JGR | Machine Learning and Computation

METHOD
10.1029/2024JH000540

Inferring End-Members From Geoscience Data Using Simplex Projected Gradient Descent-Archetypal Analysis

Zanchenling Wang[1] and **Tao Wen**[1]

[1]Department of Earth and Environmental Sciences, Syracuse University, Syracuse, NY, USA

Key Points:

- We introduce simplex projected gradient descent-archetypal analysis (SPGD-AA), an unsupervised machine learning model for inferring end-members from mixed geoscience data
- SPGD-AA is intuitive, interpretable, rigorous, broadly applicable, and outperforms conventional methods on diverse data sets

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:

Z. Wang and T. Wen,
zwang505@syr.edu;
twen08@syr.edu

Citation:

Wang, Z., & Wen, T. (2025). Inferring end-members from geoscience data using simplex projected gradient descent-archetypal analysis. *Journal of Geophysical Research: Machine Learning and Computation*, *2*, e2024JH000540. https://doi.org/10.1029/2024JH000540

Received 30 NOV 2024
Accepted 15 APR 2025

Author Contributions:

Conceptualization: Zanchenling Wang, Tao Wen
Data curation: Zanchenling Wang
Formal analysis: Zanchenling Wang
Funding acquisition: Tao Wen
Investigation: Zanchenling Wang
Methodology: Zanchenling Wang
Project administration: Tao Wen
Software: Zanchenling Wang
Supervision: Tao Wen
Validation: Zanchenling Wang
Visualization: Zanchenling Wang
Writing – original draft: Zanchenling Wang

Abstract End-member mixing analysis (EMMA) is widely used to analyze geoscience data for their end-members and mixing proportions. Many traditional EMMA methods depend on known end-members, which are sometimes uncertain or unknown. Unsupervised EMMA methods infer end-members from data, but many existing ones don't strictly follow necessary constraints and lack full mathematical interpretability. Here, we introduce a novel unsupervised machine learning method, simplex projected gradient descent-archetypal analysis (SPGD-AA), which uses the ML model archetypal analysis to infer end-members intuitively and interpretably without prior knowledge. SPGD-AA uses extreme corners in data as end-members or "archetypes," and represents data as mixtures of end-members. This method is most suitable for linear (conservative) mixing problems when samples with similar characteristics to end-members are present in data. Validation on synthetic and real data sets, including river chemistry, deep-sea sediment elemental composition, and hyperspectral imaging, shows that SPGD-AA effectively recovers end-members consistent with domain expertise and outperforms conventional approaches. SPGD-AA is applicable to a wide range of geoscience data sets and beyond.

Plain Language Summary Earth's materials (e.g., rock, soil, and water) are often mixtures of different sources. We developed a method, simplex projected gradient-archetypal analysis, which allows computers to automatically identify these sources from mixture data by identifying extreme values. We tested our method on artificial data and real-world data sets of river solutes, deep-sea sediments, and airborne images. Our method is easy to use and can be applied to various geoscience data sets and beyond.

1. Introduction

In geoscience, properties of Earth's materials (e.g., rock, soil, and water) are assessed using various field and laboratory techniques such as remote sensing and geochemical measurements. These measurements often capture the combined effects of multiple processes and sources—mathematically conceptualized here as "end-members"—that shape the materials. Quantitatively disentangling mixed data into end-members is often essential for studying each individual process or source. End-member mixing analysis (EMMA) addresses this by identifying and characterizing end-members, which represent distinct and extreme sources or processes, and expressing observational data as their mixing proportions. EMMA has been widely applied across geoscience disciplines. For example, in catchment hydrogeochemistry, EMMA is used to quantify contributions to riverine solutes from end-members like soil water, carbonate weathering, silicate weathering, and atmospheric deposition (e.g., Burns et al., 2001; Christophersen et al., 1990; Gaillardet et al., 1999; Hooper et al., 1990; Shaughnessy et al., 2021). In sedimentology, EMMA deciphers grain-size distributions to reveal sediment sources and transport processes (e.g., E. Dietze et al., 2012; M. Dietze et al., 2022; Prins & Weltje, 1999; Liu et al., 2024; Vandenberghe, 2013). Remote sensing applications, termed hyperspectral unmixing, extract material spectra (e.g., water, trees) and their abundance from multi-band images (e.g., Bioucas-Dias et al., 2012; Boardman et al., 1995; Keshava et al., 2000; Wei & Wang, 2020; Winter, 1999).

A common challenge in classical EMMA is the lack of knowledge of end-member characteristics, which is crucial for quantifying their mixing proportions. Here, we use the term, supervised EMMA, to describe cases where end-members are assigned a priori and only mixing proportions are to be evaluated. End-member characteristics can sometimes be determined from field or laboratory measurements, existing databases, literature, or modeling results. Supervised EMMA is common in geochemistry (e.g., Dymond, 1981; Gaillardet et al., 1999; Hooper et al., 1990; van Geen et al., 1988), and is also applied in sedimentology (e.g., Rea & Hovan, 1995) and remote sensing (e.g., Heinz & Chang, 2001; Khajehrayeni & Ghassemian, 2020; Wyatt & McSween, 2002; Xu

Writing – review & editing: Zanchenling Wang, Tao Wen

et al., 2018; Zhu, 2017). However, it falls short when prior end-member knowledge is incomplete, uncertain, or unavailable.

To overcome these challenges, emerging unsupervised or blind EMMA methods have been proposed to infer end-members (and their mixing proportions) solely from data without prior knowledge. These approaches are gaining traction across geoscience fields such as hydrogeochemistry (e.g., Shaughnessy et al., 2021; Xu Fei & Harman, 2022), sedimentology (e.g., E. Dietze & Dietze, 2019; Renner, 1993; Weltje, 1997; Weltje & Prins, 2007; Zhang et al., 2020), and remote sensing (e.g., Bioucas-Dias et al., 2012; Plaza et al., 2004; Wei & Wang, 2020). Among them, non-negative matrix factorization (NMF) (Lee & Seung, 1999) and its variants stand out as popular tools, particularly in hyperspectral unmixing (e.g., Hoyer, 2002; Iordache et al., 2011; Jia & Qian, 2009; Lu et al., 2013, 2014; Miao & Qi, 2007; Qian et al., 2011; Zhou et al., 2020; Zhu et al., 2014; Zhuang et al., 2019), and more recently in hydrogeochemistry (Epuna et al., 2022; Shaheen et al., 2022; Shaughnessy et al., 2021; Xu Fei & Harman, 2022). However, challenges remain with NMF-based EMMA tools: many are not self-contained, requiring other algorithms for data pre-processing, parameter tuning, and post-processing. Furthermore, they often fail to enforce sufficient optimization under strict unit simplex constraints, meaning the mixing proportions for derived end-members must sum to unity. Instead, they use penalty terms or manual transform to loosely satisfy such constraints. These issues hamper the accuracy and interpretability of these NMF methods.

Here, we propose an intuitive and interpretable unsupervised EMMA method, simplex projected gradient descent–archetypal analysis (SPGD-AA), based upon the unsupervised machine learning model, archetypal analysis (AA) (Cutler & Breiman, 1994). AA represents each observational data point as a mixture of a set of extreme points, referred to as "pure types" or "archetypes." AA and its variants (e.g., Abrol & Sharma, 2020; Alcacer et al., 2024; Dijk et al., 2019; Javadi & Montanari, 2020; Keller et al., 2019; Mørup & Hansen, 2012; Seth & Eugster, 2016) have been applied to a wide range of tasks, including galaxy spectra classification (Chan et al., 2003), biological task inference (Hart et al., 2015), document summarization (Canhasi & Kononenko, 2014), extreme climate pattern identification (Steinschneider & Lall, 2015) and hyperspectral unmixing (Zouaoui et al., 2023). However, AA remains largely underutilized in geosciences. Since AA's inception, many efforts have been made to improve its performance and robustness (Abrol & Sharma, 2020; Bauckhage & Thurau, 2009; Chen et al., 2014; Damle & Sun, 2017; Mørup & Hansen, 2012). However, many of them aim for faster but approximate solutions, often introducing complex optimization methods that make AA harder to implement and understand. Moreover, well-maintained, open-source and high-performance implementations of AA remain lacking, limiting its broader applications. Our SPGD-AA method addresses these issues by employing fast unit simplex projection (Condat, 2016) along with projected gradient descent (PGD) (Wright & Recht, 2022) in the optimization procedure of AA. The algorithm is hosted in the Python package `archetypes` (Alcacer & Wang, 2025), with performance-critical components written in Cython (Behnel et al., 2011) to maintain computational efficiency while ensuring convergence to an exact local minimum under strict unit simplex constraints. We demonstrate SPGD-AA's effectiveness in inferring end-members using synthetic data and real-world data sets from three geoscience domains: Panola Mountain stream chemistry (Hooper & Christophersen, 1992; Hooper et al., 1990), Nazca Plate sediment elemental composition (Dymond, 1981; Pisias et al., 2013), and Jasper Ridge hyperspectral image (Zhu, 2017). These data sets, extensively studied for end-member characteristics, enable us to validate SPGD-AA's ability to recover end-members derived from domain knowledge and compare its performance to other state-of-the-art unsupervised EMMA methods.

