Archetypal Analysis

Philipp Sven Lars Schäfer

April 2025

1 Problem Formulation

Let $\mathcal{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}_{n=1}^{N}$ be a data set consisting of N D-dimensional data points, and let $\mathbf{X} \in \mathbb{R}^{N \times D}$ be the matrix where each row is a data point.

In Archetypal Analysis we make two assumptions: 1) Each data point is a convex combination of K archetypes; 2) Each archetype is a convex combination of N data points.

Expressing the first assumption in matrix notation yields

$$\hat{\mathbf{X}} = \mathbf{A}\mathbf{Z} \tag{1}$$

where $\hat{\mathbf{X}} \in \mathbb{R}^{N \times D}$ is the reconstructed data matrix, $\mathbf{Z} \in \mathbb{R}^{K \times D}$ is the matrix of archetypes (i.e. each row is one archetype), and $\mathbf{A} \in \mathbb{R}^{N \times K}$ is a row-stochastic matrix that defines

Expressing the second assumption in matrix notation yields

$$\mathbf{Z} = \mathbf{B}\mathbf{X} \tag{2}$$

where $\mathbf{B} \in \mathbb{R}^{K \times N}$ is a row-stochastic matrix that defines

The reconstruction error is most commonly measured using the residual sum of squares (RSS), given by the squared Frobenius norm,

$$\|\mathbf{X} - \hat{\mathbf{X}}\|_F^2 = \|\mathbf{X} - \mathbf{A}\mathbf{Z}\|_F^2 = \|\mathbf{X} - \mathbf{A}\mathbf{B}\mathbf{X}\|_F^2 \tag{3}$$

which yields the following optimization objective

$$\begin{aligned} \mathbf{A}^\star, \mathbf{B}^\star = \underset{\substack{\mathbf{A} \in \mathbb{R}^{N \times K} \\ \mathbf{B} \in \mathbb{R}^{K \times N}}}{\arg\min} \|\mathbf{X} - \mathbf{A}\mathbf{B}\mathbf{X}\|_F^2 \quad &\text{subject to} \\ \mathbf{A} \geq 0, \mathbf{A}\mathbf{1}_K = \mathbf{1}_N& \\ \mathbf{B} \geq 0, \mathbf{B}\mathbf{1}_N = \mathbf{1}_K& \end{aligned} \tag{4}$$

Introducing the set of row-stochastic non-negative matrices,

$$F(N, K) := \left\{ \mathbf{A} \in \mathbb{R}^{N \times K} \mid \mathbf{A} \geq 0 \wedge \mathbf{A}\mathbf{1}_K = \mathbf{1}_N \right\} \tag{5}$$

we can write the objective more compactly as:

$$\mathbf{A}^\star, \mathbf{B}^\star = \underset{\substack{\mathbf{A} \in F(N,K) \\ \mathbf{B} \in F(K,N)}}{\arg\min} \|\mathbf{X} - \mathbf{A}\mathbf{B}\mathbf{X}\|_F^2 \tag{6}$$

2 Properties of the Objective

Property 1 (Translation invariance): The minimizers $\mathbf{A}^\star, \mathbf{B}^\star$ of the objective are invariant under row-wise translations of $\mathbf{X}$. Let $\tilde{\mathbf{X}} = \mathbf{X} + \mathbf{1}_N \mathbf{v}^T$ for any $\mathbf{v} \in \mathbb{R}^D$, then

$$\underset{\substack{\mathbf{A}\in F(N,K)\\ \mathbf{B}\in F(K,N)}}{\arg\min} \|\tilde{\mathbf{X}} - \mathbf{A}\mathbf{B}\tilde{\mathbf{X}}\|_F^2 = \underset{\substack{\mathbf{A}\in F(N,K)\\ \mathbf{B}\in F(K,N)}}{\arg\min} \|\mathbf{X} - \mathbf{A}\mathbf{B}\mathbf{X}\|_F^2 \tag{7}$$

