Archetypal Analysis for Sparse Nonnegative Matrix Factorization: Robustness Under Misspecification

Kayhan Behdin*[1] and Rahul Mazumder†[1,2]

[1] *Operations Research Center, Massachusetts Institute of Technology*
[2] *Sloan School of Management, Massachusetts Institute of Technology*

Abstract

We consider the problem of sparse nonnegative matrix factorization (NMF) with archetypal regularization. The goal is to represent a collection of data points as nonnegative linear combinations of a few nonnegative sparse factors with appealing geometric properties, arising from the use of archetypal regularization. We generalize the notion of robustness studied in Javadi and Montanari (2019) (without sparsity) to the notions of (a) strong robustness that implies each estimated archetype is close to the underlying archetypes and (b) weak robustness that implies there exists at least one recovered archetype that is close to the underlying archetypes. Our theoretical results on robustness guarantees hold under minimal assumptions on the underlying data, and applies to settings where the underlying archetypes need not be sparse. We propose new algorithms for our optimization problem; and present numerical experiments on synthetic and real datasets that shed further insights into our proposed framework and theoretical developments.

1 Introduction

Nonnegative Matrix Factorization (NMF) (Lee and Seung, 1999) is a well-known dimensionality reduction method where we represent a collection of data points as nonnegative linear combinations of a few nonnegative latent factors. Nonnegative factors are desirable from an interpretability standpoint in applications such as computational biology (Brunet et al., 2004; Kotliar et al., 2019), image processing (Kalayeh et al., 2014; Liu et al., 2011), text mining (Berry and Browne, 2005), and chemometrics (Lawton and Sylvestre, 1971), among others. Mathematically, given a $m \times n$ data matrix with nonnegative entries $\boldsymbol{X} \in \mathbb{R}_{\geq 0}^{m \times n}$ (the rows of $\boldsymbol{X}$ correspond to the m samples and the columns the dimensions), NMF computes nonnegative lower-dimensional latent factors $\boldsymbol{W} \in \mathbb{R}_{\geq 0}^{m \times k}, \boldsymbol{H} \in \mathbb{R}_{\geq 0}^{k \times n}$. Here, k denotes the number of latent factors with $k < m, n$ and we desire the factors to lead to a good approximation of the underlying data matrix: $\boldsymbol{X} \approx \boldsymbol{W}\boldsymbol{H}$, where

*behdink@mit.edu
†rahulmaz@mit.edu

rows of $\boldsymbol{H}$ are representatives of the data and rows of $\boldsymbol{W}$ denote the coefficient weights. In the simplest form, NMF can be formulated (Lin, 2007) as the following nonconvex optimization problem

$$\min_{\boldsymbol{H},\boldsymbol{W}} \quad \|\boldsymbol{X} - \boldsymbol{W}\boldsymbol{H}\|_F^2 \quad \text{s.t.} \quad \boldsymbol{H} \in \mathbb{R}_{\geq 0}^{k\times n}, \boldsymbol{W} \in \mathbb{R}_{\geq 0}^{m\times k} \tag{1}$$

where, $\|\cdot\|_F$ denotes the Frobenius norm of a matrix.

Archetypal Analysis. The NMF problem is inherently under-determined due to scaling issues; and additional constraints should be imposed to make it well-defined. Archetypal Analysis (AA) due to Cutler and Breiman (1994) is a regularized variant of NMF where representatives of the data (aka the *archetypes*), given by the rows of $\boldsymbol{H}$, have an appealing geometric interpretation. The archetypes are chosen so that they belong to the convex hull of the data and the data is contained within their convex hull. In practice however, it may not be possible to find such archetypes (Javadi and Montanari, 2019). To this end, Javadi and Montanari (2019) propose a regularization scheme based on AA: Among all possible sets of candidate archetypes that represent the data with an acceptable accuracy, we select the one that is the closest to the convex hull of the data. Figure 1 (a), (b) illustrate the exact AA of Cutler and Breiman (1994) and regularized AA with a toy example[1].

Robustness. If $\boldsymbol{X}$ admits an exact factorization of the form $\boldsymbol{X} = \boldsymbol{W}\boldsymbol{H}$ for some $\boldsymbol{W} \in \mathbb{R}_{\geq 0}^{m\times k}$ and $\boldsymbol{H} \in \mathbb{R}_{\geq 0}^{k\times n}$, Donoho and Stodden (2004) show that under the so-called *separability* assumption[2], it is possible to recover $\boldsymbol{W}$ and $\boldsymbol{H}$ from $\boldsymbol{X}$ (up to permutation and scaling of rows of $\boldsymbol{H}$/columns of $\boldsymbol{W}$). The separability assumption has been consequently generalized to less restrictive cases, including noisy settings. In particular, Arora et al. (2012) consider an approximately separable model, where the data is assumed to be a noisy version of a separable dataset. They show that in this model, under additional regularity conditions, a polynomial-time algorithm exists that finds a factorization that is close to the factorization of the noiseless data in a suitable metric. Their results are further improved by Recht et al. (2012), showing that noisy separable NMF can be solved by linear programs. Ge and Zou (2015) show that a relaxed version of separability, the so-called *subset separability* condition, suffices to achieve a good factorization from noisy data. Note that the separability/subset separability assumptions are usually hard to verify on real data; and in our development we do not make use of this assumption. Javadi and Montanari (2019) show that AA enjoys *robustness* to perturbation: under certain assumptions, the resulting solution from AA on the perturbed data is close to the underlying model in a suitable metric (as discussed in Section 2.2). In particular, this implies that at least one of the recovered archetypes is close to the underlying archetypes that contain and represent the noiseless data. In this paper, we generalize this notion to a stronger version of robustness which implies that *each* recovered archetype is close to some true archetype.

Some earlier works have considered NMF formulations robust to outliers (Chen et al., 2014; Kong et al., 2011). In this paper, we consider a different notion of robustness—unrelated to outliers arising in the context of robust statistics (Huber, 2004).

Sparsity. Generally speaking, due to nonnegativity constraints on the latent factors, NMF is known to produce sparse solutions, that is, $\boldsymbol{H}$ and $\boldsymbol{W}$ have some zero entries (Yang and Oja, 2010). A sparse representation of the data aids in interpretability and requires less storage space. This property of NMF has

[1] See Appendix B for details of the example.

[2] The factorization $\boldsymbol{X} = \boldsymbol{W}\boldsymbol{H}$ is called separable if rows of $\boldsymbol{X}$ are among rows of $\boldsymbol{H}$.

been utilized in different applications such as image processing (Hoyer, 2004), computational biology (Kim and Park, 2007), medical imaging (Woo et al., 2018), document clustering (Kim and Park, 2008) and audio processing (Virtanen, 2007). Several papers have proposed formulations of NMF with additional penalties and/or constraints to encourage enhanced sparsity in NMF—we refer to these as sparse NMF methods. Specifically, Hoyer (2004) consider a sparse NMF problem where they use a combination of the ℓ_1 and ℓ_2 penalty on the entries of $\boldsymbol{H}$. Peharz and Pernkopf (2012) add a constraint of the form $\|\boldsymbol{H}\|_0 \leq \ell$ to problem (1), where $\|\boldsymbol{H}\|_0$ is the ℓ_0-pseudonorm, the number of nonzero entries of $\boldsymbol{H}$ and ℓ is the desired sparsity level. Kim and Park (2007, 2008) add an ℓ_1 norm penalty on the entries of $\boldsymbol{H}$ to the cost function of (1) to impose sparsity on $\boldsymbol{H}$. To the best of our knowledge, there is limited theoretical work on sparsity in NMF—in particular, towards understanding the effect of sparsity constraints on the robustness of the representation returned by NMF, an aspect we study here. In this paper, we present a simultaneous analysis of sparsity and archetypal regularization in the form of Sparse AA (SAA). We study regularized AA (Javadi and Montanari, 2019) in the presence of additional sparsity constraints on $\boldsymbol{H}$. In other words, we look for $\boldsymbol{H}$ such that it has a few nonzero entries (i.e., $\|\boldsymbol{H}\|_0$ is small), its rows describe the data well and are close to the convex hull of data. See Figure 1 (c) for a numerical illustration of this problem. In particular, we show that sparsity constrained AA leads to robust solutions, both in the weak sense of Javadi and Montanari (2019) and the stronger notion of robustness proposed here. An important feature of our analysis is that we do not assume the underlying archetypes are sparse—i.e., we can handle model misspecification—this makes our proofs different from existing work. We also discuss how noise and sparsity affect the robustness properties of the model.

Algorithms. Due to the bilinearity of the mapping $(\boldsymbol{W}, \boldsymbol{H}) \mapsto \boldsymbol{W}\boldsymbol{H}$, most formulations of NMF end up in a nonconvex optimization problem, although some convex formulations exist (Bach et al., 2008) when the dimension of the latent factors grows to infinity. Some basic approaches to these nonconvex problems include projected gradient methods (Lin, 2007), multiplicative update rules (Gonzalez and Zhang, 2005; Lee and Seung, 1999) and alternating optimization (Chu et al., 2004; Paatero and Tapper, 1994). More sophisticated algorithms for NMF have been proposed in recent years, for example see Gillis and Vavasis (2014); Leplat et al. (2019); Mizutani (2014). In this paper we present algorithms to obtain good solutions for the regularized AA problem with sparsity constraints. To this end, we present proximal block coordinate methods, and establish that they lead to a stationary point. We discuss a useful initialization scheme based on Mixed Integer Programming (MIP) (Bertsimas et al., 2016; Wolsey and Nemhauser, 1999) that leads to high-quality solutions. To further improve the quality of solutions available from our block coordinate procedure, we present local search based methods—to this end, our framework draws inspiration from the work of Beck and Eldar (2013); Hazimeh and Mazumder (2020) and adapts it to the setting of matrix factorization problems. Note that Abrol and Sharma (2020); Elhamifar et al. (2012); Mørup and Hansen (2012) have proposed algorithms for the original AA problem without sparsity constraints. In addition, prior work has extended the notion of separability arising from NMF to address the AA problem: For example, Damle and Sun (2017) use geometric interpretations of AA and separability to develop a new algorithm for NMF. Our numerical experiments on synthetic, and real datasets validate our theoretical results, and suggest the superiority of SAA over other popular sparse NMF methods.

Our Contributions. Our contribution in this paper can be summarized as follows:

- We generalize the robustness framework of Javadi and Montanari (2019) to notions of *weak* and *strong*

robustness, introduced in this paper—these notions differ in how they describe the proximity of our estimators to the underlying archetypes. Furthermore, we prove robust solutions are good representatives of the noiseless data.

- We show how sparsity and AA can be used together to produce sparse factors that are robust to noise and perturbation in the data. Our results apply to the mis-specified setting–i.e., situations where the underlying archetypes are not necessarily sparse.
- We present algorithms[3] based on block proximal descent and local search, discuss MIP-based initialization strategies; and present convergence properties of our proposed algorithmic framework. We demonstrate via numerical experiments on synthetic and real-datasets the usefulness of our proposed approach.

Notation. For a matrix $\boldsymbol{X}$, we let $\boldsymbol{X}_{i,j}$, $\boldsymbol{X}_{i,\cdot}$ and $\boldsymbol{X}_{\cdot,j}$ denote the (i,j)-th element, i-th row and j-th column of $\boldsymbol{X}$, respectively. With a slight abuse of notation, for indexed variables, we move the subscript to superscript: that is, the i-th row of $\boldsymbol{X}_0$ is shown as $\boldsymbol{X}^0_{i,\cdot}$. The number of rows of a matrix $\boldsymbol{X}$ is denoted as $\#\text{row}(\boldsymbol{X})$. The convex hull of rows of the matrix $\boldsymbol{X} \in \mathbb{R}^{m\times n}$ is denoted by

$$\text{Conv}(\boldsymbol{X}) = \left\{ \sum_{i=1}^{m} \alpha_i \boldsymbol{X}_{i,\cdot} : \alpha_i \geq 0, \sum_{i=1}^{m} \alpha_i = 1 \right\}.$$

We define (i) the distance between a vector $\boldsymbol{x}$ and the convex hull $\text{Conv}(\boldsymbol{X})$ as $D(\boldsymbol{x}, \boldsymbol{X}) = \min_{\boldsymbol{v}\in\text{Conv}(\boldsymbol{X})} \|\boldsymbol{x} - \boldsymbol{v}\|_2^2$ (ii) the distance between a set of points (i.e., the rows of $\boldsymbol{X}$) and $\text{Conv}(\boldsymbol{Y})$ as

$$D(\boldsymbol{X}, \boldsymbol{Y}) = \sum_{i=1}^{\#\text{row}(\boldsymbol{X})} D(\boldsymbol{X}_{i,\cdot}, \boldsymbol{Y}) \quad \text{and} \quad D(\boldsymbol{X}, \boldsymbol{Y})^{1/2} = \sqrt{D(\boldsymbol{X}, \boldsymbol{Y})}.$$

The number of nonzero entries in a matrix $\boldsymbol{A}$ is denoted as $\|\boldsymbol{A}\|_0$. We define $P_\ell(\boldsymbol{H})$ to be the projection of $\boldsymbol{H} \in \mathbb{R}^{k\times n}$ onto the ℓ-sparse set $\{\boldsymbol{X} \in \mathbb{R}^{k\times n} : \|\boldsymbol{X}\|_0 \leq \ell\}$. Moreover, we define the complement as $P^\perp_\ell(\boldsymbol{H}) = \boldsymbol{H} - P_\ell(\boldsymbol{H})$. The support of a matrix $\boldsymbol{H} \in \mathbb{R}^{k\times n}$, $S(\boldsymbol{H})$, is defined as the set of its nonzero coordinates:

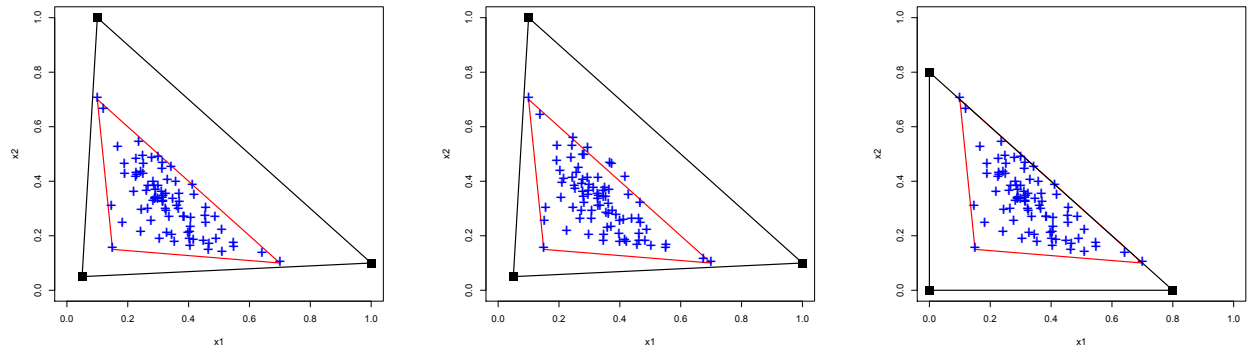
$$S(\boldsymbol{H}) = \{(i,j) \in [k] \times [n] : |\boldsymbol{H}_{i,j}| > 0\}.$$

We set $\boldsymbol{E}^{i,j} \in \mathbb{R}^{k\times n}$ to be the matrix with coordinate (i,j) equal to one and other coordinates equal to zero. Throughout this paper, we use $\sigma_{\min}(\boldsymbol{H})$ and $\sigma_{\max}(\boldsymbol{H})$ to denote the smallest and largest singular values of $\boldsymbol{H}$ (respectively). We let $\kappa(\boldsymbol{H}) := \sigma_{\max}(\boldsymbol{H})/\sigma_{\min}(\boldsymbol{H})$ denote the condition number of $\boldsymbol{H}$. For a convex and subdifferentiable function $f : \mathbb{R}^d \to \mathbb{R}$, $\partial f(\boldsymbol{x})$ denotes the set of subgradients of f at $\boldsymbol{x} \in \mathbb{R}^d$. Proofs of main results have been relegated to the appendix to improve readability.

2 Problem Formulation

Given m data points in $\mathbb{R}^n$, stacked along the rows of $\boldsymbol{X} \in \mathbb{R}^{m\times n}$, the goal of AA is to find k archetypes $\boldsymbol{H}_{1,\cdot}, \cdots, \boldsymbol{H}_{k,\cdot} \in \mathbb{R}^n$ such that: (i) the rows of $\boldsymbol{X}$ are contained in the convex hull of the rows of $\boldsymbol{H}$; and (ii) the rows of $\boldsymbol{H}$ are themselves close to the convex hull of the rows of $\boldsymbol{X}$. In SAA, we wish to learn a

[3]Implementation can be found at `https://github.com/kayhanbehdin/SparseAA`.



(a) (b) (c)

Figure 1: In these figures, blue crosses ('+') represent the data points in $\mathbb{R}^2$. We seek to find 3 archetypes such that their convex hull contains the data. Panel (a): the black convex hull (triangle) shows an arbitrary solution to NMF, while the red convex hull shows the exact AA solution that is the smallest triangle containing the data. Panel (b): the black convex hull shows a solution that describes the data with no error, while it is not close to the convex hull of the data. The red convex hull shows the regularized AA solution, which both describes the data well (but with nonzero error) and is close to the data. Panel (c): the black convex hull shows the exact AA solution which does not have any zero coordinate, while the red convex hull shows a solution which is sparse and only has 2 nonzero coordinates. In addition, no other solution with the same sparsity can be found which is closer to the data.

sparse matrix $\boldsymbol{H}$, i.e., one with few nonzero entries. Equivalently, we seek to learn $\boldsymbol{H}$ such that

$$D(\boldsymbol{X}, \boldsymbol{H}) = D(\boldsymbol{H}, \boldsymbol{X}) = 0 \tag{2}$$

where the first term ensures the data is described by $\boldsymbol{H}$ (as it implies there exists $\boldsymbol{W} \in \mathbb{R}_{\geq 0}^{m \times k}$ such that $\boldsymbol{X} = \boldsymbol{W}\boldsymbol{H}$ and each row of $\boldsymbol{W}$ sums to one). The second term ensures that rows of $\boldsymbol{H}$ are in the convex hull of rows of $\boldsymbol{X}$. As discussed earlier, the constraint (2) is too restrictive for most practical cases. Javadi and Montanari (2019) propose a relaxed version of (2), where among all the archetypes that contain the data with acceptable accuracy, they choose archetypes that are closest to the convex hull of data. We follow suit, but in addition, we impose a sparsity on $\boldsymbol{H}$ via the constraint $\|\boldsymbol{H}\|_0 \leq \ell$, where, ℓ is a budget on the number of nonzero entries. Specifically, for a pre-specified value of α, we consider:

$$\hat{\boldsymbol{H}} \in \underset{\boldsymbol{H} \in \mathbb{R}_{\geq 0}^{k \times n}}{\operatorname{argmin}} \; D(\boldsymbol{H}, \boldsymbol{X}) \quad \text{s.t.} \quad D(\boldsymbol{X}_{i,.}, \boldsymbol{H})^{1/2} \leq \alpha, i \in [m]; \quad \|\boldsymbol{H}\|_0 \leq \ell. \tag{3}$$

Above the constraint "$D(\boldsymbol{X}_{i,.}, \boldsymbol{H})^{1/2} \leq \alpha, i \in [m]$" restricts the data points (rows of $\boldsymbol{X}$) to be close to the convex hull of the rows of $\boldsymbol{H}$. We also constrain the archetypes (i.e., the right latent factors) to be sparse. Since there are multiple feasible solutions in problem (3), the objective function chooses the archetypes closest to the convex hull of data points, therefore, making the problem more well-defined.

2.1 Model setup

Suppose $\boldsymbol{X}_0 \in \mathbb{R}_{\geq 0}^{m\times n}$ with $\text{rank}(\boldsymbol{X}_0) = k$ admits a nonnegative factorization of rank k. That is, there exist $\bar{\boldsymbol{W}} \in \mathbb{R}_{\geq 0}^{m\times k}, \bar{\boldsymbol{H}} \in \mathbb{R}_{\geq 0}^{k\times n}$ such that $\boldsymbol{X}_0 = \bar{\boldsymbol{W}}\bar{\boldsymbol{H}}$ and rows of $\bar{\boldsymbol{W}}$ sum to one. This is equivalent to $D(\boldsymbol{X}_0, \bar{\boldsymbol{H}}) = 0$. However, such a factorization is not generally unique. Hence, we let

$$\boldsymbol{H}_0 \in \underset{\boldsymbol{H}\in\mathbb{R}_{\geq 0}^{k\times n}}{\text{argmin}} \quad D(\boldsymbol{H}, \boldsymbol{X}_0) \;\; \text{s.t.} \;\; D(\boldsymbol{X}_0, \boldsymbol{H}) = 0 \tag{4}$$

and assume that $\boldsymbol{H}_0$ is unique. The choice of $\boldsymbol{H}_0$ in problem (4) guarantees that $\boldsymbol{X}_0$ has an exact factorization of the form $\boldsymbol{X}_0 = \boldsymbol{W}_0\boldsymbol{H}_0$ for some $\boldsymbol{W}_0 \in \mathbb{R}_{\geq 0}^{m\times k}$ such that its rows sum to one and $\boldsymbol{H}_0$ is defined as in (4). Note that $\text{rank}(\boldsymbol{H}_0) = k$—otherwise, if $\text{rank}(\boldsymbol{H}_0) < k$, then $\text{rank}(\boldsymbol{X}_0) < k$ which contradicts our assumption on $\boldsymbol{X}_0$.

We assume that the observed data matrix $\boldsymbol{X}$, is given by $\boldsymbol{X} = \boldsymbol{X}_0 + \boldsymbol{Z}$ where $\boldsymbol{Z}$ is additive noise. In what follows, we do not make any distributional assumptions on $\boldsymbol{Z}$.

Remark 1. Our analysis of robustness is valid without the uniqueness assumption on $\boldsymbol{H}_0$. We use uniqueness for simplicity of exposition.

Remark 2. Note that we do not assume that $\boldsymbol{H}_0$ is sparse, though formulation (3) imposes an explicit cardinality constraint on $\boldsymbol{H}$. This model misspecification leads to technical challenges: The analysis presented in Javadi and Montanari (2019) does not readily generalize to our setting, and we present a new analysis technique.

