Supplement To "Non-negative Matrix Factorization via Archetypal Analysis"

Hamid Javadi* and Andrea Montanari†

March 8, 2019

A Further details on numerical experiments

The data in Figures 1, 4, and 5 were generated as follows. We retrieved infrared reflection spectra of caffeine, sucrose, lactose and trioctanoin from the NIST Chemistry WebBook dataset [LM]. We restricted these spectra to the wavenumbers between 1186 cm^{-1} and 1530 cm^{-1}, and denote by $\boldsymbol{h}_{0,1}, \dots, \boldsymbol{h}_{0,4} \in \mathbb{R}^d$, $d = 87$ the vector representations of these spectra. We then generated data $\boldsymbol{x}_i \in \mathbb{R}^d$, $i \le n = 250$ by letting

$$\boldsymbol{x}_i = \sum_{\ell=1}^{4} w_{i,\ell} \boldsymbol{h}_\ell + \boldsymbol{z}_i \,, \tag{A.1}$$

where $\boldsymbol{z}_i \sim \mathsf{N}(0, \sigma^2 \mathrm{I}_d)$ are i.i.d. Gaussian noise vectors. The weights $\boldsymbol{w}_i = (w_{i,\ell})_{\ell \le 4}$ were generated as follows. The weight vectors $\{\boldsymbol{w}_i\}_{1 \le i \le 9}$ are generated such that they have 2 nonzero entries. In other words, 9 data points are on one dimensional facets of the polytope generated by $\boldsymbol{h}_{0,1}, \dots, \boldsymbol{h}_{0,4}$. In order to randomly generate these weight vectors, for each

*Department of Electrical Engineering, Stanford University
†Department of Electrical Engineering and Statistics, Stanford University

$1 \leq i \leq 9$, a pair of indices (ℓ_1, ℓ_2) between 1 and 4 is chosen uniformly at random. Then $\{\widetilde{\boldsymbol{w}}\}_{1\leq i\leq 9}$, $\widetilde{\boldsymbol{w}} \in \mathbb{R}^2$ are generated as independent Dirichlet random vectors with parameter $(5, 5)$. Then we let $w_{i,\ell_1} = \tilde{w}_{i,1}$ and $w_{i,\ell_2} = \tilde{w}_{i,2}$ for $1 \leq i \leq 9$. The weight vectors $\{\boldsymbol{w}_i\}_{10\leq i\leq 20}$ each have 3 nonzero entries. Similar to above, for each of these weight vectors a 3-tuple of indices (ℓ_1, ℓ_2, ℓ_3) between 1 and 4 is chosen uniformly at random. Then we let $w_{i,\ell_1} = \tilde{w}_{i,1}$, $w_{i,\ell_2} = \tilde{w}_{i,2}$, $w_{i,\ell_3} = \tilde{w}_{i,4}$ for $10 \leq i \leq 20$, where $\{\widetilde{\boldsymbol{w}}\}_{10\leq i\leq 20}$, $\widetilde{\boldsymbol{w}} \in \mathbb{R}^3$ are i.i.d. Dirichlet random vectors with parameter $(5, 5, 5)$. The rest of the weight vectors have cardinality equal to 4. Hence, for $21 \leq i \leq 250$, $\boldsymbol{w}_i$ are generated as i.i.d. Dirichlet random vectors with parameter $(5, 5, 5, 5)$.

B Proof of Theorem 1

In this appendix we prove Theorem 1. We start by recalling some notations already defined in the main text, and introducing some new ones. We will then state a stronger form of the theorem (with better dependence on the problem geometry in some regimes). Finally, we will present the actual proof.

Throughout this appendix, we assume the square loss $D(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\|_2^2$.

B.1 Notations and definitions

We use bold capital letters (e.g. $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{C}$,...) for matrices, bold lower case for vectors (e.g. $\boldsymbol{x}$, $\boldsymbol{y}$, ...) and plain lower case for scalars (a, b, c and so on). In particular, $\boldsymbol{e}_i \in \mathbb{R}^d$ denotes the i'th vector in the canonical basis, $E^{r,d} = \{\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_r\}$ and for $r \leq d$, $\boldsymbol{E}_{r,d} \in \{0, 1\}^{r\times d}$ is the matrix whose i'th column is $\boldsymbol{e}_i$, and whose columns after the r-th one are equal to $\boldsymbol{0}$. For a matrix $\boldsymbol{X}$, $\boldsymbol{X}_{i,\cdot}$ and $\boldsymbol{X}_{\cdot,i}$ are its i'th row and column, respectively.

As in the main text, we denote by Δ^m the $(m-1)$-dimensional standard simplex, i.e.

$\Delta^m = \{\boldsymbol{x} \in \mathbb{R}^m_{\geq 0}, \langle \boldsymbol{x}, \mathbf{1} \rangle = 1\}$, where $\mathbf{1} \in \mathbb{R}^m$ is the all ones vector. For a matrix $\boldsymbol{H} \in \mathbb{R}^{r \times d}$, we use $\sigma_{\max}(\boldsymbol{H})$, $\sigma_{\min}(\boldsymbol{H})$ to denote its largest and smallest nonzero singular values and $\kappa(\boldsymbol{H}) = \sigma_{\max}(\boldsymbol{H})/\sigma_{\min}(\boldsymbol{H})$ to denote its condition number. We denote by $\mathrm{conv}(\boldsymbol{H}), \mathrm{aff}(\boldsymbol{H})$ the convex hull and the affine hull of the rows of $\boldsymbol{H}$, respectively. In other words,

$$\mathrm{conv}(\boldsymbol{H}) = \{\boldsymbol{x} \in \mathbb{R}^d : \boldsymbol{x} = \boldsymbol{H}^\mathsf{T}\boldsymbol{\pi}, \boldsymbol{\pi} \in \Delta^r\}, \tag{B.1}$$

$$\mathrm{aff}(\boldsymbol{H}) = \{\boldsymbol{x} \in \mathbb{R}^d : \boldsymbol{x} = \boldsymbol{H}^\mathsf{T}\boldsymbol{\alpha}, \langle \mathbf{1}, \boldsymbol{\alpha} \rangle = 1\}. \tag{B.2}$$

We denote by $Q_{r,n}$ the set of r by n row stochastic matrices. Namely,

$$Q_{r,n} = \left\{\boldsymbol{\Pi} \in \mathbb{R}^{r \times n}_{\geq 0} : \langle \boldsymbol{\Pi}_{i,\cdot}, \mathbf{1} \rangle = 1\right\}. \tag{B.3}$$

with use $Q_r \equiv Q_{r,r}$. Further, S_r is defined as

$$S_r = \left\{\boldsymbol{\Pi} \in Q_r : \Pi_{i,j} \in \{0, 1\}\right\}. \tag{B.4}$$

As a consequence, given $\boldsymbol{X} \in \mathbb{R}^{n \times d}$, $\boldsymbol{H}_1, \boldsymbol{H}_2 \in \mathbb{R}^{r \times d}$, the loss functions $\mathscr{D}(\,\cdot\,,\,\cdot\,)$ and $\mathscr{L}(\,\cdot\,,\,\cdot\,)$ take the form

$$\mathscr{D}(\boldsymbol{H}_1, \boldsymbol{X}) = \min_{\boldsymbol{\Pi} \in Q_{r,n}} \|\boldsymbol{H}_1 - \boldsymbol{\Pi}\boldsymbol{X}\|_F^2, \tag{B.5}$$

$$\mathscr{L}(\boldsymbol{H}_1, \boldsymbol{H}_2) = \min_{\boldsymbol{\Pi} \in S_r} \|\boldsymbol{H}_1 - \boldsymbol{\Pi}\boldsymbol{H}_2\|_F^2. \tag{B.6}$$

We use $\mathsf{B}_m(\rho)$ to denote the closed ball with radius ρ in m dimensions, centered at 0. In addition, for $\boldsymbol{H} \in \mathbb{R}^{m \times d}$ we define the ρ-neighborhood of $\mathrm{conv}(\boldsymbol{H})$ as

$$\mathsf{B}_r(\rho; \boldsymbol{H}) := \{\boldsymbol{x} \in \mathbb{R}^d : \mathscr{D}(\boldsymbol{x}, \boldsymbol{H}) \leq \rho^2\}. \tag{B.7}$$

For a convex set $\mathcal{C}$ we denote the set of its extremal points by $\mathrm{ext}(\mathcal{C})$ and the projection of a point $\boldsymbol{x} \in \mathbb{R}^d$ onto $\mathcal{C}$ by $\boldsymbol{\Pi}_{\mathcal{C}}(\boldsymbol{x})$. Namely,

$$\boldsymbol{\Pi}_{\mathcal{C}}(\boldsymbol{x}) = \arg\min_{\boldsymbol{y} \in \mathcal{C}} \|\boldsymbol{x} - \boldsymbol{y}\|_2. \tag{B.8}$$

Further, we use $D(\boldsymbol{x}, \mathcal{C})$ to denote the distance of $\boldsymbol{x}$ from $\mathcal{C}$, i.e.

$$D(\boldsymbol{x}, \mathcal{C}) = \|\boldsymbol{x} - \boldsymbol{\Pi}_{\mathcal{C}}(\boldsymbol{x})\|_2. \tag{B.9}$$

Also, for a matrix $\boldsymbol{X} \in \mathbb{R}^{n\times d}$, and a mapping (not necessarily linear) $\boldsymbol{P} : \mathbb{R}^d \to \mathbb{R}^d$, $\boldsymbol{P}(\boldsymbol{X}) \in \mathbb{R}^{n\times d}$ is the matrix whose i'th row is $\boldsymbol{P}(\boldsymbol{X}_{i,\cdot})$.

B.2 Theorem statement

The statement below provides more detailed result with respect to the one in Theorem 1.

Theorem 1. *Assume $\boldsymbol{X} = \boldsymbol{W}_0\boldsymbol{H}_0 + \boldsymbol{Z}$ where the factorization $\boldsymbol{X}_0 = \boldsymbol{W}_0\boldsymbol{H}_0$ satisfies the uniqueness assumption with parameter $\alpha > 0$, and that $\mathrm{conv}(\boldsymbol{X}_0)$ has internal radius $\mu > 0$. Consider the estimator $\widehat{\boldsymbol{H}}$ defined by Eq. (2.5), with $D(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\|_2^2$ (square loss) and $\delta = \max_{i\le n} \|\boldsymbol{Z}_{i,\cdot}\|_2$. If*

$$\max_{i\le n} \|\boldsymbol{Z}_{i,\cdot}\|_2 \le \frac{\alpha\mu}{30r^{3/2}}\,, \tag{B.10}$$

then, setting $\delta = \max_{i\le n} \|\boldsymbol{Z}_{i,\cdot}\|_2$ in the problem (2.5) we get

$$\mathscr{L}(\boldsymbol{H}_0, \widehat{\boldsymbol{H}}) \le \frac{C_*^2\, r^5}{\alpha^2} \max_{i\le n} \|\boldsymbol{Z}_{i,\cdot}\|_2^2\,, \tag{B.11}$$

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

Further, if

$$\max_{i\le n} \|\boldsymbol{Z}_{i,\cdot}\|_2 \le \frac{\alpha\mu}{330\kappa(\boldsymbol{H}_0)r^{5/2}}, \tag{B.12}$$

then, setting $\delta = \max_{i\le n} \|\boldsymbol{Z}_{i,\cdot}\|_2$ in the problem (2.5) we get

$$\mathscr{L}(\boldsymbol{H}_0, \widehat{\boldsymbol{H}}) \le \frac{C_{**}^2\, r^4}{\alpha^2} \max_{i\le n} \|\boldsymbol{Z}_{i,\cdot}\|_2^2\,, \tag{B.13}$$

where $C_{**} = 120\kappa(\boldsymbol{H}_0)\max(\kappa(\boldsymbol{H}_0), (\sigma_{\max}(\boldsymbol{H}_0)/r + \|\boldsymbol{z}_0\|_2)/(\mu r^{1/2}))$.

B.3 Proof

B.4 Proof strategy

Before providing a detailed proof, let us explain the proof scheme and main intuition:

1. Notice that $\widehat{\boldsymbol{H}}$ minimizes the distance $\mathscr{D}(\boldsymbol{H};\boldsymbol{X})$ from the data, among all the possible sets of archetypes that can explain the data itself (in the sense that $\boldsymbol{x}_i$ is close to $\mathrm{conv}(\boldsymbol{H})$), cf. Eq. (2.5). Using the fact that $\boldsymbol{X}$ is the perturbed version of $\boldsymbol{X}_0$, we show in Lemma B.5 below that

$$\mathscr{D}(\widehat{\boldsymbol{H}},\boldsymbol{X}_0)^{1/2} \leq \mathscr{D}(\boldsymbol{H}_0,\boldsymbol{X}_0)^{1/2} + 3\delta\sqrt{r}. \tag{B.14}$$

2. In absence of separability, there might be multiple (non-equivalent) sets of archetypes that minimize $\mathscr{D}(\boldsymbol{H};\boldsymbol{X})$, and hence reconstruction is fundamentally non-unique. We combine the separability assumption with the bound in step 1 to bound

$$\alpha(\mathscr{D}(\widehat{\boldsymbol{H}},\boldsymbol{H}_0)^{1/2} + \mathscr{D}(\boldsymbol{H}_0,\widehat{\boldsymbol{H}})^{1/2}) \tag{B.15}$$

   in terms of $\delta, \kappa(\widehat{\boldsymbol{H}}), \sigma_{\max}(\widehat{\boldsymbol{H}})$.

3. The last step gives us an error in terms of distances between convex hulls, $\mathscr{D}(\widehat{\boldsymbol{H}},\boldsymbol{H}_0)$, $\mathscr{D}(\boldsymbol{H}_0,\widehat{\boldsymbol{H}})$. Lemma B.2 translates this into an upper bound on $\mathscr{L}(\boldsymbol{H}_0,\widehat{\boldsymbol{H}})$ in terms of $\delta, \kappa(\widehat{\boldsymbol{H}}), \sigma_{\max}(\widehat{\boldsymbol{H}})$.

4. Finally, we want to translate this upper bound in terms of geometric properties of the true archetypes (as opposed to properties of the estimated archetypes, namely $\kappa(\widehat{\boldsymbol{H}}), \sigma_{\max}(\widehat{\boldsymbol{H}})$). In Lemmas B.3, B.4 we bound the quantities $\kappa(\widehat{\boldsymbol{H}}), \sigma_{\max}(\widehat{\boldsymbol{H}})$ in terms of $\mathscr{D}(\widehat{\boldsymbol{H}},\boldsymbol{H}_0)^{1/2}$, $\mathscr{D}(\boldsymbol{H}_0,\widehat{\boldsymbol{H}})^{1/2}$ as well as $\kappa(\boldsymbol{H}_0), \sigma_{\max}(\boldsymbol{H}_0)$

5. Finally, we aggregate the last three steps to finish the proof.

B.4.1 Lemmas

In the following two lemmas we bound the notion we use for estimation error $\mathscr{L}(\boldsymbol{H}_0, \widehat{\boldsymbol{H}})$, that we defined in (3.2), in terms of $\mathscr{D}(\widehat{\boldsymbol{H}}, \boldsymbol{H}_0)$, $\mathscr{D}(\boldsymbol{H}_0, \widehat{\boldsymbol{H}})$.

Lemma B.1. *Let $\mathcal{R}$ be a convex set and $\mathcal{C}$ be a convex cone. Define*

$$\gamma_{\mathcal{C}} = \max_{\|\boldsymbol{u}\|_2=1} \min_{\boldsymbol{v}\in\mathcal{C}, \|\boldsymbol{v}\|_2=1} \langle \boldsymbol{u}, \boldsymbol{v} \rangle. \tag{B.16}$$

We have

$$\min_{\boldsymbol{x}\in\mathcal{R}} \|\boldsymbol{x}\|_2 + (1+\gamma_{\mathcal{C}}) \max_{x\in\mathrm{ext}(\mathcal{R})} \|\boldsymbol{x} - \boldsymbol{\Pi}_{\mathcal{C}}(\boldsymbol{x})\|_2 \geq \gamma_{\mathcal{C}} \min_{\boldsymbol{x}\in\mathrm{ext}(\mathcal{R})} \|\boldsymbol{x}\|_2. \tag{B.17}$$

An illustration of this lemma in the case of $\mathcal{R} \subset \mathcal{C}$ is given in Figure 1. Note that, $\gamma_{\mathcal{C}}$ measures the pointedness of the cone $\mathcal{C}$. Geometrically (for $\mathcal{R} \subseteq \mathcal{C}$) the lemma states that the cosine of the angle between $\arg\min_{\boldsymbol{x}\in\mathcal{R}} \|\boldsymbol{x}\|_2$ and $\arg\min_{\boldsymbol{x}\in\mathrm{ext}(R)} \|\boldsymbol{x}\|_2$ is smaller than $\gamma_{\mathcal{C}}$.

Proof. Note that the claim is trivial in the case where $\gamma_{\mathcal{C}} \leq 0$, as by Cauchy-Schwarz inequality $\gamma_{\mathcal{C}} \geq -1$ and hence the left hand side of (B.17) is nonnegative. Therefore, we focus on the case where $\gamma_{\mathcal{C}} > 0$, i.e. $\mathcal{C}^*$, the dual cone of $\mathcal{C}$ has a nonempty interior. We write

$$\min_{\boldsymbol{x}\in\mathcal{R}} \|\boldsymbol{x}\|_2 = \min_{\boldsymbol{x}\in\mathcal{R}} \max_{\|\boldsymbol{u}\|_2=1} \langle \boldsymbol{u}, \boldsymbol{x} \rangle \geq \max_{\|\boldsymbol{u}\|_2=1} \min_{\boldsymbol{x}\in\mathcal{R}} \langle \boldsymbol{u}, \boldsymbol{x} \rangle = \max_{\|\boldsymbol{u}\|_2=1} \min_{\boldsymbol{x}\in\mathrm{ext}(\mathcal{R})} \langle \boldsymbol{u}, \boldsymbol{x} \rangle. \tag{B.18}$$

Replacing

$$\boldsymbol{x} = \boldsymbol{\Pi}_{\mathcal{C}}(\boldsymbol{x}) + (\boldsymbol{x} - \boldsymbol{\Pi}_{\mathcal{C}}(\boldsymbol{x})), \tag{B.19}$$

we get

$$\min_{\boldsymbol{x}\in\mathcal{R}} \|\boldsymbol{x}\|_2 \geq \max_{\|\boldsymbol{u}\|_2=1} \min_{\boldsymbol{x}\in\mathrm{ext}(\mathcal{R})} \langle \boldsymbol{u}, \boldsymbol{\Pi}_{\mathcal{C}}(\boldsymbol{x}) + (\boldsymbol{x} - \boldsymbol{\Pi}_{\mathcal{C}}(\boldsymbol{x})) \rangle \tag{B.20}$$

$$\geq \max_{\|\boldsymbol{u}\|_2=1} \min_{\boldsymbol{x}\in\mathrm{ext}(\mathcal{R})} \langle \boldsymbol{u}, \boldsymbol{\Pi}_{\mathcal{C}}(\boldsymbol{x}) \rangle - \max_{\boldsymbol{x}\in\mathrm{ext}(\mathcal{R})} \|\boldsymbol{x} - \boldsymbol{\Pi}_{\mathcal{C}}(\boldsymbol{x})\|_2. \tag{B.21}$$

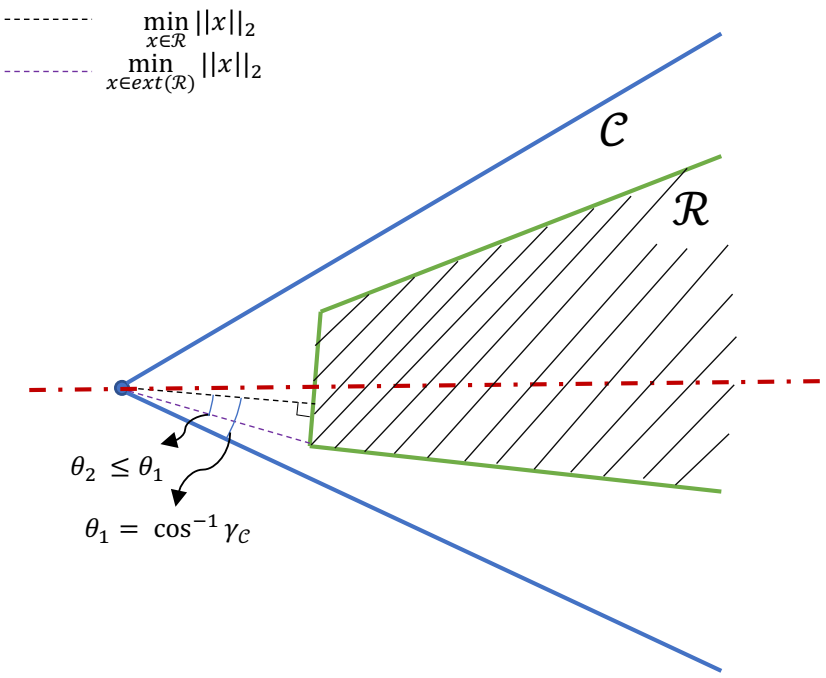

Figure 1: Picture of Lemma B.1, in the case, $\mathcal{R} \subset \mathcal{C}$.

Note that $\widetilde{\mathcal{R}}^*$, the dual cone of the set $\widetilde{\mathcal{R}} \subseteq \mathcal{C}$ where $\widetilde{\mathcal{R}} := \{\mathbf{\Pi}_{\mathcal{C}}(\boldsymbol{x})\,;\boldsymbol{x} \in \mathrm{ext}(\mathcal{R})\}$, contains $\mathcal{C}^*$ and hence it has a nonempty interior. Therefore, the maximizer of the first term in the right hand side of (B.21) is in $\widetilde{\mathcal{R}}^*$ and

$$\min_{\boldsymbol{x}\in\mathcal{R}} \|\boldsymbol{x}\|_2 \geq \max_{\boldsymbol{u}\in\widetilde{\mathcal{R}}^*;\|\boldsymbol{u}\|_2=1} \min_{\boldsymbol{x}\in\mathrm{ext}(\mathcal{R})} \langle \boldsymbol{u}, \mathbf{\Pi}_{\mathcal{C}}(\boldsymbol{x})\rangle - \max_{\boldsymbol{x}\in\mathrm{ext}(\mathcal{R})} \|\boldsymbol{x} - \mathbf{\Pi}_{\mathcal{C}}(\boldsymbol{x})\|_2 \tag{B.22}$$

$$= \max_{\boldsymbol{u}\in\widetilde{\mathcal{R}}^*;\|\boldsymbol{u}\|_2=1} \min_{\boldsymbol{x}\in\mathrm{ext}(\mathcal{R})} \left[\left\langle \boldsymbol{u}, \frac{\mathbf{\Pi}_{\mathcal{C}}(\boldsymbol{x})}{\|\mathbf{\Pi}_{\mathcal{C}}(\boldsymbol{x})\|_2}\right\rangle \|\mathbf{\Pi}_{\mathcal{C}}(\boldsymbol{x})\|_2\right] - \max_{\boldsymbol{x}\in\mathrm{ext}(\mathcal{R})} \|\boldsymbol{x} - \mathbf{\Pi}_{\mathcal{C}}(\boldsymbol{x})\|_2. \tag{B.23}$$

Since for $\boldsymbol{u} \in \widetilde{\mathcal{R}}^*, \boldsymbol{x} \in \mathrm{ext}(\mathcal{R})$ we have $\langle \boldsymbol{u}, \mathbf{\Pi}_{\mathcal{C}}(\boldsymbol{x})\rangle \geq 0$, using the definition of $\gamma_{\mathcal{C}}$, we have

$$\min_{\boldsymbol{x}\in\mathcal{R}} \|\boldsymbol{x}\|_2 \geq \gamma_{\mathcal{C}} \min_{\boldsymbol{x}\in\mathrm{ext}(\mathcal{R})} \|\mathbf{\Pi}_{\mathcal{C}}(\boldsymbol{x})\|_2 - \max_{\boldsymbol{x}\in\mathrm{ext}(\mathcal{R})} \|\boldsymbol{x} - \mathbf{\Pi}_{\mathcal{C}}(\boldsymbol{x})\|_2. \tag{B.24}$$

Note that

$$\|\boldsymbol{\Pi}_{\mathcal{C}}(\boldsymbol{x})\|_2 \geq \|\boldsymbol{x}\|_2 - \|\boldsymbol{x} - \boldsymbol{\Pi}_{\mathcal{C}}(\boldsymbol{x})\|_2. \tag{B.25}$$

Therefore,

$$\min_{\boldsymbol{x}\in\mathcal{R}} \|\boldsymbol{x}\|_2 \geq \gamma_{\mathcal{C}} \min_{\boldsymbol{x}\in\text{ext}(\mathcal{R})} \|\boldsymbol{x}\|_2 - (1+\gamma_{\mathcal{C}}) \max_{\boldsymbol{x}\in\text{ext}(\mathcal{R})} \|\boldsymbol{x} - \boldsymbol{\Pi}_{\mathcal{C}}(\boldsymbol{x})\|_2, \tag{B.26}$$

and this completes the proof. □

The next lemma is a consequence of Lemma B.1.

