in\mathbb{R}^m_{\ge 0}:\;\;\langle\boldsymbol{x},\boldsymbol{1}\rangle=1\Big\}.\tag{4}$$

In other words, denoting by $\mathrm{conv}(\boldsymbol{V})=\mathrm{conv}(\{\boldsymbol{v}_1,\dots,\boldsymbol{v}_m\})$ the convex hull of the rows of matrix $\boldsymbol{V}$, $\mathscr{D}(\boldsymbol{u};\boldsymbol{V})$ is the minimum loss between $\boldsymbol{x}$ and any point in $\mathrm{conv}(\boldsymbol{V})$. If $\boldsymbol{U}\in\mathbb{R}^{k\times d}$ is a matrix with rows $\boldsymbol{u}_1,\dots,\boldsymbol{u}_k\in\mathcal{Q}$, we generalize this definition by letting

$$\mathscr{D}(\boldsymbol{U};\boldsymbol{V})\equiv\sum_{\ell=1}^{k}\mathscr{D}(\boldsymbol{u}_\ell;\boldsymbol{V}).\tag{5}$$

While this definition makes sense more generally, we have in mind two specific examples in which $D(\boldsymbol{x};\boldsymbol{y})$ is actually separately convex in its arguments $\boldsymbol{x}$ and $\boldsymbol{y}$. (Most of our results will concern the first example.)

Example 2.1 (Square loss). In this case $\mathcal{Q}=\mathbb{R}^d$, and $D(\boldsymbol{x};\boldsymbol{y})=\|\boldsymbol{x}-\boldsymbol{y}\|_2^2$. This is the case originally studied by Cutler and Breiman Cutler and Breiman (1994).

Example 2.2 (KL divergence). We take $\mathcal{Q}=\Delta^d$, the d-dimensional simplex, and $D(\boldsymbol{x};\boldsymbol{y})$ to be the Kullback–Leibler divergence between probability distributions $\boldsymbol{x}$ and $\boldsymbol{y}$, namely $D(\boldsymbol{x};\boldsymbol{y})\equiv\sum_{i=1}^d x_i\log(x_i/y_i)$.

Given data $\boldsymbol{x}_1,\dots,\boldsymbol{x}_n$ organized in the matrix $\boldsymbol{X}\in\mathbb{R}^{n\times d}$, we estimate the archetypes by solving the problem

$$\widehat{\boldsymbol{H}}_\lambda\in\arg\min\Big\{\mathscr{D}(\boldsymbol{X};\boldsymbol{H})+\lambda\,\mathscr{D}(\boldsymbol{H};\boldsymbol{X}):\;\;\boldsymbol{H}\in\mathcal{Q}^r\Big\},\tag{6}$$

where we denote by $\mathcal{Q}^r$ the set of matrices $\boldsymbol{H}\in\mathbb{R}^{r\times d}$ with rows $\boldsymbol{h}_1,\dots,\boldsymbol{h}_r\in\mathcal{Q}$ (This problem can have multiple global minima if $\lambda=0$ or in degenerate settings. One minimizer is selected arbitrarily when this happens). A few values of λ are of special significance. If we set $\lambda=0$ and $\mathcal{Q}=\Delta^d$, we recover the standard NMF objective (2), with a more general distance function $D(\,\cdot\,,\,\cdot\,)$. As pointed out above, in general this optimization problem has no unique minimizer. If we let $\lambda\to 0+$ after the minimum is evaluated, $\widehat{\boldsymbol{H}}_\lambda$ converges to the minimizer of $\mathscr{D}(\boldsymbol{X};\boldsymbol{H})$ which is the “closest” to the convex envelope of the data $\mathrm{conv}(\boldsymbol{X})$ (in the sense of minimizing $\mathscr{D}(\boldsymbol{H};\boldsymbol{X})$). Finally for $\lambda\to\infty$, the archetypes $\boldsymbol{h}_\ell$ are forced to lie in $\mathrm{conv}(\boldsymbol{X})$ and hence we recover the method of Cutler and Breiman (1994).

Figure 2 illustrates the advantages of the estimator (6) on a small synthetic example, with $d=2$, $r=3$, and $n=500$. In this case, the data are not separable as it can be seen from the fact that no data points coincide with the extremal points of $\mathrm{conv}(\boldsymbol{H}_0)$. We first use the successive projections algorithm of Arora et al. (2013) (that is designed to deal with separable data) in order to estimate the archetypes. As expected, the reconstruction is not accurate because this algorithm assumes separability and hence estimates the archetypes by a subset of the data points. We then use these estimates as initialization in the alternate minimization algorithm of Cutler and Breiman (1994), which optimizes the objective (6) with $\lambda=\infty$. The estimates improve but not substantially: they are still constrained to belong to $\mathrm{conv}(\boldsymbol{X})$. A significant improvement is obtained by setting λ to a small value. We (approximately) minimize the cost function (6) by generalizing the alternate minimization algorithm, cf. Section 4. The optimal archetypes are no longer constrained to $\mathrm{conv}(\boldsymbol{X})$, and provide a better estimate of the true archetypes. In the last column of Figure 1, we use the same estimator, and approximately solve problem (6) by gradient descent.

In our analysis, we will consider a slightly different formulation in which the Lagrangian of equation (6) is replaced by a hard constraint:

$$\begin{aligned}&\text{minimize}\quad &&\mathscr{D}(\boldsymbol{H};\boldsymbol{X}),\\ &\text{subject to}\quad &&\mathscr{D}(\boldsymbol{x}_i;\boldsymbol{H})\le\delta^2\quad\text{for all } i\in\{1,\dots,n\}.\end{aligned}\tag{7}$$

We will use this version in the analysis presented in the next section, and denote the corresponding estimator by $\widehat{\boldsymbol{H}}$.

3. Robustness

In order to analyze the robustness properties of estimator $\widehat{\boldsymbol{H}}$, we assume that there exists an approximate factorization

$$\boldsymbol{X}=\boldsymbol{W}_0\boldsymbol{H}_0+\boldsymbol{Z},\tag{8}$$

where $\boldsymbol{W}_0\in\mathbb{R}^{n\times r}$ is a matrix of weights (with rows $\boldsymbol{w}_{0,i}\in\Delta^r$), $\boldsymbol{H}_0\in\mathbb{R}^{r\times d}$ is a matrix of archetypes (with rows $\boldsymbol{h}_{0,\ell}$), and we define $\boldsymbol{X}_0=\boldsymbol{W}_0\boldsymbol{H}_0$. The error term $\boldsymbol{Z}$ is arbitrary, with rows $\boldsymbol{z}_i$ satisfying $\max_{i\le n}\|\boldsymbol{z}_i\|_2\le\delta$. We will assume throughout r to be known.

We will quantify estimation error by the sum of distances between the true archetypes and the closest estimated archetypes

$$\mathscr{L}(\boldsymbol{H}_0,\widehat{\boldsymbol{H}})\equiv\sum_{\ell=1}^{r}\min_{\ell'\le r}D(\boldsymbol{h}_{0,\ell},\hat{\boldsymbol{h}}_{\ell'}).\tag{9}$$

In words, if $\mathscr{L}(\boldsymbol{H}_0,\widehat{\boldsymbol{H}})$ is small, then for each true archetype $\boldsymbol{h}_{0,\ell}$ there exists an estimated archetype $\hat{\boldsymbol{h}}_{\ell'}$ that is close to it in D-loss. Unless two or more of the true archetypes are close to each other, this means that there is a one-to-one correspondence between estimated archetypes and true archetypes, with small errors.

Assumption (Uniqueness). We say that the factorization $\boldsymbol{X}_0 = \boldsymbol{W}_0\boldsymbol{H}_0$ satisfies uniqueness with parameter $\alpha > 0$ (equivalently, is α-*unique*) if for all $\boldsymbol{H} \in \mathcal{Q}^r$ with $\mathrm{conv}(\boldsymbol{X}_0) \subseteq \mathrm{conv}(\boldsymbol{H})$, we have

$$\mathscr{D}(\boldsymbol{H},\boldsymbol{X}_0)^{1/2} \geq \mathscr{D}(\boldsymbol{H}_0,\boldsymbol{X}_0)^{1/2} + \alpha\left\{\mathscr{D}(\boldsymbol{H},\boldsymbol{H}_0)^{1/2} + \mathscr{D}(\boldsymbol{H}_0,\boldsymbol{H})^{1/2}\right\}. \qquad (10)$$

The rationale for this assumption is quite clear. Assume that the data lie in the convex hull of the true archetypes $\boldsymbol{H}_0$, and hence Equation (8) holds with $\boldsymbol{Z} = 0$, that is, $\boldsymbol{X} = \boldsymbol{X}_0$. We reconstruct the archetypes by demanding $\mathrm{conv}(\boldsymbol{X}_0) \subseteq \mathrm{conv}(\boldsymbol{H})$: any such $\boldsymbol{H}$ is a plausible explanation of the data. In order to make the problem well specified, we define $\boldsymbol{H}_0$ to be the matrix of archetypes that are the closest to $\boldsymbol{X}_0$, and hence $\mathscr{D}(\boldsymbol{H},\boldsymbol{X}_0) \geq \mathscr{D}(\boldsymbol{H}_0,\boldsymbol{X}_0)$ for all $\boldsymbol{H}$. In order for the reconstruction to be unique (and hence for the problem to be identifiable) we need to assume $\mathscr{D}(\boldsymbol{H},\boldsymbol{X}_0) > \mathscr{D}(\boldsymbol{H}_0,\boldsymbol{X}_0)$ strictly for $\boldsymbol{H} \neq \boldsymbol{H}_0$. The uniqueness assumption provides a quantitative version of this condition.