2.2 Robustness to noise in Archetypal Analysis

We are interested to see if a solution $\hat{\boldsymbol{H}}$ of (3) is close to $\boldsymbol{H}_0$, the underlying set of archetypes. To this end, following Javadi and Montanari (2019), we define a distance between two sets of archetypes $\boldsymbol{H}_1 \in \mathbb{R}^{k_1\times n}, \boldsymbol{H}_2 \in \mathbb{R}^{k_2\times n}$ as

$$\mathcal{L}(\boldsymbol{H}_1, \boldsymbol{H}_2) = \sum_{i=1}^{k_1} \min_{j\in[k_2]} \|\boldsymbol{H}_{i,\cdot}^1 - \boldsymbol{H}_{j,\cdot}^2\|_2^2. \tag{5}$$

Javadi and Montanari (2019) show that under the so-called *uniqueness* assumption, one has $\mathcal{L}(\boldsymbol{H}_0, \boldsymbol{H}_{\boldsymbol{X}})^{1/2} \leq C \max_{i\in[m]} \|\boldsymbol{Z}_{i,\cdot}\|_2$ for some constant $C > 0$ where $\boldsymbol{H}_{\boldsymbol{X}}$ is the solution to the relaxed AA (i.e. (3) with $\ell = nk$). Note that $\mathcal{L}(\boldsymbol{H}_1, \boldsymbol{H}_2)$ in (5) is a sum of the distances between each row of $\boldsymbol{H}_1$ and rows of $\boldsymbol{H}_2$. Observe that $\mathcal{L}(\boldsymbol{H}_1, \boldsymbol{H}_2)$ is not symmetric in its arguments. In fact, a small value of $\mathcal{L}(\boldsymbol{H}_1, \boldsymbol{H}_2)$ does not imply that $\mathcal{L}(\boldsymbol{H}_2, \boldsymbol{H}_1)$ is also small (see Section 2.3 for details). Definition 1 below presents a formal definition of weak and strong robustness.

Definition 1. (**Robustness**) An estimator $\boldsymbol{H} \in \mathbb{R}_{\geq 0}^{k\times n}$ is said to be:

(1) (*Weak robustness*): Weakly robust if $\mathcal{L}(\boldsymbol{H}_0, \boldsymbol{H})^{1/2} \leq f_1(\max_{i\in[m]} \|\boldsymbol{Z}_{i,\cdot}\|_2)$ where f_1 is an increasing real function that does not depend on $\boldsymbol{X}$.

(2) (*Strong robustness*): Strongly robust if $\mathcal{L}(\boldsymbol{H}, \boldsymbol{H}_0)^{1/2} \leq f_2(\max_{i\in[m]} \|\boldsymbol{Z}_{i,\cdot}\|_2)$ where f_2 is an increasing real function that does not depend on $\boldsymbol{X}$.

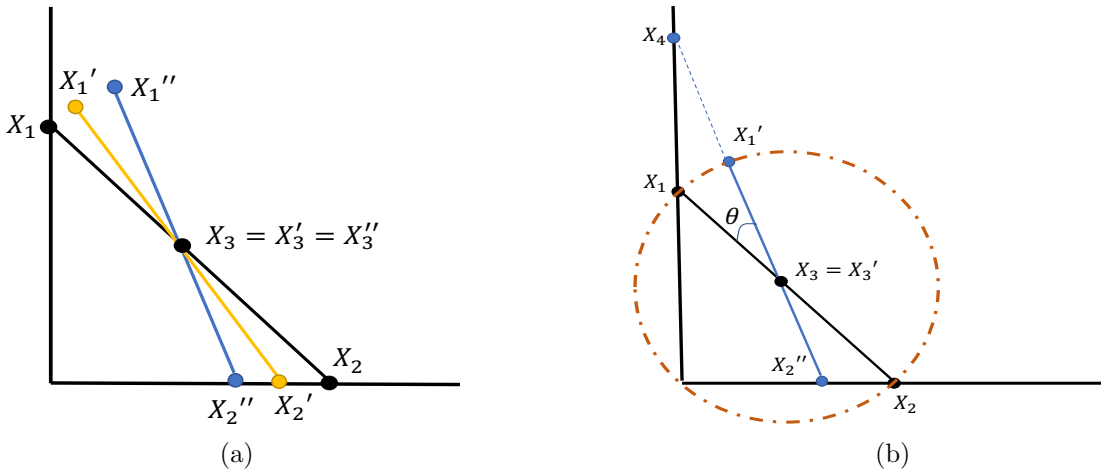


Figure 2: Illustration of Example 1: (a) The noiseless data and two examples of noisy data. (b) Details of the example for a specific value of θ.

Note that based on Definition 1, the result of Javadi and Montanari (2019) is an instance of weak robustness with $f_1(x) = Cx$ for $x \geq 0$ and some $C > 0$.

2.3 Strong robustness implies weak robustness

Here we explain the differences between weak and strong robustness and provide some intuition around the choice of terminology: strong and weak. Example 1 shows an estimator that is weakly robust but not strongly robust.

Example 1. Let $m = 3$, $n = 2$ and $k = 2$ and

$$\boldsymbol{X}_0 = \begin{bmatrix} 0 & 1 \\ 1 & 0 \\ 1/2 & 1/2 \end{bmatrix}, \qquad \boldsymbol{H}_0 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

For $\theta \in (0, \pi/4)$, we let the noisy data ($\boldsymbol{X}_\theta$) and noise matrix ($\boldsymbol{Z}_\theta$) be:

$$\boldsymbol{X}_\theta = \begin{bmatrix} \sqrt{1-\cos\theta}\cos(\frac{\pi}{4} - \frac{\theta}{2}) & 1 + \sqrt{1-\cos\theta}\sin(\frac{\pi}{4} - \frac{\theta}{2}) \\ 1 - \frac{\sin\theta}{\sqrt{2}\sin(\theta + \frac{\pi}{4})} & 0 \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix} \quad \text{and}$$

$$\boldsymbol{Z}_\theta = \begin{bmatrix} \sqrt{1-\cos\theta}\cos(\frac{\pi}{4} - \frac{\theta}{2}) & \sqrt{1-\cos\theta}\sin(\frac{\pi}{4} - \frac{\theta}{2}) \\ -\frac{\sin\theta}{\sqrt{2}\sin(\theta + \frac{\pi}{4})} & 0 \\ 0 & 0 \end{bmatrix}.$$

Figure 2 presents an illustration of Example 1. In this figure (panel (a)), X_1, X_2, X_3 correspond to the original data points (i.e., the rows of $\boldsymbol{X}_0$) and points X_1', X_2', X_3' and X_1'', X_2'', X_3'' are the noisy data points for two different values of θ. Note that, in all cases, the data points lie on a line. The lines X_1', X_2', X_3' and

X_1'', X_2'', X_3'' are obtained by rotating the noiseless data line X_1, X_2, X_3 along its center $(1/2, 1/2)$. Figure 2 (panel (b)) shows line X_1, X_2, X_3 and its rotated version X_1', X_2', X_3' (after being rotated by an angle θ). Let

$$\boldsymbol{H}_\theta = \begin{bmatrix} 0 & \left[1 - \frac{\sin\theta}{\sqrt{2}\sin(\theta+\frac{\pi}{4})}\right]\tan(\theta+\frac{\pi}{4}) \\ 1 - \frac{\sin\theta}{\sqrt{2}\sin(\theta+\frac{\pi}{4})} & 0 \end{bmatrix}.$$

Note that $\boldsymbol{H}^\theta_{2,\cdot}$ is the same as the point $\boldsymbol{X}^\theta_{2,\cdot}$. The line passing through the noisy data points intersects the y-axis at $\boldsymbol{H}^\theta_{1,\cdot}$. For the line, X_1', X_2', X_3', the point $\boldsymbol{H}^\theta_{1,\cdot}$ is given by X_4 in Figure 2, (b). As a result, $\boldsymbol{X}^\theta_{1,\cdot}, \boldsymbol{X}^\theta_{2,\cdot}, \boldsymbol{X}^\theta_{3,\cdot}$ are on the segment connecting $\boldsymbol{H}^\theta_{1,\cdot}$ and $\boldsymbol{H}^\theta_{2,\cdot}$ and $D(\boldsymbol{X}_\theta, \boldsymbol{H}_\theta) = 0$. In addition, $\max_{i\in[3]} \|\boldsymbol{Z}^\theta_{i,\cdot}\|_2 \leq \sqrt{2}$ showing the amount of noise added to the data is limited. Moreover,

$$\mathcal{L}(\boldsymbol{H}_0, \boldsymbol{H}_\theta) \leq \|\boldsymbol{H}^0_{1,\cdot} - \boldsymbol{H}^\theta_{2,\cdot}\|_2^2 + \|\boldsymbol{H}^0_{2,\cdot} - \boldsymbol{H}^\theta_{2,\cdot}\|_2^2 \leq 4$$

for all $\theta \in (0, \pi/4)$, showing $\boldsymbol{H}_\theta$ is weakly robust. However, note that

$$\begin{aligned}\mathcal{L}(\boldsymbol{H}_\theta, \boldsymbol{H}_0) &\geq \min_{i\in[2]} \|\boldsymbol{H}^\theta_{1,\cdot} - \boldsymbol{H}^0_{i,\cdot}\|_2^2 = \|\boldsymbol{H}^\theta_{1,\cdot} - \boldsymbol{H}^0_{2,\cdot}\|_2^2 \\ &= \left(\left[1 - \frac{\sin\theta}{\sqrt{2}\sin(\theta+\pi/4)}\right]\tan(\theta+\pi/4) - 1\right)^2\end{aligned}$$

and $\mathcal{L}(\boldsymbol{H}_\theta, \boldsymbol{H}_0) \to \infty$ as $\theta \to \pi/4$, showing $\boldsymbol{H}_\theta$ is not strongly robust.

$\boldsymbol{H}_\theta$ is not a strongly robust estimator in Example 1 as the first row of $\boldsymbol{H}_\theta$ can be far from $\boldsymbol{H}_0$, the set of underlying archetypes. Weak robustness implies that among recovered archetypes, there is at least one of them which is close to the correct archetypes (for example, the second row of $\boldsymbol{H}_\theta$ in Example 1 is close to the first row of $\boldsymbol{H}_0$). On the other hand, strong robustness implies that all recovered archetypes are close to some underlying archetype. This is true because considering the definition of $\mathcal{L}(\boldsymbol{H}_1, \boldsymbol{H}_2)$ in (5), strong robustness limits the distance between each recovered archetype and the set of correct archetypes. However, weak robustness limits the distance between each true archetype and the set of recovered archetypes, therefore, this distance can be small even if some recovered archetypes are far. Theorem 1 below states that strong robustness implies weak robustness (and in light of Example 1, this containment is strict).

Theorem 1 (Strong robustness implies weak robustness)**.** Let us define the quantity

$$b(\boldsymbol{H}_0) = \max_{i,j\in[k]} \|\boldsymbol{H}^0_{i,\cdot} - \boldsymbol{H}^0_{j,\cdot}\|_2. \tag{6}$$

Then for any $\boldsymbol{H} \in \mathbb{R}^{k\times n}_{\geq 0}$ we have

$$\mathcal{L}(\boldsymbol{H}_0, \boldsymbol{H}) \leq 2kb(\boldsymbol{H}_0)^2 + 2\mathcal{L}(\boldsymbol{H}, \boldsymbol{H}_0). \tag{7}$$

If $\boldsymbol{H}$ is a strongly robust estimator, $\mathcal{L}(\boldsymbol{H}, \boldsymbol{H}_0)$ is bounded and as $b(\boldsymbol{H}_0)$ depends on the underlying archetypes, $b(\boldsymbol{H}_0)$ is finite. Therefore, the right hand side in (7) is bounded and $\mathcal{L}(\boldsymbol{H}_0, \boldsymbol{H})$ is bounded, implying $\boldsymbol{H}$ is also weakly robust.

Theorem 2 shows that a weak/strong robust estimator $\boldsymbol{H}$ serves as a good approximation to the noiseless

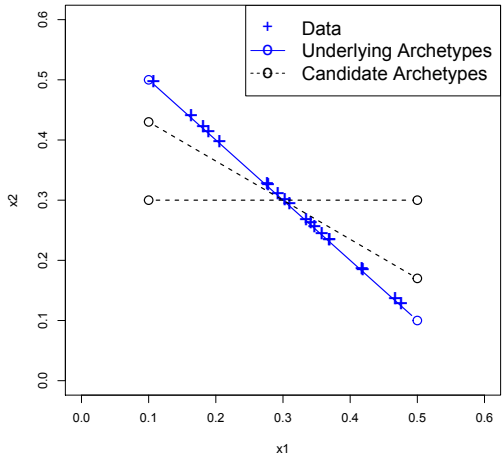


Figure 3: In this figure, blue crosses represent the data points in $\mathbb{R}^2$ and blue circles and line represent $\boldsymbol{H}_0$ and its convex hull containing the data. Black circles and lines represent two candidate set of archetypes and their convex hulls. Note that the set that is closer to $\boldsymbol{H}_0$ describes the data better as anticipated by Theorem 2.

data matrix $\boldsymbol{X}_0$.

Theorem 2. For any $\boldsymbol{H} \in \mathbb{R}^{k\times n}_{\geq 0}$ we have

$$D(\boldsymbol{X}_0, \boldsymbol{H})^{1/2} \leq \sqrt{m} \min\left\{\mathcal{L}(\boldsymbol{H}_0, \boldsymbol{H})^{1/2}, k\|\boldsymbol{H}_0\|_F + \mathcal{L}(\boldsymbol{H}, \boldsymbol{H}_0)^{1/2}\right\}. \tag{8}$$

If the right hand side in (8) is small, which means $\boldsymbol{H}$ is either a weakly or strongly robust estimator, then the left hand side in (8) is small. This shows that the noiseless data is close to the convex hull of $\boldsymbol{H}$—so rows of $\boldsymbol{H}$ are good representatives of the noiseless data. See Figure 3 for an illustrative example.

3 Sparse AA

In this section, our primary goal is to show that under certain conditions, both weak and strong robustness hold for the SAA estimator (3).

Theorem 3. Let $\boldsymbol{X}_0, \boldsymbol{H}_0$ be as described in Section 2.1 and $\hat{\boldsymbol{H}}$ be a solution of problem (3). Set $\alpha = \delta + \beta$ where $\delta = \max_{i\in[m]} \|\boldsymbol{Z}_{i,\cdot}\|_2$ and $\beta = \sqrt{m}\|P_\ell^\perp(\boldsymbol{H}_0)\|_F$. Moreover, let $\boldsymbol{X} = \boldsymbol{X}_0 + \boldsymbol{Z}$ and $\tilde{\boldsymbol{X}}_0 \in \mathbb{R}^{k\times n}$ be such that

$$\tilde{\boldsymbol{X}}^0_{i,\cdot} = \operatorname*{argmin}_{\boldsymbol{u}\in\{\boldsymbol{X}^0_{j,\cdot}: j\in[m]\}} \|\boldsymbol{u} - \boldsymbol{H}^0_{i,\cdot}\|_2.$$

There exist constants[4] $c_1, \cdots, c_{10}$ depending on $m, n, k, \kappa(\boldsymbol{H}_0), \sigma_{\min}(\boldsymbol{H}_0)$ such that the following bounds hold:

1. (Weak Robustness)

$$\mathcal{L}(\boldsymbol{H}_0, \hat{\boldsymbol{H}})^{1/2} \leq c_1 D(\boldsymbol{H}_0, \tilde{\boldsymbol{X}}_0)^{1/2} + c_2 \max_{i\in[m]} \|\boldsymbol{Z}_{i,\cdot}\|_2 + c_3\|P_\ell^\perp(\boldsymbol{H}_0)\|_F \tag{9}$$

[4]To aid readability, the precise expressions for these constants are presented in Section D.3.

2. (Strong Robustness) If

$$c_4 D(\boldsymbol{H}_0, \tilde{\boldsymbol{X}}_0)^{1/2} + c_5 \max_{i\in[m]} \|\boldsymbol{Z}_{i,\cdot}\|_2 + c_6 \|P_\ell^\perp(\boldsymbol{H}_0)\|_F \leq c_7, \tag{10}$$

we have the following strong robustness guarantee

$$\mathcal{L}(\hat{\boldsymbol{H}}, \boldsymbol{H}_0)^{1/2} \leq c_8 D(\boldsymbol{H}_0, \tilde{\boldsymbol{X}}_0)^{1/2} + c_9 \max_{i\in[m]} \|\boldsymbol{Z}_{i,\cdot}\|_2 + c_{10} \|P_\ell^\perp(\boldsymbol{H}_0)\|_F. \tag{11}$$

In Theorem 3, δ controls the amount of additive noise $\boldsymbol{Z}$; and $\beta = \sqrt{m}\|P_\ell^\perp(\boldsymbol{H}_0)\|_F$ captures the sparsity level of $\boldsymbol{H}_0$. If the data is noisy (i.e. δ is large) and/or $\boldsymbol{H}_0$ is not sparse (i.e. β is large), the value of α in problem (3) is larger. This implies that $D(\boldsymbol{X}_{i,\cdot}, \hat{\boldsymbol{H}})$ for $i \in [m]$ can be potentially larger because of the constraint $D(\boldsymbol{X}_{i,\cdot}, \boldsymbol{H})^{1/2} \leq \alpha$ in problem (3)—$\hat{\boldsymbol{H}}$ might not represent the data points well. However, this is the price we pay to guarantee robustness.

If $\boldsymbol{H}_0$ is not ℓ-sparse, or equivalently $\|P_\ell^\perp(\boldsymbol{H}_0)\|_F > 0$, problem (3) obtains a sparse estimator $\hat{\boldsymbol{H}}$, which approximates $\boldsymbol{H}_0$. This is an example of model misspecification; and even in this case, Problem (3) leads to an estimator that is weakly and strongly robust. In Theorem 3, the quantity $D(\boldsymbol{H}_0, \tilde{\boldsymbol{X}}_0)$ determines how close the underlying model is to the noiseless data; and it depends upon $\boldsymbol{H}_0$ and $\boldsymbol{X}_0$. Choosing $\boldsymbol{H}_0$ as in (4) results in a smaller value of $D(\boldsymbol{H}_0, \tilde{\boldsymbol{X}}_0)$ and improves our bounds. This constant can be zero which is a generalization of the separable case (Donoho and Stodden, 2004), where the underlying archetypes are assumed to be among noiseless data points. In addition, condition (10) ensures that the noise in the data is not too large and the underlying archetypes are suitably sparse — this suffices to derive a strong robustness guarantee for $\hat{\boldsymbol{H}}$.

In the special case where $\boldsymbol{H}_0$ is ℓ-sparse (i.e., $\|P_\ell^\perp(\boldsymbol{H}_0)\|_F = 0$) and the underlying model is separable (i.e., $D(\boldsymbol{H}_0, \tilde{\boldsymbol{X}}_0) = 0$) the results of Theorem 3 can be simplified as in Corollary 1.

Corollary 1. Let $\hat{\boldsymbol{H}}$ be the solution of problem (3). Under the assumption of Theorem 3 and assuming $\|P_\ell^\perp(\boldsymbol{H}_0)\|_F = D(\boldsymbol{H}_0, \tilde{\boldsymbol{X}}_0) = 0$, we have the following:

1. (Weak Robustness)

$$\mathcal{L}(\boldsymbol{H}_0, \hat{\boldsymbol{H}})^{1/2} \leq [4mk\kappa(\boldsymbol{H}_0) + (1+\sqrt{2})\sqrt{k}(k + k^{3/2})] \max_{i\in[m]} \|\boldsymbol{Z}_{i,\cdot}\|_2$$

2. (Strong Robustness) If $[(k + k^{3/2}) + 2mk] \max_{i\in[m]} \|\boldsymbol{Z}_{i,\cdot}\|_2 \leq \sigma_{\min}(\boldsymbol{H}_0)/(6\sqrt{k})$,

$$\mathcal{L}(\hat{\boldsymbol{H}}, \boldsymbol{H}_0)^{1/2} \leq [7\kappa(\boldsymbol{H}_0)(k + k^{3/2}) + 2(1+\sqrt{2})k^{3/2}m] \max_{i\in[m]} \|\boldsymbol{Z}_{i,\cdot}\|_2.$$

Considering that $m > k$, the results in Corollary 1 can be summarized as

$$\mathcal{L}(\boldsymbol{H}_0, \hat{\boldsymbol{H}})^{1/2} = \mathcal{O}(mk\delta) \quad \text{and} \quad \mathcal{L}(\hat{\boldsymbol{H}}, \boldsymbol{H}_0)^{1/2} = \mathcal{O}(mk^{3/2}\delta)$$

where $\delta = \max_{i\in[m]} \|\boldsymbol{Z}_{i,\cdot}\|_2$ (assuming $\kappa(\boldsymbol{H}_0)$ is constant). This shows the bound for the strong robustness quantity $\mathcal{L}(\hat{\boldsymbol{H}}, \boldsymbol{H}_0)^{1/2}$ is larger. This observation can be explained as follows: strong robustness bounds the distance between each recovered archetype and the underlying archetypes, and as the location of each recovered archetype is unknown and uncertain, the strong robustness quantity deals with more uncertainty

compared to weak robustness and is harder to bound.

Comparison with Javadi and Montanari (2019): Although Javadi and Montanari (2019) do not consider sparsity in their formulation, it is insightful to compare our results with theirs. The results of Javadi and Montanari (2019) are valid under a specific uniqueness assumption on the model—this matches with our assumption in the separable case. Therefore, to compare our results we consider the case $\beta = D(\boldsymbol{H}_0, \tilde{\boldsymbol{X}}_0) = 0$ as in Corollary 1. In this regime, the result of Javadi and Montanari (2019) is similar to the first part of the Corollary 1 without any sparsity guarantee on the solution and with different coefficients. They do not provide results similar to the second part of the corollary. Moreover, the bound of Javadi and Montanari (2019) is $\mathcal{L}(\boldsymbol{H}_0, \hat{\boldsymbol{H}})^{1/2} = \mathcal{O}(k^{9/4}\delta)$ (assuming other parameters in their bound are constant) which is loose compared to our bound if $m = \mathcal{O}(k^{5/4})$. This shows that our results are tighter when the number of data points is small. Admittedly, the uniqueness assumption of Javadi and Montanari (2019) is more general than the separable case, however, this assumption is difficult to verify except for very simple cases, as discussed by the authors.