Proof: Let $\mathbf{v} \in \mathbb{R}^D$, and let $\tilde{\mathbf{X}} = \mathbf{X} + \mathbf{1}_N \mathbf{v}^T$ be the translated matrix. Then for any feasible $\mathbf{A}, \mathbf{B}$

$$\begin{aligned} \tilde{\mathbf{X}} - \mathbf{A}\mathbf{B}\tilde{\mathbf{X}} &= \left(\mathbf{X} + \mathbf{1}_N \mathbf{v}^T\right) - \mathbf{A}\mathbf{B}\left(\mathbf{X} + \mathbf{1}_N \mathbf{v}^T\right) \\ &= \mathbf{X} + \mathbf{1}_N \mathbf{v}^T - \mathbf{A}\mathbf{B}\mathbf{X} - \mathbf{A}\mathbf{B}\mathbf{1}_N \mathbf{v}^T \end{aligned} \tag{8}$$

Since $\mathbf{B}\mathbf{1}_N = \mathbf{1}_K$ and $\mathbf{A}\mathbf{1}_K = \mathbf{1}_N$, this simplifies to

$$\begin{aligned} \tilde{\mathbf{X}} - \mathbf{A}\mathbf{B}\tilde{\mathbf{X}} &= \mathbf{X} + \mathbf{1}_N \mathbf{v}^T - \mathbf{A}\mathbf{B}\mathbf{X} - \mathbf{1}_N \mathbf{v}^T \\ &= \mathbf{X} - \mathbf{A}\mathbf{B}\mathbf{X} \end{aligned} \tag{9}$$

Therefore, the reconstruction error remains unchanged, and the minimizers $\mathbf{A}^\star, \mathbf{B}^\star$ are invariant under such translations.

Property 2 (Scale invariance): The minimizers $\mathbf{A}^\star, \mathbf{B}^\star$ of the objective are invariant under global scaling of $\mathbf{X}$. Let $\tilde{\mathbf{X}} = \lambda\mathbf{X}$ for any $\lambda \neq 0$, then

$$\underset{\substack{\mathbf{A}\in F(N,K)\\ \mathbf{B}\in F(K,N)}}{\arg\min} \|\tilde{\mathbf{X}} - \mathbf{A}\mathbf{B}\tilde{\mathbf{X}}\|_F^2 = \underset{\substack{\mathbf{A}\in F(N,K)\\ \mathbf{B}\in F(K,N)}}{\arg\min} \|\mathbf{X} - \mathbf{A}\mathbf{B}\mathbf{X}\|_F^2 \tag{10}$$

Proof: Let $\lambda \neq 0$, and let $\tilde{\mathbf{X}} = \lambda\mathbf{X}$ be the scaled matrix. Then for any feasible $\mathbf{A}, \mathbf{B}$

$$\begin{aligned} \tilde{\mathbf{X}} - \mathbf{A}\mathbf{B}\tilde{\mathbf{X}} &= \lambda\mathbf{X} - \mathbf{A}\mathbf{B}\lambda\mathbf{X} \\ &= \lambda\left(\mathbf{X} - \mathbf{A}\mathbf{B}\mathbf{X}\right) \end{aligned} \tag{11}$$

Thus the objective for the scaled matrix is given by

$$\underset{\substack{\mathbf{A}\in F(N,K)\\ \mathbf{B}\in F(K,N)}}{\arg\min} \|\tilde{\mathbf{X}} - \mathbf{A}\mathbf{B}\tilde{\mathbf{X}}\|_F^2 = \underset{\substack{\mathbf{A}\in F(N,K)\\ \mathbf{B}\in F(K,N)}}{\arg\min} \lambda^2\|\mathbf{X} - \mathbf{A}\mathbf{B}\mathbf{X}\|_F^2 \tag{12}$$