Given $\boldsymbol{X}_0$, $\boldsymbol{H}_0$, the best constant α such that Equation (10) holds for all $\boldsymbol{H}$ such that $\mathrm{conv}(\boldsymbol{X}_0) \subseteq \mathrm{conv}(\boldsymbol{H})$ is the uniqueness constant of $(\boldsymbol{H}_0,\boldsymbol{X}_0)$, denoted by $\alpha(\boldsymbol{H}_0,\boldsymbol{X}_0)$, Notice that this is a geometric property that depends on $\boldsymbol{X}_0$ only through $\mathrm{conv}(\boldsymbol{X}_0)$.

Remark 3.1. If $\boldsymbol{X}_0 = \boldsymbol{W}_0\boldsymbol{H}_0$ is a separable factorization, then it satisfies uniqueness with parameter $\alpha = 1$. Indeed, since each data point $\boldsymbol{x}_{0,i}$ belongs to $\mathrm{conv}(\boldsymbol{H}_0)$, we have $\mathrm{conv}(\boldsymbol{X}_0) \subseteq \mathrm{conv}(\boldsymbol{H}_0)$. In addition, by separability, the rows of $\boldsymbol{H}_0$ are a subset of the rows of $\boldsymbol{X}_0$. Therefore, in this case $\mathrm{conv}(\boldsymbol{H}_0) = \mathrm{conv}(\boldsymbol{X}_0)$, whence for $\boldsymbol{H}$ such that $\mathrm{conv}(\boldsymbol{X}_0) \subseteq \mathrm{conv}(\boldsymbol{H})$, $\mathscr{D}(\boldsymbol{H},\boldsymbol{X}_0) = \mathscr{D}(\boldsymbol{H},\boldsymbol{H}_0)$ and $\mathscr{D}(\boldsymbol{H}_0,\boldsymbol{X}_0) = \mathscr{D}(\boldsymbol{H}_0,\boldsymbol{H}) = 0$.

It is further possible to show that $\alpha \in [0,1]$ for the square loss $D(\boldsymbol{x},\boldsymbol{y}) = \|\boldsymbol{x}-\boldsymbol{y}\|_2^2$ (with $\mathcal{Q} = \mathbb{R}^d$) and all $\boldsymbol{H}_0, \boldsymbol{X}_0$. Indeed, $\alpha \geq 0$ follows simply by defining $\boldsymbol{H}_0$ to be the matrix of archetypes that are closest to $\boldsymbol{X}_0$, whence $\mathscr{D}(\boldsymbol{H},\boldsymbol{X}_0) \geq \mathscr{D}(\boldsymbol{H}_0,\boldsymbol{X}_0)$ and therefore $\alpha \geq 0$.

In order to see that $\alpha \leq 1$, we consider the following construction of $\boldsymbol{H}_1 \in \mathcal{Q}^r$ (see Figure 3). Let $\mathcal{C}_1$ be the cone generated by $\boldsymbol{h}_{0,2} - \boldsymbol{h}_{0,1}, \boldsymbol{h}_{0,3} - \boldsymbol{h}_{0,1}, \ldots, \boldsymbol{h}_{0,r} - \boldsymbol{h}_{0,1}$, that is, $\mathcal{C}_1 = \{\boldsymbol{v} \in \mathbb{R}^d; \boldsymbol{v} = \sum_{i=2}^r u_i(\boldsymbol{h}_{0,i} - \boldsymbol{h}_{0,1}), u_i \geq 0\}$. Let $\mathcal{C}_1^*$ be the dual cone (Recall that given a cone $\mathcal{C} \subseteq \mathbb{R}^d$, its dual is defined as $\mathcal{C}^* = \{\boldsymbol{x} \in \mathbb{R}^d : \langle \boldsymbol{x},\boldsymbol{y}\rangle \geq 0 \;\; \forall \boldsymbol{y} \in \mathcal{C}\}$) and $\widetilde{\boldsymbol{h}}$ be an arbitrary point in $\mathrm{conv}(\boldsymbol{H}_0) \cap \left(\mathcal{C}_1^* + \boldsymbol{h}_{0,1}\right)$ (It is easy to see that $\mathrm{conv}(\boldsymbol{H}_0) \cap \left(\mathcal{C}_1^* + \boldsymbol{h}_{0,1}\right)$ is nonempty. Indeed, if $\mathcal{H}$ is a half-space that contains $\mathcal{C}_1$, and $\boldsymbol{n} \in \mathcal{H}$ the normal to the corresponding hyperplane, it follows from its definition that $\boldsymbol{h}_{0,1} + t\boldsymbol{n}$ is such a point for some $t \geq 0$). Take

$$\boldsymbol{h}_{1,1} = 2\boldsymbol{h}_{0,1} - \widetilde{\boldsymbol{h}}, \qquad \boldsymbol{h}_{1,i} = \boldsymbol{h}_{0,i}, \;\; \text{for } i = 2,3,\ldots,r. \qquad (11)$$

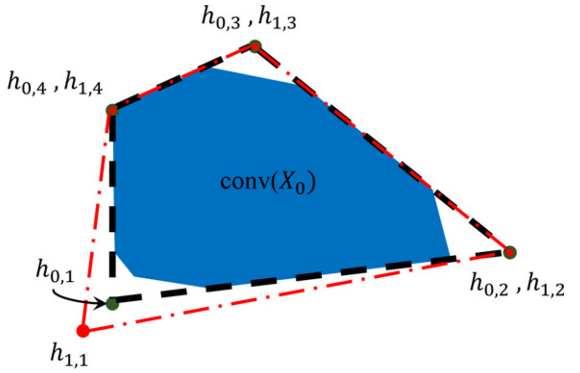


Figure 3. An example of $\boldsymbol{H}_0$, $\boldsymbol{H}_1$, $\boldsymbol{X}_0$ in Remark 3.1. The inner and outer dashed polygons have vertices –respectively– $\{\boldsymbol{h}_{0,i}\}_{i\leq r}$ and $\{\boldsymbol{h}_{1,i}\}_{i\leq r}$. The filled region represents $\mathrm{conv}(\boldsymbol{X}_0)$.

Note that $\boldsymbol{h}_{0,1} = (\widetilde{\boldsymbol{h}} + \boldsymbol{h}_{0,1})/2 \in \mathrm{conv}(\boldsymbol{H}_1)$ and therefore $\mathrm{conv}(\boldsymbol{H}_1) \supseteq \mathrm{conv}(\boldsymbol{H}_0) \supseteq \mathrm{conv}(\boldsymbol{X}_0)$, whence

$$\mathscr{D}(\boldsymbol{H}_0,\boldsymbol{H}_1)^{1/2} = 0. \qquad (12)$$

Further, $\boldsymbol{h}_{1,1} - \boldsymbol{h}_{0,1} \in \mathcal{C}_1^0$, the polar cone of $\mathcal{C}_1$ (since by definition $\mathcal{C}_1^0 = -\mathcal{C}_1^*$). Thus, the ℓ_2 projection of $\boldsymbol{h}_{1,1}$ onto $\mathrm{conv}(\boldsymbol{H}_0)$ is $\boldsymbol{h}_{0,1}$, whence

$$\mathscr{D}(\boldsymbol{H}_1,\boldsymbol{H}_0)^{1/2} = \left\|\boldsymbol{h}_{1,1} - \boldsymbol{h}_{0,1}\right\|_2. \qquad (13)$$

Further, by triangle inequality

$$\mathscr{D}(\boldsymbol{h}_{1,1},\boldsymbol{X}_0)^{1/2} \leq \mathscr{D}(\boldsymbol{h}_{0,1},\boldsymbol{X}_0)^{1/2} + \left\|\boldsymbol{h}_{1,1} - \boldsymbol{h}_{0,1}\right\|_2. \qquad (14)$$