3.1 The penalized formulation

Theorem 3 presents robustness guarantees for the constrained SAA problem (3). From an algorithmic viewpoint however, the penalized form:

$$\hat{\boldsymbol{H}}_\lambda \in \underset{\boldsymbol{H} \in \mathbb{R}^{k\times n}_{\geq 0}}{\operatorname{argmin}} \quad D(\boldsymbol{X}, \boldsymbol{H}) + \lambda D(\boldsymbol{H}, \boldsymbol{X}) \quad \text{s.t.} \quad \|\boldsymbol{H}\|_0 \leq \ell \tag{12}$$

is more appealing, and we propose algorithms for this penalized form. In (12), $D(\boldsymbol{X}, \boldsymbol{H})$ is the data fidelity term, $D(\boldsymbol{H}, \boldsymbol{X})$ is the regularization term and λ is the regularization parameter. In fact, $\lambda = 0$ is equivalent to setting $\alpha = 0$ in problem (3) (which can lead to an infeasible problem) and $\lambda \to \infty$ is equivalent to removing the data fidelity constraint all together.
We show robustness properties of estimator (12). Proposition 1 establishes both weak (9) and strong (11) robustness. For simplicity, we consider the separable case ($D(\boldsymbol{H}_0, \tilde{\boldsymbol{X}}_0) = 0$) where $\boldsymbol{H}_0$ is sparse ($\|P^\perp_\ell(\boldsymbol{H}_0)\|_F = 0$).

Proposition 1 (The penalized formulation (12))**.** Let $\hat{\boldsymbol{H}}_\lambda$ be a solution to problem (12). Suppose the assumptions of Theorem 3 hold; and in addition $D(\boldsymbol{H}_0, \tilde{\boldsymbol{X}}_0) = \|P^\perp_\ell(\boldsymbol{H}_0)\|_F = 0$. There exist constants[5] $c^1_\lambda, c^2_\lambda, c^3_\lambda$ depending on $m, k, \kappa(\boldsymbol{H}_0), \lambda$ such that $c^1_\lambda, c^2_\lambda, c^3_\lambda \to \infty$ as $\lambda \to 0$ or $\lambda \to \infty$; and the following holds:

$$\mathcal{L}(\boldsymbol{H}_0, \hat{\boldsymbol{H}}_\lambda)^{1/2} \leq c^1_\lambda \max_{i\in[m]} \|\boldsymbol{Z}_{i,.}\|_2. \tag{13}$$

Moreover, if

$$c^3_\lambda \max_{i\in[m]} \|\boldsymbol{Z}_{i,.}\|_2 \leq c_7, \tag{14}$$

then

$$\mathcal{L}(\hat{\boldsymbol{H}}_\lambda, \boldsymbol{H}_0)^{1/2} \leq c^2_\lambda \max_{i\in[m]} \|\boldsymbol{Z}_{i,.}\|_2. \tag{15}$$

[5]The values of the constants can be found in the appendix, Section D.4

Note that $c_\lambda^1, c_\lambda^2 \to \infty$ as $\lambda \to 0$ or $\lambda \to \infty$—hence, there is no trivial value of λ that guarantees robustness. In fact, if we have $\lambda = 0$ (this is equivalent to normal NMF), the archetypes need not be close to the data (which is approximated by the underlying archetypes)—they can be far from the underlying archetypes and robustness is not guaranteed. If $\lambda \to \infty$, we do not reduce the recovery error ($D(\boldsymbol{X}, \boldsymbol{H})$) which does not result in robustness. This shows the usefulness of using AA as a regularization term.

Remark 3. Results for the penalized form of AA as in Proposition 1 are not discussed in Javadi and Montanari (2019), as far as we can tell.

4 SAA Algorithm

In this section, we propose algorithms to obtain good (feasible) solutions to the penalized SAA problem (12). In Section 4.1, we present a block coordinate descent method and derive its convergence guarantees. As problem (12) is nonconvex, Section 4.2 discusses an initialization scheme based on MIP techniques to keep away from suboptimal solutions. To further improve the quality of the solution obtained from the block coordinate method, we present a heuristic local search algorithm in Section 4.3.

4.1 A Block Coordinate Algorithm

We rewrite problem (12) as follows:

$$\begin{aligned} \min_{\boldsymbol{W}, \tilde{\boldsymbol{W}}, \boldsymbol{H}} \quad & \Psi(\boldsymbol{W}, \tilde{\boldsymbol{W}}, \boldsymbol{H}) := \|\boldsymbol{X} - \boldsymbol{W}\boldsymbol{H}\|_F^2 + \lambda \|\boldsymbol{H} - \tilde{\boldsymbol{W}}\boldsymbol{X}\|_F^2 \\ \text{s.t.} \quad & \boldsymbol{H} \geq \boldsymbol{0}, \boldsymbol{W} \geq 0, \tilde{\boldsymbol{W}} \geq \boldsymbol{0} \\ & \boldsymbol{W}\boldsymbol{1}_k = \boldsymbol{1}_m, \tilde{\boldsymbol{W}}\boldsymbol{1}_m = \boldsymbol{1}_k, \|\boldsymbol{H}\|_0 \leq \ell. \end{aligned} \tag{16}$$

We propose a proximal gradient based block coordinate descent algorithm (Xu and Yin, 2017) for (16). We first note that the gradient of the objective function is Lipschitz for every block $(\boldsymbol{W}, \tilde{\boldsymbol{W}}, \boldsymbol{H})$, that is,

$$\|\nabla_{\boldsymbol{H}} \Psi(\boldsymbol{H}_1, \boldsymbol{W}, \tilde{\boldsymbol{W}}) - \nabla_{\boldsymbol{H}} \Psi(\boldsymbol{H}_2, \boldsymbol{W}, \tilde{\boldsymbol{W}})\|_F \leq L_1(\boldsymbol{W}) \|\boldsymbol{H}_1 - \boldsymbol{H}_2\|_F,$$

$$\|\nabla_{\boldsymbol{W}} \Psi(\boldsymbol{H}, \boldsymbol{W}_1, \tilde{\boldsymbol{W}}) - \nabla_{\boldsymbol{W}} \Psi(\boldsymbol{H}, \boldsymbol{W}_2, \tilde{\boldsymbol{W}})\|_F \leq L_2(\boldsymbol{H}) \|\boldsymbol{W}_1 - \boldsymbol{W}_2\|_F$$

$$\|\nabla_{\tilde{\boldsymbol{W}}} \Psi(\boldsymbol{H}, \boldsymbol{W}, \tilde{\boldsymbol{W}}_1) - \nabla_{\tilde{\boldsymbol{W}}} \Psi(\boldsymbol{H}, \boldsymbol{W}, \tilde{\boldsymbol{W}}_2)\|_F \leq L_3(\boldsymbol{X}) \|\tilde{\boldsymbol{W}}_1 - \tilde{\boldsymbol{W}}_2\|_F$$

where $L_1(\boldsymbol{W}) = 2(\lambda + \sigma_{\max}(\boldsymbol{W})^2)$, $L_2(\boldsymbol{H}) = 2\max\{\sigma_{\max}(\boldsymbol{H})^2, \varepsilon\}$ and $L_3(\boldsymbol{X}) = 2\lambda\sigma_{\max}(\boldsymbol{X})^2$ for any fixed $\varepsilon > 0$. Our algorithm follows the block proximal update of Xu and Yin (2017). Specifically, for step size

values $1/2L_1(\boldsymbol{W}_j), 1/2L_2(\boldsymbol{H}_j), 1/2L_3(\boldsymbol{X})$, at iteration j we perform the following updates:

$$\boldsymbol{H}_{j+1} = \underset{\substack{\boldsymbol{H}\geq 0 \\ \|\boldsymbol{H}\|_0 \leq \ell}}{\operatorname{argmin}} \left\| \boldsymbol{H} - \left(\boldsymbol{H}_j - \frac{1}{2L_1(\boldsymbol{W}_j)} \nabla_{\boldsymbol{H}} \Psi(\boldsymbol{H}_j, \boldsymbol{W}_j, \tilde{\boldsymbol{W}}_j) \right) \right\|_F^2 \tag{17}$$

$$\boldsymbol{W}_{j+1} = \underset{\substack{\boldsymbol{W}\geq 0 \\ \boldsymbol{W}\mathbf{1}_k = \mathbf{1}_m}}{\operatorname{argmin}} \left\| \boldsymbol{W} - \left(\boldsymbol{W}_j - \frac{1}{2L_2(\boldsymbol{H}_{j+1})} \nabla_{\boldsymbol{W}} \Psi(\boldsymbol{H}_{j+1}, \boldsymbol{W}_j, \tilde{\boldsymbol{W}}_j) \right) \right\|_F^2 \tag{18}$$

$$\tilde{\boldsymbol{W}}_{j+1} = \underset{\substack{\tilde{\boldsymbol{W}}\geq 0 \\ \tilde{\boldsymbol{W}}\mathbf{1}_m = \mathbf{1}_k}}{\operatorname{argmin}} \left\| \tilde{\boldsymbol{W}} - \left(\tilde{\boldsymbol{W}}_j - \frac{1}{2L_3(\boldsymbol{X})} \nabla_{\tilde{\boldsymbol{W}}} \Psi(\boldsymbol{H}_{j+1}, \boldsymbol{W}_{j+1}, \tilde{\boldsymbol{W}}_j) \right) \right\|_F^2 . \tag{19}$$

After a sweep across the updates (17), (18) and (19) the objective decreases:

$$\Psi(\boldsymbol{H}_{j+1}, \boldsymbol{W}_{j+1}, \tilde{\boldsymbol{W}}_{j+1}) \leq \Psi(\boldsymbol{H}_j, \boldsymbol{W}_j, \tilde{\boldsymbol{W}}_j).$$

Algorithm 1 summarizes the above procedure, where $P_{\text{simplex}}(\boldsymbol{W})$ projects each row of $\boldsymbol{W}$ onto the unit simplex. See Duchi et al. (2008) for an efficient algorithm to calculate P_{simplex}. Before proceeding to the theoretical analysis of Algorithm 1, we need to define stationarity.

Definition 2. We say a point $\boldsymbol{\theta}^* = (\boldsymbol{H}^*, \boldsymbol{W}^*, \tilde{\boldsymbol{W}}^*)$ is stationary for problem (16) if update rules (17), (18) and (19) initialized with $\boldsymbol{\theta}^*$ result in the same solution $\boldsymbol{\theta}^*$.

Remark 4. Definition 2 is a generalization of the notion of L-stationarity by Beck and Eldar (2013) to the case of the block proximal method. Moreover, Definition 2 presents a necessary condition for optimality of problem (16).

Algorithm 1: SparseAA$(\boldsymbol{H}_0, \boldsymbol{W}_0, \tilde{\boldsymbol{W}}_0, \lambda)$

$j = 0$
while *not converged* **do**
  $\boldsymbol{\pi} = \boldsymbol{H}_j - [1/L_1(\boldsymbol{W}_j)](-\boldsymbol{W}_j^T[\boldsymbol{X} - \boldsymbol{W}_j\boldsymbol{H}_j] + \lambda[\boldsymbol{H}_j - \tilde{\boldsymbol{W}}_j\boldsymbol{X}])$
  $\boldsymbol{H}_{j+1} = P_\ell\left(\max\{\boldsymbol{\pi}, 0\}\right)$
  $\boldsymbol{W}_{j+1} = P_{\text{simplex}}(\boldsymbol{W}_j - [1/L_2(\boldsymbol{H}_{j+1})](\boldsymbol{X} - \boldsymbol{W}_j\boldsymbol{H}_{j+1})\boldsymbol{H}_{j+1}^T)$
  $\tilde{\boldsymbol{W}}_{j+1} = P_{\text{simplex}}(\tilde{\boldsymbol{W}}_j - \lambda[1/L_3(\boldsymbol{X})](\boldsymbol{H}_{j+1} - \tilde{\boldsymbol{W}}_j\boldsymbol{X})\boldsymbol{X}^T)$
  $j = j + 1$
end

In Theorem 4, first we show that problem (16) satisfies the convergence conditions of Xu and Yin (2017) and therefore Algorithm 1 converges. Then, we show that the limit point of Algorithm 1 is a stationary point as in Definition 2.

Theorem 4. Suppose $\lambda > 0$ and let $\{(\boldsymbol{H}_j, \boldsymbol{W}_j, \tilde{\boldsymbol{W}}_j)\}_{j\geq 1}$ be the sequence of solutions produced by Algorithm 1. The following results hold:

1. The sequence $(\boldsymbol{H}_j, \boldsymbol{W}_j, \tilde{\boldsymbol{W}}_j)$ converges to a feasible solution $(\boldsymbol{H}^*, \boldsymbol{W}^*, \tilde{\boldsymbol{W}}^*)$ of (16)
2. Let

$$\boldsymbol{T} = \max\{0, \boldsymbol{H}^* - [1/L_1(\boldsymbol{W}^*)](-\boldsymbol{W}^{*T}[\boldsymbol{X} - \boldsymbol{W}^*\boldsymbol{H}^*] + \lambda[\boldsymbol{H}^* - \tilde{\boldsymbol{W}}^*\boldsymbol{X}])\} \tag{20}$$

and if $\|\boldsymbol{T}\|_0 > \ell$, assume $\boldsymbol{T}^\sharp_\ell > \boldsymbol{T}^\sharp_{\ell+1}$ where $\boldsymbol{T}^\sharp_1, \cdots, \boldsymbol{T}^\sharp_{kn}$ are entries of $\boldsymbol{T}$ ordered from largest to smallest. Then, the limit point $(\boldsymbol{H}^*, \boldsymbol{W}^*, \tilde{\boldsymbol{W}}^*)$ is a stationary point of (16).

The condition on $\boldsymbol{T}$ above is needed to make sure $P_\ell(\boldsymbol{T})$ is unique. Otherwise, there will be multiple possible solutions for update rule (17) when initialized with $\boldsymbol{\theta}^*$. Note however, that this condition is quite mild, and is unlikely to be violated in practice (due to noise in data).

4.2 Initialization via Mixed Integer Programming (MIP)

Problem (16) is not convex, hence having a good initialization is critical to obtain a high-quality local solution. To initialize Algorithm 1, we use a continuation method as discussed below. We first obtain a solution to (16) for a large value of λ — this leads to the following problem:

$$\min_{\tilde{\boldsymbol{W}}, \boldsymbol{H}} \quad \|\boldsymbol{H} - \tilde{\boldsymbol{W}}\boldsymbol{X}\|_F^2 \quad \text{s.t.} \quad \boldsymbol{H} \geq 0;\ \tilde{\boldsymbol{W}} \geq 0;\ \tilde{\boldsymbol{W}}\mathbf{1}_m = \mathbf{1}_k;\ \|\boldsymbol{H}\|_0 \leq \ell. \tag{21}$$

Note that Problem (21) has a convex quadratic objective in $\tilde{\boldsymbol{W}}, \boldsymbol{H}$ and the only source of nonconvexity is the cardinality constraint on $\boldsymbol{H}$. This is a Mixed Integer Quadratic Problem (MIQP)—while these problems are computationally difficult in the worst-case, recent work (Bertsimas and Van Parys, 2020; Bertsimas et al., 2016; Hazimeh et al., 2020)[6] has shown that they can be solved to near-optimality using specialized algorithms for large-scale problems. Thusly motivated, we present new algorithms to solve (21) to optimality. Once we obtain a near-optimal solution to (21), we decrease λ and use Algorithm 1 to obtain a feasible solution for (21). We continue this process by successively decreasing λ, and using a solution obtained from the previous (larger) value of λ to initialize Algorithm 1.
To formulate (21) as a MIQP, we first show that the solution of this problem is bounded.

Proposition 2. If $(\boldsymbol{H}^*, \tilde{\boldsymbol{W}}^*)$ is an optimal solution to (21), we have the following bound on $\boldsymbol{H}^*$:

$$\|\boldsymbol{H}^*\|_F^2 \leq k\left(\max_{u\in[m]} \|\boldsymbol{X}_{u,.}\|_2 + \sqrt{k}\min_{u\in[m]} \|\boldsymbol{X}_{u,.}\|_2\right)^2 := b.$$

Based on Proposition 2, we reformulate problem (21) as the following MIQP:

$$\begin{aligned} \min_{\boldsymbol{H}, \tilde{\boldsymbol{W}}, \boldsymbol{Z}} \quad & \|\boldsymbol{H} - \tilde{\boldsymbol{W}}\boldsymbol{X}\|_F^2 \\ \text{s.t.} \quad & \boldsymbol{H} \geq 0, \tilde{\boldsymbol{W}} \geq 0, \boldsymbol{Z} \in \{0,1\}^{k\times n} \\ & \tilde{\boldsymbol{W}}\mathbf{1}_m = \mathbf{1}_k, \sum_{i,j} \boldsymbol{Z}_{i,j} \leq \ell \\ & \boldsymbol{H}_{i,j} \leq \sqrt{b}\boldsymbol{Z}_{i,j} \quad \forall (i,j) \in [k]\times[n], \end{aligned} \tag{22}$$

where b is as defined in Proposition 2. Note that the last constraint in (22) does not change the optimal solution because of Proposition 2.
Problem (22) can be formulated and solved (to optimality) by off-the-shelf MIP solvers (e.g., Gurobi, CPLEX, GLPK) for small/moderate instances—however, the runtimes become long as soon as $m, n \sim 100$ or so. With

[6] Note that problem (16) does not admit a MIQP representation, unlike (21).

efficiency in mind, we present a cutting plane approach to obtain a certifiably optimal solution[7]. To this end, we rewrite (22) as a binary convex optimization problem:

$$\min_{\boldsymbol{Z}} \quad F(\boldsymbol{Z}) \quad \text{s.t.} \quad \boldsymbol{Z} \in \{0,1\}^{k\times n}, \quad \sum_{(i,j)\in[k]\times[n]} \boldsymbol{Z}_{i,j} \leq \ell \tag{23}$$

where, for any $\boldsymbol{Z} \in [0,1]^{k\times n}$, the objective $F(\boldsymbol{Z})$ is implicitly defined as the solution to the following convex optimization problem:

$$\begin{aligned} F(\boldsymbol{Z}) = \min_{\boldsymbol{H},\tilde{\boldsymbol{W}}} \quad & \|\boldsymbol{H} - \tilde{\boldsymbol{W}}\boldsymbol{X}\|_F^2 \\ \text{s.t.} \quad & \boldsymbol{H} \geq 0, \quad \tilde{\boldsymbol{W}} \geq 0, \quad \tilde{\boldsymbol{W}}\mathbf{1}_m = \mathbf{1}_k \\ & \boldsymbol{H}_{i,j} \leq \sqrt{b}\boldsymbol{Z}_{i,j} \quad \forall (i,j) \in [k]\times[n]. \end{aligned} \tag{24}$$

Proposition 3 presents some properties of the function $F(\boldsymbol{Z})$.

Proposition 3. Let $(\boldsymbol{H}^*, \tilde{\boldsymbol{W}}^*)$ be an optimal solution to the minimization problem (24). Then, we have the following:

1. The function $\boldsymbol{Z} \mapsto F(\boldsymbol{Z})$ is convex on $\boldsymbol{Z} \in [0,1]^{k\times n}$.
2. $\boldsymbol{G} = -\sqrt{b}\boldsymbol{\Lambda}$ is a subgradient of $F(\boldsymbol{Z})$, where for $(i,j) \in [k]\times[n]$,

$$\boldsymbol{\Lambda}_{i,j} = \begin{cases} 2(\tilde{\boldsymbol{W}}^*\boldsymbol{X} - \boldsymbol{H}^*)_{i,j} & \text{if } (\tilde{\boldsymbol{W}}^*\boldsymbol{X} - \boldsymbol{H}^*)_{i,j} > 0 \\ 0 & \text{if } (\tilde{\boldsymbol{W}}^*\boldsymbol{X} - \boldsymbol{H}^*)_{i,j} \leq 0 \end{cases}.$$

The function $F(\boldsymbol{Z})$ is convex and subdifferentiable. Specifically, for any $\boldsymbol{Z}_0 \in \mathbb{R}^{k\times n}$ and any subgradient $\boldsymbol{G}_0 \in \partial F(\boldsymbol{Z}_0)$,

$$F(\boldsymbol{Z}) \geq F(\boldsymbol{Z}_0) + \langle \boldsymbol{G}_0, \boldsymbol{Z} - \boldsymbol{Z}_0 \rangle. \tag{25}$$

MIP Algorithm: We present an outer approximation algorithm (Duran and Grossmann, 1986) to solve the binary convex program (23). This algorithm starts from an initial point $\boldsymbol{Z}_0$ which is feasible for (23). At iteration t, using a list of subgradient-based inequalities (25), we consider the following piecewise linear lower bound of $F(\boldsymbol{Z})$:

$$F(\boldsymbol{Z}) \geq \max\{F(\boldsymbol{Z}_0) + \langle \boldsymbol{G}_0, \boldsymbol{Z} - \boldsymbol{Z}_0 \rangle, \cdots, F(\boldsymbol{Z}_{t-1}) + \langle \boldsymbol{G}_{t-1}, \boldsymbol{Z} - \boldsymbol{Z}_{t-1} \rangle\} \tag{26}$$

where $\boldsymbol{Z}_0, \cdots, \boldsymbol{Z}_{t-1}$ are feasible for Problem (23) and $\boldsymbol{G}_t \in \partial F(\boldsymbol{Z}_t)$ for all t. We define $\boldsymbol{Z}_t$ as a minimizer of the right hand side of (26) under the constraints of Problem (23). Mathematically, this can be written as

[7] That is, along with delivering a feasible solution, we also present a dual bound (aka lower bound) on the optimal objective value of (22).

a Mixed Integer Linear Program (MILP)

$$
\begin{aligned}
(\boldsymbol{Z}_t, \eta_t) \in \underset{\boldsymbol{Z},\eta}{\operatorname{argmin}} \quad & \eta \\
\text{s.t.} \quad & \boldsymbol{Z} \in \{0,1\}^{k\times n}, \eta \in \mathbb{R} \\
& \eta \geq F(\boldsymbol{Z}_i) + \langle \boldsymbol{G}_i, \boldsymbol{Z} - \boldsymbol{Z}_i \rangle \quad i = 0, \cdots, t-1 \\
& \sum_{(i,j)\in[k]\times[n]} \boldsymbol{Z}_{i,j} \leq \ell.
\end{aligned} \tag{27}
$$

The optimal objective value of (27) is a lower bound (aka dual bound) for (23); and these lower bounds improve as the iterations progress (i.e., t increases). As the feasible set of problem (23) is finite, after finitely many iterations t, an optimal solution to (23) is found. The optimality gap (OG) of the outer approximation can be calculated as $\text{OG} = (\text{UB} - \text{LB})/\text{UB}$ where LB is the current (and the best) lower bound achieved by the piecewise approximation and UB is the best upper bound for (23) found so far. We summarize the procedure in Algorithm 2 where, 'tol' denotes a pre-specified tolerance level.