Since we assumed $\lambda \neq 0$, λ^2 will always be a positive scalar. Multiplying the objective by any positive scalar does not change the location of its minimum, since the ordering of function values is preserved. Thus we have

$$\underset{\substack{\mathbf{A}\in F(N,K)\\ \mathbf{B}\in F(K,N)}}{\arg\min} \|\tilde{\mathbf{X}} - \mathbf{A}\mathbf{B}\tilde{\mathbf{X}}\|_F^2 = \underset{\substack{\mathbf{A}\in F(N,K)\\ \mathbf{B}\in F(K,N)}}{\arg\min} \|\mathbf{X} - \mathbf{A}\mathbf{B}\mathbf{X}\|_F^2 \tag{13}$$

Property 3 (Rewrite using convex hull of $\mathbf{Z}$)

Proof,

3 Optimization

While this objective is an NP-hard Euclidean sum of square clustering problem [1], several practical optimization approaches have been developed that exploit that this objective is biconvex, meaning that it is convex in $\mathbf{A}$ if we fix $\mathbf{B}$ and vice versa. See Section 5 in Cutler & Breiman (1994) [3] or Section 2 in Mørup & Hansen (2012) [6]. One way to optimize such a biconvex objective is to initialize $\mathbf{A}$, $\mathbf{B}$, and then alternating between solving the convex optimization problem in one variable fixing the other variable, and vice versa.

3.1 Gradient of the Objective

To compute the gradient of the unconstrained objective w.r.t. $\mathbf{A}$ and $\mathbf{B}$, we first rewrite the residual sum of squares (Frobenius norm) in Equation (4) in terms of the trace

$$\begin{aligned}\mathrm{RSS} &= \|\mathbf{X}-\mathbf{ABX}\|_F^2 \\ &= \mathrm{tr}\left((\mathbf{X}-\mathbf{ABX})^T(\mathbf{X}-\mathbf{ABX})\right) \\ &= \mathrm{tr}(\mathbf{X}^T\mathbf{X}) - \mathrm{tr}(\mathbf{X}^T\mathbf{A}B\mathbf{X}) - \mathrm{tr}(\mathbf{X}^T\mathbf{B}^T\mathbf{A}^T\mathbf{X}) + \mathrm{tr}(\mathbf{X}^T\mathbf{B}^T\mathbf{A}^T\mathbf{ABX}) \\ &= \mathrm{tr}(\mathbf{X}^T\mathbf{X}) - 2\,\mathrm{tr}(\mathbf{X}^T\mathbf{ABX}) + \mathrm{tr}(\mathbf{X}^T\mathbf{B}^T\mathbf{A}^T\mathbf{ABX})\end{aligned} \tag{5}$$

where we used that for any $\mathbf{G},\mathbf{H}\in\mathbb{R}^{N\times N}$ it is true that $\mathrm{tr}(\mathbf{G}+\mathbf{H}) = \mathrm{tr}(\mathbf{G})+\mathrm{tr}(\mathbf{H})$ and $\mathrm{tr}(\mathbf{G}^T) = \mathrm{tr}(\mathbf{G})$

Next we will use Equation 101 from the Matrix Cookbook by Petersen and Pedersen (2012) [7] which states that for any matrices $G, H, J \in \mathbb{R}^{N\times N}$ we have

$$\frac{\partial}{\partial H}\mathrm{tr}(GHJ) = G^T J^T \tag{14}$$

and Equation 116 which states that for any matrices $G, H, J \in \mathbb{R}^{N\times N}$ we have

$$\frac{\partial}{\partial H}\mathrm{tr}(G^T H^T J H G) = J^T H G G^T + J H G G^T \tag{15}$$

So computing the gradient of the RSS w.r.t. A we have

$$\begin{aligned}G^{(A)} &= \nabla_A\,\mathrm{RSS} \\ &= \nabla_A\left[\mathrm{tr}(X^TX) - 2\,\mathrm{tr}(X^TABX) + \mathrm{tr}(X^TB^TA^TABX)\right] \\ &= -2\nabla_A\,\mathrm{tr}(\underbrace{X^T}_{G}\underbrace{A}_{H}\underbrace{BX}_{J}) + \nabla_A\,\mathrm{tr}(\underbrace{(BX)^T}_{G^T}\underbrace{A^T}_{H^T}\underbrace{I}_{J}\underbrace{A}_{H}\underbrace{BX}_{G}) \\ &= -2XX^TB^T + \left(I^TABXX^TB^T + IABXX^TB^T\right) \\ &= -2XX^TB^T + 2ABXX^TB^T \\ &= 2\left(ABXX^TB^T - XX^TB^T\right) \\ &= 2\left(AZZ^T - XZ^T\right)\end{aligned} \tag{16}$$