Lemma B.2. *Let $\boldsymbol{H}, \boldsymbol{H}_0 \in \mathbb{R}^{r\times d}$, $r \leq d$, be matrices with linearly independent rows. We have*

$$\mathscr{L}(\boldsymbol{H}_0, \boldsymbol{H})^{1/2} \leq \sqrt{2}\kappa(\boldsymbol{H}_0)\mathscr{D}(\boldsymbol{H}_0, \boldsymbol{H})^{1/2} + (1+\sqrt{2})\sqrt{r}\kappa(\boldsymbol{H}_0)\mathscr{D}(\boldsymbol{H}, \boldsymbol{H}_0)^{1/2}. \tag{B.27}$$

Proof. Consider the cone $\mathcal{C}_1 \subset \mathbb{R}^d$, generated by vectors $\boldsymbol{e}_2 - \boldsymbol{e}_1, \ldots, \boldsymbol{e}_r - \boldsymbol{e}_1 \in \mathbb{R}^d$, i.e,

$$\mathcal{C}_1 = \left\{ \boldsymbol{v} \in \mathbb{R}^d; \boldsymbol{v} = \sum_{i=2}^{r} v_i(\boldsymbol{e}_i - \boldsymbol{e}_1), v_i \geq 0 \right\}. \tag{B.28}$$

For $\boldsymbol{v} \in \mathcal{C}_1, \|\boldsymbol{v}\|_2 = 1$ we have

$$\boldsymbol{v} = \left(-\langle \mathbf{1}, \boldsymbol{x}\rangle, \boldsymbol{x}, 0, 0, \ldots, 0\right), \tag{B.29}$$

where $\boldsymbol{x} \in \mathbb{R}^{r-1}_{\geq 0}$ and

$$\|\boldsymbol{x}\|_2^2 + \langle \mathbf{1}, \boldsymbol{x}\rangle^2 = 1. \tag{B.30}$$

Since, $\langle \mathbf{1}, \boldsymbol{x}\rangle = \|\boldsymbol{x}\|_1 \geq \|\boldsymbol{x}\|_2$, we get $\langle \mathbf{1}, \boldsymbol{x}\rangle \geq 1/\sqrt{2}$. Thus, for $\boldsymbol{u} = -\boldsymbol{e}_1$, we have $\langle \boldsymbol{u}, \boldsymbol{v}\rangle \geq 1/\sqrt{2}$. Therefore, for $\gamma_{\mathcal{C}_1}$ defined as in Lemma B.1, we have $\gamma_{\mathcal{C}_1} \geq 1/\sqrt{2}$. In addition, by symmetry, for $i \in \{1, 2, \ldots, r\}$, for the cone $\mathcal{C}_i \subset \mathbb{R}^d$, generated by vectors

$\boldsymbol{e}_1 - \boldsymbol{e}_i, \boldsymbol{e}_2 - \boldsymbol{e}_i, \ldots, \boldsymbol{e}_r - \boldsymbol{e}_i \in \mathbb{R}^d$ we have $\gamma_{\mathcal{C}_i} = \gamma \geq 1/\sqrt{2}$. Hence, using Lemma B.1 for $\boldsymbol{H} \in \mathbb{R}^{r \times d}$, $\mathcal{R} = \mathrm{conv}(\boldsymbol{H}) - \boldsymbol{e}_j$ (the set obtained by translating $\mathrm{conv}(\boldsymbol{H})$ by $-\boldsymbol{e}_j$), $\mathcal{C} = \mathcal{C}_j$ we get for $j = 1, 2, \ldots, r$

$$\begin{aligned}\min_{\boldsymbol{q} \in \Delta^r} \|\boldsymbol{e}_j - \boldsymbol{H}^\mathsf{T}\boldsymbol{q}\|_2 &\geq \gamma \min_{\boldsymbol{q} \in E^{r,r}} \|\boldsymbol{e}_j - \boldsymbol{H}^\mathsf{T}\boldsymbol{q}\|_2 \\ &\quad - (1+\gamma) \max_{i \in [r]} \min_{\boldsymbol{q} \in \mathbb{R}^r_{\geq 0}} \|\boldsymbol{H}^\mathsf{T}_{i,\cdot} - \boldsymbol{e}_j - \boldsymbol{E}^\mathsf{T}_{r,d}\boldsymbol{q} + \boldsymbol{e}_j \langle \mathbf{1}, \boldsymbol{q} \rangle\|_2 \qquad \text{(B.31)} \\ &\geq \gamma \min_{\boldsymbol{q} \in E^{r,r}} \|\boldsymbol{e}_j - \boldsymbol{H}^\mathsf{T}\boldsymbol{q}\|_2 - (1+\gamma) \max_{i \in [r]} \min_{\boldsymbol{q} \in \Delta^r} \|\boldsymbol{H}^\mathsf{T}_{i,\cdot} - \boldsymbol{E}^\mathsf{T}_{r,d}\boldsymbol{q}\|_2. \qquad \text{(B.32)}\end{aligned}$$

Hence, for $j = 1, 2, \ldots, r$

$$\min_{\boldsymbol{q} \in \Delta^r} \|\boldsymbol{e}_j - \boldsymbol{H}^\mathsf{T}\boldsymbol{q}\|_2 + (1+\gamma) \max_{i \in [r]} \min_{\boldsymbol{q} \in \Delta^r} \|\boldsymbol{H}^\mathsf{T}_{i,\cdot} - \boldsymbol{E}_{r,d}\|_2 \geq \gamma \min_{\boldsymbol{q} \in E^{r,r}} \|\boldsymbol{e}_j - \boldsymbol{H}^\mathsf{T}\boldsymbol{q}\|_2. \qquad \text{(B.33)}$$

For simplicity, define $\boldsymbol{a}, \mathbf{b} \in \mathbb{R}^r_{\geq 0}$, $c \geq 0$ as

$$a_j \equiv \min_{\boldsymbol{q} \in \Delta^r} \|\boldsymbol{e}_j - \boldsymbol{H}^\mathsf{T}\boldsymbol{q}\|_2 \,, b_j := \gamma \min_{\boldsymbol{q} \in E^{r,r}} \|\boldsymbol{e}_j - \boldsymbol{H}^\mathsf{T}\boldsymbol{q}\|_2 \,, c := (1+\gamma) \max_{i \in [r]} \min_{\boldsymbol{q} \in \Delta^r} \|\boldsymbol{H}^\mathsf{T}_{i,\cdot} - \boldsymbol{E}_{r,d}\|_2 \,. \qquad \text{(B.34)}$$

Using triangle inequality,

$$\|\boldsymbol{a}\|_2 + c\sqrt{r} \geq \|\boldsymbol{a} + c\mathbf{1}\|_2 \,. \qquad \text{(B.35)}$$

Using definition of $\boldsymbol{a}, \mathbf{b}, c$, (B.33) implies that $a_j + c \geq b_j$. Therefore, since $a_j, b_j, c \geq 0$,

$$\|\boldsymbol{a} + c\mathbf{1}\|_2 \geq \|\mathbf{b}\|_2. \qquad \text{(B.36)}$$

Hence, using (B.35)

$$\|\boldsymbol{a}\|_2 \geq \|\mathbf{b}\|_2 - c\sqrt{r}. \qquad \text{(B.37)}$$

In other words,

$$\min_{\boldsymbol{Q}\in Q_r} \|\boldsymbol{E}_{r,d} - \boldsymbol{Q}\boldsymbol{H}\|_F \geq \gamma \min_{\boldsymbol{Q}\in S_r} \|\boldsymbol{E}_{r,d} - \boldsymbol{Q}\boldsymbol{H}\|_F - (1+\gamma)\sqrt{r} \max_{i\in[r]} \min_{\boldsymbol{q}\in\Delta^r} \|\boldsymbol{H}_{i,\cdot}^{\mathsf{T}} - \boldsymbol{E}_{r,d}^{\mathsf{T}}\boldsymbol{q}\|_2. \tag{B.38}$$

Now consider $\boldsymbol{H}_0 \in \mathbb{R}^{r\times d}$ where $\boldsymbol{H}_0 = \boldsymbol{E}_{r,d}\boldsymbol{M}$, $\boldsymbol{H} = \boldsymbol{Y}\boldsymbol{M}$, where $\boldsymbol{M} \in \mathbb{R}^{d\times d}$ is invertible. We have

$$\begin{aligned}
\mathscr{D}(\boldsymbol{H}_0, \boldsymbol{H})^{1/2} &= \min_{\boldsymbol{Q}\in Q_r} \|\boldsymbol{H}_0 - \boldsymbol{Q}\boldsymbol{H}\|_F = \min_{\boldsymbol{Q}\in Q_r} \|(\boldsymbol{E}_{r,d} - \boldsymbol{Q}\boldsymbol{Y})\boldsymbol{M}\|_F && \text{(B.39)}\\
&\geq \sigma_{\min}(\boldsymbol{M}) \min_{\boldsymbol{Q}\in Q_r} \|\boldsymbol{E}_{r,d} - \boldsymbol{Q}\boldsymbol{Y}\|_F && \text{(B.40)}\\
&\geq \gamma\sigma_{\min}(\boldsymbol{M}) \min_{\boldsymbol{Q}\in S_r} \|\boldsymbol{E}_{r,d} - \boldsymbol{Q}\boldsymbol{Y}\|_F \\
&\quad - \sigma_{\min}(\boldsymbol{M})\sqrt{r}(1+\gamma) \max_{i\in[r]} \min_{\boldsymbol{q}\in\Delta^r} \|\boldsymbol{Y}_{i,\cdot}^{\mathsf{T}} - \boldsymbol{E}_{r,d}^{\mathsf{T}}\boldsymbol{q}\|_2 && \text{(B.41)}\\
&= \gamma\sigma_{\min}(\boldsymbol{M}) \min_{\boldsymbol{Q}\in S_r} \|(\boldsymbol{H}_0 - \boldsymbol{Q}\boldsymbol{H})\boldsymbol{M}^{-1}\|_F \\
&\quad - \sigma_{\min}(\boldsymbol{M})\sqrt{r}(1+\gamma) \max_{i\in[r]} \min_{\boldsymbol{q}\in\Delta^r} \|(\boldsymbol{M}^{-1})^{\mathsf{T}}(\boldsymbol{H}_{i,\cdot}^{\mathsf{T}} - \boldsymbol{H}_0^{\mathsf{T}}\boldsymbol{q})\|_2. && \text{(B.42)}
\end{aligned}$$

Thus, using the fact that $\sigma_{\max}(\boldsymbol{M})/\sigma_{\min}(\boldsymbol{M}) = \kappa(\boldsymbol{M}) = \kappa(\boldsymbol{H}_0)$,

$$\begin{aligned}
\mathscr{D}(\boldsymbol{H}_0, \boldsymbol{H})^{1/2} &\geq \frac{\gamma}{\kappa(\boldsymbol{H}_0)}\mathscr{L}(\boldsymbol{H}_0, \boldsymbol{H})^{1/2} - (1+\gamma)\sqrt{r} \min_{\boldsymbol{q}\in\Delta^r} \|\boldsymbol{H}_{i,\cdot}^{\mathsf{T}} - \boldsymbol{H}_0^{\mathsf{T}}\boldsymbol{q}\|_2 && \text{(B.43)}\\
&\geq \frac{\gamma}{\kappa(\boldsymbol{H}_0)}\mathscr{L}(\boldsymbol{H}_0, \boldsymbol{H})^{1/2} - (1+\gamma)\sqrt{r}(\boldsymbol{H}, \boldsymbol{H}_0)^{1/2}\,. && \text{(B.44)}
\end{aligned}$$

Therefore,

$$\mathscr{L}(\boldsymbol{H}_0, \boldsymbol{H})^{1/2} \leq \frac{\kappa(\boldsymbol{H}_0)}{\gamma}\mathscr{D}(\boldsymbol{H}_0, \boldsymbol{H})^{1/2} + \frac{(1+\gamma)\kappa(\boldsymbol{H}_0)\sqrt{r}}{\gamma}\mathscr{D}(\boldsymbol{H}, \boldsymbol{H}_0)^{1/2}\,. \tag{B.45}$$

Finally, note that the function $f(x) = (1+x)/x$ is monotone decreasing over $\mathbb{R}_{>0}$. Hence, for $\gamma \geq 1/\sqrt{2}$, $(1+\gamma)/\gamma \leq 1+\sqrt{2}$. Therefore, we get

$$\mathscr{L}(\boldsymbol{H}_0, \boldsymbol{H})^{1/2} \leq \sqrt{2}\kappa(\boldsymbol{H}_0)\mathscr{D}(\boldsymbol{H}_0, \boldsymbol{H})^{1/2} + (1+\sqrt{2})\sqrt{r}\kappa(\boldsymbol{H}_0)\mathscr{D}(\boldsymbol{H}, \boldsymbol{H}_0)^{1/2} \tag{B.46}$$

and this completes the proof. □

We continue with the following lemmas on the condition number of the matrix $\boldsymbol{H}$.

Lemma B.3. *Let $\boldsymbol{H}_0, \boldsymbol{H} \in \mathbb{R}^{r\times d}$, $r \le d$,with $\boldsymbol{H}$ having full row rank. We have*

$$\sigma_{\max}(\boldsymbol{H}) \le \mathscr{D}(\boldsymbol{H}, \boldsymbol{H}_0)^{1/2} + \sqrt{r}\sigma_{\max}(\boldsymbol{H}_0), . \tag{B.47}$$

In addition, if

$$\mathscr{D}(\boldsymbol{H}_0, \boldsymbol{H})^{1/2} \le \frac{\sigma_{\min}(\boldsymbol{H}_0)}{2}, \tag{B.48}$$

then

$$\kappa(\boldsymbol{H}) \le \frac{2r\sigma_{\max}(\boldsymbol{H}_0) + 2\mathscr{D}(\boldsymbol{H}, \boldsymbol{H}_0)^{1/2}\sqrt{r}}{\sigma_{\min}(\boldsymbol{H}_0)}. \tag{B.49}$$

Further, if

$$\mathscr{D}(\boldsymbol{H}, \boldsymbol{H}_0)^{1/2} + \mathscr{D}(\boldsymbol{H}_0, \boldsymbol{H})^{1/2} \le \frac{\sigma_{\min}(\boldsymbol{H}_0)}{6\sqrt{r}}, \tag{B.50}$$

then

$$\sigma_{\max}(\boldsymbol{H}) \le 2\sigma_{\max}(\boldsymbol{H}_0), \tag{B.51}$$

$$\kappa(\boldsymbol{H}) \le (7/2)\kappa(\boldsymbol{H}_0). \tag{B.52}$$

Proof. For the sake of simplicity, we will write $\mathscr{D}_1 = \mathscr{D}(\boldsymbol{H}, \boldsymbol{H}_0)^{1/2}$, $\mathscr{D}_2 = \mathscr{D}(\boldsymbol{H}_0, \boldsymbol{H})^{1/2}$ Note that using the assumptions of Lemma B.3 we have

$$\begin{aligned} \boldsymbol{H}_0 &= \boldsymbol{P}\boldsymbol{H} + \boldsymbol{A}_2; \quad \|\boldsymbol{A}_2\|_F = \mathscr{D}_2, \\ \boldsymbol{H} &= \boldsymbol{R}\boldsymbol{H}_0 + \boldsymbol{A}_1; \quad \|\boldsymbol{A}_1\|_F = \mathscr{D}_1, \end{aligned} \tag{B.53}$$

where $\boldsymbol{P}, \boldsymbol{R} \in \mathbb{R}^{r\times r}_{\ge 0}$ are row-stochastic matrices and $\boldsymbol{A}_1, \boldsymbol{A}_2 \in \mathbb{R}^{r\times d}$. Also, $\sigma_{\max}(\boldsymbol{A}_1) \le \|\boldsymbol{A}_1\|_F = \mathscr{D}_1$, $\sigma_{\max}(\boldsymbol{A}_2) \le \|\boldsymbol{A}_2\|_F = \mathscr{D}_2$. Therefore,

$$\begin{aligned} \sigma_{\max}(\boldsymbol{P})\sigma_{\min}(\boldsymbol{H}) &\ge \sigma_{\min}(\boldsymbol{P}\boldsymbol{H}) = \sigma_{\min}(\boldsymbol{H}_0 - \boldsymbol{A}_2) \\ &\ge \sigma_{\min}(\boldsymbol{H}_0) - \sigma_{\max}(\boldsymbol{A}_2) \ge \sigma_{\min}(\boldsymbol{H}_0) - \mathscr{D}_2. \end{aligned} \tag{B.54}$$

In addition, note that for a row stochastic matrix $\boldsymbol{P} \in Q_r$, we have

$$\sigma_{\max}(\boldsymbol{P}) \leq \|\boldsymbol{P}\|_F = \left(\sum_{i=1}^{r} \|\boldsymbol{P}_{i,.}\|_2^2\right)^{1/2} \leq \left(\sum_{i=1}^{r} \|\boldsymbol{P}_{i,.}\|_1^2\right)^{1/2} \leq \sqrt{r}. \tag{B.55}$$

Hence, for $\mathscr{D}_2 \leq \sigma_{\min}(\boldsymbol{H}_0)$ we get

$$\sigma_{\min}(\boldsymbol{H}) \geq \frac{\sigma_{\min}(\boldsymbol{H}_0) - \mathscr{D}_2}{\sqrt{r}}\,. \tag{B.56}$$

In addition, using (B.53)

$$\begin{aligned}\sigma_{\max}(\boldsymbol{H}) = \sigma_{\max}(\boldsymbol{R}\boldsymbol{H}_0 + \boldsymbol{A}_1) &\leq \sigma_{\max}(\boldsymbol{R}\boldsymbol{H}_0) + \sigma_{\max}(\boldsymbol{A}_1) \\ &\leq \sigma_{\max}(\boldsymbol{R})\sigma_{\max}(\boldsymbol{H}_0) + \mathscr{D}_1 \leq \sqrt{r}\sigma_{\max}(\boldsymbol{H}_0) + \mathscr{D}_1. \end{aligned} \tag{B.57}$$

Hence, using (B.56), (B.57), for $\mathscr{D}_2 \leq \sigma_{\min}(\boldsymbol{H}_0)$ we have

$$\kappa(\boldsymbol{H}) = \frac{\sigma_{\max}(\boldsymbol{H})}{\sigma_{\min}(\boldsymbol{H})} \leq \frac{r\sigma_{\max}(\boldsymbol{H}_0) + \mathscr{D}_1\sqrt{r}}{\sigma_{\min}(\boldsymbol{H}_0) - \mathscr{D}_2}. \tag{B.58}$$

Thus, for $\mathscr{D}_2 \leq \sigma_{\min}(\boldsymbol{H}_0)/2$, by replacing $\mathscr{D}_2$ with $\sigma_{\min}(\boldsymbol{H}_0)/2$ in (B.58) we get Eqs. (B.47), (B.49).

Now assume that $\mathscr{D}_1 + \mathscr{D}_2 \leq \sigma_{\min}(\boldsymbol{H}_0)/(6\sqrt{r})$. In this case, using (B.53) we have

$$\boldsymbol{H}_0 = \boldsymbol{P}(\boldsymbol{R}\boldsymbol{H}_0 + \boldsymbol{A}_1) + \boldsymbol{A}_2. \tag{B.59}$$

Therefore,

$$(\mathbf{I} - \boldsymbol{P}\boldsymbol{R})\boldsymbol{H}_0 = \boldsymbol{P}\boldsymbol{A}_1 + \boldsymbol{A}_2, \tag{B.60}$$

hence,

$$\mathbf{I} - \boldsymbol{P}\boldsymbol{R} = (\boldsymbol{P}\boldsymbol{A}_1 + \boldsymbol{A}_2)\boldsymbol{H}_0^{\dagger} \tag{B.61}$$

and

$$\boldsymbol{P}\boldsymbol{R} = \mathbf{I} - \boldsymbol{P}\boldsymbol{A}_1\boldsymbol{H}_0^\dagger - \boldsymbol{A}_2\boldsymbol{H}_0^\dagger. \tag{B.62}$$

where $\boldsymbol{H}_0^\dagger$ is the right inverse of matrix $\boldsymbol{H}_0$. Note that

$$\sigma_{\max}(\boldsymbol{H}_0^\dagger) = \sigma_{\min}(\boldsymbol{H}_0)^{-1}. \tag{B.63}$$

By permuting the rows and columns of $\boldsymbol{H}_0$, without loss of generality, we can assume that $R_{ii} = \|\boldsymbol{R}_{\cdot,i}\|_\infty$. We can write

$$R_{ii} = \|\boldsymbol{R}_{\cdot,i}\|_\infty \geq \langle \boldsymbol{P}_{i,\cdot}, \boldsymbol{R}_{\cdot,i}\rangle = (\boldsymbol{P}\boldsymbol{R})_{ii} = 1 - (\boldsymbol{P}\boldsymbol{A}_1\boldsymbol{H}_0^\dagger)_{ii} - (\boldsymbol{A}_2\boldsymbol{H}_0^\dagger)_{ii} \tag{B.64}$$

$$\geq 1 - \|(\boldsymbol{P}\boldsymbol{A}_1\boldsymbol{H}_0^\dagger)_{i,\cdot}\|_2 - \|(\boldsymbol{A}_2\boldsymbol{H}_0^\dagger)_{i,\cdot}\|_2 \tag{B.65}$$

$$\geq 1 - \max_{\boldsymbol{u}\in\Delta^r} \|\boldsymbol{A}_1^\mathsf{T}\boldsymbol{u}\|_2\sigma_{\max}(\boldsymbol{H}_0^\dagger) - \|(\boldsymbol{A}_2)_{i,\cdot}\|_2\sigma_{\max}(\boldsymbol{H}_0^\dagger) \tag{B.66}$$

$$\geq 1 - \max_{\boldsymbol{u}\in\Delta^r} \|\boldsymbol{u}\|_2\sigma_{\max}(\boldsymbol{A}_1)\sigma_{\max}(\boldsymbol{H}_0^\dagger) - \|\boldsymbol{A}_2\|_F\sigma_{\max}(\boldsymbol{H}_0^\dagger) \tag{B.67}$$

$$\geq 1 - \frac{\mathscr{D}_1 + \mathscr{D}_2}{\sigma_{\min}(\boldsymbol{H}_0)}. \tag{B.68}$$

Hence, for all $i, j \in [r], i \neq j$, since $\boldsymbol{R}$ is row-stochastic,

$$R_{ji} \leq \frac{\mathscr{D}_1 + \mathscr{D}_2}{\sigma_{\min}(\boldsymbol{H}_0)}. \tag{B.69}$$

Thus,

$$\langle \boldsymbol{P}_{i,\cdot}, \boldsymbol{R}_{\cdot,i}\rangle = R_{ii}P_{ii} + \sum_{j\neq i} P_{ij}R_{ji} \leq R_{ii}P_{ii} + \left(\max_{j\neq i} R_{ji}\right)\sum_{j\neq i} P_{ij} \tag{B.70}$$

$$\leq P_{ii} + \frac{\mathscr{D}_1 + \mathscr{D}_2}{\sigma_{\min}(\boldsymbol{H}_0)}(1 - P_{ii}). \tag{B.71}$$

Therefore, using (B.68),

$$P_{ii} \geq \frac{\sigma_{\min}(\boldsymbol{H}_0) - 2(\mathscr{D}_1 + \mathscr{D}_2)}{\sigma_{\min}(\boldsymbol{H}_0) - (\mathscr{D}_1 + \mathscr{D}_2)}. \tag{B.72}$$

Thus, we can write

$$\boldsymbol{P} = \mathbf{I} + \Delta; \quad \|\Delta_{i,.}\|_1 \leq \frac{2(\mathscr{D}_1 + \mathscr{D}_2)}{\sigma_{\min}(\boldsymbol{H}_0) - (\mathscr{D}_1 + \mathscr{D}_2)}. \tag{B.73}$$

Therefore,

$$\sigma_{\max}(\Delta) \leq \|\Delta\|_F = \left(\sum_{i=1}^{r} \|\Delta_{i,.}\|_2^2\right)^{1/2} \leq \left(\sum_{i=1}^{r} \|\Delta_{i,.}\|_1^2\right)^{1/2} \leq \frac{2(\mathscr{D}_1 + \mathscr{D}_2)\sqrt{r}}{\sigma_{\min}(\boldsymbol{H}_0) - (\mathscr{D}_1 + \mathscr{D}_2)}. \tag{B.74}$$

Hence,

$$\sigma_{\max}(\boldsymbol{P}) \leq 1 + \frac{2\sqrt{r}(\mathscr{D}_1 + \mathscr{D}_2)}{\sigma_{\min}(\boldsymbol{H}_0) - (\mathscr{D}_1 + \mathscr{D}_2)}, \quad \sigma_{\min}(\boldsymbol{P}) \geq 1 - \frac{2\sqrt{r}(\mathscr{D}_1 + \mathscr{D}_2)}{\sigma_{\min}(\boldsymbol{H}_0) - (\mathscr{D}_1 + \mathscr{D}_2)}. \tag{B.75}$$

From (B.53) we have $\sigma_{\min}(\boldsymbol{PH}) \geq \sigma_{\min}(\boldsymbol{H}_0) - \mathscr{D}_2$. Using $\sigma_{\min}(\boldsymbol{PH}) \leq \sigma_{\max}(\boldsymbol{P})\sigma_{\min}(\boldsymbol{H})$, we get

$$\sigma_{\min}(\boldsymbol{H}) \geq \frac{(\sigma_{\min}(\boldsymbol{H}_0) - \mathscr{D}_2)(\sigma_{\min}(\boldsymbol{H}_0) - (\mathscr{D}_1 + \mathscr{D}_2))}{\sigma_{\min}(\boldsymbol{H}_0) - (\mathscr{D}_1 + \mathscr{D}_2) + 2\sqrt{r}(\mathscr{D}_1 + \mathscr{D}_2)}. \tag{B.76}$$