Algorithm 2: An outer approximation method to solve (23)

```
t = 1
while OG > tol do
    (Z_t, η_t) are solutions of (27).
    F_best = min_{i=0,···,t−1} F(Z_i)
    OG = (F_best − η_t)/F_best
    t = t + 1
end
```

The optimization Problem (24) is convex in $(\boldsymbol{H}, \tilde{\boldsymbol{W}})$ and we use an accelerated proximal gradient method (Beck and Teboulle, 2009) to solve it.

Note that our proposed algorithm is different from that of Bertsimas and Van Parys (2020) who consider the sparse linear regression problem with an additional ridge regularization. Bertsimas and Van Parys (2020) use an outer approximation algorithm to solve an equivalent convex integer program, with an explicit closed-form expression. In contrast, in our work, the function $F(\boldsymbol{Z})$ is given (implicitly) by the solution to an optimization problem. Furthermore, our formulation of (22) uses the binary variable as a linear constraint — this is different from Bertsimas and Van Parys (2020) where, the binary variable appears as a nonlinear expression within the objective function.

Numerical results are presented in Section 5.

4.3 Improving Algorithm 1 with Local Search

The block CD method (Algorithm 1) is guaranteed to deliver a stationary point for Problem (16). We present some heuristics to improve the solution quality based on local search, drawing inspiration from the work of Beck and Eldar (2013); Hazimeh and Mazumder (2020) who use local search ideas for a different problem.

Once we are at a stationary point delivered by Algorithm 1, our local search algorithm swaps a coordinate in the support of $\boldsymbol{H}$ with a coordinate from outside the support. That is, a nonzero coordinate of $\boldsymbol{H}$ is set

to zero and a zero coordinate of $\boldsymbol{H}$ is allowed to become nonzero. Then, the optimization is solely done on the coordinate entering the support. If this optimization leads to a lower objective value, we retain the new support. Mathematically, let $(\boldsymbol{H}^*, \boldsymbol{W}^*, \tilde{\boldsymbol{W}}^*)$ be a feasible solution of problem (16) with $\|\boldsymbol{H}^*\|_0 = \ell$. This solution can be an output of Algorithm 1. Suppose $(i_1, j_1) \in S(\boldsymbol{H}^*)$ (here, $S(\boldsymbol{H}^*)$ is the support of $\boldsymbol{H}^*$) leaves the support and $(i_2, j_2) \notin S(\boldsymbol{H}^*)$ enters the support. We perform an optimization on the coordinate (i_2, j_2) of $\boldsymbol{H}$ to decide whether this change in the support improves the objective function. Let $\boldsymbol{E}^{i_1,j_1}$ be a matrix with all entries equal to zero except coordinate (i_1, j_1) equal to one. We denote $\boldsymbol{H} = \boldsymbol{H}^* - \boldsymbol{H}^*_{i_1,j_1}\boldsymbol{E}^{i_1,j_1}$ as the solution with coordinate (i_1, j_1) removed from the support. The candidate solution with the new support has the form $\boldsymbol{H} + t\boldsymbol{E}^{i_2,j_2}$ for $t \geq 0$. This leads to the following problem:

$$\begin{aligned} \min_{\boldsymbol{W},\tilde{\boldsymbol{W}},t} \quad & \|\boldsymbol{X} - \boldsymbol{W}\boldsymbol{H} - t\boldsymbol{W}\boldsymbol{E}^{i_2,j_2}\|_F^2 + \lambda\|\boldsymbol{H} + t\boldsymbol{E}^{i_2,j_2} - \tilde{\boldsymbol{W}}\boldsymbol{X}\|_F^2 \\ \text{s.t.} \quad & t \geq 0, \boldsymbol{W} \geq 0, \tilde{\boldsymbol{W}} \geq 0 \\ & \boldsymbol{W}\mathbf{1}_k = \mathbf{1}_m, \tilde{\boldsymbol{W}}\mathbf{1}_m = \mathbf{1}_k \end{aligned} \tag{28}$$

where, we are optimizing over $(\boldsymbol{W}, \tilde{\boldsymbol{W}}, t)$ for a given (i_2, j_2) and (i_1, j_1). For a fixed value of t, the optimal values of $\boldsymbol{W}$ and $\tilde{\boldsymbol{W}}$ in (28) are given as

$$\hat{\boldsymbol{W}} \in \operatorname*{argmin}_{\boldsymbol{W}} \quad \|\boldsymbol{X} - \boldsymbol{W}(\boldsymbol{H} + t\boldsymbol{E}^{i_2,j_2})\|_F^2 \quad \text{s.t.} \quad \boldsymbol{W} \geq 0; \quad \boldsymbol{W}\mathbf{1}_k = \mathbf{1}_m \tag{29}$$

$$\hat{\tilde{\boldsymbol{W}}} \in \operatorname*{argmin}_{\tilde{\boldsymbol{W}}} \quad \|\boldsymbol{H} + t\boldsymbol{E}^{i_2,j_2} - \tilde{\boldsymbol{W}}\boldsymbol{X}\|_F^2 \quad \text{s.t.} \quad \tilde{\boldsymbol{W}} \geq 0; \quad \tilde{\boldsymbol{W}}\mathbf{1}_m = \mathbf{1}_k. \tag{30}$$

Problems (29) and (30) are convex and can be efficiently solved by standard first order methods such as proximal gradient. Note that these first order methods also benefit from warm-starts available from prior estimates of $(\boldsymbol{W}, \tilde{\boldsymbol{W}})$. Once $\boldsymbol{W}$ and $\tilde{\boldsymbol{W}}$ are updated by (29) and (30), the value of t that minimizes (28) with $\boldsymbol{W} = \hat{\boldsymbol{W}}$ and $\tilde{\boldsymbol{W}} = \hat{\tilde{\boldsymbol{W}}}$ is:

$$t = \max\left\{\frac{\sum_{r=1}^m \boldsymbol{U}_{r,j_2}\boldsymbol{W}_{r,i_2} - \lambda\boldsymbol{V}_{i_2,j_2}}{\lambda + \|\boldsymbol{W}_{\cdot,i_2}\|_2^2}, 0\right\} \tag{31}$$

where $\boldsymbol{U} = \boldsymbol{X} - \boldsymbol{W}\boldsymbol{H}$ and $\boldsymbol{V} = \boldsymbol{H} - \tilde{\boldsymbol{W}}\boldsymbol{X}$.

We use an alternating optimization scheme where the three updates (29),(30) and (31) are performed sequentially until convergence. These updates result in a descent method by construction, though there may not be a strict decrease in the objective value (in which case, the swap may not result in a better solution). In the discussion above, we assumed a fixed pair of indices (i_1, j_1) and (i_2, j_2). Ideally, we would like to try all possible choices of such indices and consider the one that leads to the maximal decrease in objective value (if any). As this is computationally intensive, we use a heuristic to select a suitable pair of indices. We choose (i_1, j_1) to be the smallest nonzero entry in $\boldsymbol{H}^*$:

$$(i_1, j_1) \in \operatorname*{argmin}_{(i,j)\in S(\boldsymbol{H}^*)} \boldsymbol{H}^*_{i,j}. \tag{32}$$

For the pair (i_2, j_2) from outside the current support of $\boldsymbol{H}$, we choose the coordinate of $\boldsymbol{H}^*$ that has the

Algorithm 3: A Local Search improvement for Algorithm 1

Initialize with $\boldsymbol{H}^*, \boldsymbol{W}^*, \tilde{\boldsymbol{W}}^*$.
while *not converged* **do**
  if $\|\boldsymbol{H}^*\|_0 < \ell$ **then**
    Choose (i_2, j_2) as in (33).
    $\boldsymbol{H} = \boldsymbol{H}^*$
  end
  else
    Choose (i_1, j_1) as in (32) and (i_2, j_2) as in (33).
    $\boldsymbol{H} = \boldsymbol{H}^* - \boldsymbol{H}^*_{i_1,j_1}\boldsymbol{E}^{i_1,j_1}$
  end
  while *not converged* **do**
    Update $\boldsymbol{W}$, $\tilde{\boldsymbol{W}}$ and t via (29), (30) and (31).
  end
  if $\Psi(\boldsymbol{H} + t\boldsymbol{E}^{i_2,j_2}, \boldsymbol{W}, \tilde{\boldsymbol{W}}) < \Psi(\boldsymbol{H}^*, \boldsymbol{W}^*, \tilde{\boldsymbol{W}}^*)$ **then**
    $\boldsymbol{H}^* = \boldsymbol{H} + t\boldsymbol{E}^{i_2,j_2}, \boldsymbol{W}^* = \boldsymbol{W}, \tilde{\boldsymbol{W}}^* = \tilde{\boldsymbol{W}}$
  end
end

smallest (most negative) gradient of the objective function:

$$(i_2, j_2) \in \underset{(i,j)\in[k]\times[n]\setminus S(\boldsymbol{H}^*)}{\operatorname{argmin}} \frac{\partial}{\partial \boldsymbol{H}_{i,j}} \Psi(\boldsymbol{H}^*, \boldsymbol{W}^*, \tilde{\boldsymbol{W}}^*). \tag{33}$$

The overall procedure of local search is shown in Algorithm 3.

5 Numerical Experiments

In this section, we discuss results of our numerical experiments on synthetic and real data and investigate how our framework performs in practice. Our experiments are done on a computer equipped with `Intel(R) Core(R) i7 6700HQ CPU @ 2.60GHz`, running `Microsoft(R) Windows(R) 10` and using 16GB of RAM. We have implemented all algorithms in `Julia` and we use `Gurobi(R)` to solve MILPs arising in our initialization scheme. An implementation of our framework in Julia is available at:

`https://github.com/kayhanbehdin/SparseAA`.

5.1 Synthetic Data

In this section, we consider synthetic data to validate the theory we developed in Sections 3 and 4; and gather further insights into the operating characteristics of SAA.

Dataset generation: The entries of $\boldsymbol{H}_0$ are drawn iid from Unif$[0, 1]$ and 20% of entries are set to zero at random to produce a matrix with at most $0.8nk$ nonzero entries. Independent of $\boldsymbol{H}_0$, the entries of $\boldsymbol{W}_0$ are drawn iid from Unif$[0, 1]$ and each row is normalized to sum to one. The noiseless data is $\boldsymbol{X}_0 = \boldsymbol{W}_0\boldsymbol{H}_0$ and the noisy data is produced as $\boldsymbol{X} = \max\{\boldsymbol{X}_0 + \boldsymbol{Z}, 0\}$ where entries of $\boldsymbol{Z}$ are from an independent Gaussian ensemble with mean zero and variance σ_z^2.

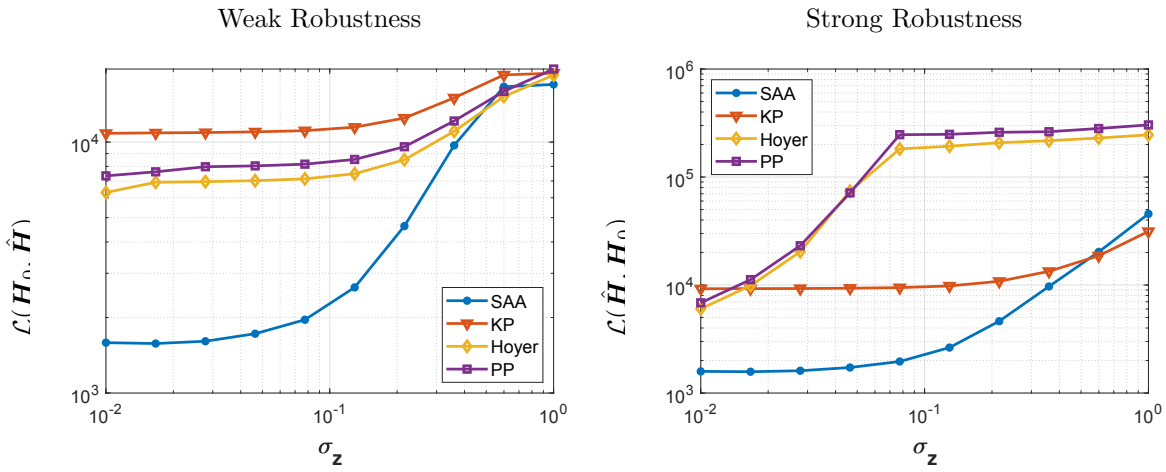


Figure 4: Figure comparing our method (SAA) with three other sparse NMF methods. We compare weak (left panel) and strong (right panel) robustness for varying values of σ_z on the synthetic data in Section 5.1.

5.1.1 Understanding robustness

We compare the performance of our algorithm with other sparse NMF algorithms in terms of robustness of the solution. We consider algorithms by Kim and Park (2007) (shown as KP), Peharz and Pernkopf (2012) (shown as PP) and Hoyer (2004) (shown as Hoyer). We set $m = 200, k = 15, n = 5000, \lambda = 1$. We consider two settings: First, we keep the sparsity level ℓ fixed and change the noise level σ_z; Second, we keep σ_z fixed and vary the sparsity level ℓ.

Robustness versus varying σ_z: First, we set $\ell/nk = 0.5$ and tune parameters for different algorithms to get solutions with $0.5nk$ nonzeros. Specifically, as KP considers an ℓ_1 regularized version of NMF, we start with a small value of the ℓ_1 regularization parameter and gradually increase it till we reach the target sparsity-level. PP uses an ℓ_0 constrained version of original NMF (without archetypal regularization) so we set the ℓ_0 sparsity level to ℓ. The sparsity constraint of Hoyer is set such that the result has ℓ nonzeros. We vary σ_z and plot the average value of weak ($\mathcal{L}(\boldsymbol{H}_0, \boldsymbol{H})$) and strong ($\mathcal{L}(\boldsymbol{H}, \boldsymbol{H}_0)$) robustness quantities. The results for this scenario are shown in Figure 4. As it can be seen, SAA almost always outperforms other algorithms in terms of strong and weak robustness of solutions. Specifically, the difference between SAA and other algorithms is most noticeable when the noise is small. This is expected as other algorithms in our experiments do not use any regularization that results in robustness. Moreover, solutions of SAA become less robust as noise is increased, as anticipated by Theorem 3 and Proposition 1. In addition, it is interesting to note that the weak robustness quantity in Figure 4 [left panel] is smaller than the strong robustness quantity in Figure 4 [right panel]. This is expected based on our discussions in Section 3.

Robustness versus varying ℓ: To compare the performance of different algorithms for varying values of ℓ, we do another set of experiments. We consider a setup similar to the previous experiment. However, we fix $\sigma_z = 0.1$ and change the value of ℓ (while keeping the underlying model sparsity fixed at $0.8nk$). This shows how well different algorithms can deal with misspecification in the underlying model. The results for this case are shown in Figure 5. We see that SAA outperforms other algorithms and provides the most robust solutions (both in terms of weak and strong robustness). In addition, we observe that as ℓ is decreased, the

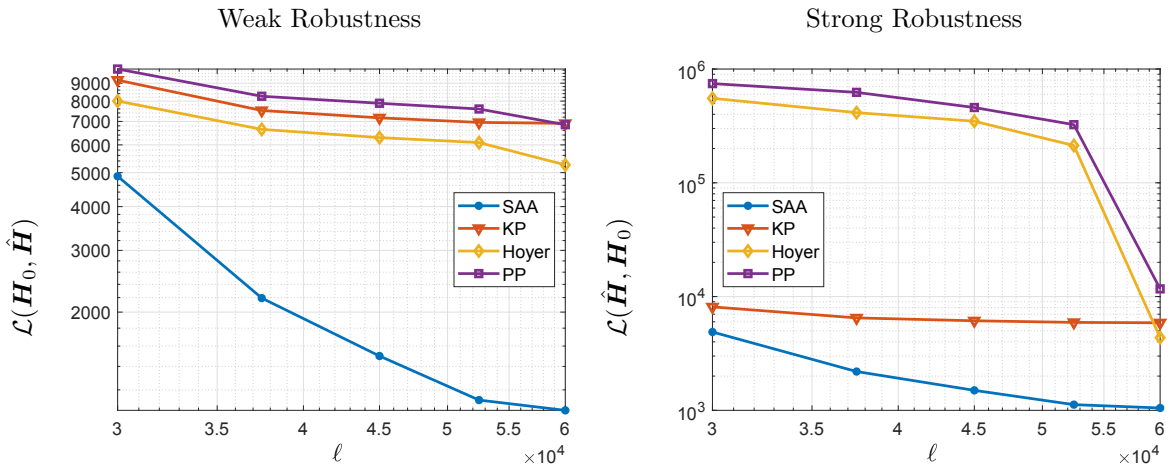


Figure 5: Comparison of robustness quantities for varying values of ℓ on the synthetic data in Section 5.1.

solution becomes less robust as anticipated by Theorem 3.

5.1.2 Usefulness of MIP-based Initialization and Local Search

We perform numerical experiments to show the usefulness of the initialization procedure in Section 4.2 and the local search scheme in Section 4.3. We show these algorithms improve upon a baseline initialization. We set $m = 200, k = 20, n = 12000$ and $\lambda = 1$ and vary σ_z and ℓ (see Table 1). We consider three cases:

(i) Algorithm 1 initialized with $\boldsymbol{H} = \boldsymbol{0}$ (shown as Zero in Table 1).

(ii) Algorithm 1 with MIP initialization (see Section 4.2) and using warm-start continuation over 8 values of λ on a logarithmic scale from 30 to 1. (Shown as SAA in Table 1).

(iii) Improving solution from (ii) with local search discussed in Section 4.3. This is denoted by SAA+LS in Table 1.

Algorithm 2 is limited to 20 minutes of maximum runtime and the best solution is returned. The average final cost function achieved by three methods explained above (over 5 independently generated datasets) are reported in Table 1. As it can be seen, our initialization scheme achieves a significantly lower objective value compared to the baseline (Zero) while being computationally feasible for such data size. In addition, our local search algorithm can improve the objective value as well as the support in a reasonable time (the number of changes in support for SAA+LS is reported in the parentheses). In our experiments, Algorithm 1 terminates in less than 15 seconds.

5.2 The Face Dataset (Samaria and Harter, 1994)

A classical application of sparse NMF is in face detection and recognition (Hoyer, 2004). The goal is to obtain a low-rank representation of a dataset of human faces under different lighting and shadow conditions and also different angles of photography. Hoyer (2004) show the effect of sparsity in finding such representations of the data. In particular, they show that sparse NMF leads to part-based representations where each factor represents one part of the face. Here, we are interested in finding the effect of AA as well as the combined effect of AA and sparsity. We use the AT&T database of faces (Samaria and Harter, 1994) which consists

σ_z	Method	$\ell/nk = 0.5$	$\ell/nk = 0.65$	$\ell/nk = 0.8$
0.01	Zero	47785	42123	37566
	SAA	32193	29766	27139
	SAA+LS	31867 (12.8)	29729 (13.4)	27091 (20.2)
0.1	Zero	69366	63776	59276
	SAA	54784	52253	49228
	SAA+LS	54465 (18.4)	52159 (19.8)	49131 (24)
0.5	Zero	448088	440974	434590
	SAA	426238	420134	416901
	SAA+LS	424776 (22.6)	419932 (26.8)	416628 (32)

Table 1: Comparison of zero initialization, SAA and SAA+LS in Section 5.1. The number in parentheses for SAA+LS shows the number of changes in the support after local search.

of 40 different people and 10 different photos of each person, 400 images in total. Each image is a grayscale 92×112 image, which is converted to a vector of length 10304. We then concatenate these 400 vectorized images into one matrix of size 400×10304 matrix. We consider $k = 25$ (following Hoyer (2004)) for this dataset and do the factorization based on problem (12) for different values of λ and ℓ. The estimated representations of the data (rows of resulting $\boldsymbol{H}$ which are reshaped into 92×112 images) are shown in Figure 6. We use the MIP initialization and continuation framework (over 8 values of λ) in Section 4.2. In the rest of this section, we discuss the connections between the robustness theory we developed and our numerical results.

We first explore the difference between AA and basic NMF. By comparing Figures 6 (a), (b), we deduce that as λ is increased, the resulting factors appear to become more similar to human faces—making it easier to recognize the people in the dataset. As λ is decreased, the results become more abstract and the images do not resemble human faces anymore. Considering that the theory we developed in Section 3 holds under mild assumptions, the representation achieved from Problem (12) with $\lambda = 0.4$ more likely corresponds to a robust solution—that is, the solution is closer to the underlying representation (we do not expect $\lambda = 0$ to result in robustness). Intuitively, we expect the underlying model that produces the face images to resemble the people in the dataset—this suggests that the solution achieved by $\lambda = 0.4$ is more robust.

As discussed by Hoyer (2004), adding sparsity to NMF produces part-based representations of the data. This can be also seen in our experimental results in Figure 6. By forcing the solution $\boldsymbol{H}$ to be sparse, we notice that each set of solutions (in Figures 6 (c), (d)) includes two groups of factors. The first group consists of complete faces (columns and rows 2 to 4 in each set) and the second group consists of parts of a face (the border factors). In fact, this can be interpreted as each face being a combination of an overall shape of a human face and additional details arising from different parts of the face. The factors that contain a complete face in Figures 6 (c), (d) represent the overall shape of a human face, while other factors represent different parts of a face, like the forehead, cheeks, eyes and also the background of images. Once again, considering that our theoretical development in Section 3 is valid in the sparse NMF case, we expect the factors recovered by $\lambda = 0.4$ to be more robust. Our hypothesis is the solution with $\lambda = 0.4$ is more robust as factors in Figure 6 (d) appear to more closely resemble human faces (compare the central columns in Figures 6 (c), (d)).