Similarly, computing the gradient of the RSS w.r.t. B we have

$$\begin{aligned}
G^{(B)} &= \nabla_B \,\mathrm{RSS} \\
&= \nabla_A \left[\mathrm{tr}(X^T X) - 2\,\mathrm{tr}(X^T ABX) + \mathrm{tr}(X^T B^T A^T ABX)\right] \\
&= -2\nabla_B \,\mathrm{tr}(\underbrace{X^T A}_{G}\underbrace{B}_{H}\underbrace{X}_{J}) + \nabla_B \,\mathrm{tr}(\underbrace{X^T}_{G^T}\underbrace{B^T}_{H^T}\underbrace{A^T A}_{J}\underbrace{B}_{H}\underbrace{X}_{G}) \\
&= -2A^T XX^T + \left(A^T ABXX^T + A^T ABXX^T\right) \\
&= -2A^T XX^T + 2A^T ABXX^T \\
&= 2\left(A^T ABXX^T - A^T XX^T\right)
\end{aligned} \tag{17}$$

3.2 Regularized Nonnegative Least Squares

Introduced in 1994 by Adele Cutler and Leo Breiman [3], this was the first algorithm to solve the archetypal analysis objective in Equation (4).

Algorithm 1 Archetypal Analysis Algorithm

Initialize $\mathbf{B}$ and compute the archetypes $\mathbf{Z} = \mathbf{BX}$
while not converged or maximum number of iterations is reached **do**
  for $n = 1$ to N **do**
    Find optimal $\mathbf{a}_n$ by solving the constrained optimization problem:

$$\mathbf{a}_n = \underset{\mathbf{a}_n \in \mathbb{R}^K}{\arg\min} \|\mathbf{x}_n - \mathbf{Z}^T \mathbf{a}_n\|_2^2 \quad \text{subject to} \quad \mathbf{a}_n \geq 0, \sum_{k=1}^{K} a_{nk} = 1$$

  end for
  Compute the optimal archetypes $\mathbf{Z}$ given $\mathbf{A}$, i.e.

$$\mathbf{Z} = \underset{\mathbf{Z} \in \mathbb{R}^{K \times D}}{\arg\min} \|\mathbf{X} - \mathbf{AZ}\|_F^2$$

  for $k = 1$ to K **do**
    Find optimal $\mathbf{b}_k$ by solving the constrained optimization problem:

$$\mathbf{b}_k = \underset{\mathbf{b}_k \in \mathbb{R}^N}{\arg\min} \|\mathbf{z}_k - \mathbf{X}^T \mathbf{b}_k\|_2^2 \quad \text{subject to} \quad \mathbf{b}_k \geq 0, \sum_{n=1}^{N} b_{kn} = 1$$

  end for
  Compute the archetypes given $\mathbf{B}$, i.e. $\mathbf{Z} = \mathbf{BX}$
end while
return $\mathbf{A}, \mathbf{B}, \mathbf{Z}$

The authors originally proposed to solve the constrained optimization problems using a Nonnegative Least Squares Problem (NNLS) solver and enforcing the convexity constraints using a penalty term with regularization parameter λ, i.e.

$$\begin{aligned}
\mathbf{a}_n &= \underset{a_n \in \mathbb{R}^K}{\arg\min} \|x_n - Z^T \mathbf{a}_n\|_2^2 + \lambda \|\mathbf{1}_K - \mathbf{a}_n\|_2^2 \quad \text{subject to} \quad \mathbf{a}_n \geq 0 \\
&= \underset{\mathbf{a}_n \in \mathbb{R}^K}{\arg\min} \left\| \begin{bmatrix} \mathbf{x}_n \\ \lambda \end{bmatrix} - \begin{bmatrix} \mathbf{Z}^T \\ \lambda \mathbf{1}_K^T \end{bmatrix} \mathbf{a}_n \right\|_2^2
\end{aligned} \tag{18}$$