2. Data and Methods

2.1. General Mathematical Form of End-Member Mixing Analysis

Let $\mathbf{X}$ denote an $m \times n$ data set matrix $\begin{bmatrix}\boldsymbol{x}_1 & \boldsymbol{x}_2 & \dots & \boldsymbol{x}_n\end{bmatrix}$, each column ($\boldsymbol{x}_i$) representing an observational data point with m features. The values in $\mathbf{X}$ may indicate observed properties (e.g., chemical composition).

In EMMA, we consider p end-members that constitute the sources of the mixed observations, each described by a m-dimensional column vector $\boldsymbol{e}_j$ ($j = 1, 2, \dots, p$). EMMA approximates each observation $\boldsymbol{x}_i$ as a mixture of end-members $\hat{\boldsymbol{x}}_i$:

$$\boldsymbol{x}_i \approx \hat{\boldsymbol{x}}_i = \sum_{j=1}^{p} \boldsymbol{e}_j a_{ji} = \mathbf{E}\boldsymbol{a}_i, \quad a_{ji} \geq 0 \text{ and } \sum_{j=1}^{p} a_{ji} = 1, \quad i = 1, 2, \dots, n, \tag{1}$$

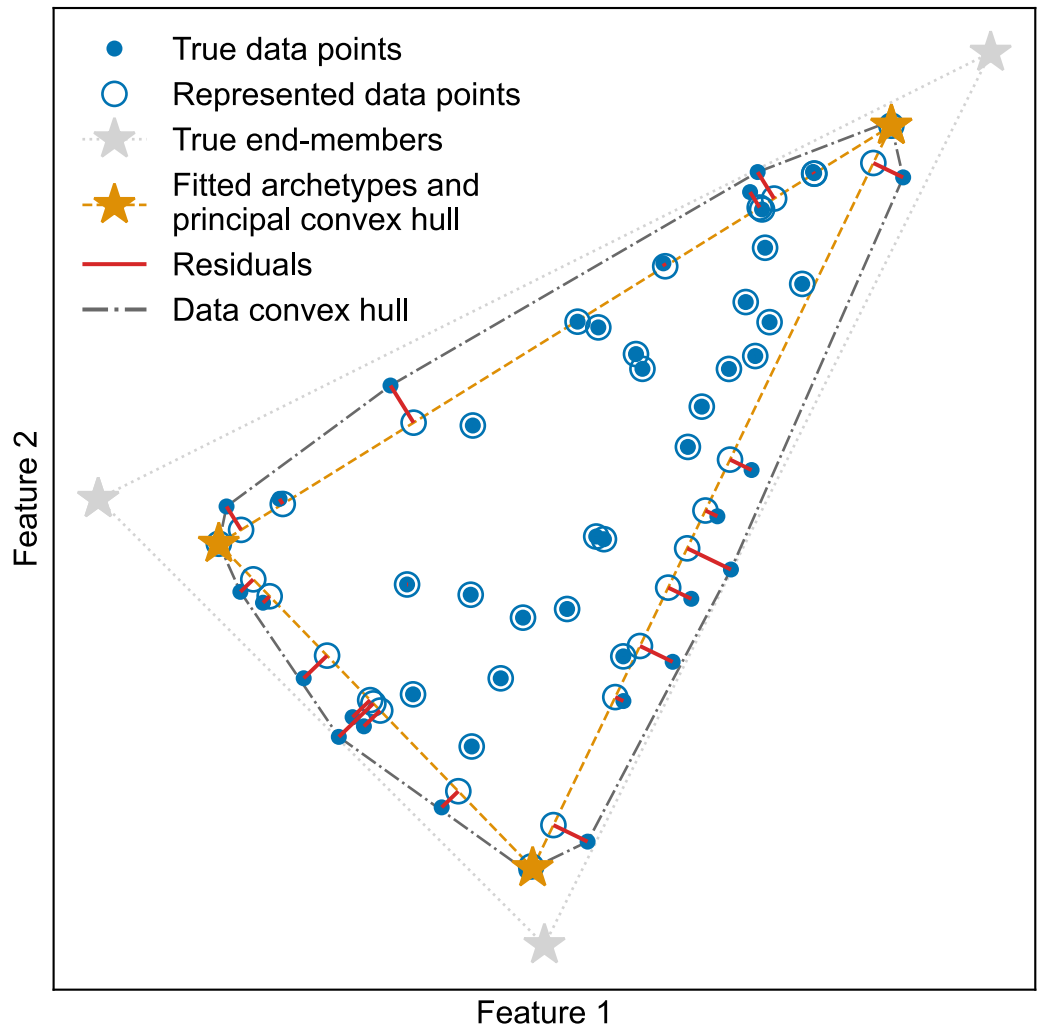

Figure 1. A schematic diagram of end-member mixing analysis and archetypal analysis (AA) (2-dimensional observational data and 3 archetypes). Observations (blue dots), as strict mixtures of end-members (gray stars), lie inside the convex hull of end-members (gray triangle). AA searches for a set of archetypes (orange stars) within the data convex hull (dash-dotted gray polygon), to approximate observations as mixtures of these archetypes (blue circles). The objective is to minimize the residual sum of squared distances (red lines) between true observations and their approximations. The best approximations are projections of true observations onto the archetypal convex hull (orange triangle).

where a_{ji} is the mixing proportion of the j-th end-member in the i-th observation, $\boldsymbol{a}_i = \begin{bmatrix} a_{1i} & a_{2i} & \dots & a_{pi} \end{bmatrix}^\top$, and $\mathbf{E}$ denotes the $m \times p$ end-member characteristic matrix $\begin{bmatrix} \boldsymbol{e}_1 & \boldsymbol{e}_2 & \dots & \boldsymbol{e}_p \end{bmatrix}$. Mixing proportions must be non-negative and sum to 1, that is, $\boldsymbol{a}_i$ is constrained within the $p-1$-dimensional unit simplex Δ^{p-1}. With this constraint, the approximation $\hat{\boldsymbol{x}}_i$ is a convex combination of end-members $\{\boldsymbol{e}_1, \boldsymbol{e}_2, \dots, \boldsymbol{e}_p\}$.

Let $\mathbf{A}$ denote the $p \times n$ mixing proportion matrix $\begin{bmatrix} \boldsymbol{a}_1 & \boldsymbol{a}_2 & \dots & \boldsymbol{a}_n \end{bmatrix}$. Equation 1 can be rewritten in a more compact form:

$$\mathbf{X} \approx \mathbf{EA}. \tag{2}$$

Any convex combinations of end-members must lie within their convex hull (the smallest convex polygon or polytope enclosing a point set, e.g., the gray triangle in Figure 1). A mixing model is valid only when most observations $\mathbf{X}$ lie within or near the convex hull of end-members. In supervised EMMA, the end-member matrix $\mathbf{E}$ is known or assumed, and mixing proportions $\mathbf{A}$ are typically estimated using ordinary or constrained least

squares, for example, non-negative least squares (NNLS) (Bro & De Jong, 1997; Lawson & Hanson, 1995) and fully constrained least squares (Heinz & Chang, 2001).

2.2. Archetypal Analysis

In unsupervised EMMA, the only information available is that mixture data lie inside or at least close to the convex hull of end-members, which is to be reconstructed. When all data points are far from pure end-members, the convex hull of data most likely does not resemble the end-member convex hull. This is to say, information is insufficient to determine the true end-members solely from data without prior knowledge. When the requirements of unsupervised EMMA are satisfied, AA provides an interpretable and reasonable solution of finding end-members by searching for a set of p "archetypes" $\{\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_p\}$ that represent extreme patterns outlines the principal convex hull of the data set (Mørup & Hansen, 2012), and representing data points as mixtures of archetypes (Figure 1). In this way, AA aligns naturally with the goal of inferring end-members from data. Classic AA restricts archetypes to the data convex hull, which means archetypes are either observations themselves or mixtures of observations:

$$\boldsymbol{e}_j = \sum_{i=1}^{n} \boldsymbol{x}_i b_{ij} = \mathbf{X}\boldsymbol{b}_j, \quad b_{ij} \geq 0 \text{ and } \sum_{i=1}^{n} b_{ij} = 1, \quad j = 1, 2, \ldots, p, \tag{3}$$

where $\boldsymbol{b}_j = \begin{bmatrix} b_{1j} & b_{2j} & \ldots & b_{nj} \end{bmatrix}^\top \in \Delta^{n-1}$. Define $n \times p$ matrix $\mathbf{B} = \begin{bmatrix} \boldsymbol{b}_1 & \boldsymbol{b}_2 & \ldots & \boldsymbol{b}_p \end{bmatrix}$ then we get

$$\mathbf{E} = \mathbf{XB}. \tag{4}$$

Combining Equations 2 and 4, we have

$$\mathbf{X} \approx \hat{\mathbf{X}} = \mathbf{XBA}, \tag{5}$$

where $\hat{\mathbf{X}}$ is the reconstructed approximations of observations.