Thus,

$$\begin{aligned}\mathscr{D}(\boldsymbol{h}_{1,1},\boldsymbol{X}_0) &\leq \mathscr{D}(\boldsymbol{h}_{0,1},\boldsymbol{X}_0) + \left\|\boldsymbol{h}_{1,1} - \boldsymbol{h}_{0,1}\right\|_2^2 \\ &\quad + 2\mathscr{D}(\boldsymbol{h}_{0,1},\boldsymbol{X}_0)^{1/2}\left\|\boldsymbol{h}_{1,1} - \boldsymbol{h}_{0,1}\right\|_2, \qquad (15)\\ \mathscr{D}(\boldsymbol{H}_1,\boldsymbol{X}_0) &\leq \mathscr{D}(\boldsymbol{H}_0,\boldsymbol{X}_0) + \left\|\boldsymbol{h}_{1,1} - \boldsymbol{h}_{0,1}\right\|_2^2 \\ &\quad + 2\mathscr{D}(\boldsymbol{h}_{0,1},\boldsymbol{X}_0)^{1/2}\left\|\boldsymbol{h}_{1,1} - \boldsymbol{h}_{0,1}\right\|_2 \qquad (16)\\ &\leq \mathscr{D}(\boldsymbol{H}_0,\boldsymbol{X}_0) + \left\|\boldsymbol{h}_{1,1} - \boldsymbol{h}_{0,1}\right\|_2^2 \\ &\quad + 2\mathscr{D}(\boldsymbol{H}_0,\boldsymbol{X}_0)^{1/2}\left\|\boldsymbol{h}_{1,1} - \boldsymbol{h}_{0,1}\right\|_2. \qquad (17)\end{aligned}$$

Therefore, by Equation (12)

$$\begin{aligned}\mathscr{D}(\boldsymbol{H}_1,\boldsymbol{X}_0)^{1/2} &\leq [\mathscr{D}(\boldsymbol{H}_0,\boldsymbol{X}_0) + \mathscr{D}(\boldsymbol{H}_1,\boldsymbol{H}_0) \\ &\quad + 2\mathscr{D}(\boldsymbol{H}_0,\boldsymbol{X}_0)^{1/2}\mathscr{D}(\boldsymbol{H}_1,\boldsymbol{H}_0)^{1/2}]^{1/2} \qquad (18)\\ &= \mathscr{D}(\boldsymbol{H}_0,\boldsymbol{X}_0)^{1/2} + \mathscr{D}(\boldsymbol{H}_1,\boldsymbol{H}_0)^{1/2} \qquad (19)\\ &= \mathscr{D}(\boldsymbol{H}_0,\boldsymbol{X}_0)^{1/2} + \mathscr{D}(\boldsymbol{H}_1,\boldsymbol{H}_0)^{1/2} \\ &\quad + \mathscr{D}(\boldsymbol{H}_0,\boldsymbol{H}_1)^{1/2}. \qquad (20)\end{aligned}$$

Hence, $\alpha \leq 1$.

We say that the convex hull $\mathrm{conv}(\boldsymbol{X}_0)$ has *internal radius* (at least) μ if it contains an $r-1$-dimensional ball of radius μ, that is, if there exists $\boldsymbol{z}_0 \in \mathbb{R}^d$, $\boldsymbol{U} \in \mathbb{R}^{d\times(r-1)}$, with $\boldsymbol{U}^\mathsf{T}\boldsymbol{U} = \mathbf{I}_d$, such that $\boldsymbol{z}_0 + \boldsymbol{U}\mathsf{B}_{r-1}(\mu) \subseteq \mathrm{conv}(\boldsymbol{X}_0)$. We further denote by $\kappa(\boldsymbol{M})$ the condition number of matrix $\boldsymbol{M}$.

Theorem 1. Assume $\boldsymbol{X} = \boldsymbol{W}_0\boldsymbol{H}_0 + \boldsymbol{Z}$ where the factorization $\boldsymbol{X}_0 = \boldsymbol{W}_0\boldsymbol{H}_0$ satisfies the uniqueness assumption with parameter $\alpha > 0$, and that $\mathrm{conv}(\boldsymbol{X}_0)$ has internal radius $\mu > 0$. Consider the estimator $\widehat{\boldsymbol{H}}$ defined by Equation (7), with $D(\boldsymbol{x},\boldsymbol{y}) = \|\boldsymbol{x}-\boldsymbol{y}\|_2^2$ (square loss) and $\delta = \max_{i\leq n}\|\boldsymbol{Z}_{i,\cdot}\|_2$. If

$$\max_{i\leq n}\|\boldsymbol{Z}_{i,\cdot}\|_2 \leq \frac{\alpha\mu}{30r^{3/2}}, \qquad (21)$$

then, we have

$$\mathscr{L}(\boldsymbol{H}_0,\widehat{\boldsymbol{H}}) \leq \frac{C_*^2 r^5}{\alpha^2}\max_{i\leq n}\|\boldsymbol{Z}_{i,\cdot}\|_2^2, \qquad (22)$$

where C_* is a coefficient that depends uniquely on the geometry of $\boldsymbol{H}_0, \boldsymbol{X}_0$, namely $C_* = 120(\sigma_{\max}(\boldsymbol{H}_0)\kappa_{\max}(\boldsymbol{H}_0)/\mu)$.

Remark 3.2. Let us emphasize that C_* is part of the bound on the robustness and its knowledge is not required by the algorithm. (It is analogous to the restricted eigenvalue constant in sparse regression.)

Remark 3.3. Robustness to outliers. Note that although the result of Theorem 1 establishes robustness of the estimator $\widehat{\boldsymbol{H}}$, the error bound depends on the maximum of norms of the rows of the noise matrix. While such a bound provides a guarantee against adversarial (nonrandom) noise, it can be overly pessimistic when the noise is mostly small, with a few outlier data points. In this scenario, better performances would probably be achieved by replacing the square loss $D(\boldsymbol{x},\boldsymbol{y}) = \|\boldsymbol{x}-\boldsymbol{y}\|_2^2$ by a more robust distance, for instance the ℓ_2 loss, $D(\boldsymbol{x},\boldsymbol{y}) = \|\boldsymbol{x}-\boldsymbol{y}\|_2$. The study of such robust distances is a promising avenue for future work.

4. Algorithms

While our main focus is on structural properties of nonnegative matrix factorization, we provide evidence that the optimization problem, we defined can be solved in practical scenarios (Our code is available online at *http://web.stanford.edu/~hrhakim/NMF/*). A more detailed study is left to future work.

From a computational point of view, the Lagrangian formulation (6) is more appealing. For the sake of simplicity, we denote the regularized risk by

$$\mathscr{R}_\lambda(\boldsymbol{H}) \equiv \mathscr{D}(\boldsymbol{X};\boldsymbol{H}) + \lambda\,\mathscr{D}(\boldsymbol{H};\boldsymbol{X})\,. \tag{23}$$

The notation $\mathscr{R}_\lambda(\boldsymbol{H})$ leaves implicit the dependence on the data $\boldsymbol{X}$. Notice that this function is nonconvex and indeed has multiple global minima. In particular, permuting the rows of a minimizer $\boldsymbol{H}$ yields other minimizers. We will describe two greedy optimization algorithms: one based on gradient descent, and one based on alternating minimization. In both cases, it is helpful to use a good initialization: two initialization methods are introduced in the next section.

4.1. Initialization

We experimented with two initialization methods, described below.

1. *Spectral initialization.* Under the assumption that the archetypes $\{\boldsymbol{h}_{0,\ell}\}_{\ell\le r}$ are linearly independent (and for a matrix of weights $\boldsymbol{W}$ with full rank), the "noiseless" matrix $\boldsymbol{X}_0$ has rank exactly r. This motivates the following approach. We compute the singular value decomposition $\boldsymbol{X} = \sum_{i=1}^{n\wedge d}\sigma_i\boldsymbol{u}_i\boldsymbol{v}_i^{\mathsf{T}}$, $\sigma_1 \ge \sigma_2 \ge \cdots \ge \sigma_{n\wedge d}$, and initialize $\widehat{\boldsymbol{H}}$ as the matrix $\widehat{\boldsymbol{H}}^{(0)}$ with rows $\hat{\boldsymbol{h}}_1^{(0)} = \boldsymbol{v}_1,\dots,\hat{\boldsymbol{h}}_r^{(0)} = \boldsymbol{v}_r$.

2. *Successive projections initialization.* In this method, we initialize $\widehat{\boldsymbol{H}}^{(0)}$ by choosing a set of archetypes $\{\hat{\boldsymbol{h}}_\ell^{(0)}\}_{1\le\ell\le r}$ that are a subset of the data $\{\boldsymbol{x}_i\}_{1\le i\le n}$, selected as follows. The first archetype $\hat{\boldsymbol{h}}_1^{(0)}$ is the data point which is farthest from the origin. For each subsequent archetype, we choose the point that is farthest from the affine subspace spanned by the previous ones, as detailed in the pseudocode below.

```
ARCHETYPE INITIALIZATION ALGORITHM
Input : Data $\{x_i\}_{i\le n}$, $x_i \in \mathbb{R}^d$; integer $r$;
Output : Initial archetypes $\{\hat{h}_\ell^{(0)}\}_{1\le\ell\le r}$;
Set $i(1) = \arg\max\{D(x_i;\mathbf{0}) \,:\, i \le n\}$;
Set $\hat{h}_1^{(0)} = x_{i(1)}$;
For $\ell \in \{1,\dots,r\}$
    Define $V_\ell \equiv \mathrm{aff}(\hat{h}_1^{(0)}, \hat{h}_2^{(0)}, \dots, \hat{h}_\ell^{(0)})$;
    Set $i(\ell+1) = \arg\max\{\mathscr{D}(x_i;V_\ell) \,:\, i \le n\}$;
    Set $\hat{h}_{\ell+1}^{(0)} = x_{i(\ell+1)}$;
End For;
Return $\{\hat{h}_\ell^{(0)}\}_{1\le\ell\le r}$;
```

This coincides with the successive projections algorithm of Araújo et al. (2001), with the minor difference that V_ℓ is the affine subspace spanned by the first ℓ vectors, instead of the linear subspace (The same modification is also used in Arora et al. (2013), but we do not apply the full algorithm of Arora et al. (2013).) This method can be proved to return the exact archetypes if data are separable and the archetypes are affinely independent Arora et al. (2013); Gillis and Vavasis (2014). When data are not separable, it provides nevertheless a good initial assignment.