Equivalently, for $\mathbf{B}$ we have

$$\begin{aligned} \mathbf{b}_k &= \underset{b_k \in \mathbb{R}^N}{\arg\min} \|\mathbf{z}_k - \mathbf{X}^T \mathbf{b}_k\|_2^2 + \lambda \|\mathbf{1}_N - \mathbf{b}_k\|_2^2 \quad \text{subject to} \quad \mathbf{b}_k \geq 0 \\ &= \underset{\mathbf{a}_n \in \mathbb{R}^K}{\arg\min} \left\| \begin{bmatrix} \mathbf{z}_k \\ \lambda \end{bmatrix} - \begin{bmatrix} X^T \\ \lambda \mathbf{1}_N^T \end{bmatrix} b_k \right\|_2^2 \end{aligned} \tag{19}$$

3.3 Principal Convex Hull Algorithm (PCHA)

Inspired by the projected gradient method for NMF [5] and normalization invariance approach introduced for NMF [4], the PCHA algorithm was introduced by Morten Mørup and Lars Kai Hansen in 2012 to solve the archetypal analysis objective.

The idea is to use a projected gradient algorithm to solve the objective in Equation (4).

First, we recast the optimization problem in terms of the l1-normalization invariant variables $\tilde{a}_n$ and $\tilde{b}_k$ (called invariant because these variables won't change if one applies l1-normalization)

$$\tilde{a}_{nk} = \frac{a_{nk}}{\sum_{k''=1}^{K} a_{nk''}}, \quad \tilde{b}_{kn} = \frac{b_{kn}}{\sum_{n''=1}^{N} b_{kn''}} \tag{20}$$

Then the gradient of the RSS wrt to a_n is obtained using the chain rule which yields

$$\begin{aligned} \frac{\partial \text{RSS}}{\partial a_n} &= \frac{\partial \text{RSS}}{\partial \tilde{a}_n} \frac{\partial \tilde{a}_n}{\partial a_n} \\ &= \left(\tilde{g}_n^{(A)}\right)^T \left(\frac{\left(\sum_{k''=1}^{K} a_{nk''}\right) \mathbf{I}_K - a_n \mathbf{1}_K^T}{\left(\sum_{k''=1}^{K} a_{nk''}\right)^2} \right) \\ &= \frac{\left(\sum_{k''=1}^{K} a_{nk''}\right) \left(\tilde{g}_n^{(A)}\right)^T \mathbf{I}_K - \left(\tilde{g}_n^{(A)}\right)^T a_n \mathbf{1}_K^T}{\left(\sum_{k''=1}^{K} a_{nk''}\right)^2} \end{aligned} \tag{21}$$

So for a single element we have

$$\begin{aligned} \frac{\partial \text{RSS}}{\partial a_{nk}} &= \frac{\partial \text{RSS}}{\partial \tilde{a}_n} \frac{\partial \tilde{a}_n}{\partial a_{nk}} \\ &= \frac{\left(\sum_{k''=1}^{K} a_{nk''}\right) \tilde{g}_{nk}^{(A)} - \left(\tilde{g}_n^{(A)}\right)^T a_n}{\left(\sum_{k''=1}^{K} a_{nk''}\right)^2} \\ &= \frac{\left(\sum_{k''=1}^{K} a_{nk''}\right) \tilde{g}_{nk}^{(A)} - \sum_{k''=1}^{K} \tilde{g}_{nk''}^{(A)} a_{nk''}}{\left(\sum_{k''=1}^{K} a_{nk''}\right)^2} \end{aligned} \tag{22}$$