Further, from (B.53) we have $\sigma_{\max}(\boldsymbol{PH}) \leq \sigma_{\max}(\boldsymbol{H}_0) + \mathscr{D}_2$. Therefore, using the fact that $\sigma_{\max}(\boldsymbol{PH}) \geq \sigma_{\min}(\boldsymbol{P})\sigma_{\max}(\boldsymbol{H})$, we get

$$\sigma_{\max}(\boldsymbol{H}) \leq \frac{(\sigma_{\max}(\boldsymbol{H}_0) + \mathscr{D}_2)(\sigma_{\min}(\boldsymbol{H}_0) - (\mathscr{D}_1 + \mathscr{D}_2))}{\sigma_{\min}(\boldsymbol{H}_0) - (\mathscr{D}_1 + \mathscr{D}_2) - 2\sqrt{r}(\mathscr{D}_1 + \mathscr{D}_2)}. \tag{B.77}$$

Hence, for $\mathscr{D}_1 + \mathscr{D}_2 \leq \sigma_{\min}(\boldsymbol{H}_0)/(6\sqrt{r})$, we have $\sigma_{\max}(\boldsymbol{H}) \leq 35\sigma_{\max}(\boldsymbol{H}_0)/18 < 2\sigma_{\max}(\boldsymbol{H}_0)$. In addition,

$$\kappa(\boldsymbol{H}) \leq \left(\frac{\sigma_{\max}(\boldsymbol{H}_0) + \mathscr{D}_2}{\sigma_{\min}(\boldsymbol{H}_0) - \mathscr{D}_2}\right)\left(1 + \frac{4\sqrt{r}(\mathscr{D}_1 + \mathscr{D}_2)}{\sigma_{\min}(\boldsymbol{H}_0) - (\mathscr{D}_1 + \mathscr{D}_2) - 2\sqrt{r}(\mathscr{D}_1 + \mathscr{D}_2)}\right) \tag{B.78}$$

$$\leq \frac{6\kappa(\boldsymbol{H}_0) + 1}{5}\left(1 + \frac{4}{3}\right) \leq \frac{42\kappa(\boldsymbol{H}_0) + 7}{15} < \frac{7\kappa(\boldsymbol{H}_0)}{2}, \tag{B.79}$$

and this completes the proof. □

Lemma B.4. *Let $\boldsymbol{X}_0 = \boldsymbol{W}_0\boldsymbol{H}_0 \in \mathbb{R}^{n\times d}$ be such that $\mathrm{conv}(\boldsymbol{X}_0)$ has internal radius at least $\mu > 0$, and $\boldsymbol{X} = \boldsymbol{X}_0 + \boldsymbol{Z}$ with $\max_{i\le n}\|\boldsymbol{Z}_{i,\cdot}\|_2 \le \delta$. If $\boldsymbol{H} \in \mathbb{R}^{r\times d}$, $\boldsymbol{H}_{i,\cdot} \in \mathrm{aff}(\boldsymbol{H}_0)$ is feasible for problem (2.5) and has linearly independent rows, then we have*

$$\sigma_{\min}(\boldsymbol{H}) \ge \sqrt{2}(\mu - 2\delta)\,. \tag{B.80}$$

Proof. Let

$$\boldsymbol{X}'_{i,\cdot} = \boldsymbol{\Pi}_{\mathrm{conv}(\boldsymbol{H})}(\boldsymbol{X}_{i,\cdot}) \equiv \arg\min_{\boldsymbol{x}\in\mathrm{conv}(\boldsymbol{H})} \|\boldsymbol{X}_{i,\cdot} - \boldsymbol{x}\|_2\,. \tag{B.81}$$

Note that since $\boldsymbol{H}$ is feasible for problem (2.5) and $\max_{i\le n}\|\boldsymbol{Z}_{i,\cdot}\|_2 \le \delta$

$$\|(\boldsymbol{X}_0)_{i,\cdot} - \boldsymbol{X}'_{i,\cdot}\|_2 \le \|(\boldsymbol{X}_0)_{i,\cdot} - \boldsymbol{X}_{i,\cdot}\|_2 + \|\boldsymbol{X}_{i,\cdot} - \boldsymbol{X}'_{i,\cdot}\|_2 \le 2\delta. \tag{B.82}$$

Therefore, for any $\boldsymbol{x}_0 \in \mathrm{conv}(\boldsymbol{X}_0)$, writing $\boldsymbol{x}_0 = \boldsymbol{X}_0^\mathsf{T}\boldsymbol{a}_0$, $\boldsymbol{a}_0 \in \Delta^n$, we have

$$\mathscr{D}(\boldsymbol{x}_0, \boldsymbol{X}')^{1/2} = \min_{\boldsymbol{a}\in\Delta^n} \left\|\boldsymbol{X}_0^\mathsf{T}\boldsymbol{a}_0 - \boldsymbol{X}'^\mathsf{T}\boldsymbol{a}\right\|_2 \le \left\|\boldsymbol{X}_0^\mathsf{T}\boldsymbol{a}_0 - \boldsymbol{X}'^\mathsf{T}\boldsymbol{a}_0\right\|_2 \tag{B.83}$$

$$\le \left(\sum_{i=1}^n (a_0)_i\right) \|(\boldsymbol{X}_0)_{i,\cdot} - \boldsymbol{X}'_{i,\cdot}\|_2 \le 2\delta. \tag{B.84}$$

Since $\mathrm{conv}(\boldsymbol{X}_0)$ has internal radius at least μ, there exists $\boldsymbol{z}_0 \in \mathbb{R}^d$, and an orthogonal matrix $\boldsymbol{U} \in \mathbb{R}^{d\times r'}$, $r' = r-1$, such that $\boldsymbol{z}_0 + \boldsymbol{U}\mathsf{B}_{r'}(\mu) \subseteq \mathrm{conv}(\boldsymbol{X}_0)$. Hence, for every $\boldsymbol{z} \in \mathbb{R}^{r'}$, $\|\boldsymbol{z}\|_2 = 1$ there exists $\boldsymbol{a} \in \Delta^n$ such that

$$\mu\boldsymbol{U}\boldsymbol{z} + \boldsymbol{z}_0 = \boldsymbol{X}_0^\mathsf{T}\boldsymbol{a}. \tag{B.85}$$

Therefore, for any unit vector $\boldsymbol{u}$ in column space of $\boldsymbol{U}$, for the line segment

$$l_{\boldsymbol{u},\mu} = \{\boldsymbol{z} : \boldsymbol{z} = \boldsymbol{z}_0 + \alpha\boldsymbol{u}, |\alpha| \le \mu\} \subseteq \mathrm{conv}(\boldsymbol{X}_0)\,. \tag{B.86}$$

Thus,

$$l_{\boldsymbol{u},\mu} \subseteq \boldsymbol{P}_{\boldsymbol{u}}(\mathrm{conv}(\boldsymbol{X}_0)) \tag{B.87}$$

where $\boldsymbol{P}_{\boldsymbol{u}}$ is the orthogonal projection onto the line containing $l_{\boldsymbol{u},\mu}$. Note that since $\boldsymbol{P}_{\boldsymbol{u}}(\cdot)$ is a nonexpansive mapping, using (B.84), for any $\boldsymbol{x}_0 \in \mathrm{conv}(\boldsymbol{X}_0)$ we have

$$D(\boldsymbol{P}_{\boldsymbol{u}}(\boldsymbol{x}_0), \boldsymbol{P}_{\boldsymbol{u}}(\mathrm{conv}(\boldsymbol{X}')))^{1/2} \leq \mathscr{D}(\boldsymbol{x}_0, \boldsymbol{X}')^{1/2} \leq 2\delta. \tag{B.88}$$

In other words, for any $\boldsymbol{x}_0 \in \boldsymbol{P}_{\boldsymbol{u}}(\mathrm{conv}(\boldsymbol{X}_0))$, $D(\boldsymbol{x}_0, \boldsymbol{P}_{\boldsymbol{u}}(\mathrm{conv}(\boldsymbol{X}'))) \leq 2\delta$. Therefore, using (B.86) for any $\boldsymbol{u}$ in column space of $\boldsymbol{U}$, we have

$$l_{\boldsymbol{u},\mu-2\delta} \subseteq \boldsymbol{P}_u(\mathrm{conv}(\boldsymbol{X}')). \tag{B.89}$$

This implies that

$$\boldsymbol{z}_0 + \boldsymbol{U}\mathsf{B}_{r'}(\mu - 2\delta) \subseteq \mathrm{conv}(\boldsymbol{X}') \subseteq \mathrm{conv}(\boldsymbol{H}). \tag{B.90}$$

Hence, for every $\boldsymbol{z} \in \mathbb{R}^{r'}$, $\|\boldsymbol{z}\|_2 = 1$ there exists $\boldsymbol{a} \in \Delta^r$ such that

$$(\mu - 2\delta)\boldsymbol{U}\boldsymbol{z} + \boldsymbol{z}_0 = \boldsymbol{H}^{\mathsf{T}}\boldsymbol{a}. \tag{B.91}$$

Note that $\boldsymbol{H}^{\mathsf{T}}$ has linearly independent columns. Multiplying the previous equation by $(\boldsymbol{H}^{\mathsf{T}})^{\dagger}$ the left inverse of $\boldsymbol{H}^{\mathsf{T}}$, we get

$$(\mu - 2\delta)(\boldsymbol{H}^{\mathsf{T}})^{\dagger}\boldsymbol{U}\boldsymbol{z} + (\boldsymbol{H}^{\mathsf{T}})^{\dagger}\boldsymbol{z}_0 = \boldsymbol{a}. \tag{B.92}$$

Let

$$\boldsymbol{a}_1 = (\mu - 2\delta)(\boldsymbol{H}^{\mathsf{T}})^{\dagger}\boldsymbol{U}\boldsymbol{v} + (\boldsymbol{H}^{\mathsf{T}})^{\dagger}\boldsymbol{z}_0\,, \tag{B.93}$$

$$\boldsymbol{a}_2 = -(\mu - 2\delta)(\boldsymbol{H}^{\mathsf{T}})^{\dagger}\boldsymbol{U}\boldsymbol{v} + (\boldsymbol{H}^{\mathsf{T}})^{\dagger}\boldsymbol{z}_0\,, \tag{B.94}$$

where $\boldsymbol{v}$ is the right singular vector corresponding to the largest singular value of $(\boldsymbol{H}^{\mathsf{T}})^{\dagger}\boldsymbol{U}$. Therefore, we have

$$\boldsymbol{a}_1 = (\mu - 2\delta)\sigma_{\max}((\boldsymbol{H}^{\mathsf{T}})^{\dagger}\boldsymbol{U})\boldsymbol{v} + (\boldsymbol{H}^{\mathsf{T}})^{\dagger}\boldsymbol{z}_0, \tag{B.95}$$

$$\boldsymbol{a}_2 = -(\mu - 2\delta)\sigma_{\max}((\boldsymbol{H}^{\mathsf{T}})^{\dagger}\boldsymbol{U})\boldsymbol{v} + (\boldsymbol{H}^{\mathsf{T}})^{\dagger}\boldsymbol{z}_0. \tag{B.96}$$

Thus, for $\boldsymbol{a}_1, \boldsymbol{a}_2 \in \Delta^r$

$$\|\boldsymbol{a}_1 - \boldsymbol{a}_2\|_2 = 2(\mu - 2\delta)\sigma_{\max}((\boldsymbol{H}^\mathsf{T})^\dagger \boldsymbol{U}). \tag{B.97}$$

Note that

$$\|\boldsymbol{a}_1 - \boldsymbol{a}_2\|_2 \leq \sqrt{2}. \tag{B.98}$$

Thus,

$$2(\mu - 2\delta)\sigma_{\max}((\boldsymbol{H}^\mathsf{T})^\dagger \boldsymbol{U}) = \frac{2(\mu - 2\delta)}{\sigma_{\min}(\boldsymbol{H})} \leq \sqrt{2}. \tag{B.99}$$

Hence,

$$\sigma_{\min}(\boldsymbol{H}) \geq \sqrt{2}(\mu - 2\delta). \tag{B.100}$$

□

The following lemma states an important property of $\widehat{\boldsymbol{H}}$ the optimal solution of problem (2.5).

Lemma B.5. *If* $\max_i \|\boldsymbol{Z}_{i,\cdot}\|_2 \leq \delta$ *and* $\widehat{\boldsymbol{H}}$ *is the optimal solution of problem* (2.5)*, then we have*

$$\mathscr{D}(\widehat{\boldsymbol{H}}, \boldsymbol{X}_0)^{1/2} \leq \mathscr{D}(\boldsymbol{H}_0, \boldsymbol{X}_0)^{1/2} + 3\delta\sqrt{r}. \tag{B.101}$$

Proof. First note that $\boldsymbol{X} = \boldsymbol{W}_0\boldsymbol{H}_0 + \boldsymbol{Z}$ and the rows of $\boldsymbol{W}_0\boldsymbol{H}_0$ are in $\mathrm{conv}(\boldsymbol{H}_0)$. Thus, since $\max_i \|\boldsymbol{Z}_{i,\cdot}\|_2 \leq \delta$, we have

$$\max_{i\leq n} \mathscr{D}(\boldsymbol{X}_{i,\cdot}, \boldsymbol{H}_0)^{1/2} \leq \max_{i\leq n} \left\|(\boldsymbol{X}_0 - \boldsymbol{W}_0\boldsymbol{H}_0)_{i,\cdot}\right\|_2 = \max_{i\leq n} \|\boldsymbol{Z}_{i,\cdot}\|_2 \leq \delta. \tag{B.102}$$

Hence, $\boldsymbol{H}_0$ is a feasible solution for the problem (2.5). Therefore, we have

$$\mathscr{D}(\widehat{\boldsymbol{H}}, \boldsymbol{X}) \leq \mathscr{D}(\boldsymbol{H}_0, \boldsymbol{X}). \tag{B.103}$$

Letting $\tilde{\boldsymbol{\alpha}}_i = \arg\min_{\boldsymbol{\alpha} \in \Delta^n} \|\widehat{\boldsymbol{H}}_{i,\cdot}^{\mathsf{T}} - \boldsymbol{X}^{\mathsf{T}}\boldsymbol{\alpha}\|_2$, we have

$$\begin{aligned}
\mathscr{D}(\widehat{\boldsymbol{H}}, \boldsymbol{X}) &= \sum_{i=1}^{r} \min_{\boldsymbol{\alpha}_i \in \Delta^r} \|\widehat{\boldsymbol{H}}_{i,\cdot}^{\mathsf{T}} - \boldsymbol{X}^{\mathsf{T}}\boldsymbol{\alpha}_i\|_2^2 && \text{(B.104)}\\
&= \sum_{i=1}^{r} \min_{\boldsymbol{\alpha}_i \in \Delta^r} \|\widehat{\boldsymbol{H}}_{i,\cdot}^{\mathsf{T}} - \boldsymbol{X}_0^{\mathsf{T}}\boldsymbol{\alpha}_i - \boldsymbol{Z}^{\mathsf{T}}\boldsymbol{\alpha}_i\|_2^2 && \text{(B.105)}\\
&= \sum_{i=1}^{r} \min_{\boldsymbol{\alpha}_i \in \Delta^r} \left(\|\widehat{\boldsymbol{H}}_{i,\cdot}^{\mathsf{T}} - \boldsymbol{X}_0^{\mathsf{T}}\boldsymbol{\alpha}_i\|_2^2 - 2\left\langle \boldsymbol{Z}^{\mathsf{T}}\boldsymbol{\alpha}_i, \widehat{\boldsymbol{H}}_{i,\cdot}^{\mathsf{T}} - \boldsymbol{X}_0^{\mathsf{T}}\boldsymbol{\alpha}_i \right\rangle + \|\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{\alpha}_i\|_2^2 \right) && \text{(B.106)}\\
&= \sum_{i=1}^{r} \left(\|\widehat{\boldsymbol{H}}_{i,\cdot}^{\mathsf{T}} - \boldsymbol{X}_0^{\mathsf{T}}\tilde{\boldsymbol{\alpha}}_i\|_2^2 - 2\left\langle \boldsymbol{Z}^{\mathsf{T}}\tilde{\boldsymbol{\alpha}}_i, \widehat{\boldsymbol{H}}_{i,\cdot}^{\mathsf{T}} - \boldsymbol{X}_0^{\mathsf{T}}\tilde{\boldsymbol{\alpha}}_i \right\rangle + \|\boldsymbol{Z}^{\mathsf{T}}\tilde{\boldsymbol{\alpha}}_i\|_2^2 \right). && \text{(B.107)}
\end{aligned}$$

Using the fact that (by triangle inequality) $\|\boldsymbol{Z}^{\mathsf{T}}\tilde{\boldsymbol{\alpha}}_i\|_2 \leq (\max_i \|\boldsymbol{Z}_{i,\cdot}\|_2) \|\tilde{\boldsymbol{\alpha}}_i\|_1 \leq \delta$, we have

$$\begin{aligned}
\mathscr{D}(\widehat{\boldsymbol{H}}, \boldsymbol{X}) &\geq \sum_{i=1}^{r} \left(\|\widehat{\boldsymbol{H}}_{i,\cdot}^{\mathsf{T}} - \boldsymbol{X}_0^{\mathsf{T}}\tilde{\boldsymbol{\alpha}}_i\|_2^2 - 2\delta \|\widehat{\boldsymbol{H}}_{i,\cdot}^{\mathsf{T}} - \boldsymbol{X}_0^{\mathsf{T}}\tilde{\boldsymbol{\alpha}}_i\|_2 \right) && \text{(B.108)}\\
&\geq U^2 - 2\delta\sqrt{r}U && \text{(B.109)}
\end{aligned}$$

where $U^2 = \sum_{i=1}^{r} \|\widehat{\boldsymbol{H}}_{i,\cdot}^{\mathsf{T}} - \boldsymbol{X}_0^{\mathsf{T}}\tilde{\boldsymbol{\alpha}}_i\|_2^2$. In addition, since $U \geq 0$, $\mathscr{D}(\widehat{\boldsymbol{H}}, \boldsymbol{X}) \geq 0$,

$$\mathscr{D}(\widehat{\boldsymbol{H}}, \boldsymbol{X}) \geq (U^2 - 2\delta\sqrt{r}U)\mathbb{I}_{U \geq 2\delta\sqrt{r}}. \tag{B.110}$$

Note that for $U \geq 2\delta\sqrt{r}$, the function $U^2 - 2\delta\sqrt{r}U$ is increasing. Hence, since

$$U \geq \left(\sum_{i=1}^{r} \min_{\boldsymbol{\alpha}_i} \|\widehat{\boldsymbol{H}}_{i,\cdot}^{\mathsf{T}} - \boldsymbol{X}_0^{\mathsf{T}}\boldsymbol{\alpha}_i\|_2^2 \right)^{1/2} = \mathscr{D}(\widehat{\boldsymbol{H}}, \boldsymbol{X}_0)^{1/2}, \tag{B.111}$$

we have

$$\mathscr{D}(\widehat{\boldsymbol{H}}, \boldsymbol{X}) \geq (U^2 - 2\delta\sqrt{r}U)\mathbb{I}_{U \geq 2\delta\sqrt{r}} \geq \mathscr{D}(\widehat{\boldsymbol{H}}, \boldsymbol{X}_0) - 2\delta\sqrt{r}\mathscr{D}(\widehat{\boldsymbol{H}}, \boldsymbol{X}_0)^{1/2}. \tag{B.112}$$

Therefore, letting $(x)_+ := \max\{x, 0\}$ be the positive part of x, using the fact that for all $x, a \geq 0$ we have $(x^2 - ax)_+^{1/2} \geq x - a$, we have

$$\mathscr{D}(\widehat{\boldsymbol{H}}, \boldsymbol{X})^{1/2} \geq \left(\mathscr{D}(\widehat{\boldsymbol{H}}, \boldsymbol{X}_0) - 2\delta\sqrt{r}\mathscr{D}(\widehat{\boldsymbol{H}}, \boldsymbol{X}_0)^{1/2}\right)_+^{1/2} \geq \mathscr{D}(\widehat{\boldsymbol{H}}, \boldsymbol{X}_0)^{1/2} - 2\delta\sqrt{r}. \quad \text{(B.113)}$$

In addition,

$$\begin{aligned}
\mathscr{D}(\boldsymbol{H}_0, \boldsymbol{X}) &= \sum_{i=1}^{r} \min_{\boldsymbol{\alpha}_i \in \Delta^n} \|(\boldsymbol{H}_0)_{i,\cdot} - \boldsymbol{X}_0^\mathsf{T}\boldsymbol{\alpha}_i - \boldsymbol{Z}^\mathsf{T}\boldsymbol{\alpha}_i\|_2^2 && \text{(B.114)}\\
&\leq \sum_{i=1}^{r} \min_{\boldsymbol{\alpha}_i \in \Delta^n} \left\{\|(\boldsymbol{H}_0)_{i,\cdot} - \boldsymbol{X}_0^\mathsf{T}\boldsymbol{\alpha}_i\|_2 + \|\boldsymbol{Z}^\mathsf{T}\boldsymbol{\alpha}_i\|_2\right\}^2 && \text{(B.115)}\\
&\leq \sum_{i=1}^{r} \left\{\min_{\boldsymbol{\alpha}_i \in \Delta^n} \|(\boldsymbol{H}_0)_{i,\cdot} - \boldsymbol{X}_0^\mathsf{T}\boldsymbol{\alpha}_i\|_2 + \max_{\boldsymbol{\alpha}_i \in \Delta^n} \|\boldsymbol{Z}^\mathsf{T}\boldsymbol{\alpha}_i\|_2\right\}^2 && \text{(B.116)}\\
&\leq \left\{\left(\sum_{i=1}^{r} \min_{\boldsymbol{\alpha}_i \in \Delta^n} \|(\boldsymbol{H}_0)_{i,\cdot} - \boldsymbol{X}_0^\mathsf{T}\boldsymbol{\alpha}_i\|_2^2\right)^{1/2} + \delta\sqrt{r}\right\}^2 && \text{(B.117)}\\
&\leq \left(\mathscr{D}(\boldsymbol{H}_0, \boldsymbol{X}_0)^{1/2} + \delta\sqrt{r}\right)^2. && \text{(B.118)}
\end{aligned}$$

Hence,

$$\mathscr{D}(\boldsymbol{H}_0, \boldsymbol{X})^{1/2} \leq \mathscr{D}(\boldsymbol{H}_0, \boldsymbol{X}_0)^{1/2} + \delta\sqrt{r}. \quad \text{(B.119)}$$

Combining equations (B.113), (B.119), and (B.103), we get

$$\mathscr{D}(\widehat{\boldsymbol{H}}, \boldsymbol{X}_0)^{1/2} \leq \mathscr{D}(\boldsymbol{H}_0, \boldsymbol{X}_0)^{1/2} + 3\delta\sqrt{r}. \quad \text{(B.120)}$$

This completes the proof of lemma. □

Lemma B.6. *Let $\boldsymbol{X}_0$ be such that the uniqueness assumption holds with parameter $\alpha > 0$, and $\mathrm{conv}(\boldsymbol{X}_0)$ has internal radius at least $\mu > 0$. In particular, we have $\boldsymbol{z}_0 + \boldsymbol{U}\mathsf{B}_{r-1}(\mu) \subseteq$*

$\mathrm{conv}(\boldsymbol{X}_0)$ *for* $\boldsymbol{z}_0 \in \mathbb{R}^d$*, and an orthogonal matrix* $\boldsymbol{U} \in \mathbb{R}^{d\times(r-1)}$*. Finally assume that* $\max_{i\le n} \|\boldsymbol{Z}_{i,.}\|_2 \le \delta$*. Then for* $\widehat{\boldsymbol{H}}$ *the optimal solution of problem* (2.5)*, we have*

$$\alpha(\mathscr{D}(\widehat{\boldsymbol{H}}, \boldsymbol{H}_0)^{1/2} + \mathscr{D}(\boldsymbol{H}_0, \widehat{\boldsymbol{H}})^{1/2}) \le 2(1+2\alpha)\left[r^{3/2}\delta\kappa(\boldsymbol{P}_0(\widehat{\boldsymbol{H}})) + \frac{\delta\sqrt{r}}{\mu}\sigma_{\max}(\widehat{\boldsymbol{H}} - \mathbf{1}\boldsymbol{z}_0^{\mathsf{T}})\right] + 3\delta\sqrt{r} \tag{B.121}$$

where $\boldsymbol{P}_0 : \mathbb{R}^d \to \mathbb{R}^d$ *is the orthogonal projector onto* $\mathrm{aff}(\boldsymbol{H}_0)$ *(in particular,* $\boldsymbol{P}_0$ *is an affine map).*

Proof. Let $\widetilde{\boldsymbol{H}}$ be such that $\mathrm{conv}(\boldsymbol{X}_0) \subseteq \mathrm{conv}(\widetilde{\boldsymbol{H}})$. The uniqueness assumption implies

$$\mathscr{D}(\widetilde{\boldsymbol{H}}, \boldsymbol{X}_0)^{1/2} \ge \mathscr{D}(\boldsymbol{H}_0, \boldsymbol{X}_0)^{1/2} + \alpha\big(\mathscr{D}(\widetilde{\boldsymbol{H}}, \boldsymbol{H}_0)^{1/2} + \mathscr{D}(\boldsymbol{H}_0, \widetilde{\boldsymbol{H}})^{1/2}\big). \tag{B.122}$$

Note that Lemma B.5 implies

$$\mathscr{D}(\widehat{\boldsymbol{H}}, \boldsymbol{X}_0)^{1/2} \le \mathscr{D}(\boldsymbol{H}_0, \boldsymbol{X}_0)^{1/2} + 3\delta\sqrt{r}. \tag{B.123}$$

Therefore,

$$\mathscr{D}(\widetilde{\boldsymbol{H}}, \boldsymbol{X}_0)^{1/2} \ge \mathscr{D}(\widehat{\boldsymbol{H}}, \boldsymbol{X}_0)^{1/2} - 3\delta\sqrt{r} + \alpha\big(\mathscr{D}(\widetilde{\boldsymbol{H}}, \boldsymbol{H}_0)^{1/2} + \mathscr{D}(\boldsymbol{H}_0, \widetilde{\boldsymbol{H}})^{1/2}\big). \tag{B.124}$$