4.2. Proximal Alternating Linearized Minimization

Bolte, Sabach, and Teboulle (2014) developed a proximal alternating linearized minimization algorithm (PALM) to solve the problems of the form

$$\text{minimize} \qquad \Psi(\boldsymbol{x},\boldsymbol{y}) = f(\boldsymbol{x}) + g(\boldsymbol{y}) + h(\boldsymbol{x},\boldsymbol{y}), \tag{24}$$

where $f : \mathbb{R}^m \to (-\infty,+\infty]$ and $g : \mathbb{R}^n \to (-\infty,\infty]$ are lower semicontinuous and $h \in C^1(\mathbb{R}^m \times \mathbb{R}^n)$. PALM is guaranteed to converge to critical points of the function Ψ Bolte, Sabach, and Teboulle (2014).

We apply this algorithm to minimize the cost function (23), with $D(\boldsymbol{x},\boldsymbol{y}) = \|\boldsymbol{x}-\boldsymbol{y}\|_2^2$, which we write as

$$\mathscr{R}_\lambda(\boldsymbol{H}) = \min_{\boldsymbol{W}} \Psi(\boldsymbol{H},\boldsymbol{W}) = \min_{\boldsymbol{W}} \Big\{ f(\boldsymbol{H}) + g(\boldsymbol{W}) + h(\boldsymbol{H},\boldsymbol{W}) \Big\}\,. \tag{25}$$

where

$$f(\boldsymbol{H}) = \lambda\,\mathscr{D}(\boldsymbol{H},\boldsymbol{X}), \tag{26}$$

$$g(\boldsymbol{W}) = \sum_{i=1}^{n} \mathbf{I}\left(\boldsymbol{w}_i \in \Delta^r\right), \tag{27}$$

$$h(\boldsymbol{H},\boldsymbol{W}) = \|\boldsymbol{X} - \boldsymbol{W}\boldsymbol{H}\|_F^2\,. \tag{28}$$

In above equations $\boldsymbol{w}_i$ are the rows of $\boldsymbol{W}$, and the indicator function $\mathbf{I}(\boldsymbol{x} \in \Delta^r)$ is equal to zero if $\boldsymbol{x} \in \Delta^r$ and is equal to infinity otherwise.

By using this decomposition, the iterations of the PALM algorithm read

$$\widetilde{H}^k = H^k - \frac{1}{\gamma_1^k}(W^k)^{\mathsf{T}}\left(W^kH^k - X\right), \tag{29}$$

$$H^{k+1} = \widetilde{H}^k - \frac{\lambda}{\lambda+\gamma_1^k}\left(\widetilde{H}^k - \Pi_{\mathrm{conv}(X)}\left(\widetilde{H}^k\right)\right), \tag{30}$$

$$W^{k+1} = \Pi_{\Delta^r}\left(W^k - \frac{1}{\gamma_2^k}\left(W^kH^{k+1} - X\right)(H^{k+1})^{\mathsf{T}}\right), \tag{31}$$

where γ_1^k, γ_2^k are step sizes and, for $M \in \mathbb{R}^{m_1\times m_2}$, and $\mathcal{S} \subseteq \mathbb{R}^{m_2}$ a closed convex set, $\Pi_{\mathcal{S}}(M)$ is the matrix obtained by projecting the rows of M onto the set $\mathcal{S}$.

Proposition 4.1. Consider the risk (23), with loss $D(x, y) = \|x - y\|_2^2$, and the corresponding cost function $\Psi(H, W)$. If the step sizes are chosen such that $\gamma_1^k > \left\|W^{k^{\mathsf{T}}}W^k\right\|_F$, $\gamma_2^k > \max\left\{\left\|H^{k+1}H^{k+1^{\mathsf{T}}}\right\|_F, \varepsilon\right\}$ for some constant $\varepsilon > 0$, then (H^k, W^k) converges to a stationary point of the function $\Psi(H, W)$.

The proof of this statement is deferred to Appendix D.

It is also useful to notice that the gradient of $\mathscr{R}_\lambda(H)$ can be computed explicitly (this can be useful to devise a stopping criterion).

Proposition 4.2. Consider the risk (23), with loss $D(x, y) = \|x - y\|_2^2$, and assume that the rows of H are affinely independent. Then, $\mathscr{R}_\lambda$ is differentiable at H with gradient

$$\nabla\mathscr{R}_\lambda(H) = 2\sum_{i=1}^n \alpha_i^*\left(\Pi_{\mathrm{conv}(H)}(x_i) - x_i\right) + 2\lambda\left(H - \Pi_{\mathrm{conv}(X)}(H)\right), \tag{32}$$

$$\alpha_i^* = \arg\min_{\alpha\in\Delta^r}\left\|H^{\mathsf{T}}\alpha - x_i^{\mathsf{T}}\right\|_2. \tag{33}$$

Here we recall that $\Pi_{\mathrm{conv}(X)}(H)$ denotes the matrix whose rows are equal to $\Pi_{\mathrm{conv}(X)}(H_{1,\cdot}), \dots, \Pi_{\mathrm{conv}(X)}(H_{r,\cdot})$.

The proof of this proposition is given in Appendix C. Appendix E also discusses two alternative algorithms.

Remark 4.1. Computational complexity and running time. Each iteration of the PALM algorithm consists of three steps in Equations (29) − (31). For r a fixed and small number, the first step takes $O\left(n^2 d\right)$ operations. In practice, the costliest step is the second one, which requires to compute the projection of r rows of $\widetilde{H}^k$ onto conv(X). This projection can be computed using the Frank-Wolfe algorithm (Frank and Wolfe 1956), which takes $O\left(n^2/\varepsilon\right)$ operations to output an ε-approximate solution (Jaggi 2013; Frandi, Ñanculef, and Suykens 2014). However, these projections lie on low-dimensional faces of the polytope conv(X). Using active-set methods can reduce the complexity to $O(Kn/\varepsilon)$, where $K \ll n$ is the dimension of these faces (Frandi, Ñanculef, and Suykens 2014). Finally, the last step takes $O\left(nd^2\right)$ operations. Hence, for each iteration, the overall dependence of the complexity on the dimensions is $O(nd(n + d))$.

4.3. Selecting the Regularization Parameter λ

We can view the parameter λ in Equation (23) as a regularization parameter that controls the degree to which the estimated archetypes $\widehat{H}$ are close to the convex hull of the data points. Selecting the value of the regularization parameter is a notoriously difficult problem in high-dimensional statistics. In some applications, sweeping the full regularization path can be informative, and prior knowledge about the archetypes can be used to find an appropriate value of λ. Here, we describe a possible data-driven method to select a value of λ. While this method appears to perform surprisingly well in simulations (see Table 1 and Appendix F), we keep our discussion at heuristic level, deferring a more complete study to future work.

The basic intuition is that for small λ, the objective $\mathscr{R}_\lambda(H)$ is minimized by archetypes that are fairly unconstrained, but whose convex hull contains the data as well as possible: this is an overfitting regime. For large λ, the archetypes are constrained to be in the convex hull of the data, but might not reproduce them accurately. Our proposal tries to select a λ that balances these objectives.

While Theorem 1 holds for nonrandom noise, for the present heuristic discussion it is more convenient to consider random noise. Namely, we assume that the rows z_i of matrix Z in Equation (8) are roughly isotropic vectors in $\mathbb{R}^d$, that is, for a unit vector $u \in \mathbb{R}^d$, the average of the expression $n^{-1}\sum_{i=1}^n \mathbb{E}\{\langle z_i, u\rangle^2\} \approx \sigma^2$ is roughly independent of the u.

The basic idea is to use the singular value decomposition of the data X as a measure of the noise level. Let $X = U\Sigma V^{\mathsf{T}}$ be the singular value decomposition, and denote by $P_r = V_r V_r^{\mathsf{T}}$ the projector onto the space spanned by the top-r right singular vectors. We then define the baseline error

$$\mathscr{D}_{\mathrm{LB}}^{(r)} \equiv \sum_{i=1}^n D(x_i, P_r x_i) = \sum_{i=1}^n \|x_i - P_r x_i\|_2^2. \tag{34}$$

Note that, for any reconstruction $\widehat{H}$, $\mathscr{D}(X, \widehat{H})$ is the sum of square distances of the data points from a polytope with dimension at most r. Hence $\mathscr{D}(X, \widehat{H}) \geq \mathscr{D}_{\mathrm{LB}}^{(r)}$. Further, $\mathscr{D}(X, \widehat{H}_\lambda)$ increases with λ. Figure 7 shows the behavior of the function $\lambda \mapsto \mathscr{D}(X, \widehat{H}_\lambda) - \mathscr{D}_{\mathrm{LB}}^{(r)}$ in a small simulation.