AA simultaneously searches for optimal archetypes and mixing proportions that minimize the reconstruction error, defined as the residual sum of squares (RSS) between the original observations and approximations:

$$\text{RSS} = \|\mathbf{X} - \mathbf{XBA}\|_F^2, \tag{6}$$

subject to column-wise unit simplex constraints on $\mathbf{B}$ and $\mathbf{A}$, where $\|\cdot\|_F$ denotes the Frobenius norm. For any set of archetypes, RSS is minimized when all approximations are projections of observations onto the archetypal convex hull. Thus, AA minimizes the sum of squared distances from outliers to the archetypal convex hull (dashed gray lines in Figure 1). When $p = 1$, the optimal archetype is simply the center of mass of the data (Cutler & Breiman, 1994). When $p > 1$, AA becomes a constrained non-convex optimization problem, typically solved by alternatingly updates of $\mathbf{B}$ and $\mathbf{A}$, starting from initial guesses $\mathbf{B}_0$ and $\mathbf{A}_0$. The two optimization subproblems are often solved with NNLS (Bauckhage & Thurau, 2009; Cutler & Breiman, 1994; Damle & Sun, 2017; Eugster & Leisch, 2009, 2011) or other approximate algorithms (Abrol & Sharma, 2020; Bauckhage et al., 2015; Mørup & Hansen, 2012), which usually don't strictly enforce unit simplex constraints. Our focus is to provide an approach that converges to an exact local minimum of RSS.

2.3. Simplex Projected Gradient Descent Method

Though some other precise optimization methods for AA have been proposed (Chen et al., 2014; Zouaoui et al., 2023), a generalizable and intuitive approach is PGD (Wright & Recht, 2022). PGD iteratively updates the solution by moving against the direction of the gradient of the target function, and projects the new solution onto the feasible set (the unit simplex, in AA's case) after each update:

$$\boldsymbol{x}_{i+1} = \mathcal{P}(\boldsymbol{x}_i - \alpha \nabla f(\boldsymbol{x}_i)), \quad i = 1, 2, \ldots \tag{7}$$

where $f(\boldsymbol{x})$ is the target function (RSS) to be minimized, $\mathcal{P}$ denotes the projection operator onto the feasible set, α is the step size, and $\nabla f(\boldsymbol{x})$ is the gradient of f at $\boldsymbol{x}$. A pseudo-PGD method for AA is first proposed by Mørup and Hansen (2012), using ℓ_1 normalization instead of unit simplex projection and modifying gradients accordingly. This method works in practice but lacks theoretical guarantees. Our SPGD-AA adopts a fast unit simplex projection algorithm (Condat, 2016) for a standard alternating PGD that converges to a local RSS minimum, and is expected to converge to a global minimum with a proper initialization. We initialize **A** as a zero matrix with one randomly placed one per column, and **B** using a Furthest-sum method (Mørup & Hansen, 2012). Step sizes are dynamically adjusted during optimization, also following Mørup and Hansen (2012). Further details are provided in Text S1 in Supporting Information S1.

2.4. Data Sets

2.4.1. Synthetic Data Sets

We generated two synthetic data sets based on four hypothesized end-members (Table S1 in Supporting Information S1) with and without added noise, each with 1,000 samples. Mixing proportions were sampled from a flat Dirichlet distribution (a Dirichlet distribution with all $\alpha = 1$, the "uniform" distribution on the unit simplex) (Bishop, 2016), ensuring equal likelihood for all feasible proportions. In the noise-free data set, all data points are strict mixtures of end-members. The noisy data set mimics real-world scenarios where end-members fluctuate, and observations are subject to uncertainties due to undetected sources, nonlinear mixing, or measurement errors. To generate this data set, we added Gaussian noise (standard deviation $\sigma = 0.1$) to each feature of each end-member before mixing, followed by additional Gaussian noise ($\sigma = 0.1$) applied to each feature of resulting mixtures. A few negative values in the noisy data set were replaced with zeros to maintain compatibility with compared unsupervised EMMA methods like NMF. Additionally, we generated two noise-free data sets with the same end-members but larger values (2 and 4) for all α in the Dirichlet distribution. Increasing α alters the uniformly distributed characteristics, suppresses the likelihood of extreme mixing proportions, thereby reducing the probability of sampling near end-members. When the sample size is limited, larger α values make it more likely that no samples will be drawn close to the end-members.

2.4.2. Panola Mountain Stream Chemistry

Panola Mountain Research Watershed (PMRW), located in Georgia, USA, is a long-term research site for small catchment biogeochemistry. This 41-ha catchment is underlain by granodiorite, covered with old, highly weathered soil, entirely forested, and situated in a warm temperate subtropical climate (Hooper, 2001; Hooper & Christophersen, 1992). The data set includes the concentration of six solutes (alkalinity, sulfate, sodium, magnesium, calcium and dissolved silica) in 905 stream water samples collected in PMRW from 1 October 1985 to 30 September 1988 (Hooper, 2001; Hooper et al., 1990). Hooper et al. (1990) suggested that the streamwater chemistry of these solutes can be explained as conservative (non-reactive) mixtures of three soil water end-members: groundwater (dominant during dry summer months), hillslope (dominant during wet winter and storms) and organic horizon (influencing during high-flow conditions). These end-members were characterized from soil solutions collected at PMRW, showing distinct and stable chemical compositions across space and time (Christophersen et al., 1990; Hooper et al., 1990). This mixing model is further investigated and validated in later studies (e.g., Christophersen & Hooper, 1992; Hooper & Christophersen, 1992; Xu Fei & Harman, 2022).

2.4.3. Nazca Plate Sediments

The Nazca Plate sediment data set includes the abundance of eight elements (Al, Si, Fe, Mn, Cu, Ni, Zn, and Ba) in 327 deep-sea surface sediment samples, reported on a carbonate-free basis (Dymond, 1981; Pisias et al., 2013). These samples were collected across Nazca Plate, an oceanic tectonic plate in the southeastern Pacific Ocean basin off the west coast of South America. Dymond (1981) used linear programming on the data set to quantify the contributions from five end-members: hydrothermal, detrital, biogenic, authigenic (hydrogenous) sediments, and dissolution residue. Elemental ratios of the hydrothermal end-member were derived from samples near the crest of East Pacific Rise where hydrothermal sediments dominate. The dissolution residue end-member was based on domain knowledge, while others were determined from literature. The model succeeded in explaining sources of major elements in sediments, though some discrepancies in modeled versus observed values for Mn,

Ni, Cu, and Zn suggested that these minor elements may deviate from the constant-value assumption of end-member modeling (Dymond, 1981).

2.4.4. Jasper Ridge Hyperspectral Image

Jasper Ridge data set is a 100 × 100 pixel hyperspectral image of the Jasper Ridge biological preserve, California, USA, captured by the AVIRIS sensor (Zhu, 2017). Each pixel originally records reflectance spectra at 224 bands (380–2,500 nm). After removing 26 bands affected by water vapor and atmospheric interference, 198 bands remain. Four end-members—road, soil, water, and tree—were identified by selecting "pure" pixels from the image that closely match reference spectra from spectral libraries (Zhu, 2017; Zhu et al., 2014). This data set is widely used as a benchmark in hyperspectral unmixing studies (e.g., Li & Tan, 2024; Ozkan et al., 2019; Xiong et al., 2022; Xu et al., 2020; Zhang et al., 2018; Zhou et al., 2020; Zhu et al., 2014).

2.5. Experimental Setup

For a just comparison with other unsupervised EMMA methods in previous studies, we rescaled each feature (solute) in the Panola data set to unit variance (and rescale the results back) following Xu Fei and Harman (2022), and converted the Nazca data set's elemental abundances to weight fractions relative to the total of all eight elements, following Leinen and Pisias (1984). The only required parameter for SPGD-AA is the number of archetypes (end-members) p, which is also adopted from previous studies: 3 for the Panola data set, 4 for the Jasper data set, and 5 for the Nazca data set. In all our experiments, to balance accuracy and runtime, we set additional parameters for the number of AA instances and step size adjustment (see Text S1 in Supporting Information S1). For synthetic data sets, we also ran three other unsupervised EMMA algorithms, NMF by Shaughnessy et al. (2021), convex hull end-member mixing analysis (CHEMMA) (Xu Fei & Harman, 2022), and entropic descent archetypal analysis (EDAA) (Zouaoui et al., 2023) for comparison with SPGD-AA.

3. Results and Discussion

3.1. Synthetic Data Sets

For both noise-free and noisy data sets with mixing proportions drawn from a flat Dirichlet distribution, SPGD-AA end-members closely approximate the true ones (Figure 2). This shows that, by restricting the end-members within the data convex hull, SPGD-AA gains some robustness against noise. Modeled mixing proportions also align well with true ones, with a R^2 of 0.995 for the noise-free data set and 0.851 for the noisy data set (Figure S1 and Table S3 in Supporting Information S1). Adding noise to synthetic samples alters their characteristics, impacting the accurate estimation of mixing proportions. In contrast, NMF, whose results tend to be sparse and extend beyond the data convex hull, failed to precisely recover end-members, especially in the presence of noise. EDAA was able to partially recover some features of end-members with less precision, yet the inferred end-members do not coincide with the corners of data point cloud, possibly due to improper default parameters or model selection procedure. CHEMMA produced results comparable to SPGD-AA, but all SPGD-AA end-members in both data sets have smaller Euclidean distances to the true ones, that is, higher accuracy (Table S2 in Supporting Information S1). CHEMMA's lower precision likely arises from its reliance on 2D projection to subsample the data convex hull (Thurau et al., 2011), so that only a subset of data convex hull is used in determining end-members with most internal points ignored. In SPGD-AA, all original data points are taken into account without subsampling the data convex hull, better preserving the integrity of data. Both CHEMMA and SPGD-AA require multiple runs with different random initializations, but they differ in how the outputs are handled. CHEMMA aggregates these outputs by constrained clustering to estimate the distribution and mean/median position of end-members, which could serve as a prior distribution in Bayesian-based methods. In SPGD-AA, we selected only the result with the smallest RSS as the final point estimate of end-members, disregarding all other suboptimal local minima to increase accuracy and maintain interpretability. This practice also enables SPGD-AA to infer end-members and mixing proportions simultaneously, whereas in CHEMMA, mixing proportions must be evaluated under looser constraints after clustering. Finally, CHEMMA's dependent packages are outdated, with some of them incompatible with Python 3. SPGD-AA is more straightforward and easier to access.