Hence,

$$\alpha\big(\mathscr{D}(\widetilde{\boldsymbol{H}}, \boldsymbol{H}_0)^{1/2} + \mathscr{D}(\boldsymbol{H}_0, \widetilde{\boldsymbol{H}})^{1/2}\big) \le \mathscr{D}(\widetilde{\boldsymbol{H}}, \boldsymbol{X}_0)^{1/2} - \mathscr{D}(\widehat{\boldsymbol{H}}, \boldsymbol{X}_0)^{1/2} + 3\delta\sqrt{r}. \tag{B.125}$$

In addition, for a convex set S, by triangle inequality we have

$$\left[\sum_{i=1}^n D(\widehat{\boldsymbol{H}}_{i,.}, S)\right]^{1/2} - \left[\sum_{i=1}^n D(\widetilde{\boldsymbol{H}}_{i,.}, S)\right]^{1/2} \le \left[\sum_{i=1}^n \left(D(\widehat{\boldsymbol{H}}_{i,.}, S)^{1/2} - D(\widetilde{\boldsymbol{H}}_{i,.}, S)^{1/2}\right)^2\right]^{1/2}. \tag{B.126}$$

In addition, note that for $\boldsymbol{x} \in \mathbb{R}^d$, $S \subseteq \mathbb{R}^d$, $D(\boldsymbol{x}, S)^{1/2} = \inf_{\boldsymbol{z} \in S} \|\boldsymbol{x} - \boldsymbol{z}\|_2$. Hence, Using triangle inequality, for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$, $\boldsymbol{z} \in S$,

$$\|\boldsymbol{x} - \boldsymbol{z}\|_2 \leq \|\boldsymbol{y} - \boldsymbol{z}\|_2 + \|\boldsymbol{x} - \boldsymbol{y}\|_2. \tag{B.127}$$

Taking infimum over $\boldsymbol{z} \in S$ above and rearranging terms, we get

$$D(\boldsymbol{x}, S)^{1/2} - D(\boldsymbol{y}, S)^{1/2} \leq \|\boldsymbol{x} - \boldsymbol{y}\|_2. \tag{B.128}$$

Similarly, by replacing the roles of $\boldsymbol{x}, \boldsymbol{y}$ above we conclude that for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$, $S \subseteq \mathbb{R}^d$

$$\left|D(\boldsymbol{x}, S)^{1/2} - D(\boldsymbol{y}, S)^{1/2}\right| \leq \|\boldsymbol{x} - \boldsymbol{y}\|_2. \tag{B.129}$$

In particular, we have

$$|D(\widetilde{\boldsymbol{H}}_{i,\cdot}, S)^{1/2} - D(\widehat{\boldsymbol{H}}_{i,\cdot}, S)^{1/2}| \leq \|\widetilde{\boldsymbol{H}}_{i,\cdot} - \widehat{\boldsymbol{H}}_{i,\cdot}\|_2. \tag{B.130}$$

Therefore, using (B.126) we have

$$\left[\sum_{i=1}^{n} D(\widehat{\boldsymbol{H}}_{i,\cdot}, S)\right]^{1/2} - \left[\sum_{i=1}^{n} D(\widetilde{\boldsymbol{H}}_{i,\cdot}, S)\right]^{1/2} \leq \left[\sum_{i=1}^{n} \|\widehat{\boldsymbol{H}}_{i,\cdot} - \widetilde{\boldsymbol{H}}_{i,\cdot}\|_2^2\right]^{1/2} = \|\widehat{\boldsymbol{H}} - \widetilde{\boldsymbol{H}}\|_F. \tag{B.131}$$

Hence,

$$|\mathscr{D}(\widetilde{\boldsymbol{H}}, \boldsymbol{X}_0)^{1/2} - \mathscr{D}(\widehat{\boldsymbol{H}}, \boldsymbol{X}_0)^{1/2}| \leq \|\widetilde{\boldsymbol{H}} - \widehat{\boldsymbol{H}}\|_F \tag{B.132}$$

and

$$|\mathscr{D}(\widetilde{\boldsymbol{H}}, \boldsymbol{H}_0)^{1/2} - \mathscr{D}(\widehat{\boldsymbol{H}}, \boldsymbol{H}_0)^{1/2}| \leq \|\widetilde{\boldsymbol{H}} - \widehat{\boldsymbol{H}}\|_F. \tag{B.133}$$

In addition, similarly to the proof of Lemma B.5, we can write

$$\mathscr{D}(\boldsymbol{H}_0, \widetilde{\boldsymbol{H}}) = \sum_{i=1}^{r} \min_{\boldsymbol{\alpha}_i \in \Delta^r} \|(\boldsymbol{H}_0)_{i,\cdot} - \widetilde{\boldsymbol{H}}^{\mathsf{T}} \boldsymbol{\alpha}_i\|_2^2 \tag{B.134}$$

$$= \sum_{i=1}^{r} \min_{\boldsymbol{\alpha}_i \in \Delta^r} \|(\boldsymbol{H}_0)_{i,\cdot} - \widehat{\boldsymbol{H}}^{\mathsf{T}} \boldsymbol{\alpha}_i - (\widehat{\boldsymbol{H}} - \widetilde{\boldsymbol{H}})^{\mathsf{T}} \boldsymbol{\alpha}_i\|_2^2 \tag{B.135}$$

$$\leq \sum_{i=1}^{r} \min_{\boldsymbol{\alpha}_i \in \Delta^r} \left\{ \|(\boldsymbol{H}_0)_{i,\cdot} - \widehat{\boldsymbol{H}}^{\mathsf{T}} \boldsymbol{\alpha}_i\|_2 + \|(\widehat{\boldsymbol{H}} - \widetilde{\boldsymbol{H}})^{\mathsf{T}} \boldsymbol{\alpha}_i\|_2 \right\}^2 \tag{B.136}$$

$$\leq \sum_{i=1}^{r} \left\{ \min_{\boldsymbol{\alpha} \in \Delta^r} \|(\boldsymbol{H}_0)_{i,\cdot} - \widehat{\boldsymbol{H}}^{\mathsf{T}} \boldsymbol{\alpha}\|_2 + \max_{\boldsymbol{\alpha} \in \Delta^r} \|(\widehat{\boldsymbol{H}} - \widetilde{\boldsymbol{H}})^{\mathsf{T}} \boldsymbol{\alpha}\|_2 \right\}^2 \tag{B.137}$$

$$\leq \left\{ \left(\sum_{i=1}^{r} \min_{\boldsymbol{\alpha} \in \Delta^r} \|(\boldsymbol{H}_0)_{i,\cdot} - \widehat{\boldsymbol{H}}^{\mathsf{T}} \boldsymbol{\alpha}_i\|_2^2 \right)^{1/2} + \sqrt{r} \max_{i \in [r]} \|\widehat{\boldsymbol{H}}_{i,\cdot} - \widetilde{\boldsymbol{H}}_{i,\cdot}\|_2 \right\}^2 \tag{B.138}$$

$$\leq \left(\mathscr{D}(\boldsymbol{H}_0, \widehat{\boldsymbol{H}})^{1/2} + \sqrt{r} \max_{i \in [r]} \|\widehat{\boldsymbol{H}}_{i,\cdot} - \widetilde{\boldsymbol{H}}_{i,\cdot}\|_2 \right)^2 . \tag{B.139}$$

Thus,

$$|\mathscr{D}(\boldsymbol{H}_0, \widetilde{\boldsymbol{H}})^{1/2} - \mathscr{D}(\boldsymbol{H}_0, \widehat{\boldsymbol{H}})^{1/2}| \leq \sqrt{r} \max_{i \in [r]} \|\widetilde{\boldsymbol{H}}_{i,\cdot} - \widehat{\boldsymbol{H}}_{i,\cdot}\|_2 \tag{B.140}$$

Therefore, combining (B.125), (B.132), (B.133), (B.140), we get

$$\alpha\big(\mathscr{D}(\widehat{\boldsymbol{H}}, \boldsymbol{H}_0)^{1/2} + \mathscr{D}(\boldsymbol{H}_0, \widehat{\boldsymbol{H}})^{1/2}\big) \leq (1+\alpha)\|\widetilde{\boldsymbol{H}} - \widehat{\boldsymbol{H}}\|_F + \alpha\sqrt{r} \max_{i \in [r]} \|\widetilde{\boldsymbol{H}}_{i,\cdot} - \widehat{\boldsymbol{H}}_{i,\cdot}\|_2 + 3\delta\sqrt{r}. \tag{B.141}$$

Now, we would like to bound the terms $\|\widetilde{\boldsymbol{H}} - \widehat{\boldsymbol{H}}\|_F$, $\max_{i \in [r]} \|\widetilde{\boldsymbol{H}}_{i,\cdot} - \widehat{\boldsymbol{H}}_{i,\cdot}\|_2$. Note that using the fact that $\widehat{\boldsymbol{H}}$ is feasible for Problem (2.5), we have

$$\mathscr{D}(\boldsymbol{X}_{i,\cdot}, \widehat{\boldsymbol{H}}) \leq \delta^2 . \tag{B.142}$$

Thus by triangle inequality,

$$\mathscr{D}((\boldsymbol{X}_0)_{i,\cdot}, \widehat{\boldsymbol{H}})^{1/2} \leq \|\boldsymbol{X}_{i,\cdot} - (\boldsymbol{X}_0)_{i,\cdot}\|_2 + \mathscr{D}(\boldsymbol{X}_{i,\cdot}, \widehat{\boldsymbol{H}})^{1/2} \leq 2\delta. \tag{B.143}$$

In addition, we know that $(\boldsymbol{X}_0)_{i,\cdot} \in \mathrm{aff}(\boldsymbol{H}_0)$, where $\mathrm{aff}(\boldsymbol{H}_0)$ is a $r-1$ dimensional affine subspace. Therefore, $\mathrm{conv}(\boldsymbol{X}_0) \subseteq \mathrm{aff}(\boldsymbol{H}_0)$ and, by convexity of $\mathsf{B}_d(2\delta, \widehat{\boldsymbol{H}})$, we get

$$\mathrm{conv}(\boldsymbol{X}_0) \subseteq \mathsf{B}_d(2\delta, \widehat{\boldsymbol{H}}) \cap \mathrm{aff}(\boldsymbol{H}_0). \tag{B.144}$$

First consider the case in which $\widehat{\boldsymbol{H}}_{i,\cdot} \in \mathrm{aff}(\boldsymbol{H}_0)$. for all $i \in \{1, 2, \ldots, r\}$. By a perturbation argument, we can assume that the rows of $\widehat{\boldsymbol{H}}$ are linearly independent, and hence $\mathrm{aff}(\widehat{\boldsymbol{H}}) = \mathrm{aff}(\boldsymbol{H}_0)$. Consider $\tilde{\boldsymbol{Q}} \in \mathbb{R}^{r \times d}$ defined by

$$\widetilde{Q}_{ii} = 1 + \xi, \qquad \text{if } \ i = j \in \{1, 2, \ldots, r\}, \tag{B.145}$$

$$\widetilde{Q}_{ij} = -\frac{\xi}{r-1}, \quad \text{if } \ i \neq j \in \{1, 2, \ldots, r\}, \tag{B.146}$$

$$\widetilde{Q}_{ij} = 0, \qquad \text{if } \ j \in \{r+1, r+2, \ldots, d\} \tag{B.147}$$

where $\xi = 2r\delta_0$. Note that for every $\boldsymbol{y} \in \mathsf{B}_d(2\delta_0; \boldsymbol{E}_{r,d}) \cap \mathrm{aff}(\boldsymbol{E}_{r,d})$, we have $\mathscr{D}(\boldsymbol{y}, \boldsymbol{E}_{r,d})^{1/2} \leq 2\delta_0$. In addition, since $\boldsymbol{y} \in \mathrm{aff}(\boldsymbol{E}_{r,d})$, $\langle \boldsymbol{y}, \mathbf{1} \rangle = 1$. Hence, for $\boldsymbol{y} \in \mathsf{B}_d(2\delta_0; \boldsymbol{E}_{r,d}) \cap \mathrm{aff}(\boldsymbol{E}_{r,d})$, we can write

$$\boldsymbol{y} = \boldsymbol{\pi} + \boldsymbol{x} \tag{B.148}$$

where $\boldsymbol{\pi} \in \mathrm{conv}(\boldsymbol{E}_{r,d})$, $\boldsymbol{x} \in \mathbb{R}^d$, $\langle \mathbf{1}, \boldsymbol{x} \rangle = 0$, $\|\boldsymbol{x}\|_2 \leq 2\delta_0$. It is easy to check that for this $\boldsymbol{y}$ we have

$$\boldsymbol{y} = \sum_{i=1}^{r} \beta_i \widetilde{\boldsymbol{Q}}_{i,\cdot} \tag{B.149}$$

where $\boldsymbol{\beta} \in \mathbb{R}^r$ is such that for $i = 1, 2, \ldots, r$,

$$\beta_i = \frac{r-1}{r-1+\xi r}(\pi_i + x_i) + \frac{\xi}{r-1+\xi r}. \tag{B.150}$$

Further, note that since $\boldsymbol{\pi} \in \mathrm{conv}(\boldsymbol{E}_{r,d})$, $\pi_i \geq 0$ and $x_i \geq -\|\boldsymbol{x}\|_2 \geq -2\delta_0$, we have $\pi_i + x_i \geq -2\delta_0$. Hence, for $i \in \{1, 2, \ldots, r\}$,

$$\beta_i \geq \frac{-2\delta_0(r-1) + \xi}{r-1+\xi r} = \frac{2\delta_0}{r-1+\xi r} \geq 0. \tag{B.151}$$

In addition,

$$\sum_{i=1}^{r}\beta_i = \frac{r\xi}{r-1+\xi r} + \frac{r-1}{r-1+\xi r}\left(\sum_{i=1}^{r}(\pi_i + x_i)\right) = 1. \tag{B.152}$$

Therefore, every $\boldsymbol{y} \in \mathsf{B}_d(2\delta_0; \boldsymbol{E}_{r,d}) \cap \mathrm{aff}(\boldsymbol{E}_{r,d})$ can be written as a convex combination of the rows of $\widetilde{\boldsymbol{Q}}$. Hence,

$$\mathsf{B}_d(2\delta_0; \boldsymbol{E}_{r,d}) \cap \mathrm{aff}(\boldsymbol{E}_{r,d}) \subseteq \mathrm{conv}(\widetilde{\boldsymbol{Q}}). \tag{B.153}$$

Let $\widehat{\boldsymbol{H}} = \boldsymbol{E}_{r,d}\boldsymbol{M}$, $\boldsymbol{M} \in \mathbb{R}^{d\times d}$. Since $\mathrm{aff}(\widehat{\boldsymbol{H}}) = \mathrm{aff}(\boldsymbol{H}_0)$, by taking $\widetilde{\boldsymbol{H}} = \widetilde{\boldsymbol{Q}}\boldsymbol{M}$, we have

$$\begin{aligned}
\mathrm{conv}(\widetilde{\boldsymbol{H}}) &\supseteq \left[\cup_{\boldsymbol{x}\in\mathrm{conv}(\boldsymbol{E}_{r,d})}\boldsymbol{M}^{\mathsf{T}}\mathsf{B}_d(2\delta_0;\boldsymbol{x})\right] \cap \mathrm{aff}(\widehat{\boldsymbol{H}}) && \text{(B.154)}\\
&\supseteq \left[\cup_{\boldsymbol{x}\in\mathrm{conv}(\widehat{\boldsymbol{H}})}\mathsf{B}_d(2\delta_0\sigma_{\min}(\boldsymbol{M});\boldsymbol{x})\right] \cap \mathrm{aff}(\widehat{\boldsymbol{H}}) && \text{(B.155)}\\
&\supseteq \mathsf{B}_d(2\delta;\widehat{\boldsymbol{H}}) \cap \mathrm{aff}(\boldsymbol{H}_0), && \text{(B.156)}
\end{aligned}$$

provided that $\delta_0 = \delta/\sigma_{\min}(\boldsymbol{M}) = \delta/\sigma_{\min}(\widehat{\boldsymbol{H}})$. Hence, using (B.144) for this δ_0, $\mathrm{conv}(\boldsymbol{X}_0) \subseteq \mathrm{conv}(\widetilde{\boldsymbol{H}})$. Note that for $\widetilde{\boldsymbol{Q}}$, we have $\|\widetilde{\boldsymbol{Q}}_{i,\cdot} - \boldsymbol{e}_i\|_2 \le 2r\delta_0$. Thus,

$$\|\widetilde{\boldsymbol{Q}} - \boldsymbol{E}_{r,d}\|_F \le 2r^{3/2}\delta_0. \tag{B.157}$$

Therefore, there exists $\widetilde{\boldsymbol{H}} \in \mathbb{R}^{r\times d}$ such that $\mathrm{conv}(\boldsymbol{X}_0) \subseteq \mathrm{conv}(\widetilde{\boldsymbol{H}})$ and

$$\begin{aligned}
&\|\widetilde{\boldsymbol{H}} - \widehat{\boldsymbol{H}}\|_F = \|(\widetilde{\boldsymbol{Q}} - \boldsymbol{E}_{r,d})\boldsymbol{M}\|_F \le 2r^{3/2}\delta_0\sigma_{\max}(\boldsymbol{M}) = 2r^{3/2}\delta_0\sigma_{\max}(\widehat{\boldsymbol{H}}) = 2r^{3/2}\delta\kappa(\widehat{\boldsymbol{H}}),\\
&\max_{i\in[r]}\|\widetilde{\boldsymbol{H}}_{i,\cdot} - \widehat{\boldsymbol{H}}_{i,\cdot}\|_2 = \max_{i\in[r]}\|(\widetilde{\boldsymbol{Q}}_{i,\cdot} - \boldsymbol{e}_i)\boldsymbol{M}\|_2 \le 2r\delta_0\sigma_{\max}(\boldsymbol{M}) = 2r\delta_0\sigma_{\max}(\widehat{\boldsymbol{H}}) = 2r\delta\kappa(\widehat{\boldsymbol{H}}).
\end{aligned}$$

Now consider the general case in which $\mathrm{aff}(\widehat{\boldsymbol{H}}) \neq \mathrm{aff}(\boldsymbol{H}_0)$. Let $\boldsymbol{H}' \in \mathbb{R}^{r\times d}$ be such that $\boldsymbol{H}'_{i,\cdot}$ is the projection of $\widehat{\boldsymbol{H}}_{i,\cdot}$ onto $\mathrm{aff}(\boldsymbol{H}_0)$. Assuming that the rows of $\boldsymbol{H}'$ are linearly independent, $\mathrm{aff}(\boldsymbol{H}') = \mathrm{aff}(\boldsymbol{H}_0)$. Note that since $\mathrm{conv}(\boldsymbol{X}_0) \in \mathrm{aff}(\boldsymbol{H}_0)$, for every point $\boldsymbol{x} \in \mathrm{conv}(\boldsymbol{X}_0)$, $\mathscr{D}(\boldsymbol{x}, \boldsymbol{H}')^{1/2} \le \mathscr{D}(\boldsymbol{x}, \widehat{\boldsymbol{H}})^{1/2} \le 2\delta$. Thus,

$$(\boldsymbol{X}_0)_{i,\cdot} \in \mathsf{B}_d(2\delta, \boldsymbol{H}') \cap \mathrm{aff}(\boldsymbol{H}'). \tag{B.158}$$

Therefore, using the above argument for the case where $\mathrm{aff}(\widehat{\boldsymbol{H}}) = \mathrm{aff}(\boldsymbol{H}_0)$, we can find $\widetilde{\boldsymbol{H}}$ such that $\mathrm{conv}(\boldsymbol{X}_0) \subseteq \mathrm{conv}(\widetilde{\boldsymbol{H}})$ and

$$\|\widetilde{\boldsymbol{H}} - \boldsymbol{H}'\|_F \leq 2r^{3/2}\delta\kappa(\boldsymbol{H}'), \tag{B.159}$$

$$\max_{i\in[r]} \|\widetilde{\boldsymbol{H}}_{i,\cdot} - \boldsymbol{H}'_{i,\cdot}\|_2 \leq 2r\delta\kappa(\boldsymbol{H}'). \tag{B.160}$$

Hence, for every $i = 1, 2, \ldots, r$,

$$\begin{aligned} \|\widetilde{\boldsymbol{H}}_{i,\cdot} - \widehat{\boldsymbol{H}}_{i,\cdot}\|_2 &\leq \|\widetilde{\boldsymbol{H}}_{i,\cdot} - \boldsymbol{H}'_{i,\cdot}\|_2 + \|\boldsymbol{H}'_{i,\cdot} - \widehat{\boldsymbol{H}}_{i,\cdot}\|_2 \\ &\leq 2r\delta\kappa(\boldsymbol{H}') + \|\boldsymbol{P}_0(\widehat{\boldsymbol{H}}_{i,\cdot}) - \widehat{\boldsymbol{H}}_{i,\cdot}\|_2 \end{aligned} \tag{B.161}$$

where $\boldsymbol{P}_0$ is the orthogonal projection onto $\mathrm{aff}(\boldsymbol{H}_0)$. We next use the assumption on the internal radius of $\mathrm{conv}(\boldsymbol{X}_0)$ to upper bound the term $\|\boldsymbol{P}_0(\widehat{\boldsymbol{H}}_{i,\cdot}) - \widehat{\boldsymbol{H}}_{i,\cdot}\|_2$. Note that since $\mathrm{conv}(\boldsymbol{X}_0) \subseteq \mathsf{B}_d(2\delta, \widehat{\boldsymbol{H}})$, letting $\bar{\boldsymbol{H}} = \widehat{\boldsymbol{H}} - \mathbf{1}\boldsymbol{z}_0^\mathsf{T}$, for some orthogonal matrix $\boldsymbol{U} \in \mathbb{R}^{d\times r'}$, $r' = r - 1$, we have

$$\max_{\|\boldsymbol{z}\|_2\leq\mu} \min_{\langle\boldsymbol{a},\mathbf{1}\rangle=1,\boldsymbol{a}\geq 0} \|\boldsymbol{U}\boldsymbol{z} - \bar{\boldsymbol{H}}^\mathsf{T}\boldsymbol{a}\|_2^2 = \max_{\|\boldsymbol{z}\|_2\leq\mu} \min_{\langle\boldsymbol{a},\mathbf{1}\rangle=1,\boldsymbol{a}\geq 0} \|\boldsymbol{U}\boldsymbol{z} - (\widehat{\boldsymbol{H}} - \mathbf{1}\boldsymbol{z}_0^\mathsf{T})^\mathsf{T}\boldsymbol{a}\|_2^2 \tag{B.162}$$

$$\leq \max_{\|\boldsymbol{z}\|_2\leq\mu} \min_{\langle\boldsymbol{a},\mathbf{1}\rangle=1,\boldsymbol{a}\geq 0} \|\boldsymbol{U}\boldsymbol{z} + \boldsymbol{z}_0 - \widehat{\boldsymbol{H}}^\mathsf{T}\boldsymbol{a}\|_2^2 \leq 4\delta^2. \tag{B.163}$$

Now, using Cauchy-Schwarz inequality we can write

$$\max_{\|\boldsymbol{z}\|_2\leq\mu} \min_{\|\boldsymbol{a}\|_2\leq 1} \|\boldsymbol{U}\boldsymbol{z} - \bar{\boldsymbol{H}}^\mathsf{T}\boldsymbol{a}\|_2^2 \leq \max_{\|\boldsymbol{z}\|_2\leq\mu} \min_{\langle\boldsymbol{a},\mathbf{1}\rangle=1,\boldsymbol{a}\geq 0} \|\boldsymbol{U}\boldsymbol{z} - \bar{\boldsymbol{H}}^\mathsf{T}\boldsymbol{a}\|_2^2 \leq 4\delta^2. \tag{B.164}$$

Note that,

$$\min_{\|\boldsymbol{a}\|_2\leq 1} \|\boldsymbol{U}\boldsymbol{z} - \bar{\boldsymbol{H}}^\mathsf{T}\boldsymbol{a}\|_2^2 = \max_{\rho\geq 0} \min_{\boldsymbol{a}} \left\{ \|\boldsymbol{z}\|_2^2 - 2\left\langle \boldsymbol{z}, \boldsymbol{U}^\mathsf{T}\bar{\boldsymbol{H}}^\mathsf{T}\boldsymbol{a}\right\rangle + \left\langle \boldsymbol{a}, (\bar{\boldsymbol{H}}\bar{\boldsymbol{H}}^\mathsf{T} + \rho\mathbf{I})\boldsymbol{a}\right\rangle - \rho \right\} \tag{B.165}$$

$$= \max_{\rho\geq 0} \left\{ \|\boldsymbol{z}\|_2^2 - \left\langle \bar{\boldsymbol{H}}\boldsymbol{U}\boldsymbol{z}, (\bar{\boldsymbol{H}}\bar{\boldsymbol{H}}^\mathsf{T} + \rho\mathbf{I})^{-1}\bar{\boldsymbol{H}}\boldsymbol{U}\boldsymbol{z}\right\rangle - \rho \right\} \tag{B.166}$$

Hence, using (B.164)

$$\mu^2 \max_{\rho \geq 0} \left\{ \lambda_{\max}(\mathbf{I} - \boldsymbol{U}^\mathsf{T} \bar{\boldsymbol{H}}^\mathsf{T} (\bar{\boldsymbol{H}} \bar{\boldsymbol{H}}^\mathsf{T} + \rho \mathbf{I})^{-1} \bar{\boldsymbol{H}} \boldsymbol{U}) - \rho \right\} \leq 4\delta^2. \tag{B.167}$$