IF we additionally assume that a_n has been l1 normalized in the previous iteration we get

$$\frac{\partial \text{RSS}}{\partial a_{nk}} = \tilde{g}_{nk}^{(A)} - \sum_{k''=1}^{K} \tilde{g}_{nk''}^{(A)} a_{nk''} \tag{23}$$

which is exactly the same as in Section 2.2. of Mørup & Hansen (2012) [6]

To write down the algorithm we define P_{Σ_M}, a function that projects the rows of any matrix $\mathbf{H} \in \mathbb{R}^{N\times M}$ onto the M simplex

$$\tilde{\mathbf{H}} = P_{\Sigma_M}(\mathbf{H}) \quad \text{with} \\ \tilde{\mathbf{H}}_{nm} = \frac{\max(\mathbf{H}_{nm}, 0)}{\sum_{m'=1}^{M} \max(\mathbf{H}_{nm'}, 0)} \tag{24}$$

Putting everything together, the algorithm in matrix notation is shown in Algorithm 3

3.4 Frank-Wolfe Algorithm

The idea of the Frank-Wolfe algorithm for archetypal analysis is to use gradient information, but to avoid the costly projection step of the PCHA.

As described above, the objective is convex in $\mathbf{A}$ when fixing $\mathbf{B}$ and vice versa. Furthermore, in this alternating optimization setting, the rows of $\mathbf{A}$ and $\mathbf{B}$ are constrained to the Σ_K and Σ_N simplex, respectively, which are convex sets. Thus, we have a convex minimization problem over a convex set which can be tackled using the efficient Frank-Wolfe algorithm [2]

4 Initialization

4.1 Furthest Sum

5 References

[1] Daniel Aloise et al. "NP-Hardness of Euclidean Sum-of-Squares Clustering". In: *Machine Learning* 75.2 (May 2009), pp. 245–248. ISSN: 0885-6125, 1573-0565. DOI: 10.1007/s10994-009-5103-0. http://link.springer.com/10.1007/s10994-009-5103-0 (visited on 12/07/2024). http://link.springer.com/10.1007/s10994-009-5103-0.

[2] Kenneth L. Clarkson. "Coresets, Sparse Greedy Approximation, and the Frank-Wolfe Algorithm". In: *ACM Trans. Algorithms* 6.4 (Sept. 3, 2010), 63:1–63:30. ISSN: 1549-6325. DOI: 10.1145/1824777.1824783. https://doi.org/10.1145/1824777.1824783 (visited on 02/11/2025). https://doi.org/10.1145/1824777.1824783.

[3] Adele Cutler and Leo Breiman. "Archetypal Analysis". In: *Technometrics : a journal of statistics for the physical, chemical, and engineering sciences* 36.4 (1994), pp. 338–347. ISSN: 0040-1706. DOI: 10.1080/00401706.1994.10485840.

[4] J. Eggert and E. Korner. "Sparse Coding and NMF". In: *2004 IEEE International Joint Conference on Neural Networks (IEEE Cat. No.04CH37541)*. Vol. 4. July 2004, 2529–2533 vol.4. DOI: 10.1109/IJCNN.2004.1381036. https://ieeexplore.ieee.org/document/1381036 (visited on 03/30/2025). https://ieeexplore.ieee.org/document/1381036.

[5] Chih-Jen Lin. "Projected Gradient Methods for Nonnegative Matrix Factorization". In: *Neural Computation* 19.10 (Oct. 2007), pp. 2756–2779. ISSN: 0899-7667. DOI: 10.1162/neco.2007.19.10.2756. https://ieeexplore.ieee.org/document/6795860 (visited on 03/30/2025). https://ieeexplore.ieee.org/document/6795860.

[6] Morten Mørup and Lars Kai Hansen. "Archetypal Analysis for Machine Learning and Data Mining". In: *Neurocomputing* 80 (Mar. 2012), pp. 54–63. ISSN: 09252312. DOI: 10.1016/j.neucom.2011.06.033. https://linkinghub.elsevier.com/retrieve/pii/S0925231211006060 (visited on 12/07/2024). https://linkinghub.elsevier.com/retrieve/pii/S0925231211006060.