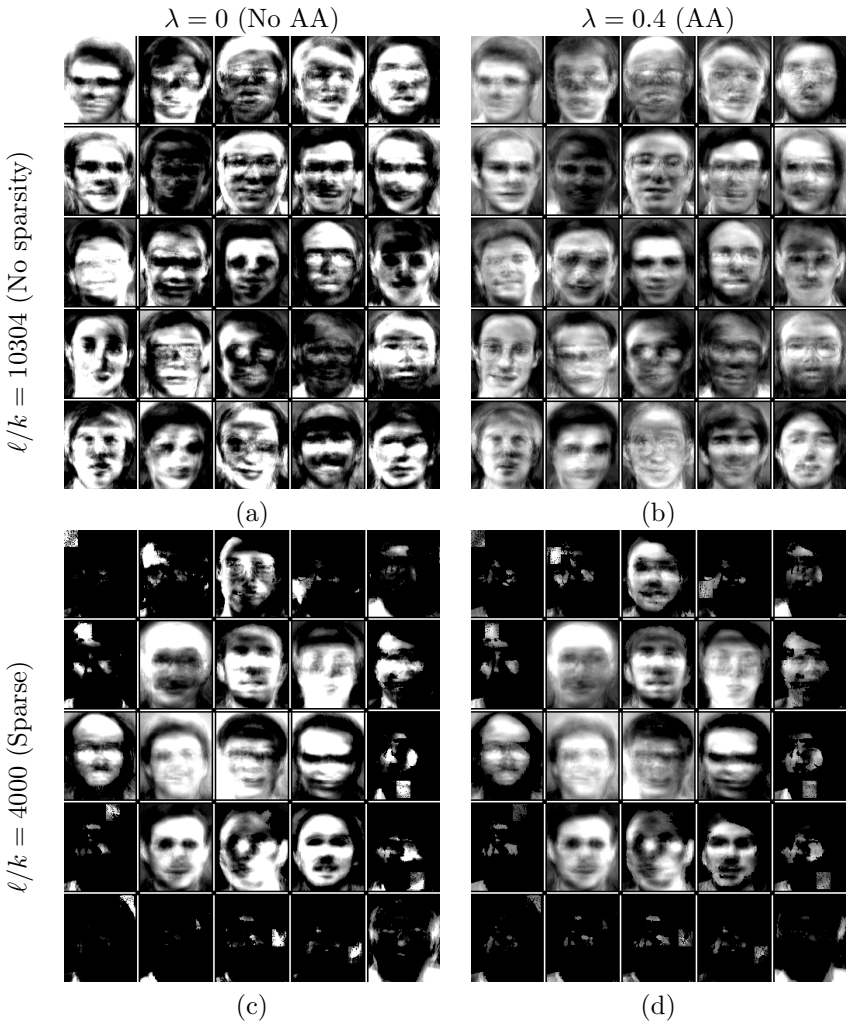


Figure 6: The resulting face images in Section 5.2: (a) $\lambda = 0$, $\ell/k = 10304$ (b) $\lambda = .4$, $\ell/k = 10304$ (c) $\lambda = 0$, $\ell/k = 4000$ (d) $\lambda = .4$, $\ell/k = 4000$

5.3 Cancer Gene Expression Example (Ramaswamy et al., 2001)

It is well-known that a primary goal of AA is to find a few representative points for a collection of data points. These representative points are useful in cluster analysis, where data points are put into a few clusters based on a suitable similarity/dissimilarity measure. In AA, each archetype (i.e., a row of $\boldsymbol{H}$) can be considered as a cluster center and data points are assigned to clusters based on their proximity to different archetypes. As a result, each row of matrix $\boldsymbol{H}$ (or each archetype) is considered as a center and each data point is assigned to the closest row of $\boldsymbol{H}$ (Mørup and Hansen, 2012).

An important application area of sparse NMF for clustering is in computational biology. Specifically, this problem has been considered by Kim and Park (2007) where the authors provide biological interpretations of sparsity in the context of NMF and do an extensive analysis of sparse NMF for microarray data. Here, we are interested in the clustering performance of our method. To this end, we consider a real dataset: the 14 Cancers Gene Expression dataset (Ramaswamy et al., 2001). This dataset consists of gene expression

data of 198 samples and 14 different types of tumors. There are 16,063 features in the dataset, however, the data is not nonnegative. Therefore, we use a trick introduced by Kim and Tidor (2003) to transform the data: each feature is divided into two new features where one contains nonnegative coordinates (and zero elsewhere) and the other one contains the absolute value of negative coordinates (and zero elsewhere). Consequently, this leads to 32,126 nonnegative features—the data matrix is given by $\boldsymbol{X} \in \mathbb{R}_{\geq 0}^{198 \times 32126}$. The rank of the factorization is 14, the number of different types of tumors in the dataset. The i-th data point ($i \in [198]$) belongs to the cluster j_i:

$$j_i = \underset{j \in [14]}{\operatorname{argmin}} \|\boldsymbol{X}_{i,\cdot} - \boldsymbol{H}_{j,\cdot}\|_2$$

where $\boldsymbol{H}$ is the resulting matrix of archetypes. To compare the performance of different algorithms (discussed below), we use two metrics, Purity and Entropy (Kim and Park, 2008). Let for $i \in [198]$, $j_i^* \in [14]$ denote the true cluster of point i and $j_i \in [14]$ denote the estimated cluster for the same point. Let m_r^u be the number of samples that belong to the true cluster u but are estimated to be in cluster r. Equivalently,

$$m_r^u = |\{i \in [m] : j_i^* = u, j_i = r\}|.$$

The metrics Purity and Entropy are defined as

$$\text{Purity} = \frac{1}{m} \sum_{r=1}^{k} \max_{u \in [k]} m_r^u \quad \text{and} \quad \text{Entropy} = -\frac{1}{m \log_2 k} \sum_{r=1}^{k} \sum_{u=1}^{k} m_r^u \log_2 \frac{m_r^u}{m_r}$$

where $m_r = \sum_{u=1}^{k} m_r^u$. A larger value of Purity and a smaller value of Entropy imply a better clustering performance.

The results of clustering performance of different algorithms is reported in Table 2. SAA is our proposed framework (we use continuation over 8 values of λ from 30 to 1 as in Section 4.2), AA is the algorithm proposed by Javadi and Montanari (2019) which does not enforce any sparsity. KP and PP are as introduced before. We also include Kmeans in our experiments as a baseline for the clustering performance. Hoyer did not provide interpretable results on this dataset and therefore is not included in the table. As it can be seen, among algorithms that enforce sparsity, SAA performs the best in terms of clustering. In fact, SAA is at par with Kmeans, while providing a solution that is two times more sparse. AA has the best clustering performance, however, it fails to provide a sparse solution. Other algorithms provide sparse solutions, but their clustering performance is not as good as SAA. In our experiments in this section, all NMF-based methods terminated in less than a minute.

	SAA	AA	Kmeans	KP	PP
$\|\boldsymbol{H}\|_0/kn$	0.350	0.649	0.710	0.365	0.385
Purity	0.660	0.868	0.654	0.446	0.477
Entropy	0.361	0.216	0.375	0.723	0.690

Table 2: Performance of different algorithms for the gene expression dataset in Section 5.3

5.4 Scene Categorization dataset (Xiao et al., 2010)

Finally, we consider another popular application of AA arising in the context of image categorization (Abrol and Sharma, 2020; Chen et al., 2014). Given a collection of photos, AA can be used to identify a small subset of photos as their representatives (archetypes). Based on the mathematical formulation of AA, archetypes are expected to represent extreme scenes and objects present in the data, for instance, they should categorize different indoor/outdoor settings or city/nature scenes.

As far as we can tell, NMF with an explicit ℓ_0 regularization has not been used before to address the problem of scene categorization. However, in view of Theorem 3, we do not anticipate that additional sparsity will reduce the robustness properties of the estimator if the underlying matrix of archetypes is sparse. In addition, a sparse model may be desirable in terms of compressed storage. We apply SAA on the Scene Categorization dataset (Xiao et al., 2010). We select 12 different scenes that consist of different indoor and outdoor settings (2617 images in total). These scenes are toll plaza, hospital exterior, harbor, electricity station, underwater, youth hostel, valley, ski resort, football stadium, residential neighborhood, vineyard and iceberg. We extract and concatenate GIST and HOG features (Xiao et al., 2010) and implement different sparse NMF algorithms on the data with $k = 12$. As estimated archetypes in the feature space cannot be visualized, we use the closest data point to each archetype to visualize the result.

First, we consider AA and SAA with $\ell/kn = 0.5$ (we use the continuation framework in Section 4.2 and choose the value of λ that maximizes purity). The resulting visualization of archetypes for these two cases is the same. The visualization of estimated archetypes is shown in Figure 7 (a). We observe that the resulting archetypes appear to span the 12 different scenes in the dataset. Figure 7 (b) shows the resulting visualization for PP where the resulting archetypes matrix is set to have at most $0.65nk$ nonzeros. As it can be seen, PP can identify 10 distinct scenes and chooses the electricity station and the toll plaza twice. Figure 7 (c) shows the results for KP with the same sparsity as PP. This algorithm only identifies 5 distinct scenes. In summary, our SAA algorithm works as well as AA in terms of identifying different scenes while providing a sparse solution. This shows a sparse solution can be achieved without losing categorization performance.

6 Conclusion

In this paper, we consider the problem of sparse NMF with archetypal regularization where the goal is to represent a collection of data points as nonnegative linear combinations of a few nonnegative sparse factors. Javadi and Montanari (2019) recently showed that NMF (without sparsity) with archetypal regularization leads to robustness—factors learnt from noisy data are close to the underlying factors that generate the noiseless data. We generalize the notion of robustness to (a) strong robustness that implies each estimated archetype is close to the underlying archetypes and (b) weak robustness that implies there exists at least one recovered archetype that is close to the underlying archetypes. Javadi and Montanari (2019) is an instance of the notion of weak robustness presented herein. We show that under minimal assumptions, robustness in sparse NMF can be achieved by considering a sparsity constrained regularized AA problem, even if the underlying archetypes are not sparse. We present a block coordinate algorithm to get a good solution to the sparse AA problem and also an initialization framework using mixed integer programming that leads to better numerical results. We also present a local search algorithm that improves the quality of the solution of

SAA (this paper)

PP (Peharz and Pernkopf, 2012)

KP (Kim and Park, 2007)

Figure 7: The visualization of archetypes achieved by different algorithm for the scene categorization data in Section 5.4: (a) SAA (b) PP (c) KP.

our block coordinate algorithm. Numerical experiments on synthetic and real datasets shed further insights into the theoretical developments pursued in this paper.

7 Acknowledgements

Rahul Mazumder would like to thank Kushal Dey for helpful discussions in the initial stages of this work. This research was partially supported by grants from the Office of Naval Research (ONR-N000141812298), National Science Foundation (NSF-IIS-1718258), IBM and Liberty Mutual Insurance.

Proofs and Technical Details

A Additional Notation

We define

$$\tilde{D}(\boldsymbol{X},\boldsymbol{Y}) = \sum_{i=1}^{\#\text{row}(\boldsymbol{X})} \sqrt{D(\boldsymbol{X}_{i,.},\boldsymbol{Y})} \text{ and } \tilde{\mathcal{L}}(\boldsymbol{X},\boldsymbol{Y}) = \sum_{i=1}^{\#\text{row}(\boldsymbol{X})} \min_{j\in[\#\text{row}(\boldsymbol{Y})]} \|\boldsymbol{X}_{i,.} - \boldsymbol{Y}_{j,.}\|_2. \tag{34}$$

For a set S, S^c denotes the complement of the set. We use $\mathbb{I}_k$ to denote the identity matrix of size k.

B Technical Details of The Toy Example

For Figure 1 Panel (a), we let

$$\boldsymbol{H}_0 = \begin{bmatrix} 0.15 & 0.15 \\ 0.1 & 0.7 \\ 0.7 & 0.1 \end{bmatrix}$$

and produce 50 data points. To do so, each entry of $\boldsymbol{W}_0$ is drawn from an independent uniform distribution in $[0,1]$ and each row is normalized to sum to one. The noiseless data matrix is $\boldsymbol{X}_0 = \boldsymbol{W}_0\boldsymbol{H}_0$ and we add three rows of $\boldsymbol{H}_0$ to this to a obtain separable problem. In this case, we can see $D(\boldsymbol{X}_0,\boldsymbol{H}_0) = D(\boldsymbol{H}_0,\boldsymbol{X}_0) = 0$ which is the exact AA solution of Cutler and Breiman (1994). The red convex hull is $\text{Conv}(\boldsymbol{H}_0)$. Let

$$\boldsymbol{H}_1 = \begin{bmatrix} 0.05 & 0.05 \\ 1 & 0.1 \\ 0.1 & 1 \end{bmatrix}.$$

The black convex hull is $\text{Conv}(\boldsymbol{H}_1)$ for which we have $D(\boldsymbol{X}_0,\boldsymbol{H}_1) = 0$ but $D(\boldsymbol{H}_1,\boldsymbol{X}_0) > 0$.
For Figure 1 Panel (b), the data $\boldsymbol{X}$ is produced by $\boldsymbol{X} = \boldsymbol{X}_0 + \boldsymbol{Z}$ where $\boldsymbol{Z}$ has zero-mean iid normal coordinates with variance of 0.1. In this case, we have for every data point i, $D(\boldsymbol{X}_{i,.},\boldsymbol{H}_0) \leq 0.1$ making $\boldsymbol{H}_0$ a feasible solution for the regularized AA of Javadi and Montanari (2019) and as $D(\boldsymbol{H}_0,\boldsymbol{X}) = 0$, this is the solution of the regularized AA problem (the red convex hull). However, we have $D(\boldsymbol{X},\boldsymbol{H}_1) = 0$ and $D(\boldsymbol{H}_1,\boldsymbol{X}) > 0$, so the black convex hull is not an optimal solution.
Finally, let

$$\boldsymbol{H}_2 = \begin{bmatrix} 0 & 0 \\ 0 & 0.8 \\ 0.8 & 0 \end{bmatrix}.$$

In Figure 1 Panel (c), the red convex hull is $\text{Conv}(\boldsymbol{H}_2)$ and the black one is $\text{Conv}(\boldsymbol{H}_0)$. We have $\|\boldsymbol{H}_2\|_0 = 2$ and $\|\boldsymbol{H}_0\|_0 = 6$ so the black convex hull is not sparse. In addition, among all solutions that have $\|\boldsymbol{H}\|_0 = 2$ and $D(\boldsymbol{X}_0,\boldsymbol{H}) = 0$, the quantity $D(\boldsymbol{H},\boldsymbol{X}_0)$ is minimized for the red convex hull, making it the sparse archetypal solution.

C Technical Lemmas

Lemma 1. For any two matrices $\boldsymbol{X} \in \mathbb{R}^{m_1 \times n}$ and $\boldsymbol{Y} \in \mathbb{R}^{m_2 \times n}$, we have:

$$\frac{1}{\sqrt{m_1}} \tilde{D}(\boldsymbol{X}, \boldsymbol{Y}) \leq D(\boldsymbol{X}, \boldsymbol{Y})^{1/2} \leq \tilde{D}(\boldsymbol{X}, \boldsymbol{Y}) \tag{35}$$

$$\frac{1}{\sqrt{m_1}} \tilde{\mathcal{L}}(\boldsymbol{X}, \boldsymbol{Y}) \leq \mathcal{L}(\boldsymbol{X}, \boldsymbol{Y})^{1/2} \leq \tilde{\mathcal{L}}(\boldsymbol{X}, \boldsymbol{Y}), \tag{36}$$

where, recall $\tilde{D}$ and $\tilde{\mathcal{L}}$ are as defined in (34).

Proof. Define the vector $\boldsymbol{u} \in \mathbb{R}^{m_1}$ such that $\boldsymbol{u}_i = \sqrt{D(\boldsymbol{X}_{i,.}, \boldsymbol{Y})}$. Note that

$$\frac{1}{\sqrt{m_1}} \tilde{D}(\boldsymbol{X}, \boldsymbol{Y}) = \frac{1}{\sqrt{m_1}} \|\boldsymbol{u}\|_1 \leq \|\boldsymbol{u}\|_2 = D(\boldsymbol{X}, \boldsymbol{Y})^{1/2} \leq \|\boldsymbol{u}\|_1 = \tilde{D}(\boldsymbol{X}, \boldsymbol{Y})$$

which establishes (35). The proof of (36) is similar. □

Lemma 2. If $D(\boldsymbol{A}_1, \boldsymbol{A}_2) \leq D(\boldsymbol{B}_1, \boldsymbol{B}_2)$ where $\boldsymbol{A}_1, \boldsymbol{A}_2 \in \mathbb{R}^{m \times n}$,$\boldsymbol{B}_1, \boldsymbol{B}_2 \in \mathbb{R}^{k \times n}$, we have

$$\tilde{D}(\boldsymbol{A}_1, \boldsymbol{A}_2) \leq \sqrt{m} \tilde{D}(\boldsymbol{B}_1, \boldsymbol{B}_2).$$

Proof. We make use of Lemma 1. The proof follows from:

$$\tilde{D}(\boldsymbol{A}_1, \boldsymbol{A}_2) \leq \sqrt{m} D(\boldsymbol{A}_1, \boldsymbol{A}_2)^{1/2} \leq \sqrt{m} D(\boldsymbol{B}_1, \boldsymbol{B}_2)^{1/2} \leq \sqrt{m} \tilde{D}(\boldsymbol{B}_1, \boldsymbol{B}_2).$$

□

Lemma 3. Suppose $\boldsymbol{X} \in \mathbb{R}^{m_1 \times n}, \boldsymbol{Y} \in \mathbb{R}^{m_2 \times n}, \boldsymbol{Z} \in \mathbb{R}^{m_3 \times n}$, then

$$\tilde{D}(\boldsymbol{X}, \boldsymbol{Z}) \leq \tilde{\mathcal{L}}(\boldsymbol{X}, \boldsymbol{Y}) + m_1 \tilde{D}(\boldsymbol{Y}, \boldsymbol{Z}).$$

Proof. Fix $i \in [m_1]$. For any $\boldsymbol{u} \in \mathbb{R}^n$, we have

$$\begin{aligned} \sqrt{D(\boldsymbol{X}_{i,.}, \boldsymbol{Z})} &= \min_{\boldsymbol{v} \in \mathrm{Conv}(\boldsymbol{Z})} \|\boldsymbol{X}_{i,.} - \boldsymbol{v}\|_2 \\ &\leq \min_{\boldsymbol{v} \in \mathrm{Conv}(\boldsymbol{Z})} \{\|\boldsymbol{X}_{i,.} - \boldsymbol{u}\|_2 + \|\boldsymbol{u} - \boldsymbol{v}\|_2\} \\ &= \|\boldsymbol{X}_{i,.} - \boldsymbol{u}\|_2 + \min_{\boldsymbol{v} \in \mathrm{Conv}(\boldsymbol{Z})} \|\boldsymbol{u} - \boldsymbol{v}\|_2. \end{aligned} \tag{37}$$

Let

$$\boldsymbol{u} = \operatorname*{argmin}_{\boldsymbol{p} \in \{\boldsymbol{Y}_{j,.} : j \in [\#\mathrm{row}(\boldsymbol{Y})]\}} \|\boldsymbol{X}_{i,.} - \boldsymbol{p}\|_2.$$

As a result, noting that $\tilde{D}(\boldsymbol{Y}, \boldsymbol{Z}) = \sum_{j=1}^{m_2} \sqrt{D(\boldsymbol{Y}_{j,.}, \boldsymbol{Z})}$ and $\boldsymbol{u}$ is a row of $\boldsymbol{Y}$,

$$\min_{\boldsymbol{v} \in \mathrm{Conv}(\boldsymbol{Z})} \|\boldsymbol{u} - \boldsymbol{v}\|_2 = \sqrt{D(\boldsymbol{u}, \boldsymbol{Z})} \leq \sum_{j=1}^{m_2} \sqrt{D(\boldsymbol{Y}_{j,.}, \boldsymbol{Z})} = \tilde{D}(\boldsymbol{Y}, \boldsymbol{Z}). \tag{38}$$

Using the definition of $\boldsymbol{u}$ and (38) we can bound the rhs in (37) to get:

$$\sqrt{D(\boldsymbol{X}_{i,.}, \boldsymbol{Z})} \leq \min_{j \in m_2} \|\boldsymbol{X}_{i,.} - \boldsymbol{Y}_{j,.}\|_2 + \tilde{D}(\boldsymbol{Y}, \boldsymbol{Z}).$$

By summing the above over i, we have:

$$\begin{aligned}
\tilde{D}(\boldsymbol{X}, \boldsymbol{Z}) &= \sum_{i=1}^{m_1} \sqrt{D(\boldsymbol{X}_{i,.}, \boldsymbol{Z})} \\
&\leq \sum_{i=1}^{m_1} \min_{j \in m_2} \|\boldsymbol{X}_{i,.} - \boldsymbol{Y}_{j,.}\|_2 + \sum_{i=1}^{m_1} \tilde{D}(\boldsymbol{Y}, \boldsymbol{Z}) \\
&= \tilde{\mathcal{L}}(\boldsymbol{X}, \boldsymbol{Y}) + m_1 \tilde{D}(\boldsymbol{Y}, \boldsymbol{Z})
\end{aligned}$$

which completes the proof. □

Lemma 4 (Javadi and Montanari (2019))**.** If $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{m \times n}$, $m \leq n$ are matrices with linearly independent rows, we have

$$\mathcal{L}(\boldsymbol{A}, \boldsymbol{B})^{1/2} \leq 2\kappa(\boldsymbol{A}) D(\boldsymbol{A}, \boldsymbol{B})^{1/2} + (1 + \sqrt{2})\sqrt{m} D(\boldsymbol{B}, \boldsymbol{A})^{1/2}$$

where recall that $\kappa(\boldsymbol{H}) := \sigma_{\max}(\boldsymbol{H})/\sigma_{\min}(\boldsymbol{H})$ denotes the condition number of $\boldsymbol{H}$.