For small λ, $\mathscr{D}(X, H)$ is the dominant term in the objective function, and we expect $\mathscr{D}(X, \widehat{H}_\lambda)$ to be close to its lower bound. Data points x_i will lie outside of conv$(\widehat{H}_\lambda)$ only by a distance of the order of the noise level (in each direction). In this regime, noise dominates the reconstruction error. However, as λ increases, the term $\mathscr{D}(H, X)$ eventually becomes dominant in the cost function, and the estimated archetypes $\hat{h}_i$ move inside conv(X) even in the absence of noise. In this regime, bias dominates the reconstruction error.

It is natural to choose λ to be near the transition between the noise- and bias-dominated regimes. In practice we start with a small value $\lambda = \lambda_0$ (the method is pretty insensitive to the choice of this value and we find it very easy to set it by scanning a few values), and fix a constant $c_0 > 1$ (in practice we take $c_0 = 1.2$ but again, the precise value is not important). We then

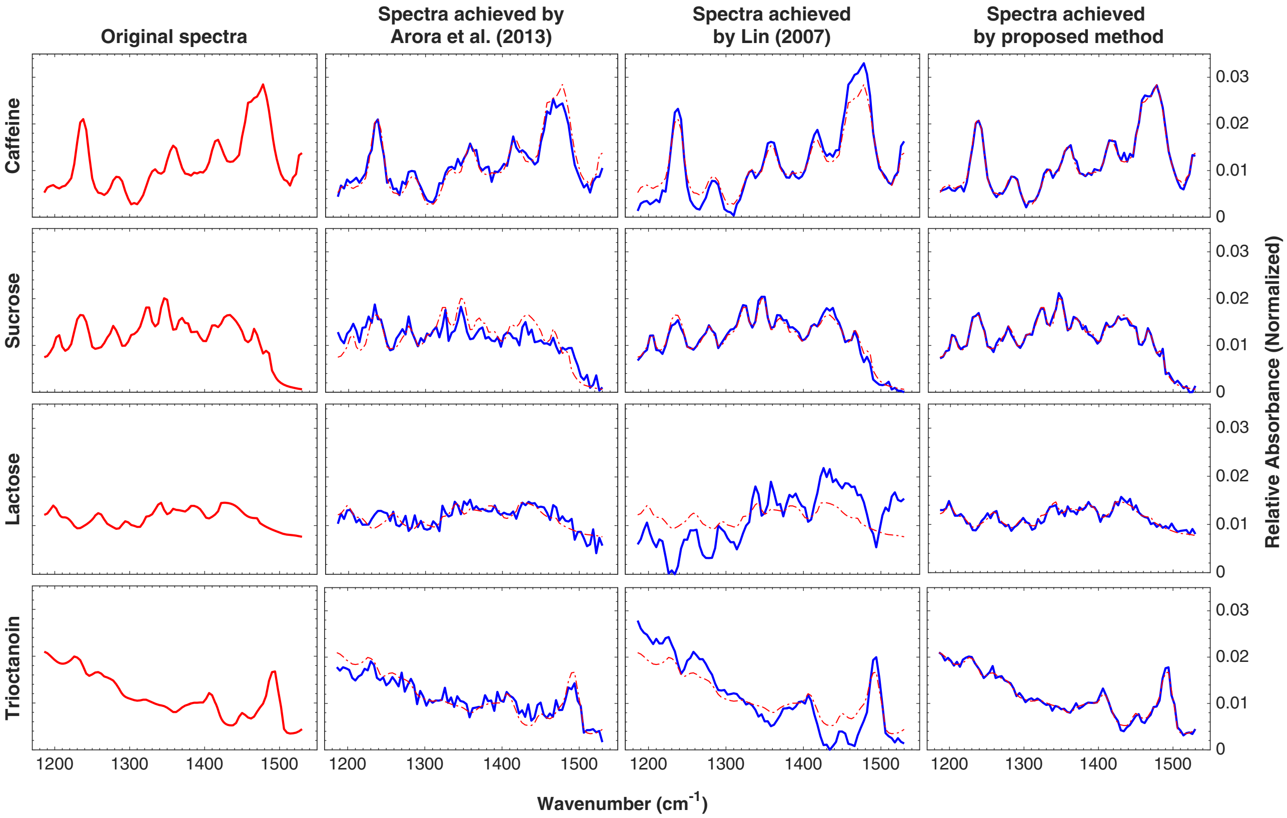


Figure 4. Reconstructing infrared spectra of four molecules, from noisy random convex combinations. Noise level $\sigma = 10^{-3}$. Left column: original spectra. The other columns correspond to different reconstruction methods.

select $\lambda = \lambda_*$ where

$$\lambda_* = \min\Big\{\lambda \geq \lambda_0 \text{ such that } \mathscr{D}\left(X, \widehat{H}_\lambda\right) - \mathscr{D}_{\rm LB}^{(r)} \geq c_0\left(\mathscr{D}\left(X, \widehat{H}_{\lambda_0}\right) - \mathscr{D}_{\rm LB}^{(r)}\right)\Big\}. \quad (35)$$

In practice we sweep over a grid of values of λ for checking the above condition.

The experiments in the next section illustrate the performances achieved by this data-driven selection method.

4.4. Numerical Experiments

We implemented both the PALM algorithm described in the previous section, and the two algorithms described in Appendix E. The outcomes are generally similar.

Figures 4 and 5 repeat the experiment already described in the introduction, in the presence of noise. We generate $n = 250$ convex combinations of $r = 4$ spectra $h_{0,1}, \dots, h_{0,4} \in \mathbb{R}^d$, $d = 87$. Unlike in the introduction, we add Gaussian noise with variance σ^2, independent across coordinates). We use the successive projection initialization discussed in Section 4.1 and minimize the Lagrangian $\mathscr{R}_\lambda(H)$, with $\lambda = 4$ (for Figure 4) and $\lambda = 0.8$ (for Figure 5). (These values were chosen as to approximately minimize the estimation error). The reconstructed spectra appear to be accurate and robust to noise.

In Figure 6, we repeated the same experiment systematically for 10 noise realizations for each noise level σ, and report the resulting average reconstruction error $\mathscr{L}(H_0, \widehat{H})$. In Table 1, we report the average reconstruction error for a grid of values of the noise level σ. We compare our method with estimates obtained with nine state-of-the-art algorithms from the literature. A few remarks are in order:

- In our approach, we did not observe significant difference in the performances achieved with different initializations. We report here the results obtained using the successive projection initialization of Section 4.1.
- In most cases, our approach achieves the smallest reconstruction error. In all cases, the error achieved is close to the smallest. In particular, the present approach appears to overperform alternative algorithms by a large factor in the small-noise regime.
- We do not notice significant degradation in performances by using the data-driven regularization parameter of Section 4.3.
- In Appendix F, we report the result of simulations with other noise models, and other choices for the archetypes. The qualitative conclusions presented here are confirmed by those results as well.

4.5. Hyperspectral Unmixing: An Experiment With Real Data

In this section, we evaluate the performance of the proposed method on a hyperspectral unmixing dataset. In these applications, the spectrum of each pixel of an input image is the result of a mixture of reflection spectra from a number of materials. The goal is to *blindly unmix* the input image to recover a set of spectra