Algorithm 2 Principal Convex Hull Algorithm (PCHA)

Initialize $\tilde{\mathbf{A}}$, $\tilde{\mathbf{B}}$
Initialize $\mu_{\mathbf{A}} \leftarrow 1$, $\mu_{\mathbf{B}} \leftarrow 1$
while not converged or maximum number of iterations is reached **do**
  Update $\mathbf{A}$ using projected gradient descent
  $\mathbf{Z} \leftarrow \tilde{\mathbf{B}}\mathbf{X}$
  $\mathrm{RSS}_{\mathrm{old}} \leftarrow \|\mathbf{X} - \mathbf{A}\mathbf{Z}\|_F^2$
  for $t = 1$ to T **do**
    $\tilde{\mathbf{G}}^{(\mathbf{A})} \leftarrow 2\left(\tilde{\mathbf{A}}\mathbf{Z}\mathbf{Z}^T - \mathbf{X}\mathbf{Z}^T\right)$
    $\mathbf{G}^{(\mathbf{A})} \leftarrow \tilde{\mathbf{G}}^{(\mathbf{A})} - \left(\tilde{\mathbf{G}}^{(\mathbf{A})} \odot \mathbf{A}\right)\mathbf{1}_K\mathbf{1}_K^T$
    for $j = 1$ to $100T$ **do** ▷ line search
      $\mathbf{A} \leftarrow \mathbf{A} - \mu_{\mathbf{A}}\mathbf{G}^{(\mathbf{A})}$
      $\tilde{\mathbf{A}} \leftarrow P_{\Sigma_K}(\mathbf{A})$
      $\mathrm{RSS}_{\mathrm{new}} \leftarrow \|\mathbf{X} - \tilde{\mathbf{A}}\mathbf{Z}\|_F^2$
      if $\mathrm{RSS}_{\mathrm{new}} < \mathrm{RSS}_{\mathrm{old}} + (1+\epsilon)$ **then**
        $\mu_{\mathbf{A}} \leftarrow 1.2 \cdot \mu_{\mathbf{A}}$
        break
      else
        $\mu_{\mathbf{A}} \leftarrow 0.5 \cdot \mu_{\mathbf{A}}$
      end if
    end for
  end for
  Update $\mathbf{B}$ using projected gradient descent
  $\mathrm{RSS}_{\mathrm{old}} \leftarrow \|\mathbf{X} - \mathbf{A}\mathbf{B}\mathbf{X}\|_F^2$
  for $t = 1$ to T **do**
    $\tilde{\mathbf{G}}^{(\mathbf{B})} \leftarrow 2\left(\tilde{\mathbf{A}}^T\tilde{\mathbf{A}}\tilde{\mathbf{B}}\mathbf{X}\mathbf{X}^T - \tilde{\mathbf{A}}^T\mathbf{X}\mathbf{X}^T\right)$
    $\mathbf{G}^{(\mathbf{B})} \leftarrow \tilde{\mathbf{G}}^{(\mathbf{B})} - \left(\tilde{\mathbf{G}}^{(\mathbf{B})} \odot \mathbf{B}\right)\mathbf{1}_N\mathbf{1}_N^T$
    for $j = 1$ to $100T$ **do** ▷ line search
      $\mathbf{B} \leftarrow \mathbf{B} - \mu_{\mathbf{B}}\mathbf{G}^{(\mathbf{B})}$
      $\tilde{\mathbf{B}} \leftarrow P_{\Sigma_N}(\mathbf{B})$
      $\mathrm{RSS}_{\mathrm{new}} \leftarrow \|\mathbf{X} - \tilde{\mathbf{A}}\tilde{\mathbf{B}}\mathbf{X}\|_F^2$
      if $\mathrm{RSS}_{\mathrm{new}} < \mathrm{RSS}_{\mathrm{old}} + (1+\epsilon)$ **then**
        $\mu_{\mathbf{B}} \leftarrow 1.2 \cdot \mu_{\mathbf{B}}$
        break
      else
        $\mu_{\mathbf{B}} \leftarrow 0.5 \cdot \mu_{\mathbf{B}}$
      end if
    end for
  end for
  Check for Convergence
  $\mathbf{Z} \leftarrow \tilde{\mathbf{B}}\mathbf{X}$
  $\mathrm{RSS} \leftarrow \|\mathbf{X} - \tilde{\mathbf{A}}\mathbf{Z}\|_F^2$
  if RSS reduction is sufficient **then**
    break
  end if
end while
return $\tilde{\mathbf{A}}, \tilde{\mathbf{B}}, \mathbf{Z}$