Lemma 5. Suppose $\boldsymbol{H} \in \mathbb{R}^{k \times n}$, $\boldsymbol{X}_0 \in \mathbb{R}^{m \times n}$ and $\boldsymbol{Z} \in \mathbb{R}^{m \times n}$ is such that $\max_{i \in [m]} \|\boldsymbol{Z}_{i,.}\|_2 \leq \delta$. Let $\boldsymbol{X} = \boldsymbol{X}_0 + \boldsymbol{Z}$. We have

$$\tilde{D}(\boldsymbol{H}, \boldsymbol{X}) \leq \tilde{D}(\boldsymbol{H}, \boldsymbol{X}_0) + k\delta \tag{39}$$

$$\tilde{D}(\boldsymbol{X}, \boldsymbol{H}) \leq \tilde{D}(\boldsymbol{X}_0, \boldsymbol{H}) + m\delta. \tag{40}$$

Proof. Let us denote the m-dimensional unit simplex by:

$$\Delta^m = \{\boldsymbol{\alpha} \in \mathbb{R}^m : \sum_{i=1}^{m} \boldsymbol{\alpha}_i = 1, \boldsymbol{\alpha}_i \geq 0\}.$$

We have for any $i \in [k]$

$$\begin{aligned}
\tilde{D}(\boldsymbol{H}_{i,.}, \boldsymbol{X}) &= \min_{\boldsymbol{\alpha} \in \Delta^m} \|\boldsymbol{H}_{i,.} - \sum_{j=1}^{m} \boldsymbol{\alpha}_j \boldsymbol{X}_{j,.}\|_2 \\
&= \min_{\boldsymbol{\alpha} \in \Delta^m} \|\boldsymbol{H}_{i,.} - \sum_{j=1}^{m} \boldsymbol{\alpha}_j \boldsymbol{X}^0_{j,.} - \sum_{j=1}^{m} \boldsymbol{\alpha}_j \boldsymbol{Z}_{j,.}\|_2 \\
&\leq \min_{\boldsymbol{\alpha} \in \Delta^m} \left\{ \|\boldsymbol{H}_{i,.} - \sum_{j=1}^{m} \boldsymbol{\alpha}_j \boldsymbol{X}^0_{j,.}\|_2 + \|\sum_{j=1}^{m} \boldsymbol{\alpha}_j \boldsymbol{Z}_{j,.}\|_2 \right\} \\
&\leq \min_{\boldsymbol{\alpha} \in \Delta^m} \left\{ \|\boldsymbol{H}_{i,.} - \sum_{j=1}^{m} \boldsymbol{\alpha}_j \boldsymbol{X}^0_{j,.}\|_2 + \sum_{j=1}^{m} \boldsymbol{\alpha}_j \|\boldsymbol{Z}_{j,.}\|_2 \right\} \\
&\leq \min_{\boldsymbol{\alpha} \in \Delta^m} \left\{ \|\boldsymbol{H}_{i,.} - \sum_{j=1}^{m} \boldsymbol{\alpha}_j \boldsymbol{X}^0_{j,.}\|_2 + (\sum_{j=1}^{m} \boldsymbol{\alpha}_j) \max_{j \in [m]} \|\boldsymbol{Z}_{j,.}\|_2 \right\} \\
&\leq \min_{\boldsymbol{\alpha} \in \Delta^m} \|\boldsymbol{H}_{i,.} - \sum_{j=1}^{m} \boldsymbol{\alpha}_j \boldsymbol{X}^0_{j,.}\|_2 + \delta \\
&= \tilde{D}(\boldsymbol{H}_{i,.}, \boldsymbol{X}_0) + \delta.
\end{aligned}$$

Using the above bound, we have:

$$\tilde{D}(\boldsymbol{H}, \boldsymbol{X}) = \sum_{i=1}^{k} \tilde{D}(\boldsymbol{H}_{i,.}, \boldsymbol{X}) \leq \sum_{i=1}^{k} \tilde{D}(\boldsymbol{H}_{i,.}, \boldsymbol{X}_0) + k\delta = \tilde{D}(\boldsymbol{H}, \boldsymbol{X}_0) + k\delta$$

which establishes (39). We now show (40). For a fixed $i \in [m]$, we have

$$\begin{aligned}
\tilde{D}(\boldsymbol{X}_{i,.}, \boldsymbol{H}) &= \min_{\boldsymbol{\alpha} \in \Delta^k} \|\boldsymbol{X}_{i,.} - \sum_{j=1}^{k} \boldsymbol{\alpha}_j \boldsymbol{H}_{j,.}\|_2 \\
&= \min_{\boldsymbol{\alpha} \in \Delta^k} \|\boldsymbol{X}^0_{i,.} - \sum_{j=1}^{k} \boldsymbol{\alpha}_j \boldsymbol{H}_{j,.} + \boldsymbol{Z}_{i,.}\|_2 \\
&\leq \min_{\boldsymbol{\alpha} \in \Delta^k} \|\boldsymbol{X}^0_{i,.} - \sum_{j=1}^{k} \boldsymbol{\alpha}_j \boldsymbol{H}_{j,.}\|_2 + \|\boldsymbol{Z}_{i,.}\|_2 \\
&\leq \min_{\boldsymbol{\alpha} \in \Delta^k} \|\boldsymbol{X}^0_{i,.} - \sum_{j=1}^{k} \boldsymbol{\alpha}_j \boldsymbol{H}_{j,.}\|_2 + \delta \\
&= \tilde{D}(\boldsymbol{X}^0_{i,.}, \boldsymbol{H}) + \delta.
\end{aligned}$$

By summing the above bound over i, we arrive at (40). □

Lemma 6. Suppose $\boldsymbol{H} \in \mathbb{R}^{k \times n}$, $k \leq n$, has full row rank and $\boldsymbol{X} \in \mathbb{R}^{m \times n}$ is such that $\text{Conv}(\boldsymbol{X}) \subseteq \text{Conv}(\boldsymbol{H})$.

Let $\tilde{\boldsymbol{X}} \in \mathbb{R}^{k\times n}$ be a matrix with its i-row given by:

$$\tilde{\boldsymbol{X}}_{i,\cdot} = \operatorname*{argmin}_{\boldsymbol{u}\in\{\boldsymbol{X}_{j,\cdot}:j\in[m]\}} \|\boldsymbol{u} - \boldsymbol{H}_{i,\cdot}\|_2 \tag{41}$$

for all $i \in [k]$. Then,

$$\mathcal{L}(\boldsymbol{H}, \boldsymbol{X})^{1/2} \leq 2k\kappa(\boldsymbol{H}) D(\boldsymbol{H}, \tilde{\boldsymbol{X}})^{1/2}. \tag{42}$$

Proof. Fix any $\epsilon > 0$. There is a matrix $\boldsymbol{Z}_\epsilon \in \mathbb{R}^{k\times n}$ such that $\|\boldsymbol{Z}_\epsilon\|_F \leq \epsilon$ (therefore, $\max_{i\in[m]} \|\boldsymbol{Z}^\epsilon_{i,\cdot}\|_2 \leq \epsilon$) and $\tilde{\boldsymbol{X}} + \boldsymbol{Z}_\epsilon$ has full row rank. We have for any $i \in [k]$

$$\begin{aligned}
\tilde{\mathcal{L}}(\boldsymbol{H}_{i,\cdot}, \boldsymbol{X}) &= \tilde{\mathcal{L}}(\boldsymbol{H}_{i,\cdot}, \tilde{\boldsymbol{X}}) \\
&= \min_{j\in[k]} \|\boldsymbol{H}_{i,\cdot} - \tilde{\boldsymbol{X}}_{j,\cdot}\|_2 \\
&= \min_{j\in[k]} \|\boldsymbol{H}_{i,\cdot} - \tilde{\boldsymbol{X}}_{j,\cdot} + (\boldsymbol{Z}_\epsilon)_{j,\cdot} - (\boldsymbol{Z}_\epsilon)_{j,\cdot}\|_2 \\
&\leq \min_{j\in[k]} \|\boldsymbol{H}_{i,\cdot} - \tilde{\boldsymbol{X}}_{j,\cdot} - (\boldsymbol{Z}_\epsilon)_{j,\cdot}\|_2 + \|(\boldsymbol{Z}_\epsilon)_{j,\cdot}\|_2 \\
&\leq \tilde{\mathcal{L}}(\boldsymbol{H}_{i,\cdot}, \tilde{\boldsymbol{X}} + \boldsymbol{Z}_\epsilon) + \epsilon.
\end{aligned} \tag{43}$$

Note that we have the following inequalities:

$$\begin{aligned}
\mathcal{L}(\boldsymbol{H}, \boldsymbol{X})^{1/2} &\overset{(a)}{\leq} \tilde{\mathcal{L}}(\boldsymbol{H}, \boldsymbol{X}) \\
&= \sum_{i=1}^{k} \tilde{\mathcal{L}}(\boldsymbol{H}_{i,\cdot}, \boldsymbol{X}) \\
&\overset{(b)}{\leq} \tilde{\mathcal{L}}(\boldsymbol{H}, \tilde{\boldsymbol{X}} + \boldsymbol{Z}_\epsilon) + k\epsilon \\
&\overset{(c)}{\leq} \sqrt{k}\mathcal{L}(\boldsymbol{H}, \tilde{\boldsymbol{X}} + \boldsymbol{Z}_\epsilon)^{1/2} + k\epsilon,
\end{aligned} \tag{44}$$

where, $(a), (c)$ are results of Lemma 1 and (b) is a result of (43). In addition, by Lemmas 1 and 5, we have:

$$D(\boldsymbol{H}, \tilde{\boldsymbol{X}} + \boldsymbol{Z}_\epsilon)^{1/2} \leq \tilde{D}(\boldsymbol{H}, \tilde{\boldsymbol{X}} + \boldsymbol{Z}_\epsilon) \leq \tilde{D}(\boldsymbol{H}, \tilde{\boldsymbol{X}}) + k\epsilon \tag{45}$$

and

$$D(\tilde{\boldsymbol{X}} + \boldsymbol{Z}_\epsilon, \boldsymbol{H})^{1/2} \leq \tilde{D}(\tilde{\boldsymbol{X}} + \boldsymbol{Z}_\epsilon, \boldsymbol{H}) \leq \tilde{D}(\tilde{\boldsymbol{X}}, \boldsymbol{H}) + m\epsilon = m\epsilon, \tag{46}$$

where the last equality in (46) follows from the fact that $\tilde{D}(\tilde{\boldsymbol{X}}, \boldsymbol{H}) = 0$ (as $\mathrm{Conv}(\boldsymbol{X}) \subseteq \mathrm{Conv}(\boldsymbol{H})$).

Starting with (44), we have

$$\begin{aligned}
\mathcal{L}(\boldsymbol{H}, \boldsymbol{X})^{1/2} &\leq \sqrt{k}\mathcal{L}(\boldsymbol{H}, \tilde{\boldsymbol{X}} + \boldsymbol{Z}_\epsilon)^{1/2} + k\epsilon && (47)\\
&\leq 2\sqrt{k}\kappa(\boldsymbol{H}) D(\boldsymbol{H}, \tilde{\boldsymbol{X}} + \boldsymbol{Z}_\epsilon)^{1/2} + k(1+\sqrt{2}) D(\tilde{\boldsymbol{X}} + \boldsymbol{Z}_\epsilon, \boldsymbol{H})^{1/2} + k\epsilon && (48)\\
&\leq 2\sqrt{k}\kappa(\boldsymbol{H}) \tilde{D}(\boldsymbol{H}, \tilde{\boldsymbol{X}}) + \epsilon(k + mk(1+\sqrt{2}) + 2\sqrt{k^3}\kappa(\boldsymbol{H})) && (49)\\
&\leq 2k\kappa(\boldsymbol{H}) D(\boldsymbol{H}, \tilde{\boldsymbol{X}})^{1/2} + \epsilon(k + mk(1+\sqrt{2}) + 2\sqrt{k^3}\kappa(\boldsymbol{H})) && (50)
\end{aligned}$$

where above, inequality (48) is a result of Lemma 4; (49) follows from (45) and (46). Finally, inequality (50) is a result of Lemma 1. As inequality (50) is true for any $\epsilon > 0$, taking the limit $\epsilon \downarrow 0+$, we arrive at (42). □

D Proofs of Main Results

D.1 Proof of Theorem 1

Proof. For any i, let us denote:

$$j_i = \operatorname*{argmin}_{j\in[k]} \|\boldsymbol{H}_{i,.} - \boldsymbol{H}^0_{j,.}\|_2.$$

We have

$$\begin{aligned}
\mathcal{L}(\boldsymbol{H}_0, \boldsymbol{H}) &= \sum_{i=1}^{k} \min_{j\in[k]} \|\boldsymbol{H}^0_{i,.} - \boldsymbol{H}_{j,.}\|_2^2 \\
&\le \sum_{i=1}^{k} \|\boldsymbol{H}^0_{i,.} - \boldsymbol{H}_{i,.}\|_2^2 \\
&= \sum_{i=1}^{k} \|\boldsymbol{H}^0_{i,.} - \boldsymbol{H}^0_{j_i,.} + \boldsymbol{H}^0_{j_i,.} - \boldsymbol{H}_{i,.}\|_2^2 \\
&\le 2\sum_{i=1}^{k} \|\boldsymbol{H}^0_{i,.} - \boldsymbol{H}^0_{j_i,.}\|_2^2 + 2\sum_{i=1}^{k} \|\boldsymbol{H}^0_{j_i,.} - \boldsymbol{H}_{i,.}\|_2^2 \\
&\le 2kb(\boldsymbol{H}_0)^2 + 2\mathcal{L}(\boldsymbol{H}, \boldsymbol{H}_0),
\end{aligned}$$

where, the last line follows from the definition of $b(\boldsymbol{H})$ in (6). □

D.2 Proof of Theorem 2

Proof. For any given matrices $\boldsymbol{H}$ and $\boldsymbol{H}_0$, there exists a matrix $\boldsymbol{U}_1 \in \{0,1\}^{k\times k}$ such that it has exactly one 1 in each row, such that

$$\mathcal{L}(\boldsymbol{H}, \boldsymbol{H}_0) = \|\boldsymbol{H} - \boldsymbol{U}_1\boldsymbol{H}_0\|_F^2. \tag{51}$$

In fact, for any $i \in [k]$, we have $\boldsymbol{U}^1_{i,j_i} = 1$ where

$$j_i = \operatorname*{argmin}_{j\in[k]} \|\boldsymbol{H}_{i,.} - \boldsymbol{H}^0_{j,.}\|_2.$$

Noting the noiseless data $\boldsymbol{X}_0$ is given as $\boldsymbol{X}_0 = \boldsymbol{W}_0\boldsymbol{H}_0$ where $\boldsymbol{W}_0 \ge 0$ and $\boldsymbol{W}_0\mathbf{1}_k = \mathbf{1}_m$, one has

$$\begin{aligned}
D(\boldsymbol{X}_0, \boldsymbol{H})^{1/2} &= \min_{\substack{\boldsymbol{W}\ge 0 \\ \boldsymbol{W}\mathbf{1}_k = \mathbf{1}_m}} \|\boldsymbol{X}_0 - \boldsymbol{W}\boldsymbol{H}\|_F \le \|\boldsymbol{X}_0 - \boldsymbol{W}_0\boldsymbol{H}\|_F \\
&= \|\boldsymbol{W}_0\boldsymbol{H}_0 - \boldsymbol{W}_0(\boldsymbol{H} - \boldsymbol{U}_1\boldsymbol{H}_0 + \boldsymbol{U}_1\boldsymbol{H}_0)\|_F \\
&= \|\boldsymbol{W}_0(\mathbb{I}_k - \boldsymbol{U}_1)\boldsymbol{H}_0 + \boldsymbol{W}_0(\boldsymbol{H} - \boldsymbol{U}_1\boldsymbol{H}_0)\|_F \\
&\le \|\boldsymbol{W}_0(\mathbb{I}_k - \boldsymbol{U}_1)\boldsymbol{H}_0\|_F + \|\boldsymbol{W}_0(\boldsymbol{H} - \boldsymbol{U}_1\boldsymbol{H}_0)\|_F.
\end{aligned} \tag{52}$$

Note that each row of $\boldsymbol{W}_0$ sums to 1 and is nonnegative. Therefore, for each row, $\|\boldsymbol{W}^0_{i,.}\|_2^2 \leq 1$ and

$$\|\boldsymbol{W}_0\|_F^2 = \sum_{i=1}^{m} \|\boldsymbol{W}^0_{i,.}\|_2^2 \leq m. \tag{53}$$

Using (52), we have the following inequalities:

$$\begin{aligned} D(\boldsymbol{X}_0, \boldsymbol{H})^{1/2} &\leq \|\boldsymbol{W}_0(\mathbb{I}_k - \boldsymbol{U}_1)\boldsymbol{H}_0\|_F + \|\boldsymbol{W}_0(\boldsymbol{H} - \boldsymbol{U}_1\boldsymbol{H}_0)\|_F \\ &\leq \sqrt{m}\|\boldsymbol{H}_0\|_F\|\mathbb{I}_k - \boldsymbol{U}_1\|_F + \sqrt{m}\|\boldsymbol{H} - \boldsymbol{U}_1\boldsymbol{H}_0\|_F \\ &\leq k\sqrt{m}\|\boldsymbol{H}_0\|_F + \sqrt{m}\mathcal{L}(\boldsymbol{H}, \boldsymbol{H}_0)^{1/2} \end{aligned}$$

where the second inequality makes use of (53); and the last inequality uses (51) and $\|\mathbb{I}_k - \boldsymbol{U}_1\|_F \leq k$ (since, $\mathbb{I}_k - \boldsymbol{U}_1$ is a matrix with every entry between -1 and 1).

Similar to $\boldsymbol{U}_1$ in the definition (51), we can consider $\boldsymbol{U}_2 \in \{0,1\}^{k\times k}$ such that

$$\mathcal{L}(\boldsymbol{H}_0, \boldsymbol{H}) = \|\boldsymbol{H}_0 - \boldsymbol{U}_2\boldsymbol{H}\|_F^2. \tag{54}$$

Since $\boldsymbol{U}_2\mathbf{1}_k = \mathbf{1}_k$; and using the fact that every row of $\boldsymbol{W}_0$ sums to one, we have: $\boldsymbol{W}_0\boldsymbol{U}_2\mathbf{1}_k = \mathbf{1}_m$ and

$$\boldsymbol{W}_0\boldsymbol{U}_2 \in \{\boldsymbol{W} \ : \ \boldsymbol{W} \geq 0, \boldsymbol{W}\mathbf{1}_k = \mathbf{1}_m\}. \tag{55}$$

Note that we have the following:

$$\begin{aligned} D(\boldsymbol{X}_0, \boldsymbol{H})^{1/2} &= \min_{\substack{\boldsymbol{W}\geq 0 \\ \boldsymbol{W}\mathbf{1}_k = \mathbf{1}_m}} \|\boldsymbol{X}_0 - \boldsymbol{W}\boldsymbol{H}\|_F \overset{(a)}{\leq} \|\boldsymbol{X}_0 - \boldsymbol{W}_0\boldsymbol{U}_2\boldsymbol{H}\|_F \\ &= \|\boldsymbol{W}_0\boldsymbol{H}_0 - \boldsymbol{W}_0\boldsymbol{U}_2\boldsymbol{H}\|_F \\ &\overset{(b)}{\leq} \sqrt{m}\|\boldsymbol{H}_0 - \boldsymbol{U}_2\boldsymbol{H}\|_F \overset{(c)}{=} \sqrt{m}\mathcal{L}(\boldsymbol{H}_0, \boldsymbol{H})^{1/2}, \end{aligned}$$

where, (a) uses feasibility condition (55), (b) is due to (53) and (c) uses (54). □

D.3 Proof of Theorem 3

Note that the constants in this theorem are listed below.

$$\begin{aligned}
c_1 =& 4\sqrt{k^3}\kappa^2(\boldsymbol{H}_0) + (1+\sqrt{2})\sqrt{k^3}\\
c_2 =& 4mk\kappa(\boldsymbol{H}_0) + (1+\sqrt{2})\sqrt{k}(k+\sqrt{k^3})\\
c_3 =& 2\sqrt{m^3}k\kappa(\boldsymbol{H}_0) + (1+\sqrt{2})k^2\\
c_4 =& k + 2\sqrt{k^3}\kappa(\boldsymbol{H}_0)\\
c_5 =& (k+\sqrt{k^3}) + 2mk\\
c_6 =& \sqrt{k^3} + k\sqrt{m^3}\\
c_7 =& \frac{\sigma_{\min}(\boldsymbol{H}_0)}{6\sqrt{k}}\\
c_8 =& 7k\kappa(\boldsymbol{H}_0) + 2(1+\sqrt{2})k^2\kappa(\boldsymbol{H}_0)\\
c_9 =& 7\kappa(\boldsymbol{H}_0)(k+\sqrt{k^3}) + 2(1+\sqrt{2})\sqrt{k^3}m\\
c_{10} =& 7\kappa(\boldsymbol{H}_0)\sqrt{k^3} + (1+\sqrt{2})\sqrt{k^3}\sqrt{m^3}.
\end{aligned} \tag{56}$$

We first present Lemmas 7 and 8 useful for the proof of Theorem 3.

Lemma 7. Under the assumptions of Theorem 3, $P_\ell(\boldsymbol{H}_0)$ is feasible for problem (3).