In particular, for $\rho = 0$ we get

$$\mu^2 \lambda_{\max}(\mathbf{I} - \boldsymbol{U}^\mathsf{T} \bar{\boldsymbol{H}}^\mathsf{T} (\bar{\boldsymbol{H}} \bar{\boldsymbol{H}}^\mathsf{T})^{-1} \bar{\boldsymbol{H}} \boldsymbol{U}) \leq 4\delta^2. \tag{B.168}$$

Taking $\bar{\boldsymbol{H}} = \tilde{\boldsymbol{U}} \boldsymbol{\Sigma} \tilde{\boldsymbol{V}}^\mathsf{T}$, the singular value decomposition of $\bar{\boldsymbol{H}}$, we have $\sigma_{\max}(\bar{\boldsymbol{H}}) = \sigma_{\max}(\widehat{\boldsymbol{H}} - \mathbf{1}\boldsymbol{z}_0^\mathsf{T}) = \max_i \Sigma_{ii}$. Letting $\boldsymbol{U}^\mathsf{T} \tilde{\boldsymbol{V}} = \boldsymbol{Q}$, we get

$$\max_{\rho \geq 0} \lambda_{\max} \left(\mathbf{I} - \boldsymbol{Q} \boldsymbol{Q}^\mathsf{T} \right) \leq \frac{4\delta^2}{\mu^2}. \tag{B.169}$$

Letting $q = \sigma_{\min}(\boldsymbol{Q})$, this results in

$$1 - q^2 \leq \frac{4\delta^2}{\mu^2}. \tag{B.170}$$

In addition, note that, by the internal radius assumption, for any $\boldsymbol{z} \in \mathbb{R}^{r'}$, $\boldsymbol{z}_0 + \boldsymbol{U}\boldsymbol{z} \in \mathrm{aff}(\boldsymbol{H}_0)$. Further, since $\boldsymbol{z}_0 \in \mathrm{aff}(\boldsymbol{H}_0)$,

$$\max_{i \in [r]} \| \boldsymbol{P}_0(\widehat{\boldsymbol{H}}_{i,.}) - \widehat{\boldsymbol{H}}_{i,.} \|_2 = \max_{i \in [r]} \| \boldsymbol{P}_{\boldsymbol{U}}(\bar{\boldsymbol{H}}_{i,.}) - \bar{\boldsymbol{H}}_{i,.} \|_2 \tag{B.171}$$

$$\leq \max_{\|\boldsymbol{a}\|_2 \leq 1} \| \boldsymbol{P}_{\boldsymbol{U}}(\bar{\boldsymbol{H}}^\mathsf{T} \boldsymbol{a}) - \bar{\boldsymbol{H}}^\mathsf{T} \boldsymbol{a} \|_2 \tag{B.172}$$

$$\leq \max_{\|\boldsymbol{a}\|_2 \leq 1} \| \boldsymbol{P}_{\boldsymbol{U}}(\bar{\boldsymbol{H}}^\mathsf{T} \boldsymbol{a}) - \bar{\boldsymbol{H}}^\mathsf{T} \boldsymbol{a} \|_2 \tag{B.173}$$

$$\leq \max_{\|\boldsymbol{a}\|_2 \leq 1} \min_{\boldsymbol{z}} \| \boldsymbol{U}\boldsymbol{z} - \bar{\boldsymbol{H}}^\mathsf{T} \boldsymbol{a} \|_2^2 \tag{B.174}$$

where $\boldsymbol{P}_{\boldsymbol{U}}$ is the projector onto the column space of $\boldsymbol{U}$. Note that,

$$\max_{\|\boldsymbol{a}\|_2\leq 1}\min_{\boldsymbol{z}}\|\boldsymbol{U}\boldsymbol{z}-\bar{\boldsymbol{H}}^{\mathsf{T}}\boldsymbol{a}\|_2^2=\max_{\|\boldsymbol{a}\|_2\leq 1}\left\{-\left\langle\boldsymbol{a},\bar{\boldsymbol{H}}\boldsymbol{U}\boldsymbol{U}^{\mathsf{T}}\bar{\boldsymbol{H}}^{\mathsf{T}}\boldsymbol{a}\right\rangle+\left\langle\boldsymbol{a},\bar{\boldsymbol{H}}\bar{\boldsymbol{H}}^{\mathsf{T}}\boldsymbol{a}\right\rangle\right\} \tag{B.175}$$

$$=\lambda_{\max}(\bar{\boldsymbol{H}}\bar{\boldsymbol{H}}^{\mathsf{T}}-\bar{\boldsymbol{H}}\boldsymbol{U}\boldsymbol{U}^{\mathsf{T}}\bar{\boldsymbol{H}}^{\mathsf{T}}) \tag{B.176}$$

$$=\lambda_{\max}(\boldsymbol{\Sigma}(\mathbf{I}-\boldsymbol{Q}^{\mathsf{T}}\boldsymbol{Q})\boldsymbol{\Sigma}) \tag{B.177}$$

$$\leq\sigma_{\max}(\bar{\boldsymbol{H}})^2\lambda_{\max}(\mathbf{I}-\boldsymbol{Q}^{\mathsf{T}}\boldsymbol{Q}) \tag{B.178}$$

$$\leq\sigma_{\max}(\bar{\boldsymbol{H}})^2(1-q^2)\leq\frac{4\sigma_{\max}(\bar{\boldsymbol{H}})^2\delta^2}{\mu^2} \tag{B.179}$$

where the last inequality follows from (B.170). This results in

$$\max_{i\in[r]}\|\boldsymbol{P}_0(\widehat{\boldsymbol{H}}_{i,\cdot})-\widehat{\boldsymbol{H}}_{i,\cdot}\|_2\leq\frac{2\sigma_{\max}(\bar{\boldsymbol{H}})\delta}{\mu}=\frac{2\sigma_{\max}(\widehat{\boldsymbol{H}}-\mathbf{1}\boldsymbol{z}_0^{\mathsf{T}})\delta}{\mu}. \tag{B.180}$$

Therefore, $\|\boldsymbol{P}_0(\widehat{\boldsymbol{H}})-\widehat{\boldsymbol{H}}\|_F\leq 2\sigma_{\max}(\widehat{\boldsymbol{H}}-\mathbf{1}\boldsymbol{z}_0^{\mathsf{T}})\delta\sqrt{r}/\mu$. Hence, using (B.161) we get

$$\max_{i\in[r]}\|\widehat{\boldsymbol{H}}_{i,\cdot}-\widetilde{\boldsymbol{H}}_{i,\cdot}\|_2\leq 2r\delta\kappa(\boldsymbol{P}_0(\widehat{\boldsymbol{H}}))+\frac{2\sigma_{\max}(\widehat{\boldsymbol{H}}-\mathbf{1}\boldsymbol{z}_0^{\mathsf{T}})\delta}{\mu}, \tag{B.181}$$

$$\|\widehat{\boldsymbol{H}}-\widetilde{\boldsymbol{H}}\|_F\leq 2r^{3/2}\delta\kappa(\boldsymbol{P}_0(\widehat{\boldsymbol{H}}))+\frac{2\sigma_{\max}(\widehat{\boldsymbol{H}}-\mathbf{1}\boldsymbol{z}_0^{\mathsf{T}})\delta\sqrt{r}}{\mu}. \tag{B.182}$$

Replacing this in (B.141) completes the proof. □

B.4.2 Proof of Theorem 1

For simplicity, let $\mathscr{D}=\alpha(\mathscr{D}(\widehat{\boldsymbol{H}},\boldsymbol{H}_0)^{1/2}+\mathscr{D}(\boldsymbol{H}_0,\widehat{\boldsymbol{H}})^{1/2})$. First note that under the assumption of Theorem 1 we have

$$\boldsymbol{z}_0+\boldsymbol{U}\mathsf{B}_{r'}(\mu)\subseteq\operatorname{conv}(\boldsymbol{X}_0)\subseteq\operatorname{conv}(\boldsymbol{H}_0). \tag{B.183}$$

Therefore, using Lemma B.4 with $\boldsymbol{H}=\boldsymbol{H}_0$ and $\delta=0$, we have

$$\mu\sqrt{2}\leq\sigma_{\min}(\boldsymbol{H}_0)\leq\sigma_{\max}(\boldsymbol{H}_0). \tag{B.184}$$

In addition, since $\boldsymbol{z}_0 \in \mathrm{conv}(\boldsymbol{H}_0)$ we have $\boldsymbol{z}_0 = \boldsymbol{H}_0^\mathsf{T}\boldsymbol{\alpha}_0$ for some $\boldsymbol{\alpha}_0 \in \Delta^r$. Therefore,

$$\|\boldsymbol{z}_0\|_2 \leq \sigma_{\max}(\boldsymbol{H}_0)\|\boldsymbol{\alpha}_0\|_2 \leq \sigma_{\max}(\boldsymbol{H}_0). \tag{B.185}$$

Note that

$$\sigma_{\max}(\widehat{\boldsymbol{H}} - \mathbf{1}\boldsymbol{z}_0^\mathsf{T}) \leq \sigma_{\max}(\widehat{\boldsymbol{H}}) + \sigma_{\max}(\mathbf{1}\boldsymbol{z}_0^\mathsf{T}) = \sigma_{\max}(\widehat{\boldsymbol{H}}) + \sqrt{r}\|\boldsymbol{z}_0\|_2. \tag{B.186}$$

Therefore, using Lemma B.6 we have

$$\mathscr{D} \leq 2(1+2\alpha)\left(r^{3/2}\delta\kappa(\boldsymbol{P}_0(\widehat{\boldsymbol{H}})) + \frac{\sigma_{\max}(\widehat{\boldsymbol{H}})\delta r^{1/2}}{\mu} + \frac{r\delta\|\boldsymbol{z}_0\|_2}{\mu}\right) + 3\delta r^{1/2}. \tag{B.187}$$

In addition, Lemma B.2 implies that

$$\mathscr{L}(\boldsymbol{H}_0, \widehat{\boldsymbol{H}})^{1/2} \leq \frac{\kappa(\boldsymbol{H}_0)}{\alpha}(1+\sqrt{2})\sqrt{r}\mathscr{D}\,. \tag{B.188}$$

Further, let $\boldsymbol{P}_0$ denote the orthogonal projector on $\mathrm{aff}(\boldsymbol{H}_0)$. Hence, $\boldsymbol{P}_0$ is a non-expansive mapping: for $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$, $D(\boldsymbol{P}_0(\boldsymbol{x}), \boldsymbol{P}_0(\boldsymbol{y})) \leq D(\boldsymbol{x}, \boldsymbol{y})$. Therefore, since $\mathrm{conv}(\boldsymbol{H}_0) \subset \mathrm{aff}(\boldsymbol{H}_0)$, for any $\boldsymbol{h} \in \mathbb{R}^d$

$$\mathscr{D}(\boldsymbol{P}_0(\boldsymbol{h}), \boldsymbol{H}_0) \leq D(\boldsymbol{P}_0(\boldsymbol{h}), \boldsymbol{P}_0(\boldsymbol{\Pi}_{\mathrm{conv}(\boldsymbol{H}_0)}(\boldsymbol{h}))) \leq D(\boldsymbol{h}, \boldsymbol{\Pi}_{\mathrm{conv}(\boldsymbol{H}_0)}(\boldsymbol{h})) = \mathscr{D}(\boldsymbol{h}, \boldsymbol{H}_0). \tag{B.189}$$

Therefore,

$$\mathscr{D}(\boldsymbol{P}_0(\widehat{\boldsymbol{H}}), \boldsymbol{H}_0) \leq \mathscr{D}(\widehat{\boldsymbol{H}}, \boldsymbol{H}_0). \tag{B.190}$$

First consider the case in which

$$\delta \leq \frac{\alpha\mu}{30\, r^{3/2}}. \tag{B.191}$$

Note that in this case $\delta \leq \mu/2$. Hence, using Lemma B.3 to upper bound $\sigma_{\max}(\widehat{\boldsymbol{H}})$, $\sigma_{\max}(\boldsymbol{P}_0(\widehat{\boldsymbol{H}}))$ and Lemma B.4 to lower bound $\sigma_{\min}(\boldsymbol{P}_0(\widehat{\boldsymbol{H}}))$, by (B.190), we get

$$\sigma_{\max}(\widehat{\boldsymbol{H}}) \leq \mathscr{D}(\widehat{\boldsymbol{H}}, \boldsymbol{H}_0)^{1/2} + r^{1/2}\sigma_{\max}(\boldsymbol{H}_0) \leq \frac{\mathscr{D}}{\alpha} + r^{1/2}\sigma_{\max}(\boldsymbol{H}_0), \tag{B.192}$$

$$\begin{aligned}\kappa(\boldsymbol{P}_0(\widehat{\boldsymbol{H}})) = \frac{\sigma_{\max}(\boldsymbol{P}_0(\widehat{\boldsymbol{H}}))}{\sigma_{\min}(\boldsymbol{P}_0(\widehat{\boldsymbol{H}}))} &\leq \frac{\mathscr{D}(\boldsymbol{P}_0(\widehat{\boldsymbol{H}}), \boldsymbol{H}_0)^{1/2} + r^{1/2}\sigma_{\max}(\boldsymbol{H}_0)}{\sqrt{2}(\mu - 2\delta)} \\ &\leq \frac{\mathscr{D}(\widehat{\boldsymbol{H}}, \boldsymbol{H}_0)^{1/2} + r^{1/2}\sigma_{\max}(\boldsymbol{H}_0)}{\sqrt{2}(\mu - 2\delta)} \leq \frac{\mathscr{D}}{\alpha(\mu - 2\delta)\sqrt{2}} + \frac{r^{1/2}\sigma_{\max}(\boldsymbol{H}_0)}{(\mu - 2\delta)\sqrt{2}}.\end{aligned} \tag{B.193}$$

Replacing these in (B.187) we have

$$\begin{aligned}\mathscr{D} \leq 2(1+2\alpha)\Bigg[&\frac{r^{3/2}\mathscr{D}\delta}{\alpha(\mu - 2\delta)\sqrt{2}} + \frac{r^2\sigma_{\max}(\boldsymbol{H}_0)\delta}{(\mu - 2\delta)\sqrt{2}} + \frac{\mathscr{D}r^{1/2}\delta}{\alpha\mu} \\ &+ \frac{r\sigma_{\max}(\boldsymbol{H}_0)\delta}{\mu} + \frac{r\|\boldsymbol{z}_0\|_2\delta}{\mu}\Bigg] + 3\delta\sqrt{r}.\end{aligned} \tag{B.194}$$

Therefore,

$$\begin{aligned}&\mathscr{D}\left[1 - \frac{\sqrt{2}(1+2\alpha)r^{3/2}\delta}{\alpha(\mu - 2\delta)} - \frac{2(1+2\alpha)r^{1/2}\delta}{\alpha\mu}\right] \\ &\quad\leq 2(1+2\alpha)\left[\frac{r^2\sigma_{\max}(\boldsymbol{H}_0)\delta}{(\mu - 2\delta)\sqrt{2}} + \frac{r\sigma_{\max}(\boldsymbol{H}_0)\delta}{\mu} + \frac{r\|\boldsymbol{z}_0\|_2\delta}{\mu}\right] + 3\delta\sqrt{r}\end{aligned} \tag{B.195}$$

Notice that condition (B.191) implies that $\mu - 2\delta \geq \mu/2$ and

$$\frac{\sqrt{2}(1+2\alpha)r^{3/2}\delta}{\alpha(\mu - 2\delta)} + \frac{2(1+2\alpha)r^{1/2}\delta}{\alpha\mu} \leq \frac{1}{2}. \tag{B.196}$$

Using the previous two equations, under condition (B.191) we have

$$\begin{aligned}\mathscr{D} &\leq \frac{4(1+2\alpha)r\delta}{\mu}\left[\frac{5r\sigma_{\max}(\boldsymbol{H}_0)}{2} + \|\boldsymbol{z}_0\|_2\right] + 3\delta\sqrt{r} \\ &\leq \frac{4(1+2\alpha)r^2}{\mu}\left[\frac{5\sigma_{\max}(\boldsymbol{H}_0)}{2} + \frac{\|\boldsymbol{z}_0\|_2}{r} + \frac{3\mu}{4(1+2\alpha)r^{3/2}}\right]\delta.\end{aligned} \tag{B.197}$$

Combining this with (B.188), and using the fact that $1+2\alpha \leq 3$, we have under condition (B.191)

$$\mathscr{L}(\boldsymbol{H}_0, \widehat{\boldsymbol{H}})^{1/2} \leq \frac{12(1+\sqrt{2})r^{5/2}\kappa(\boldsymbol{H}_0)}{\mu\alpha}\left(\frac{5\sigma_{\max}(\boldsymbol{H}_0)}{2} + \frac{\|\boldsymbol{z}_0\|_2}{r} + \frac{3\mu}{4(1+2\alpha)r^{3/2}}\right)\delta \tag{B.198}$$

$$\leq \frac{29\sigma_{\max}(\boldsymbol{H}_0)\kappa(\boldsymbol{H}_0)r^{5/2}}{\alpha\mu}\left(\frac{5}{2} + \frac{\|\boldsymbol{z}_0\|_2}{r\sigma_{\max}(\boldsymbol{H}_0)} + \frac{3\mu}{4(1+2\alpha)r^{3/2}\sigma_{\max}(\boldsymbol{H}_0)}\right)\delta. \tag{B.199}$$

Note that using (B.184), (B.185) and since $\alpha \geq 0$

$$\frac{\|\boldsymbol{z}_0\|_2}{r\sigma_{\max}(\boldsymbol{H}_0)} \leq 1, \qquad \frac{3\mu}{4(1+2\alpha)r^{3/2}\sigma_{\max}(\boldsymbol{H}_0)} \leq \frac{3}{4\sqrt{2}}. \tag{B.200}$$

Therefore,

$$\mathscr{L}(\boldsymbol{H}_0, \widehat{\boldsymbol{H}})^{1/2} \leq \frac{120\sigma_{\max}(\boldsymbol{H}_0)\kappa(\boldsymbol{H}_0)r^{5/2}}{\alpha\mu}\delta. \tag{B.201}$$

Thus,

$$\mathscr{L}(\boldsymbol{H}_0, \widehat{\boldsymbol{H}}) \leq \frac{C_*^2\, r^5}{\alpha^2}\max_{i\leq n}\|\boldsymbol{Z}_{i,\cdot}\|_2^2\,, \tag{B.202}$$

where C_* is defined in Theorem 1.

Next, consider the case in which

$$\delta = \max_{i\leq n}\|\boldsymbol{Z}_{i,\cdot}\|_2 \leq \frac{\alpha\mu}{330\kappa(\boldsymbol{H}_0)r^{5/2}}, \tag{B.203}$$

Note that using (B.184),(B.185) and since $1+2\alpha \leq 3$, this condition on δ implies that

$$\delta \leq \frac{\alpha\mu\sigma_{\min}(\boldsymbol{H}_0)}{12r(1+2\alpha)(5r^{3/2}\sigma_{\max}(\boldsymbol{H}_0) + 2\|\boldsymbol{z}_0\|_2 r^{1/2} + 3\mu)}. \tag{B.204}$$

In particular, condition (B.191) holds. Hence, using equation (B.197) we get

$$\mathscr{D} \leq \frac{4(1+2\alpha)r^2}{\mu}\left[\frac{5\sigma_{\max}(\boldsymbol{H}_0)}{2} + \frac{\|\boldsymbol{z}_0\|_2}{r} + \frac{3\mu}{4(1+2\alpha)r^{3/2}}\right]\delta \leq \frac{\alpha\sigma_{\min}(\boldsymbol{H}_0)}{6\sqrt{r}}. \tag{B.205}$$

Further, note that since $\boldsymbol{P}_0$ is a projection onto an affine subspace, for $\boldsymbol{x} \in \mathbb{R}^d$, $\boldsymbol{P}_0(\boldsymbol{x}) = \widetilde{\boldsymbol{P}_0}\boldsymbol{x} + \boldsymbol{x}_0$ for some $\widetilde{\boldsymbol{P}_0} \in \mathbb{R}^{d\times d}, \boldsymbol{x}_0 \in \mathbb{R}^d$. Hence, for any $\boldsymbol{\pi} \in \Delta^r$, $\boldsymbol{h} = \widehat{\boldsymbol{H}}^\mathsf{T}\boldsymbol{\pi} \in \mathrm{conv}(\widehat{\boldsymbol{H}})$, we have

$$\begin{aligned}\boldsymbol{P}_0(\boldsymbol{h}) = \widetilde{\boldsymbol{P}_0}\boldsymbol{h} + \boldsymbol{x}_0 = \widetilde{\boldsymbol{P}_0}\widehat{\boldsymbol{H}}^\mathsf{T}\boldsymbol{\pi} + \boldsymbol{x}_0 &= \sum_{i=1}^r \pi_i \left(\widetilde{\boldsymbol{P}_0}\widehat{\boldsymbol{H}}^\mathsf{T}\boldsymbol{e}_i + \boldsymbol{x}_0\right) \\ &= \sum_{i=1}^r \pi_i \boldsymbol{P}_0(\widehat{\boldsymbol{h}_i}) \in \mathrm{conv}(\boldsymbol{P}_0(\widehat{\boldsymbol{H}}))\end{aligned} \tag{B.206}$$

where $\boldsymbol{e}_i$ is the i'th standard unit vector. Hence,

$$\boldsymbol{P}_0(\mathrm{conv}(\widehat{\boldsymbol{H}})) \subseteq \mathrm{conv}(\boldsymbol{P}_0(\widehat{\boldsymbol{H}})). \tag{B.207}$$

Thus, for $\boldsymbol{h}_0 \in \mathbb{R}^d$ an arbitrary row of $\boldsymbol{H}_0$, we have

$$\begin{aligned}\mathscr{D}(\boldsymbol{h}_0, \boldsymbol{P}_0(\widehat{\boldsymbol{H}})) = D(\boldsymbol{h}_0, \mathrm{conv}(\boldsymbol{P}_0(\widehat{\boldsymbol{H}}))) &\le D(\boldsymbol{h}_0, \boldsymbol{P}_0(\mathrm{conv}(\widehat{\boldsymbol{H}}))) \\ &\le D(\boldsymbol{h}_0, \boldsymbol{P}_0(\boldsymbol{\Pi}_{\mathrm{conv}(\widehat{\boldsymbol{H}})}(\boldsymbol{h}_0))).\end{aligned} \tag{B.208}$$

In addition, using non-expansivity of $\boldsymbol{P}_0$, we have

$$D(\boldsymbol{h}_0, \boldsymbol{P}_0(\boldsymbol{\Pi}_{\mathrm{conv}(\widehat{\boldsymbol{H}})}(\boldsymbol{h}_0))) \le D(\boldsymbol{h}_0, \boldsymbol{\Pi}_{\mathrm{conv}(\widehat{\boldsymbol{H}})}(\boldsymbol{h}_0)) = D(\boldsymbol{h}_0, \mathrm{conv}(\widehat{\boldsymbol{H}})) = \mathscr{D}(\boldsymbol{h}_0, \widehat{\boldsymbol{H}}). \tag{B.209}$$

This implies that

$$\mathscr{D}(\boldsymbol{H}_0, \boldsymbol{P}_0(\widehat{\boldsymbol{H}})) \le \mathscr{D}(\boldsymbol{H}_0, \widehat{\boldsymbol{H}}). \tag{B.210}$$

Therefore, using (B.190), (B.210) and (B.205) we get

$$\begin{aligned}\mathscr{D}(\boldsymbol{H}_0, \boldsymbol{P}_0(\widehat{\boldsymbol{H}}))^{1/2} + \mathscr{D}(\boldsymbol{P}_0(\widehat{\boldsymbol{H}}), \boldsymbol{H}_0)^{1/2} &\le \mathscr{D}(\boldsymbol{H}_0, \widehat{\boldsymbol{H}})^{1/2} + \mathscr{D}(\widehat{\boldsymbol{H}}, \boldsymbol{H}_0)^{1/2} \\ &\le \frac{\mathscr{D}}{\alpha} \le \frac{\sigma_{\min}(\boldsymbol{H}_0)}{6\sqrt{r}}.\end{aligned} \tag{B.211}$$

Hence, in this case Lemma B.3 implies that

$$\sigma_{\max}(\widehat{\boldsymbol{H}}) \leq 2\sigma_{\max}(\boldsymbol{H}_0), \tag{B.212}$$

$$\kappa(\boldsymbol{P}_0(\widehat{\boldsymbol{H}})) \leq \frac{7\kappa(\boldsymbol{H}_0)}{2}. \tag{B.213}$$

Replacing this in (B.187), we have

$$\mathscr{D} \leq (1+2\alpha)r^{1/2}\left(7r\delta\kappa(\boldsymbol{H}_0) + \frac{4\sigma_{\max}(\boldsymbol{H}_0)\delta + 2\sqrt{r}\|\boldsymbol{z}_0\|_2\delta}{\mu}\right) + 3\delta r^{1/2} \tag{B.214}$$

$$\leq 3\delta\sqrt{r}\left(8r\kappa(\boldsymbol{H}_0) + \frac{4\sigma_{\max}(\boldsymbol{H}_0) + 2\sqrt{r}\|\boldsymbol{z}_0\|_2}{\mu}\right) \tag{B.215}$$

Hence, using (B.188) under assumption (B.203), we have

$$\mathscr{L}(\boldsymbol{H}_0, \widehat{\boldsymbol{H}})^{1/2} \leq 3(1+\sqrt{2})\kappa(\boldsymbol{H}_0)r\left(8r\kappa(\boldsymbol{H}_0) + \frac{4\sigma_{\max}(\boldsymbol{H}_0) + 2\sqrt{r}\|\boldsymbol{z}_0\|_2}{\mu}\right)\frac{\delta}{\alpha} \tag{B.216}$$

$$\leq 120\kappa(\boldsymbol{H}_0)\max\left\{r\kappa(\boldsymbol{H}_0), \frac{\sigma_{\max}(\boldsymbol{H}_0) + \sqrt{r}\|\boldsymbol{z}_0\|_2}{\mu}\right\}\frac{r\delta}{\alpha}. \tag{B.217}$$

Hence, for C''_* as defined in the statement of the theorem, we get

$$\mathscr{L}(\boldsymbol{H}_0, \widehat{\boldsymbol{H}})^{1/2} \leq \frac{C''_* r}{\alpha}\max_{i\leq n}\|\boldsymbol{Z}_{i,\cdot}\|_2 \tag{B.218}$$

This completes the proof.