For additional data sets with higher concentration parameters α in the Dirichlet distribution, most unsupervised EMMA becomes less or even not feasible as data are more concentrated to the center of the end-member polytope, and SPGD-AA results move further away from true end-members to the center of the data cloud, as shown in

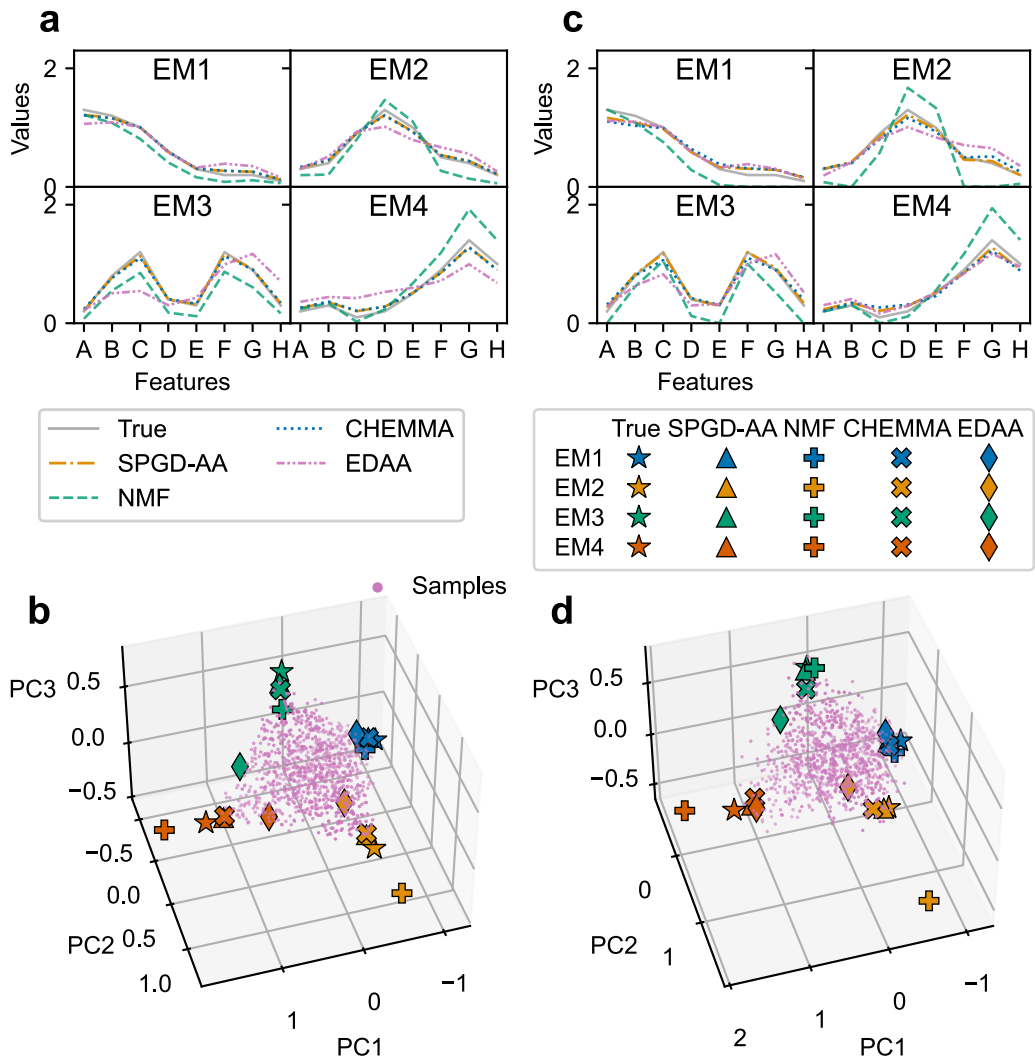

Figure 2. Results of simplex projected gradient descent-archetypal analysis (SPGD-AA) on synthetic data sets, compared with non-negative matrix factorization (NMF) (Shaughnessy et al., 2021), convex hull end-member mixing analysis (CHEMMA) (Xu Fei & Harman, 2022) and EDAA (Zouaoui et al., 2023). (a) True end-members of noise-free synthetic data sets, compared with those inferred by SPGD-AA, NMF, CHEMMA, and EDAA. (b) Noise-free synthetic data, with true and inferred end-members, visualized in the 3D principal component subspace of the data set. Panels (c) and (d) are similar to panels (a) and (b), but for the noisy synthetic data set. Detailed end-member values are provided in Table S2 in Supporting Information S1.

Figure S2 in Supporting Information S1. Though NMF seems to outperform other methods sometimes, it actually relies on selecting results and ranking end-members based on prior knowledge or assumptions (see the footnote of Table S2 in Supporting Information S1), forcing the results to exhibit certain properties. Combined with NMF's lack of a convex hull constraint and its inherent sparsity, these factors may occasionally lead to better results. However, this outperformance is not consistent or generalizable. As shown in Figure S3 in Supporting Information S1, When all features of each end-member are uniformly shifted by $+10$, SPGD-AA remains stable and produces consistent results, whereas NMF fails and returns empty outputs.

3.2. Panola Mountain Stream Chemistry

The inferred end-members share similar characteristics with the soil solutions from field sampling Hooper et al. (1990) (Figures 3a and 3d), indicating that SPGD-AA successfully recovered the end-members determined from field studies. Our results also support the conclusion of Hooper et al. (1990) that PMRW streamwater chemistry can be modeled as mixtures of soil water end-members. The recovered organic end-member somewhat

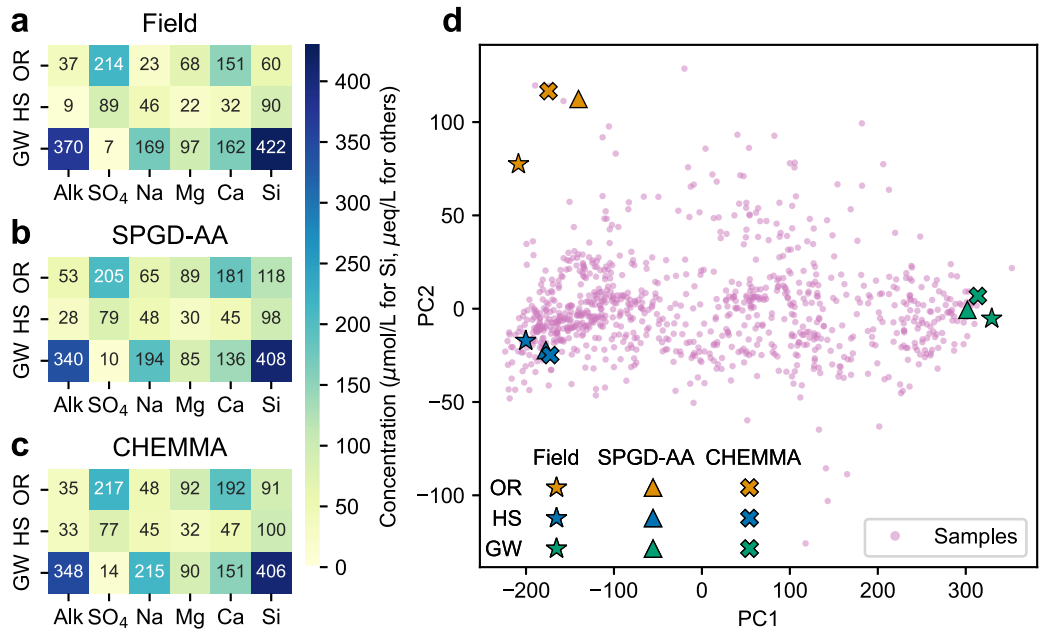

Figure 3. Results of simplex projected gradient descent-archetypal analysis (SPGD-AA) on Panola Mountain Stream Chemistry data set, compared with field samples and the convex hull end-member mixing analysis (CHEMMA) results. (a) Median solute concentrations of three soil water end-members determined from field sampling (Hooper et al., 1990): groundwater (GW), hillslope (HS) and organic horizon (OR). (b) End-members inferred from SPGD-AA, rearranged and renamed to match (a). (c) Mean end-members inferred from CHEMMA (Xu Fei & Harman, 2022). (d) Streamwater observations and end-members from panels (a) to (c) shown in the 2D principal component subspace of data set. Details of end-members are available in Table S4 in Supporting Information S1.

deviates from field values but better captures the boundary condition of streamwater observations in the 2D principal component (PC) subspace of the data set (Figure 3d). The field organic end-member is insufficient to explain a few extreme samples, which may be due to some undescribed end-members, non-conservative mixing or dilution/concentration. Compared to CHEMMA-derived end-members (Xu Fei & Harman, 2022) (Figures 3c and 3d), SPGD-AA produces similar results but in a more concise and interpretable way as discussed. We evaluated the goodness of fit of SPGD-AA, CHEMMA and field end-members by computing their relative reconstruction error RSS/RSS(1), where RSS is the RSS for a given set of end-members, and RSS(1) is the RSS when using only the data centroid as the sole end-member (Cutler & Breiman, 1994). The mixing proportions of CHEMMA and field end-members were estimated by solving simplex constrained least squares using simplex projected gradient descent (SPGD). The results are 0.074 for SPGD-AA, 0.087 for CHEMMA and 0.108 for field end-members. All values are much smaller than 1 and close to 0, meaning that the overall projection distances between outliers and the end-member polytope are relative small, so that most of the variance in the data can be effectively explained by the mixing of either set of end-members. The smallest reconstruction error of SPGD-AA indicates that it best captures the mixing space. CHEMMA end-members has a smaller reconstruction error than field end-members, which agrees with the result of Xu Fei and Harman (2022).