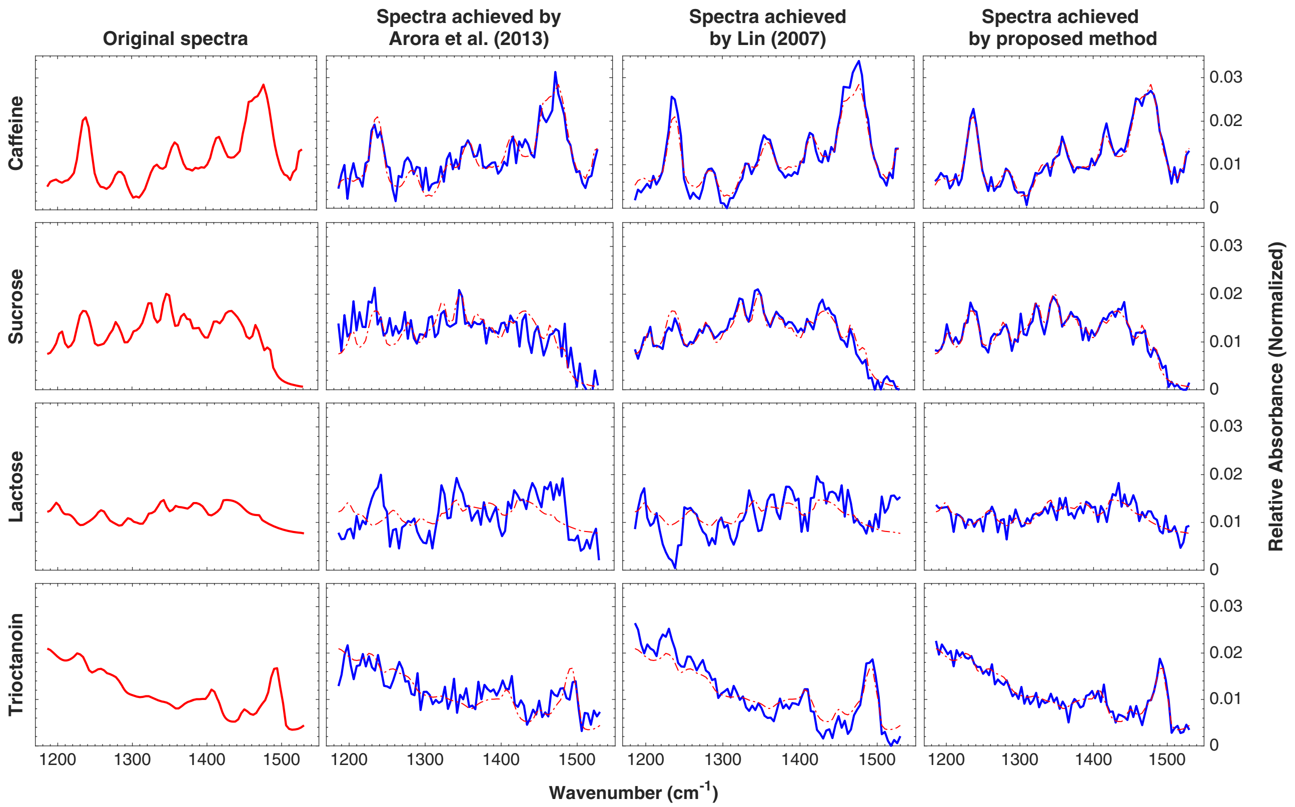


Figure 5. As in Figure 4, with $\sigma = 2 \cdot 10^{-3}$ (in blue).

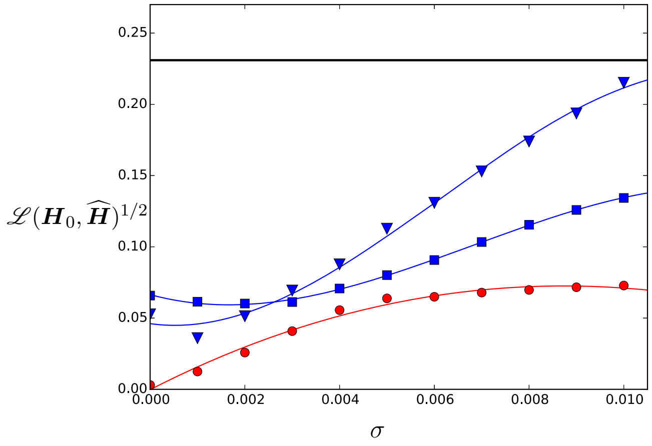


Figure 6. Risk $\mathscr{L}(\boldsymbol{H}_0, \widehat{\boldsymbol{H}})^{1/2}$ vs. σ for different reconstruction methods. Triangles (blue): anchor words algorithm from Arora et al. (2013). Squares (blue): minimizing the objective function (2) using the projected gradient algorithm of Lin (2007). Circles (red): archetypal reconstruction approach in this paper. Interpolating lines are just guides for the eye. The thick horizontal line corresponds to the trivial estimator $\widehat{\boldsymbol{H}} = 0$.

for the constituent materials called *endmembers*, as well as a set of weight vectors, one per pixel, representing the percentage of each endmember in that pixel, called *abundances* (Bioucas-Dias et al. 2012).

We use the Samson dataset from Zhu et al. (2014b), which contains the hyperspectral image of a terrain. This image has 95 × 95 pixels ($n = 9025$) each one recorded at 156 channels with wavelengths from 401 to 889 nm ($d = 156$). A photographic aerial image of the terrain is shown in Figure 8(a) (this is used for checking the results, and not as input to the algorithm). The goal is to recover the spectra of three endmembers, soil, trees, and water ($r = 3$), as well as the abundances at each pixel. The output of the algorithm for $\lambda = 1000$ can be seen in Figures 8(b) (abundances), 9 (spectra of the endmembers).

We make a few remarks:

- As it can be seen in Figure 8, the recovered abundances are consistent with the photographic image, including many details.
- In Figure 9, we compare the set of reflectance spectra recovered by our algorithm, with the corresponding spectra suggested in the literature (Zhu et al. 2014d,b,c). Despite the fact that the latter are not uniquely determined (soil, trees, and water are themselves composite), the reconstruction is quite accurate.

5. Discussion

We introduced a new optimization formulation of the nonnegative matrix factorization problem. In its Lagrangian formulation, our approach amounts to minimizing the cost function $\mathscr{R}_\lambda(\boldsymbol{H})$ defined in Equation (23). This encompasses applications in which only one of the factors is required to be non negative. A special case of this formulation ($\lambda \to \infty$) corresponds to the "archetypal analysis" of Cutler and Breiman (1994). In this case, the archetype estimates fall inside the convex hull of the data points which is appropriate only under the separability assumption of Donoho and Stodden (2003).

Our main technical result (Theorem 1) is a robustness guarantee for the reconstructed archetypes, under a certain

Table 1. Risk $\mathscr{L}(\boldsymbol{H}_0, \widehat{\boldsymbol{H}})^{1/2}$ for reconstruction of the 4 spectra in 1 using some construction methods in different noise magnitudes. The trivial estimator $\widehat{\boldsymbol{H}} = 0$ achieves $\mathscr{L}(\boldsymbol{H}_0, \widehat{\boldsymbol{H}})^{1/2} = 0.231$. For the data driven row, parameter λ is chosen as in Section 4.3 with $c_0 = 1.2$.

σ	0	0.001	0.002	0.003	0.004	0.005	0.006	0.007	0.008	0.009	0.01
Projected gradient (Lin 2007)	0.066	0.061	0.060	0.061	0.071	0.080	0.091	0.103	0.115	0.126	0.134
Multiplicative update (Lee and Seung 2001)	0.071	0.071	0.069	0.073	0.078	0.088	0.093	0.104	0.117	0.125	0.137
Fast Anchor Words (Arora et al. 2013)	0.053	0.036	0.051	0.069	0.088	0.113	0.131	0.153	0.174	0.194	0.215
Block coordinate descent (Gillis and Kumar 2015)	0.069	0.067	0.063	0.066	0.070	0.069	**0.071**	**0.08**	**0.079**	**0.085**	**0.088**
HALS (Cichocki et al. 2009)	0.076	0.075	0.078	0.081	0.079	0.084	0.093	0.104	0.119	0.124	0.135
GNMF (Cai et al. 2011) (Frobenius)	0.062	0.085	0.114	0.135	0.141	0.147	0.151	0.151	0.151	0.155	0.165
GNMF Cai et al. (2011) (KL)	0.07	0.074	0.074	0.091	0.098	0.118	0.130	0.132	0.15	0.0151	0.165
Recursive method (Gillis and Vavasis 2014)	0.038	0.043	0.054	0.073	0.091	0.110	0.134	0.150	0.175	0.192	0.21
Conical hull (Kumar, Sindhwani, and Kambadur 2013)	0.038	0.042	0.054	0.073	0.091	0.110	0.134	0.150	0.175	0.192	0.21
Our method (oracle λ)	**0.005**	**0.014**	**0.023**	**0.038**	**0.052**	**0.057**	**0.070**	**0.078**	0.092	0.094	0.102
Our method (data driven λ)	**0.008**	**0.019**	**0.027**	**0.041**	**0.056**	**0.060**	0.075	0.086	0.105	0.119	0.125