C Proof of Proposition 4.2

The proof follows immediately from the following two propositions.

Proposition C.1. *Let $\boldsymbol{X} \in \mathbb{R}^{n\times d}$ and $D(\boldsymbol{x},\boldsymbol{y}) = \|\boldsymbol{x}-\boldsymbol{y}\|_2^2$. Then the gradient of the function $\boldsymbol{u} \mapsto \mathscr{D}(\boldsymbol{u}, \boldsymbol{X})$ is given by*

$$\nabla_{\boldsymbol{u}}\mathscr{D}(\boldsymbol{u}, \boldsymbol{X}) = 2(\boldsymbol{u} - \boldsymbol{\Pi}_{\mathrm{conv}(\boldsymbol{X})}(\boldsymbol{u}))\,. \tag{C.1}$$

Proof. Note that $\mathscr{D}(\boldsymbol{u}, \boldsymbol{X})$ is the solution of the following convex optimization problem.

$$\begin{aligned} \text{minimize} \quad & \|\boldsymbol{u} - \boldsymbol{y}\|_2^2 , \\ \text{subject to} \quad & \boldsymbol{y} = \boldsymbol{X}^\mathsf{T} \boldsymbol{\pi}, \\ & \boldsymbol{\pi} \geq 0, \\ & \langle \boldsymbol{\pi}, \mathbf{1} \rangle = 1. \end{aligned} \tag{C.2}$$

The Lagrangian for this problem is

$$\mathcal{L}(\boldsymbol{y}, \boldsymbol{\pi}, \boldsymbol{\rho}, \tilde{\rho}, \boldsymbol{\lambda}) = \|\boldsymbol{u} - \boldsymbol{y}\|_2^2 + \big\langle \boldsymbol{\rho}, (\boldsymbol{y} - \boldsymbol{X}^\mathsf{T} \boldsymbol{\pi}) \big\rangle - \langle \boldsymbol{\lambda}, \boldsymbol{\pi} \rangle + \tilde{\rho}(1 - \langle \boldsymbol{\pi}, \mathbf{1} \rangle). \tag{C.3}$$

The KKT condition implies that at the minimizer $(\boldsymbol{y}^*, \boldsymbol{\pi}^*, \boldsymbol{\rho}^*, \tilde{\rho}^*, \boldsymbol{\lambda}^*)$, we have

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{y}} = 0 \, , \tag{C.4}$$

and therefore

$$\boldsymbol{\rho}^* = 2(\boldsymbol{u} - \boldsymbol{y}^*) \tag{C.5}$$

and the dual of the above optimization problem is

$$\begin{aligned} \text{maximize} \quad & -\frac{1}{4}\|\boldsymbol{\rho}\|_2^2 + \langle \boldsymbol{\rho}, \boldsymbol{u} \rangle + \tilde{\rho}, \\ \text{subject to} \quad & \boldsymbol{\lambda} \geq 0, \\ & \boldsymbol{X}\boldsymbol{\rho} + \tilde{\rho}\mathbf{1} + \boldsymbol{\lambda} = 0. \end{aligned} \tag{C.6}$$

Note that since (C.2) is strictly feasible, Slater condition holds and by strong duality the optimal value of (C.6) is equal to $f(\boldsymbol{u})$. Hence, we have written $f(\boldsymbol{u})$ as pointwise supremum of functions. Therefore, subgradient of $f(\boldsymbol{u})$ can be achieved by taking the derivative of the objective function in (C.6) at the optimal solution (see Section 2.10 in [MN13]). Note that the derivative of this objective function at the optimal solution is

equal to $\boldsymbol{\rho}^* = 2(\boldsymbol{u} - \boldsymbol{y}^*) = 2(\boldsymbol{u} - \boldsymbol{\Pi}_{\mathrm{conv}(\boldsymbol{X})}(\boldsymbol{u}))$ (where we used Eq. (C.5)). Since the dual optimum is unique (by strong convexity in $\boldsymbol{\rho}$), the function $\boldsymbol{u} \mapsto \mathscr{D}(\boldsymbol{u}, \boldsymbol{X})$ is differentiable with gradient given by Eq. (C.1). □

Proposition C.2. *Let $\boldsymbol{u} \in \mathbb{R}^d$ and $D(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\|_2^2$, and assume that the rows of $\boldsymbol{H}_0 \in \mathbb{R}^{r \times d}$ are affine independent. Then the function $\boldsymbol{H} \mapsto \mathscr{D}(\boldsymbol{u}, \boldsymbol{H})$ is differentiable at $\boldsymbol{H}_0$ with gradient*

$$\nabla_{\boldsymbol{H}} \mathscr{D}(\boldsymbol{u}, \boldsymbol{H}_0) = 2\boldsymbol{\pi}_0(\boldsymbol{\Pi}_{\mathrm{conv}(\boldsymbol{H}_0)}(\boldsymbol{u}) - \boldsymbol{u})^{\mathsf{T}}, \qquad \boldsymbol{\pi}_0 = \arg\min_{\boldsymbol{\pi} \in \Delta^r} \left\| \boldsymbol{H}_0^{\mathsf{T}} \boldsymbol{\pi} - \boldsymbol{u} \right\|_2^2 . \tag{C.7}$$

Proof. We will denote by $\boldsymbol{G}$ the right hand side of Eq. (C.7). For $\boldsymbol{V} \in \mathbb{R}^{r \times d}$, we have

$$\mathscr{D}(\boldsymbol{u}, \boldsymbol{H}_0 + \boldsymbol{V}) = \min_{\boldsymbol{\pi} \in \Delta^r} \left\| (\boldsymbol{H}_0 + \boldsymbol{V})^{\mathsf{T}} \boldsymbol{\pi} - \boldsymbol{u} \right\|_2^2 . \tag{C.8}$$

Note that $(\boldsymbol{H}_0 + \boldsymbol{V})$ has affinely independent rows for $\boldsymbol{V}$ in a neighborhood of $\mathbf{0}$, and hence has a unique minimizer there, that we will denote by $\boldsymbol{\pi}_{\boldsymbol{V}}$. By optimality of $\boldsymbol{\pi}_{\boldsymbol{V}}$, we have

$$\begin{aligned}
\mathscr{D}(\boldsymbol{u}, \boldsymbol{H}_0 + \boldsymbol{V}) - \mathscr{D}(\boldsymbol{u}, \boldsymbol{H}_0) &= \left\| (\boldsymbol{H}_0 + \boldsymbol{V})^{\mathsf{T}} \boldsymbol{\pi}_{\boldsymbol{V}} - \boldsymbol{u} \right\|_2^2 - \left\| (\boldsymbol{H}_0 + \boldsymbol{V})^{\mathsf{T}} \boldsymbol{\pi}_0 - \boldsymbol{u} \right\|_2^2 && \text{(C.9)} \\
&\le \left\| (\boldsymbol{H}_0 + \boldsymbol{V})^{\mathsf{T}} \boldsymbol{\pi}_0 - \boldsymbol{u} \right\|_2^2 - \left\| (\boldsymbol{H}_0 + \boldsymbol{V})^{\mathsf{T}} \boldsymbol{\pi}_0 - \boldsymbol{u} \right\|_2^2 && \text{(C.10)} \\
&= \langle \boldsymbol{G}, \boldsymbol{V} \rangle + \|\boldsymbol{V} \boldsymbol{\pi}_0\|_2^2 . && \text{(C.11)}
\end{aligned}$$

On the other hand, by optimality of $\boldsymbol{\pi}_0$,

$$\begin{aligned}
\mathscr{D}(\boldsymbol{u}, \boldsymbol{H}_0 + \boldsymbol{V}) - \mathscr{D}(\boldsymbol{u}, \boldsymbol{H}_0) &\ge \left\| (\boldsymbol{H}_0 + \boldsymbol{V})^{\mathsf{T}} \boldsymbol{\pi}_{\boldsymbol{V}} - \boldsymbol{u} \right\|_2^2 - \left\| (\boldsymbol{H}_0 + \boldsymbol{V})^{\mathsf{T}} \boldsymbol{\pi}_{\boldsymbol{V}} - \boldsymbol{u} \right\|_2^2 && \text{(C.12)} \\
&= \langle 2\boldsymbol{\pi}_{\boldsymbol{V}} (\boldsymbol{\Pi}_{\mathrm{conv}(\boldsymbol{H}_0)}(\boldsymbol{u}) - \boldsymbol{u})^{\mathsf{T}}, \boldsymbol{V} \rangle + + \|\boldsymbol{V} \boldsymbol{\pi}_{\boldsymbol{V}}\|_2^2 && \text{(C.13)} \\
&= \langle \boldsymbol{G}, \boldsymbol{V} \rangle + 2\langle (\boldsymbol{\pi}_{\boldsymbol{V}} - \boldsymbol{\pi}_0)(\boldsymbol{\Pi}_{\mathrm{conv}(\boldsymbol{H}_0)}(\boldsymbol{u}) - \boldsymbol{u})^{\mathsf{T}}, \boldsymbol{V} \rangle + \|\boldsymbol{V} \boldsymbol{\pi}_{\boldsymbol{V}}\|_2^2 . && \text{(C.14)}
\end{aligned}$$

Letting $R(\boldsymbol{V}) = |\mathscr{D}(\boldsymbol{u}, \boldsymbol{H}_0 + \boldsymbol{V}) - \mathscr{D}(\boldsymbol{u}, \boldsymbol{H}_0) - \langle \boldsymbol{G}, \boldsymbol{V} \rangle|$ denote the residual, we get

$$\frac{R(\boldsymbol{V})}{\|\boldsymbol{V}\|_F} \le \|\boldsymbol{\Pi}_{\mathrm{conv}(\boldsymbol{H}_0)}(\boldsymbol{u}) - \boldsymbol{u}\|_2 \|\boldsymbol{\pi}_{\boldsymbol{V}} - \boldsymbol{\pi}_0\|_2 + \|\boldsymbol{V}\|_F (\|\boldsymbol{\pi}_{\boldsymbol{V}}\|_2 + \|\boldsymbol{\pi}_0\|_2) . \tag{C.15}$$

Note that $\boldsymbol{\pi}_{\boldsymbol{V}}$ must converge to $\boldsymbol{\pi}_0$ as $\boldsymbol{V} \to 0$ because $\boldsymbol{\pi}_0$ is the unique minimizer for $\boldsymbol{V} = \mathbf{0}$. Hence we get $R(\boldsymbol{V})/\|\boldsymbol{V}\|_F \to 0$ as $\|\boldsymbol{V}\|_F \to 0$, which proves our claim. □

D Proof of Proposition 4.1

We use the results of [BST14] to prove Proposition 4.1. We refer the reader to [BST14] for the definitions of the technical terms in this section. First, consider the function

$$f(\boldsymbol{H}) = \lambda \mathscr{D}(\boldsymbol{H}, \boldsymbol{X}). \tag{D.1}$$

Note that using the main theorem of polytope theory (Theorem 1.1 in [Zie12]), we can write

$$\text{conv}(\boldsymbol{X}) = \left\{ \boldsymbol{x} \in \mathbb{R}^d \,|\, \langle \boldsymbol{a}_i, \boldsymbol{x} \rangle \leq b_i \;\text{ for }\; 1 \leq i \leq m \right\} \tag{D.2}$$

for some $\boldsymbol{a}_i \in \mathbb{R}^d$, $b_i \in \mathbb{R}$ and a finite m. Hence, using the definition of the semi-algebraic sets (see Definition 5 in [BST14]), the set $\text{conv}(\boldsymbol{X})$ is semi-algebraic. Therefore, the function $f(\boldsymbol{H})$ which is proportional to the sum of squared ℓ_2 distances of the rows of $\boldsymbol{H}$ from a semi-algebraic set, is a semi-algebraic function (See Appendix in [BST14]). Further, the function

$$g(\boldsymbol{W}) = \sum_{i=1}^{n} \mathbf{I}\left(\boldsymbol{w}_i \in \Delta^r\right) \tag{D.3}$$

is the sum of indicator functions of semi-algebraic sets (Note that using the same argument used for $\text{conv}(\boldsymbol{X})$, Δ^r is semi-algebraic). Therefore, the function g is semi-algebraic (See Appendix in [BST14]). In addition, the function

$$h(\boldsymbol{H}, \boldsymbol{W}) = \|\boldsymbol{X} - \boldsymbol{W}\boldsymbol{H}\|_F^2 \tag{D.4}$$

is a polynomial. Hence, it is semi-algebraic. Therefore, we deduce that the function

$$\Psi(\boldsymbol{H}, \boldsymbol{W}) = f(\boldsymbol{H}) + g(\boldsymbol{W}) + h(\boldsymbol{H}, \boldsymbol{W}) \tag{D.5}$$

is semi-algebraic. In addition, since Δ^r is closed, Ψ is proper and lower semi-continuous. Therefore, $\Psi(\boldsymbol{H}, \boldsymbol{W})$ is a KL function (See Theorem 3 in [BST14]).

Now, we will show that the Assumptions 1,2 in [BST14] hold for our algorithm. First, note that since Δ^r is closed, the functions $f(\boldsymbol{H})$ and $g(\boldsymbol{W})$ are proper and lower semi-continuous. Further, $f(\boldsymbol{H}) \geq 0$, $g(\boldsymbol{W}) \geq 0$, $h(\boldsymbol{H}, \boldsymbol{W}) \geq 0$ for all $\boldsymbol{H} \in \mathbb{R}^{r\times d}$, $\boldsymbol{W} \in \mathbb{R}^{n\times r}$. In addition, the function $h(\boldsymbol{H}, \boldsymbol{W})$ is C^2. Therefore, it is Lipschitz continuous over the bounded subsets of $\mathbb{R}^{r\times d} \times \mathbb{R}^{n\times r}$. Also, the partial derivatives of $h(\boldsymbol{H}, \boldsymbol{W})$ are

$$\nabla_{\boldsymbol{H}} h(\boldsymbol{H}, \boldsymbol{W}) = 2\boldsymbol{W}^\mathsf{T}(\boldsymbol{W}\boldsymbol{H} - \boldsymbol{X}), \tag{D.6}$$

$$\nabla_{\boldsymbol{W}} h(\boldsymbol{H}, \boldsymbol{W}) = 2(\boldsymbol{W}\boldsymbol{H} - \boldsymbol{X})\boldsymbol{H}^\mathsf{T}. \tag{D.7}$$

It can be seen that for any fixed $\boldsymbol{W}$, the function $\boldsymbol{H} \mapsto \nabla_{\boldsymbol{H}} h(\boldsymbol{H}, \boldsymbol{W})$ is Lipschitz continuous with moduli $L_1(\boldsymbol{W}) = 2\|\boldsymbol{W}^\mathsf{T}\boldsymbol{W}\|_F$. Similarly, for any fixed $\boldsymbol{H}$, the function $\boldsymbol{W} \mapsto \nabla_{\boldsymbol{W}} h(\boldsymbol{H}, \boldsymbol{W})$ is Lipschitz continuous with moduli $L_2(\boldsymbol{H}) = 2\|\boldsymbol{H}\boldsymbol{H}^\mathsf{T}\|_F$. Note that since in each iteration of the algorithm the rows of $\boldsymbol{W}^k$ are in Δ^r. Hence,

$$\inf\left\{L_1(\boldsymbol{W}^k) : k \in \mathbb{N}\right\} \geq \lambda_1^-, \qquad \sup\left\{L_1(\boldsymbol{W}^k) : k \in \mathbb{N}\right\} \leq \lambda_1^+ \tag{D.8}$$

for some positive constants λ_1^-, λ_1^+. In addition, note that because the PALM algorithm is a descent algorithm, i.e., $\Psi(\boldsymbol{H}^k, \boldsymbol{W}^k) \leq \Psi(\boldsymbol{H}^{k-1}, \boldsymbol{W}^{k-1})$ for $k \in \mathbb{N}$, and since $f(\boldsymbol{H}) \to \infty$ as $\|\boldsymbol{H}\|_F \to \infty$, the value of $L_2(\boldsymbol{H}^k) = \|\boldsymbol{H}^k\boldsymbol{H}^{k^\mathsf{T}}\|_F$ remains bounded in every iteration. Finally, note that by taking $\gamma_2^k > \max\left\{\left\|\boldsymbol{H}^{k+1}\boldsymbol{H}^{k+1^\mathsf{T}}\right\|_F, \varepsilon\right\}$ for some constant $\varepsilon > 0$, we make sure that the steps in the PALM algorithm remain well defined (See Remark 3(iii) in [BST14]). Hence, we have shown that the assumptions of Theorem 1 in [BST14] hold.

Therefore, using this theorem, the sequence $\left\{\boldsymbol{H}^k, \boldsymbol{W}^k\right\}_{k\in\mathbb{N}}$ generated by the iterations in (4.7) - (4.9) has a finite length and it converges to a stationary point $(\boldsymbol{H}^*, \boldsymbol{W}^*)$ of Ψ.

E Other optimization algorithms

Apart from the proximal alternating linearized minimization discussed in Section 4.2, we experimented with two other algorithms, obtaining comparable results. For the sake of completeness, we describe these algorithms here.

E.1 Stochastic gradient descent

Using any of the initializations discussed in Section 4.1 we iterate

$$\boldsymbol{H}^{(t+1)} = \boldsymbol{H}^{(0)} - \gamma_t \boldsymbol{G}^{(t)} \,. \tag{E.1}$$

The step size γ_t is selected by backtracking line search. Ideally, the direction $\boldsymbol{G}^{(t)}$ can be taken to be equal to $\nabla\mathscr{R}_\lambda(\boldsymbol{H}^{(t)})$. However, for large datasets this is computationally impractical, since it requires to compute the projection of each data point onto the set $\mathrm{conv}(\boldsymbol{H}^{(t)})$. In order to reduce the complexity of the direction calculation, we estimate this sum by subsampling. Namely, we draw a uniformly random set $S_t \subseteq [n]$ of fixed size $|S_t| = s \leq n$, and compute

$$\boldsymbol{G}^{(t)} = \frac{2n}{|S_t|} \sum_{i\in S_t} \boldsymbol{\alpha}_i^* \left(\boldsymbol{\Pi}_{\mathrm{conv}(\boldsymbol{H})}\left(\boldsymbol{x}_i\right) - \boldsymbol{x}_i\right) + 2\lambda\left(\boldsymbol{H} - \boldsymbol{\Pi}_{\mathrm{conv}(\boldsymbol{X})}\left(\boldsymbol{H}\right)\right) \,, \tag{E.2}$$

$$\boldsymbol{\alpha}_i^* = \arg\min_{\boldsymbol{\alpha}\in\Delta^r} \left\|\boldsymbol{H}^\mathsf{T}\boldsymbol{\alpha} - \boldsymbol{x}_i^\mathsf{T}\right\|_2 \,. \tag{E.3}$$

E.2 Alternating minimization

This approach generalizes the original algorithm of [CB94]. We rewrite the objective as a function of $\boldsymbol{W} = (\boldsymbol{w}_i)_{i\le n}$, $\boldsymbol{w}_i \in \Delta^r$,$\boldsymbol{H} = (\boldsymbol{h}_i)_{i\le r}$, $\boldsymbol{h}_i \in \mathbb{R}^d$ and $\boldsymbol{A} = (\boldsymbol{\alpha}_\ell)_{\ell\le r}$, $\boldsymbol{\alpha}_\ell \in \Delta^n$

$$\mathscr{R}_\lambda(\boldsymbol{H}) = \min_{\boldsymbol{W},\boldsymbol{A}} F(\boldsymbol{H},\boldsymbol{W},\boldsymbol{A})\,, \tag{E.4}$$

$$F(\boldsymbol{H},\boldsymbol{W},\boldsymbol{A}) = \sum_{i=1}^n \Big\|\boldsymbol{x}_i - \sum_{\ell=1}^r w_{i\ell}\boldsymbol{h}_\ell\Big\|_2^3 + \lambda\sum_{\ell=1}^r \Big\|\boldsymbol{h}_\ell - \sum_{i=1}^n \alpha_{\ell,i}\boldsymbol{x}_i\Big\|_2^2. \tag{E.5}$$

The algorithm alternates between minimizing with respect to the weights $(\boldsymbol{w}_i)_{i\le n}$ (this can be done independently across $i \in \{1,\dots,n\}$) and minimizing over $(\boldsymbol{h}_\ell,\boldsymbol{\alpha}_\ell)$, which is done sequentially by cycling over $\ell \in \{1,\dots,r\}$. Minimization over $\boldsymbol{w}_i$ can be performed by solving a non-negative least squares problem. As shown in [CB94], minimization over $(\boldsymbol{h}_\ell,\boldsymbol{\alpha}_\ell)$ is also equivalent to non-negative least squares. Indeed, by a simple calculation

$$F(\boldsymbol{H},\boldsymbol{W},\boldsymbol{A}) = w_\ell^{\rm tot}\big\|\boldsymbol{h}_\ell - \boldsymbol{v}_\ell\big\|_2^2 + \lambda\Big\|\boldsymbol{h}_\ell - \sum_{i=1}^n \alpha_{\ell,i}\boldsymbol{x}_i\Big\|_2^2 + \widetilde{F}(\boldsymbol{H},\boldsymbol{W},\boldsymbol{A}) \tag{E.6}$$

$$= f_\ell(\boldsymbol{h}_\ell,\boldsymbol{\alpha}_\ell;\boldsymbol{H}_{\neq\ell},\boldsymbol{W},\boldsymbol{A}) + \widetilde{F}(\boldsymbol{H},\boldsymbol{W},\boldsymbol{A})\,. \tag{E.7}$$

where $\boldsymbol{H}_{\neq\ell} = (\boldsymbol{h}_i)_{i\neq\ell,i\le r}$, $\widetilde{F}(\boldsymbol{H},\boldsymbol{W},\boldsymbol{A})$ does not depend on $(\boldsymbol{h}_\ell,\boldsymbol{\alpha}_\ell)$, and we defined

$$w_\ell^{\rm tot} \equiv \sum_{i=1}^n w_{i\ell}^2\,, \tag{E.8}$$

$$\boldsymbol{v}_\ell \equiv \frac{1}{w_\ell^{\rm tot}}\sum_{i=1}^n w_{i,\ell}\left\{\boldsymbol{x}_i - \sum_{j\neq\ell,j\le r} w_{ij}\boldsymbol{h}_j\right\}. \tag{E.9}$$

It is therefore sufficient to minimize $f_\ell(\boldsymbol{h}_\ell,\boldsymbol{\alpha}_\ell;\boldsymbol{H}_{\neq\ell},\boldsymbol{W},\boldsymbol{A})$ with respect to its first two arguments, which is equivalent to a non-negative least squares problem. This can be seen by minimizing $f_\ell(\cdots)$ explicitly with respect to $\boldsymbol{h}_\ell$ and writing the resulting objective function.

The pseudocode for this algorithm is given below.

ALTERNATING MINIMIZATION

Input: Data $\{\boldsymbol{x}_i\}_{i\leq n}$, $\boldsymbol{x}_i \in \mathbb{R}^d$; integer r; initial archetypes $\{\boldsymbol{h}_\ell^{(0)}\}_{1\leq \ell\leq r}$; number of iterations T;

Output: Archetype estimates $\{\boldsymbol{h}_\ell^{(T)}\}_{1\leq \ell\leq r}$;

For $\ell \in \{1,\dots,r\}$:
    Set $\boldsymbol{\alpha}_\ell^{(0)} = \arg\min_{\boldsymbol{\alpha}\in\Delta^n} \|\boldsymbol{h}_\ell^{(0)} - \boldsymbol{X}\boldsymbol{\alpha}_\ell\|_2$;
For $t \in \{1,\dots,T\}$:
    Set $\boldsymbol{W}^t = \arg\min_{\boldsymbol{W}} F(\boldsymbol{H}^{t-1}, \boldsymbol{W}, \boldsymbol{A}^{t-1})$
    For $\ell \in \{1,\dots,r\}$:
    Set $\boldsymbol{h}_\ell^{(t)}, \boldsymbol{\alpha}_\ell^{(t)} = \arg\min_{\boldsymbol{h}_\ell,\boldsymbol{\alpha}_\ell} f_\ell(\boldsymbol{h}_\ell, \boldsymbol{\alpha}_\ell; \boldsymbol{H}^t_{<\ell}, \boldsymbol{H}^{t-1}_{>\ell}, \boldsymbol{W}^t, \boldsymbol{A}^t_{<\ell}, \boldsymbol{A}^{t-1}_{>\ell})$;
End For;
Return $\{\hat{\boldsymbol{h}}_\ell^{(T)}\}_{1\leq \ell\leq r}$;

Here $\boldsymbol{H}_{<\ell} = (\boldsymbol{h}_i)_{i<\ell}$, $\boldsymbol{H}_{>\ell} = (\boldsymbol{h}_i)_{\ell<i\leq r}$, and similarly for $\boldsymbol{A}$.