3.3. Nazca Plate Sediments

At least four of the five end-members inferred by SPGD-AA closely resemble those identified by Dymond (1981), particularly in terms of major elemental composition (Figure 4). Minor elements, with low values and potential violations of EMMA assumptions, have little impact on SPGD-AA results and were not further investigated. The dissolution residue end-member determined by Dymond (1981) shows more discrepancy with our results, as no sample is close to its pure form (refer to the 3.5 section for details). However, SPGD-AA successfully identified an archetype near this end-member within the data convex hull (yellow triangle in Figure 4d). Leinen and Pisias (1984) used *Q*-mode factor analysis (QFA) to infer end-members from the same data set. Compared to QFA results (Figures 4c and 4d), the SPGD-AA end-members are generally closer the ones of Dymond (1981) (except for the dissolution residue) in the 3D PC subspace of the data set, and better capture the

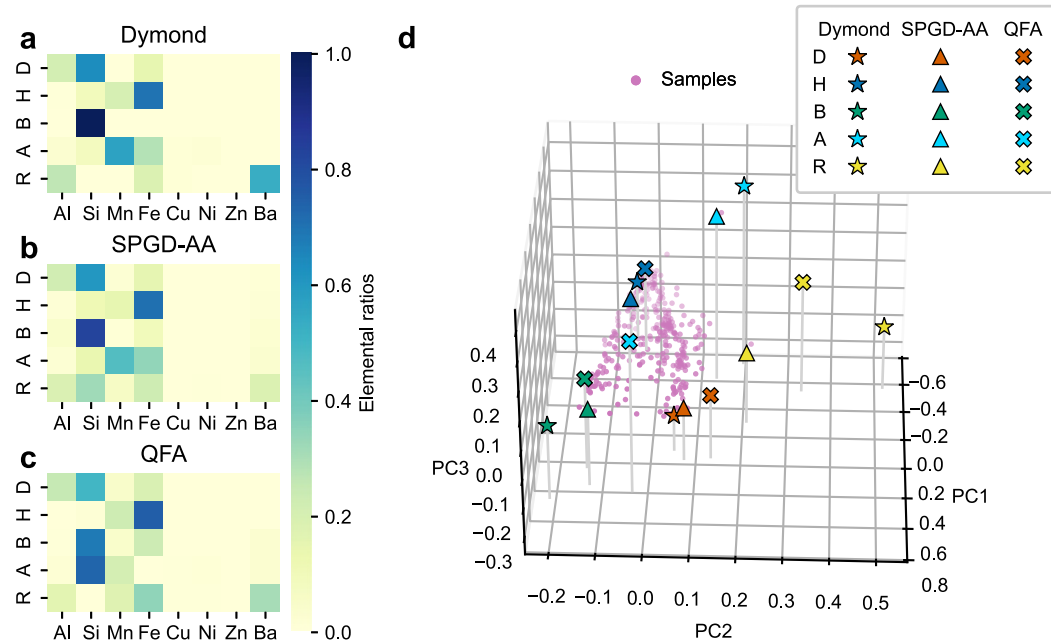

Figure 4. Results of simplex projected gradient descent-archetypal analysis (SPGD-AA) on Nazca Plate sediment data set. Panels (a–c) show the elemental fractions of end-members (detrital or D, hydrothermal or H, biogenic or B, authigenic or A, dissolution residue or R) determined by Dymond (1981), inferred by SPGD-AA (rearranged to match (a)) and identified by *Q*-mode factor analysis (Leinen & Pisias, 1984), respectively. Panel (d) visualizes sediment samples and end-members from panels (a) to (c) in the 3D principal component subspace of the elemental fraction data. Additional angles of view of panel (d) are provided in Figure S4 in Supporting Information S1. Detailed end-member characteristics are in Table S5 in Supporting Information S1.

data set's boundary conditions. QFA produced an unrealistic estimation of the authigenic end-member (Leinen & Pisias, 1984), which is far from both SPGD-AA and Dymond (1981) results. Additionally, QFA requires each sample to sum to a constant (e.g., values representing weight percentages) (Miesch, 1976), limiting is broader applications.

3.4. Jasper Ridge Hyperspectral Image

SPGD-AA successfully recovered all four end-members in the Jasper Ridge image (Figure 5a). The inferred end-members are closer to the expert-determined ground truth (GT) than those derived from classic NMF methods like ℓ_1-NMF (Iordache et al., 2011) and $\ell_{1/2}$-NMF (Qian et al., 2011), and are comparable to recent neural network approaches such as SNMF-Net (Xiong et al., 2022), in terms of spectral angle distance (SAD, angle between GT and corresponding inferred end-member vectors) metrics (Figure 5b). Zouaoui et al. (2023) used AA with active-set (Chen et al., 2014) and entropic descent (Beck & Teboulle, 2003) optimizers to unmix the Jasper Ridge image. Their results suggest that AA outperform NMF-based and neural network methods, which failed to precisely identify the road end-member. Interestingly, other studies reported better performance for these competitive methods on the same data set (e.g., Ozkan et al., 2019; Qi et al., 2023). This discrepancy can be explained by the uncommon ℓ_2-normalization applied by Zouaoui et al. (2023). This preprocessing step, which maps observations onto the unit hypersphere, alters the data set's geometry, and potentially violates the EMMA assumption that data lie within a convex polytope of end-members. SPGD-AA results show that normalization is not needed for this task.

3.5. Limitations

It is worth noting that SPGD-AA has several limitations. First, as stated above, the feasibility and accuracy of SPGD-AA and other unsupervised EMMA methods depend on the proximity of data convex hull to the polytope of end-members. When there are data points with higher purity in terms of mixing proportions, SPGD-AA generally gets better results. Though there are relaxed variants of AA (e.g., Javadi & Montanari, 2020; Mørup