Proof. By our model-setup, there is a nonnegative $\boldsymbol{W}_0$ with $\boldsymbol{W}_0\mathbf{1}_k = \mathbf{1}_m$ such that $\boldsymbol{X}_0 = \boldsymbol{W}_0\boldsymbol{H}_0$. For $i \in [m]$, let

$$\boldsymbol{v}_i = \boldsymbol{W}^0_{i,\cdot} P_\ell(\boldsymbol{H_0}) \in \mathrm{Conv}(P_\ell(\boldsymbol{H_0})).$$

Using the definition of $\boldsymbol{v}_i$ above, we have

$$\begin{aligned}
D(\boldsymbol{X}_0, P_\ell(\boldsymbol{H_0})) &= \sum_{i=1}^m \min_{\boldsymbol{u}\in\mathrm{Conv}(P_\ell(\boldsymbol{H_0}))} \|\boldsymbol{X}^0_{i,\cdot} - \boldsymbol{u}\|_2^2 \le \sum_{i=1}^m \|\boldsymbol{X}^0_{i,\cdot} - \boldsymbol{v}_i\|_2^2\\
&= \|\boldsymbol{X}_0 - \boldsymbol{W}_0 P_\ell(\boldsymbol{H_0})\|_F^2.
\end{aligned} \tag{57}$$

For $i \in [m]$, we have:

$$\begin{aligned}
D(\boldsymbol{X}^0_{i,\cdot}, P_\ell(\boldsymbol{H}_0))^{1/2} &\overset{(a)}{\le} D(\boldsymbol{X}_0, P_\ell(\boldsymbol{H}_0))^{1/2}\\
&\overset{(b)}{\le} \|\boldsymbol{X}_0 - \boldsymbol{W}_0 P_\ell(\boldsymbol{H}_0)\|_F\\
&= \|\boldsymbol{W}_0(\boldsymbol{H}_0 - P_\ell(\boldsymbol{H}_0))\|_F\\
&= \|\boldsymbol{W}_0 P_\ell^\perp(\boldsymbol{H}_0)\|_F \le \|\boldsymbol{W}_0\|_F \|P_\ell^\perp(\boldsymbol{H}_0)\|_F
\end{aligned} \tag{58}$$

where (a) uses the definition of $D(\boldsymbol{X}, \boldsymbol{H})$ and (b) is a result of (57). By (53) and (58), we have:

$$D(\boldsymbol{X}^0_{i,\cdot}, P_\ell(\boldsymbol{H}_0))^{1/2} \le \sqrt{m}\|P_\ell^\perp(\boldsymbol{H}_0)\|_F = \beta. \tag{59}$$

In what follows, for notational convenience, we denote $\boldsymbol{H} = P_\ell(\boldsymbol{H}_0)$. Note that $\boldsymbol{H}$ satisfies the sparsity

constraint in (3). We have the following:

$$\begin{aligned} D(\boldsymbol{X}_{i,.}, \boldsymbol{H})^{1/2} &\leq \tilde{D}(\boldsymbol{X}^0_{i,.} + \boldsymbol{Z}_{i,.}, \boldsymbol{H}) \leq \tilde{D}(\boldsymbol{X}^0_{i,.}, \boldsymbol{H}) + \|\boldsymbol{Z}_{i,.}\|_2 \\ &\leq D(\boldsymbol{X}^0_{i,.}, \boldsymbol{H})^{1/2} + \underbrace{\max_i \|\boldsymbol{Z}_{i,.}\|_2}_{\delta} \\ &\leq \sqrt{m}\|P^{\perp}_{\ell}(\boldsymbol{H}_0)\|_F + \delta = \beta + \delta. \end{aligned} \tag{60}$$

where the first inequality is a result of Lemma 1 and the fact $\boldsymbol{X} = \boldsymbol{X}_0 + \boldsymbol{Z}$; the second inequality is a result of Lemma 5 (we use (40) with $m = 1$); and the third inequality uses (59). Bound (60) shows that $\boldsymbol{H} = P_{\ell}(\boldsymbol{H}_0)$ is a feasible solution for problem (3). □

Lemma 8. Under the assumptions of Theorem 3, one has

$$D(\hat{\boldsymbol{H}}, \boldsymbol{H}_0)^{1/2} \leq \sqrt{k} D(\boldsymbol{H}_0, \tilde{\boldsymbol{X}}_0)^{1/2} + \sqrt{k^3/m}\beta + (k + \sqrt{k^3})\delta \tag{61}$$

$$D(\boldsymbol{H}_0, \hat{\boldsymbol{H}})^{1/2} \leq 2\sqrt{k^3}\kappa(\boldsymbol{H}_0) D(\boldsymbol{H}_0, \tilde{\boldsymbol{X}}_0)^{1/2} + 2mk\delta + mk\beta.(k + \sqrt{k^3})\delta. \tag{62}$$

Proof. We first prove (61). To this end, note that $\mathrm{Conv}(\boldsymbol{X}_0) \subseteq \mathrm{Conv}(\boldsymbol{H}_0)$, so

$$D(\hat{\boldsymbol{H}}, \boldsymbol{H}_0)^{1/2} \leq D(\hat{\boldsymbol{H}}, \boldsymbol{X}_0)^{1/2} \leq \tilde{D}(\hat{\boldsymbol{H}}, \boldsymbol{X}_0)$$

where the second inequality is a result of Lemma 1. In addition, as $\hat{\boldsymbol{H}}$ is the optimal solution of (3) and $P_{\ell}(\boldsymbol{H}_0)$ is feasible for problem (3) (by Lemma 7), we have $D(\hat{\boldsymbol{H}}, \boldsymbol{X}) \leq D(P_{\ell}(\boldsymbol{H}_0), \boldsymbol{X})$ and by Lemma 2,

$$\tilde{D}(\hat{\boldsymbol{H}}, \boldsymbol{X}) \leq \sqrt{k}\tilde{D}(P_{\ell}(\boldsymbol{H}_0), \boldsymbol{X}). \tag{63}$$

Therefore, one can write

$$\begin{aligned} D(\hat{\boldsymbol{H}}, \boldsymbol{H}_0)^{1/2} &\overset{(a)}{\leq} \tilde{D}(\hat{\boldsymbol{H}}, \boldsymbol{X}_0) = \tilde{D}(\hat{\boldsymbol{H}}, \boldsymbol{X} - \boldsymbol{Z}) \overset{(b)}{\leq} \tilde{D}(\hat{\boldsymbol{H}}, \boldsymbol{X}) + k\delta \\ &\overset{(c)}{\leq} \sqrt{k}\tilde{D}(P_{\ell}(\boldsymbol{H}_0), \boldsymbol{X}) + k\delta = \sqrt{k}\tilde{D}(\boldsymbol{H}_0 - P^{\perp}_{\ell}(\boldsymbol{H}_0), \boldsymbol{X}) + k\delta \\ &\overset{(d)}{\leq} \sqrt{k}\tilde{D}(\boldsymbol{H}_0, \boldsymbol{X}) + \sqrt{k^3}\max_{j\in[k]} \|P^{\perp}_{\ell}(\boldsymbol{H}_0)_{j,.}\|_2 + k\delta \\ &\leq \sqrt{k}\tilde{D}(\boldsymbol{H}_0, \boldsymbol{X}) + \sqrt{k^3}\|P^{\perp}_{\ell}(\boldsymbol{H}_0)\|_F + k\delta \\ &\leq \sqrt{k}\tilde{D}(\boldsymbol{H}_0, \boldsymbol{X}) + \sqrt{k^3/m}\beta + k\delta \\ &\overset{(e)}{\leq} \sqrt{k}\tilde{D}(\boldsymbol{H}_0, \boldsymbol{X}_0) + \sqrt{k^3/m}\beta + (k + \sqrt{k^3})\delta \\ &\overset{(f)}{\leq} \sqrt{k}\tilde{D}(\boldsymbol{H}_0, \tilde{\boldsymbol{X}}_0) + \sqrt{k^3/m}\beta + (k + \sqrt{k^3})\delta \end{aligned}$$

where (a) is a result of Lemma 1, $(b), (d), (e)$ are results of Lemma 5, (c) is a result of (63), and (f) is true as $\sqrt{k}\tilde{D}(\boldsymbol{H}_0, \boldsymbol{X}_0) \leq \sqrt{k}\tilde{D}(\boldsymbol{H}_0, \tilde{\boldsymbol{X}}_0)$ because $\mathrm{Conv}(\tilde{\boldsymbol{X}}_0) \subseteq \mathrm{Conv}(\boldsymbol{X}_0)$. This establishes (61).

We now proceed to show (62). To this end, using Lemma 3,

$$D(\boldsymbol{H}_0, \hat{\boldsymbol{H}})^{1/2} \leq \tilde{D}(\boldsymbol{H}_0, \hat{\boldsymbol{H}}) \leq \tilde{\mathcal{L}}(\boldsymbol{H}_0, \boldsymbol{X}_0) + k\tilde{D}(\boldsymbol{X}_0, \hat{\boldsymbol{H}}). \tag{64}$$

Note that by Lemma 6,

$$\tilde{\mathcal{L}}(\boldsymbol{H}_0, \boldsymbol{X}_0) \leq 2\sqrt{k^3}\kappa(\boldsymbol{H}_0) D(\boldsymbol{H}_0, \tilde{\boldsymbol{X}}_0)^{1/2}. \tag{65}$$

In addition,

$$\begin{aligned}
\tilde{D}(\boldsymbol{X}_0, \hat{\boldsymbol{H}}) = \tilde{D}(\boldsymbol{X} - \boldsymbol{Z}, \hat{\boldsymbol{H}}) &\overset{(a)}{\leq} \tilde{D}(\boldsymbol{X}, \hat{\boldsymbol{H}}) + m\delta \\
&\overset{(b)}{\leq} \sqrt{m} D(\boldsymbol{X}, \hat{\boldsymbol{H}})^{1/2} + m\delta \\
&\overset{(c)}{\leq} \sqrt{m}\sqrt{\sum_{i=1}^{m} (\delta + \beta)^2} + m\delta \\
&\leq m(\delta + \beta) + m\delta = 2m\delta + m\beta
\end{aligned} \tag{66}$$

where (a) is a result of Lemma 5, (b) is a result of Lemma 1 and (c) is due to the constraint $D(\boldsymbol{X}_{i,.}, \boldsymbol{H})^{1/2} \leq \delta + \beta$ in Problem (3). Therefore, by (64), (65) and (66),

$$D(\boldsymbol{H}_0, \hat{\boldsymbol{H}})^{1/2} \leq 2\sqrt{k^3}\kappa(\boldsymbol{H}_0) D(\boldsymbol{H}_0, \tilde{\boldsymbol{X}}_0)^{1/2} + 2mk\delta + mk\beta$$

which establishes (62). □

Proof of Theorem 3. Part 1) If $\hat{\boldsymbol{H}}$ has linearly independent rows, the desired result is achieved by substituting (61) and (62) into Lemma 4 with $\boldsymbol{A} = \boldsymbol{H}_0$ and $\boldsymbol{B} = \hat{\boldsymbol{H}}$. If $\hat{\boldsymbol{H}}$ does not have linearly independent rows, for $\epsilon > 0$ there exists $\hat{\boldsymbol{H}}_\epsilon$ with linearly independent rows such that $\|\hat{\boldsymbol{H}}_\epsilon - \hat{\boldsymbol{H}}\|_F \leq \epsilon$. Following a path similar to the proof of Lemma 6 and taking the limit $\epsilon \downarrow 0+$ the desired result is achieved.
Part 2) By summing (61) and (62),

$$\begin{aligned}
D(\hat{\boldsymbol{H}}, \boldsymbol{H}_0)^{1/2} + D(\boldsymbol{H}_0, \hat{\boldsymbol{H}})^{1/2} &\leq c_4 D(\boldsymbol{H}_0, \tilde{\boldsymbol{X}}_0)^{1/2} + c_5 \max_{i \in [m]} \|\boldsymbol{Z}_{i,.}\|_2 + c_6 \|P_\ell^\perp(\boldsymbol{H}_0)\|_F \\
&\overset{(a)}{\leq} \frac{\sigma_{\min}(\boldsymbol{H}_0)}{6\sqrt{k}}
\end{aligned}$$

where (a) is a result of condition (10). As a result,

$$D(\hat{\boldsymbol{H}}, \boldsymbol{H}_0)^{1/2} + D(\boldsymbol{H}_0, \hat{\boldsymbol{H}})^{1/2} \leq \frac{\sigma_{\min}(\boldsymbol{H}_0)}{6\sqrt{k}}.$$

Therefore, condition (B.42) of Javadi and Montanari (2019) holds and by Lemma B.3 of Javadi and Montanari (2019), we have

$$\kappa(\hat{\boldsymbol{H}}) \leq \frac{7}{2}\kappa(\boldsymbol{H}_0), \tag{67}$$

which shows $\hat{\boldsymbol{H}}$ has linearly independent rows. The rest of the proof is achieved by substituting (67), (61) and (62) into Lemma 4 with $\boldsymbol{A} = \hat{\boldsymbol{H}}$ and $\boldsymbol{B} = \boldsymbol{H}_0$. □

D.4 Proof of Proposition 1

The constants in this proposition are listed below:

$$\begin{aligned} c_\lambda^1 &= \left(2\kappa(\boldsymbol{H}_0)\left[k\sqrt{m}\sqrt{m+\lambda k^2}+mk\right]+(1+\sqrt{2})\sqrt{k}\left[\sqrt{mk/\lambda+k^3}+k\right]\right) \\ c_\lambda^2 &= \left(7\kappa(\boldsymbol{H}_0)\left[\sqrt{mk/\lambda+k^3}+k\right]+(1+\sqrt{2})\sqrt{k}\left[k\sqrt{m}\sqrt{m+\lambda k^2}+mk\right]\right) \\ c_\lambda^3 &= \left[(1+m)k+k\sqrt{m}\sqrt{m+\lambda k^2}+\sqrt{mk/\lambda+k^3}\right]. \end{aligned} \tag{68}$$

Lemma 9. Under the assumptions of Proposition 1, one has

$$D(\hat{\boldsymbol{H}}_\lambda, \boldsymbol{H}_0)^{1/2} \le \sqrt{k}\sqrt{\frac{m\delta^2}{\lambda}+k^2\delta^2}+k\delta, \tag{69}$$

$$D(\boldsymbol{H}_0, \hat{\boldsymbol{H}}_\lambda)^{1/2} \le k\sqrt{m}\sqrt{m\delta^2+\lambda k^2\delta^2}+mk\delta. \tag{70}$$

Proof. Recall that in this proposition, we assume $P_\ell(\boldsymbol{H}_0)=\boldsymbol{H}_0$ ($\beta=0$) and

$$D(\boldsymbol{H}_0, \tilde{\boldsymbol{X}}_0) = \tilde{D}(\boldsymbol{H}_0, \boldsymbol{X}_0) = 0. \tag{71}$$

From (60), we have

$$D(\boldsymbol{X}, \boldsymbol{H}_0) = \sum_{i=1}^m D(\boldsymbol{X}_{i,.}, \boldsymbol{H}_0) \le m\delta^2. \tag{72}$$

In addition, we have

$$D(\boldsymbol{H}_0, \boldsymbol{X})^{1/2} \overset{(a)}{\le} \tilde{D}(\boldsymbol{H}_0, \boldsymbol{X}) = \tilde{D}(\boldsymbol{H}_0, \boldsymbol{X}_0+\boldsymbol{Z}) \overset{(b)}{\le} \tilde{D}(\boldsymbol{H}_0, \boldsymbol{X}_0)+k\delta \overset{(c)}{=} k\delta \tag{73}$$

where (a) is due to Lemma 1, (b) is due to Lemma 5 and (c) is true because of (71). Therefore, from (73) we have:

$$D(\boldsymbol{H}_0, \boldsymbol{X}) \le [\tilde{D}(\boldsymbol{H}_0, \boldsymbol{X})]^2 \le k^2\delta^2.$$

Let $u=m\delta^2$ and $v=k^2\delta^2$. Note that as $\hat{\boldsymbol{H}}_\lambda$ is the optimal solution of the penalized problem (12), by (72) and (73) we have

$$D(\boldsymbol{X}, \hat{\boldsymbol{H}}_\lambda) + \lambda D(\hat{\boldsymbol{H}}_\lambda, \boldsymbol{X}) \le u+\lambda v. \tag{74}$$

Note that we have the following:

$$\begin{aligned} D(\hat{\boldsymbol{H}}_\lambda, \boldsymbol{H}_0)^{1/2} &\overset{(a)}{\le} D(\hat{\boldsymbol{H}}_\lambda, \boldsymbol{X}_0)^{1/2} \overset{(b)}{\le} \tilde{D}(\hat{\boldsymbol{H}}_\lambda, \boldsymbol{X}_0) \\ &= \tilde{D}(\hat{\boldsymbol{H}}_\lambda, \boldsymbol{X}-\boldsymbol{Z}) \overset{(c)}{\le} \tilde{D}(\hat{\boldsymbol{H}}_\lambda, \boldsymbol{X})+k\delta \\ &\overset{(d)}{\le} \sqrt{k}D(\hat{\boldsymbol{H}}_\lambda, \boldsymbol{X})^{1/2}+k\delta \overset{(e)}{\le} \sqrt{k}\sqrt{\frac{u}{\lambda}+v}+k\delta, \end{aligned} \tag{75}$$

where (a) is because $\mathrm{Conv}(\boldsymbol{X}_0) \subseteq \mathrm{Conv}(\boldsymbol{H}_0)$, $(b), (d)$ are due to Lemma 1, (c) is due to Lemma 5; and in (e) we use the observation $D(\hat{\boldsymbol{H}}_\lambda, \boldsymbol{X}) \le u/\lambda+v$ (which follows from (74)). This proves (69).

We will now prove (70). We obtain the following set of inequalities:

$$\begin{aligned} D(\boldsymbol{H}_0, \hat{\boldsymbol{H}}_\lambda)^{1/2} &\overset{(a)}{\leq} \tilde{\mathcal{L}}(\boldsymbol{H}_0, \boldsymbol{X}_0) + k\tilde{D}(\boldsymbol{X}_0, \hat{\boldsymbol{H}}) \\ &\overset{(b)}{\leq} 2\sqrt{k^3}\kappa(\boldsymbol{H}_0) D(\boldsymbol{H}_0, \tilde{\boldsymbol{X}}_0)^{1/2} + k\tilde{D}(\boldsymbol{X}_0, \hat{\boldsymbol{H}}) \\ &\overset{(c)}{=} k\tilde{D}(\boldsymbol{X}_0, \hat{\boldsymbol{H}}_\lambda) \\ &\overset{(d)}{\leq} k\tilde{D}(\boldsymbol{X}, \hat{\boldsymbol{H}}_\lambda) + mk\delta \\ &\overset{(e)}{\leq} k\sqrt{m} D(\boldsymbol{X}, \hat{\boldsymbol{H}}_\lambda)^{1/2} + mk\delta \\ &\overset{(f)}{\leq} k\sqrt{m}\sqrt{u + \lambda v} + mk\delta \end{aligned} \tag{76}$$

where (a) is a result of (64), (b) is due to (65), (c) is due to (71), (d) is due to Lemma 5 with $\boldsymbol{X}_0 = \boldsymbol{X} - \boldsymbol{Z}$, (e) is true by Lemma 1; and (f) is a result of (74). This establishes (70). □

Proof of Proposition 1. Part 1) If $\hat{\boldsymbol{H}}_\lambda$ has linearly independent rows, this part of the proposition is a direct result of (69) and (70) together with Lemma 4 with $\boldsymbol{A} = \boldsymbol{H}_0$ and $\boldsymbol{B} = \hat{\boldsymbol{H}}_\lambda$. If $\hat{\boldsymbol{H}}_\lambda$ does not have linearly independent rows, a perturbation argument similar to the proof of Theorem 3 Part 1 suffices.
Part 2) Similar to the proof of Theorem 3, condition (14) guarantees $\kappa(\hat{\boldsymbol{H}}_\lambda) \leq (7/2)\kappa(\boldsymbol{H}_0)$. The rest of the proof follows from (69) and (70) together with Lemma 4 with $\boldsymbol{A} = \hat{\boldsymbol{H}}_\lambda$ and $\boldsymbol{B} = \boldsymbol{H}_0$. □

D.5 Proof of Theorem 4

Proof. Part 1) The proof of convergence is based on Theorem 2 of Xu and Yin (2017). Following Xu and Yin (2017), we define the maximum and minimum of Lipschitz constants across three blocks $(\boldsymbol{H}, \boldsymbol{W}, \tilde{\boldsymbol{W}})$ at iteration j as

$$L_j = \max\{L_1(\boldsymbol{W}_j), L_2(\boldsymbol{H}_j), L_3(\boldsymbol{X})\} \quad \text{and} \quad \ell_j = \min\{L_1(\boldsymbol{W}_j), L_2(\boldsymbol{H}_j), L_3(\boldsymbol{X})\}$$

respectively. By substituting the values of $L_1(\boldsymbol{W}_j), L_2(\boldsymbol{H}_j), L_3(\boldsymbol{X})$,

$$L_j = \max\{2(\|\boldsymbol{W}_j^T \boldsymbol{W}_j\|_2 + \lambda), 2\max\{\|\boldsymbol{H}_j \boldsymbol{H}_j^T\|_2, \varepsilon\}, 2\lambda\|\boldsymbol{X}\boldsymbol{X}^T\|_2\} \tag{77}$$

$$\ell_j = \min\{2(\|\boldsymbol{W}_j^T \boldsymbol{W}_j\|_2 + \lambda), 2\max\{\|\boldsymbol{H}_j \boldsymbol{H}_j^T\|_2, \varepsilon\}, 2\lambda\|\boldsymbol{X}\boldsymbol{X}^T\|_2\}. \tag{78}$$

As $\boldsymbol{W}_j, \tilde{\boldsymbol{W}}_j$ are simplex matrices and bounded, and considering the cost function $\Psi(.,.,.)$ is bounded from above, $\boldsymbol{H}_j$ needs to be bounded. Consequently, L_j is uniformly bounded from above. In addition, by the assumption $\lambda > 0$, ℓ_j is uniformly bounded away from zero. As a result, the Lipschitz constants across three blocks are uniformly bounded from above and bounded away from zero from below—a condition of Theorem 2 of Xu and Yin (2017). Other conditions of Theorem 2 of Xu and Yin (2017) are satisfied and therefore, this implies the convergence of Algorithm 1.
Part 2) Let

$$\boldsymbol{T}_j = \max\{0, \boldsymbol{H}_j - [1/L_1(\boldsymbol{W}_j)](-{\boldsymbol{W}_j}^T[\boldsymbol{X} - \boldsymbol{W}_j\boldsymbol{H}_j] + \lambda[\boldsymbol{H}_j - \tilde{\boldsymbol{W}}_j\boldsymbol{X}])\}. \tag{79}$$

Note that $\boldsymbol{T}_j \to \boldsymbol{T}$ where $\boldsymbol{T}$ is defined in (20). In addition by the assumption of the second part of the theorem on $\boldsymbol{T}$, $P_\ell(\boldsymbol{T})$ is unique. First, we show that $P_\ell(\boldsymbol{T}_j) \to P_\ell(\boldsymbol{T})$. Let us consider two cases:

1. $\|\boldsymbol{T}\|_0 > \ell$: In this case, there exists $i^* \geq 1$ such that the support of $P_\ell(\boldsymbol{T}_i)$ and $P_\ell(\boldsymbol{T})$ are the same for $i \geq i^*$. Therefore, for $i \geq i^*$, for $(r,u) \in S(P_\ell(\boldsymbol{T}_i)) = S(P_\ell(\boldsymbol{T}))$ [recall that $S(\boldsymbol{T})$ is the support of $\boldsymbol{T}$],

$$P_\ell(\boldsymbol{T}_i)_{r,u} = \boldsymbol{T}^i_{r,u} \to \boldsymbol{T}_{r,u} = P_\ell(\boldsymbol{T})_{r,u}$$

and for $(r,u) \in S(P_\ell(\boldsymbol{T}_i))^c = S(P_\ell(\boldsymbol{T}))^c$,

$$P_\ell(\boldsymbol{T}_i)_{r,u} = 0 = P_\ell(\boldsymbol{T})_{r,u}.$$

2. $\|\boldsymbol{T}\|_0 \leq \ell$: In this case, $S(\boldsymbol{T}) = S(P_\ell(\boldsymbol{T}))$ and there exists $i^* \geq 1$ such that for $i \geq i^*$, $S(\boldsymbol{T}) \subseteq S(P_\ell(\boldsymbol{T}_i))$. As a result, for $i \geq i^*$, for $(r,u) \in S(\boldsymbol{T})$,

$$P_\ell(\boldsymbol{T}_i)_{r,u} = \boldsymbol{T}^i_{r,u} \to \boldsymbol{T}_{r,u} = P_\ell(\boldsymbol{T})_{r,u}$$

and for $(r,u) \in S(\boldsymbol{T})^c$,

$$P_\ell(\boldsymbol{T}_i)_{r,u} \in \{0, \boldsymbol{T}^i_{r,u}\}$$

and as $\boldsymbol{T}^i_{r,u} \to \boldsymbol{T}_{r,u} = 0$,

$$P_\ell(\boldsymbol{T}_i)_{r,u} \to P_\ell(\boldsymbol{T})_{r,u} = 0.$$

In addition, note that for any bounded convex set $C \subseteq \mathbb{R}^m$, if $\boldsymbol{x}_i \to \boldsymbol{x}^*$, $P_C(\boldsymbol{x}_i) \to P_C(\boldsymbol{x}^*)$ where P_C is the projection onto C. This along with the fact that $P_\ell(\boldsymbol{T}_j) \to P_\ell(\boldsymbol{T})$, is sufficient to show stationarity:

$$\begin{aligned}
\boldsymbol{H}^* &= \lim_{j\to\infty} \boldsymbol{H}_{j+1} \\
&\overset{(a)}{=} \lim_{j\to\infty} P_\ell\left(\boldsymbol{T}_j\right) \\
&\overset{(b)}{=} P_\ell\left(\boldsymbol{T}\right) \\
&= P_\ell(\max\{0, \boldsymbol{H}^* - [1/L_1(\boldsymbol{W}^*)](-{\boldsymbol{W}^*}^T[\boldsymbol{X} - \boldsymbol{W}^*\boldsymbol{H}^*] + \lambda[\boldsymbol{H}^* - \tilde{\boldsymbol{W}}^*\boldsymbol{X}])\}) \\
&= \underset{\substack{\boldsymbol{H}\geq 0 \\ \|\boldsymbol{H}\|_0 \leq \ell}}{\operatorname{argmin}} \left\| \boldsymbol{H} - \left(\boldsymbol{H}^* - \frac{1}{2L_1(\boldsymbol{W}^*)} \nabla_{\boldsymbol{H}} \Psi(\boldsymbol{H}^*, \boldsymbol{W}^*, \tilde{\boldsymbol{W}}^*) \right) \right\|_F^2
\end{aligned}$$

where (a) is by the definition of the iterate $\boldsymbol{H}_{j+1}$ and (b) was proved above, showing stationarity for the block $\boldsymbol{H}$. □

D.6 Proof of Proposition 2

Proof. Let $\phi(\boldsymbol{H}, \tilde{\boldsymbol{W}}) = \|\boldsymbol{H} - \tilde{\boldsymbol{W}}\boldsymbol{X}\|_F^2$ be the objective function of (21). Suppose $\boldsymbol{H}^*, \tilde{\boldsymbol{W}}^*$ are optimal solutions to problem (21) and let $\boldsymbol{H}' = \boldsymbol{0}$. Suppose j is such that $\|\boldsymbol{X}_{j,.}\|_2 = \min_{u\in[m]} \|\boldsymbol{X}_{u,.}\|_2$ and $\boldsymbol{e}_j \in \mathbb{R}^m$ is the vector with all coordinates equal to zero except coordinate j equal to one. Let $\tilde{\boldsymbol{W}}' = \boldsymbol{1}_k \boldsymbol{e}_j^T$. Hence, $\tilde{\boldsymbol{W}}'\boldsymbol{X} = \boldsymbol{1}_k \boldsymbol{X}_{j,.}$. Note that $\boldsymbol{H}', \tilde{\boldsymbol{W}}'$ are feasible for (21).

We prove the statement of this proposition by the method of contradiction. Suppose there exists $i_0 \in [k]$ such that

$$\|\boldsymbol{H}^*_{i_0,.}\|_2 > \max_{u\in[m]} \|\boldsymbol{X}_{u,.}\|_2 + \sqrt{k} \min_{u\in[m]} \|\boldsymbol{X}_{u,.}\|_2. \tag{80}$$

Note that for any $\boldsymbol{v} \in \text{Conv}(\boldsymbol{X})$, there exists $\alpha_1, \cdots, \alpha_m \geq 0$ such that they sum to one and $\boldsymbol{v} = \sum_{i=1}^m \alpha_i \boldsymbol{X}_{i,.}$. As a result,

$$\|\boldsymbol{v}\|_2 = \|\sum_{i=1}^m \alpha_i \boldsymbol{X}_{i,.}\|_2 \leq \sum_{i=1}^m \alpha_i \|\boldsymbol{X}_{i,.}\|_2 \leq \sum_{i=1}^m \alpha_i \max_{u\in[m]} \|\boldsymbol{X}_{u,.}\|_2 = \max_{u\in[m]} \|\boldsymbol{X}_{u,.}\|_2$$

which when used with (80) leads to:

$$\|\boldsymbol{H}^*_{i_0,.}\|_2 \geq \|\boldsymbol{v}\|_2. \tag{81}$$

One has

$$\begin{aligned}
k \min_{u\in[m]} \|\boldsymbol{X}_{u,.}\|_2^2 &\overset{(a)}{=} \phi(\boldsymbol{H}', \tilde{\boldsymbol{W}}') \\
&\overset{(b)}{\geq} \phi(\boldsymbol{H}^*, \tilde{\boldsymbol{W}}^*) \\
&\geq \|\boldsymbol{H}^*_{i_0,.} - \tilde{\boldsymbol{W}}_{i_0,.}\boldsymbol{X}\|_2^2 \\
&\geq D(\boldsymbol{H}^*_{i_0,.}, \boldsymbol{X}) = \min_{\boldsymbol{v}\in\text{Conv}(\boldsymbol{X})} \|\boldsymbol{H}^*_{i_0,.} - \boldsymbol{v}\|_2^2 \\
&\overset{(c)}{\geq} \min_{\boldsymbol{v}\in\text{Conv}(\boldsymbol{X})} |\|\boldsymbol{H}^*_{i_0,.}\|_2 - \|\boldsymbol{v}\|_2|^2 \overset{(d)}{\geq} |\|\boldsymbol{H}^*_{i_0,.}\|_2 - \max_{\boldsymbol{v}\in\text{Conv}(\boldsymbol{X})} \|\boldsymbol{v}\|_2|^2 \\
&= (\|\boldsymbol{H}^*_{i_0,.}\|_2 - \max_{u\in[m]} \|\boldsymbol{X}_{u,.}\|_2)^2 \\
&\overset{(e)}{>} k \min_{u\in[m]} \|\boldsymbol{X}_{u,.}\|_2^2,
\end{aligned}$$

where (a) is true by definition of $\boldsymbol{H}', \tilde{\boldsymbol{W}}'$, (b) is due to the optimality of $\boldsymbol{H}^*, \tilde{\boldsymbol{W}}^*$, (c) is true as for any two vectors $\boldsymbol{a}, \boldsymbol{b}$, $\|\boldsymbol{a}-\boldsymbol{b}\|_2^2 \geq (\|\boldsymbol{a}\|_2 - \|\boldsymbol{b}\|_2)^2$, (d) is due to (81) and (e) is because of (80). This is a contradiction. Hence, for any $i \in [k]$,

$$\|\boldsymbol{H}^*_i\|_2 \leq \max_{u\in[m]} \|\boldsymbol{X}_{u,.}\|_2 + \sqrt{k} \min_{u\in[m]} \|\boldsymbol{X}_{u,.}\|_2.$$

□

D.7 Proof of Proposition 3

Proof. Part 1) The cost function of (24) is jointly convex in $\boldsymbol{H}, \tilde{\boldsymbol{W}}, \boldsymbol{Z}$ and the feasible set of (24) is convex. Therefore, F is a marginal minimization of a jointly convex function (w.r.t. $\boldsymbol{H}, \tilde{\boldsymbol{W}}$) over a convex set and is convex (see Section 3.2.5 of Boyd et al. (2004)).

Part 2) For the second part, note that we can rewrite (24) as:

$$\begin{aligned} F(\boldsymbol{Z}) = \min_{\boldsymbol{H},\tilde{\boldsymbol{W}},\boldsymbol{U}} \quad & \|\boldsymbol{H}-\boldsymbol{U}\|_F^2 \\ \text{s.t.} \quad & \boldsymbol{H} \geq 0, \quad \tilde{\boldsymbol{W}} \geq 0, \quad \tilde{\boldsymbol{W}}\mathbf{1}_m = \mathbf{1}_k \\ & \boldsymbol{H}_{i,j} \leq \sqrt{b}\boldsymbol{Z}_{i,j} \quad \forall (i,j) \in [k]\times[n] \\ & \boldsymbol{U} = \tilde{\boldsymbol{W}}\boldsymbol{X}. \end{aligned} \tag{82}$$

We start by obtaining the dual problem of (82). Note that by enhanced Slater's condition (Boyd et al., 2004), strong duality holds for this problem. The Lagrangian of (82) can be written as

$$\begin{aligned} L(\boldsymbol{H},\tilde{\boldsymbol{W}},\boldsymbol{U},\boldsymbol{M},\boldsymbol{\Lambda},\boldsymbol{\mu}) =& \|\boldsymbol{H}-\boldsymbol{U}\|_F^2 - \langle \boldsymbol{M}_1,\boldsymbol{H}\rangle - \langle \boldsymbol{M}_2,\tilde{\boldsymbol{W}}\rangle + \langle \boldsymbol{M}_3,\boldsymbol{U}-\tilde{\boldsymbol{W}}\boldsymbol{X}\rangle \\ & + \langle \boldsymbol{\mu},\tilde{\boldsymbol{W}}\mathbf{1}_m - \mathbf{1}_k\rangle + \langle \boldsymbol{\Lambda},\boldsymbol{H}-\sqrt{b}\boldsymbol{Z}\rangle \\ =& \left[\|\boldsymbol{H}-\boldsymbol{U}\|_F^2 - \langle \boldsymbol{M}_1-\boldsymbol{\Lambda},\boldsymbol{H}\rangle + \langle \boldsymbol{M}_3,\boldsymbol{U}\rangle\right] \\ & + \left[-\langle \boldsymbol{M}_2,\tilde{\boldsymbol{W}}\rangle - \langle \boldsymbol{M}_3,\tilde{\boldsymbol{W}}\boldsymbol{X}\rangle + \langle \boldsymbol{\mu},\tilde{\boldsymbol{W}}\mathbf{1}_m\rangle\right] \\ & + \left[-\langle \boldsymbol{\mu},\mathbf{1}_k\rangle - \langle \boldsymbol{\Lambda},\sqrt{b}\boldsymbol{Z}\rangle\right], \end{aligned} \tag{83}$$

where $\boldsymbol{M}_1,\boldsymbol{M}_2,\boldsymbol{M}_3,\boldsymbol{\mu},\boldsymbol{\Lambda}$ are the corresponding Lagrangian variables. By considering the optimality conditions wrt $\boldsymbol{H},\boldsymbol{U},\tilde{\boldsymbol{W}}$, we achieve

$$2(\boldsymbol{H}-\boldsymbol{U}) = \boldsymbol{M}_1 - \boldsymbol{\Lambda}, \tag{84}$$

$$2(\boldsymbol{H}-\boldsymbol{U}) = \boldsymbol{M}_3, \tag{85}$$

$$\boldsymbol{M}_2 + \boldsymbol{M}_3\boldsymbol{X}^T = \boldsymbol{\mu}\mathbf{1}_m^T. \tag{86}$$

Using (84), (85), (86) in (83), we get the dual of (82):

$$\begin{aligned} F(\boldsymbol{Z}) = \max_{\boldsymbol{M}_1,\boldsymbol{M}_2,\boldsymbol{\Lambda}\geq 0,\boldsymbol{M}_3,\boldsymbol{\mu}} \quad & -\frac{1}{4}\|\boldsymbol{M}_3\|_F^2 - \langle \boldsymbol{\mu},\mathbf{1}_k\rangle - \langle \boldsymbol{\Lambda},\sqrt{b}\boldsymbol{Z}\rangle \\ \text{s.t.} \quad & \boldsymbol{M}_2 + \boldsymbol{M}_3\boldsymbol{X}^T = \boldsymbol{\mu}\mathbf{1}_m^T \\ & \boldsymbol{M}_1 - \boldsymbol{\Lambda} = \boldsymbol{M}_3. \end{aligned} \tag{87}$$

At optimality, note that for $(i,j) \in [k]\times[n]$, $\boldsymbol{\Lambda}_{i,j} = -\boldsymbol{M}^3_{i,j}$ if $\boldsymbol{M}^3_{i,j} < 0$ and otherwise $\boldsymbol{\Lambda}_{i,j} = 0$ as $\Lambda_{i,j}, \boldsymbol{Z}_{i,j} \geq 0$ and the cost function is higher if $\Lambda_{i,j}$ is smaller. By using Danskin's Theorem (Bertsekas, 1997), if $\boldsymbol{\Lambda}$ is the optimal solution to (87), $-\sqrt{b}\boldsymbol{\Lambda}$ is a subgradient of F. In addition, at optimality, based on KKT conditions, $\boldsymbol{M}_3 = 2(\boldsymbol{H}-\boldsymbol{U})$ by (85) and $\boldsymbol{U} = \tilde{\boldsymbol{W}}\boldsymbol{X}$ by feasibility, completing the proof. □

References

Vinayak Abrol and Pulkit Sharma. A geometric approach to archetypal analysis via sparse projections. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 42–51. PMLR, 13–18 Jul 2020.

Sanjeev Arora, Rong Ge, Ravindran Kannan, and Ankur Moitra. Computing a nonnegative matrix factorization–provably. In *Proceedings of the forty-fourth annual ACM symposium on Theory of computing*, pages 145–162, 2012.

Francis Bach, Julien Mairal, and Jean Ponce. Convex sparse matrix factorizations. *arXiv preprint arXiv:0812.1869*, 2008.

Amir Beck and Yonina C. Eldar. Sparsity constrained nonlinear optimization: Optimality conditions and algorithms. *SIAM Journal on Optimization*, 23(3):1480–1509, 2013.

Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Img. Sci.*, 2(1):183–202, March 2009.

Michael W Berry and Murray Browne. Email surveillance using non-negative matrix factorization. *Computational & Mathematical Organization Theory*, 11(3):249–264, 2005.

Dimitri P Bertsekas. Nonlinear programming. *Journal of the Operational Research Society*, 48(3):334–334, 1997.

Dimitris Bertsimas and Bart Van Parys. Sparse high-dimensional regression: Exact scalable algorithms and phase transitions. *The Annals of Statistics*, 48(1):300–323, 2020.

Dimitris Bertsimas, Angela King, and Rahul Mazumder. Best subset selection via a modern optimization lens. *The annals of statistics*, pages 813–852, 2016.

Stephen Boyd, Stephen P Boyd, and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

Jean-Philippe Brunet, Pablo Tamayo, Todd R Golub, and Jill P Mesirov. Metagenes and molecular pattern discovery using matrix factorization. *Proceedings of the national academy of sciences*, 101(12):4164–4169, 2004.

Yuansi Chen, Julien Mairal, and Zaid Harchaoui. Fast and robust archetypal analysis for representation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1478–1485, 2014.

M. Chu, F. Diele, R. Plemmons, and S. Ragni. Optimality, computation, and interpretation of nonnegative matrix factorizations. *SIAM Journal on Matrix Analysis*, pages 4–8030, 2004.

Adele Cutler and Leo Breiman. Archetypal analysis. *Technometrics*, 36(4):338–347, 1994.

Anil Damle and Yuekai Sun. A geometric approach to archetypal analysis and nonnegative matrix factorization. *Technometrics*, 59(3):361–370, 2017.

David Donoho and Victoria Stodden. When does non-negative matrix factorization give a correct decomposition into parts? In *Advances in Neural Information Processing Systems 16*, pages 1141–1148. MIT Press, 2004.

John Duchi, Shai Shalev-Shwartz, Yoram Singer, and Tushar Chandra. Efficient projections onto the l1-ball for learning in high dimensions. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, page 272–279, New York, NY, USA, 2008. Association for Computing Machinery. ISBN 9781605582054.

Marco A Duran and Ignacio E Grossmann. An outer-approximation algorithm for a class of mixed-integer nonlinear programs. *Mathematical programming*, 36(3):307–339, 1986.

E. Elhamifar, G. Sapiro, and R. Vidal. See all by looking at a few: Sparse modeling for finding representative objects. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1600–1607, 2012.

Rong Ge and James Zou. Intersecting faces: Non-negative matrix factorization with new guarantees. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2295–2303, Lille, France, 07–09 Jul 2015. PMLR.

N. Gillis and S. A. Vavasis. Fast and robust recursive algorithmsfor separable nonnegative matrix factorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(4):698–714, 2014. doi: 10.1109/TPAMI.2013.226.

Edward F. Gonzalez and Yin Zhang. Accelerating the lee-seung algorithm for nonnegative matrix factorization, 2005.

Hussein Hazimeh and Rahul Mazumder. Fast best subset selection:: Coordinate descent and local combinatorial optimization algorithms. *Operations Research*, 68(5):1517–1537, 2020.

Hussein Hazimeh, Rahul Mazumder, and Ali Saab. Sparse regression at scale: Branch-and-bound rooted in first-order optimization. *arXiv preprint arXiv:2004.06152*, 2020.

Patrik O Hoyer. Non-negative matrix factorization with sparseness constraints. *Journal of machine learning research*, 5(Nov):1457–1469, 2004.

Peter J Huber. *Robust statistics*, volume 523. John Wiley & Sons, 2004.

Hamid Javadi and Andrea Montanari. Nonnegative matrix factorization via archetypal analysis. *Journal of the American Statistical Association*, pages 1–22, 2019.

Mahdi M Kalayeh, Haroon Idrees, and Mubarak Shah. Nmf-knn: Image annotation using weighted multi-view non-negative matrix factorization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 184–191, 2014.

Hyunsoo Kim and Haesun Park. Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis. *Bioinformatics*, 23(12):1495–1502, 05 2007. ISSN 1367-4803.

Jingu Kim and Haesun Park. Sparse nonnegative matrix factorization for clustering. Technical report, Georgia Institute of Technology, 2008.

Philip M Kim and Bruce Tidor. Subsystem identification through dimensionality reduction of large-scale gene expression data. *Genome research*, 13(7):1706–1718, 2003.

Deguang Kong, Chris Ding, and Heng Huang. Robust nonnegative matrix factorization using l21-norm. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 673–682, 2011.

Dylan Kotliar, Adrian Veres, M Aurel Nagy, Shervin Tabrizi, Eran Hodis, Douglas A Melton, and Pardis C Sabeti. Identifying gene expression programs of cell-type identity and cellular activity with single-cell rna-seq. *Elife*, 8, 2019.

William H Lawton and Edward A Sylvestre. Self modeling curve resolution. *Technometrics*, 13(3):617–633, 1971.

Daniel D Lee and H Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

V. Leplat, A. M. S. Ang, and N. Gillis. Minimum-volume rank-deficient nonnegative matrix factorizations. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3402–3406, 2019. doi: 10.1109/ICASSP.2019.8682280.

C. Lin. Projected gradient methods for nonnegative matrix factorization. *Neural Computation*, 19(10): 2756–2779, 2007.

Haifeng Liu, Zhaohui Wu, Xuelong Li, Deng Cai, and Thomas S Huang. Constrained nonnegative matrix factorization for image representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(7):1299–1311, 2011.

Tomohiko Mizutani. Ellipsoidal rounding for nonnegative matrix factorization under noisy separability. *The Journal of Machine Learning Research*, 15(1):1011–1039, 2014.

Morten Mørup and Lars Kai Hansen. Archetypal analysis for machine learning and data mining. *Neurocomputing*, 80:54 – 63, 2012. ISSN 0925-2312. Special Issue on Machine Learning for Signal Processing 2010.

Pentti Paatero and Unto Tapper. Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*, 5(2):111–126, 1994.

Robert Peharz and Franz Pernkopf. Sparse nonnegative matrix factorization with l0-constraints. *Neurocomputing*, 80:38–46, 2012.

Sridhar Ramaswamy, Pablo Tamayo, Ryan Rifkin, Sayan Mukherjee, Chen-Hsiang Yeang, Michael Angelo, Christine Ladd, Michael Reich, Eva Latulippe, Jill P Mesirov, et al. Multiclass cancer diagnosis using tumor gene expression signatures. *Proceedings of the National Academy of Sciences*, 98(26):15149–15154, 2001.

Ben Recht, Christopher Re, Joel Tropp, and Victor Bittorf. Factoring nonnegative matrices with linear programs. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1214–1222. Curran Associates, Inc., 2012.

F. S. Samaria and A. C. Harter. Parameterisation of a stochastic model for human face identification. In *Proceedings of 1994 IEEE Workshop on Applications of Computer Vision*, pages 138–142, 1994.

T. Virtanen. Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(3):1066–1074, 2007. doi: 10.1109/TASL.2006.885253.

Laurence A Wolsey and George L Nemhauser. *Integer and combinatorial optimization*, volume 55. John Wiley & Sons, 1999.

Jonghye Woo, Jerry L Prince, Maureen Stone, Fangxu Xing, Arnold D Gomez, Jordan R Green, Christopher J Hartnick, Thomas J Brady, Timothy G Reese, Van J Wedeen, et al. A sparse non-negative matrix factorization framework for identifying functional units of tongue behavior from mri. *IEEE transactions on medical imaging*, 38(3):730–740, 2018.

Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3485–3492. IEEE, 2010.

Yangyang Xu and Wotao Yin. A globally convergent algorithm for nonconvex optimization based on block coordinate update. *Journal of Scientific Computing*, 72(2):700–734, 2017.

Z. Yang and E. Oja. Linear and nonlinear projective nonnegative matrix factorization. *IEEE Transactions on Neural Networks*, 21(5):734–749, 2010. doi: 10.1109/TNN.2010.2041361.