F Further simulation results

In this section, we evaluate the performances of the proposed method by comparing it with a number of algorithms for non-negative matrix factorization under different settings.

F.1 Further comparisons with algorithms from the literature

Figures 2 to 7 extend the comparison of Figures 1, 4 and 5 to seven alternative reconstruction methods in the literature. Namely:

- *No noise.* Figures 2 and 3 repeat the experiment of Figure 1 for seven new methods.

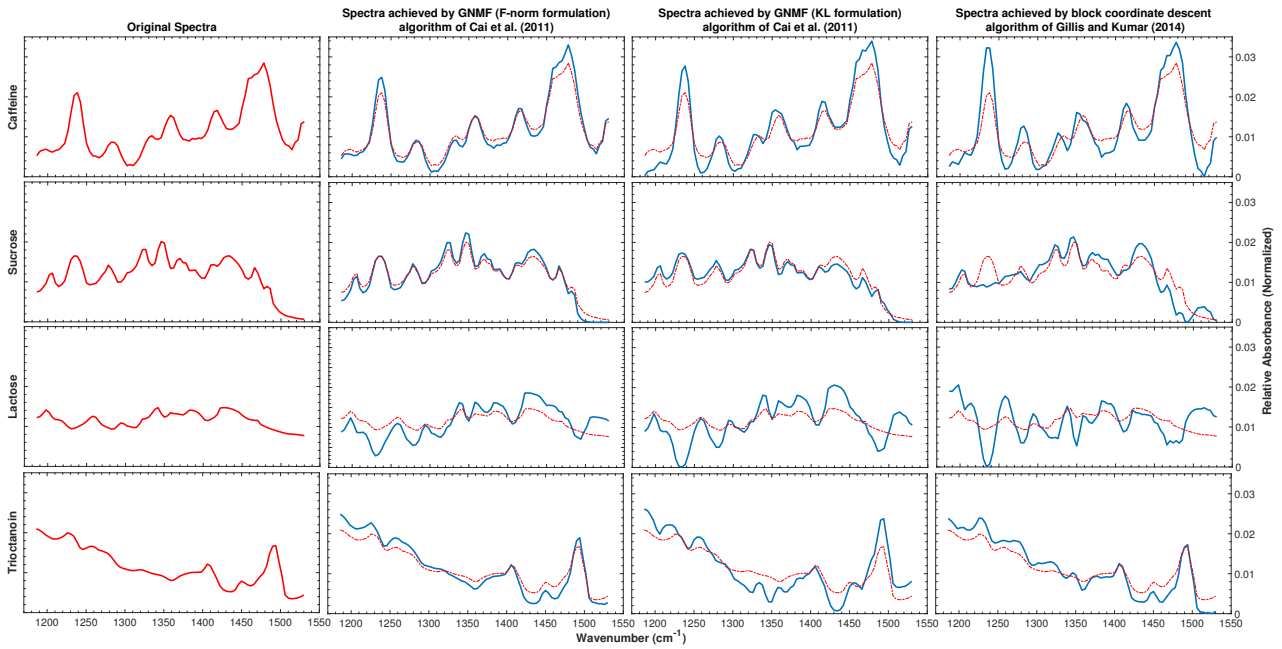

Figure 2: As in Figure 1 for three other methods.

- *Low noise.* Figures 4, 5 repeat the experiment of Figure 4 for the same seven new methods.

- *High noise.* Figures 6, 7 repeat the experiment of Figure 5 for the same seven new methods.

F.2 Simulations with non-Gaussian correlated noise

In order to show the robustness of the proposed method to the noise model, we have repeated the experiments of Table 1 with dependent, non-Gaussian noise (see caption for a definition of the noise model). In Table 1, we report the average reconstruction error achieved by the same nine algorithms for non-negative matrix factorization. For each noise

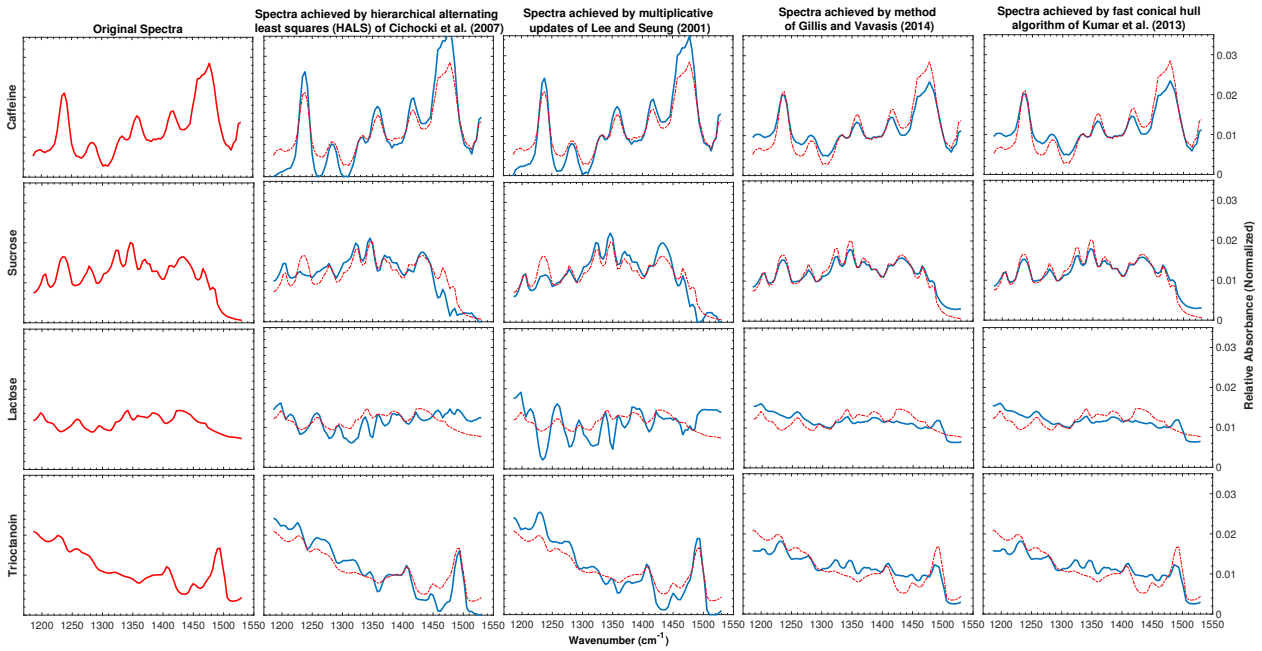

Figure 3: As in Figure 1 for four other methods.

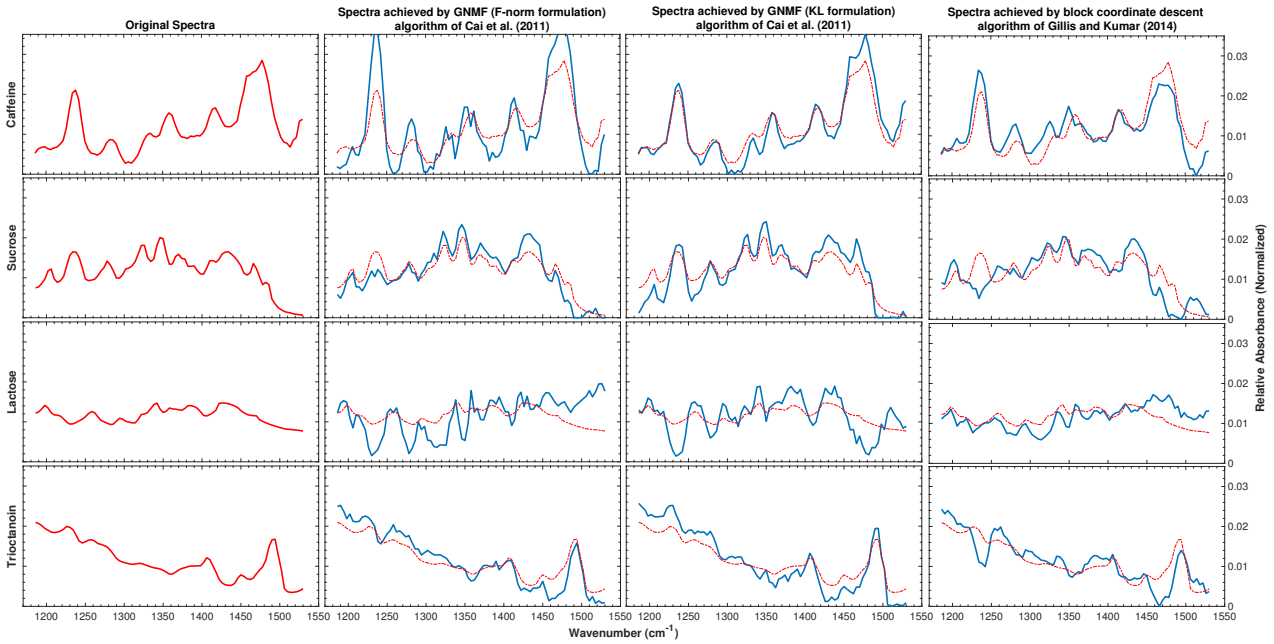

Figure 4: As in Figure 4 for three other methods.

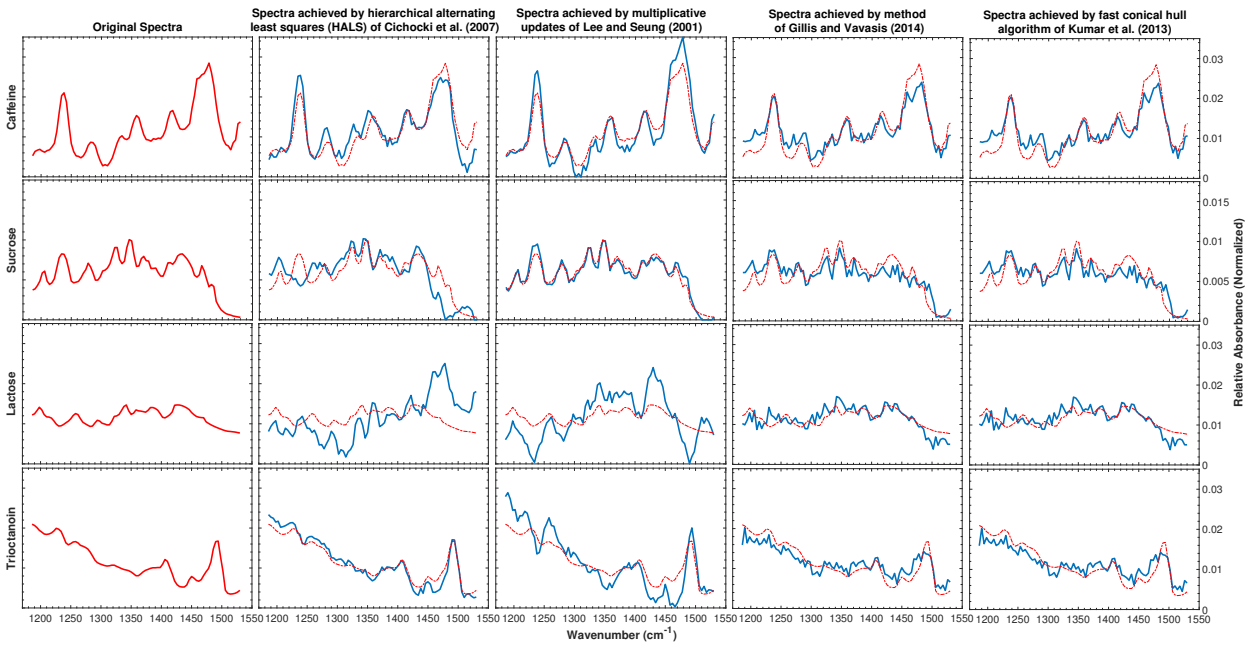

Figure 5: As in Figure 4 for four other methods.

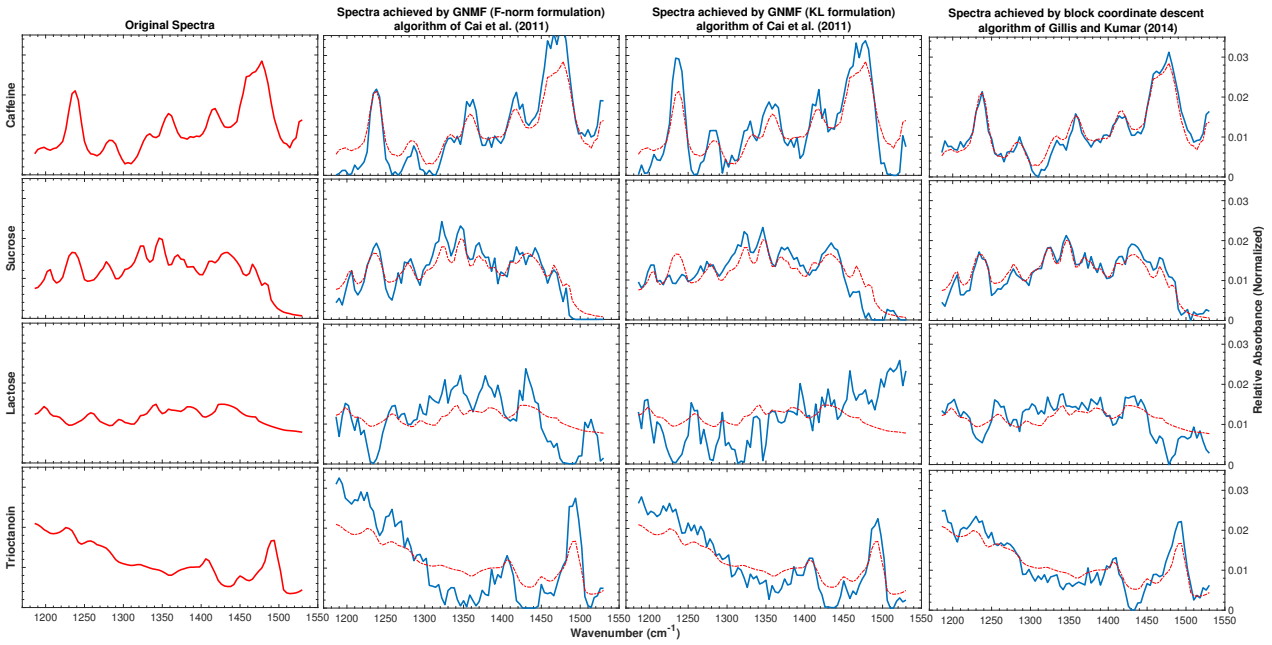

Figure 6: As in Figure 5 for three other methods.

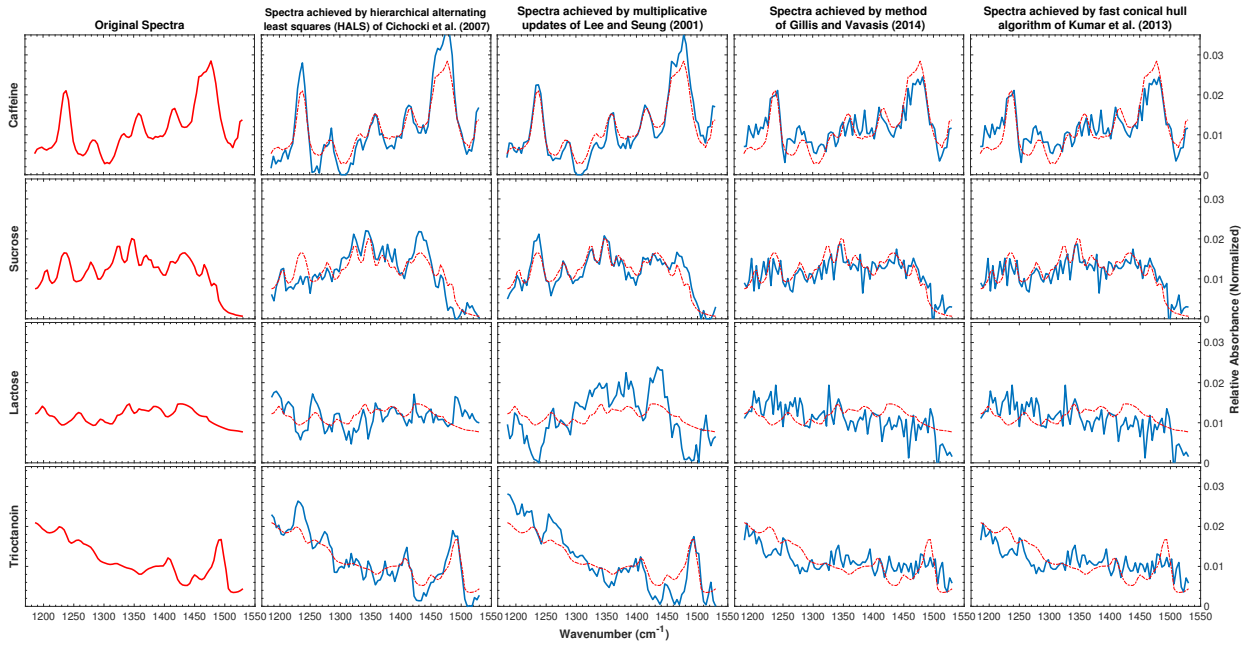

Figure 7: As in Figure 5 for four other methods.

level we run the various algorithms on 10 noise realizations for each noise level σ. We show in bold the smallest achieved error and the smallest error with data driven choice of the algorithm parameters.

σ	0	0.001	0.002	0.003	0.004	0.005	0.006	0.007	0.008	0.009	0.01
Projected gradient [Lin07]	0.062	0.058	0.06	0.07	0.089	0.101	0.122	0.133	0.142	0.143	0.151
Multiplicative update[LS01]	0.069	0.062	0.069	0.078	0.092	0.114	0.125	0.136	0.145	0.143	0.15
Fast Anchor Words [AGH$^+$13]	0.041	0.047	0.055	0.075	0.093	0.113	0.134	0.156	0.176	0.195	0.22
Block coordinate descent [GK15]	0.067	0.068	0.067	0.069	**0.077**	**0.086**	**0.092**	**0.095**	**0.096**	**0.098**	**0.099**
HALS [CZPA09]	0.073	0.077	0.074	0.077	0.95	0.112	0.117	0.131	0.140	0.145	0.151
GNMF [CHHH11] (Frobenius)	0.065	0.081	0.095	0.102	0.111	0.121	0.131	0.140	0.143	0.143	0.15
GNMF [CHHH11] (KL)	0.066	0.075	0.081	0.087	0.099	0.121	0.128	0.138	0.141	0.140	0.151
Recursive method [GV14]	0.034	0.04	0.053	0.068	0.089	0.111	0.13	0.15	0.17	0.19	0.21
Conical hull [KSK13]	0.034	0.04	0.052	0.068	0.088	0.111	0.13	0.15	0.17	0.19	0.21
Our method (oracle λ)	**0.005**	**0.015**	**0.039**	**0.06**	0.081	**0.091**	0.102	0.113	0.122	0.132	0.14
Our method (data driven λ)	**0.006**	**0.021**	**0.041**	**0.067**	0.094	0.108	0.124	0.134	0.147	0.161	0.178

Table 1: Risk $\mathscr{L}(\boldsymbol{H}_0, \widehat{\boldsymbol{H}})^{1/2}$ for reconstruction of the 4 spectra in Figure 1 under dependent, non-gaussian noise using some construction methods in different noise magnitudes. The trivial estimator $\widehat{\boldsymbol{H}} = 0$ achieves $\mathscr{L}(\boldsymbol{H}_0, \widehat{\boldsymbol{H}})^{1/2} = 0.231$. For this table I have generated the data as $\boldsymbol{X} = \boldsymbol{W}_0\boldsymbol{H}_0 + \mu\boldsymbol{Q}\boldsymbol{Z}$ where Z_{ij}, $1 \le i, j \le n$ are i.i.d Laplace $(0, 1)$ random variables and $\boldsymbol{Q} \in \mathbb{R}^{n\times n}$ is a circulant matrix with first row equal to $[1\,, \mathbf{0}_{n-(n_1+1)/2}\,, \rho\mathbf{1}_{(n_1-1)/2}]$ and $\mu = (\sigma/\sqrt{2})(1 + (n_1 - 1)\rho^2/2)^{-1/2}$. In these simulations, I have taken $\rho = 0.5$, $n_1 = 21$. For the data driven row, parameter λ is chosen as in Section 4.3 with $c_0 = 1.2$.

F.3 Simulations with other ground truth signals

We repeated the synthetic datasets experiments of Section 4 with using four other spectra as the ground truth signals. In these experiments we have used the reflection spectra of chalk, Maltose, Acetaminophen and baking soda from the NIST Chemistry WebBook dataset [LM]. For these signals we have $d = 107$ and we generate $n = 250$ data points as in Appendix A. We have also used different parameters to generate weight vectors in

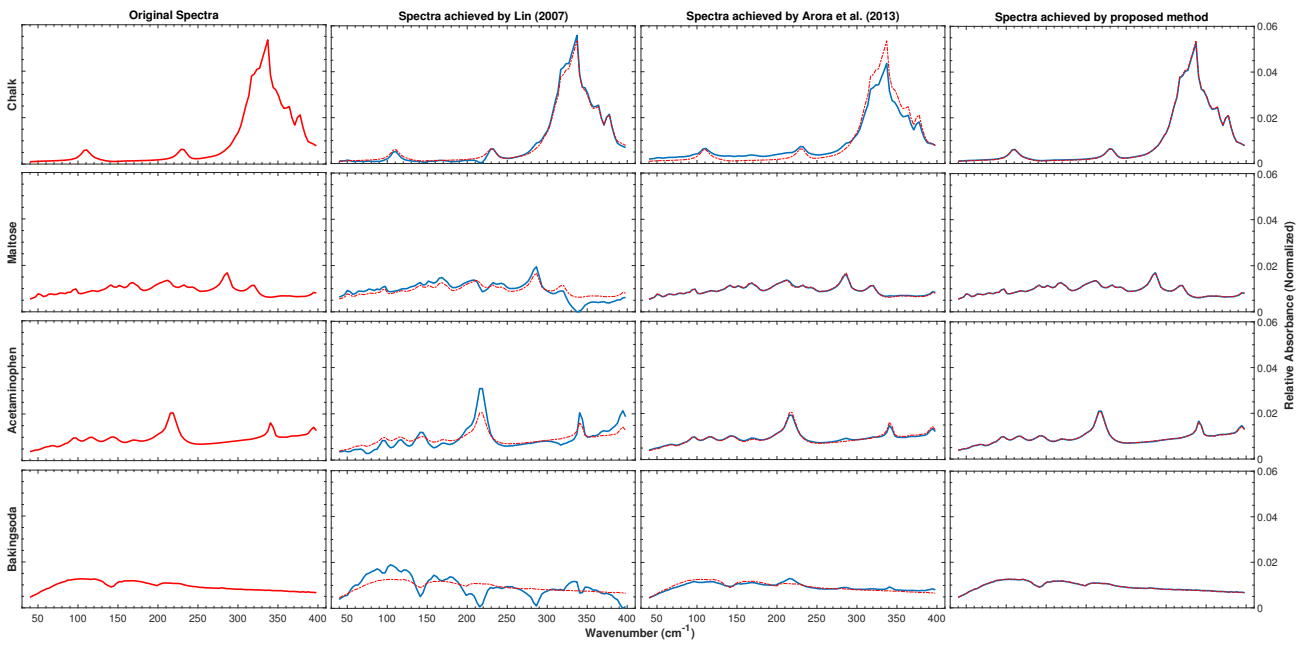

Figure 8: As in Figure 1 for four other ground truth spectra.

this case. Dirichlet parameters are chosen to be equal to one (instead of 5), the number of weight vectors with cardinality equal to 2 is equal to 12 and the number of weight vectors with cardinality equal to 3 is equal to 8. Other weight vectors have cardinality equal to 4. The recovered spectra of different algorithms in the noiseless and noisy settings are reported in Figures 8 to 16.

In Table 2, we report the average reconstruction error achieved by the same nine algorithms for non-negative matrix factorization on this dataset. For each noise level we use 10 noise realizations for each noise level σ.

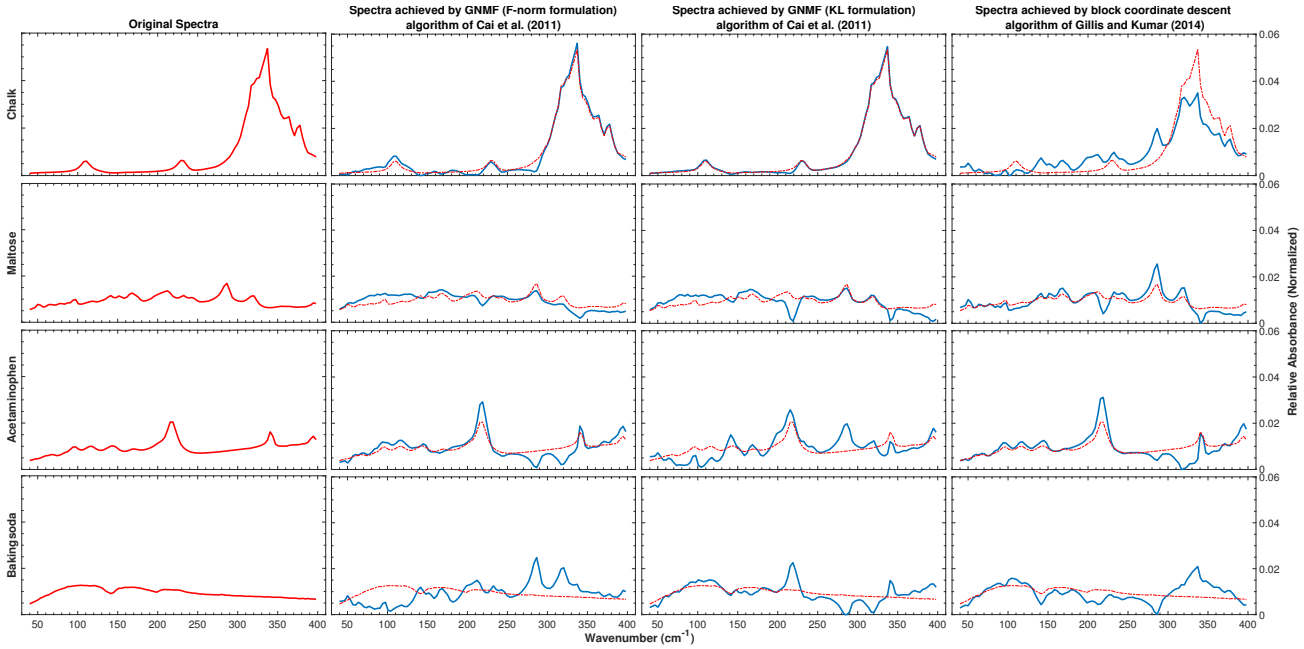

Figure 9: As in Figure 2 for four other ground truth spectra.