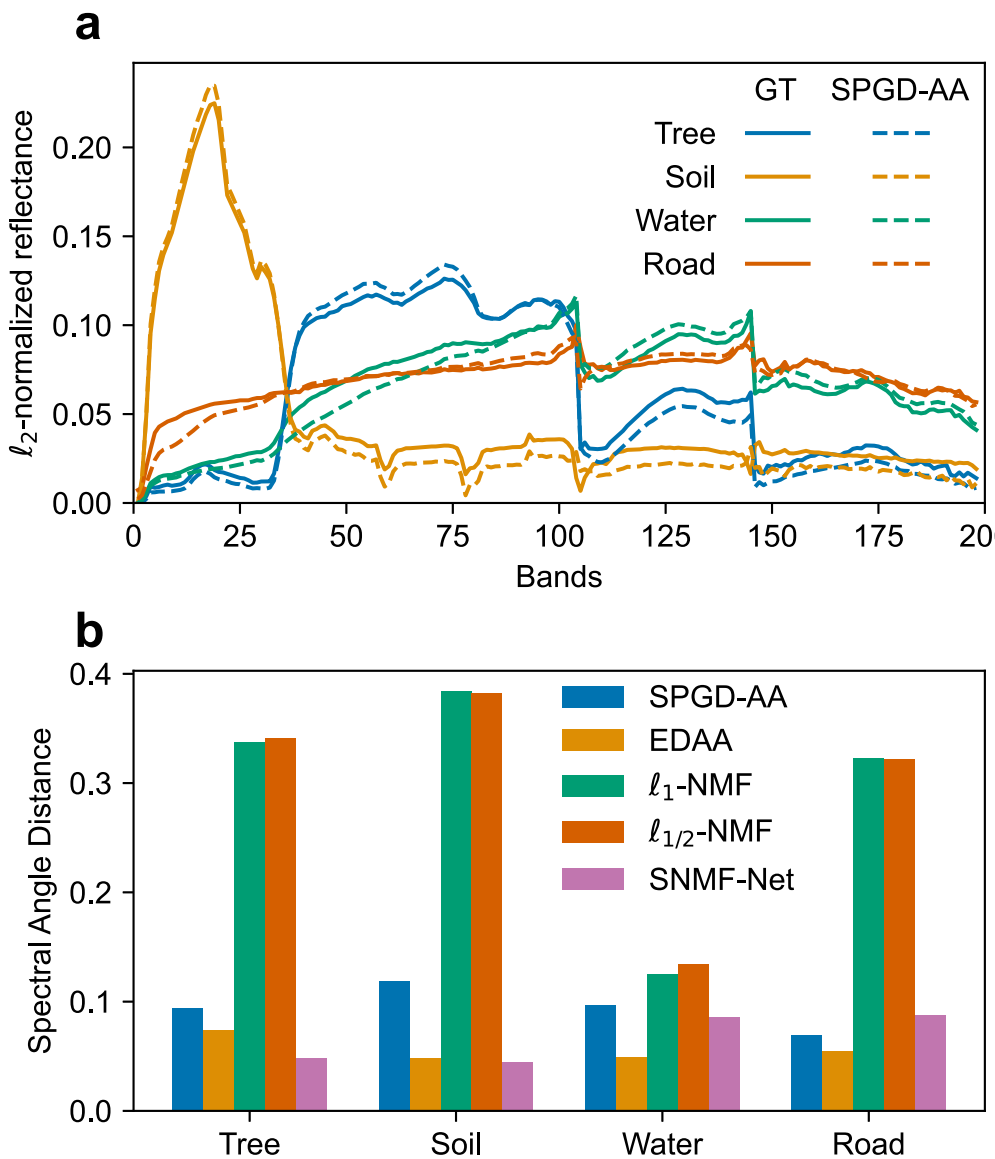

Figure 5. Results of simplex projected gradient descent-archetypal analysis (SPGD-AA) on the Jasper Ridge hyperspectral image. (a) ℓ_1-normalized reflectance spectra of expert-determined ground truth (GT, solid lines) and SPGD-AA end-members (dashed lines): tree, soil, water and road. (b) Spectral angle distances (SADs, in radians) between GT and end-members inferred by SPGD-AA and other methods. Comparative results are compiled from Zouaoui et al. (2023) (entropic descent archetypal analysis with ℓ_2-normalization), Xu et al. (2020) (ℓ_1-NMF and $\ell_{1/2}$-NMF), and Xiong et al. (2022) (SNMF-Net). SPGD-AA end-member values are stored in Table S6.

& Hansen, 2012) and other EMMA variants (e.g., Zhang et al., 2020) that allows extrapolating archetypes or end-members outside the data convex hull, this may yield unrealistic results. Besides, these methods have to introduce additional tunable parameters that need to be determined with prior knowledge, making them less interpretable and less generalizable.

The second is related to the determination of the number of end-members p. In our experiments, p is assumed to be known, and determining p automatically from data remains a challenge. In AA, increasing p reduces RSS(p)/RSS(1) and improves goodness of fit, but may also introduce excessive and uninterpretable end-members. In the absence of prior information, a common heuristic is to run AA with different p values, and manually select the "elbow" point where the RSS(p)/RSS(1) curve starts to flatten (e.g., Cutler & Breiman, 1994; Epifanio et al., 2013; Seth & Eugster, 2016).

We find, however, some real geoscience data sets show no discernible elbows in the curve that match our domain knowledge, as these data sets can be noisy and highly imbalanced, with data primarily clustered around certain end-members while others are underrepresented. For example, consider a system with three end-members where most observations lie along the segment between two end-members, and only a few are close to the third. Increasing p from 2 to 3 may not significantly reduce RSS(p)/RSS(1), as most observations can already be effectively explained with two end-members. Whether to include the third end-member or treat it as noise or an outlier depends on domain knowledge. Therefore, it is recommended to use the elbow criterion only as a reference, run AA with different values of p and decide on an appropriate number based on consistency with domain expertise and interpretability.

4. Conclusions

Experimental results on synthetic data and real-world geoscience data sets demonstrate SPGD-AA's potential as an interpretable and widely applicable method for unsupervised EMMA. When applying SPGD-AA, users must be mindful of the assumptions underlying EMMA: end-members have distinct characteristics whose variability is much smaller than that of the mixed observations, and the mixing should be linear and conservative (non-reactive) for observations to be modeled effectively as mixtures of end-members. SPGD-AA, like other unsupervised EMMA methods, would fail if no observations are extreme enough to represent true end-members. We recommend validating SPGD-AA results with domain expertise to ensure inferred end-members are meaningful.

SPGD-AA can complement supervised EMMA by identifying overlooked end-members in prior knowledge and guiding targeted field sampling to further constrain end-member characteristics. The concepts of RSS and PGD can be adapted to supervised EMMA to improve computational accuracy and assess model validity. Future work could focus on improving optimization robustness (e.g., using stochastic gradient descent to avoid local minima), developing AA variants like kernel-AA (e.g., Abrol & Sharma, 2020; Javadi & Montanari, 2020; Mørup & Hansen, 2012), and exploring the uncertainty of inferred end-members as in Xu Fei and Harman (2022). It is essential to unify terminology and establish common cyberinfrastructure across domains, to facilitate interdisciplinary collaboration and maximize SPGD-AA's potential to address geoscience questions.

Data Availability Statement

All data and scripts for performing SPGD-AA, along with end-members inferred by SPGD-AA are available on GitHub (https://github.com/WEN-Research-Group/endmembers) and Zenodo (Wang, 2025). The source code of SPGD-AA (in the `archetypes` Python package) is available either on Github (https://github.com/aleixalcacer/archetypes) or Zenodo (Alcacer & Wang, 2025).

Acknowledgments
This material is based upon work supported by the U.S. National Science Foundation under Grant OAC-2209864 to TW. We also would like to acknowledge Dr. Aleix Alcacer for developing the original `archetypes` Python package, the two anonymous reviewers for their constructive comments, Dr. Jianghui Du for suggesting the Nazca sediment data set and Haejo Kim, Liyang Qin, Claire Bush, and Dr. Linda Ivany for their suggestions on the manuscript.

References

Abrol, V., & Sharma, P. (2020). A geometric approach to archetypal analysis via sparse projections. In *Proceedings of the 37th International Conference on Machine Learning* (pp. 42–51). PMLR.

Alcacer, A., Epifanio, I., & Gual-Arnau, X. (2024). Biarchetype analysis: Simultaneous learning of observations and features based on extremes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *46*, 1–12. https://doi.org/10.1109/TPAMI.2024.3400730

Alcacer, A., & Wang, Z. (2025). Wangzcl/archetypes: Zenodo 963ec9b [Software]. *Zenodo*. https://doi.org/10.5281/zenodo.15198941

Bauckhage, C., Kersting, K., Hoppe, F., & Thurau, C. (2015). Archetypal analysis as an autoencoder. In *Workshop new challenges in neural computation* (Vol. 2015, pp. 8–16). Citeseer.

Bauckhage, C., & Thurau, C. (2009). Making archetypal analysis practical. In J. Denzler, G. Notni, & H. Süße (Eds.), *Pattern recognition* (pp. 272–281). Springer. https://doi.org/10.1007/978-3-642-03798-6_28

Beck, A., & Teboulle, M. (2003). Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, *31*(3), 167–175. https://doi.org/10.1016/S0167-6377(02)00231-6

Behnel, S., Bradshaw, R., Citro, C., Dalcin, L., Seljebotn, D. S., & Smith, K. (2011). Cython: The best of both worlds. *Computing in Science and Engineering*, *13*(2), 31–39. https://doi.org/10.1109/MCSE.2010.118

Bioucas-Dias, J. M., Plaza, A., Dobigeon, N., Parente, M., Du, Q., Gader, P., & Chanussot, J. (2012). Hyperspectral unmixing overview: Geometrical, statistical, and sparse regression-based approaches. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *5*(2), 354–379. https://doi.org/10.1109/JSTARS.2012.2194696

Bishop, C. M. (2016). *Pattern recognition and machine learning* (Softcover reprint of the original 1st edition 2006 (corrected at 8th printing 2009)). Springer.

Boardman, J. W., Kruse, F. A., & Green, R. O. (1995). Mapping target signatures via partial unmixing of AVIRIS data.