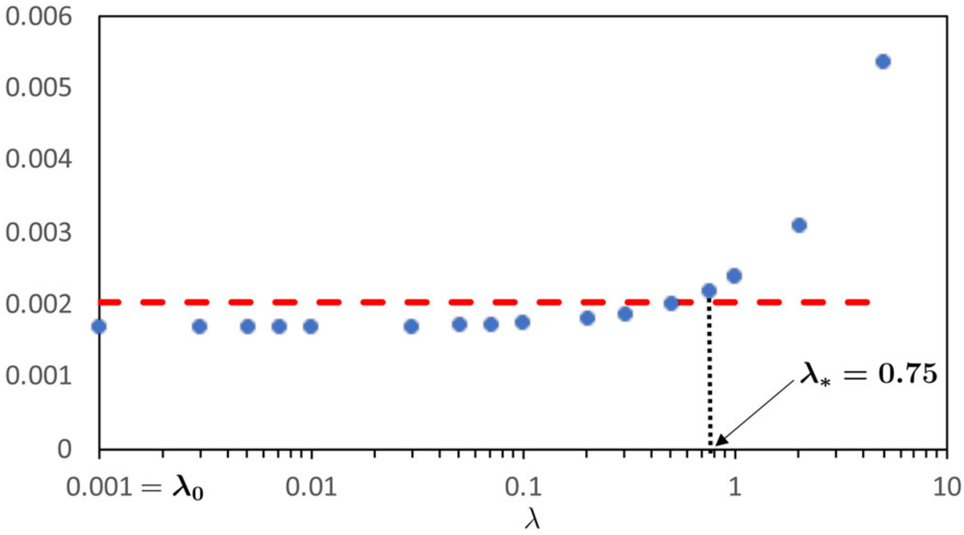
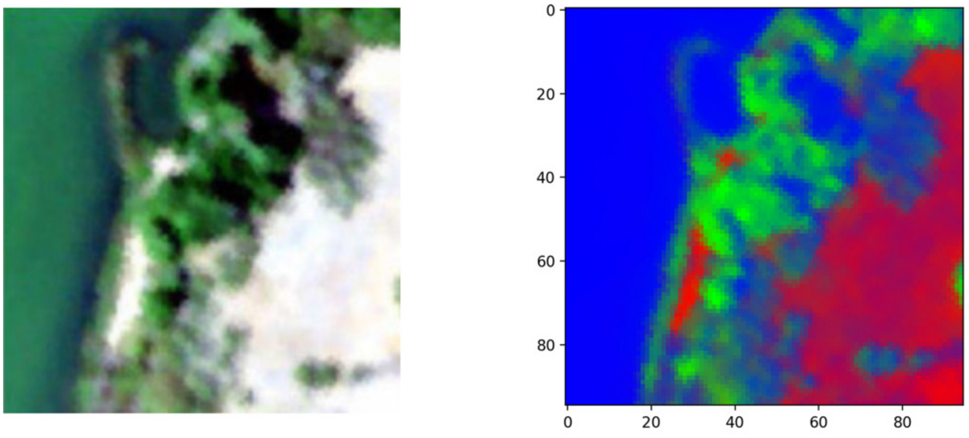


Figure 7. Procedure described in subsection 4.3 for selecting λ_*. Blue (solid) dots show the value of $\mathscr{D}(\boldsymbol{X}, \widehat{\boldsymbol{H}}_\lambda) - \mathscr{D}_{\rm LB}^{(r)}$ versus λ for λ in a grid of values. Red (dashed) line is $c_0 \left(\mathscr{D}(\boldsymbol{X}, \widehat{\boldsymbol{H}}_\lambda) - \mathscr{D}_{\rm LB}^{(r)}\right)$ for $c_0 = 1.2$. λ_0 is equal to 0.001 and λ_* is chosen as the smallest value of λ in the grid for which the blue (solid) curve is above the red (dashed) line.



(a) The original image

(b) Recovered abundances

Figure 8. The Samson image Zhu et al. (2014b) used for hyperspectral unmixing experiment.

uniqueness assumption. Uniqueness appears to hold for generic datasets. In particular, while separability implies uniqueness (with optimal constant $\alpha = 1$), uniqueness holds for nonseparable data as well. To the best of our knowledge, similar robustness results have been obtained in the past only under the more restrictive separability assumption (Recht et al. 2012; Arora et al. 2013; Gillis and Vavasis 2014, 2015) (albeit these works obtain a better dependence on r). The only exception is the recent work of Ge and Zou (2015) who proved robustness under a "subset separability" condition, which provides a significant relaxation of separability. Under this condition, Ge and Zou (2015) developed a polynomial-time algorithm to

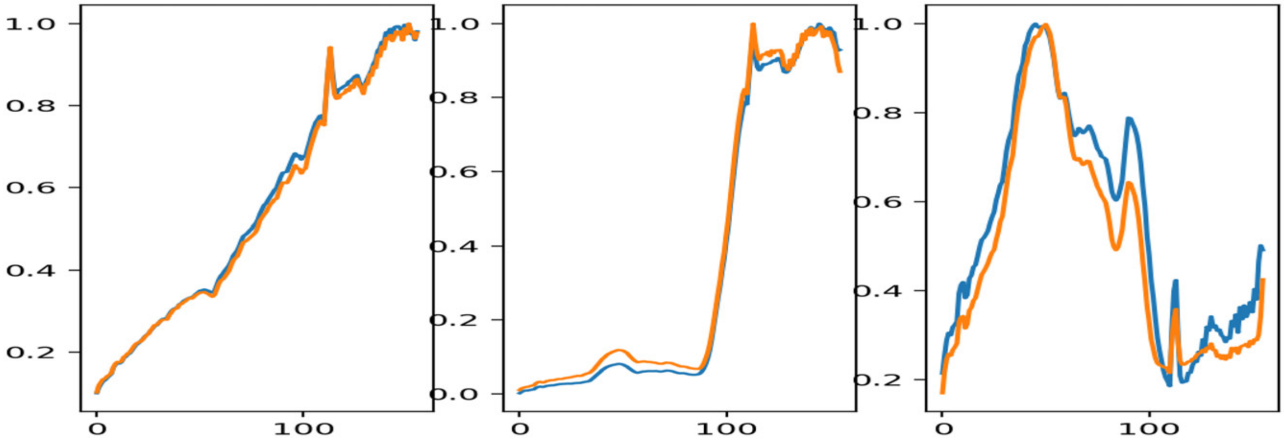


Figure 9. Reflectance spectra of the endmembers. Blue: recovered spectra of the proposed method. Orange: spectra suggested in Zhu et al. (2014d,b,c). Left: soil, middle: tree, right: water.

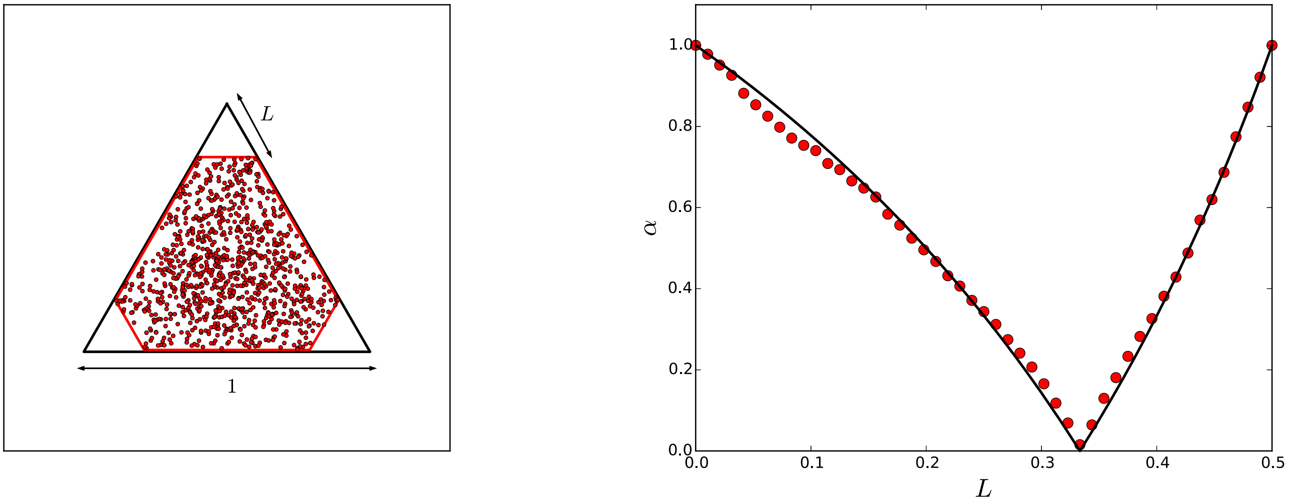


Figure 10. Numerical computation of the uniqueness parameter α. Left: data geometry. In this scenario, we take $r = 3$ and the data points are randomly generated inside a hexagon. The distance from the hexagon vertices to their closest vertex of the triangle is equal to L. In particular, the red hexagon corresponds to $\text{conv}(\boldsymbol{X})$, and the black (equilateral) triangle to the archetypes $\boldsymbol{H}_0$, for $L < 1/3$. For $L = 1/3$ the archetypes are not unique, and for $L \in (1/3, 1/2]$, they are given by an equilateral triangle rotated by $\pi/3$ (pointing down). Right: numerical evaluation of the uniqueness constant (red circles). The continuous line corresponds to an analytical upper bound (triangle rotated by $\pi/3$ with respect to $\boldsymbol{H}_0$).

estimate the archetypes by identifying and intersecting the faces of $\text{conv}(\boldsymbol{H}_0)$. However, the algorithm of Ge and Zou (2015) exploits collinearities to identify the faces, and this requires additional "genericity" assumptions.

Admittedly, the uniqueness constant α is difficult to evaluate analytically, even for simple geometries of the data. However, by definition it does not vanish except in the case of multiple minimizers, and we expect it typically to be of order one. Figure 10 illustrates this point by computing numerically α for a simple one-parameter family of geometries with $r = 3$, $d = 2$. The parameter α vanishes at a single point, corresponding to a degenerate problem with multiple solutions.