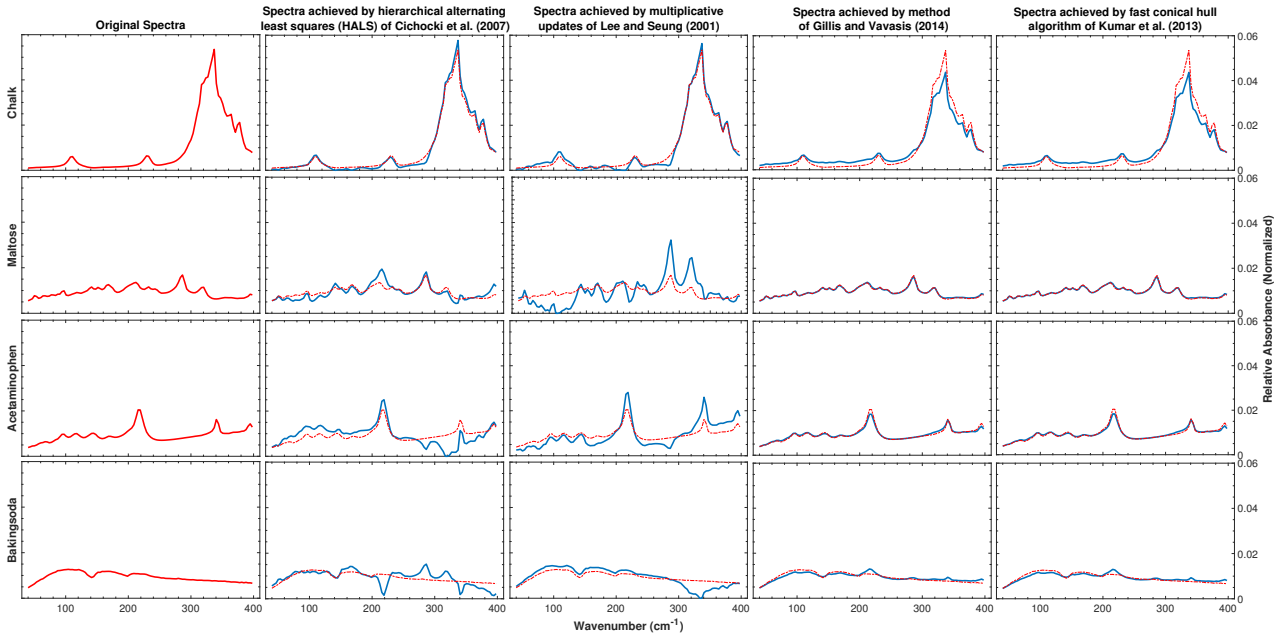

Figure 10: As in Figure 3 for four other ground truth spectra.

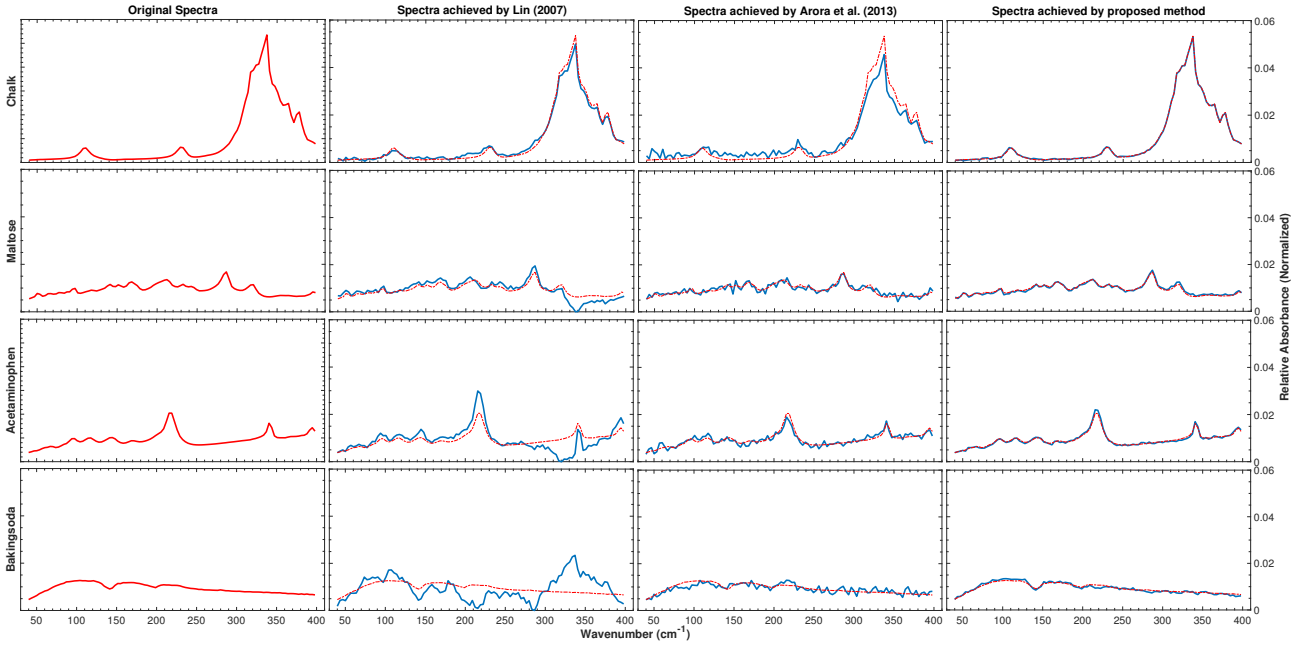

Figure 11: As in Figure 4 for four other ground truth spectra.

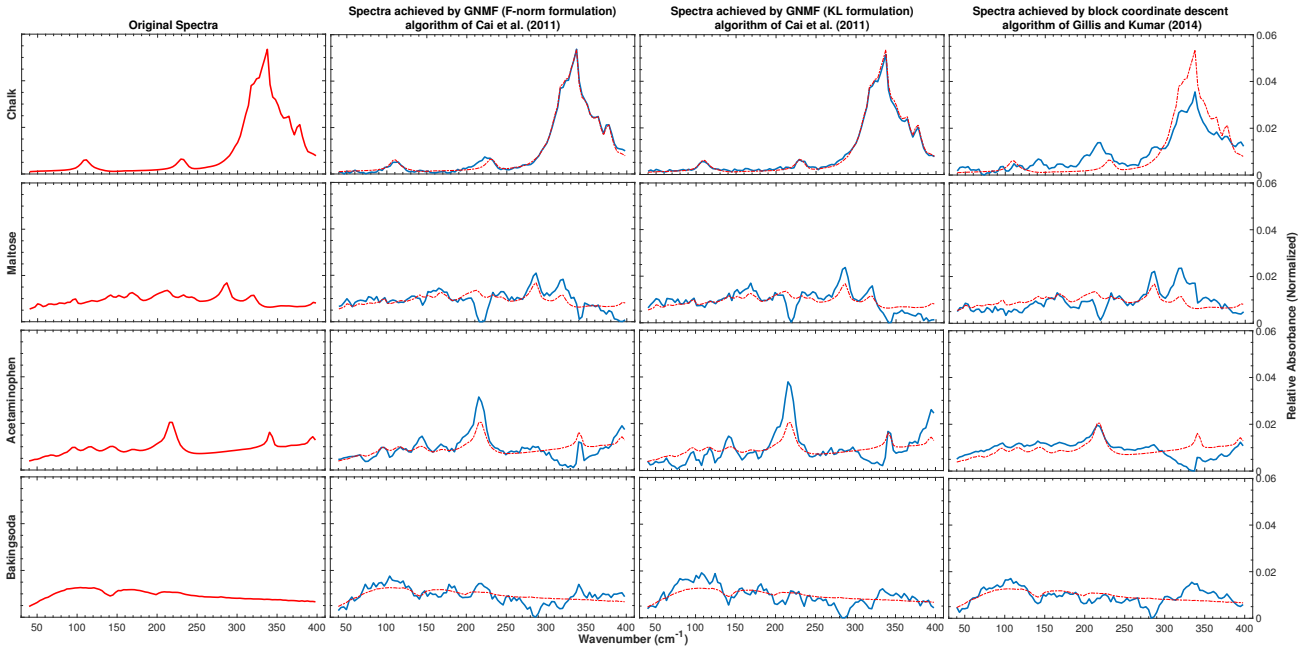

Figure 12: As in Figure 4 for four other ground truth spectra.

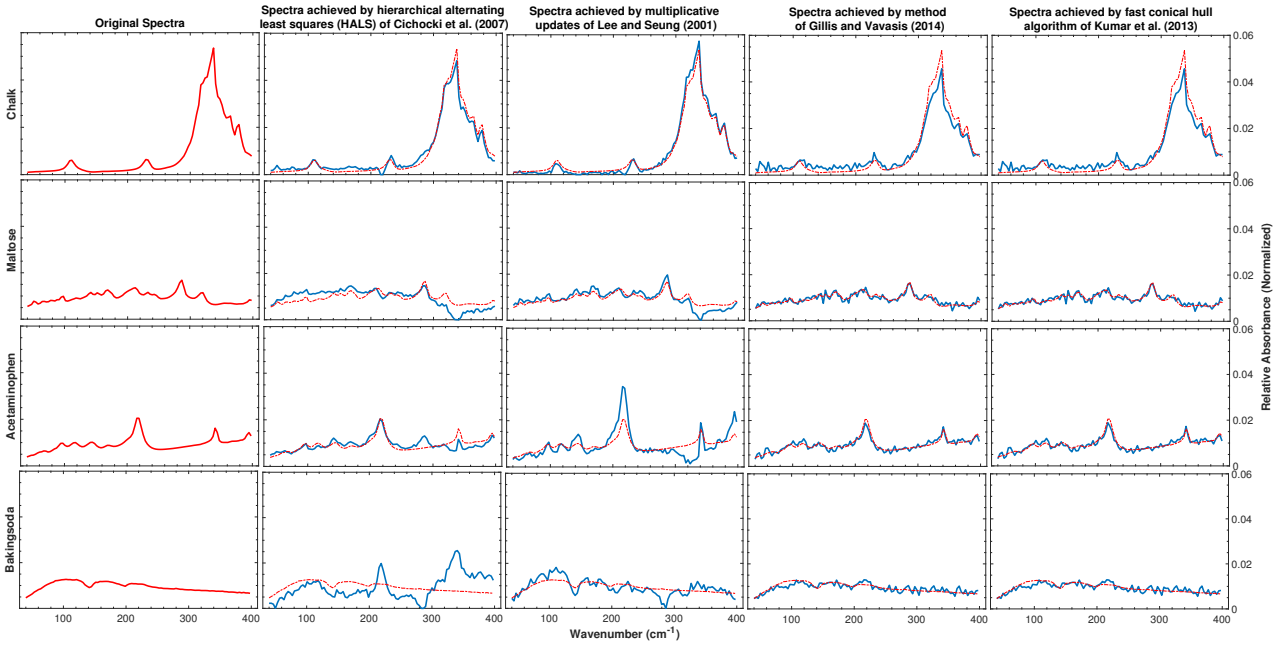

Figure 13: As in Figure 5 for four other ground truth spectra.

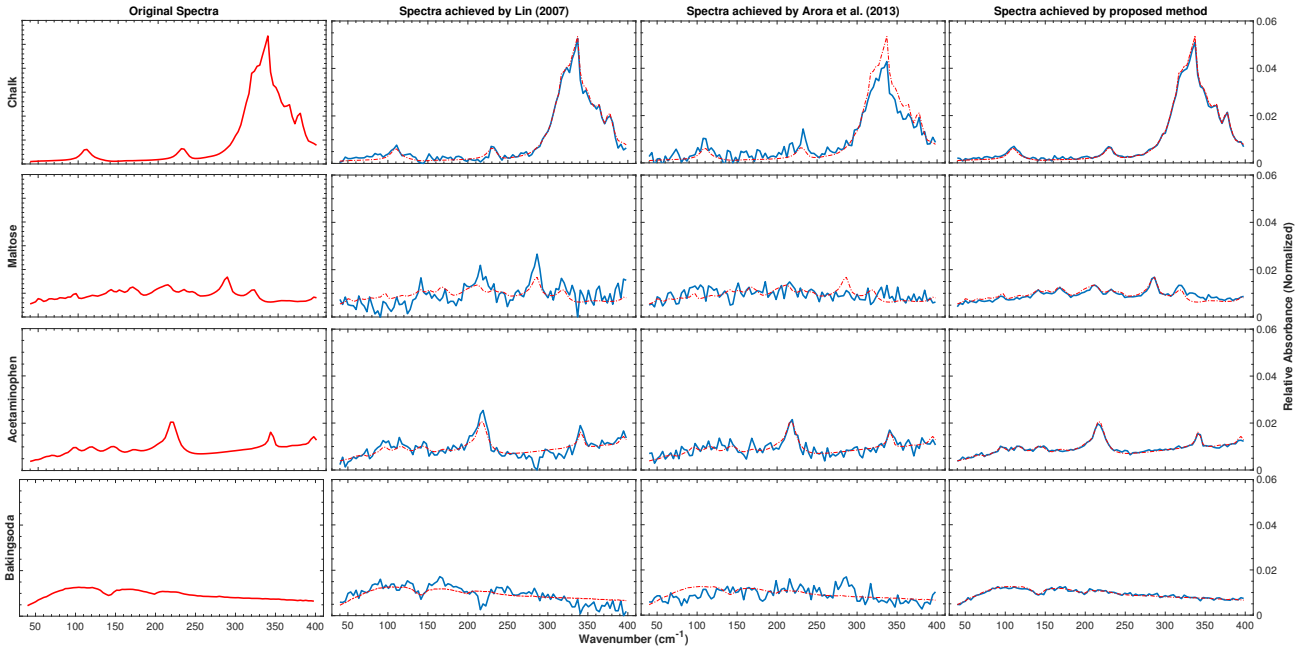

Figure 14: As in Figure 5 for four other ground truth spectra.

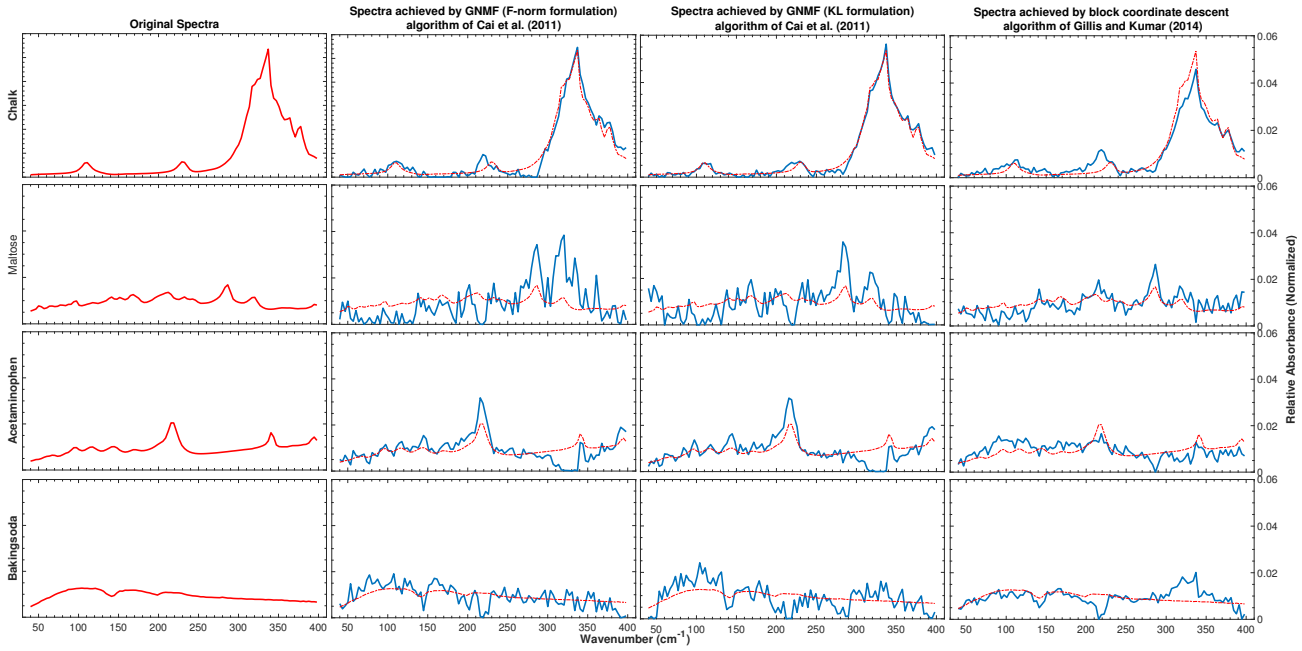

Figure 15: As in Figure 6 for four other ground truth spectra.

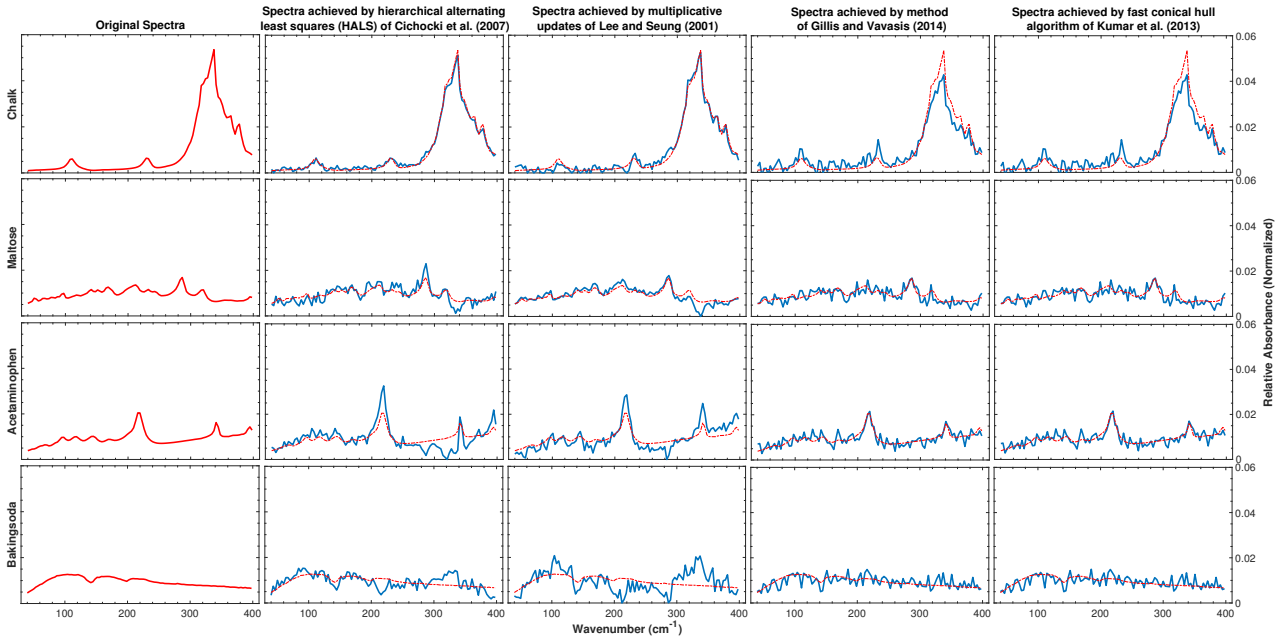

Figure 16: As in Figure 7 for four other ground truth spectra.

σ	0	0.001	0.002	0.003	0.004	0.005	0.006	0.007	0.008	0.009	0.01
Projected gradient [Lin07]	0.061	0.062	0.061	0.063	0.075	0.085	0.097	0.100	0.106	0.114	0.119
Multiplicative update [LS01]	0.064	0.064	0.064	0.069	0.075	0.086	0.092	0.099	0.110	0.115	0.119
Fast Anchor Words [AGH$^+$13]	0.045	0.041	0.056	0.075	0.094	0.119	0.140	0.161	0.183	0.208	0.231
Block coordinate descent[GK15]	0.089	0.089	0.085	0.086	0.083	0.091	0.095	0.088	**0.092**	**0.092**	**0.093**
HALS [CZPA09]	0.0601	0.059	0.064	0.067	0.073	0.089	0.095	0.100	0.109	0.115	0.120
GNMF [CHHH11] (Frobenius)	0.057	0.081	0.089	0.109	0.121	0.133	0.123	0.125	0.123	0.125	0.123
GNMF [CHHH11] (KL)	0.066	0.078	0.092	0.097	0.107	0.115	0.123	0.126	0.132	0.131	0.136
Recursive method [GV14]	0.031	0.039	0.054	0.074	0.093	0.117	0.138	0.159	0.183	0.205	0.229
Conical hull [KSK13]	0.031	0.039	0.054	0.074	0.093	0.117	0.138	0.159	0.183	0.205	0.229
Our method (oracle λ)	**0.002**	**0.007**	**0.014**	**0.023**	**0.040**	**0.049**	**0.068**	**0.064**	**0.068**	**0.072**	**0.076**
Our method (Data driven λ)	**0.003**	**0.009**	**0.014**	**0.027**	**0.052**	**0.06**	**0.075**	**0.087**	0.095	0.114	0.133

Table 2: Risk $\mathscr{L}(\boldsymbol{H}_0, \widehat{\boldsymbol{H}})^{1/2}$ for reconstruction of the 4 spectra in Figure 8 using some reconstruction methods in different noise magnitudes. The trivial estimator $\widehat{\boldsymbol{H}} = 0$ achieves $\mathscr{L}(\boldsymbol{H}_0, \widehat{\boldsymbol{H}})^{1/2} = 0.237$. For the data driven row, parameter λ is chosen as in subsection 4.3 with $c = 1.2$ in $\{0.001, 0.002, 0.005, 0.007, 0.01, 0.03, 0.05, 0.07, 0.1, 0.2, 0.3, 0.5, 0.75, 1, 2, 5\}$.

References

[AGH$^+$13] Sanjeev Arora, Rong Ge, Yonatan Halpern, David M Mimno, Ankur Moitra, David Sontag, Yichen Wu, and Michael Zhu, *A practical algorithm for topic modeling with provable guarantees.*, ICML (2), 2013, pp. 280–288.

[BST14] Jérôme Bolte, Shoham Sabach, and Marc Teboulle, *Proximal alternating linearized minimization for nonconvex and nonsmooth problems*, Mathematical Programming **146** (2014), no. 1-2, 459–494.

[CHHH11] Deng Cai, Xiaofei He, Jiawei Han, and Thomas S Huang, *Graph regularized nonnegative matrix factorization for data representation*, IEEE Transactions on Pattern Analysis and Machine Intelligence **33** (2011), no. 8, 1548–1560.

[CZPA09] Andrzej Cichocki, Rafal Zdunek, Anh Huy Phan, and Shun-ichi Amari, *Nonnegative matrix and tensor factorizations: applications to exploratory multiway data analysis and blind source separation*, John Wiley & Sons, 2009.

[CB94] Adele Cutler and Leo Breiman, *Archetypal analysis*, Technometrics **36** (1994), no. 4, 338–347.

[GV14] Nicolas Gillis and Stephen A Vavasis, *Fast and robust recursive algorithms for separable nonnegative matrix factorization*, IEEE transactions on pattern analysis and machine intelligence **36** (2014), no. 4, 698–714.

[GK15] Nicolas Gillis and Abhishek Kumar, *Exact and heuristic algorithms for semi-nonnegative matrix factorization*, SIAM Journal on Matrix Analysis and Applications **36** (2015), no. 4, 1404–1424.

[KSK13] Abhishek Kumar, Vikas Sindhwani, and Prabhanjan Kambadur, *Fast conical hull algorithms for near-separable non-negative matrix factorization*, International Conference on Machine Learning, 2013, pp. 231–239.

[LS01] ———, *Algorithms for non-negative matrix factorization*, Advances in neural information processing systems, 2001, pp. 556–562.

[Lin07] Chih-Jen Lin, *Projected gradient methods for nonnegative matrix factorization*, Neural computation **19** (2007), no. 10, 2756–2779.

[LM] P.J. Linstrom and W.G. Mallard (eds.), *Nist chemistry webbook, nist standard reference database number 69*, National Institute of Standards and Technology, Gaithersburg MD, 20899, `http://webbook.nist.gov`, (retrieved January 5, 2017).

[MN13] Boris S Mordukhovich and Nguyen Mau Nam, *An easy path to convex analysis and applications*, Synthesis Lectures on Mathematics and Statistics **6** (2013), no. 2, 1–218.

[Zie12] Günter M Ziegler, *Lectures on polytopes*, vol. 152, Springer Science & Business Media, 2012.