Bro, R., & De Jong, S. (1997). A fast non-negativity-constrained least squares algorithm. *Journal of Chemometrics*, *11*(5), 393–401. https://doi.org/10.1002/(SICI)1099-128X(199709/10)11:5%3C393::AID-CEM483%3E3.0.CO;2-L

Burns, D. A., McDonnell, J. J., Hooper, R. P., Peters, N. E., Freer, J. E., Kendall, C., & Beven, K. (2001). Quantifying contributions to storm runoff through end-member mixing analysis and hydrologic measurements at the Panola Mountain Research Watershed (Georgia, USA). *Hydrological Processes*, *15*(10), 1903–1924. https://doi.org/10.1002/hyp.246

Canhasi, E., & Kononenko, I. (2014). Weighted archetypal analysis of the multi-element graph for query-focused multi-document summarization. *Expert Systems with Applications*, *41*(2), 535–543. https://doi.org/10.1016/j.eswa.2013.07.079

Chan, B. H. P., Mitchell, D. A., & Cram, L. E. (2003). Archetypal analysis of galaxy spectra. *Monthly Notices of the Royal Astronomical Society*, *338*(3), 790–795. https://doi.org/10.1046/j.1365-8711.2003.06099.x

Chen, Y., Mairal, J., & Harchaoui, Z. (2014). Fast and robust archetypal analysis for representation learning. In *2014 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1478–1485). https://doi.org/10.1109/CVPR.2014.192

Christophersen, N., & Hooper, R. P. (1992). Multivariate analysis of stream water chemical data: The use of principal components analysis for the end-member mixing problem. *Water Resources Research*, *28*(1), 99–107. https://doi.org/10.1029/91WR02518

Christophersen, N., Neal, C., Hooper, R. P., Vogt, R. D., & Andersen, S. (1990). Modelling streamwater chemistry as a mixture of soilwater end-members—A step towards second-generation acidification models. *Journal of Hydrology*, *116*(1), 307–320. https://doi.org/10.1016/0022-1694(90)90130-P

Condat, L. (2016). Fast projection onto the simplex and the l_1 ball. *Mathematical Programming*, *158*(1), 575–585. https://doi.org/10.1007/s10107-015-0946-6

Cutler, A., & Breiman, L. (1994). Archetypal analysis. *Technometrics*, *36*(4), 338–347. https://doi.org/10.1080/00401706.1994.10485840

Damle, A., & Sun, Y. (2017). A geometric approach to archetypal analysis and nonnegative matrix factorization. *Technometrics*, *59*(3), 361–370. https://doi.org/10.1080/00401706.2016.1247017

Dietze, E., & Dietze, M. (2019). Grain-size distribution unmixing using the R package EMMAgeo. *E&G Quaternary Science Journal*, *68*(1), 29–46. https://doi.org/10.5194/egqsj-68-29-2019

Dietze, E., Hartmann, K., Diekmann, B., IJmker, J., Lehmkuhl, F., Opitz, S., et al. (2012). An end-member algorithm for deciphering modern detrital processes from lake sediments of Lake Donggi Cona, NE Tibetan Plateau, China. *Sedimentary Geology*, *243–244*, 169–180. https://doi.org/10.1016/j.sedgeo.2011.09.014

Dietze, M., Schulte, P., & Dietze, E. (2022). Application of end-member modelling to grain-size data: Constraints and limitations. *Sedimentology*, *69*(2), 845–863. https://doi.org/10.1111/sed.12929

Dijk, D. V., Burkhardt, D. B., Amodio, M., Tong, A., Wolf, G., & Krishnaswamy, S. (2019). Finding archetypal spaces using neural networks. In *2019 IEEE International Conference on Big Data (Big Data)* (pp. 2634–2643). IEEE. https://doi.org/10.1109/BigData47090.2019.9006484

Dymond, J. (1981). Geochemistry of nazca plate surface sediments: An evaluation of hydrothermal, biogenic, detrital, and hydrogenous sources. In L. V. D. Kulm, J. Dymond, E. J. Dasch, D. M. Hussong, & R. Roderick (Eds.), *Nazca plate: Crustal formation and Andean convergence*. Geological Society of America. https://doi.org/10.1130/MEM154-p133

Epifanio, I., Vinué, G., & Alemany, S. (2013). Archetypal analysis: Contributions for estimating boundary cases in multivariate accommodation problem. *Computers and Industrial Engineering*, *64*(3), 757–765. https://doi.org/10.1016/j.cie.2012.12.011

Epuna, F., Shaheen, S. W., & Wen, T. (2022). Road salting and natural brine migration revealed as major sources of groundwater contamination across regions of northern Appalachia with and without unconventional oil and gas development. *Water Research*, *225*, 119128. https://doi.org/10.1016/j.watres.2022.119128

Eugster, M. J. A., & Leisch, F. (2009). From spider-man to hero—Archetypal analysis in R. *Journal of Statistical Software*, *30*(8), 1–23. https://doi.org/10.18637/jss.v030.i08

Eugster, M. J. A., & Leisch, F. (2011). Weighted and robust archetypal analysis. *Computational Statistics and Data Analysis*, *55*(3), 1215–1225. https://doi.org/10.1016/j.csda.2010.10.017

Gaillardet, J., Dupré, B., Louvat, P., & Allègre, C. (1999). Global silicate weathering and CO_2 consumption rates deduced from the chemistry of large rivers. *Chemical Geology*, *159*(1–4), 3–30. https://doi.org/10.1016/S0009-2541(99)00031-5

Hart, Y., Sheftel, H., Hausser, J., Szekely, P., Ben-Moshe, N. B., Korem, Y., et al. (2015). Inferring biological tasks using Pareto analysis of high-dimensional data. *Nature Methods*, *12*(3), 233–235. https://doi.org/10.1038/nmeth.3254

Heinz, D., & Chang, C.-I. (2001). Fully constrained least squares linear spectral mixture analysis method for material quantification in hyperspectral imagery. *IEEE Transactions on Geoscience and Remote Sensing*, *39*(3), 529–545. https://doi.org/10.1109/36.911111

Hooper, R. P. (2001). Applying the scientific method to small catchment studies: A review of the Panola Mountain experience. *Hydrological Processes*, *15*(10), 2039–2050. https://doi.org/10.1002/hyp.255

Hooper, R. P., & Christophersen, N. (1992). Predicting episodic stream acidification in the southeastern United States: Combining a long-term acidification model and the end-member mixing concept. *Water Resources Research*, *28*(7), 1983–1990. https://doi.org/10.1029/92WR00706

Hooper, R. P., Christophersen, N., & Peters, N. E. (1990). Modelling streamwater chemistry as a mixture of soilwater end-members—An application to the Panola Mountain catchment, Georgia, U.S.A. *Journal of Hydrology*, *116*(1), 321–343. https://doi.org/10.1016/0022-1694(90)90131-G

Hoyer, P. (2002). Non-negative sparse coding. In *Proceedings of the 12th IEEE Workshop on Neural Networks for Signal Processing* (pp. 557–565). https://doi.org/10.1109/NNSP.2002.1030067

Iordache, M.-D., Bioucas-Dias, J. M., & Plaza, A. (2011). Sparse unmixing of hyperspectral data. *IEEE Transactions on Geoscience and Remote Sensing*, *49*(6), 2014–2039. https://doi.org/10.1109/TGRS.2010.2098413

Javadi, H., & Montanari, A. (2020). Nonnegative matrix factorization via archetypal analysis. *Journal of the American Statistical Association*, *115*(530), 896–907. https://doi.org/10.1080/01621459.2019.1594832

Jia, S., & Qian, Y. (2009). Constrained nonnegative matrix factorization for hyperspectral unmixing. *IEEE Transactions on Geoscience and Remote Sensing*, *47*(1), 161–173. https://doi.org/10.1109/TGRS.2008.2002882

Keller, S. M., Samarin, M., Wieser, M., & Roth, V. (2019). Deep archetypal analysis. In G. A. Fink, S. Frintrop, & X. Jiang (Eds.), *Pattern recognition* (pp. 171–185). Springer International Publishing. https://doi.org/10.1007/978-3-030-33676-9_12

Keshava, N., Kerekes, J. P., Manolakis, D. G., & Shaw, G. A. (2000). Algorithm taxonomy for hyperspectral unmixing. In *Algorithms for Multispectral, Hyperspectral, and Ultraspectral Imagery VI* (Vol. 4049, pp. 42–63). SPIE Proceedings. https://doi.org/10.1117/12.410362

Khajehrayeni, F., & Ghassemian, H. (2020). Hyperspectral unmixing using deep convolutional autoencoders in a supervised scenario. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *13*, 567–576. https://doi.org/10.1109/JSTARS.2020.2966512

Lawson, C. L., & Hanson, R. J. (1995). *Solving least squares problems* (Vol. 15). SIAM. https://doi.org/10.1137/1.9781611971217

Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, *401*(6755), 788–791. https://doi.org/10.1038/44565

Leinen, M., & Pisias, N. (1984). An objective technique for determining end-member compositions and for partitioning sediments according to their sources. *Geochimica et Cosmochimica Acta*, *48*(1), 47–62. https://doi.org/10.1016/0016-7037(84)90348-X

Li, Y., & Tan, T. (2024). A new convex model for linear hyperspectral unmixing. *Journal of Computational and Applied Mathematics*, *441*, 115708. https://doi.org/10.1016/j.cam.2023.115708

Liu, Y., Wang, T., Wen, T., Zhang, J., Liu, B., Li, Y., et al. (2024). Deep learning-based grain-size decomposition model: A feasible solution for dealing with methodological uncertainty. *Sedimentology*, *71*(6), 1873–1894. https://doi.org/10.1111/sed.13195

Lu, X., Wu, H., & Yuan, Y. (2014). Double constrained NMF for hyperspectral unmixing. *IEEE Transactions on Geoscience and Remote Sensing*, *52*(5), 2746–2758. https://doi.org/10.1109/TGRS.2013.2265322

Lu, X., Wu, H., Yuan, Y., Yan, P., & Li, X. (2013). Manifold regularized sparse NMF for hyperspectral unmixing. *IEEE Transactions on Geoscience and Remote Sensing*, *51*(5), 2815–2826. https://doi.org/10.1109/TGRS.2012.2213825

Miao, L., & Qi, H. (2007). Endmember extraction from highly mixed data using minimum volume constrained nonnegative matrix factorization. *IEEE Transactions on Geoscience and Remote Sensing*, *45*(3), 765–777. https://doi.org/10.1109/TGRS.2006.888466

Miesch, A. T. (1976). *Q*-mode factor analysis of compositional data. *Computers and Geosciences*, *1*(3), 147–159. https://doi.org/10.1016/0098-3004(76)90003-0