[7] Kaare Brandt Petersen and Michael Syskind Pedersen. *The Matrix Cookbook*. Technical University of Denmark, 2012. https://www.math.uwaterloo.ca/~hwolkowi/matrixcookbook.pdf. https://www.math.uwaterloo.ca/~hwolkowi/matrixcookbook.pdf.

6 Appendix

6.1 Notation

- $N \in \mathbb{N}$ is the number of samples
- $D \in \mathbb{N}$ is the number of dimensions
- $K \leq \min(N, D)$ is the number of archetypes
- $\mathcal{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}_{n=1}^{N}$ is our dataset, where each $\mathbf{x}_n \in \mathbb{R}^D$
- $\mathbf{X} \in \mathbb{R}^{N \times D}$ is our data matrix where each row is one sample
- $\mathbf{Z} \in \mathbb{R}^{K \times D}$ is our matrix of archetypes where each row is one archetype

6.2 Algorithms

Algorithm 3 Principal Convex Hull Algorithm (PCHA)

Require: Data matrix $\mathbf{X} \in \mathbb{R}^{N \times D}$, learning rates $\mu_{\mathbf{A}} > 0$, $\mu_{\mathbf{B}} > 0$

Initialize $\mathbf{A}, \mathbf{B}$
$\text{RSS}_{\text{old}} \leftarrow \|\mathbf{X} - \mathbf{ABX}\|_F^2$
while not converged **do**
Update A coefficients:
    $\mathbf{Z} \leftarrow \mathbf{BX}$ ▷ compute archetypes
    $\mathbf{G}^{(\mathbf{A})} \leftarrow 2(\mathbf{AZZ}^T - \mathbf{XZ}^T)$ ▷ gradient of RSS w.r.t. **A**
    $\mathbf{A} \leftarrow \mathbf{A} - \mu_{\mathbf{A}}\mathbf{G}^{(\mathbf{A})}$ ▷ gradient descent step
    $\mathbf{A} \leftarrow P_{\Sigma_K}(\mathbf{A})$ ▷ project rows of **A** onto K-simplex
Update B coefficients:
    $\mathbf{G}^{(\mathbf{B})} \leftarrow 2(\mathbf{A}^T\mathbf{ABXX}^T - \mathbf{A}^T\mathbf{XX}^T)$ ▷ gradient of RSS w.r.t. **B**
    $\mathbf{B} \leftarrow \mathbf{B} - \mu_{\mathbf{B}}\mathbf{G}^{(\mathbf{B})}$ ▷ gradient descent step
    $\mathbf{B} \leftarrow P_{\Sigma_N}(\mathbf{B})$ ▷ project rows of **B** onto N-simplex
Check convergence:
    $\text{RSS}_{\text{new}} \leftarrow \|\mathbf{X} - \mathbf{ABX}\|_F^2$
    $\text{rel\_decrease} \leftarrow \frac{\text{RSS}_{\text{old}} - \text{RSS}_{\text{new}}}{\text{RSS}_{\text{old}}}$ ▷ relative decrease in RSS
    if $\text{rel\_decrease} < \epsilon$ **then**
        break ▷ convergence criterion met
    end if
    $\text{RSS}_{\text{old}} \leftarrow \text{RSS}_{\text{new}}$
end while
return $\mathbf{A}, \mathbf{B}, \mathbf{Z}$