We conclude by mentioning three important problems that are not addressed by this paper: (1) Are there natural condition under which the risk function $\mathscr{R}_\lambda(\boldsymbol{H})$ of Equation (23) can be optimized in polynomial time? We only presented an algorithm that is guaranteed to converge to a critical point. (2) We assumed the rank r to be known. In practice it will need to be estimated from the data. (3) We proposed a data-driven method to select the regularization parameter λ. While this method performs well in numerical experiments, it would be interesting to develop rigorous guarantees.

Supplementary Materials

Supplementary material contains more details on the numerical experiments presented in the paper and further simulation results. Moreover, the proof of the theorems and theoretical results presented in the paper, in addition to a discussion on alternative optimization algorithms other than the PALM algorithm proposed in section 4.2 can be found in the supplementary material.

Funding

This work was partially supported by the NSF grants CCF-1319979, CCF-1714305, IIS-1741162, and a Stanford Graduate Fellowship.

References

Arora, S., Ge, R., Halpern, Y., Mimno, D. M., Moitra,A., Sontag, D., Wu, Y., Zhu, M. (2013), "A Practical Algorithm for Topic Modeling With Provable Guarantees," *ICML*, 280–288. volume: 28 issue number: 2 publisher: PMLR [2,3,4,6,9,10]

Arora, S., Ge, R., Kannan, R., and Moitra, A. (2012), "Computing a Nonnegative Matrix Factorization–Provably," in *Proceedings of the Forty-Fourth Annual ACM Symposium on Theory of Computing*, New York, NY: ACM, pp. 145–162. [2]

Araújo, M. C. U, Bezerra Saldanha, T. C., Harrop Galvao, R. K., Yoneyama, T., Chame, H. C., and Visani, V. (2001), "The Successive Projections Algorithm for Variable Selection in Spectroscopic Multicomponent Analysis," *Chemometrics and Intelligent Laboratory Systems*, 57, 65–73. [6]

Bioucas-Dias, J. M., Plaza, A., Dobigeon, N., Parente, M., Du, Q., Gader, P., and Chanussot, J. (2012), "Hyperspectral Unmixing Overview: Geometrical, Statistical, and Sparse Regression-based Approaches," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 5, 354–379. [9]

Bolte, J., Sabach, S., and Teboulle, M. (2014), "Proximal Alternating Linearized Minimization for Nonconvex and Nonsmooth Problems," *Mathematical Programming*, 146, 459–494. [6]

Cai, D., He, X., Han, J., and Huang, T. S. (2011), "Graph Regularized Nonnegative Matrix Factorization for Data Representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33, 1548–1560. [10]

Cichocki, A., Zdunek, R., Phan, A. H., and Amari, S-i. (2009), *Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-way Data Analysis and Blind Source Separation*, Hoboken, NJ: Wiley. [10]

Cutler, A., and Breiman, L. (1994), "Archetypal Analysis," *Technometrics*, 36, 338–347. [3,4,9]

Del Buono, N., and Pio, G. (2015), "Non-negative Matrix Tri-factorization for Co-clustering: An Analysis of the Block Matrix," *Information Sciences*, 301, 13–26. [1]

Donoho, D. L., and Stodden, V. (2003), "When Does Non-negative Matrix Factorization Give a Correct Decomposition Into Parts?" *Proceedings of the 16th International Conference of Neural Information Processing Systems Series: NIPS'03*, 1141–1148, Cambridge, MA: MIT Press [2,9]

Frandi, E., Ñanculef, R., and Suykens, J. (2014), "Complexity Issues and Randomization Strategies in Frank-Wolfe Algorithms for Machine Learning," arXiv preprint arXiv:1410.4062. [7]

Frank, M., and Wolfe, P. (1956), "An Algorithm for Quadratic Programming," *Naval Research Logistics*, 3, 95–110. [7]

Ge, R., and Zou, J. (2015), "Intersecting Faces: Non-negative Matrix Factorization With New Guarantees," *Proceedings of the 32nd International Conference on Machine Learning* (*ICML-15*), pp. 2295–2303. [2,10,11]

Gillis, N., and Kumar, A. (2015), "Exact and Heuristic Algorithms for Semi-nonnegative Matrix Factorization," *SIAM Journal on Matrix Analysis and Applications*, 36, 1404–1424. [10]

Gillis, N., and Vavasis, S. A. (2014), "Fast and Robust Recursive Algorithms for Separable Nonnegative Matrix Factorization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36, 698–714. [6,10]

——— (2015), "Semidefinite Programming Based Preconditioning for More Robust Near-separable Nonnegative Matrix Factorization," *SIAM Journal on Optimization*, 25, 677–698. [10]

Jaggi, M. (2013), "Revisiting Frank-Wolfe: Projection-free Sparse Convex Optimization," *ICML*, 427–435. [7]

Kim, D., Sra, S., and Dhillon, I. S. (2008), "Fast Projection-based Methods for the Least Squares Nonnegative Matrix Approximation Problem," *Statistical Analysis and Data Mining*, 1, 38–51. [2]

Kumar, A., Sindhwani, V., and Kambadur, P. (2013), "*Fast Conical Hull Algorithms for Near-Separable Non-negative Matrix Factorization*, International Conference on Machine Learning, pp. 231–239. [10]

Lin, C-J. (2007), "Projected Gradient Methods for Nonnegative Matrix Factorization," *Neural Computation*, 19, 2756–2779. [3,9,10]

Linstrom, P.J., and Mallard, W.G. (eds.), *Nist Chemistry Webbook, Nist Standard Reference Database Number 69*. Gaithersburg, MD: National Institute of Standards and Technology. Available at *http://webbook.nist.gov* (retrieved January 5, 2017). [1]

Lee, D. D., and Seung, H. S. (1999), "Learning the Parts of Objects by Non-negative Matrix Factorization," *Nature*, 401, 788–791. [1,2]

——— (2001), "Algorithms for Non-negative Matrix Factorization," *Advances in Neural Information Processing Systems*, 13, 556–562. [2,10]

Ma, W-K., Bioucas-Dias, J. M., Chan, T-H, Gillis, N., Gader, P., Plaza, A. J., Ambikapathi, A.M. and Chi, C.-Y. (2014), "A Signal Processing Perspective on Hyperspectral Unmixing: Insights From Remote Sensing," *IEEE Signal Processing Magazine*, 31, 67–81. [1]

Long, B., Zhang, Z. M., and Yu, P. S. (2005), *Co-clustering by Block Value Decomposition*, in *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, pp. 635–640. Chicago, IL: ACM [1]

Miao, L., and Qi, H. (2007), "Endmember Extraction From Highly Mixed Data Using Minimum Volume Constrained Nonnegative Matrix Factorization," *IEEE Transactions on Geoscience and Remote Sensing*, 45, 765–777. [3]

Mørup, M., and Hansen, L. K. (2012), "Archetypal Analysis for Machine Learning and Data Mining," *Neurocomputing*, 80, 54–63. [3]

Paatero, P. (1997), "Least Squares Formulation of Robust Non-negative Factor Analysis," *Chemometrics and Intelligent Laboratory Systems*, 37, 23–35. [2]

Paatero, P., and Tapper, U. (1994), Positive Matrix Factorization: A Non-negative Factor Model With Optimal Utilization of Error Estimates of Data Values, *Environmetrics*, 5, 111–126. [1,2]

Roux, J. L., Cheveigné, A. D., and Parra, L. C. (2009), "Adaptive Template Matching With Shift-invariant Semi-nmf," *Advances in Neural Information Processing Systems*, 21, 921–928. [2]

Recht, B., Re, C., Tropp, J., and Bittorf, V. (2012), "Factoring Nonnegative Matrices With Linear Programs," *Advances in Neural Information Processing Systems*, 25, 1214–1222. [2,10]

Wang, F., and Li, P. (2010), "Efficient Nonnegative Matrix Factorization With Random Projections," in *Proceedings of the 2010 SIAM International Conference on Data Mining*, Columbus, OH: SIAM, pp. 281–292. [2]

Wang, H., Nie, F., Huang, H., and Makedon, F. (2011), "Fast Non-negative Matrix Tri-factorization for Large-scale Data Co-clustering," *IJCAI Proceedings-International Joint Conference on Artificial Intelligence*, vol. 22, p. 1553. [1]

Xu, W., Liu, X., and Gong, Y. (2003), "Document Clustering Based on Non-negative Matrix Factorization," in *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, pp. 267–273. Toronto, Canada: ACM. [1]

Zhu, F., Wang, Y., Fan, B., Xiang, S., Meng, G., and Pan, C. (2014b), "Spectral Unmixing Via Data-guided Sparsity," *IEEE Transactions on Image Processing*, 23, 5412–5427. [9,10,11]

Zhu, F., Wang, Y., Fan, B., Meng, G., and Pan, C. (2014c), "Effective Spectral Unmixing Via Robust Representation and Learning-based Sparsity," arXiv:1409.0685. [9,11]

Zhu, F., Wang, Y., Xiang, S., Fan, B., and Pan, C. (2014d), "Structured Sparse Method for Hyperspectral Unmixing," *ISPRS Journal of Photogrammetry and Remote Sensing*, 88, 101–118. [9,11]

Ziegler, G. M. (2012), *Lectures on Polytopes*, vol. 152, New York, NY: Springer Science & Business Media.