Mørup, M., & Hansen, L. K. (2012). Archetypal analysis for machine learning and data mining. *Neurocomputing*, *80*, 54–63. https://doi.org/10.1016/j.neucom.2011.06.033

Ozkan, S., Kaya, B., & Akar, G. B. (2019). Endnet: Sparse autoencoder network for endmember extraction and hyperspectral unmixing. *IEEE Transactions on Geoscience and Remote Sensing*, *57*(1), 482–496. https://doi.org/10.1109/TGRS.2018.2856929

Pisias, N. G., Murray, R. W., & Scudder, R. P. (2013). Multivariate statistical analysis and partitioning of sedimentary geochemical data sets: General principles and specific MATLAB scripts. *Geochemistry, Geophysics, Geosystems*, *14*(10), 4015–4020. https://doi.org/10.1002/ggge.20247

Plaza, A., Martinez, P., Perez, R., & Plaza, J. (2004). A quantitative and comparative analysis of endmember extraction algorithms from hyperspectral data. *IEEE Transactions on Geoscience and Remote Sensing*, *42*(3), 650–663. https://doi.org/10.1109/TGRS.2003.820314

Prins, M. A., & Weltje, G. J. (1999). End-member modeling of siliciclastic grain-size distributions: The late quaternary record of eolian and fluvial sediment supply to the Arabian Sea and its paleoclimatic significance. In J. W. Harbaugh, W. L. Watney, E. C. Rankey, R. Slingerland, R. H. Goldstein, & E. K. Franseen (Eds.), *Numerical experiments in stratigraphy: Recent advances in stratigraphic and sedimentologic computer simulations* (Vol. 62). SEPM Society for Sedimentary Geology. https://doi.org/10.2110/pec.99.62.0091

Qi, L., Chen, Z., Gao, F., Dong, J., Gao, X., & Du, Q. (2023). Multiview spatial–spectral two-stream network for hyperspectral image unmixing. *IEEE Transactions on Geoscience and Remote Sensing*, *61*, 1–16. https://doi.org/10.1109/TGRS.2023.3237556

Qian, Y., Jia, S., Zhou, J., & Robles-Kelly, A. (2011). Hyperspectral unmixing via $L_{1/2}$ sparsity-constrained nonnegative matrix factorization. *IEEE Transactions on Geoscience and Remote Sensing*, *49*(11), 4282–4297. https://doi.org/10.1109/TGRS.2011.2144605

Rea, D. K., & Hovan, S. A. (1995). Grain size distribution and depositional processes of the mineral component of abyssal sediments: Lessons from the North Pacific. *Paleoceanography*, *10*(2), 251–258. https://doi.org/10.1029/94PA03355

Renner, R. M. (1993). The resolution of a compositional data set into mixtures of fixed source compositions. *Journal of the Royal Statistical Society Series C: Applied Statistics*, *42*(4), 615–631. https://doi.org/10.2307/2986179

Seth, S., & Eugster, M. J. A. (2016). Probabilistic archetypal analysis. *Machine Learning*, *102*(1), 85–113. https://doi.org/10.1007/s10994-015-5498-8

Shaughnessy, A. R., Gu, X., Wen, T., & Brantley, S. L. (2021). Machine learning deciphers co_2 sequestration and subsurface flowpaths from stream chemistry. *Hydrology and Earth System Sciences*, *25*(6), 3397–3409. https://doi.org/10.5194/hess-25-3397-2021

Shaheen, S. W., Wen, T., Herman, A., & Brantley, S. L. (2022). Geochemical evidence of potential groundwater contamination with human health risks where hydraulic fracturing overlaps with extensive legacy hydrocarbon extraction. *Environmental Science & Technology*, *56*(14), 10010–10019. https://doi.org/10.1021/acs.est.2c00001

Steinschneider, S., & Lall, U. (2015). Daily precipitation and tropical moisture exports across the eastern United States: An application of archetypal analysis to identify spatiotemporal structure. https://doi.org/10.1175/JCLI-D-15-0340.1

Thurau, C., Kersting, K., Wahabzada, M., & Bauckhage, C. (2011). Convex non-negative matrix factorization for massive datasets. *Knowledge and Information Systems*, *29*(2), 457–478. https://doi.org/10.1007/s10115-010-0352-6

Vandenberghe, J. (2013). Grain size of fine-grained windblown sediment: A powerful proxy for process identification. *Earth-Science Reviews*, *121*, 18–30. https://doi.org/10.1016/j.earscirev.2013.03.001

van Geen, A., Rosener, P., & Boyle, E. (1988). Entrainment of trace-metal-enriched Atlantic-shelf water in the inflow to the Mediterranean Sea. *Nature*, *331*(6155), 423–426. https://doi.org/10.1038/331423a0

Wang, Z. (2025). *Wen-research-group/endmembers: V0.0.0-alpha [collection]*. Zenodo. https://doi.org/10.5281/zenodo.15178390

Wei, J., & Wang, X. (2020). An overview on linear unmixing of hyperspectral data. *Mathematical Problems in Engineering*, *2020*, e3735403–e3735412. https://doi.org/10.1155/2020/3735403

Weltje, G. J. (1997). End-member modeling of compositional data: Numerical-statistical algorithms for solving the explicit mixing problem. *Mathematical Geology*, *29*(4), 503–549. https://doi.org/10.1007/BF02775085

Weltje, G. J., & Prins, M. A. (2007). Genetically meaningful decomposition of grain-size distributions. *Sedimentary Geology*, *202*(3), 409–424. https://doi.org/10.1016/j.sedgeo.2007.03.007

Winter, M. E. (1999). N-FINDR: An algorithm for fast autonomous spectral end-member determination in hyperspectral data. In *Imaging Spectrometry V* (Vol. 3753, pp. 266–275). SPIE Proceedings. https://doi.org/10.1117/12.366289

Wright, S. J., & Recht, B. (2022). *Optimization for data analysis*. Cambridge University Press. https://doi.org/10.1017/9781009004282

Wyatt, M. B., & McSween, H. Y. (2002). Spectral evidence for weathered basalt as an alternative to andesite in the northern lowlands of Mars. *Nature*, *417*(6886), 263–266. https://doi.org/10.1038/417263a

Xiong, F., Zhou, J., Tao, S., Lu, J., & Qian, Y. (2022). SNMF-Net: Learning a deep alternating neural network for hyperspectral unmixing. *IEEE Transactions on Geoscience and Remote Sensing*, *60*, 1–16. https://doi.org/10.1109/TGRS.2021.3081177

Xu, X., Li, J., Li, S., & Plaza, A. (2020). Curvelet transform domain-based sparse nonnegative matrix factorization for hyperspectral unmixing. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *13*, 4908–4924. https://doi.org/10.1109/JSTARS.2020.3017023

Xu, X., Shi, Z., & Pan, B. (2018). A supervised abundance estimation method for hyperspectral unmixing. *Remote Sensing Letters*, *9*(4), 383–392. https://doi.org/10.1080/2150704X.2017.1415471

Xu Fei, E., & Harman, C. J. (2022). A data-driven method for estimating the composition of end-members from stream water chemistry time series. *Hydrology and Earth System Sciences*, *26*(8), 1977–1991. https://doi.org/10.5194/hess-26-1977-2022

Zhang, X., Sun, Y., Zhang, J., Wu, P., & Jiao, L. (2018). Hyperspectral unmixing via deep convolutional neural networks. *IEEE Geoscience and Remote Sensing Letters*, *15*(11), 1755–1759. https://doi.org/10.1109/LGRS.2018.2857804

Zhang, X., Wang, H., Xu, S., & Yang, Z. (2020). A basic end-member model algorithm for grain-size data of marine sediments. *Estuarine, Coastal and Shelf Science*, *236*, 106656. https://doi.org/10.1016/j.ecss.2020.106656

Zhou, L., Zhang, X., Wang, J., Bai, X., Tong, L., Zhang, L., et al. (2020). Subspace structure regularized nonnegative matrix factorization for hyperspectral unmixing. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *13*, 4257–4270. https://doi.org/10.1109/JSTARS.2020.3011257

Zhu, F. (2017). Hyperspectral unmixing: Ground truth labeling, datasets, benchmark performances and survey. *arXiv*. https://doi.org/10.48550/ARXIV.1708.05125

Zhu, F., Wang, Y., Xiang, S., Fan, B., & Pan, C. (2014). Structured sparse method for hyperspectral unmixing. *ISPRS Journal of Photogrammetry and Remote Sensing*, *88*, 101–118. https://doi.org/10.1016/j.isprsjprs.2013.11.014

Zhuang, L., Lin, C.-H., Figueiredo, M. A. T., & Bioucas-Dias, J. M. (2019). Regularization parameter selection in minimum volume hyperspectral unmixing. *IEEE Transactions on Geoscience and Remote Sensing*, *57*(12), 9858–9877. https://doi.org/10.1109/TGRS.2019.2929776

Zouaoui, A., Muhawenayo, G., Rasti, B., Chanussot, J., & Mairal, J. (2023). Entropic descent archetypal analysis for blind hyperspectral unmixing. *IEEE Transactions on Image Processing*, *32*, 4649–4663. https://doi.org/10.1109/TIP.2023.3301769